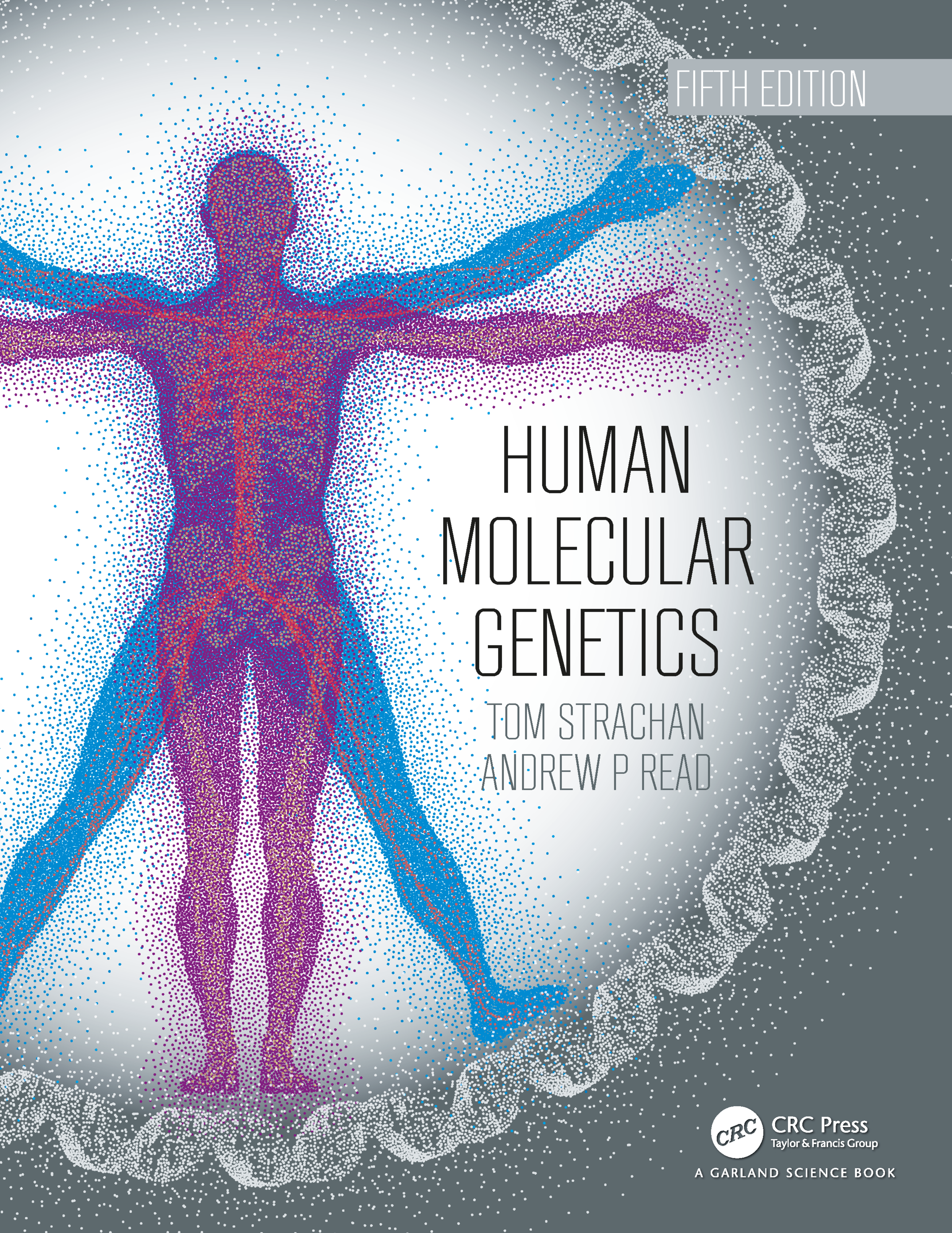


FIFTH EDITION

HUMAN MOLECULAR GENETICS

TOM STRACHAN
ANDREW P READ

CRC Press
Taylor & Francis Group
A GARLAND SCIENCE BOOK

FIFTH EDITION

HUMAN MOLECULAR GENETICS

FIFTH EDITION

HUMAN MOLECULAR GENETICS

TOM STRACHAN
BSc, PhD, FRSE, FMedSci
Emeritus Professor of Human Molecular Genetics
Newcastle University, Newcastle, UK

ANDREW P READ
MA, PhD, FRCPath, FMedSci
Emeritus Professor of Human Genetics
University of Manchester, Manchester, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A GARLAND SCIENCE BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-0-815-34589-3 (Paperback)
978-0-367-00250-3 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

PART 2 UNDERSTANDING GENOMES

PART 3 GENETIC VARIATION BETWEEN INDIVIDUALS AND SPECIES

PART 5 APPLIED HUMAN MOLECULAR GENETICS

Preface

Much has changed since the fourth edition of *Human Molecular Genetics* (HMG4) appeared in 2011, so this fifth edition has seen a comprehensive rewrite and reorganization. Few of the chapters retain their identity from HMG4, but our aims throughout the book remain the same: to provide a framework of principles rather than to list facts (which are better found in online resources), to provide a bridge between basic textbooks and the research literature, and to communicate our excitement and enthusiasm for this very fast-moving area of science.

The biggest single development since HMG4 has been the massive expansion of DNA sequencing in every area of human genetics. In response, we have provided a much-extended and updated coverage of massively-parallel sequencing technology, including the exciting new field of single-cell genomics. In some respects, the sequencing revolution has made things simpler. Many specific techniques covered in HMG4 have been largely or completely superseded by sequencing. The reader will note, however, that we still illustrate karyotypes. We make no apology for showing these where appropriate because they have educational value—it is often easier to understand what is going on by looking at a karyotype rather than sequence data, even though nowadays most laboratories would use sequencing rather than microscopy for these purposes.

In the preface to HMG4 we wrote, "we can confidently expect that the genomes of huge numbers of organisms and individuals will have been completed before the next edition of this book," and that expectation has been amply realized. Human genetics is now firmly in the world of Big Data and big international collaborations—and our coverage reflects this.

Among new or rearranged chapters we were particularly gratified when Mark Jobling and the team who produced the excellent *Human Evolutionary Genetics* (2nd edn, Garland Science, 2013) agreed to contribute a chapter on human evolution. Analysis of contemporary and ancient DNA has progressed enormously in the past few years and is revealing fascinating insights into our origins and history—but neither of us felt qualified to write with sufficient authority on this important topic.

Other developments include:

- A radical revision of coverage of early mammalian development and stem cells, with a detailed explanation of the origins of cellular differentiation and an in-depth survey of pluripotent stem cells as well as tissue stem cells and cell reprogramming.
- A specific chapter on genetic manipulation of mammalian cells, bringing together material from different HMG4 chapters and tracing the evolution of genome editing from a focus on simply using homologous recombination to the modern emphasis on using programmable nucleases.
- A chapter that deals with both the architecture of the human genome and also the ENCODE Project and other new initiatives to understand how our genome functions.
- A chapter giving unified coverage of gene regulation and epigenetics.
- A chapter giving an overview of human genetic variation that includes the origins of DNA sequence variation, DNA repair mechanisms, variant classes, population genomics, and functional genetic variation.

- A new chapter on human population genetics—a topic that we felt received inadequate coverage in previous editions.
- A specific chapter on molecular pathology, bringing together and expanding material from HMG4.
- Greatly expanded discussion of the achievements and limitations of genome-wide association studies (GWAS) in identifying susceptibility factors for common complex conditions. As the GWAS era is giving way to large-scale sequencing approaches, this seems a particularly apposite time for a critical analysis.
- New coverage of DNA diagnostics reflecting the major changes that have come with the routine use of whole exome and whole genome sequencing.
- Revised discussion of cancer genetics and genomics reflecting developments in multiplatform analyses, liquid biopsies, and targeted treatments.
- A new chapter that brings together model organisms and disease modeling, including the fast-moving new field of cellular disease modeling, especially organoid models that arose from basic developmental studies.

Apart from these specific topics, every page of the text has been revised and updated to provide an overview of human molecular genetics in 2018.

This book has only been possible because of the work of the team under Joanna Koster at Taylor & Francis who have converted our drafts and sketches into the finished product—Paul Bennett, Jordan Wearing, Matt McClements, Ruth Maxwell, Becky Hainz-Baxter, and probably others who have worked from time to time on the project. As ever, we are deeply grateful to our wives, Meryl and Gilly, for their forbearance and support during the long gestation of this book.

Tom Strachan
Andrew P Read

About the authors

Tom Strachan is Emeritus Professor of Human Molecular Genetics at Newcastle University, Newcastle, UK, and is a Fellow of the Royal Society of Edinburgh and a Fellow of the Academy of Medical Sciences. He was the founding Head of Institute at Newcastle University's Institute of Human Genetics (now the Institute of Genetic Medicine) and its Scientific Director from 2001 to 2009. Tom's early research interests were in multigene family evolution and interlocus sequence exchange, notably in the HLA and 21-hydroxylase gene clusters. While pursuing the latter, he became interested in medical genetics. His most recent research has focused on certain developmental disorders and developmental control genes.

Andrew Read is Emeritus Professor of Human Genetics at Manchester University, Manchester, UK, and a Fellow of the Academy of Medical Sciences. Andrew has been particularly concerned with making the benefits of DNA technology available to people with genetic problems. He established one of the first DNA diagnostic laboratories in the UK over 35 years ago (which became one of two National Genetics Reference Laboratories), and was founder chairman of the British Society for Human Genetics (now the British Society for Genetic Medicine), the main UK professional body in this area. His own research is on the molecular pathology of various hereditary syndromes, especially hereditary hearing loss.

Tom Strachan and **Andrew Read** were recipients of the European Society of Human Genetics Education Award in 2007.

Contributors

The authors are grateful to the following who contributed Chapter 14, *Human evolution*, in this edition:

Mark A Jobling DPhil

Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

Edward J Hollox PhD

Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

Toomas Kivisild PhD

Department of Human Genetics, KU Leuven, Leuven, Belgium

Chris Tyler-Smith PhD

Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

BASICS OF DNA, CHROMOSOMES, CELLS, DEVELOPMENT AND INHERITANCE

PART ONE

1 Basic principles of nucleic acid structure and gene expression

Molecular genetics is largely defined by the interplay between three classes of macromolecule: the nucleic acid molecules, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), and proteins. In organisms and cells, DNA is the genetic (hereditary) material that is transmitted to daughter cells when cells replicate, and from one generation to the next when organisms reproduce. Viruses also have genetic material that is transmitted to viral progeny; according to the type of virus, the genetic material may be DNA or RNA. The term **genome** is the collective name for the set of different DNA molecules in an organism, cell, or DNA virus, or of RNA molecules in an RNA virus. All proteins have a polypeptide core that is synthesized using genetic information within DNA molecules (or within the hereditary RNA molecules of an RNA virus).

RNA may have been the hereditary material at a very early stage of evolution, but now, except in certain viruses, it no longer serves this role. Instead, the genetic information in cells came to be stored in DNA molecules (which are more chemically stable than RNA and can be copied accurately and transmitted to daughter cells, and from one generation to the next). In eukaryotes, DNA molecules are found mainly in the chromosomes of the nucleus, but the mitochondria of all eukaryotic cells also have small DNA molecules, as do the chloroplasts of plant cells.

Genes are segments of hereditary DNA or RNA molecules that are used to make one or both of two types of functional end product: a polypeptide or a mature functional RNA. Both types of product are then subject to processing reactions to make a working molecule. For example, a polypeptide may be subject to cleavage and/or to minor chemical changes to its constituent components, and may often also be complexed with other molecules including carbohydrates, lipids, or other polypeptides in order to make a working protein.

In simple organisms, the DNA is packed with genes (bacteria typically have from several hundred up to a few thousand different genes packed within 1–10 Mb [megabases] of DNA). In the more complex cells of eukaryotes, the genes are usually much more sparsely distributed within the DNA, and in complex multicellular eukaroytes much of the DNA consists of highly-repetitive sequences (whose functions are often not so readily easily identified).

There are many different types of RNA molecule, but according to their function they can be divided into two broad classes. A **coding RNA** sequence, popularly called a **messenger RNA** (**mRNA**), carries genetic information from DNA to the protein synthesis machinery. Messenger RNA made in the nucleus needs to be exported to the cytoplasm to make proteins, but the messenger RNA synthesized in mitochondria and chloroplasts is used to make certain proteins within these organelles.

Mature **noncoding RNA** sequences are the second broad class of RNA. They are not used as a template to make polypeptides. Instead, they often assist the expression of other genes, sometimes acting in a fairly general way and sometimes by regulating the expression of a small set of target genes. Because most gene expression is ultimately dedicated to making polypeptides, either directly or by regulating how they are produced, proteins represent the major functional endpoint of the information stored in DNA.

The central dogma of molecular biology

Genetic information generally flows in a one-way direction: DNA is decoded to make RNA, and then coding RNA (messenger RNA) is used to make polypeptides that subsequently form proteins. Because of its universality, this flow of genetic information has been described as the central dogma of molecular biology. Two sequential processes are essential in all cellular organisms:

1. Transcription, by which a sequence of bases on a DNA strand is used as a template by an RNA polymerase to synthesize an RNA; the RNA product is processed to make a messenger RNA (coding RNA) or noncoding RNA;
2. Translation, by which a messenger RNA is decoded to make polypeptides at ribosomes, large RNA–protein complexes found in the cytoplasm and also in mitochondria and chloroplasts.

Genetic information is encoded in the linear sequence of nucleotides in DNA. That information is copied during transcription to specify a linear sequence of nucleotides in the RNA product. In the case of a coding RNA, groups of three nucleotides at a time (codons) are read in a linear sequence to specify a linear sequence of amino acids in the polypeptide product.

The central dogma is now recognized to be not strictly valid. A class of RNA virus known as retroviruses provided the first evidence. These viruses have an RNA genome with a gene that makes a reverse transcriptase, a DNA polymerase that uses an RNA template to make a DNA sequence copy. Thereafter, it became clear that cellular reverse transcriptases also exist. We now know that many DNA sequences in our cells specify reverse transcriptases to allow DNA copies to be made from different RNAs. This reverse flow of genetic information from RNA to DNA has been important in the evolution of our genome (as described in Chapter 13), and in replicating the DNA sequences at the very ends of linear chromosomes (described in Chapter 2).

1.1 COMPOSITION OF NUCLEIC ACIDS AND POLYPEPTIDES

We describe below the structure of nucleic acids and proteins. All proteins have a linear polypeptide backbone (encoded by a gene) to which carbohydrate, lipid, and small chemical groups may be added at the post-translational level. Here we describe the component units of nucleic acids and polypeptides, and the different types of chemical bonding within these macromolecules.

Nucleic acids and polypeptides are linear sequences of simple repeat units

DNA and RNA strands are large polymers that have very similar structures. Each has a linear sugar–phosphate backbone that has alternating residues of a five-carbon sugar and a phosphate, with a nitrogenous base attached to each sugar residue (**Figure 1.1A**). The sugars are ribose in RNA and deoxyribose in DNA, and they differ in either lacking or possessing an –OH group at their 2′-carbon positions (**Figure 1.1B**).

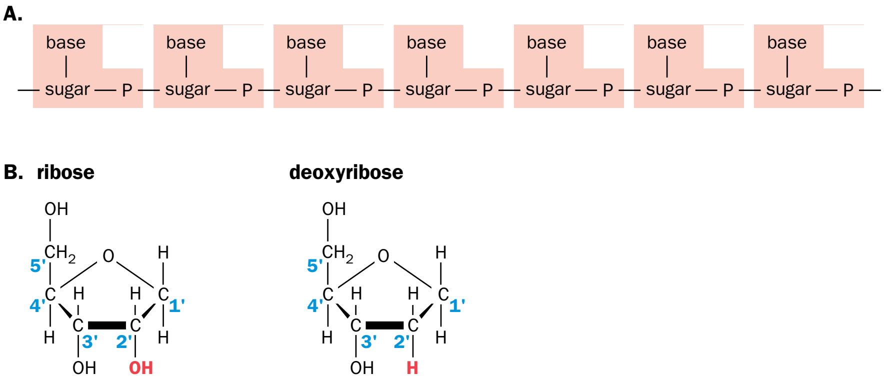


Figure 1.1 Repeat units in nucleic acids. (**A**) The linear backbone of nucleic acids consists of alternating phosphate (P) and sugar residues. Attached to each sugar is a base. The basic repeat unit (pale peach shading) consists of a base + sugar + phosphate = a nucleotide. (**B**) Ribose, the sugar in RNA, and deoxyribose, the sugar in DNA, both have five carbon atoms numbered 1′ to 5′. Deoxyribose lacks the hydroxyl (OH) group attached to carbon 2 of ribose (the proper name is 2′-deoxyribose).

Unlike the sugar and phosphate residues, the bases of a nucleic acid molecule vary, and it is the sequence of bases that identifies the nucleic acid and determines its function. The bases of a nucleic acid each consist of heterocyclic rings of carbon and nitrogen atoms and can be divided into two structural classes: **purines**, which have two interlocked rings, and **pyrimidines**, which have a single ring. In both DNA and RNA there are four principal types of base, two purines and two pyrimidines. Three types of base adenine (A), cytosine (C), and guanine (G) are common to both DNA and RNA. The fourth base is thymine (T) in DNA and the closely related uracil (U) in RNA. Uracil lacks the 5-methyl group found in thymine (Figure 1.2A).

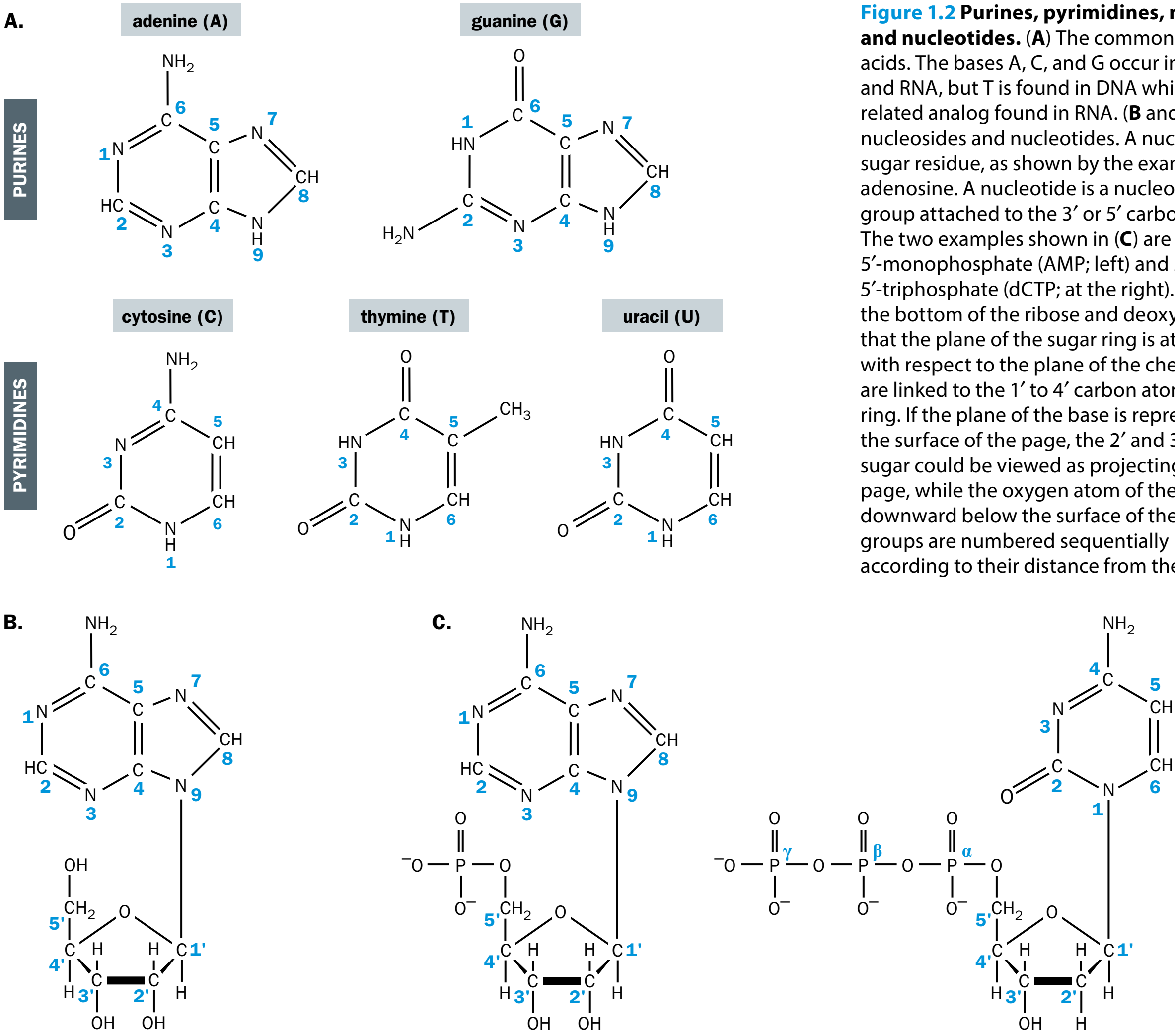


Figure 1.2 Purines, pyrimidines, nucleosides, and nucleotides. (**A**) The common bases in nucleic acids. The bases A, C, and G occur in both DNA and RNA, but T is found in DNA while U is a closely related analog found in RNA. (**B** and **C**) Examples of nucleosides and nucleotides. A nucleoside is a base + sugar residue, as shown by the example in (**B**), which is adenosine. A nucleotide is a nucleoside + phosphate group attached to the 3′ or 5′ carbon of the sugar. The two examples shown in (**C**) are adenosine 5′-monophosphate (AMP; left) and 2′-deoxycytidine 5′-triphosphate (dCTP; at the right). The bold lines at the bottom of the ribose and deoxyribose rings mean that the plane of the sugar ring is at an angle of 90° with respect to the plane of the chemical groups that are linked to the 1′ to 4′ carbon atoms within the ring. If the plane of the base is represented as lying on the surface of the page, the 2′ and 3′ carbons of the sugar could be viewed as projecting upward out of the page, while the oxygen atom of the sugar ring projects downward below the surface of the page. Phosphate groups are numbered sequentially (α, β, γ, etc.), according to their distance from the sugar ring.

In nucleic acids, each base is covalently attached to the sugar by an ***N*-glycosidic bond** that joins a nitrogen atom (nitrogen 1 of a pyrimidine or nitrogen 9 of a purine) to the carbon 1′ (one prime) of the sugar. A sugar with an attached base is called a nucleoside (Figure 1.2B). A nucleoside with a phosphate group attached at the 5′ or 3′ carbon of the sugar is the basic repeat unit of a DNA strand, and is called a **nucleotide** (Figure 1.2C and Table 1.1).

As described below, DNA also contains a few types of minor base produced by chemical modification, but base modification is much more common in RNA where a large variety of chemical modifications of both bases and ribose sugars are known to occur.

Polypeptides

Proteins are composed of one or more **polypeptide** chains that may be modified by the addition of carbohydrate side chains or other chemical groups. Like DNA and RNA, polypeptides are polymers that have a linear sequence of repeating units. The basic repeat unit is called an **amino acid**.

TABLE 1.1 NOMENCLATURE FOR BASES, NUCLEOSIDES, AND NUCLEOTIDES

		Nucleoside (= base + sugar)		Nucleotide (= nucleoside + phosphate)		
	Base	**Ribose**	**Deoxyribose**	**Monophosphate**	**Diphosphate**	**Triphosphate**
PURINE	Adenine	Adenosine	Deoxyadenosine	Adenosine monophosphate (AMP)[a,b] Deoxyadenosine monophosphate (dAMP)[a]	Adenosine diphosphate (ADP) Deoxyadenosine diphosphate (dADP)	Adenosine triphosphate (ATP) Deoxyadenosine triphosphate (dATP)
	Guanine	Guanosine	Deoxyguanosine	Guanosine monophosphate (GMP)[b] Deoxyguanosine monophosphate (dGMP)	Guanosine diphosphate (GDP) Deoxyguanosine diphosphate (dGDP)	Guanosine triphosphate (GTP) Deoxyguanosine triphosphate (dGTP)
PYRIMIDINE	Cytosine	Cytidine	Deoxycytidine	Cytidine monophosphate (CMP)[b] Deoxycytidine monophosphate (dCMP)	Cytidine diphosphate (CDP) Deoxycytidine diphosphate (dCDP)	Cytidine triphosphate (CTP) Deoxycytidine triphosphate (dCTP)
	Thymine	Thymidine	Deoxythymidine	Thymidine monophosphate (TMP)[b] Deoxythymidine monophosphate (dTMP)	Thymidine diphosphate (TDP) Deoxythymidine diphosphate (dTDP)	Thymidine triphosphate (TTP) Deoxythymidine triphosphate (dTTP)
	Uracil	Uridine	Deoxyuridine	Uridine monophosphate (UMP)[b] Deoxyuridine monophosphate (dUMP)	Uridine diphosphate (UDP) Deoxyuridine diphosphate (dUDP)	Uridine triphosphate (UTP) Deoxyuridine triphosphate (dUTP)

[a] Where the sugar is ribose, the nucleotide is AMP; where the sugar is deoxyribose, the nucleotide is dAMP. This pattern applies throughout the table. Note that TMP, TDP, and TTP are not normally found in cells.

[b] Nucleoside monophosphates are alternatively named as follows: AMP, adenylate; GMP, guanylate; CMP, cytidylate; TMP, thymidylate; UMP, uridylate.

Amino acids get their name because in its electrically neutral form a single unbound amino acid has an amino group (–NH_2) connected by a central α-carbon atom to a carboxyl group (–COOH). The central carbon atom also bears an identifying side chain that determines the chemical nature of the amino acid. At physiological pH, the amino group acquires a proton and becomes positively charged and the carboxyl group loses a proton and becomes negatively charged (**Figure 1.3A**). According to the type of amino acid, the side chain may or may not have a charge, as detailed below.

Polypeptides are formed by sequential condensation reactions between the amino group of one amino acid and the carboxyl group of the next amino acid to be incorporated into the polymer. As a result, a polypeptide has a repeating backbone where the amino acid residues are linked by amide groups (–CO–NH–) that are referred to as **peptide bonds** (**Figure 1.3B**), and where the side chain (generally called an R-group) can differ from one amino acid to another (**Figure 1.4**).

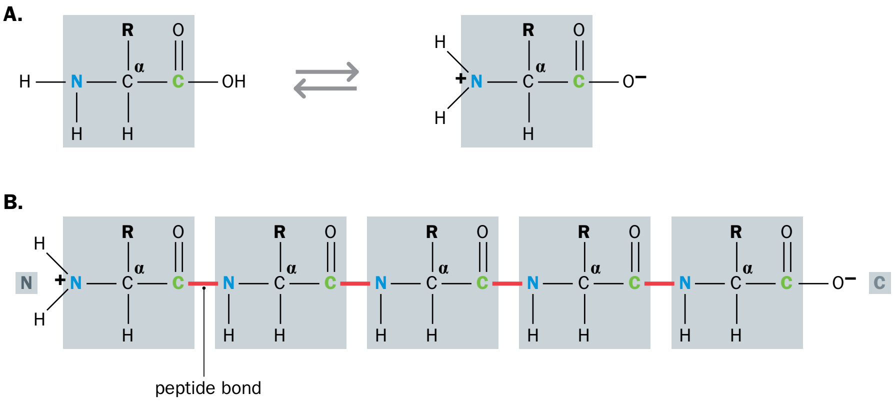


Figure 1.3 The general structure of an amino acid and a polypeptide. (**A**) Amino acid structure. At the left is the uncharged form of a generalized individual amino acid. A central α carbon is linked to three major groups: an amino group (NH_2), a carboxyl group (COOH), and a side chain R, giving the general formula $H_2N–CH(R)–COOH$. At physiological pH, as shown at the right, the end groups are ionized: the amino group acquires a positive charge and the carboxyl group acquires a negative charge. The gray shading shows an amino acid repeating unit as found in a polypeptide. (**B**) Polypeptide structure. A polypeptide forms by sequential addition of amino acid monomers in a condensation reaction involving the carboxyl group of the last amino acid to be incorporated and the amino group of the next amino acid to be incorporated. The amino acid monomers (highlighted by gray shading) are therefore connected by amide bonds (–CO–NH–), known in this context as peptide bonds. One end of the polypeptide backbone will retain the charged amino group of the original amino acid and is known as the N-terminal end; the other end has the charged carboxyl group of the last amino acid to be incorporated, and is the C-terminal end.

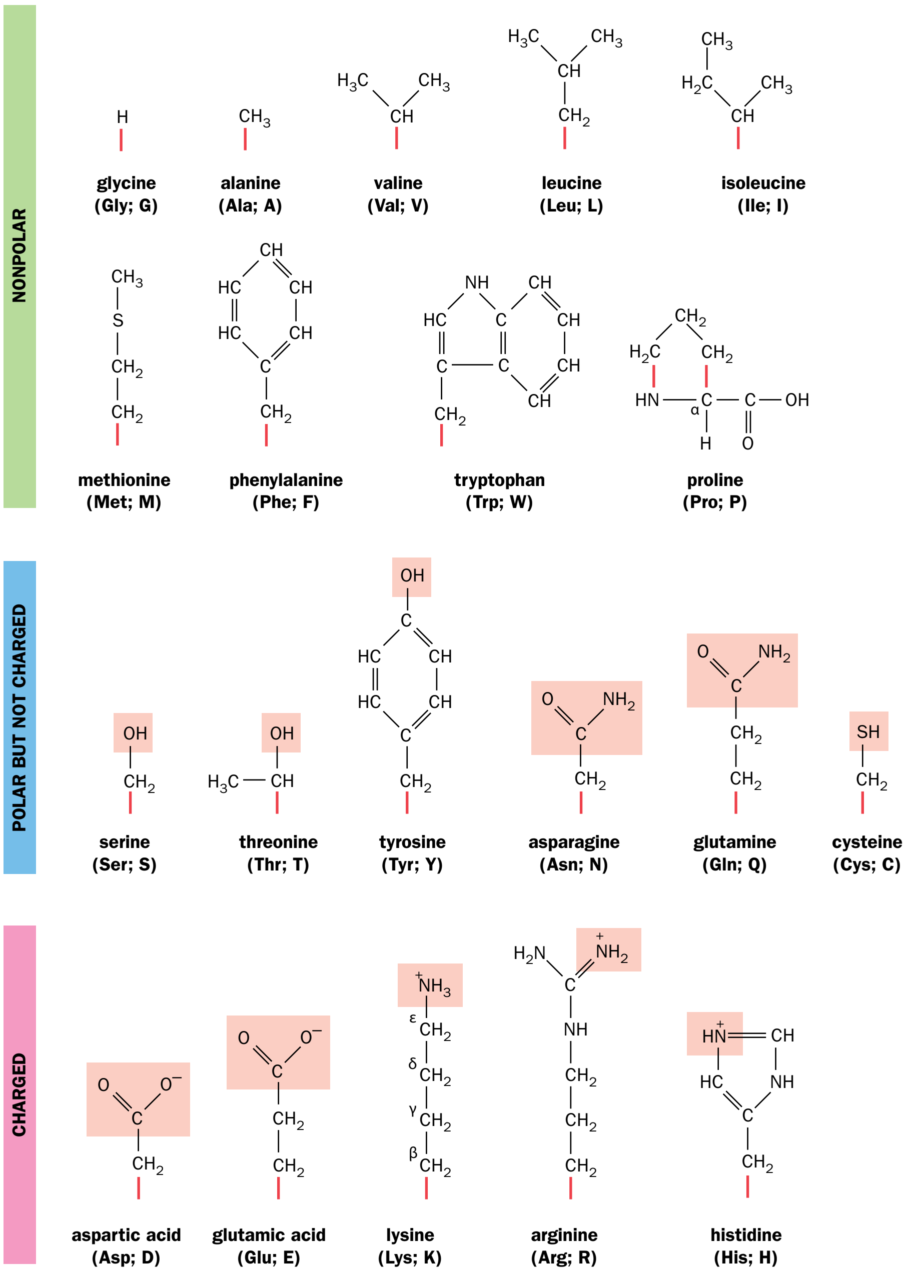


Figure 1.4 Side chains of the 20 common amino acids, grouped according to chemical class. In 19 of the 20 common amino acids the side chain is connected by a single covalent bond (red) to the α-carbon atom of the amino acid backbone; for these, we give the structure of the side chain only. Proline is the exception and we give its full structure here. Its side chain (-CH$_2$-CH$_2$-CH$_2$-) is connected to the backbone by two covalent bonds (red), with one end joined to the central α carbon atom, and the other end to the nitrogen atom of the backbone amino group. The convention for naming carbon atoms in a side chain is to use sequential Greek letters, counting out from the central α carbon atom (β, γ, δ, and so on; for example, in lysine's side chain, the carbon atom joined to the amino group, is the ε or epsilon carbon atom). Some amino acids have side chains with polar groups (pale peach shading) that may be uncharged or charged. The uncharged polar amino acids comprise three with a free hydroxyl group (serine, threonine, and tyrosine), two with amide groups (asparagine and glutamine), plus cysteine (which is only weakly polar). The charged amino acids comprise two acidic amino acids, aspartic acid (= aspartate) and glutamic acid (= glutamate), with a negatively-charged carboxyl ion on their side chain at physiological pH, plus three basic amino acids. The latter include two strongly basic amino acids, lysine and arginine, each with a positively-charged nitrogen atom in the side chain at physiological pH, plus the very weakly basic histidine. Note: at physiological pH histidines are predominantly neutral, but at low pH they can be positively charged (as shown here).

Twenty different amino acids are common in nature and can be classified into three main groups according to their side chains (see Figure 1.4). Nine amino acids have a nonpolar side chain. In most of these cases the side chain is a simple aliphatic group, but phenylalanine and tryptophan have aromatic side chains and proline has a very unusual side chain that connects the central carbon atom to the N-terminal amino group (see Figure 1.4). The nonpolar neutral amino acids are hydrophobic (repel water), often interacting with one another and with other hydrophobic groups.

Six amino acids are polar but electrically neutral overall. Their side chains carry polar groups with fractional electrical charges (often denoted as δ^+ or δ^-). Five amino acids have a charged side chain that either has a negative charge at physiological pH (acidic) or a net positive charge (basic, see Figure 1.4). In general, charged and uncharged polar amino acids are hydrophilic while nonpolar amino acids are hydrophobic. However, glycine and cysteine occupy intermediate positions on the hydrophilic–hydrophobic scale (glycine has just a single hydrogen as its side chain, and the –SH group is not so polar as an –OH group).

The amino acids of proteins often undergo chemical modification of the side chains. Quite often a very simple chemical group is added to the side chain of the amino acid, but sometimes a large carbohydrate, lipid, or even another protein is joined to the side chain, as described below.

The role of chemical bonding in the stability and function of macromolecules

The stability of nucleic acid and protein polymers is primarily dependent on strong covalent bonds between the atoms of their linear backbones. In addition to covalent bonds, weak noncovalent bonds (**Table 1.2**) are important in stabilizing molecules and in allowing a variety of transient interactions between diverse molecules within cells. Whereas covalent bonds are comparatively stable, and require a high input of energy to break them, individual noncovalent bonds are typically >10 times weaker than individual covalent bonds. As a result, they are constantly being made and broken at physiological temperatures.

TABLE 1.2 WEAK NONCOVALENT BONDS AND FORCES

Type of bond	Nature of bond
Hydrogen	Hydrogen bonds form when a hydrogen atom interacts with electron-attracting atoms, usually oxygen or nitrogen atoms
Ionic	Ionic interactions occur between charged groups. They can be very strong in crystals but in an aqueous environment the charged groups are shielded by both H_2O molecules and ions in solution and so are quite weak. Nevertheless, they can be very important in biological function, as in enzyme–substrate recognition
Van der Waals forces	Any two atoms in close proximity show a weak attractive bonding interaction due to their fluctuating electrical charges (van der Waals attraction). When atoms become extremely close, they repel each other very strongly (van der Waals repulsion). Although the forces are individually very weak, van der Waals attraction can be important when there is a very good fit between the surfaces of two macromolecules
Hydrophobic forces	Water is a polar molecule. Hydrophobic molecules or chemical groups in an aqueous environment tend to cluster. This minimizes their disruptive effects on the complex network of hydrogen bonds between water molecules. Hydrophobic groups are said to be held together by hydrophobic bonds, although the basis of their attraction is their common repulsion by water molecules

The cellular environment is an aqueous one and the structure of water is particularly complex, with a rapidly fluctuating network of noncovalent bonding occurring between water molecules. The predominant force in this structure is the **hydrogen bond**, a weak electrostatic bond between fractionally positive hydrogen atoms and fractionally negative atoms (oxygen atoms, in the case of water molecules).

Charged molecules are highly soluble in water. Because of the phosphate groups in their component nucleotides, both DNA and RNA are negatively-charged polyanions. Depending on their amino acid composition, proteins may be electrically neutral, or they may carry a net positive charge (**basic protein**) or a net negative charge (**acidic protein**). All of these molecules can form multiple interactions with the water during their solubilization. Even electrically neutral proteins are readily soluble if they contain sufficient charged or neutral polar amino acids. In contrast, membrane-bound proteins with many hydrophobic amino acids are thermodynamically more stable in a hydrophobic environment.

Although individually weak, the combined action of numerous noncovalent bonds can make large contributions to the stability of the **conformation** (structure) of macromolecules and are important for specifying their shape. We describe in the next section how hydrogen bonds between pairs of bases are essential for maintaining the structure of DNA and RNA molecules; and in the final section of this chapter we illustrate the central role of hydrogen bonding in determining the shape of diverse structural motifs in proteins, including the classic α-helices, β-sheets, and so on.

Because noncovalent bonds are fragile and able to be broken and remade easily, they also allow transient interactions between different molecules. Hydrogen bonding is especially important in allowing transient interactions between different nucleic acids, facilitating the recognition by regulatory RNAs of target sequences in other RNAs or in DNA. We provide examples in different chapters, notably when we consider gene regulation.

1.2 BASE PAIRING IN DNA AND RNA, THE DOUBLE HELIX, AND DNA REPLICATION

As described above, nucleic acids have a sugar–phosphate backbone with alternating sugar residues and phosphate groups. Neighboring sugar residues are linked by **3′–5′ phosphodiester bonds**, in which a phosphate group links the 3′ carbon atom of one sugar to the 5′ carbon atom of the next sugar in the sugar–phosphate backbone (Figure 1.5).

The genetic material of certain viruses is single-stranded DNA, but each cellular DNA species has two DNA strands (a DNA duplex). The two DNA strands are structured as a **double helix**: they curve around each other and each base on one DNA strand is non-covalently linked (by hydrogen bonding) to a laterally opposed base on the opposite DNA strand, forming a **base pair** (Figure 1.6).

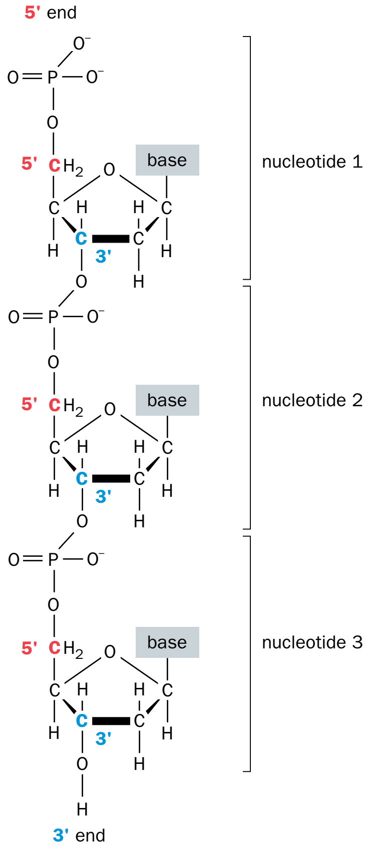


Figure 1.5 Repeating structure and asymmetric 5′ and 3′ ends of a nucleic acid strand. The repeat unit of a DNA or RNA strand is a nucleotide, consisting of a sugar with an attached base and phosphate group and, for simplicity, we show here a trinucleotide in which the 5′ carbons (red) and 3′ carbons (blue) are highlighted. There is asymmetry in how neighboring sugars are joined by the intervening phosphate group. That is, a phosphodiester bond connecting two sugars in a nucleic acid joins the carbon 3′ of one sugar to the carbon 5′ of a neighbor (a 3′–5′ phosphodiester bond). This results in asymmetry at the linear ends of the strand where the terminal nucleotides will have a sugar with either a carbon 3′ or a carbon 5′ atom that is not joined to a neighboring sugar. At the 5′ end of a nucleic acid strand the carbon 5′ of the sugar has a free phosphate group, and at the 3′ end the carbon 3′ of the sugar is attached to a hydroxyl group only.

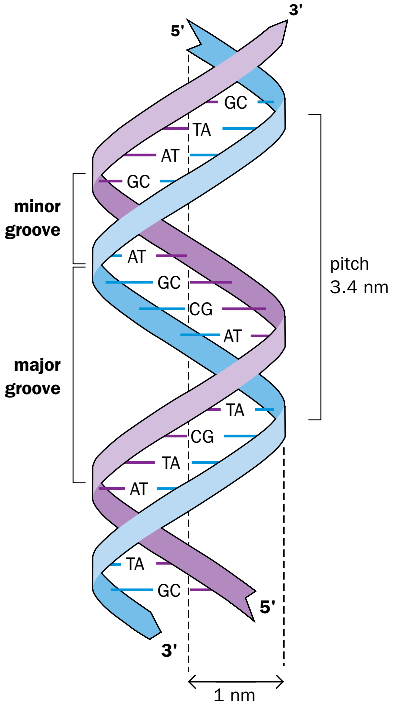


Figure 1.6 Structural features of B-DNA, the most common form of a DNA double helix. The two DNA strands of a double helix wind round each other. Under physiological conditions the B-form of a double helix is the most common form in bacterial and eukaryotic cells. It is a right-handed helix (imagine looking from one end of the helix as it spirals away from you into the distance: if the DNA strands spiral away from you in a clockwise direction you have a right-handed helix; if they spiral away in an anticlockwise direction, the helix is left-handed). B-DNA has a pitch of 3.4 nm, a radius of 1 nm per turn, and 10 base pairs per turn, and has a minor groove and a broader major groove (which facilitates access to DNA-binding proteins). See **Figure 1.13** for alternative structures for a double helix.

Base pairing involves certain purine–pyrimidine combinations only, and in double-stranded DNA the two DNA strands fit together correctly only if opposite every A on one strand is a T on the other strand, and opposite every G is a C. That is, only two types of base pair are tolerated in DNA: A-T and G-C base pairs. In standard Watson–Crick base pairing, the G-C base pairs are held together by three hydrogen bonds and are stronger than A-T base pairs, which are held together by two hydrogen bonds (**Figure 1.7**).

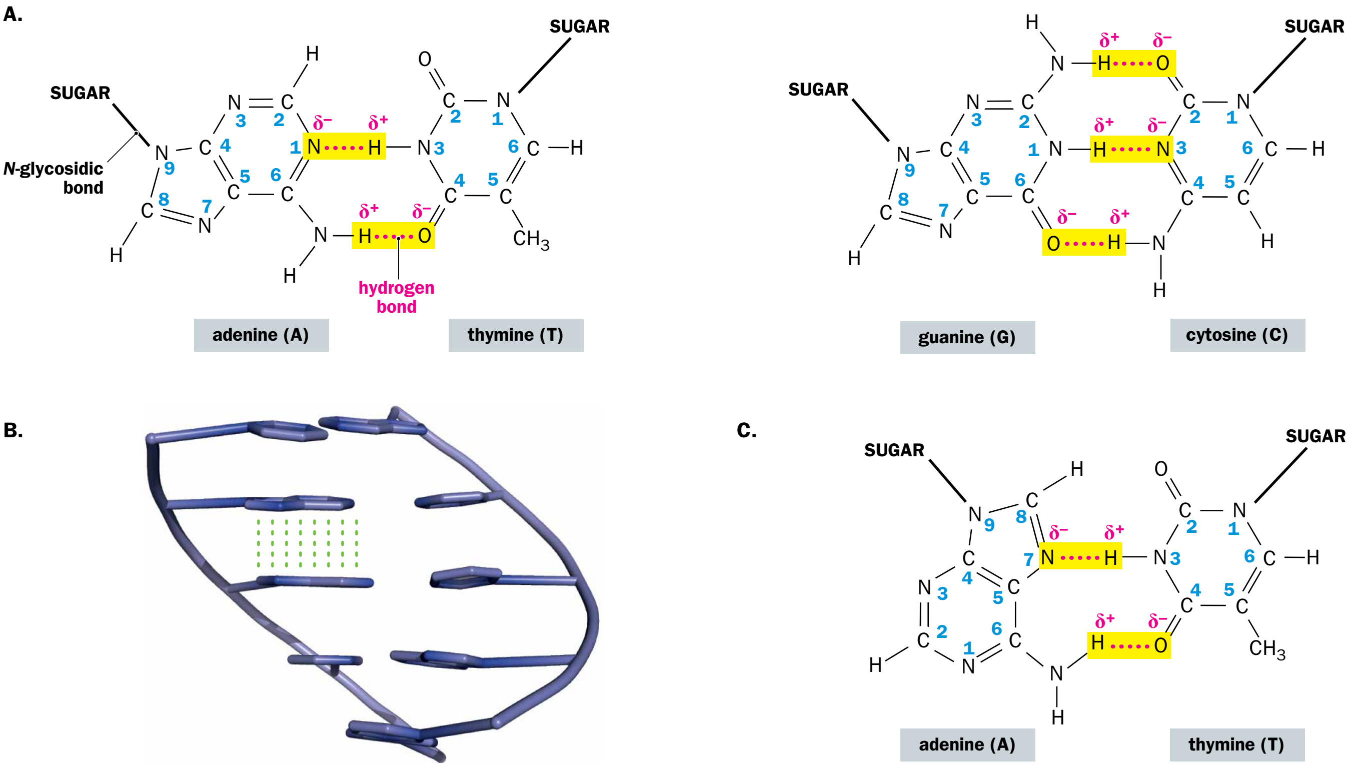


Figure 1.7 Base pairing in DNA. (**A**) Watson–Crick base pairing. A-T base pairs (left) have two connecting hydrogen bonds (dotted red lines); G-C base pairs have three hydrogen bonds. Fractional positive charges and fractional negative charges are shown by δ^+ and δ^-, respectively. (**B**) When bases pair in DNA they are arranged in the same plane, perpendicular to the long axis of the DNA helix. Van der Waals attractions between neighboring bases on each strand (shown by green dashed lines for one set of neighboring bases, as an example) are also important in the stability of the double helix. (**C**) Hoogsteen base pairing. Here a Hoogsteen A-T base pair is shown. It arises by flipping of the Watson–Crick A-T base pair via rotation of the *N*-glycosidic bond linking the sugar to nitrogen atom 9 of the adenine. As a result, the hydrogen atom attached to nitrogen atom 3 of thymine now hydrogen-bonds to nitrogen atom 7 of adenine, instead of the nitrogen atom 1 in the Watson–Crick base pair.

Because base pairing in DNA is restricted to A-T and G-C base pairs, the base composition of DNA is not random: the amount of A equals that of T, and the amount of G equals that of C. The base composition of DNA is often specified by quoting the %GC composition (= %G + %C). For example, a DNA with 42% GC has the following base composition: G, 21%; C, 21%; A, 29%; T, 29%.

In addition to base pairing, there is one other restriction on how two single-stranded nucleic acids must be arranged to form a stable duplex, and it depends on the asymmetry of the strands imposed by 3′, 5′-phosphodiester bonding. Recall from **Figure 1.5** that a nucleic acid strand has asymmetric ends: a **5′ end** that has a terminal sugar residue in which carbon atom number 5′ is not linked to another sugar residue, and a **3′ end** that has a terminal sugar residue whose 3′ carbon is not involved in phosphodiester bonding. So, each nucleic acid strand has a 5′ → 3′ orientation, but if two such strands are to form a duplex they must be in an antiparallel arrangement: the 5′ → 3′ direction of one DNA strand must be the opposite to that of its partner (**Figure 1.8**).

The interstrand hydrogen bonds formed in base pairing are crucially important in forming a double helix. In addition, base-stacking forces are important contributors to the stability of the helix, notably van der Waals attractions between adjacent planar bases (see **Figure 1.7B**).

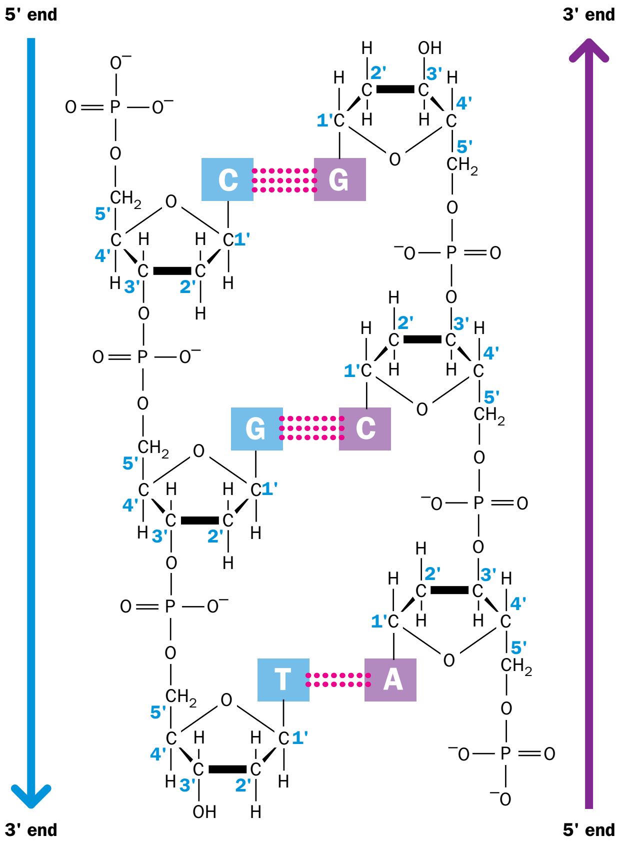


Figure 1.8 Antiparallel nature of the DNA double helix. The two antiparallel DNA strands run in opposite directions in linking 3′ to 5′ carbon atoms in the sugar residues. This double-stranded trinucleotide has the sequence 5′ pCpGpT–OH 3′/5′ pApCpG–OH 3′, where p stands for a phosphate group and –OH 3′ represents the 3′ terminal hydroxyl group. This is conventionally abbreviated to give the 5′ → 3′ sequence of nucleotides on only one strand, either as 5′ CGT 3′ (blue strand) or as 5′ ACG 3′ (purple strand).

Methylated bases in DNA

As described below, the nucleotides in RNA are chemically modified in very many different ways. In DNA, however, chemical modification is limited to methylation of bases. In bacterial cells, for example, certain adenines and cytosines are methylated as a way of distinguishing the host-cell DNA from invading viruses (virus DNA that is not methylated in the same way as the host-cell DNA will be cleaved by a cellular restriction endonuclease).

In vertebrates, nucleotide modification in DNA is directed at the 5′ carbon of certain cytosines, forming 5-methylcytosine (5-meC), which can be hydroxylated to form 5-hydroxymethylcytosine (**Figure 1.9**). Base pairing is not affected: both 5-meC and 5-hydroxymethylcytosine base-pair as normal with guanine. As detailed in Chapter 10, these modifications are **epigenetic marks** that serve as a reversible switch to regulate transcriptional activity.

The CpG dinucleotide (cytosine with a guanine as its 3′ neighbor) can be a target sequence for methylation of cytosines in vertebrate DNA, but it has become clear that cytosine methylation in vertebrate DNA can also occur at cytosines with a different 3′ neighbor (adenine, cytosine, or thymine), notably in brain cells and pluripotent cells (see He & Ecker [2015] PMID 26077819; Further Reading).

Alternative base pairing and DNA conformation

The canonical double helix is portrayed as a rigid, rather uniform structure, but it can undergo local changes in conformation and alternative types of base pairing. Runs of A-T base pairs can make the minor groove even narrower, for example, and under some circumstances and according to the sequence, alternative types of double helix can form, such as the Z-type DNA double helix shown below in **Figure 1.13**. The base pairs in DNA can even adopt different geometries: the standard Watson–Crick base pairing shown in **Figure 1.7A** can flip to form Hoogsteen base pairs after rotation of individual bases through close to 180° around the *N*-glycosidic bond (see **Figure 1.7C**).

As well as being important in many protein–DNA interactions, Hoogsteen base pairing enables certain types of higher-level DNA folding. Short regions of guanine-rich DNA

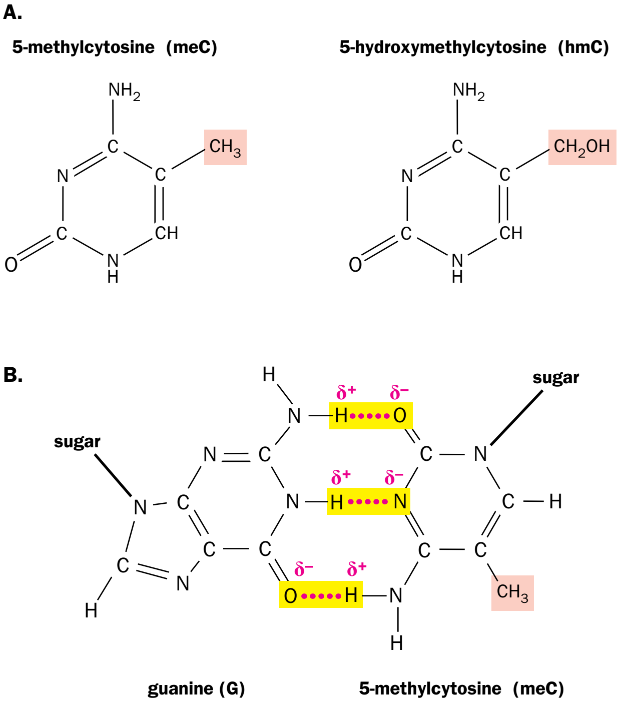


Figure 1.9 Structures of 5-methylcytosine (meC) and 5-hydroxymethylcytosine (hmC). (**A**) The carbon 5 of cytosine has an attached hydrogen that has been replaced by the groups highlighted in pale peach to give 5-methylcytosine and 5-hydroxymethylcytosine. (**B**) In terms of base pairing, the modified cytosines behave as normal cytosines and base-pair normally with guanine (the carbon-5 atom of cytosine is directed away from the laterally opposed guanine, as shown in this example of base pairing of meC to G; compare with the G-C base pairing in **Figure 1.7A**). The cytosine modifications are epigenetic marks that are important for regulating gene expression.

(or RNA) can form a four-stranded DNA structure (G-quadruplex) as a result of the formation of Hoogsteen bonds between sets of four guanines (**Figure 1.10**). DNA sequences with the potential of forming G-quadruplexes are significantly enriched in various functionally important DNA regions. The DNA replication origins in complex eukaryotes, 5′ untranslated regions, and splicing sites are enriched in such sequences and they have also been implicated in the DNA of telomeres, the specialized sequences at the ends of linear chromosomes.

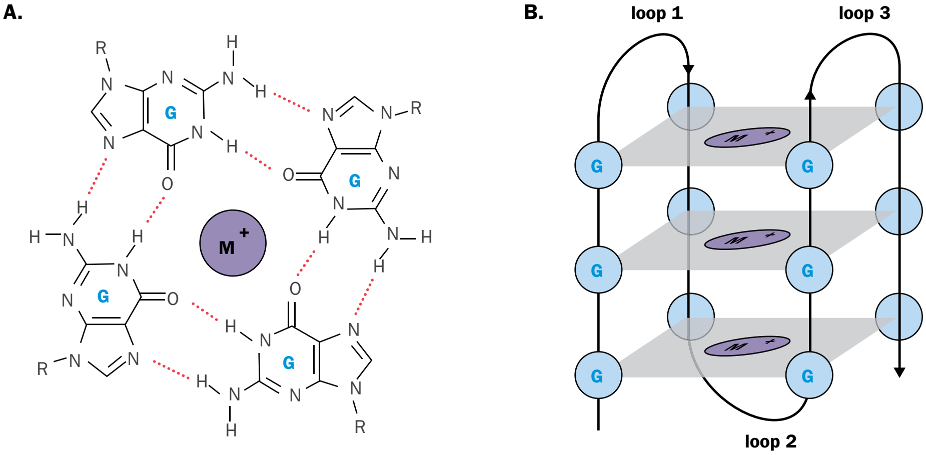


Figure 1.10 G-quadruplex structure. G-quadruplexes are four-DNA-strand structures that can form within guanine-rich DNA (or RNA) sequences. They contain two or more G-tetrads (**A**) that form when four guanines are held in a planar array through Hoogsteen hydrogen bonding coordination of a monovalent cation (M^+) to the lone electron pairs at O6 of each guanine provides additional stabilization. (**B**) An example of a G-quadruplex structure with three stacked G-tetrads. The four-DNA-strand structure is formed after looping of G-rich DNA.

Complementary sequences and sequence notation

Barring new mutation, double-stranded DNAs within cells show perfect base-matching over extremely long distances, such as over 249 million nucleotides in the case of the DNA duplex within human chromosome 1. The two DNA strands of a duplex are said to exhibit base complementarity and to have **complementary** (**base**) **sequences**.

Because genetic information is encoded by the linear sequence of bases in the DNA strands it is customary to define nucleic acids by their base sequences, which are always written in the 5′ → 3′ direction, which is the direction of synthesis of new DNA (or RNA) from a DNA template. The sequence of a single-stranded oligonucleotide might be written accurately as 5′ p-C-p-G-p-A-p-C-p-C-p-A-p-T-OH 3′, where p = phosphate, but it is simpler to write it just as CGACCAT.

When it comes to double-stranded DNA it is sufficient to write the sequence of just one of the two strands; the sequence of the complementary strand can immediately be predicted by the base-pairing rules given above. For example, if a given DNA strand has the sequence CGACCAT, the sequence of the complementary strand can easily be predicted to be ATGGTCG (in the 5′ → 3′ direction, as shown below, where A-T base pairs are shown in green and C-G base pairs in blue).

given DNA strand: 5′ CGACCAT 3′
|||||||
→ complementary strand: 3′ GCTGGTA 5′

DNA replication is semi-conservative

Before cells divide, each double-stranded DNA must be replicated to generate two identical double helices, one each for the two daughter cells. First, the two DNA strands of the original double helix are unwound using a DNA helicase. Then, each of the original DNA strands is used as a template by a DNA polymerase to make a complementary DNA strand, using the four deoxynucleoside triphosphates (dNTPs; that is, a combination of dATP, dCTP, dGTP, and dTTP).

The DNA polymerase cleaves the dNTP precursors between the first (α) and second (β) phosphates. The resulting deoxynucleotides have a single phosphate group (deoxynucleoside monophosphates; dNMPs) and are added to the 3′ end of the growing DNA chain; the residual pyrophosphate (containing the β and γ phosphates that belonged to the dNTP) is discarded (Figure 1.11A).

Figure 1.11B shows how the two parental strands of the original DNA duplex are unwound, allowing the separated strands to each act as templates for a DNA polymerase

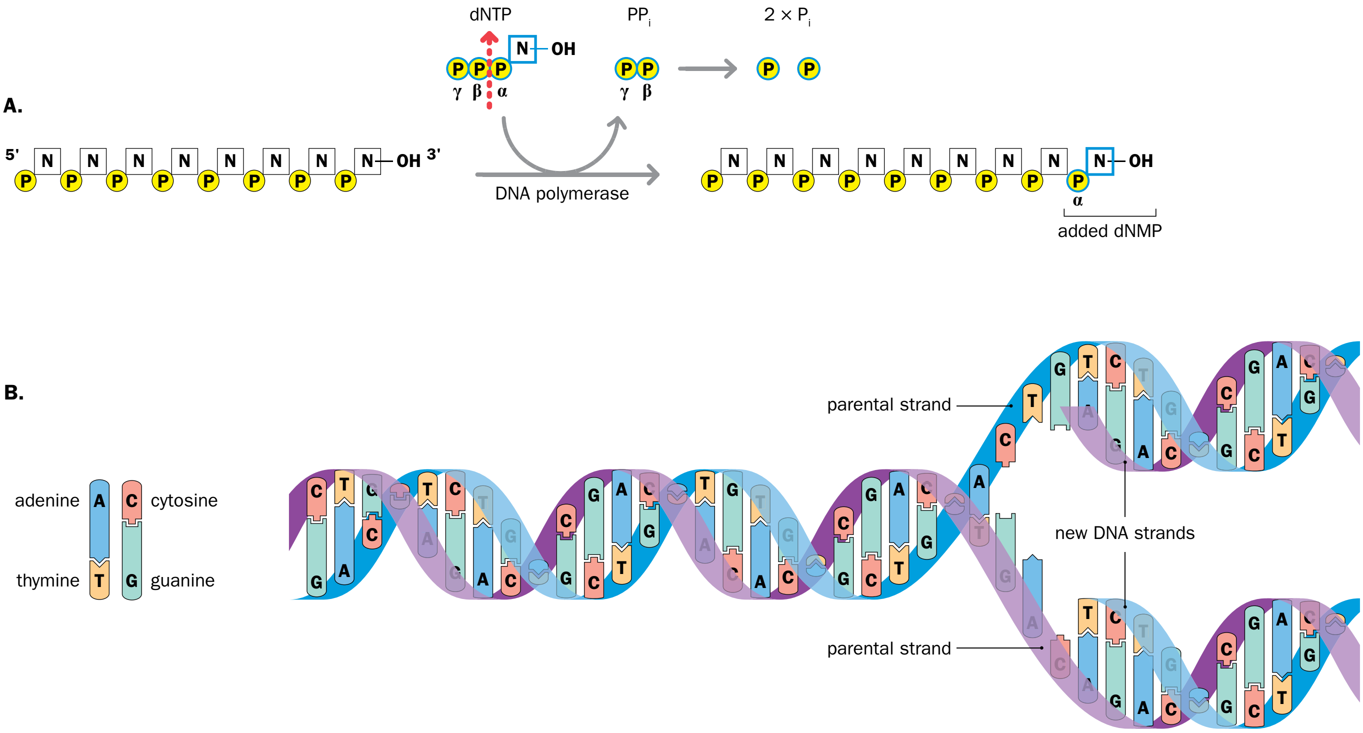


Figure 1.11 Elongation of a DNA chain and replication of a DNA duplex. (**A**) The basic process of elongation of a DNA chain. DNA synthesis is initiated using an RNA primer and a specialized primase enzyme (see text) but after that the new DNA chain grows by adding nucleotides one at a time to the 3′ end using a DNA-dependent DNA polymerase and the four deoxynucleoside triphosphates (dNTPs) as substrates. The reaction is driven by the change in free energy when the substrate dNTP is cleaved at the phosphoanhydride bond between the first (α) and second (β) phosphates. That liberates a dNMP to be added onto the hydroxyl group at the 3′ end of the growing DNA chain and a pyrophosphate group (PP_i) that is then cleaved to give two phosphate groups, with a further change in free energy. (**B**) Replication of a DNA duplex. The parental duplex consists of two complementary DNA strands that unwind to serve as templates for the synthesis of new complementary DNA strands. Each completed daughter DNA duplex contains one of the two parental DNA strands plus one newly synthesized DNA strand, and is structurally identical to the original parental DNA duplex.

to synthesize a new complementary DNA strand. As a result, two daughter DNA duplexes are formed that are identical to each other and to the original parental DNA duplex. As each daughter duplex contains one strand from the original DNA duplex and one newly synthesized DNA strand, DNA replication is said to be semi-conservative.

The semi-discontinuous nature of DNA replication

DNA replication is initiated at specific points, called **origins of replication**, generating Y-shaped replication forks where the parental DNA duplex is opened up. The antiparallel parental DNA strands serve as templates for the synthesis of complementary daughter strands that run in opposite directions.

The overall direction of chain growth is 5′ → 3′ for one newly synthesized daughter strand, the **leading strand**, but 3′ → 5′ for the other daughter strand, the **lagging strand** (**Figure 1.12**). The reactions catalyzed by DNA polymerase involve adding dNMP residues to the free 3′ hydroxyl group of the growing DNA strand. However, only the leading strand always has a free 3′ hydroxyl group that allows continuous elongation in the same direction in which the replication fork moves.

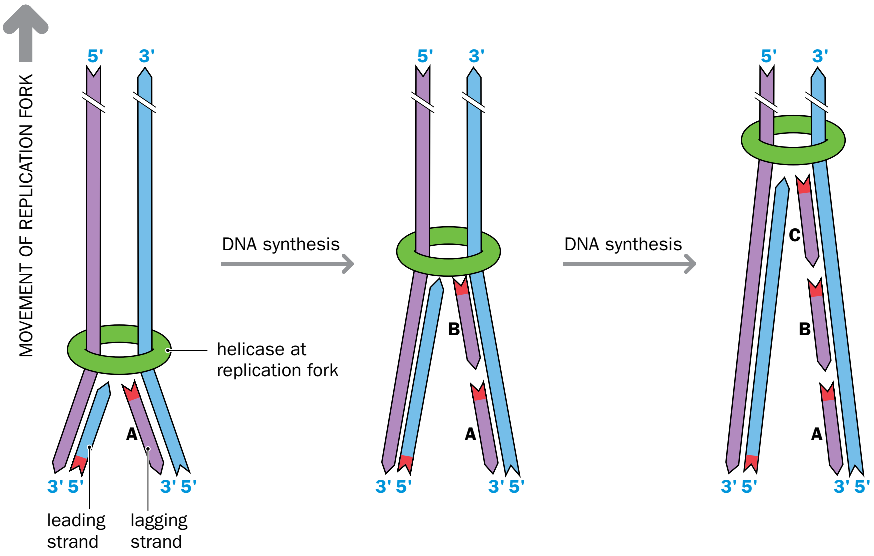


Figure 1.12 Semi-discontinuous DNA replication. A DNA helicase is needed to open up a replication fork, allowing synthesis of new daughter DNA strands to begin. The overall direction of movement of the replication fork matches that of the continuous 5′ → 3′ synthesis of the leading daughter DNA strand. Replication is semi-discontinuous because the lagging strand, which is synthesized in the opposite direction, is built up in pieces (Okazaki fragments, shown here as fragments A, B, and C that are synthesized in the order: A, B, then C). They will later be stitched together using a DNA ligase. Note that the initiation of the leading strand and of each Okazaki fragment of the lagging strand requires a short RNA primer (shown in red) that must be synthesized and base-paired with the original DNA strand. The RNA primers will be subsequently removed and replaced by a corresponding DNA sequence.

The direction of synthesis of the lagging strand is opposite to that in which the replication fork moves. As a result, synthesis of the lagging strand must be discontinuous: it is made as a progressive series of DNA fragments typically 100–1000 nucleotides long (**Okazaki fragments**) that are covalently joined by the enzyme DNA ligase to make the complete lagging strand (**Figure 1.12**). Because only the leading strand is synthesized continuously, DNA synthesis is said to be semi-discontinuous.

Unlike RNA polymerases, DNA polymerases cannot initiate synthesis unless provided with a short oligonucleotide primer with a free 3′ hydroxyl end (from which it can extend chain synthesis). In cells, short RNA primers are used for this purpose and are synthesized by a DNA primase. Just one RNA primer is needed when synthesizing the leading strand, but because synthesis of the lagging strand is discontinuous, an RNA primer is needed to initiate the synthesis of each Okazaki fragment (see **Figure 1.12**). Subsequently, the incorporated short RNA primer sequences will be excised (using a 5′ → 3′ exonuclease) and replaced by the corresponding DNA sequence.

The diversity of mammalian DNA polymerases

The machinery for DNA replication relies on a variety of proteins (**Box 1.1**) and RNA primers, and has been highly conserved during evolution. However, the complexity of the process is greater in complex eukaryotes. Mammalian cells have, for example, close to 20 different multisubunit DNA polymerases with specialized functions.

Most of the DNA polymerases made in our cells are DNA-directed DNA polymerases: they use an individual DNA strand as a template for synthesizing a complementary DNA strand. We also have some RNA-directed DNA polymerases (**reverse transcriptases**) that use RNA templates to make complementary DNA sequences. We introduce one of these, a component of the enzyme telomerase, in Chapter 2; another major source of reverse transcriptases are certain transposon repeats in our genome, as described in Chapter 9.

The family of DNA-directed DNA polymerases includes the classical DNA polymerases δ (delta) and ε (epsilon) that are highly accurate in copying DNA sequences. They are responsible for replicating most of the nuclear DNA of our cells: polymerase δ synthesizes the lagging strand and polymerase ε synthesizes the leading strand. These enzymes have low error rates because they have an associated 3′-5′ exonuclease activity responsible for **proofreading**: if a mistake is made and the wrong base is inserted at the 3′ hydroxyl group of the growing DNA chain, the 3′-5′ exonuclease snips it out so that the correct base can be inserted.

Initiation of DNA replication and of Okazaki fragments requires DNA polymerase α, a complex of a polymerase and a primase. It lacks its own proofreading function: errors that it makes in base incorporation are corrected instead by DNA polymerase δ. An additional DNA polymerase γ, with an intrinsic proofreading exonuclease activity, is dedicated to synthesizing mitochondrial DNA.

Many other DNA polymerases, including some low-fidelity polymerases with comparatively high error rates for base incorporation, have dedicated roles in recombination and DNA repair. We consider them in Chapter 10 within the context of cellular mechanisms used to maximize genetic variation and to repair DNA.

BOX 1.1 MAJOR CLASSES OF PROTEINS INVOLVED IN DNA REPLICATION

- Topoisomerases—start the process of DNA unwinding by breaking a single DNA strand, releasing the tension holding the helix in its coiled and supercoiled form.
- Helicases—unwind the double helix at the replication fork (after supercoiling has been eliminated by a topoisomerase).
- Single-strand-DNA binding proteins—maintain the stability of the replication fork. Single-stranded DNA is very vulnerable to enzymatic attack; the bound proteins protect it from being degraded.
- Primases—attach a small complementary RNA sequence (a **primer**) to single-stranded DNA at the replication fork. The RNA primer provides the 3′ hydroxyl (OH) group needed by DNA polymerase to begin synthesis (unlike RNA polymerases, DNA polymerases cannot initiate new strand synthesis from a bare single-stranded template but require an initiating molecule with a free 3′ OH group onto which dNMPs**,** provided from the appropriate dNTP substrates, can be sequentially attached to build a complementary strand).
- DNA polymerases—synthesize new DNA strands. New cellular DNA synthesis normally depends on an existing DNA strand template that is read by a DNA-directed DNA polymerase. This complex aggregate of protein subunits often also provides DNA proofreading and DNA repair functions. This means that any wrongly incorporated bases can be identified, removed, and repaired. DNA can also be synthesized in cells from an RNA template, using an RNA-directed DNA polymerase (a reverse transcriptase, see text).
- DNA ligases—seal nicks that remain in newly synthesized DNA after the RNA primers are removed and the small gaps filled in by DNA polymerase. The DNA ligases catalyze the formation of a phosphodiester bond between unattached but adjacent 3′ OH and 5′ phosphate groups.

RNA structure and RNA genomes

Unlike DNA, RNA is normally single-stranded. The exception is provided by certain RNA viruses that have a double-stranded RNA genome. Double-stranded RNA is also transiently formed when some RNA viruses with a single-stranded RNA genome transcribe RNA sequences to make a complementary RNA. See **Box 1.2** for the extraordinary variety of viral genomes.

To perform certain cell functions, different cellular RNA molecules may need to transiently associate by RNA–RNA base pairing, and some RNA sequences also engage with DNA sequences to form RNA–DNA duplexes. RNA–RNA and RNA–DNA duplexes can occur transiently within the context of gene regulation and transcription: many regulatory noncoding RNAs work by base pairing with target sequences in an RNA or DNA strand, and RNA–DNA helices are transiently formed during RNA synthesis when one DNA strand is used to make a complementary RNA, as described in the next section. The presence of a 2′ hydroxyl group on ribose sugars provides a structural constraint on the double helix of an RNA–RNA or an RNA–DNA duplex, so that an A-type double helix is formed rather than the common B-type double helix (**Figure 1.13**).

In addition to permitting RNA–RNA duplexes, and RNA–DNA duplexes, hydrogen bonding is crucially important in shaping the structure of a single-stranded RNA, and in allowing functional, short, double-stranded sequence motifs that can be recognized by specific RNA-binding proteins. Hairpin structures are the most common type of secondary structure motif (**Figure 1.14A**). Intrachain base pairing is extremely important in determining the shape of noncoding RNA and complex structures can form as a result, both in classical noncoding RNAs (**Figure 1.14B**) and in the noncoding untranslated regions of mRNAs.

BOX 1.2 THE EXTRAORDINARY VARIETY OF VIRAL GENOMES

The genomes of all cells contain one or more types of double-stranded DNA that may be circular (mitochondria, chloroplasts, and almost all prokaryotic DNAs) or linear (nuclear DNAs in eukaryotes and the DNA of a very few bacteria, such as *Borrelia* spp.). Viruses, however, have developed extraordinary diversity in genome organization.

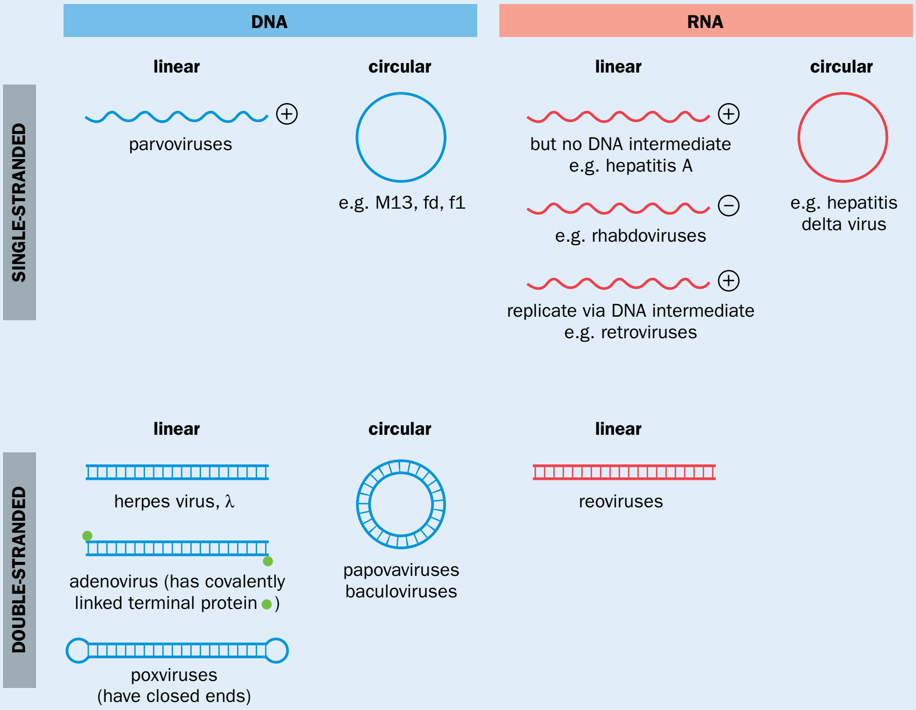
In many viruses, the genome is made up of one type of nucleic acid that may be single-stranded DNA, double-stranded DNA, single-stranded RNA, or double-stranded RNA, and the genome may be linear or circular (**Figure 1**).

Viruses that have a single-stranded genome can be classified as (+)-viruses with a positive-strand genome or (–)-viruses with a negative-strand genome, depending on the *sense* of the RNA used to make polypeptides. That is, if the polypeptide-making RNA has the same sense as the genome, the virus is a (+)-virus; if instead it is complementary in sequence to the genome, the virus is a (–)-virus. Although virus genomes often consist of one type of nucleic acid, some viruses have genomes consisting of multiple different nucleic acids that can be circular or linear (**Figure 2**).

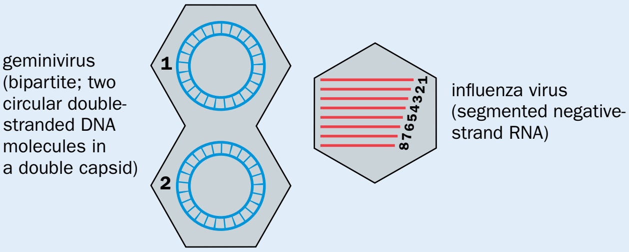
RNA VIRUSES

Whereas DNA viruses generally replicate in the nucleus, many RNA viruses replicate in the cytoplasm. The genomes of RNA viruses are generally small in size, and also have higher mutation rates than those of DNA viruses (because RNA replication has a much higher error rate than DNA replication). The elevated mutation rate allows RNA viruses to adapt rapidly to changing environmental conditions.

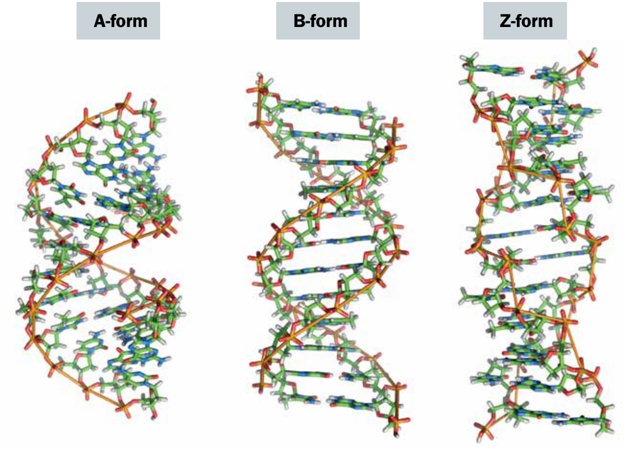
Retroviruses are unusual RNA viruses because they replicate in the nucleus. (Conversely, some DNA viruses replicate though an RNA intermediate; see the legend to **Figure 1**.) The single-stranded RNA genome of retroviruses is first converted into a single-stranded complementary DNA (cDNA) using a viral reverse transcriptase. Next, the cDNA is converted into a double-stranded DNA using a DNA polymerase from the host cell. Dedicated viral proteins then help insert this double-stranded DNA into the genome of the host cell, where it can remain for long periods or be used to synthesize new viral RNA genomes that are packaged as new virus particles.



Box 1.2 Figure 1 The extraordinary diversity of simple virus genomes. The + and – symbols refer to positive or negative single-stranded genomes, as defined in the text. Note that some single-stranded RNA viruses, such as HIV (human immunodeficiency virus), replicate through a DNA intermediate by having a reverse transcriptase that is used to make a complementary DNA copy of their genome; such viruses are known as *retroviruses*. Conversely, some DNA viruses, such as the hepatitis B virus, replicate through an RNA intermediate.



Box 1.2 Figure 2 Examples of segmented and multipartite virus genomes. The single-stranded (–) RNA genome of the influenza virus is segmented into eight separate pieces that specify different polypeptides. The double-stranded DNA genome of a geminivirus is segmented into two different pieces, and because they are separately enclosed by a capsid, this genome is also said to be bipartite.



	A-form	B-form	Z-form
handedness	right-handed	right-handed	left-handed
pitch (nm)	3.2	3.4	4.5
base pairs per turn	11	10	12

Figure 1.13 Diversity in the structure of nucleic acid double helices. Shown here are three alternative structures of a double helix, in this case a DNA duplex. In addition to the B-form, the predominant form of DNA in cells under physiological conditions, the rare Z-form of DNA can also form in the case of certain sequences under certain physiological conditions (high salt concentrations). It has a narrower, more elongated helix. The A-form helix is shorter and wider than the B-form helix and has a deep, narrow major groove, which makes it less accessible to proteins. A-DNA only exists under nonphysiological conditions, but the presence of a 2′ hydroxyl group on the ribose sugar means that under physiological conditions both RNA–RNA and RNA–DNA duplexes are constrained to form an A-form double helix. Pitch, distance per complete turn.

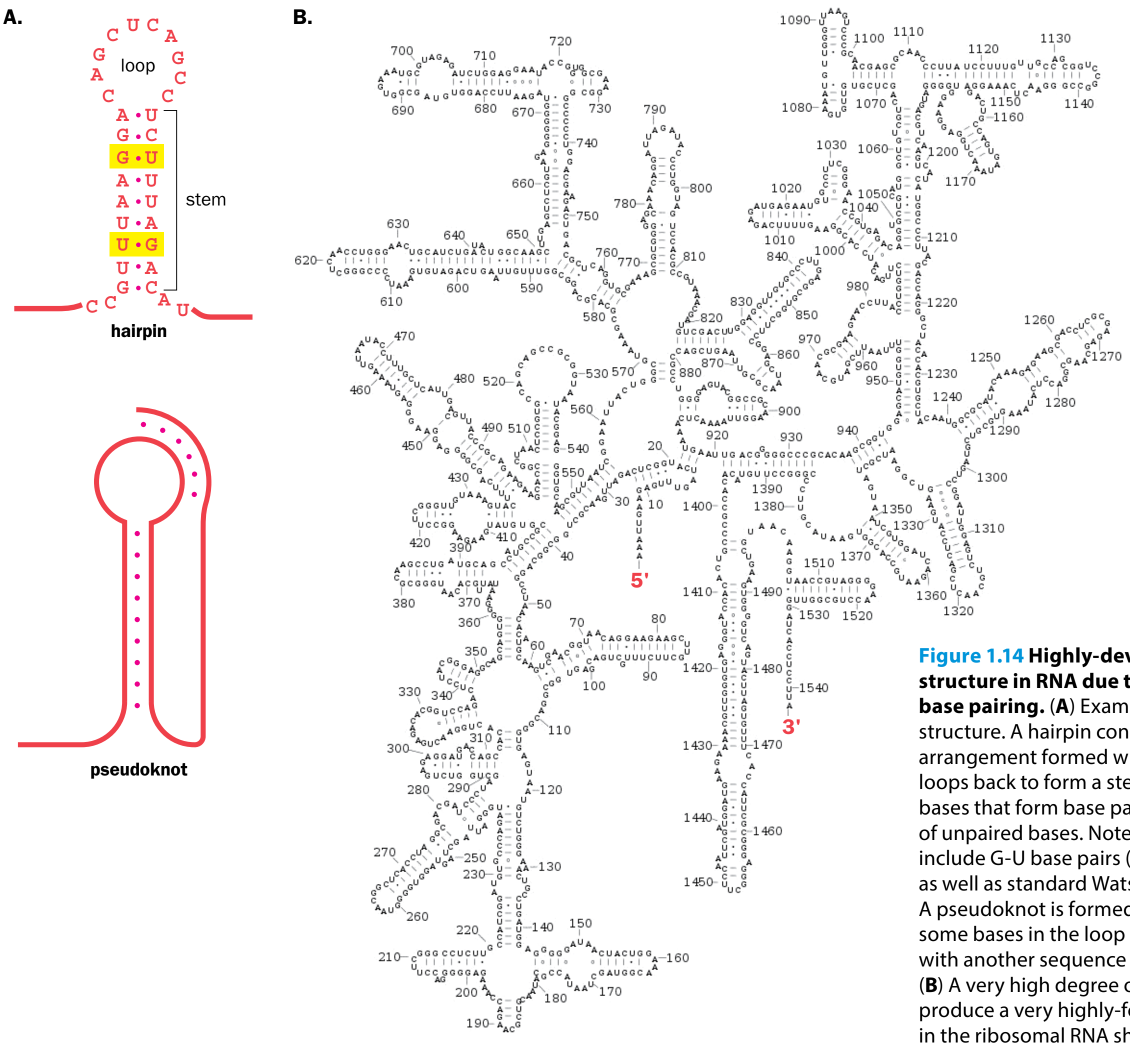


Figure 1.14 Highly-developed secondary structure in RNA due to intramolecular base pairing. (**A**) Examples of RNA secondary structure. A hairpin consists of a stem-and-loop arrangement formed when the RNA strand loops back to form a stem of laterally opposed bases that form base pairs, leaving a small loop of unpaired bases. Note that base pairing can include G-U base pairs (highlighted in yellow) as well as standard Watson–Crick base pairs. A pseudoknot is formed from a hairpin in which some bases in the loop engage in base pairing with another sequence on the RNA strand. (**B**) A very high degree of secondary structure can produce a very highly-folded structure, evident in the ribosomal RNA shown here (the 5′ and 3′ ends of the RNA are highlighted for clarity). Note that this is necessarily a two-dimensional representation of a complex three-dimensional structure. See also the example in the transfer RNA shown in **Figure 1.26**.

Note that in addition to the expected A-U and G-C base pairs, G-U base pairing can occur in short regions of double-stranded RNA within an RNA strand (see, for example, the hairpin in **Figure 1.14A**). Although not particularly stable, G-U base pairing does not significantly distort the RNA–RNA helix.

1.3 RNA TRANSCRIPTION AND GENE EXPRESSION

As well as having global roles in storing and transmitting genetic information and supporting chromosome function, DNA can have cell-type-specific functions because it contains sequences that can be used to make RNA and polypeptides in ways that differ from one cell type to another. **Genes** are discrete DNA segments that are spaced at irregular intervals along a DNA strand and serve as templates for making complementary RNA sequences (transcription). The initial primary RNA transcript must then undergo a series of maturation steps that ultimately result in a mature, functional noncoding RNA or a messenger RNA that will in turn serve as a template to make a polypeptide. Some of the gene products are needed by essentially all cells for a variety of vital cell processes (DNA replication, protein synthesis, and so on). But other RNA and protein products are made in some cell types but not others and may even be specific for individual cells in some exceptional cases (for example, individual B and T lymphocytes can make cell-specific immunoglobulins and T-cell receptors, respectively).

The DNA compositions of the different cell types in a multicellular organism are essentially identical. The variation between cells happens because of differences in gene expression, primarily at the level of transcription: different genes are transcribed in different cells according to the needs of the cells. Some genes, known as housekeeping genes, need to be expressed in essentially all cells, but other genes show tissue-specific gene expression or they may be expressed at specific times (for example, at specific stages of development or of the cell cycle).

The basic process of transcription

We consider the details of transcription more fully in Chapter 10 when we consider how gene expression is regulated. Here, we are concerned with the most basic features of transcription. At its most fundamental level, transcription means that RNA is synthesized using DNA-directed RNA polymerases. That is, DNA strands serve as templates for RNA synthesis. In eukaryotic cells, the bulk of cellular RNA is synthesized in the nucleus by transcribing nuclear genes, but a small amount is synthesized in mitochondria and the chloroplasts of plant cells by transcribing the DNA found in these organelles.

During transcription the DNA helix needs to be unwound locally to allow the RNA polymerase to gain access to a separated DNA strand from which it will make a complementary RNA sequence. The RNA transcript has a complementary sequence to that of the template strand of the DNA, and has the same 5′ → 3′ direction and base sequence (except that U replaces T) as the other, nontemplate DNA strand. The nontemplate strand is often called the **sense strand**, and the template strand is often called the **antisense strand** (**Figure 1.15**).

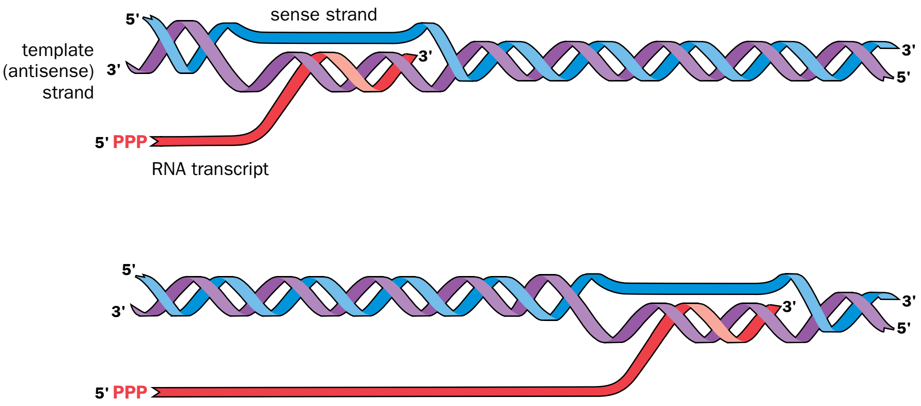


Figure 1.15 Transcribed RNA is complementary in sequence to one strand of DNA. During transcription, the DNA double helix is locally unwound as the RNA polymerase advances. One DNA strand, the template strand, is used by the polymerase as a template to synthesize a complementary RNA strand, and the RNA is synthesized from ribonucleoside triphosphate precursors (rNTPs). The polymerase cleaves rNTPs to give ribonucleoside monophosphates (rNMPs) that are inserted one nucleotide at a time by joining to the 3′ hydroxyl group of the previous nucleotide according to base-pairing rules. (However, the primary RNA transcript will have a triphosphate [PPP] group at its 5′ end that has not been cleaved.) The base sequence of the RNA transcript will be complementary in sequence to the template strand and will therefore be identical to the sense strand of DNA (except that U replaces T). During transcription, a very short region of DNA–RNA double helix is formed transiently. As the polymerase advances, the area with the DNA–RNA helix advances behind it, with the DNA double helix re-forming behind that, displacing the RNA.

RNA polymerases synthesize RNA from four nucleotide precursors: ATP, CTP, GTP, and UTP. Elongation involves the addition of the appropriate ribonucleotide monophosphate residue (AMP, CMP, GMP, or UMP) to the free 3′ hydroxyl group at the 3′ end of the growing RNA strand. These nucleotides are derived by splitting a pyrophosphate residue (PP_i) from their appropriate ribonucleoside triphosphate (rNTP) precursors. Note that the initiator nucleotide at the extreme 5′ end of a primary transcript retains its 5′ triphosphate group.

In documenting gene sequences, it is customary to show only the DNA sequence of the sense strand. The orientation of sequences relative to a gene normally refers to the sense strand. For example, the 5′ end of a gene refers to sequences at the 5′ end of the sense strand, and **upstream** or **downstream sequences** flank the gene at its 5′ or 3′ ends, respectively, with reference to the sense strand.

For transcription to proceed efficiently, various proteins (transcription factors) must bind to particular DNA sequence elements (collectively called a *promoter*) that are often located close to and upstream of a gene. The bound transcription factors serve to position and guide the RNA polymerase. Additional DNA sequence elements and transcription factors are also important as described below and in later chapters.

RNA polymerase classes in eukaryotic cells

There are four classes of DNA-dependent RNA polymerase in eukaryotic cells. As described below, three of them are large multisubunit enzymes that are used in transcribing nuclear genes. A distantly related, single-subunit RNA polymerase is devoted to transcribing mitochondrial DNA. Note that the mitochondrial RNA polymerase is, however, encoded by a nuclear gene: it makes an mRNA that is translated on cytoplasmic ribosomes and then imported into mitochondria. We detail how human mitochondrial DNA is transcribed in Chapter 9.

Unlike DNA polymerases, RNA polymerases do not need an oligonucleotide primer to initiate the RNA synthesis, but the RNA polymerases cannot initiate transcription by themselves. Instead, protein regulators known as transcription factors must activate the process by binding to certain regulatory DNA sequence elements within the gene or its vicinity. A crucial regulatory element is the **promoter**, a collection of closely spaced, short DNA sequence elements in the immediate vicinity of a gene. Promoters are recognized and bound by transcription factors that then guide and activate the polymerase. Transcription factors are said to be *trans*-acting, because they are produced by remote genes and need to migrate to their sites of action. In contrast, promoter sequences are *cis*-acting, because they are located on the same DNA molecule as the genes they regulate.

RNA polymerase II and transcription of protein-coding genes in the nucleus

As an illustration of how the nuclear RNA polymerases work in cells, we take the important example of RNA polymerase II, which is responsible for transcribing all the protein-coding genes in the nucleus plus many important genes encoding different noncoding RNAs (**Table 1.3**). It relies on both general transcription factors (operating in all cells and important for expressing genes in diverse types of cell) plus tissue-specific and sometimes even cell-specific factors (to permit some genes to be expressed in only certain types of cell).

TABLE 1.3 THE FOUR CLASSES OF EUKARYOTIC RNA POLYMERASE

RNA polymerase	Location	RNA synthesized
I	Nucleolus	The bulk of the cytoplasmic ribosomal RNAs (28S, 18S, and 5.8S ribosomal RNAs)[a]
II	Nucleoplasm	All mRNAs from nuclear genes plus many noncoding RNAs: snoRNAs, miRNAs, lncRNAs, and most snRNAs
III	Nucleoplasm	Various small noncoding RNAs: 5S ribosomal RNA, cytosolic tRNAs, U6 snRNA, and others
mt	Mitochondria	All RNAs transcribed from mitochondrial DNA

[a] Ribosomal RNAs are named using Svedberg (S) units, which refer to their rates of sedimentation in the ultracentrifuge the larger the S value, the larger is the size of the RNA. mRNA, messenger RNA; miRNA, microRNA; lncRNAs, long noncoding RNAs; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA. U6 snRNA is a component of the spliceosome, an RNA–protein complex that removes unwanted noncoding sequences from newly formed RNA transcripts.

For a gene to be transcribed by RNA polymerase II, the DNA at the transcription initiation site must first be bound by general transcription factors, to form a pre-initiation complex. The complex that is required to initiate transcription by an RNA polymerase is known as the basal transcription apparatus and consists of the polymerase plus all of its associated general transcription factors. (Note that although there are fixed transcription initiation sites, termination of RNA polymerase II transcripts is not regulated at the DNA level, but instead depends on RNA processing, as described in Section 1.4).

General transcription factors required by RNA polymerase II include TFIIA (**tr**anscription **f**actor for RNA polymerase **II**, complex **A**), TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. Of these, TFIID and TFIIE are known to bind specific core sequence elements within the promoter, such as the elements with the consensus sequences given in Figure 1.16. Note that these transcription factors may themselves comprise a number of components. For example, TFIID consists of the TATA box-binding protein (TBP) plus various TBP-associated factors (TAF proteins) that regulate how TBP binds to the TATA box.

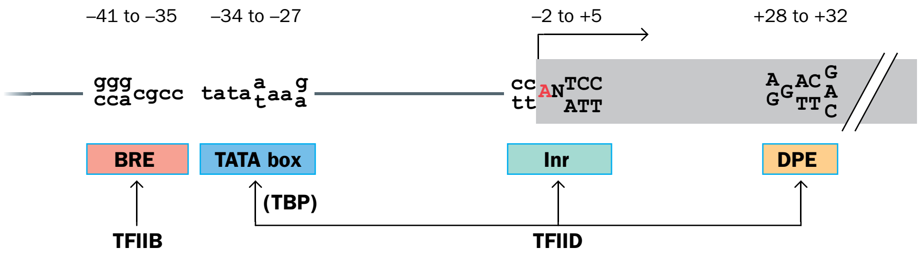


Figure 1.16 Consensus sequences for some core promoter elements often found in genes transcribed by RNA polymerase II. The TATA box is bound by the TATA-binding protein (TBP) subunit of transcription factor TFIID. The initiator (Inr) element defines the transcription start site (the A highlighted in red) when located 25–30 base pairs (bp) from a TATA box. The downstream core promoter element (DPE) is only functional when placed precisely at +28 to +32 bp relative to the highlighted A of an Inr element. Both Inr and DPE are bound by TFIID. Transcription factor TFIIB binds to BRE (TFII**B** **r**ecognition **e**lement) and accurately positions RNA polymerase at the transcription start site. However, none of these elements is either necessary or sufficient for promoter activity, and many active RNA polymerase II promoters lack all of them. N represents any nucleotide. (Modified with permission from Smale ST & Kadonaga JT [2003] *Annu Rev Biochem* **72**:449–479; PMID 12651739. © 2003 by Annual Reviews, http://www.annualreviews.org.)

TFIIF regulates the interaction between RNA polymerase II and TBP and helps attract TFIIE so that TFIIH can be recruited. The latter performs key tasks. Notably, it unwinds the DNA at the transcription start point, and it activates RNA polymerase II, releasing it from the promoter.

In addition to the general transcription factors required by RNA polymerase II, specific recognition elements are recognized by tissue-restricted transcription factors. For example, an **enhancer** is a cluster of *cis*-acting short sequence elements that can enhance the transcriptional activity of a small subset of genes. Unlike a promoter, which has a relatively constant position with regard to the transcriptional initiation site, enhancers are located at variable (often considerable) distances from their transcriptional start sites. Furthermore, their function is independent of their orientation. Enhancers do, however, also bind gene regulatory proteins. The DNA between the promoter and enhancer sites loops out, which brings the two different DNA sequences together and allows the proteins bound to the enhancer to interact with the transcription factors bound to the promoter, or with the RNA polymerase.

A **silencer** has similar properties to an enhancer but it inhibits, rather than stimulates, the transcriptional activity of a specific gene.

Different sets of RNA genes are transcribed by the three eukaryotic RNA polymerases

The protein-coding genes in nuclear DNA are always transcribed by RNA polymerase II. However, nuclear RNA genes (genes that make functional noncoding RNA) may be transcribed by RNA polymerases I, II, or III, depending on the type of RNA (Table 1.3). RNA polymerase I is unusual because it is dedicated to transcribing RNA from a single transcription unit, generating a large transcript that is then processed to yield three types of ribosomal RNA (see below).

RNA polymerase II synthesizes various types of small noncoding RNA in addition to mRNA. They include many types of small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) that are involved in different RNA processing events. In addition, it synthesizes many microRNAs (miRNAs) and long noncoding RNAs (lncRNAs) that can show tissue-specific expression and typically regulate expression of distinctive sets of target genes.

RNA polymerase III transcribes a variety of small noncoding RNAs that are typically expressed in almost all cells (Table 1.3). In many cases, the genes are known to have internal promoters that lie downstream of the transcriptional start site (Figure 1.17). Internal promoters are possible because the job of a promoter is simply to attract transcription

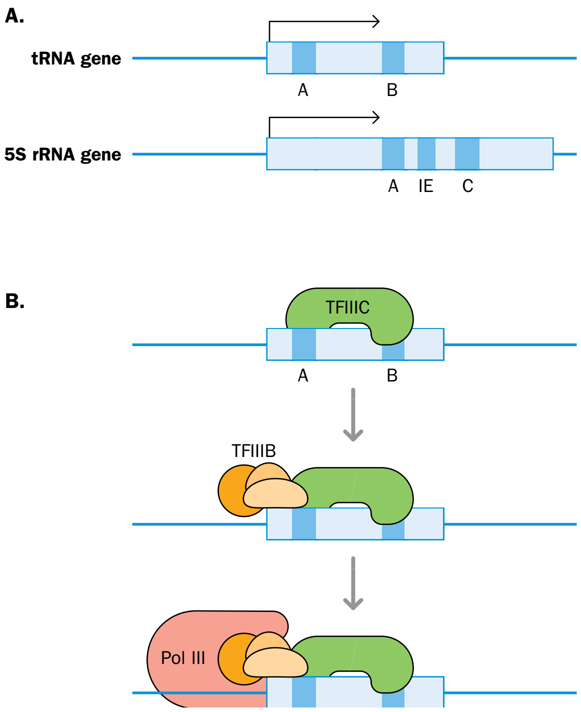


Figure 1.17 Internal promoter elements in many genes transcribed by RNA polymerase III. (**A**) tRNA genes have an internal promoter consisting of an A box (located within the D arm of the tRNA; see **Figure 1.26**) and a B box that is usually found in the TψC arm of tRNA. The promoter of the *Xenopus* 5S rRNA gene has three components: an A box (+50 to +60), an intermediate element (IE; +67 to +72), and the C box (+80 to +90). Arrows mark the +1 position. (**B**) Transcription factor binding to allow expression of a tRNA gene. TFIIIC binds to the A and B boxes of the internal promoter of a tRNA gene then guides the binding of another transcription factor, TFIIIB, to a position upstream of the transcriptional start site. TFIIIB guides RNA polymerase III (Pol III) to bind to the transcriptional start site. Thereafter, TFIIIC is no longer required; any bound TFIIIC will subsequently be removed from the internal promoter to allow unhindered transcription.

factors that will guide the RNA polymerase to the correct transcriptional start site. By the time the polymerase is in place and ready to initiate transcription, any transcription factors previously bound to downstream promoter elements will be removed from the template strand (see **Figure 1.17**).

1.4 RNA PROCESSING

The primary RNA transcripts of most eukaryotic genes undergo a series of processing reactions in order to make a mature mRNA or noncoding RNA. Various types of RNA processing event can be involved (**Table 1.4**). We touch on several of these in this section, but we will cover one type of processing known as RNA editing within the context of gene regulation in Chapter 10.

TABLE 1.4 MAJOR CLASSES OF RNA PROCESSING EVENTS

RNA processing class	Examples	Illustrations
RNA cleavage	During RNA splicing	**Figure 1.18**
	Cleavage of transcripts after polyadenylation signal	**Figure 1.23**
	Cleavage of precursors to various types of noncoding RNAs, including rRNA, tRNA, miRNA	**Figures 1.24** and **10.33**
	Cleavage of large, multigenic mitochondrial RNA transcripts	**Figure 9.1**
Addition of specialized nucleotides at ends of RNA	Capping	**Figure 1.22**
	Polyadenylation	**Figure 1.23**
	CCA addition to 3′ end of tRNA	**Figure 1.26**
Chemical modification of nucleotides	Base modifications, sugar modifications, changes to glycosidic bond	**Figure 1.25**
Substitution of nucleotides	RNA editing (covered in Chapter 10)	**Figures 10.29** and **10.30**

RNA splicing removes unwanted sequences from the primary transcript

For most vertebrate genes, almost all polypeptide-encoding genes, and many RNA genes only a small portion of the gene sequence is eventually decoded to give the final product. In these cases the genetic instructions for making an mRNA or mature noncoding RNA occur in **exon** segments that are separated by intervening **intron** sequences that do not contribute genetic information to the final product.

Transcription of a gene initially produces a primary RNA transcript that is complementary to the entire length of the gene, including both exons and introns. This primary transcript then undergoes **RNA splicing**, whereby the intronic RNA segments are removed and discarded while the remaining exonic RNA segments are joined end-to-end, to give a shorter RNA product (**Figure 1.18**).

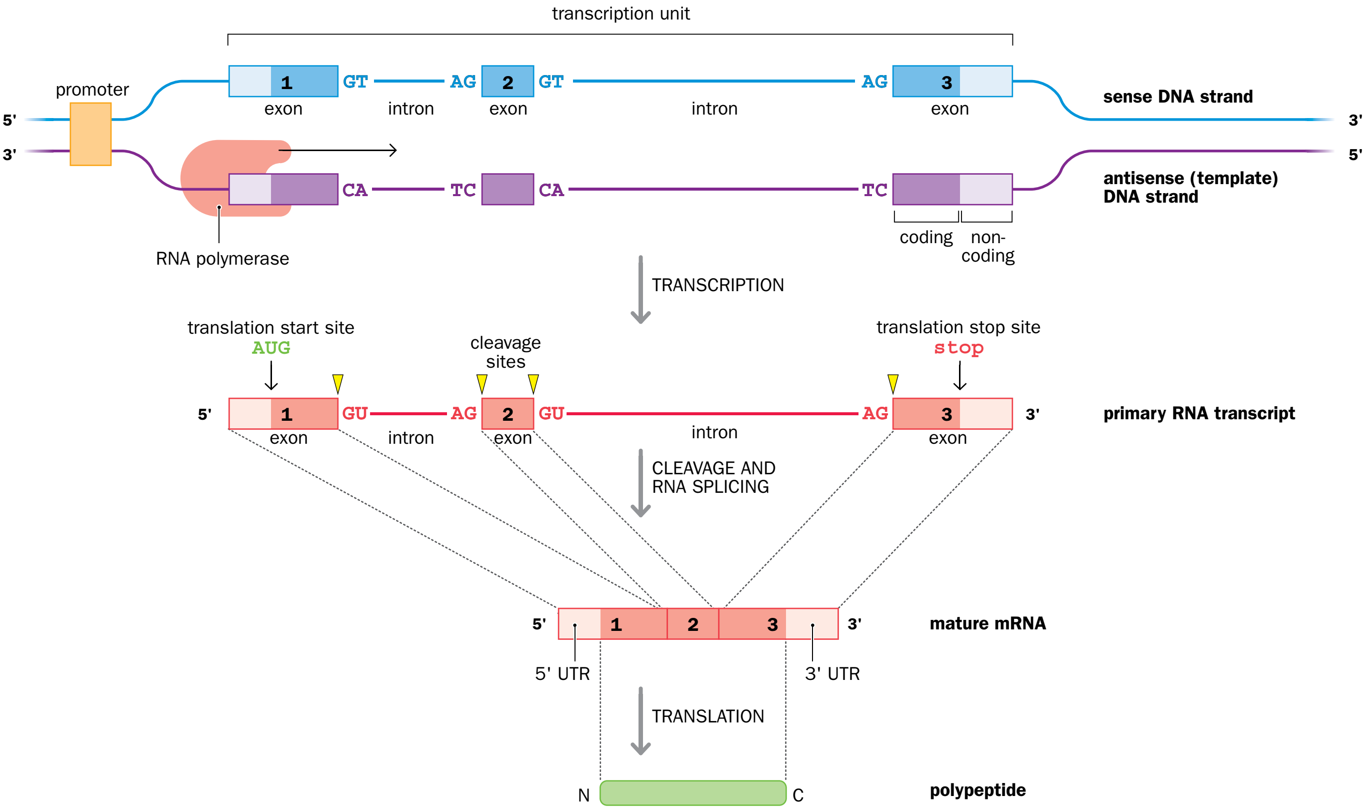


Figure 1.18 RNA splicing brings transcribed exon sequences together. Most of our protein-coding genes (and many RNA genes) undergo RNA splicing. In this generalized example, a protein-coding gene is illustrated with an upstream promoter and three exons separated by two introns that each begin with the dinucleotide GT and end in the dinucleotide AG. The central exon (exon 2) is composed entirely of coding DNA, but exons 1 and 3 have noncoding DNA sequences (that will eventually be used to make untranslated sequences in the mRNA). The three exons and the two separating introns are transcribed together to give a large primary RNA transcript. The RNA transcript is cleaved at positions corresponding to exon–intron boundaries. The two transcribed intron sequences that are excised are each degraded, but the transcribed exon sequences are joined (*spliced*) together to form a contiguous mature RNA that has noncoding sequences at both the 5′ and 3′ ends. In the mature mRNA these terminal sequences will not be translated and so are known as *untranslated* regions (UTRs). The central coding sequence of the mRNA is defined by a translation start site (which is almost always the trinucleotide AUG) and a translation stop site, and is read (*translated*) to produce a polypeptide. N, N-terminus; C, C-terminus.

RNA splicing requires recognition of the nucleotide sequences at the boundaries of transcribed exons and introns (**splice junctions**). The dinucleotides at the ends of introns are highly conserved: the vast majority of introns start with a GT (becoming GU in intronic RNA) and end with an AG (the GT–AG rule).

Although the conserved GT and AG dinucleotides are crucial for splicing, they are not sufficient to mark the limits of an intron. The nucleotide sequences that are immediately adjacent to them are also quite highly conserved, constituting splice junction consensus sequences (**Figure 1.19**). A third conserved intronic sequence that is also important in splicing is known as the branch site and is typically located no more than 40 nucleotides upstream of the intron's 3′-terminal AG (see **Figure 1.19**). Other exonic and intronic

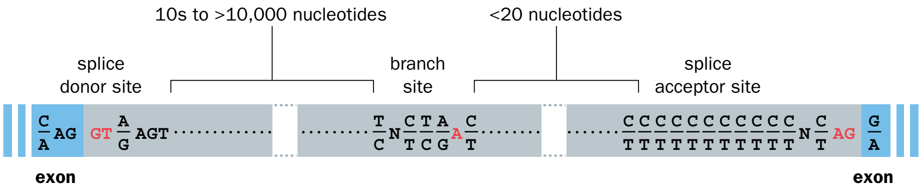


Figure 1.19 **Three consensus DNA sequences in introns of complex eukaryotes.** Most introns in eukaryotic genes contain conserved sequences that correspond to three functionally important regions. Two of the regions, the splice donor site and the splice acceptor site, span the 5′ and 3′ boundaries of the intron. The branch site is an additional important region that typically occurs less than 20 nucleotides upstream of the splice acceptor site. The nucleotides shown in red in these three consensus sequences are almost invariant. The other nucleotides detailed in both the intron and the exons are those most commonly found at each position. In some instances, two nucleotides may be equally common, as in the case of C and T near the 3′ end of the intron. Where N appears, any of the four nucleotides may occur.

sequences can promote splicing (splice enhancer sequences) or inhibit it (splice silencer sequences), and mutations in these sequences can cause disease.

As illustrated in Figure 1.20, the essential steps in splicing are as follows:

1. Nucleophilic attack of the intron's 5′ terminal G nucleotide by the invariant A of the branch-site consensus sequence, to form a lariat-shaped structure;
2. Cleavage of the exon–intron junction at the splice donor site;
3. Nucleophilic attack by the 3′ end of the upstream exon of the splice acceptor site, leading to cleavage and release of the intronic RNA in the form of a lariat, and the splicing together of the two exonic RNA segments.

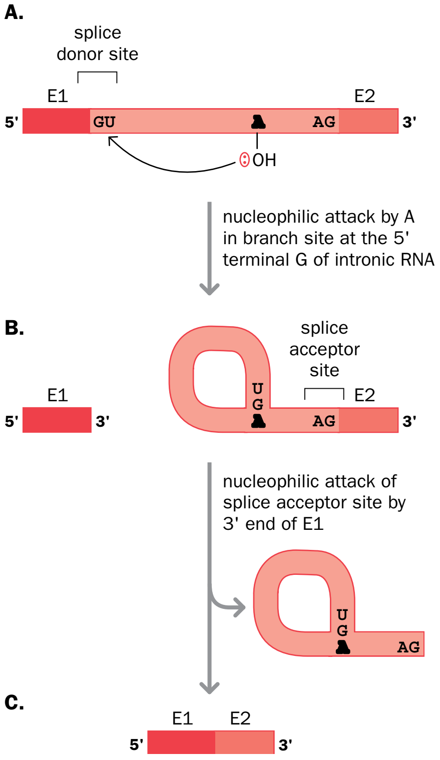

A.
splice donor site
E1
E2
5'
GU
AG
3'
OH
nucleophilic attack by A in branch site at the 5' terminal G of intronic RNA
B.
splice acceptor site
E1
5'
3'
U G
E2
AG
3'
nucleophilic attack of splice acceptor site by 3' end of E1
U G
AG
C.
E1
E2
5'
3'

Figure 1.20 **The mechanism of RNA splicing.** (**A**) The unprocessed primary RNA transcript with intronic RNA separating sequences E1 and E2 that correspond to exons in DNA. The splicing mechanism involves a nucleophilic attack on the G of the 5′ GU dinucleotide. This is carried out by the 2′ hydroxyl (OH) group on the conserved A of the branch site and results in (**B**) formation of a lariat structure, and cleavage of the splice donor site. The 3′ OH at the 3′ end of the E1 sequence carries out a nucleophilic attack on the splice acceptor site causing release of the intronic RNA (as a lariat-shaped structure) and (**C**) fusion (splicing) of E1 and E2.

In the case of genes residing in eukaryotic nuclei, RNA splicing is mediated by a large RNA–protein complex called the **spliceosome**. Spliceosomes have five types of snRNA and more than 50 proteins. The snRNA molecules associate with proteins to form small nuclear ribonucleoprotein (snRNP, or "snurp") particles. The specificity of the splicing reaction is established by RNA–RNA base pairing between the RNA transcript to be spliced and snRNA molecules within the spliceosome. There are two types of spliceosome:

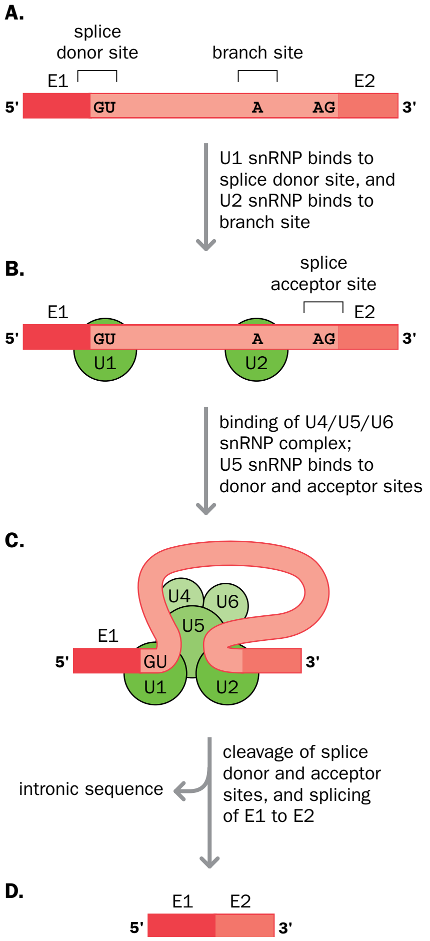
- The major (GU-AG) spliceosome processes transcripts that have classical GU-AG introns. It contains five types of snRNA: U1 and U2 snRNAs recognize and bind the splice donor and branch sites, respectively; U4, U5, and U6 snRNAs subsequently bind to cause looping out of the intronic RNA (Figure 1.21);
- The minor (AU-AC) spliceosome processes transcripts that have rare AU-AC introns. It also has five snRNAs but uses U11 and U12 snRNA instead of U1 and U2 and has variants of U4 and U6 snRNA.

Once a splice donor site is recognized by the spliceosome, it scans the RNA sequence until it meets the next splice acceptor site (signaled as a target by the upsteam presence of the branch-site consensus sequence).

Specific nucleotide additions are made at the ends of RNA polymerase II transcripts and transfer RNA precursors

Many mRNA transcripts and mature noncoding RNAs have nucleotides at their ends that were not directed by transcription of the sense strand of the respective gene. Instead, the specific end nucleotides or oligonucleotides were covalently attached by processing enzymes. The effect is known, or thought, to be to enhance the stability of the RNA or assist it in carrying out an important function.

In addition to RNA splicing, the ends of RNA polymerase II transcripts undergo modifications. The 5′ end is capped by adding a specific variant guanine that is added by means of an unusual and distinctive type of phosphodiester bond, and a long sequence of adenines is added to the 3′ end. As well as protecting the ends of the RNA from cellular exonucleases, these modifications may assist the correct functioning of the RNA transcripts. The addition of the trinucleotide CCA to the 3′ end of transfer RNAs is important for tRNA function, as described below.


A.
splice donor site
branch site
E1
E2
5'
GU
A
AG
3'
U1 snRNP binds to splice donor site, and U2 snRNP binds to branch site
B.
splice acceptor site
E1
E2
5'
GU
A
AG
3'
U1
U2
binding of U4/U5/U6 snRNP complex; U5 snRNP binds to donor and acceptor sites
C.
U4
U6
U5
E1
5'
GU
3'
U1
U2
cleavage of splice donor and acceptor sites, and splicing of E1 to E2
intronic sequence
D.
E1
E2
5'
3'

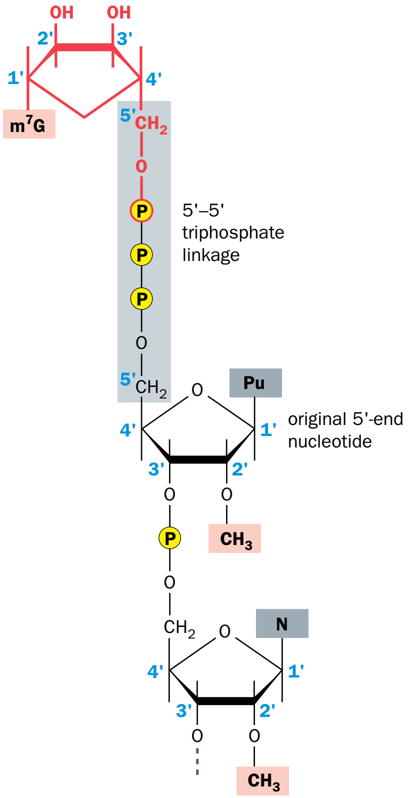


Figure 1.21 Role of small nuclear ribonucleoproteins (snRNPs) in RNA splicing. (**A**) The unprocessed primary RNA transcript as in **Figure 1.20**. (**B**) Within the spliceosome, part of the U1 snRNA is complementary in sequence to the splice donor-site consensus sequence, and binds to it by RNA–RNA base pairing. The U2 snRNA similarly binds to the branch site. Interaction between the splice donor and splice acceptor sites is stabilized by (**C**) the binding of a multi-snRNP particle that contains the U4, U5, and U6 snRNAs. The U5 snRNP binds simultaneously to both the splice donor and splice acceptor sites. Their cleavage releases the intronic sequence and allows (**D**) E1 and E2 to be spliced together.

Figure 1.22 The 5′ cap of a eukaryotic mRNA. The 5′ end of a eukaryotic mRNA has a specialized cap that provides protection against exonucleases and has various other functions, including assisting initiation of translation (see text). The capping process involves: (i) removal of the gamma (γ) phosphate of the original terminal 5′ nucleotide, which is normally a purine (Pu); (ii) addition of a GMP (derived from a GTP precursor) through a *5′-5′ triphosphate linkage* (gray shading); (iii) methylation of nitrogen atom 7 of the new 5′ terminal G to produce 7-methylguanosine (m^{7}G). In mRNAs synthesized in vertebrate cells, the 2′ carbon atom of the ribose of each of the two adjacent nucleotides is also methylated, as illustrated by pink shading. N, any nucleotide.

5′ Capping

Shortly after transcriptional initiation by RNA polymerase II, a methylated nucleoside (7-methylguanosine, m^{7}G) is added and linked by a 5′–5′ phosphodiester bond to the first 5′ nucleotide. This is a major feature of an end-addition form of RNA processing known as **capping** of the 5′ end of the transcript (Figure 1.22). 5′ Capping is thought to serve diverse functions, including:

- Protecting the transcript from 5′ → 3′ exonuclease attack (the uncapped RNA transcripts are rapidly degraded);
- Facilitating transport of mRNAs from the nucleus to the cytoplasm;
- Facilitating RNA splicing; and
- Facilitating attachment of the 40S subunit of cytoplasmic ribosomes to mRNA during translation.

3′ Polyadenylation

Transcription by both RNA polymerase I and III stops after the enzyme recognizes a specific transcription termination site. However, the 3′ ends of mRNA molecules are determined by a post-transcriptional cleavage reaction. As RNA polymerase II advances to transcribe a gene, it carries at its rear two multiprotein complexes, CPSF (cleavage and polyadenylation specificity factor) and CStF (cleavage and stimulation factor), that co-operate to identify a specific hexanucleotide polyadenylation signal, often AAUAAA, located downstream of the termination codon in the RNA transcript. Thereafter, the RNA is cleaved at a specific site 15–30 nucleotides downstream of the AAUAAA sequence (although the primary transcript may continue for hundreds or even thousands of nucleotides past the cleavage point).

After cleavage has occurred, in mammalian cells about 200 adenylate (AMP) residues are added sequentially by the enzyme poly(A) polymerase. This polyadenylation reaction (Figure 1.23) produces a **poly(A) tail** that is thought to:

- Help transport mRNA to the cytoplasm;
- Stabilize at least some mRNA molecules in the cytoplasm; and
- Enhance recognition of mRNA by the ribosomal machinery.

Histone genes are unique in producing mRNA that does not become polyadenylated; termination of their transcription nevertheless also involves 3′ cleavage of the primary transcript.

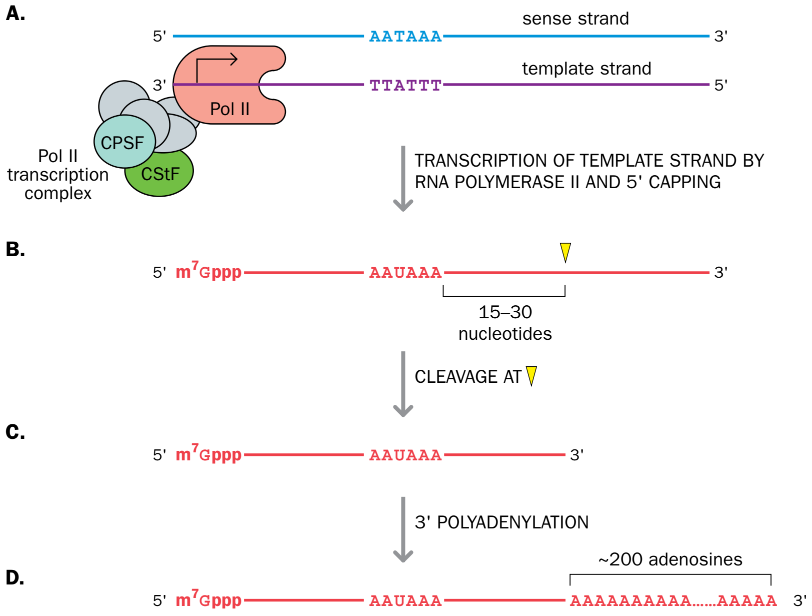


Figure 1.23 Polyadenylation of 3′ ends of eukaryotic mRNAs. (**A** and **B**) RNA polymerase II transcribes the template strand of a gene, and as it does this it carries at its rear multiprotein complexes including two required for polyadenylation: CPSF (cleavage and polyadenylation specificity factor) and CStF (cleavage and stimulation factor). They co-operate to identify a polyadenylation signal downstream of the termination codon in the RNA transcript and to cut the transcript at the point marked by the yellow dart. The polyadenylation signal comprises an AAUAAA sequence or close variant and some poorly understood downstream signals. (**C**) Cleavage occurs normally about 15–30 nucleotides downstream of the AAUAAA element and (**D**) AMP residues are subsequently added by poly(A) polymerase to form a poly(A) tail.

3′ CCA addition to tRNAs

All mature transfer RNAs have at their 3′ terminus the sequence CCA, but in eukaryotes this sequence is not copied from the sense strand of the tRNA gene. Instead, it is added by a nucleotidyltransferase in the nucleus. This sequence is vital for the function of a tRNA and ensures correct recognition of the mature tRNA by an enzyme that will covalently link the correct amino acid to the end adenosine, as described in Section 1.5. Any tRNA that has not been correctly processed in this way will not be permitted export to the cytoplasm.

Noncoding RNAs are formed after a series of cleavage events and chemical modification of individual nucleotides

RNA splicing involves cleavages of the primary RNA transcripts to generate exon sequences that are then stitched together. Most protein-coding genes undergo RNA splicing, and so do many RNA genes. But additional types of cleavage occur in the processing of most types of noncoding RNA including ribosomal RNAs, tRNAs, miRNAs, and so on.

Ribosomal RNA synthesis provides exceptional examples of RNA cleavages. Four major classes of eukaryotic ribosomal RNA (rRNA) have been identified: 28S, 18S, 5.8S, and 5S rRNA (S is the Svedberg coefficient, a measure of how fast large molecular structures sediment in an ultracentrifuge, corresponding directly to size and shape). 18S rRNA is found in the small subunits of ribosomes; the other three are components of the large subunit. Very large amounts of rRNA are required for cells to carry out protein synthesis, and many genes are devoted to making rRNA in the nucleolus, a visibly distinct compartment of the nucleus.

In human cells, a cluster of about 250 genes synthesizes 5S rRNA using RNA polymerase III. However, the 28S, 18S, and 5.8S rRNAs are encoded by consecutive genes on a common 13 kilobase (kb) transcription unit (**Figure 1.24**) that is transcribed by a dedicated polymerase, RNA polymerase I. A compound unit of the 13 kb transcription unit and an adjacent 27 kb nontranscribed spacer is tandemly repeated about 30–40 times at the **nucleolar organizer regions** on the short arms of each of the five human acrocentric chromosomes (13, 14, 15, 21, and 22). Although on different chromosomes, these five clusters of rRNA genes, each about 1.5 Mb long, are brought into close proximity within nucleoli, where they are transcribed in concert; the aggregate cluster of rRNA genes is sometimes referred to as **ribosomal DNA (rDNA)**.

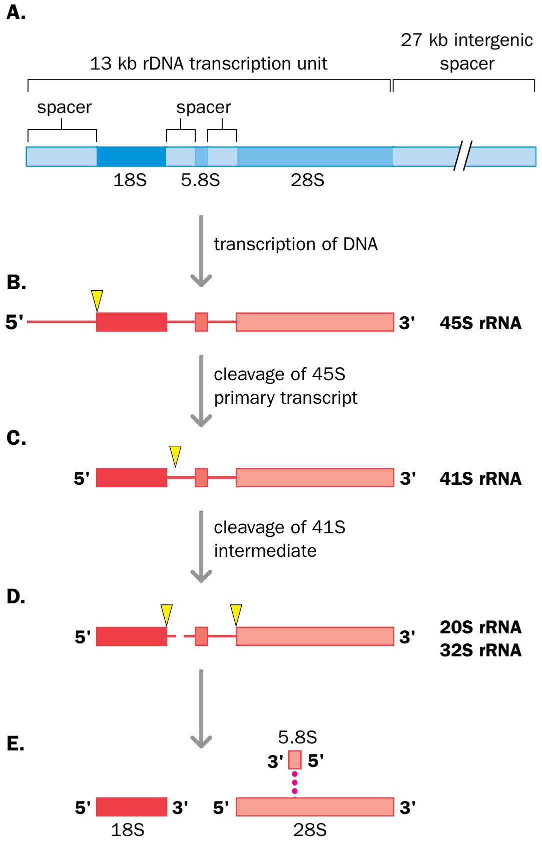


Figure 1.24 Three major rRNA classes are synthesized by cleavage of a shared primary transcript. (**A**) In human cells, the 18S, 5.8S, and 28S rRNAs are encoded by a single transcription unit that is 13 kb long. It occurs within tandem repeat units of about 40 kb that also include a roughly 27 kb nontranscribed (intergenic) spacer. (**B**) Transcription by RNA polymerase I produces a 13 kb primary transcript (45S rRNA) that then undergoes a complex series of post-transcriptional cleavages. (**C**–**E**) Ultimately, individual 18S, 28S, and 5.8S rRNA molecules are released. The 18S rRNA will form part of the small ribosomal subunit. The 5.8S rRNA binds to a complementary segment of the 28S rRNA; the resulting complex will form part of the large ribosomal subunit. The latter also contains 5S rRNA, which is encoded separately by dedicated genes transcribed by RNA polymerase III.

In the case of transfer RNAs, cleavages are required to remove both a 5′ leader sequence of eight nucleotides and a 3′ trailer sequence of four nucleotides, prior to addition of the terminal CCA sequence at the 3′ end. When we consider the details of gene regulation in Chapter 10, we will also illustrate how miRNAs are formed from larger precursors by RNA cleavage.

Chemically modified nucleosides in RNA

In humans (and other vertebrates), chemical modification of nucleotides in DNA is limited to methylation at the 5′ carbon of certain cytosines, as described above. However, maturation of RNA involves frequent and highly varied modification of nucleotides (more than 100 different types of modification are recorded in the RNA Modification Database). The modifications occur at the level of nucleosides and they can be of three types: base modifications, sugar modifications, and altered glycosidic bonding (see **Figure 1.25** for examples).

At least 16 different modified nucleosides occur naturally in mRNA (see Figure 1 of Li & Mason [2014], PMID 24898039; Further Reading). 7-Methylguanosine and 2′-*O*-ribosyl methylations occur at the 5′ end of vertebrate mRNAs (see **Figure 1.22**). In addition, various modifications can occur in internal nucleotides in mRNA. The most common is N^6-methyladenosine (a methyl group is attached to the N6 position of adenine); the methylated adenosine often occurs in the sequence (A/G)

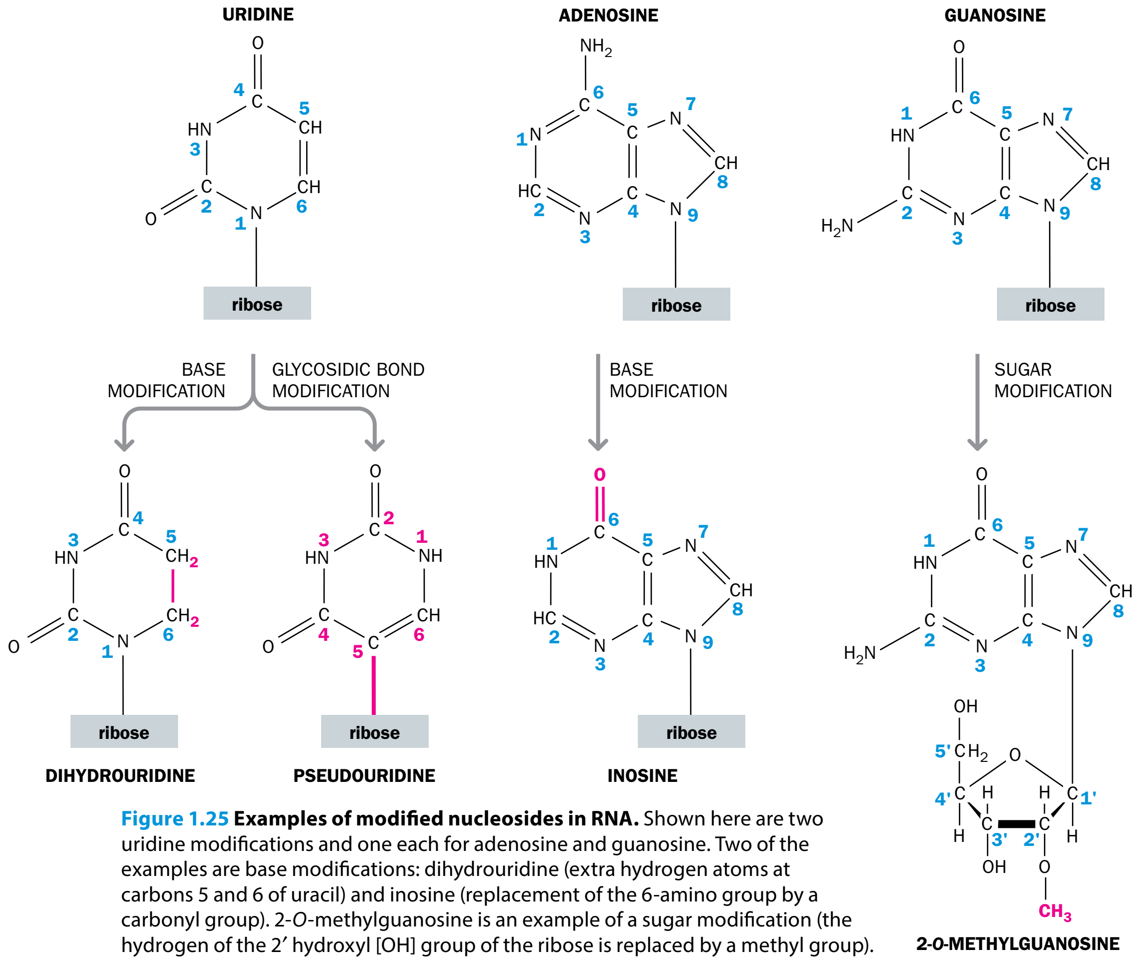


Figure 1.25 Examples of modified nucleosides in RNA. Shown here are two uridine modifications and one each for adenosine and guanosine. Two of the examples are base modifications: dihydrouridine (extra hydrogen atoms at carbons 5 and 6 of uracil) and inosine (replacement of the 6-amino group by a carbonyl group). 2-*O*-methylguanosine is an example of a sugar modification (the hydrogen of the 2′ hydroxyl [OH] group of the ribose is replaced by a methyl group). Pseudouridine is an isomer of uridine and is produced by altering the glycosidic bond that connects the base to the sugar (in the normal *N*-glycosidic bond uracil is attached through nitrogen 1, but in pseudouridine the glycosidic bond forms with carbon 5 of uracil instead of nitrogen 1).

mAC the methylated adenine is immediately preceded by a purine and followed by a cytosine. Other internal modified nucleosides in mRNA include 5-methylcytidine and *N*^{6},2′-*O*-dimethyladenosine.

Nucleoside modifications are especially prevalent in stable noncoding RNAs, including rRNAs (where the RNA modifications are undertaken using different snoRNAs) and particularly so in transfer RNAs. Approximately 12% of the nucleotides in tRNAs are modified, mostly as a result of some type of methylation (in the tRNA illustrated in **Figure 1.26**, nine out of the 74 nucleotides show chemical modifications, and seven of the nine involve methylation). The patterns of nucleotide modification are highly conserved in different tRNAs. That is, the same types of modification tend to occur at specific nucleotide positions—the TψC and D arms of a tRNA are even defined by the occurrence of pseudouridine (ψ) and dihydrouridine (D) at defined nucleotide positions (see **Figure 1.26**).

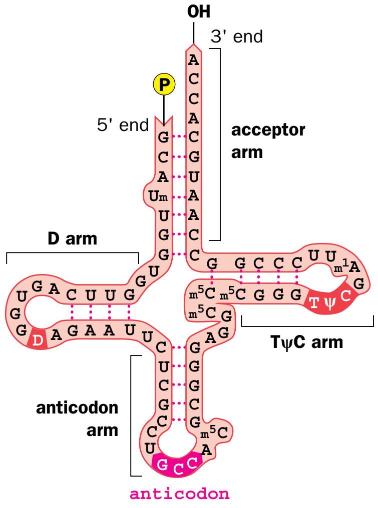


Figure 1.26 Extensive intramolecular base pairing and nucleoside modification in transfer RNA. The tRNA^{Gly} shown here has the classical cloverleaf tRNA structure, with three stem-loops (the D arm, the anticodon arm, and the TψC arm) plus a stretch of base pairing between the 5′ and 3′ terminal sequences (called the acceptor arm because the 3′ end is where an amino acid is attached). Note that different tRNAs tend to have the same number of base pairs in the stems of the different arms of their cloverleaf structure and that the anticodon at the center of the middle loop identifies the tRNA according to the amino acid it will bear. Nine of the 74 nucleotides have been subjected to a nucleoside modification, including four 5-methylcytidines (m^{5}C) and one each of 1-methyladenosine (m^{1}A), 2′-*O*-methyluridine (Um), 5,6-dihydrouridine (D), and pseudouridine (ψ). In addition, a uracil was methylated at its carbon 5 to give thymine (T). See **Figure 1.25** for the structures of some of the modified nucleotides.

The precise purpose of the nucleotide modifications remains to be clarified. Many of the modifications in the main body of tRNA are known to affect tRNA folding and stability, and several of the modifications around the anticodon loop can affect translation or cell growth (for example, in wobble base pairing and stabilization of codon–anticodon interactions). In mRNA, the comparatively abundant 6-methyladenosine has been implicated in regulating alternative splicing, and the same modification is used to signal that precursor miRNAs are ready to be processed into miRNAs.

1.5 TRANSLATION, POST-TRANSLATIONAL PROCESSING, AND PROTEIN STRUCTURE

The different mRNAs produced by genes in the nucleus migrates to the cytoplasm. Here they engage with large (80S) ribosomes (located on the outer membrane of the nuclear envelope and the connecting rough endoplasmic reticulum, and also in the cytosol). With the help of various other components, translation is initiated to produce polypeptides. Messenger RNAs transcribed from genes in the mitochondria and chloroplasts are translated on comparatively small (55S) ribosomes within these organelles.

Only a central segment of a eukaryotic mRNA molecule is translated to make a polypeptide. The flanking **untranslated regions** (the 5′ UTR and 3′ UTR) are transcribed from exon sequences present at the 5′ and 3′ ends of the gene. They assist in binding and stabilizing the mRNA on the ribosomes, and promote efficient translation (Figure 1.27).

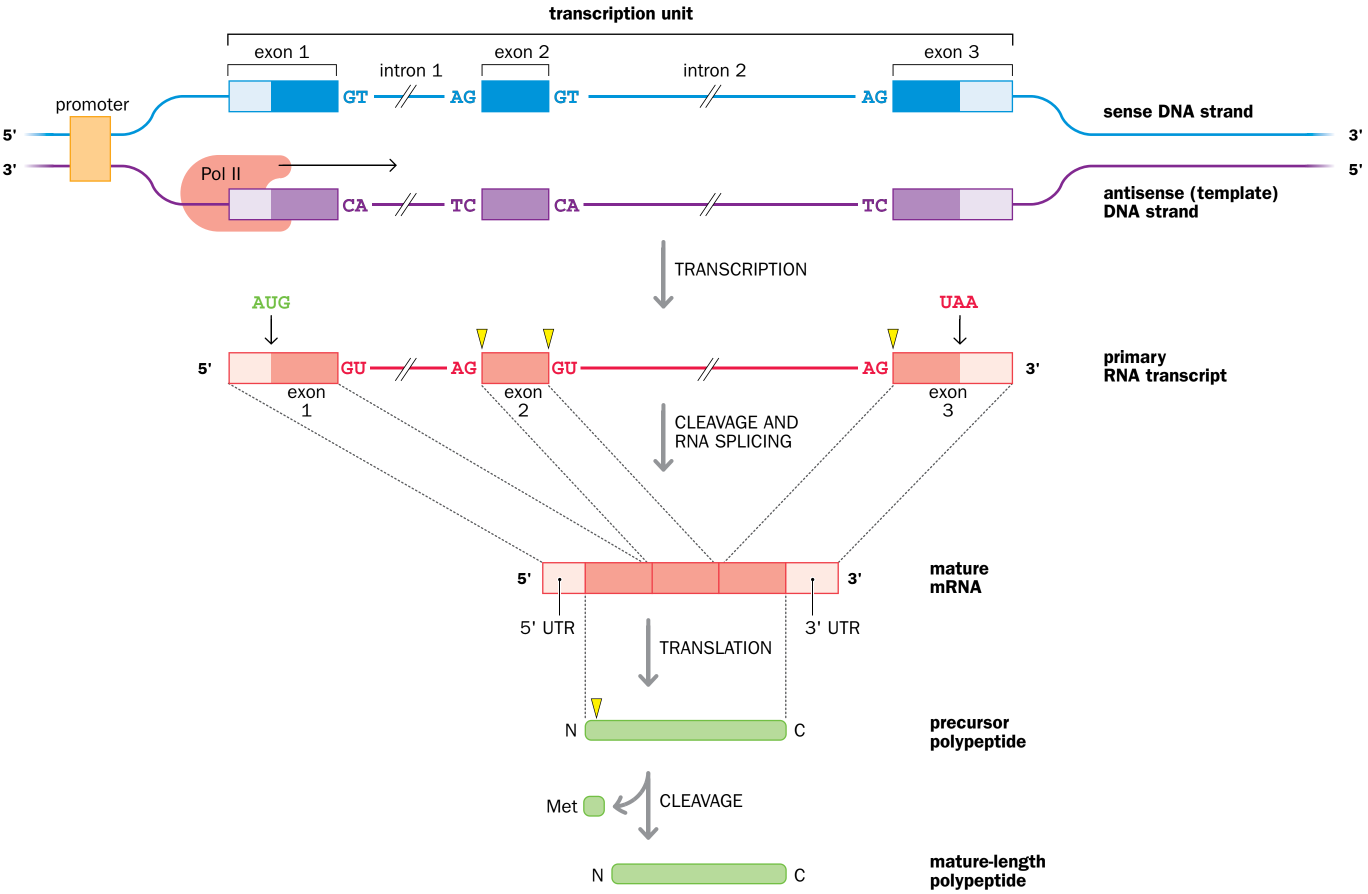


Figure 1.27 Transcription and translation of the human β-globin gene. The β-globin gene comprises three exons and two introns. The 5′ end sequence of exon 1 and the 3′ end sequence of exon 3 are noncoding sequences (unshaded sections). The primary RNA transcript represents a faithful RNA copy of the transcription unit of the sense strand, having an identical base sequence (except that thymines are replaced by uracils). After cleavage to remove intron sequences, the exons are spliced together to give a mature mRNA with a central coding sequence that will be translated (red). The noncoding sequences at the ends of the mRNA (unshaded) will not be translated, and are described as the 5′ untranslated region (5′ UTR) and the 3′ untranslated region (3′ UTR). During translation, as is customary, the initiating codon is AUG, which specifies a methionine, and translation is halted upon encountering an in-frame stop codon, in this case the stop codon UAA. The 147-amino-acid precursor polypeptide undergoes cleavage to remove the methionine at its N-terminus (N), generating the mature 146-amino-acid polypeptide. For the sake of clarity, capping and polyadenylation of the RNA has been omitted.

Ribosomes are large RNA–protein complexes composed of two subunits. In eukaryotes, the 80S ribosomes have a large 60S subunit that contains three types of rRNA molecule, 28S rRNA, 5.8S rRNA, and 5S rRNA, and a smaller 40S subunit containing a single 18S rRNA. In humans, the 80S ribosomes have 80 proteins, 47 in the large subunit and 33 in the small subunit.

Ribosomes provide the structural framework for polypeptide synthesis. The RNA components are predominantly responsible for the catalytic function of the ribosome; the protein components are thought to enhance the function of the rRNA molecules, although a surprising number of them do not appear to be essential for ribosome function.

Messenger RNA is decoded to specify polypeptides

The assembly of a new polypeptide from its constituent amino acids is governed by a triplet genetic code. Within an mRNA, the central nucleotide sequence that is used to make the polypeptide is scanned from 5′ to 3′ on the ribosome in groups of three nucleotides, called **codons** (the corresponding groups of three nucleotides on the sense strand of DNA are called triplets).

Each codon specifies an amino acid and the decoding process uses a collection of different tRNA molecules, each of which binds one type of amino acid. An amino acid–tRNA complex is known as an aminoacyl tRNA and is formed when a dedicated aminoacyl tRNA synthetase covalently links the required amino acid to the terminal adenosine in the conserved CCA trinucleotide at the 3′ end of the tRNA.

Each tRNA has its own **anticodon**, a trinucleotide at the center of the anticodon arm (see **Figure 1.26**) that provides the necessary specificity to interpret the genetic code. For an amino acid to be added to a growing polypeptide, the relevant codon of the mRNA molecule must be recognized by base pairing with a complementary anticodon on the appropriate aminoacyl tRNA molecule. This happens on the ribosome. The small ribosomal subunit binds the mRNA, while the large subunit has two sites for binding aminoacyl tRNAs: a P (peptidyl) site and an A (aminoacyl) site (**Figure 1.28**).

The cap at the 5′ end of messenger RNA molecules is important in initiating translation. It is recognized by certain key proteins that bind the small ribosomal subunit and these initiation factors hold the mRNA in place. In cap-dependent translation initiation, the ribosome scans the 5′ UTR of the mRNA in the 5′ → 3′ direction to find a suitable **initiation codon**, an AUG that is found within the Kozak consensus sequence 5′GCC**Pu**CC**AUG**G3′ (where Pu = purine). The most important determinants are the G at position +4 (immediately following the AUG codon) and the purine (preferably A) at –3 (three nucleotides upstream of the AUG codon).

When a suitable initation codon is identified, an initiating tRNA^{Met} with its attached methionine binds to the P site on the large ribosomal subunit so that its anticodon base-pairs with the AUG initiator codon on the mRNA (see **Figure 1.28**). Once this happens, the transcriptional reading frame is established and codons are interpreted as successive groups of three nucleotides continuing in the 5′ → 3′ direction downstream of the initiating AUG codon. An aminoacyl tRNA for the second codon (a tRNA^{Gly} to recognize GGG in the example of **Figure 1.28**) binds to the neighboring A site in the large subunit.

Once the P and A sites are occupied by aminoacyl tRNAs, the 28S rRNA within the large ribosomal subunit acts as a peptidyltransferase. (An RNA like this that works as an enzyme is said to be a **ribozyme**.) The 28S rRNA catalyzes formation of a peptide bond by a condensation reaction between the amino group of the amino acid held by the tRNA in the A site and the carboxyl group of the methionine held by the tRNA^{Met}. The net result is to detach the initiator methionine from its tRNA and attach it to the second amino acid, forming a dipeptide (see **Figure 1.28**). Now without any attached amino acid, the tRNA^{Met} migrates away from the P site and its place is taken by the tRNA with the attached dipeptide that formerly occupied the A site. The liberated A site is now filled by an aminoacyl tRNA carrying an anticodon that is complementary to the third codon, and a new peptide bond is formed to make a tripeptide, and so on.

After a ribosome initiates translation of an mRNA and then moves along the mRNA, other ribosomes can engage with the same mRNA. The resulting polyribosome structures (polysomes) make multiple copies of a polypeptide from the one mRNA molecule. Polypeptide chain elongation occurs until a **termination codon** is met. In the case of mRNA transcribed from nuclear genes, termination codons come in three varieties: UAA (ochre); UAG (amber); and UGA (opal), but there are some differences in the case of mitochondrial mRNA as described in the next section.

In response to a termination codon, a protein release factor enters the A site instead of an aminoacyl tRNA to signal that the polypeptide should disengage from the ribosome.

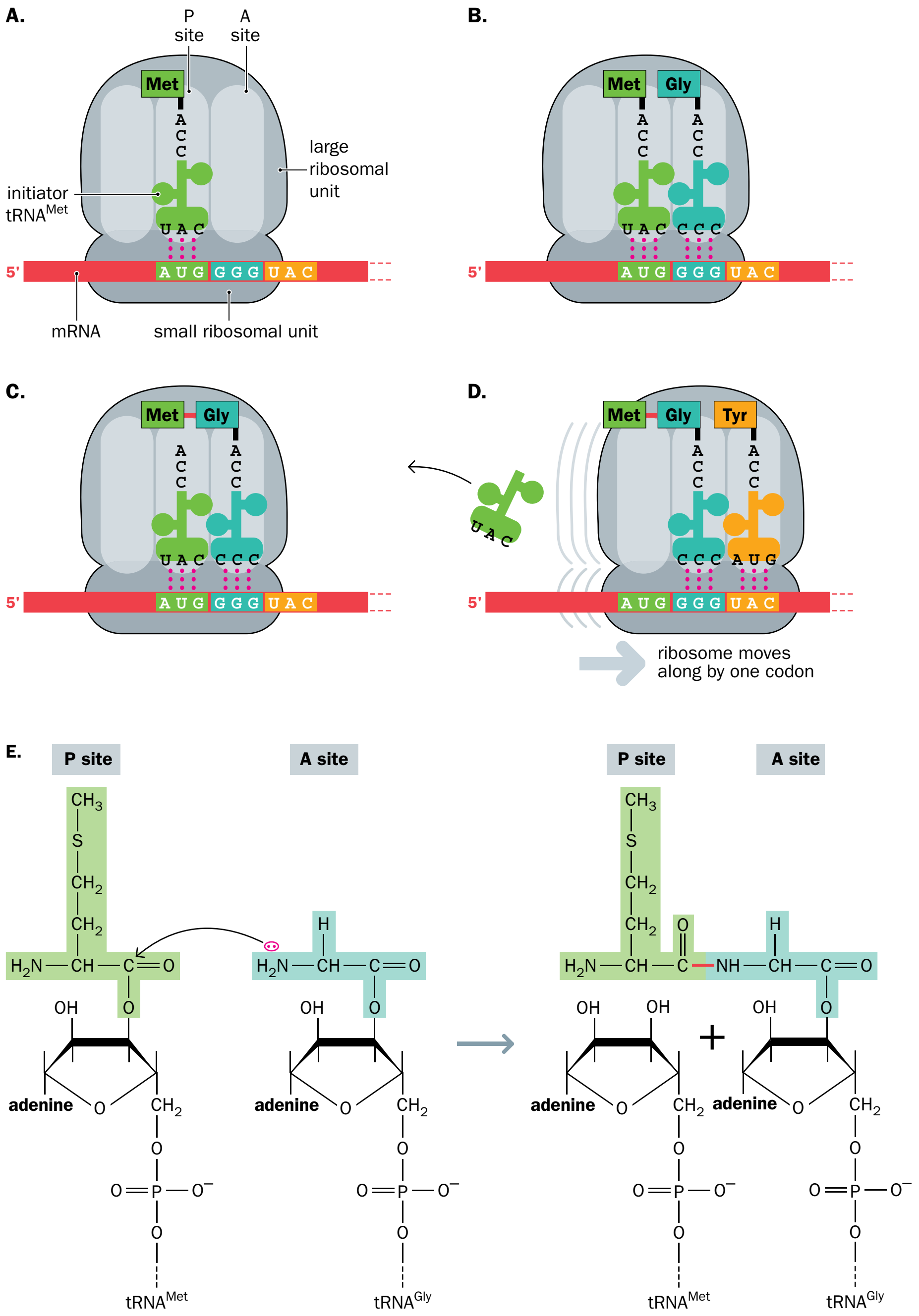


Figure 1.28 The genetic code is deciphered on ribosomes by codon–anticodon recognition. (**A**) The large subunit (60S in eukaryotes) has two sites for binding an aminoacyl tRNA (a transfer RNA with its attached amino acid): the P (peptidyl) site and the A (aminoacyl) site. The small ribosomal subunit (40S in eukaryotes) binds mRNA, which is scanned along its 5′ UTR in a 5′ → 3′ direction until the start codon is identified, an AUG located within a larger consensus sequence (see text). An initiator tRNA^{Met} carrying a methionine residue binds to the P site with its anticodon in register with the AUG start codon. (**B**) The appropriate aminoacyl tRNA is bound to the A site with its anticodon base pairing with the next codon (GGG in this case, specifying glycine). (**C**) The rRNA in the large subunit catalyzes peptide bond formation, resulting in the methionine detaching from its tRNA and being bound instead to the glycine attached to the tRNA held at the A site. (**D**) The ribosome translocates along the mRNA so that the tRNA bearing the Met-Gly dipeptide is bound by the P site. The next aminoacyl tRNA (here, carrying Tyr) binds to the A site in preparation for new peptide bond formation. (**E**) Peptide bond formation. The N atom of the amino group of the amino acid bound to the tRNA in the A site makes a nucleophilic attack on the carboxyl C atom of the amino acid held by the tRNA bound to the P site.

The completed polypeptide will then undergo processing that can include cleavage and modification of the side chains. Its backbone will have a free amino group at one end (the N-terminal end) and a free carboxyl group at the other end (C-terminal end).

The genetic code is degenerate and not quite universal

The genetic code is a three-letter code, and there are four possible bases to choose from at each of the three base positions in a codon. There are therefore $4^3 = 64$ possible codons, which is more than sufficient to encode the 20 major types of amino acid. The genetic code is degenerate because, on average, each amino acid is specified by about three different codons. Some amino acids (such as leucine, serine, and arginine) are specified by as many as six codons; others are much more poorly represented (**Figure 1.29**). The degeneracy of the genetic code most often involves the third base of the codon.

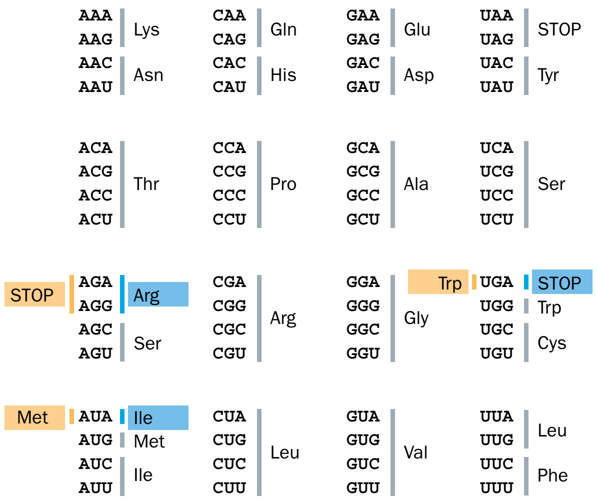


Figure 1.29 The genetic code. Pale gray bars to the right of the codons identify the 60 codons that are interpreted in the same way for mRNA from genes in our nuclear and mitochondrial DNA (mtDNA). Four codons, AGA, AGG, AUA, and UGA are interpreted differently. They are flanked on the right by pale blue bars showing the interpretation for nuclear genes, and on the left by pale orange bars showing how they are interpreted for genes in mtDNA. For nuclear genes, the "universal" genetic code has 61 codons that specify 20 different amino acids, with different levels of redundancy from unique codons (Met, Trp) at one extreme to as much as six-fold redundancy (Arg, Leu, Ser). The remaining three codons, UAA, UAG, and UGA normally act as stop codons (according to the surrounding sequence context; however, UGA can occasionally specify a 21st amino acid, selenocysteine, and UAG can occasionally specify glutamine). For genes in mtDNA, 60 codons specify an amino acid and there are four stop codons (AGA, AGG, UAA, and UAG).

Although over 60 codons can specify an amino acid, the number of different cytoplasmic tRNA molecules is quite a bit less and only 22 types of mitochondrial tRNA are made. The interpretation of over 60 sense codons with a much smaller number of different tRNAs is possible because base pairing in RNA is more flexible than in DNA. Pairing of codon and anticodon follows the normal A-U and G-C rules for the first two base positions in a codon. However, at the third position there is some flexibility (base wobble) and G-U base pairs are tolerated here (**Table 1.5**).

The genetic code is the same for virtually all prokaryotes and the nuclear genomes of eukaryotes. However, mitochondria and chloroplasts have a very limited capacity for protein synthesis and during evolution their genetic codes have diverged a little from that used at cytoplasmic ribosomes. Translation of nuclear-encoded mRNA continues until one of three stop codons is encountered (UAA, UAG, or UGA) but, in mammalian mitochondria, there are four possibilities (UAA, UAG, AGA, and AGG; see **Figure 1.29**).

The meaning of a codon can also be dependent upon the sequence context; that is, the nature of the nucleotide sequence in which it is embedded. Depending on the surrounding sequence, some codons in a few types of nuclear-encoded mRNA can be interpreted differently to normal. For example, in a wide variety of cells the stop codon UGA in some nuclear-encoded mRNAs is decoded to give the rare amino acid selenocysteine, and UAG can sometimes be interpreted to give glutamine.

TABLE 1.5 RULES FOR BASE PAIRING CAN BE RELAXED (WOBBLE) AT THE THIRD POSITION OF A CODON

Base at 5′ end of tRNA anticodon	Base recognized at 3′ end of mRNA codon
A	U only
C	G only
G (or I)	C or U
U	A or G

I = inosine, a deaminated form of guanosine (see **Figure 1.25** for the structure).

Post-translational processing: chemical modification of amino acids and polypeptide cleavage

Polypeptides frequently undergo a variety of enzymatic chemical modifications, during or after translation. The modifications involve covalent attachment of simple or complex chemical groups, and they may be reversible (as a way of changing the behavior of the protein toward different outcomes) or be irreversible. The dbPTM database provides a searchable database of protein translation modifications.

Many proteins have one or more simple chemical groups attached to specific amino acids. In addition, more complex chemical groups are covalently attached to the polypeptide backbone of certain proteins. In the latter case, large groups may be added irreversibly: carbohydrates to secreted proteins, and lipids and glycolipids to many membrane proteins. In other cases, ADP-ribose units and certain types of proteins can be reversibly attached to proteins to regulate them in some way (**Table 1.6**).

Another type of post-translational processing involves specific cleavage of a precursor polypeptide to yield one or more active polypeptide products.

TABLE 1.6 MAJOR TYPES OF CHEMICAL MODIFICATION IN PROTEINS

Type of modification (group added)		Target amino acid(s)	Notes
SIMPLE CHEMICAL GROUPS			
Phosphorylation (PO_4^-)		Tyr, Ser, Thr	Achieved by specific kinases; reversed by phosphatases
Methylation (CH_3)		Lys	Achieved by methylases; reversed by demethylases
Hydroxylation (OH)		Pro, Lys, Asp	Hydroxyproline (Hyp) and hydroxylysine (Hyl) are particularly common in collagens
Acetylation (CH_3CO)		Lys	Achieved by an acetylase; reversed by deacetylase
Carboxylation (COOH)		Glu	Achieved by γ-carboxylase
COMPLEX CHEMICAL GROUPS			
Carbohydrate	*N*-glycosylation (monosaccharides to complex carbohydrates)	Asn[a]	Takes place initially in the endoplasmic reticulum, with later additional changes occurring in the Golgi apparatus
	O-glycosylation (sugars and complex carbohydrates)	Ser, Thr, Hyl[b]	Takes place in the Golgi apparatus; less common than *N*-glycosylation
Glycolipid	Attachment of glycosylphosphatidylinositol	Asp[c]	Anchors protein to outer layer of plasma membrane (see **Figure 1.31A**)
Lipid	Myristoylation (C_{14} fatty acyl)	Gly[d]	Serve as membrane anchors
	Palmitoylation (C_{16} fatty acyl)	Cys[e]	
	Farnesylation (C_{15} prenyl)	Cys[c]	
	Geranylgeranylation (C_{20} prenyl)	Cys[c]	
Protein	Sumoylation (SUMO, or small ubiquitin-like modifier protein)	Lys	SUMO proteins are ~100 amino acids long and mature by cleavage of four amino acids to leave a C-terminal glycine that is joined by an isopeptide bond to the protein target
	Ubiquitylation (ubiquitin)	Lys	Ubiquitins are highly-conserved, 76-amino-acid-long proteins. A C-terminal glycine forms the isopeptide bond with the terminal amino group of a lysine side chain of the target protein
Other	ADP-ribosylation and poly (ADP-ribosyl)ation	Arg	ADP-ribosyl groups are donated by NAD^+ (nicotine adenine dinucleotide). Sometimes multiple ADP-ribosyl groups are combined to form a complex poly(ADP-ribose) structure (see **Figure 1.31B**)

[a] Especially common when Asn is in the sequence Asn-X-(Ser/Thr), where X is any amino acid other than Pro.
[b] Hydroxylysine.
[c] At C-terminus of polypeptide.
[d] At N-terminus of polypeptide.
[e] To form *S*-palmitoyl link.

Addition of simple chemical groups

The characteristics and functions of proteins are often regulated by reversible attachment of very simple chemical groups, such as phosphate, methyl, acetyl, hydroxyl, and carboxyl groups. Small armies of enzymes are needed to add or remove these groups from specific proteins, or protein classes, such as designated kinases, methylases, and acetylases to add phosphate, methyl, and acetyl groups, respectively, and phosphatases, demethylases, and deacetylases to remove them.

These modifications are essential for diverse cell functions including regulation of chromatin structure, transcription, cell signaling, and so on, as described in later chapters, and key proteins are known to undergo a wide range of post-translational modifications (**Figure 1.30**).

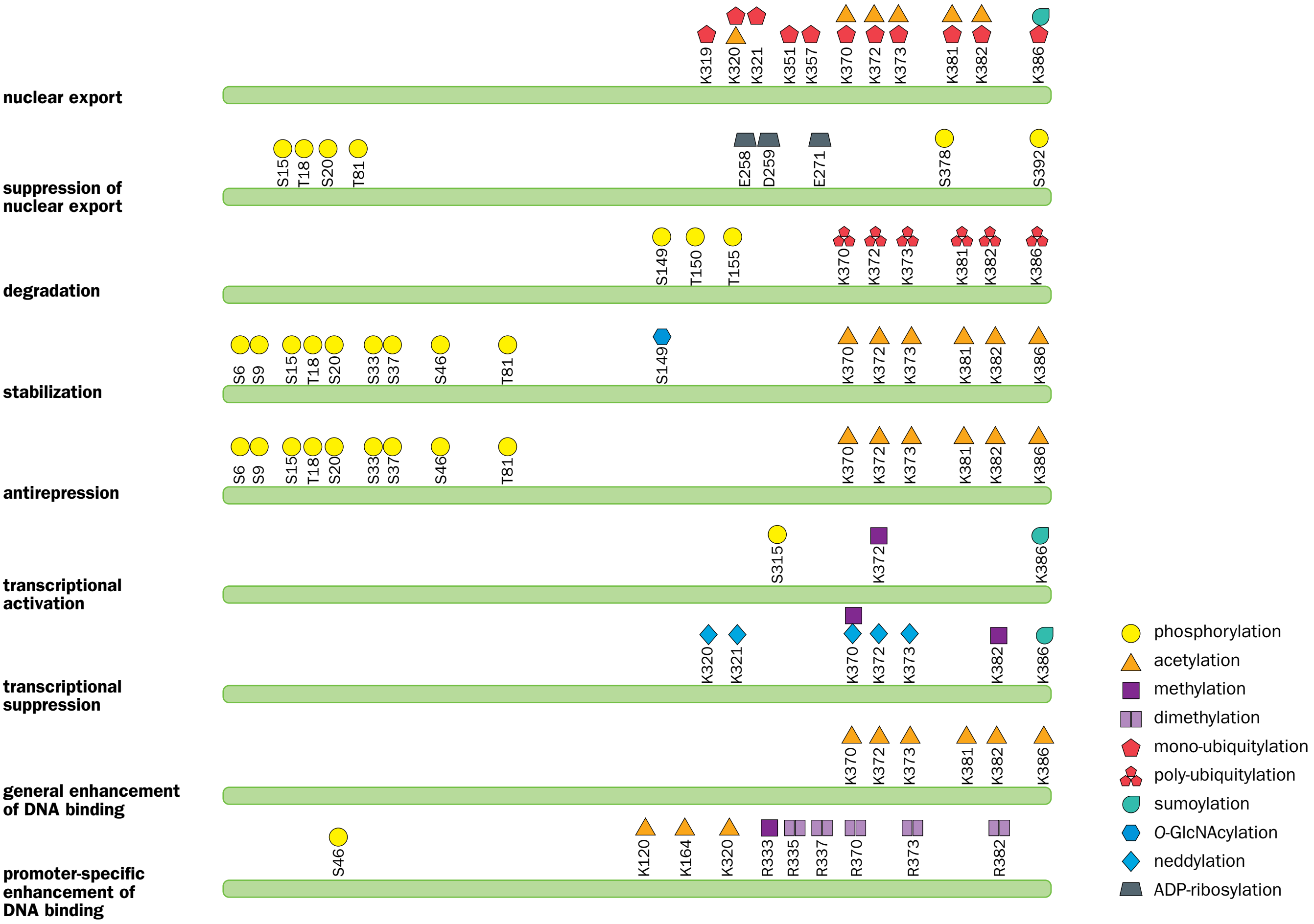


Figure 1.30 Post-translational modifications in the p53 protein that are known to be responsible for specific changes in its behaviour. The p53 protein is known to undergo many different post-translational modifications, but only the ones known to be directly responsible for the biological effects listed to the left are shown here. GlcNAc, *N*-acetylglucosamine. (Adapted from Gu B & Zhu WG [2012] *Int J Biol Sci* **8**:672–684; PMID 22606048. With permission from Ivyspring International Publisher.)

Addition of carbohydrate groups

Glycoproteins have oligosaccharides covalently attached to the side chains of certain amino acids. Few proteins in the cytosol are glycosylated (carry an attached carbohydrate); if they are, they have a single sugar residue, *N*-acetylglucosamine, attached to a serine or threonine residue. However, proteins that are secreted from cells or transported to lysosomes, the Golgi apparatus, or the plasma membrane are routinely glycosylated. In these cases, the sugars are assembled as oligosaccharides before being attached to the protein.

Two major types of glycosylation are recognized. Carbohydrate *N*-glycosylation involves attaching a carbohydrate group to the nitrogen atom of an asparagine side chain and *O*-glycosylation entails adding a carbohydrate to the oxygen atom of an OH group carried by the side chains of certain amino acids (see **Table 1.6**). Some modifications simply involve adding monosaccharides or disaccharides, but in others more complex carbohydrates are attached that may be linear or branched polymers.

Proteoglycans are proteins with attached glycosaminoglycans (polysaccharides) that usually include repeating disaccharide units containing glucosamine or galactosamine. The best-characterized proteoglycans are components of the extracellular matrix, a complex network of macromolecules secreted by, and surrounding, cells in tissues or in culture systems.

Addition of lipid and glycolipid groups

Some proteins, notably membrane proteins, are modified by the addition of fatty acyl or prenyl groups. The added groups typically serve as membrane anchors, hydrophobic amino acid sequences that secure a newly synthesized protein within either a plasma membrane or the endoplasmic reticulum (see **Table 1.6**).

Anchoring of a protein to the outer layer of the plasma membrane often involves attaching a glycosylphosphatidylinositol (GPI) group (**Figure 1.31A**). This glycolipid group contains fatty acyl groups that serve as the membrane anchor; they are linked successively to a glycerophosphate unit, an oligosaccharide unit, and finally through a phosphoethanolamine unit to the C-terminus of the protein. The entire protein, except for the GPI anchor, is located in the extracellular space.

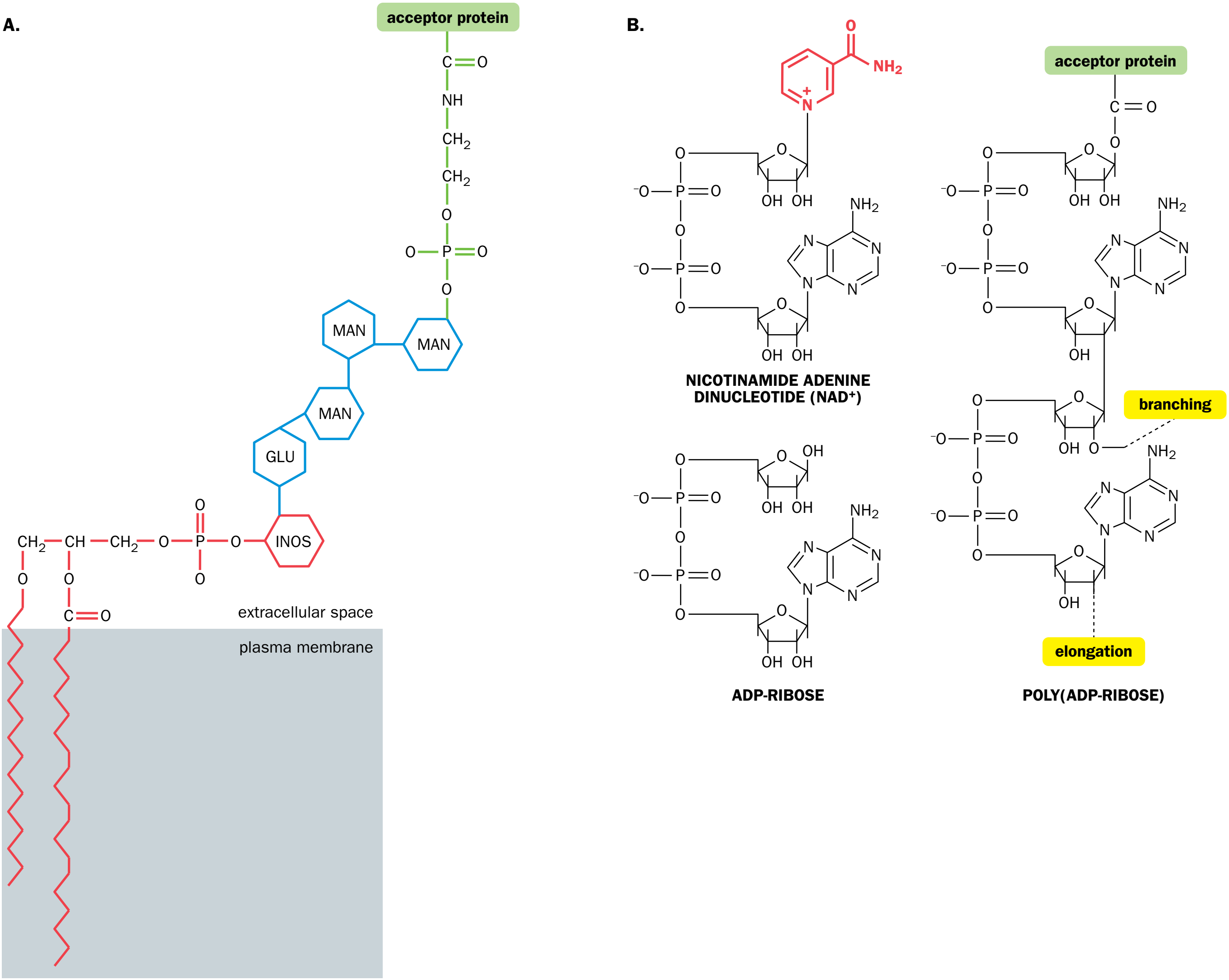


Figure 1.31 Examples of protein modification by attachment of complex groups. (**A**) Addition of glycosylphosphatidylinositol (GPI) to anchor a protein to the surface of the plasma membrane. The C-terminal carboxyl group of the protein is attached by an ethanolamine group (green) to a glycan group (blue), based on mannose (MAN) and glucose (GLU), that is linked in turn to a phosphatidylinositol group (red). The latter has long acyl side groups that insert into the plasma membrane to provide anchorage. (**B**) Addition of ADP-ribose units. The figure shows the chemical structures of nicotinamide, nicotinamide adenine dinucleotide (NAD$^+$), and poly(ADP-ribose). Addition of ADP-ribose units is achieved using NAD$^+$ as a donor of ADP-ribose. Poly(ADP-ribose) is a branched polymer synthesized on acceptor proteins by poly(ADP-ribose) polymerases (PARPs) using NAD$^+$ as a donor of ADP-ribose units. (A, modified from Rosse WF & Ware RE [1995] *Blood* **86**:3277–3286; PMID 7579428; B, adapted from Luo X & Kraus WL [2012] *Genes Dev* **26**:417–432; PMID 22391446. With permission from Cold Spring Harbor Laboratory Press.)

Addition of proteins

Some proteins are regulated by attachment of specialized proteins. Sumoylation is a protein modification in which SUMO (**s**mall **u**biquitin-like **mo**difier) proteins are reversibly attached to proteins. Sumoylation induces the proteins to change their behavior (via altered binding properties, change of localization within cells, and so on), and is important in regulating certain processes, such as transcription, chromatin structure, DNA repair, protein stability, transport between the nucleus and cytoplasm, and so on.

Ubiquitin proteins resemble SUMO proteins and can also be reversibly attached to proteins. A major purpose of this modification is to target proteins for destruction (when the protein needs to be replaced, or is a threat to the cell). Adding a small chain of ubiquitin residues (polyubiquitin) to a protein marks that protein for proteolytic degradation in the proteasome, whereupon the polyubiquitin is recycled as single ubiquitin units. In other cases, a single ubiquitin residue can be attached to a protein and this type of ubiquitin modification can have different regulatory roles.

ADP-ribosylation

Another reversible protein modification involves the enzymatic addition of ADP-ribose units, which are donated by nicotinamide adenine dinucleotide (NAD^+). Mono-ADP-ribosyltransferases catalyze the addition of a single ADP-ribose unit from NAD^+ to an arginine side chain of the target protein. Poly(ADP-ribose) polymerases (PARPs) catalyze the addition of multiple ADP-ribose units provided by donor NAD^+ molecules (Figure 1.31B).

Post-translational cleavage

The primary translation product may also undergo internal cleavage to generate a smaller mature product. Occasionally the initiating methionine is cleaved from the primary translation product, as during the synthesis of β-globin (see Figure 1.27). Secreted proteins (plasma proteins, polypeptide hormones, neuropeptides, growth factors, and so on) are typically produced with a short N-terminal **signal sequence** that serves as a destination tag and is cleaved prior to export from the cell. Post-translational cleavage is sometimes also employed to generate active polypeptides from a larger polypeptide precursor (see the example of insulin production in Figure 1.32).

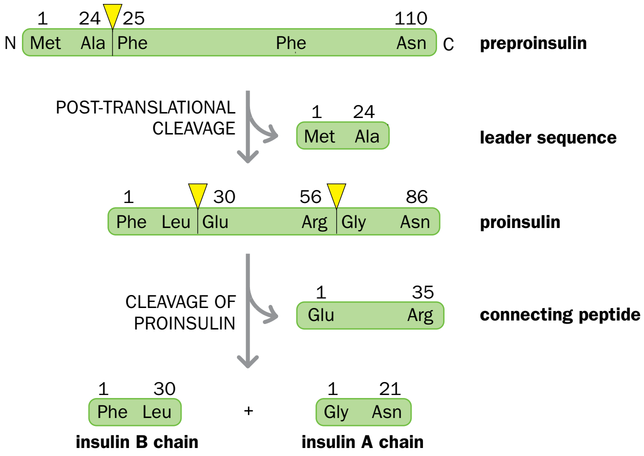


Figure 1.32 Insulin synthesis involves multiple post-translational cleavages of polypeptide precursors. Human insulin mRNA is translated to give a 110-amino-acid (aa) preproinsulin that has a 24-aa N-terminal leader sequence, an address tag required for the protein to be exported from the cell. In processing, the leader sequence is cleaved off and discarded. The remaining 86-aa proinsulin precursor contains a central segment (the connecting peptide) that may maintain the conformation of the A and B chains of insulin in readiness for making the final insulin protein. At the last moment, the connecting peptide is excised and the A and B chains are then covalently bonded together by disulfide bridges as illustrated in **Figure 1.35**.

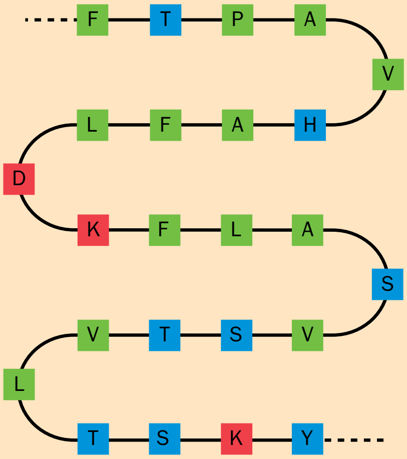
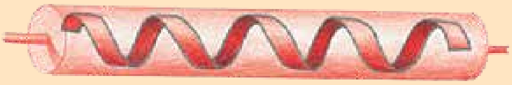
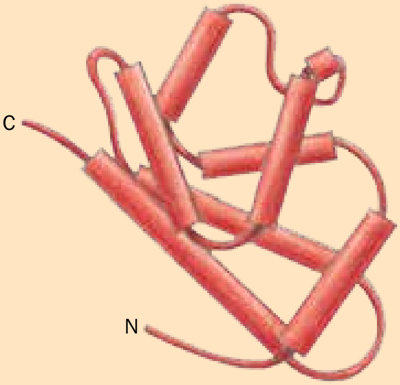
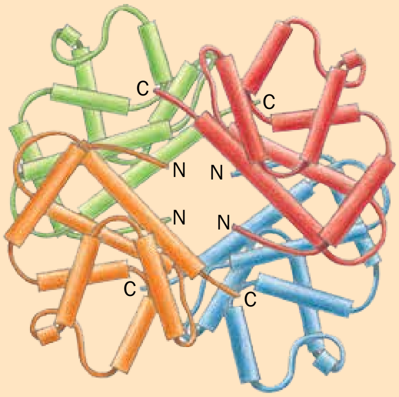
The complex relationship between amino acid sequence and protein structure

Proteins can be composed of one or more polypeptides, each of which may be subject to post-translational modification. Interactions between a protein and either of the following may substantially alter the conformation of that protein:

- A **co-factor**, such as a divalent cation (like Ca^{2+}, Fe^{2+}, Cu^{2+}, and Zn^{2+}), or a small molecule required for functional enzyme activity, for example NAD^+);
- A **ligand** (any molecule that a protein binds specifically).

Four different levels of structural organization in proteins have been distinguished and defined (Table 1.7).

TABLE 1.7 LEVELS OF PROTEIN STRUCTURE

Level	Definition	Notes
Primary  F T P A V L F A H D K F L A S V T S V L T S K Y	The linear sequence of amino acids in a polypeptide	Can vary enormously in length from a few to thousands of amino acids
Secondary 	The path that a polypeptide backbone follows within local regions of the primary structure	Varies along the length of the polypeptide; common elements of secondary structure include the α-helix and β-pleated sheet
Tertiary  	The overall three-dimensional structure of a polypeptide, arising from the combination of all of the secondary structures	Can take various forms (globular, rodlike, tube, coil, sheet)
Quaternary  	The aggregate structure of a multimeric protein (comprising >1 subunit, which may be of more than one type)	Can be stabilized by disulfide bridges between subunits or ligand binding, and other factors

Even within a single polypeptide there is ample scope for hydrogen bonding between different amino acid residues. This stabilizes the partial polar charges along the backbone of the polypeptide and has profound effects on that protein's overall shape. With regard to a protein's conformation, the most significant hydrogen bonds are those that occur between the oxygen of one peptide bond's carbonyl (CO) group and the hydrogen of the amino (NH) group of another peptide bond. Several fundamental structural patterns (motifs) stabilized by hydrogen bonding within a single polypeptide have been identified, the most fundamental of which are described below.

The α-helix

This is a rigid cylinder that is stabilized by hydrogen bonding between the carbonyl oxygen of a peptide bond and the hydrogen atom of the amino nitrogen of a peptide bond located four amino acids away (**Figure 1.33**). α-Helices often occur in proteins that perform key cellular functions (such as transcription factors, where they are usually represented in the DNA-binding domains). Identical α-helices with a repeating arrangement of nonpolar side chains can coil round each other to form a particularly stable coiled coil. Coiled coils occur in many fibrous proteins, such as collagen of the extracellular matrix, the muscle protein tropomyosin, α-keratin in hair, and fibrinogen in blood clots.

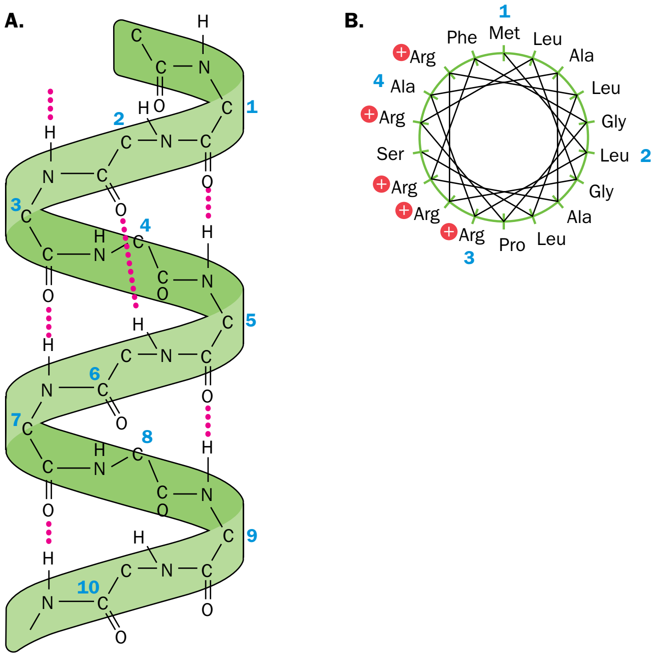


Figure 1.33 The structure of a standard α-helix and an amphipathic α-helix. (**A**) The structure of an α-helix is stabilized by hydrogen bonding between the oxygen of the carbonyl group (C=O) of each peptide bond and the hydrogen on the peptide bond amide group (NH) of the fourth amino acid away, making the helix have 3.6 amino acids per turn. The side chains of each amino acid are located on the outside of the helix; there is almost no free space within the helix. Note: only the backbone of the polypeptide is shown and some bonds have been omitted for clarity. (**B**) An amphipathic α-helix has tighter packing and has charged amino acids and hydrophobic amino acids located on different surfaces.

The β-sheet

β-Sheets (also called β-pleated sheets) are also stabilized by hydrogen bonding but, in this case, the bonds occur between opposed peptide bonds in parallel or antiparallel segments of the same polypeptide chain (**Figure 1.34A**). β-Sheets occur, often together with α-helices, at the core of most globular proteins, and can form complex structures such as the β-barrel (**Figure 1.34B**).

The β-turn

Hydrogen bonding can occur between amino acids that are even nearer to each other within a polypeptide. When this arises between the peptide bond CO group of one amino acid residue and the peptide bond NH group of an amino acid residue three places farther along, this results in a hairpin β-turn. Abrupt changes in the direction of a polypeptide enable compact globular shapes to be achieved. These β-turns can connect parallel or antiparallel strands in β-pleated sheets.

Higher-order structures

Many more complex structural motifs, consisting of combinations of the above structural modules, form **protein domains**. Such domains are often crucial to a protein's overall shape and stability and often represent functional units involved in binding

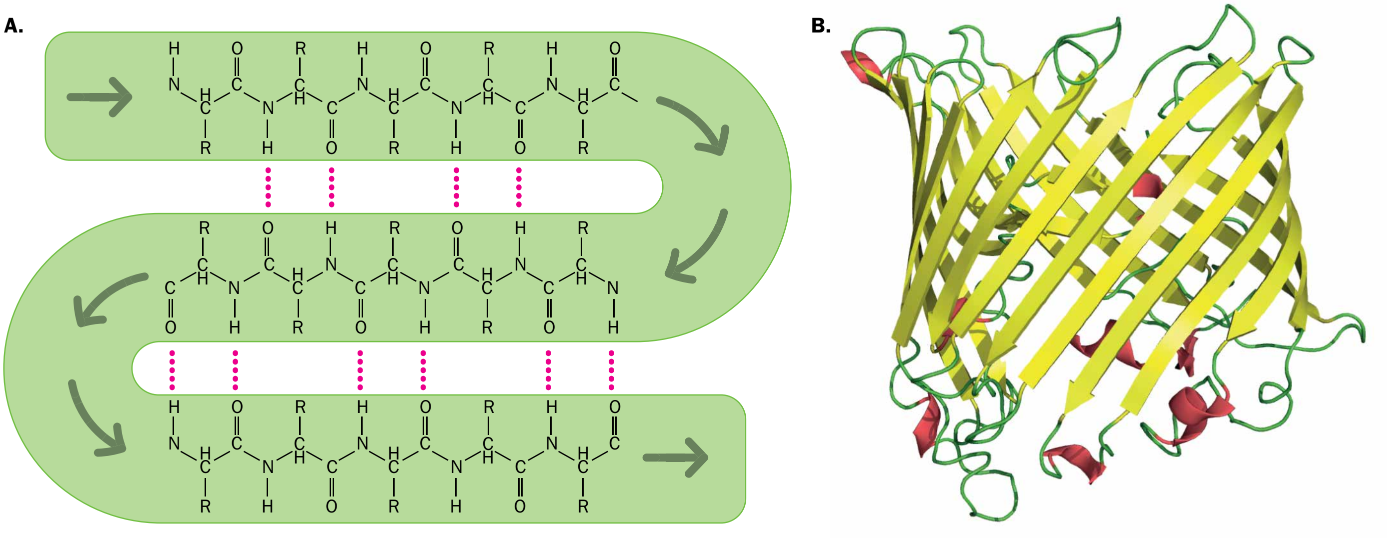


Figure 1.34 **The structure of β-sheets and β-barrels.** (**A**) In a β-sheet (also called a β-pleated sheet), hydrogen bonding occurs between the carbonyl oxygens and amide hydrogens on adjacent segments of a sheet that may be composed either of parallel segments of the polypeptide chain or, as shown here, of antiparallel segments (arrows mark the direction of travel from N-terminus to C-terminus). (**B**) A β-barrel is a large β-sheet that forms a closed structure in which the first β-strand is hydrogen bonded to the last (the β-strands are typically arranged in an antiparallel arrangement). The barrel structure provides an insulating internal environment and is often found in proteins that span the hydrophobic cell membrane (allowing passage of small molecules that are charged or polar), and in proteins that bind hydrophobic ligands (in the center of the barrel). This example shows a side view of a single monomer of a sucrose porin protein from *Salmonella typhimurium* that facilitates transfer of the polar sucrose molecule (which cannot simply diffuse through the hydrophobic cell membrane). (Created from PDB 1D 1A0S using PyMol by Opabinia regalis and reproduced under the Creative Commons BY 3.0 license.)

other molecules. Another important determinant of the structure (and function) of a protein are **disulfide bridges**. They can form between the sulfur atoms of sulfhydryl (–SH) groups on two amino acids that may reside on a single polypeptide chain or on two polypeptide chains (**Figure 1.35**).

In general, the primary structure of a protein determines the set of secondary structures that, together, generates the protein's tertiary structure. Secondary structural motifs can be predicted from analysis of the primary structure, but the overall tertiary structure cannot easily be accurately predicted. Finally, some proteins form complex aggregates of polypeptide subunits, giving an arrangement known as the quaternary structure.

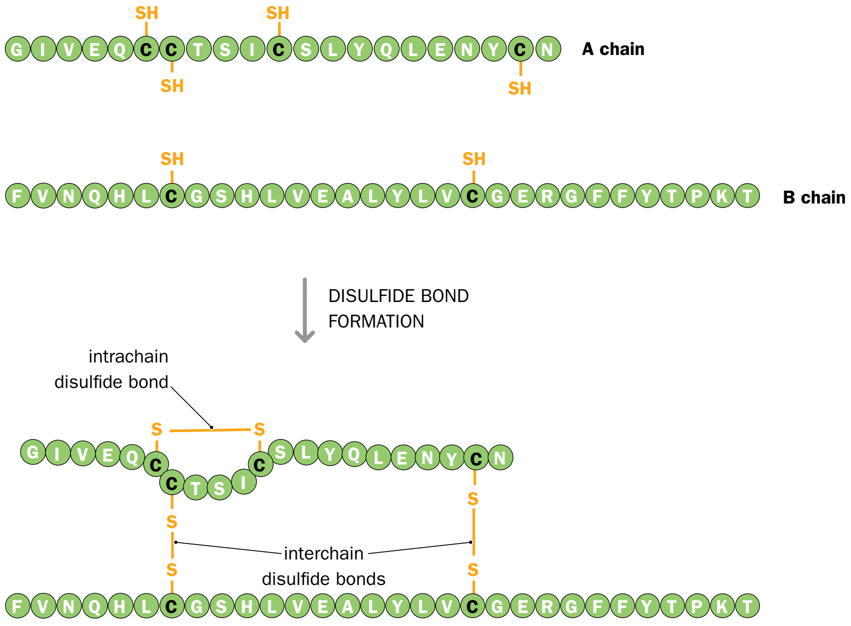

SH
SH
G I V E Q C C T S I C S L Y Q L E N Y C N
A chain
SH
SH
SH
SH
F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T
B chain
DISULFIDE BOND FORMATION
intrachain disulfide bond
S
S
G I V E Q C C T S I C S L Y Q L E N Y C N
S
S
S
S
interchain disulfide bonds
F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T

Figure 1.35 **Intrachain and interchain disulfide bridges in human insulin.** When the insulin A and B chains are first formed by cleavage (see **Figure 1.32**) the cysteine residues have a free sulfhydryl group (–SH), but because the chains have been held in close proximity, disulfide bridges (–S–S–) can form by a condensation reaction between the sulfhydryl groups. One disulfide bridge forms between residues 6 and 11 of the A chain. Two disulfide bridges hold the A and B chains physically together (connecting Cys-7 of the A chain to Cys-7 of the B chain, and Cys-20 of the A chain to Cys-19 of the B chain).

SUMMARY

- Biological nucleic acids come in two forms, DNA and RNA. In all cells and some viruses, DNA serves as the hereditary material. Cellular RNAs play different roles in gene expression and function, but in some viruses RNA is the genetic material.

- In cells, each DNA molecule consists of two long DNA strands that are polymers of nucleotide repeat units. Each nucleotide consists of a sugar (deoxyribose), a phosphate, and a nitrogenous base that is one of four types: adenine, cytosine, guanine, and thymine.

- A DNA strand has a sugar–phosphate backbone with the bases attached to the sugars. It is the sequence of the bases that determines the identity and genetic function of any DNA sequence.

- RNA molecules are also polymers of nucleotides but, unlike DNA, cellular RNAs are usually single-stranded. Just as for DNA, an RNA strand has a sugar–phosphate backbone with four types of base, but in RNA the sugar is ribose and the base uracil replaces thymine.

- DNA normally occurs as a double helix: the two DNA strands wrap round each other, stabilized by hydrogen bonds between opposed bases on the two DNA strands (base pairs).

- There are two types of highly-stable base pair in DNA: adenine bonds to thymine, and cytosine bonds to guanine. A DNA double-helix structure has highly-stable base pairs between the two DNA strands across their lengths, and the strands are said to have complementary sequences.

- Cellular RNAs are single-stranded and much more flexible than double-stranded DNA. Hydrogen bonding between bases on the same strand shapes RNA structure, and a large diversity of structures are seen in the RNA populations of a cell.

- To transmit genetic information as cells divide, the two DNA strands of each double helix must first be unwound. The DNA is then replicated by using the original DNA strands as templates to synthesize new, complementary DNA strands, forming two identical double helices that are then shared equally between the two daughter cells.

- Cellular genes are discrete segments of DNA that are used to make a functional RNA or polypeptide. In each case, there must be local unwinding of the two DNA strands so that an RNA polymerase can gain access to a sequence on one of the DNA strands, using it as a template to synthesize a functional, complementary RNA strand (transcription).

- Protein-coding genes make a coding RNA (messenger RNA) that in turn will serve as a template for making a polypeptide.

- Many other genes make functional noncoding RNAs that do not encode polypeptide; instead, they often serve to assist or regulate the expression of other genes.

- In eukaryotes, the great bulk of DNA (and genes) is found in the nucleus, but each mitochondrion and plant chloroplast contains a tiny amount of DNA with a very few genes.

- To become functional, a newly synthesized RNA must undergo a series of maturation steps, typically including cleavage steps that excise unwanted intervening sequences. Chemical modifications (of bases and/or the attached ribose sugar) are also common.

- Polypeptide synthesis occurs at ribosomes, either in the cytoplasm or inside mitochondria and chloroplasts.

- The sequential information encoded in a messenger RNA is interpreted at the ribosome via a triplet genetic code, determining the basic structure of the polypeptide.

- Chemically modified nucleotides are rare in the DNA of vertebrate cells (being predominantly restricted to methylation of the 5′ carbon of certain cytosines only). However, over 100 different types of modified nucleotide exist in RNA, some of which are known to be important for the structure, stability, or function of the RNA.

- Polypeptides are often formed by cleavage from larger precursors, and also undergo a wide variety of chemical modifications that regulate their behavior and/or are essential for their function. Regulation is most commonly carried out by reversible covalent attachment of simple chemical groups to specific amino acids in the protein.

- Addition of complex carbohydrates and lipid groups is necessary for the functions of many secreted proteins and membrane proteins. Sometimes small proteins and other complex groups are covalently attached to proteins to modify their behavior.

- Noncoding RNA molecules are mostly important in regulating gene expression, but proteins are the most important executors of cell functions and display extraordinary structural and functional diversity.

- Protein structure is determined by the linear sequence of amino acids, interactions between subunits (if any), and interactions with the environment. The conformation of a single polypeptide chain is largely dependent on hydrogen bonding between specific polar chemical groups on different amino acids of the chain.

FURTHER READING

Agris PF *et al*. (2007) tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol* **366**:1–13; PMID 17187822.

Alberts B *et al*. (2014) *Molecular Biology of the Cell*, 6th edn. Garland Science.

Big Picture Book of Viruses. http://www.virology.net/Big_Virology/BVHomePage.html

Calladine CR *et al*. (2004) *Understanding DNA. The Molecule and How It Works*, 3rd edn. Academic Press.

dbPTM. Database of protein post-translational modification. http://dbptm.mbc.nctu.edu.tw/

He Y & Ecker JR (2015) Non-CG methylation in the human genome. *Annu Rev Genomics Hum Genet* **16**:55–77; PMID 26077819.

Johansson E & Dixon N (2013) Replicative DNA polymerases. *Cold Spring Harb Perspect Biol* **5**:a012799; PMID 23732474.

Li S & Mason CE (2014) The pivotal regulatory landscape of RNA modifications. *Annu Rev Genomics Hum Genet* **15**:127–150; PMID 24898039.

Whitford D (2005) *Proteins. Structure and Function*. John Wiley & Sons.

2

Fundamentals of cells and chromosomes

The underlying structure and fundamental functions of DNA, replication and transcription, were introduced in the previous chapter. But our DNA functions in a context. The immensely long DNA molecules in the nucleus are complexed with a variety of structural and regulatory proteins, and structured into linear chromosomes. The DNA molecules in mitochondria are different: they are very short and circular, and have comparatively little protein attached to them.

The nucleus and mitochondria are two of many discrete bodies known as **organelles** that occur within eukaryotic cells. In this chapter, we explore eukaryotic cells, focusing on details of their structure, how they are related to prokaryotic cells, and how they evolved from ancestral prokaryotes. In addition, we consider the evolution of multicellularity and cell diversity. We do not, however, consider aspects of intracellular transport in this chapter: they will be covered in Chapter 8 when we cover genetic manipulation of human and animal cells.

All cells originate from other cells. This almost always happens by a process of cell division: a mother cell divides to give two, usually identical, daughter cells (the division is said to be symmetrical, but in some situations cell division is programmed to be asymmetrical, producing daughter cells that are different). In order to maintain cell size, a period of cell growth is required so that before the next round of cell division, the daughter cells have reached the same size as the mother cell that produced them (**Figure 2.1**).

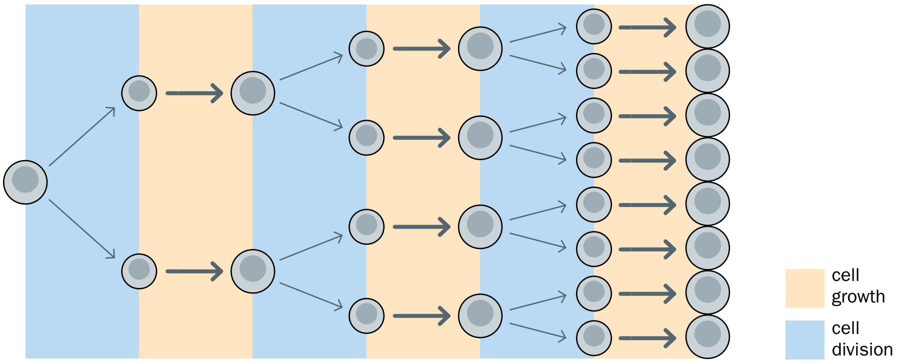


Figure 2.1 Cell proliferation involves alternating rounds of cell division and cell growth. In symmetrical cell division, when a mother cell divides, the two identical daughter cells will each have half the mass of the mother cell. They need some time to grow before the next round of cell division in order to attain the same size as the mother cell that produced them. Note that some cell divisions can be asymmetrical and produce daughter cells that differ in size and/or function, as in the case of some types of stem cell division (see **Figure 4.14**).

During development, successive cell divisions lead to an expansion of cell numbers (**cell proliferation**). However, when we reach maturity, cell proliferation is largely confined to cells of certain tissues where there is a need to renew cells on a regular basis (such as intestinal epithelial cells, skin cells, blood cells, and so on). There must be a balance between cell proliferation and cell death to ensure stability of cell number in adult organisms.

The process of cell division is a small component of the **cell cycle**. That involves replication of both nuclear DNA molecules in the chromosomes and mitochondrial DNA molecules, after which copies of each type of DNA molecule are divided (segregated) between the two daughter cells. There are important differences between how this occurs in routine cell division compared to the specialized reductive cell division that gives rise to sperm and egg cells.

A feature common to both types of cell division is the importance to the cell of chromosome condensation. This affects the expression of information encoded in the DNA and makes long and fragile DNA strands resilient to breakage during the dramatic rearrangements that occur in cell division.

Note that this chapter focuses on general aspects of eukaryotic chromosomes, and additional chromosome topics are covered later. In Chapter 10, we examine how chromosome structure relates to genetic and epigenetic regulation of mammalian gene expression. In Chapter 15, we explore the causes and detection of chromosome abnormalities, introducing structural features of human chromosomes, and how human chromosomes are studied.

2.1 CELL STRUCTURE AND DIVERSITY, AND CELL EVOLUTION

Cells are the cornerstone of biology, and for most biologists they are the defining characteristic of life forms (but viruses might also be considered life forms, according to how one defines life). Although cells show enormous diversity, varying hugely in size, function, and complexity, they all have the same basic form: an aqueous cytoplasm enclosed by a membrane, a lipid bilayer that contains diverse proteins. All cells have a common origin from a primordial cell.

Prokaryotes and eukaryotes represent a fundamental division of cellular life forms

According to differences in their internal organization and functions, cells can be classified into broad taxonomic groups. A major division, founded on fundamental differences in cell architecture, distinguishes prokaryotes (which are always unicellular) and eukaryotes (which may be unicellular or multicellular; **Figure 2.2**).

Prokaryotes have a simple internal organization, with a single, membrane-bound compartment that is not usually subdivided by any internal membranes. Under the electron microscope, prokaryotic cells typically appear relatively featureless. (However, prokaryotes are far from primitive: they have been through many more generations of evolution than humans.)

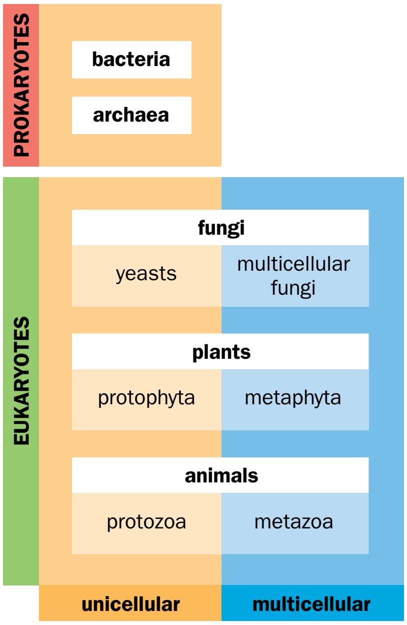


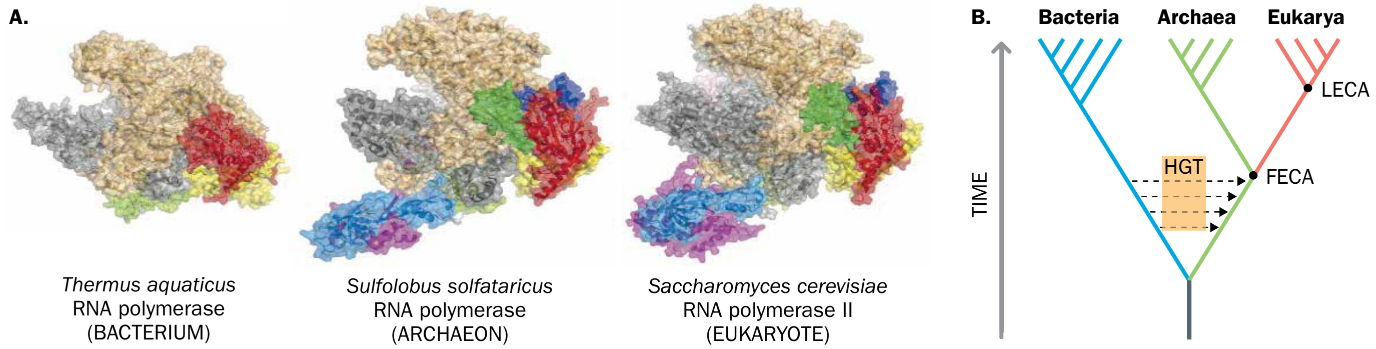
Figure 2.2 Classification of unicellular and multicellular organisms. The first two domains of life, bacteria and archaea, are also known as prokaryotes as they lack internal membranes. Eukaryotes, which form the third domain of life, have membrane-bound organelles. This domain is further subdivided into unicellular or multicellular fungi, plants, and animals. Unicellular eukaryotes are occasionally classified as protists, but the term protist is often used in a more limited sense to describe those unicellular eukaryotes that cannot easily be classified as animals, plants, or fungi.

Unlike in eukaryotes, the chromosomal DNA of prokaryotes is neither enclosed by a nuclear membrane nor highly structured; instead, it exists as a simple nucleoprotein complex, the nucleoid (see **Box 2.1 Figure 1**). The typical prokaryote has a single, circular chromosome containing a few Mb of DNA (usually from 1 to 10 Mb), but there are exceptions: some prokaryotes have two or three circular chromosomes, and a few have a linear chromosome, or a mix of linear and circular chromosomes.

Until quite recently, prokaryotes were simply considered to comprise diverse types of bacteria, but in 1977 phylogenetic studies by Carl Woese indicated that prokaryotes comprise two very different kingdoms or domains of life, as distinct from each other as they are from eukaryotes.

- Bacteria (formerly called eubacteria) are found in many environments. Some cause disease; others perform tasks that are useful or essential to human survival. Huge numbers of bacteria inhabit our bodies. Comprising about 500–1000 different species, the vast majority of such commensal bacteria live in the gut and many are beneficial, as explained below.
- Archaea (formerly called archaebacteria) are a poorly understood group of organisms that superficially resemble bacteria. They are often found in extreme environments, and different groups survive in extremes of heat, salt, and acidity, for example. Other species are found in more convivial locations, such as in the guts of cows and on our skin.

The division of prokaryotes into bacteria and archaea was first suggested by phylogenetic taxonomy studies that compared 16S rRNA (the highly-conserved RNA of the small ribosomal subunit) in different prokaryotes. Supportive evidence came, however, from many other areas, such as in the structure of RNA polymerase (**Figure 2.3A**). Large-scale protein comparisons have supported a comparatively close relationship between archaea and eukaryotes (**Figure 2.3B**).



Thermus aquaticus
RNA polymerase
(BACTERIUM)

Sulfolobus solfataricus
RNA polymerase
(ARCHAEON)

Saccharomyces cerevisiae
RNA polymerase II
(EUKARYOTE)

Figure 2.3 RNA polymerase structure and evolutionary relationships for the three domains of life. (**A**) RNA polymerase structure. Color coding identifies homologous subunits. Here, the archaeal RNA polymerase is closely related to the eukaryotic RNA polymerase and much more complex than the bacterial RNA polymerase. In the past, the greater complexity of eukaryotic RNA polymerases when compared to bacterial RNA polymerases was thought to reflect the greater complexity of transcriptional processes in eukaryotic cells. However, the close resemblance between archaeal and eukaryotic RNA polymerases points to a different explanation, while supporting generally close similarity in information processing systems in eukaryotes and archaea (see text). (**B**) A phylogenetic tree illustrating that eukaryotes arose by periodic horizontal gene transfers (HGT) from bacterial progenitors to a lineage of archaeal progenitor cells (green line). The node labeled FECA, the first eukaryote common ancestor, marks eukaryogenesis, the point at which eukaryotes originated and diverged from the lineage leading to present-day archaea (the key event was endosymbiosis of an α-proteobacterium, as described in **Figures 2.5** and **2.6**). Prior to eukaryogenesis, various HGT events allowed transfer of genes from bacterial cells to the archaeal lineage leading to the FECA, which might have caused the archaeal ancestor to increase in complexity (as illustrated in **Figure 2.6** and accompanying text). LECA, last eukaryote common ancestor. (A, adapted from Werner F [2008] *Trends Microbiol* **16**:247–250; PMID 18468900. With permission from Elsevier.)

Eukaryotes are thought to have first appeared more than 2.1 billion years ago. They have a much more complex organization than their prokaryote counterparts, with many internal membranes and membrane-bound organelles. The membranes surrounding and dividing the cell are selectively permeable, regulating transport of a variety of ions and small molecules into and out of the cell and between compartments.

All eukaryotes belong to a single domain of life, the eukarya, but comprise both unicellular organisms and multicellular fungi, plants, and animals. Eukaryotic cells are distinguished by having a nucleus (containing most of the cell's DNA) plus many other organelles in the cytoplasm with diverse functions, including ribosomes, the protein synthesis factories (see **Box 2.1**). A eukaryotic cell differs from a prokaryotic cell, therefore, in having physical separation of transcription (within the nucleus) and translation (in the cytoplasm). Even the cytosol, the soluble portion of the cytoplasm, is highly organized. It has an internal scaffold of protein filaments, the **cytoskeleton**, that provides stability, generates the forces needed for movement and changes in cell shape, facilitates the intracellular transport of organelles, and allows communication between the cell and its environment.

Partitioning of the genome into different cell compartments in eukaryotes

Most of the DNA in a prokaryotic cell is accounted for by just one type of circular DNA molecule that is bound by a few proteins to form a **nucleoid** (sometimes also called a chromosome; a few types of small, extrachromosomal, circular DNA molecule, known as plasmids, may also be found). By comparison, the DNA of eukaryotic cells is much more complex: there are often many different DNA molecules, and they can be immensely long and have diverse functional capabilities. The **genome** (the collection of different DNA molecules) of a eukaryotic cell is partitioned between at least two types of organelle: a single nucleus and multiple mitochondria. (Plant cells and algae have additional DNA-containing chloroplasts).

BOX 2.1 INTRACELLULAR ORGANIZATION WITHIN ANIMAL CELLS

Eukaryotic cells have many different types of membrane-enclosed structures within their cytoplasm. Not all the structures listed below are present in every cell type; some human cells are so specialized to perform a single function that the nucleus and other organelles are discarded and the cells rely on pre-synthesized gene products.

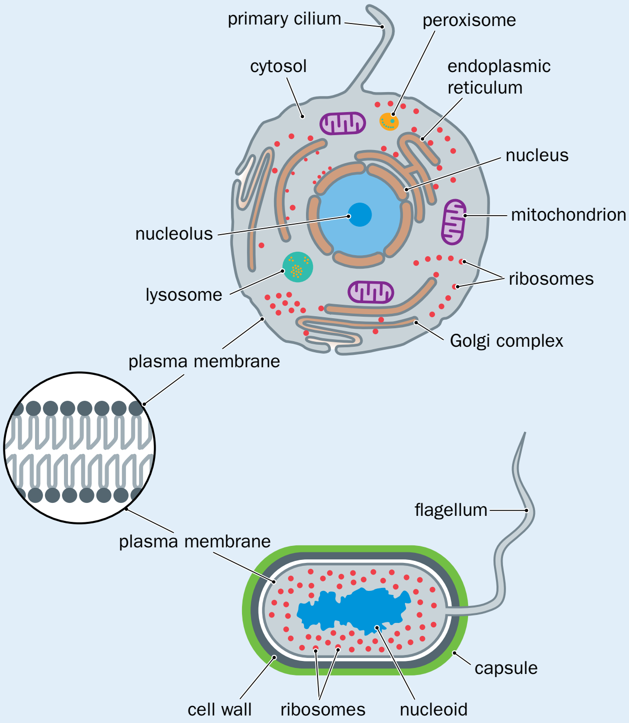
The **plasma membrane** provides a protective barrier around the cell. It is based on a double layer of phospholipids (**Figure 1**). Hydrophobic lipid "tails" are sandwiched between hydrophilic phosphate groups that are in contact with polar aqueous environments, the cytoplasm and extracellular environment. The plasma membrane is selectively permeable, regulating the transport of a variety of ions and small molecules into and out of the cell.

The **cytosol**, the aqueous component of the cytoplasm, makes up about half the volume of the cell and is the site of major metabolic activity, including most protein synthesis. The cytosol is very highly organized by a series of protein filaments, collectively called the **cytoskeleton**, that have a major role in cell movement, cell shape, and intracellular transport. There are three types of cytoskeletal filament:

- **Microfilaments** are polymers of the protein actin (and so are also known as actin filaments). The cytoskeletal network existing beneath the plasma membrane of most cells in the body is known as the **cell cortex** and is rich in actin filaments that are attached to the membrane in diverse ways. The cortex provides mechanical support to the cell but can be rapidly remodeled, allowing controlled changes to cell shape, such as **endocytosis** (the general process in which a portion of the plasma membrane invaginates to form a pit and then pinches off to form an endocytic vesicle, enclosing some of the extracellular fluid), and facilitating cell movement by forming transient filopodia and lamellipodia (extensions to the cell that allow it to crawl along surfaces).
- **Microtubules** are much more rigid than actin filaments and are polymers of tubulin proteins. They are important constituents of the centrosome and mitotic spindle and also form the core of cilia and flagella.
- **Intermediate filaments** have predominantly structural roles. Examples include the neurofilaments of nervous system cells and keratins in epithelial cells.

The **endoplasmic reticulum** (**ER**) consists of flattened, single-membrane vesicles whose inner compartments (cisternae) are interconnected to form channels throughout the cytoplasm. The ER has two key functions: the intracellular storage of Ca^{2+}, which is widely used in cell signaling; and the synthesis, folding, and modification of proteins and lipids destined for the cell membrane or for secretion. The **rough endoplasmic reticulum** is studded with ribosomes that synthesize proteins that will cross the membrane into the intracisternal space before being transported to the periphery of the cell. Here, they can be incorporated into the plasma membrane, retrieved to the ER, or secreted from the cell. The attachment of the sugar residues (glycosylation) that adorn many human proteins begins in the ER.

The **Golgi complex** consists of flattened, single-membrane vesicles, which are often stacked. Its primary function is to secrete cell products, such as proteins, to the exterior and to help form the plasma membrane and the membranes of lysosomes. Some of the small vesicles that arise peripherally by a pinching-off process (secretory vacuoles) contain secretory products. Glycoproteins arriving from the ER are further modified in the Golgi complex.



Box 2.1 Figure 1 Prokaryotic and eukaryotic cell anatomy. Prokaryotic cells are much smaller than eukaryotic cells and lack the internal organelles found in the latter. The eukaryotic cell shown at the top of this figure is a generic vertebrate cell.

The **nucleus** contains the chromosomes and the vast majority of the DNA of an animal cell. It is surrounded by a **nuclear envelope**, composed of two membranes, the outer one being studded with ribosomes. The membranes are separated by a narrow space and are continuous with the ER. Openings in the nuclear envelope (*nuclear pores*) are lined with specialized protein complexes that act as specific transporters of macromolecules between the nucleus and cytoplasm. Within the nucleus, the chromosomes are arranged in a highly-ordered way. Some evidence supports the existence of a *nuclear matrix*, or scaffold, a protein network to which chromosomes are attached. The nucleus also contains various suborganelles that lack membranes (see **Figure 2.4A** in the main text).

Mitochondria are sites of oxidative phosphorylation, generating ATP to power the different functions of a cell by oxidizing organic nutrients. They have a comparatively smooth outer membrane, and a complex, highly-folded inner mitochondrial membrane to which the mitochondrial nucleoid (mtDNA plus bound proteins) is attached. The inner compartment, the mitochondrial matrix, contains enzymes and chemical intermediates involved in energy metabolism. Mitochondria also have their own ribosomes that translate mRNA transcribed from mitochondrial DNA. Note that **Figure 1** gives the standard portrayal of mitochondria as separate organelles, but usually they form a highly connected and dynamic reticular network (see **Figures 2.4B** and **D** in the main text).

Peroxisomes (**microbodies**) are small, single-membrane vesicles containing enzymes that use molecular oxygen to oxidize their substrates and generate hydrogen peroxide.

Lysosomes are small, membrane-enclosed vesicles containing hydrolytic enzymes that digest materials brought into the cell by **phagocytosis** (absorption of solid objects) or pinocytosis (absorption of liquids). Lysosomes also help in the degradation of cell components after cell death.

Cilia are small structures containing microtubule filaments that extend from the plasma membrane and beat backward and forward, or rotate. In vertebrates, a very few specialized cell types use multiple cilia to generate movement. They include epithelial cells lining the lungs and oviduct, where the cilia beat together to move mucus away from the lungs or the egg toward the uterus. Most vertebrate cells, however, have a single cilium, known as the **primary cilium**, whose function is not involved in generating movement. Instead, it is packed with many different kinds of receptor molecule and acts as a sensor of the cell's environment. Sperm cells have a single, rather large and much longer version of a cilium, known as a flagellum. The flagellum moves in a whiplike fashion to propel the sperm cell forward.

Nuclear (chromosomal) DNA

Most of the DNA in a eukaryotic cell is present in the nucleus, distributed between multiple linear chromosomes. The soluble part of the nucleus, the nucleosol, has a nuclear matrix that contains a variety of subnuclear structures lacking membranes (**Figure 2.4A**). Pores in the nuclear envelope provide a regulated passageway for molecules to transit between the nucleus and cytoplasm.

Each chromosome is a single, very long, negatively-charged DNA molecule intricately packaged with positively-charged histone proteins (see below) and various other proteins. The number of different chromosomes (and associated DNA content) varies greatly between species. Because each type of chromosome in a cell contains its own type of DNA molecule there are many different linear DNA molecules in eukaryotic nuclei, each present in a very few copies (most of our cells, for example, have two copies of each chromosome and two copies of the associated DNA molecule).

Mitochondrial DNA

The remainder of the DNA is housed in the mitochondria. Mitochondria are often portrayed as static structures, such as in the image in **Box 2.1 Figure 1**, but in reality they seem to be part of a reticular network with long tubules, and there is a continuous process of division and fusion, even in resting cells (**Figure 2.4B**).

The mitochondrial DNA (mtDNA) differs from nuclear DNA in many respects. Thus, mtDNA molecules are comparatively small, circular DNA molecules and have much less bound protein. Additionally, there is only one type of mtDNA and it is present in many copies per cell (the copy number varies, but many human cells have several thousand copies of a mtDNA molecule).

Whereas chromosomes are units for segregating nuclear DNA, the unit of segregation for mtDNA is the **nucleoid**, a complex of from one to a few protein-bound mtDNA molecules that binds to the inner face of the inner mitochondrial membrane (**Figure 2.4C**). Nucleoids can be seen to be distributed along mitochondrial tubules in suitably stained cells (**Figure 2.4D**).

The extraordinary diversity of cells in the body

Although estimates are necessarily approximate, the adult human body is thought to have somewhere in the region of 10^{13} to 10^{14} cells (a recent calculation is provided in Bianconi E *et al*. [2013]; PMID 23829164). In addition to our body cells, we each host our own personal **microbiome**, a diverse mixture of microbial organisms and viruses with a total of more than 10 times as many cells as in our bodies! The bulk of these cells are located in our intestines, and many of the bacterial cells are beneficial: some ferment complex indigestible carbohydrates, for example, and others synthesize various vitamins upon which we depend, including folic acid, vitamin K, and biotin.

While the average diameter of eukaryotic cells is ~10–30 μm, some specialized cells can grow much larger. Mammalian egg cells are ~100 μm in diameter but other eggs that store nutrients required for development can be much larger, such as an ostrich egg. Some cells are very long. Human muscle-fiber cells can extend as long as 30 cm, and individual human neurons can reach up to 1 meter in length.

Complex animals have many highly-specialized cells, and histology textbooks recognize over 200 different cell types in adult humans. Histology is a comparatively crude way of classifying cells, however, relying heavily on differences in cell size, morphology, and ability to take up certain stains. These limitations, plus the recognition that some of our cells are difficult to access and study, means that cell diversity has been grossly underestimated.

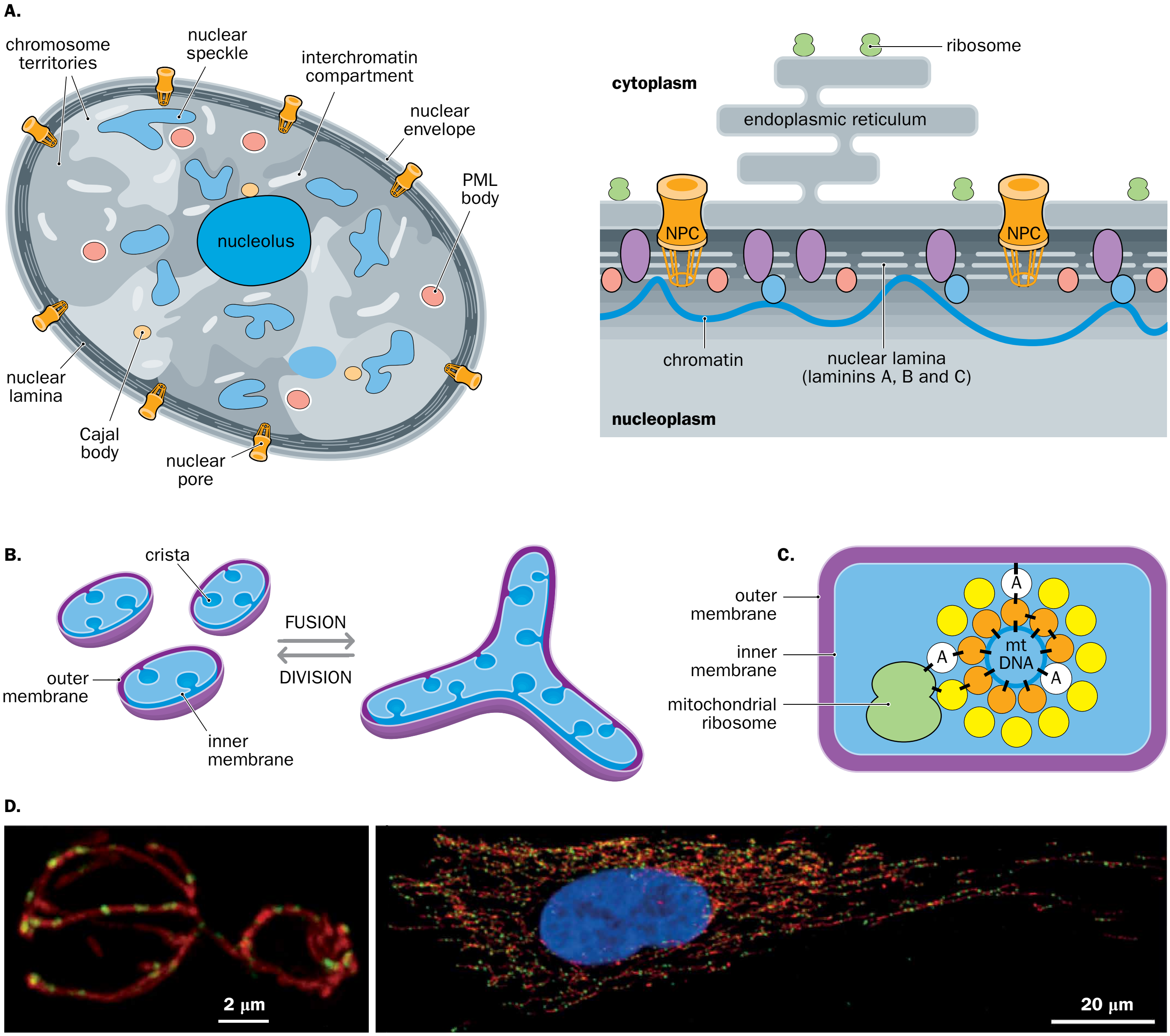


Figure 2.4 The structures of the two genome repositories of animal cells. (**A**) Nuclear structure (left) and details of nuclear lamina and nucleo-cytoplasmic connections (right). Individual chromosomes tend to occupy discrete chromosome territories in interphase nuclei (see **Figure 2.20**). Readily visible in each nucleus are one or a few nucleoli, regions where chromosomal segments containing rRNA genes are brought together to synthesize and process rRNA. Cajal bodies are sites for assembling small nuclear/nucleolar ribonucleoprotein (snRNP/snoRNP) particles. Fully mature snRNPs assemble at nuclear speckles in readiness for splicing pre-mRNA. PML bodies are composed predominantly of the premyelocytic leukemia protein and are thought to be involved in post-translational control and stress pathways (PMID 2972366). The nuclear lamina, consisting of intermediate filaments (lamins) and lamin-associated membrane proteins, lies on the inner surface of the inner nuclear membrane. It maintains nuclear stability, organizes chromatin, and binds nuclear pore complexes (NPC), nuclear envelope proteins (purple), and transcription factors (pink). Certain chromatin-associated proteins (blue) bind to nuclear envelope proteins. Note that the nuclear envelope is continuous with the endoplasmic reticulum (ER), and that the outer membrane of both of these is studded with ribosomes. (**B**) Mitochondria have a dynamic structure: they continually fuse and divide, and can develop tubular structures. Their inner membrane forms multiple folds known as cristae (*singular* crista). (**C**) Mitochondrial nucleoid organization. Highly-schematic representation of a nucleoid (which often contains multiple mtDNA copies rather than the single mtDNA shown here for clarity). Core proteins (shown in orange) are mtDNA-binding proteins that are dominated by the mtDNA packaging protein TFAM. Proteins shown in yellow cannot bind mtDNA directly, but bind to other proteins of the nucleoid. White circles A represent the ATAD3 protein that binds mtDNA directly, and also binds the major core protein to anchor the nucleoid to the inner membrane and to mitochondrial ribosomes. (**D**) Distribution of nucleoids (green) within tubular mitochondria (red) of a yeast cell (left) and a human fibroblast with its nucleus stained in blue (right). (A [left], reprinted from Lanctô C *et al*. [2007] *Nat Rev Genet* **8**:104–115; PMID 17230197. With permission from Springer Nature. Copyright © 2007; B and D [left], adapted from Friedman JR & Nunnari J [2014] *Nature* **505**:335–343; PMID 24429632. With permission from Springer Nature. Copyright © 2014; C, adapted from Gilkerson R *et al*. [2013] *Cold Spring Harb Perspect Biol* **5**:a011080; PMID 23637282. With permission from Cold Spring Harbor Laboratory Press; D [right], from Kukat C *et al*. [2011] *Proc Natl Acad Sci USA* **108**:13534–13539; PMID 21808029. With permission from the National Academy of Sciences. Copyright 2011, National Academy of Sciences, USA.)

More recent molecular and functional studies indicate that human cell diversity is orders of magnitude greater than suggested by histological classification. Neurons, for example, are now known to be extremely diverse. Recent estimates suggest that there may be >10,000 types of human neuron and they are linked to each other by astonishingly complex connections (some individual neurons can be connected to 100,000 other neurons). Certain types of somatic DNA rearrangements are frequent during neurogenesis and may contribute to neuron diversity.

B and T lymphocytes display a special type of diversity. As they mature, they undergo cell-specific DNA rearrangements, so that individual B cells from a single person can produce different immunoglobulins, and similarly individual T cells can exhibit different T-cell receptors. We will consider the different types of somatic DNA rearrangements in lymphocytes and neurons in detail in Chapter 11 when we consider the general principles that underlie genetic variation.

Variability in cell life span and turnover

The average life span of a human cell is thought to be of the order of 7–10 years, but life spans vary enormously according to cell type (**Table 2.1**). Some cells are very long-lived. At the other extreme are cells that live for only a few days or weeks, being replaced by new cells generated ultimately from stem cells. The cells with the shortest life spans are those that are continually subjected to external challenges and stresses and/or high workloads.

TABLE 2.1 EXAMPLES OF LIFE SPAN/TURNOVER OF HUMAN CELLS

Human cell type	Approximate life span/turnover rate
Intestinal epithelium	3–5 days (the fastest turnover of any cell type)
Neutrophil	5–6 days
Plasma cell[a]	4 days–5 weeks
Red blood cell	10–20 weeks
Fat cell	~10% of adult fat cells are renewed annually
Other intestinal cells	~15 years
Cardiomyocyte	Decades (fewer than 50% are replaced in a normal person's life span)
Neuron	Very long-lived: many are not replaced in the life span of a human

[a] Or antibody-producing B cell. The most accurate measurements come from retrospective carbon dating of cells, as in the study by Spalding KL *et al.* (2005) *Cell* **122**:133–143; PMID 16009139.

Different methods have been used to assess cell turnover (the rate at which cells die off and are replaced by new cells). Easily accessible blood and epithelial cells have been readily studied, but a more general approach involves administering labeled nucleotides and following the incorporation of the label into DNA as cells divide in animal models, such as mice (the approach is not used in humans for safety reasons, and short-lived mice are never going to be good models in this case). It has, however, been possible to measure the turnover of our cells using a retrospective carbon-dating procedure in humans (atmospheric ^{14}C was introduced into the food chain at the time of extensive nuclear bomb testing from the late 1950s to the time of the test ban treaty in 1963; see the legend to **Table 2.1**).

Germ cells are specialized for reproductive functions

In multicellular organisms, development and growth are separated from reproductive functions. During growth and development there is a need for cell division to create increasing numbers of cells or to replace defective or worn-out cells by new cells. That is done by a standard type of cell division known as **mitosis**, in which a cell divides to give two daughter cells that are usually identical to each other and to the parent cell, and we will illustrate how the DNA replicates and is distributed equally between the daughter cells later in this chapter. However, sometimes during development and differentiation, cell division is asymmetric and results in two daughter cells with different properties (see **Box 2.2**).

BOX 2.2 ASYMMETRIC CELL DIVISIONS

In symmetrical cell divisions, the parent cell produces two daughter cells that have the same properties as each other and the parent cell. But some cell divisions are asymmetric: the parent cell gives rise to two daughter cells that are different in some way from each other. In oogenesis, for example, the primary oocyte gives rise to a secondary oocyte and a much smaller polar body, and when the secondary oocyte divides, it gives rise to an egg and a much smaller polar body (see **Figure 2.12**). When stem cells divide symmetrically, they produce identical daughter cells; but they can also divide asymmetrically to give one daughter cell that is like its parent and another daughter cell that is more differentiated.

Asymmetric cell division can occur in different ways. Recall that in some organisms the egg develops with asymmetric localization of a regulatory protein; cleavage of the egg results in daughter cells with significantly different amounts of the regulatory protein. Asymmetric cell division can also involve positioning the mitotic spindle away from the center of the cell, producing one large and one small daughter cell (as in the case of oogenesis, when small polar bodies are formed after the primary and secondary oocytes divide). Another possibility is to change the orientation of the mitotic spindle and centrosomes. That may result in the two daughter cells being exposed to different microenvironments where they receive different chemical signals from neighboring cells. As described in Chapter 4, signals received from neighboring cells can be directed to one side of the recipient cell and cause components of a signalling pathway to accumulate at that end of the cell in such a way as to re-orient the mitotic spindle and centrosomes.

A specialized population of **germ cells** is set aside to carry out reproductive functions; in evolutionary terms, the remaining **somatic cells** provide a vessel to carry these reproductive cells in order to achieve reproduction. In plants and primitive animals, ordinary somatic cells can give rise to germ cells throughout the life of the organism. However, in most of the animals that we understand in detail—insects, nematodes, and vertebrates, the germ cells are set aside very early in development as a dedicated **germ line** and represent the sole source of gametes.

The germ cells are the only cells in the body capable of **meiosis**, the specialized cell divisions that give rise to mature sperm and egg cells (allowing the genetic material to be transmitted from one generation to the next). In mammals, germ-line cells derive from primordial germ cells that are induced in the early embryo, as described in Chapter 4.

How eukaryotic cells evolved, and the origin of mitochondria, the nucleus, and the cytoskeleton

"*Omnis cellula e cellula*," the idea that all cells come from cells, was first popularized by the German cell pathologist Rudolf Virchow in the middle of the nineteenth century. Each cell in organisms living today originated through countless sequential cell divisions, and is connected through an unbroken chain of cells to a primordial cell. How that cell developed is an interesting question!

The origination of eukaryotes from prokaryotic cell ancestors signified a large evolutionary leap. Present-day eukaryotic cells are significantly larger than prokaryotic cells (being 10–30 times larger in linear dimensions, and roughly 1000–10,000 times larger in volume, than a typical bacterium such as *Escherichia coli*). The genome sizes and total gene numbers of eukaryotes increased, allowing increased functional complexity. That extended to the architecture of the cell (development of internal membrane systems, organelles, and a highly-sophisticated cytoskeleton) and to information processing/gene regulation (the physical separation of transcription from translation, and the development of RNA splicing, RNA interference, and so on). The subsequent evolution of multicellularity allowed extraordinary functional complexity.

The first eukaryotic cell is thought to have originated about 1.5–2 billion years ago. It is now widely accepted to have occurred by a special type of cell fusion, **endosymbiosis**, in which one cell engulfs another cell without destroying it (so that functions of both cells are retained). Because each of the original cells had a genome and a protein-synthesizing capacity (ribosomes, transfer RNAs, and associated translation factors), the cell fusion resulted in a stable cell with two genomes and two sets of protein-synthesis machinery (**Figure 2.5**).

Endosymbiosis can explain the origin of the two eukaryotic organelles that have their own independent genomes and protein-synthesis capacity: mitochondria and chloroplasts. Mitochondria are found in all eukaryotic cells, and originated when an anaerobic eukaryotic precursor cell engulfed an aerobic cell (the endosymbiont became a protomitochondrion). At that time, oxygen levels were known to be rising rapidly in the atmosphere, and engulfing an aerobic cell would have been advantageous for the anaerobic host cell. (Chloroplasts are thought to have evolved from secondary symbiosis in which a photosynthesizing cell was engulfed by an early eukaryotic cell, giving rise to a eukaryotic lineage leading to plants and algae.)

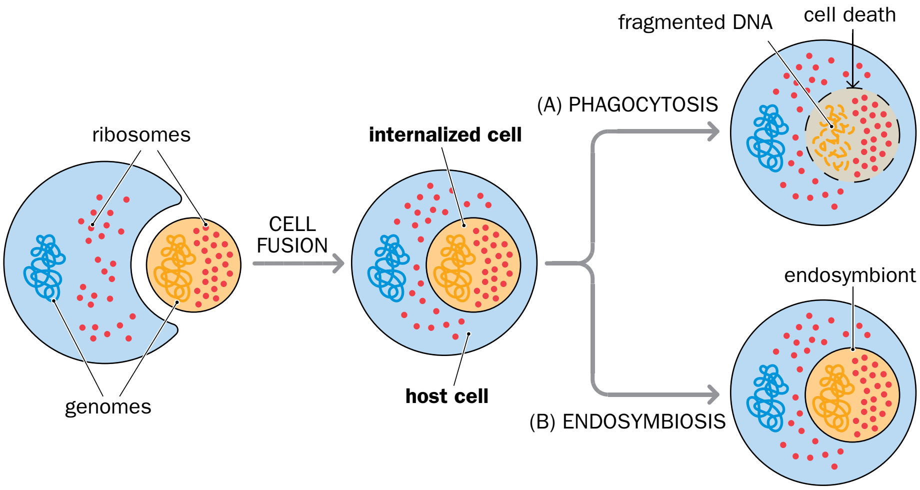


Figure 2.5 Cell fusion by cell engulfment usually results in phagocytosis, but a rare alternative is co-operative symbiosis. Cell engulfment is a form of cell fusion in which one cell, the host cell, has the flexibility to send out cytoplasmic processes to surround another cell and internalize it. (**A**) The internalized cell is typically destroyed by the host cell (phagocytosis): the genome is cleaved into small fragments and other components are degraded. This process has intermittently allowed a form of horizontal gene transfer during evolution, notably between prokaryotic cells, whereby DNA fragments from the genome of the internalized cell are incorporated into the host-cell genome (see **Figure 2.6A**). (**B**) A rare alternative fate for the internalized cell is that it is not destroyed; instead it co-operates with the host cell in a symbiotic relationship known as *endosymbiosis* (the internalized cell is known as an endosymbiont). In this case, the fusion cell continues with two genomes and two sets of protein-synthesis machinery (including ribosomes and a complement of transfer RNAs), one of each donated by the host cell and one by the internalized cell. The eukaryote lineage arose by this process; see **Figure 2.6C**.

During the long period of evolutionary time that has elapsed since these endosymbiotic events, there has been an expansion in the size of the host-cell genome, giving rise to the nuclear genome of eukaryotic cells. At the same time, the mitochondrial and chloroplast genomes evolved through a process where most of the original DNA present in the endosymbiont was shed from the genome. The human mitochondrial genome, for example, is only 16.6 kb (about 0.3% of the size of an average prokaryotic genome).

Both archaeal and bacterial DNA sequences contributed to the evolution of eukaryotic genomes. That became clear when inferred eukaryotic protein sequences (much more evolutionarily conserved than the corresponding gene sequences) were compared against translated coding sequences from prokaryotes in order to identify significantly related sequences. Many eukaryotic genes do not have any recognizable equivalent (**homolog**) in prokaryotic genomes, but some eukaryotic genes have clear homologs in archaea, and some others have clear homologs in bacteria. The archaea–eukaryote homologs are especially common in information processing systems such as DNA replication, transcription, recombination, and DNA repair. By contrast, the bacteria–eukaryote homologs are more likely to have operational functions, working in metabolic pathways, as membrane components, and so on.

Phylogenetic analyses of genes in current mitochondria clearly indicate that the endosymbiont that gave rise to the mitochondrial genome was an α-proteobacterium. The host cell was a complex type of archaeon (archaea, but not bacteria, have clear homologs of important eukaryotic proteins working in nuclear DNA replication, and also homologs of eukaryotic histones and of actins that work in the eukaryotic cytoskeleton). However, the nuclear genome of eukaryotes is a mosaic of DNA sequences that originated from both archaeal and bacterial genomes.

The hybrid origins of the nuclear genome are due to repeated rounds of **horizontal gene transfer** in which sequence components of the genome of various bacterial cells were integrated into the genome of the archaeal eukaryotic precursor cell and its descendants. First, the archaeal eukaryotic precursor cell is envisaged to have developed complexity by repeated phagocytosis of diverse bacterial cells: after degrading each engulfed bacterial cell and cleaving its DNA, bacterial DNA fragments integrated into the host cell's genome, extending its size and ultimately its functional capabilities. Then, after the endosymbiosis that gave rise to the proto-mitochondrion, many DNA sequences were intermittently shed from the α-proteobacterial genome and transferred to the archaeal host-cell genome (**Figure 2.6**). That may have happened over a short period of evolutionary time, but even in modern times mtDNA sequences are occasionally transferred from the mitochondrial to the nuclear genome. We describe examples of these nuclear mitochondrial (NUMT) sequences when surveying the architecture of the human genome in Section 9.1.

The pivotal development of multicellular organisms might have arisen through simple gene mutations

The development of multicellularity was a pivotal point in eukaryote evolution: different cells in a multicellular organism may become specialized to carry out different tasks, enabling much greater functional complexity. The task of ensuring that the organism is able to reproduce is delegated to specialized germ cells. The remaining somatic cells can

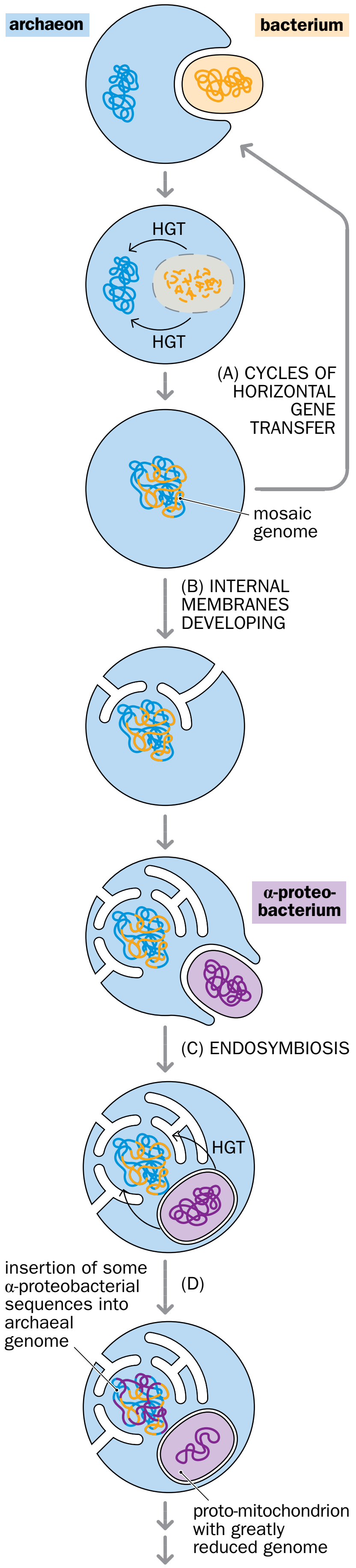

archaeon
bacterium
HGT
HGT
(A) CYCLES OF HORIZONTAL GENE TRANSFER
mosaic genome
(B) INTERNAL MEMBRANES DEVELOPING
α-proteo-bacterium
(C) ENDOSYMBIOSIS
HGT
insertion of some α-proteobacterial sequences into archaeal genome
(D)
proto-mitochondrion with greatly reduced genome

Figure 2.6 A key step in eukaryote evolution was an endosymbiotic event in which a complex anaerobic archaeon engulfed an aerobic α-proteobacterium. The archaeon is imagined to have had some internal membrane structure (not shown here for clarity). (**A**) It may have achieved complexity through multiple previous events in which it phagocytosed bacterial cells, leading to destruction of the internalized cell and the release of short DNA sequences that were then incorporated into the archaeal genome. After several cycles of internalizing bacterial cells and horizontal gene transfer (HGT) from the ingested bacterial genomes, the host cell genome became a mix of archaeal (blue) and bacterial (orange) DNA sequences. (**B**) The resulting increase in genome complexity could have speeded-up evolution, leading to the development of additional internal membranes. (**C**) The key endosymbiosis event involved internalizing an α-proteobacterium. Thereafter, DNA sequences were periodically shed from the internalized cell's genome over a long period of evolutionary time: the genome became much smaller, giving rise to present-day mitochondrial genomes. Some of the discarded α-proteobacterial sequences were degraded and lost, but others were incorporated into the archaeal genome (**D**), which subsequently became more complex, giving rise to the nuclear genome of eukaryotes. This model has been supported by the recent discovery of Lokiarchaeota, complex archaea whose genomes encode an expanded repertoire of eukaryotic signature proteins that suggest sophisticated membrane-remodeling capabilities (see Spang A *et al.* [2015] *Nature* **521**:173–179; PMID 25945739). (Adapted from Martijn J & Ettema TJ [2013] *Biochem Soc Trans* **41**:451–457; PMID 23356327. With permission from Biochemical Society.)

be programmed to become diverse types of cell, such as neurons, lymphocytes, and so on, in the case of animals.

Repeated cell division allows the organism to grow from a single fertilized egg cell, and complex cell–cell interactions and cell–environment interactions during development permit progressive development of cell specialization. Cells can then co-operate to build complex tissues and organs. We cover these areas in detail in Chapter 4.

The development of multicellularity allowed stunning evolutionary advances: complex fungi, plants, and animals evolved, and one species went on to develop sufficient cognitive powers that it began to radically transform its environment, rather than simply adapt to it. One might expect that the development of multicellularity would have required progressive changes to the genome. Genetic studies in yeast, however, have suggested that single gene mutations that prevent mother cells detaching from daughter cells might have been highly significant (Figure 2.7).

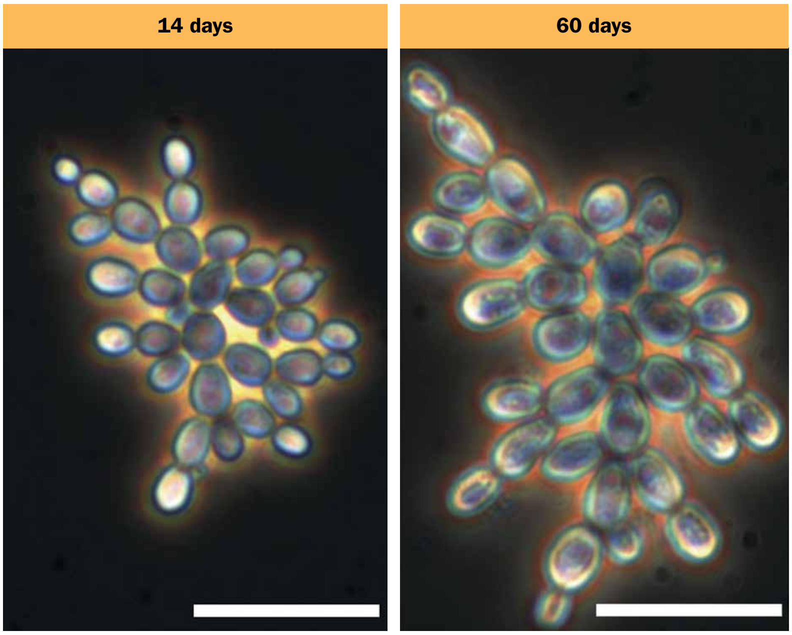


Figure 2.7 Might multicellularity have evolved by a single gene mutation? Artificial manipulation of the yeast *Saccharomyces cerevisiae* can produce a mutation inactivating the *ACE2* gene (which makes a type of transcription factor). Disrupting the production of just this one transcription factor is enough to prevent mother–daughter cell separation, generating multicellular “snowflake” yeast. The cells in the cluster showed evidence of coordination and a high-broad-sense heritability for multicellular traits. By 60 days, the cells had evolved in concert to be 2.2-fold larger than an early snowflake yeast from the same population (14 days). Scale bars, 50 μm. (From Ratcliff WC *et al.* [2015] *Nat Commun* **6**:6102; PMID 25600558. With permission from Springer Nature. Copyright © 2015.)

2.2 DNA AND CHROMOSOME COPY NUMBER DURING THE CELL CYCLE

The chromosome and DNA content of cells is defined by the number (n) of different chromosomes (the **chromosome set**) and the associated DNA content (C). For human cells, $n = 23$ and C is ~3.5 pg (3.5×10^{-12} g). Different cell types in an organism, however, may differ in DNA content and in **ploidy**, the number of copies they have of the chromosome set.

Cells differ in DNA content between organisms, between individuals within a species, and within an individual. For any species, the reference DNA content of cells, the **C value**, is the amount of DNA in cells that have a single chromosome set. C values vary widely for different organisms, but there is no direct relationship between the C value and biological complexity. The human C value is only 19% of that of an onion, for example. We return to consider this in detail in Chapter 13 when we consider our place in the Tree of Life.

Different cells within a single individual show differences in ploidy

The DNA content of cells within a single individual can vary in different ways. We will consider minor differences due to genetic variation in later chapters. Here, we are concerned with differences in chromosome and DNA copy number. In animals, the gametes (sperm and egg cells) may be viewed as reference cells because they carry a single chromosome set; they are said to be **haploid** (with *n* chromosomes and a DNA content of C).

Most human and mammalian somatic cells carry two copies of the chromosome set and are **diploid** (with 2*n* chromosomes and a DNA content of 2C). Note, however, that in several non-mammalian animal species, most somatic cells are not diploid, but are usually either haploid or polyploid. In the latter case, some are tetraploid (4*n*) and others have a ploidy >4*n* (triploidy is less common in animals because it can give rise to problems in producing sperm and egg cells).

Although the majority of human somatic cells are diploid, some cells, for example erythrocytes, platelets, and mature keratinocytes, lose their nucleus and so are nulliploid. Others are naturally polyploid, and are formed in one of two ways. Some cells may become polyploid after undergoing several rounds of DNA replication without cell division (Figure 2.8A). Examples are hepatocytes (<8C) in the liver, cardiomyocytes (4C–8C) in heart muscle, and megakaryocytes (16C–64C; Figure 2.8B and C. In other cases, cells may become polyploid through cell fusions. For example, skeletal muscle-fiber cells are polyploid as a result of going through multiple rounds of cell fusion. The individual muscle-fiber cells can become very long and contain very many diploid nuclei. Multinucleated cells like this are known as **syncytial cells** (Figure 2.8D).

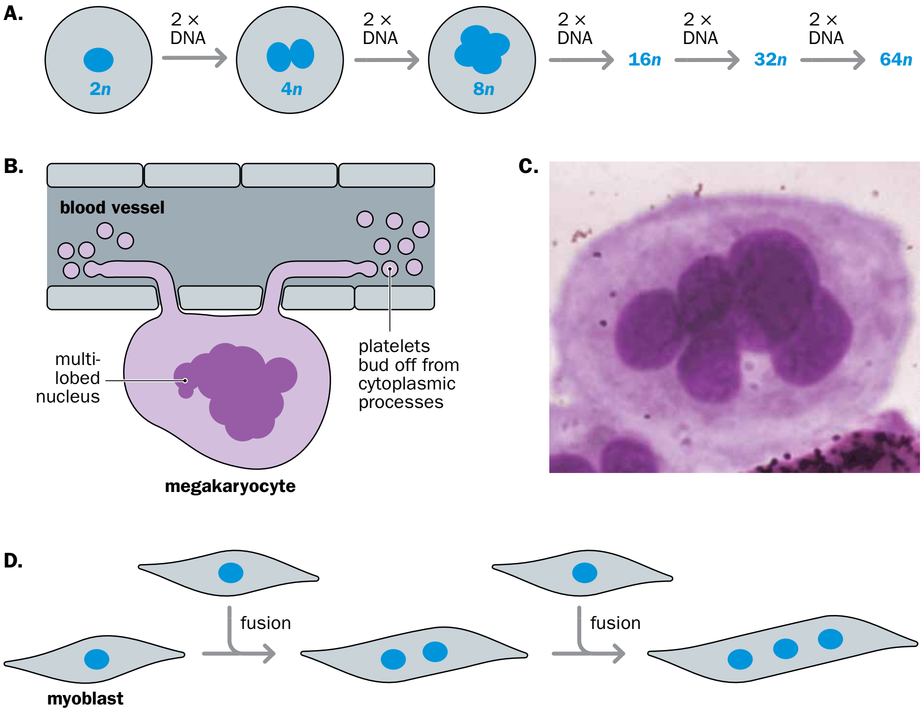


Figure 2.8 Polyploid somatic cells can arise from endomitosis or cell fusion. (**A**) Principle of endomitosis in which the DNA of a cell replicates but without cell division. (**B**) The megakaryocyte is a giant polyploid bone marrow cell (often 16C–64C) that is responsible for producing the thrombocytes (platelets) needed for blood clotting. It has a large multilobed nucleus as a result of undergoing multiple rounds of endomitosis. Multiple platelets are formed by budding from cytoplasmic processes of the megakaryocyte and so have no nucleus. (**C**) Example of a megakaryocyte showing the multilobed nucleus in the center (courtesy of Centers for Disease Control and Prevention [CDC]/Kathy Keller). (**D**) Skeletal muscle-fiber cells are polyploid because they are formed by fusion of large numbers of myoblast cells to produce extremely long multinucleated cells. A multinucleated cell is known as a syncytium.

Cells with the normal chromosome number for that type of cell are said to be **euploid**. However, cells can develop an abnormal number of chromosomes, and are then said to be **aneuploid**. That can happen either as a result of abnormalities in chromosome segregation (detailed in Chapter 15), but also occurs by different mechanisms in cancer cells (described in Chapter 19).

Differences in ploidy and DNA content during the cell cycle

The cells of our body are all derived ultimately from a single diploid cell, the **zygote**, that is formed when a sperm fertilizes an egg. Starting from the zygote, organisms grow by repeated rounds of cell division. Each round of cell division is a **cell cycle** and comprises a brief M phase, during which cell division occurs, and the much longer intervening **interphase**, which has three parts (**Figure 2.9**). They are S phase (when DNA synthesis occurs), the G_1 phase (gap between M phase and S phase), and G_2 phase (gap between S phase and M phase).

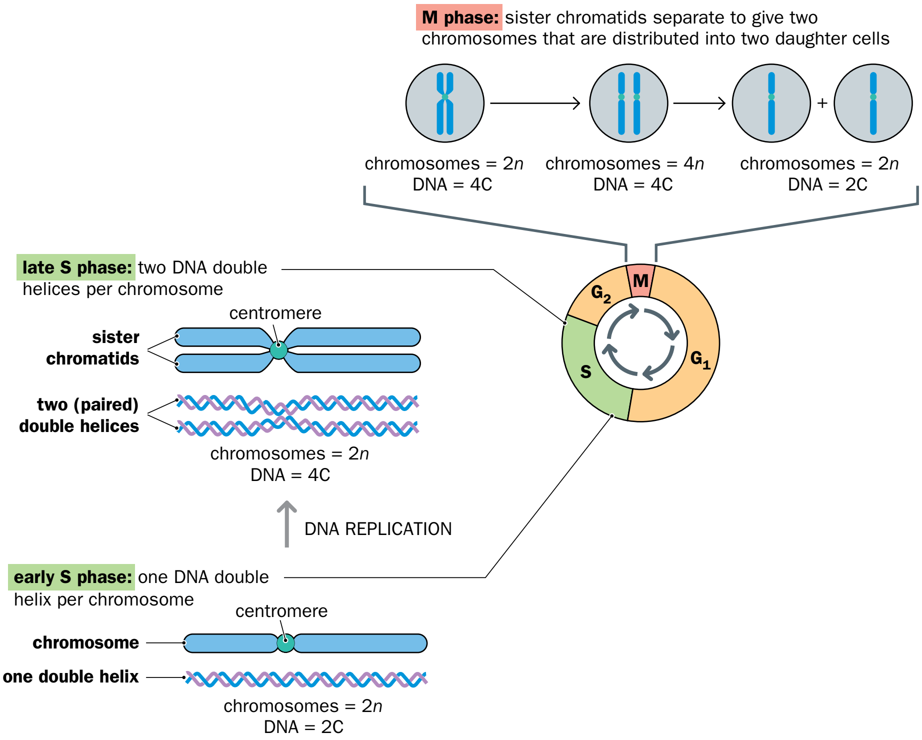

M phase: sister chromatids separate to give two chromosomes that are distributed into two daughter cells
+
chromosomes = 2*n*
DNA = 4C
chromosomes = 4*n*
DNA = 4C
chromosomes = 2*n*
DNA = 2C
late S phase: two DNA double helices per chromosome
centromere
sister chromatids
two (paired) double helices
chromosomes = 2*n*
DNA = 4C
DNA REPLICATION
early S phase: one DNA double helix per chromosome
centromere
chromosome
one double helix
chromosomes = 2*n*
DNA = 2C
G_2
M
G_1
S

Figure 2.9 Changes in chromosomes and DNA content during the cell cycle. The cell cycle shown at the right includes a very short M phase, when the chromosomes become extremely highly condensed in preparation for nuclear and cell division. Afterward, cells enter a long period of growth called interphase, during which chromosomes are enormously extended so that genes can be expressed. Interphase is divided into three phases: G_1, S (when the DNA replicates), and G_2. Chromosomes contain one DNA double helix from the end of M phase right through until just before the DNA duplicates in S phase. After the DNA double helix duplicates, the two resulting double helices are held tightly together along their lengths (by specialized protein complexes called cohesins) until the M phase. As the chromosomes condense at M phase they are now seen to consist of two sister chromatids, each containing a DNA duplex, that are bound together only at the centromeres. During M phase the two sister chromatids separate to form two independent chromosomes that are then equally distributed into the daughter cells.

We will describe the cell biology underlying the phases of the cell cycle in Chapter 19. Here, we are mostly concerned with the chromosome and DNA copy number. During each cell cycle, chromosomes undergo profound changes to their structure, number, and distribution within the cell. From the end of the M phase, right through until before DNA duplication in S phase, a chromosome of a diploid cell contains a single DNA double helix, and the total DNA content is 2C (see **Figure 2.9**). After DNA duplication, the total DNA content is 4C, but the duplicated double helices are held together along their lengths so that each chromosome has double the DNA content of a chromosome in early S phase. During M phase, the duplicated double helices separate, generating two daughter chromosomes, giving 4*n* chromosomes. After equal distribution of the chromosomes to the two daughter cells, both cells will have 2*n* chromosomes and a DNA content of 2C (see **Figure 2.9**).

G_1 is the normal state of a cell, and the long-term end state of nondividing cells. Cells enter S phase only if they are committed to mitosis; as will be described in more detail in Chapter 3, nondividing cells remain in a modified G_1 stage, sometimes called the G_0 phase. The cell-cycle diagram can give the impression that all the interesting action happens in S and M phases, but this is an illusion. A cell spends most of its life in the G_0 or G_1 phase, and that is where the genome does most of its work.

A small subset of diploid body cells constitutes the **germ line** that gives rise to **gametes** (sperm cells or egg cells). In humans, where *n* = 23, each gamete contains one sex chromosome plus 22 nonsex chromosomes (**autosomes**). In eggs, the sex chromosome is always an X; in sperm, it may be either an X or a Y. After a haploid sperm fertilizes a haploid egg, the resulting diploid zygote and almost all of its descendent cells have the chromosome constitution 46,XX (female) or 46,XY (male) (**Figure 2.10**).

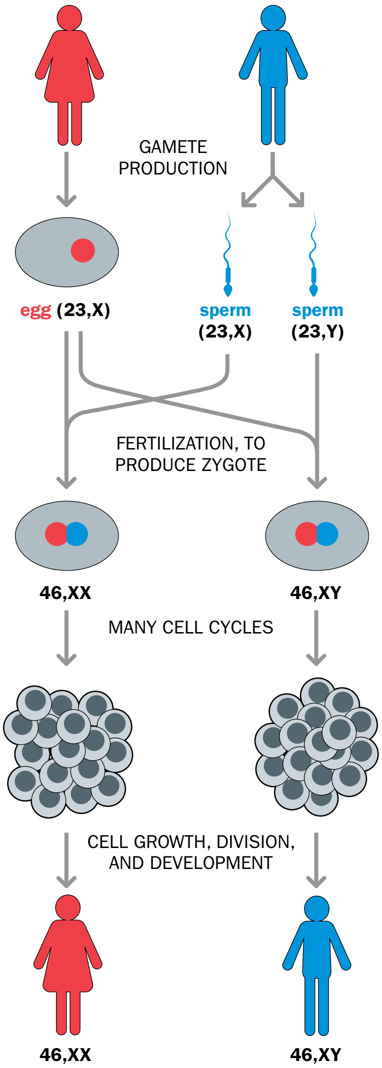


Figure 2.10 The human life cycle, from a chromosomal viewpoint. Haploid egg and sperm cells originate from diploid precursors in the ovary and testis in women and men, respectively. All eggs have a 23,X chromosome constitution, representing 22 autosomes plus a single X sex chromosome. A sperm can carry either sex chromosome, so that the chromosome constitution is 23,X (50%) and 23,Y (50%). After fertilization and fusion of the egg and sperm nuclei, the diploid zygote will have a chromosome constitution of either 46,XX or 46,XY, depending on which sex chromosome the fertilizing sperm carried. After many cell cycles, this zygote gives rise to all cells of the adult body, almost all of which will have the same chromosome complement as the zygote from which they originated.

Cells outside the germ line are **somatic cells**. Human somatic cells are usually diploid but, as described above, there are notable exceptions, ranging from nulliploid cells (erythroid cells, terminally differentiated skin cells) to polyploid cells (as in **Figure 2.8**).

2.3 CELL DIVISION AND TRANSMISSION OF DNA TO DAUGHTER CELLS

Mitosis and meiosis both involve chromosome replication prior to cell division. However, the products of mitosis have the same ploidy as the initiating cell, while meiosis halves the cell's ploidy. Furthermore, while mitosis gives rise to genetically identical products, meiosis generates genetic diversity to ensure that offspring are genetically different to their parents.

When it comes to cell division, most attention is particularly focused on what happens to chromosomes and chromosomal DNA in mitosis and meiosis. That is understandable because almost all of a cell's DNA and genes are located in the chromosomes of the nucleus, and because there are very tight controls on the duplication and segregation of chromosomal DNA. In addition, however, mtDNA also undergoes replication, after which the mtDNA molecules need to be partitioned (segregated) between the daughter cells. As detailed below, the control over both the copy number and segregation of mtDNA is much less stringent than for chromosomal DNA.

Mitosis is the normal form of cell division

As an embryo develops through fetus, infant, and child to adult, many cell cycles are needed to generate the required number of cells. As many cells have a limited life span, there is also a continuous requirement to generate new cells, even in an adult organism. All of these cell divisions occur by **mitosis**, which is the normal process of cell division throughout the human life cycle. Mitosis ensures that a single parent cell gives rise to two daughter cells that are both genetically identical to the parent cell, barring any errors that might have occurred during DNA replication. During a human lifetime, there may be something like 10^{17} mitotic divisions.

The M phase of the cell cycle includes various stages of nuclear division (prophase, prometaphase, metaphase, anaphase, and telophase), and also cell division (*cytokinesis*), which overlaps the final stages of mitosis (**Figure 2.11**). In preparation for cell division, the previously highly-extended, duplicated chromosomes contract and condense so that, by the metaphase stage of mitosis, they are readily visible when viewed under the microscope.

The chromosomes of early S phase have one DNA double helix but following DNA replication, two identical DNA double helices are produced (see **Figure 2.9**). The two DNA helices are held together along their lengths by cohesins, protein complexes resembling the condensin proteins that compact chromatin. Precisely how the sister chromatids are held together by cohesins is uncertain. Three of the cohesin subunits can interact to form a large protein ring, and some models envisage cohesin rings encircling the two double helices to entrap them; other models imagine that cohesin rings form round the individual double helices and then interact to ensure that the two double helices are held tightly together.

Later, when the chromosomes undergo compaction in preparation for cell division, the cohesins are removed from all parts of the chromosomes apart from the centromeres. As a result, by prometaphase, when the chromosomes can now be viewed under the light microscope, individual chromosomes can now be seen to comprise two **sister**

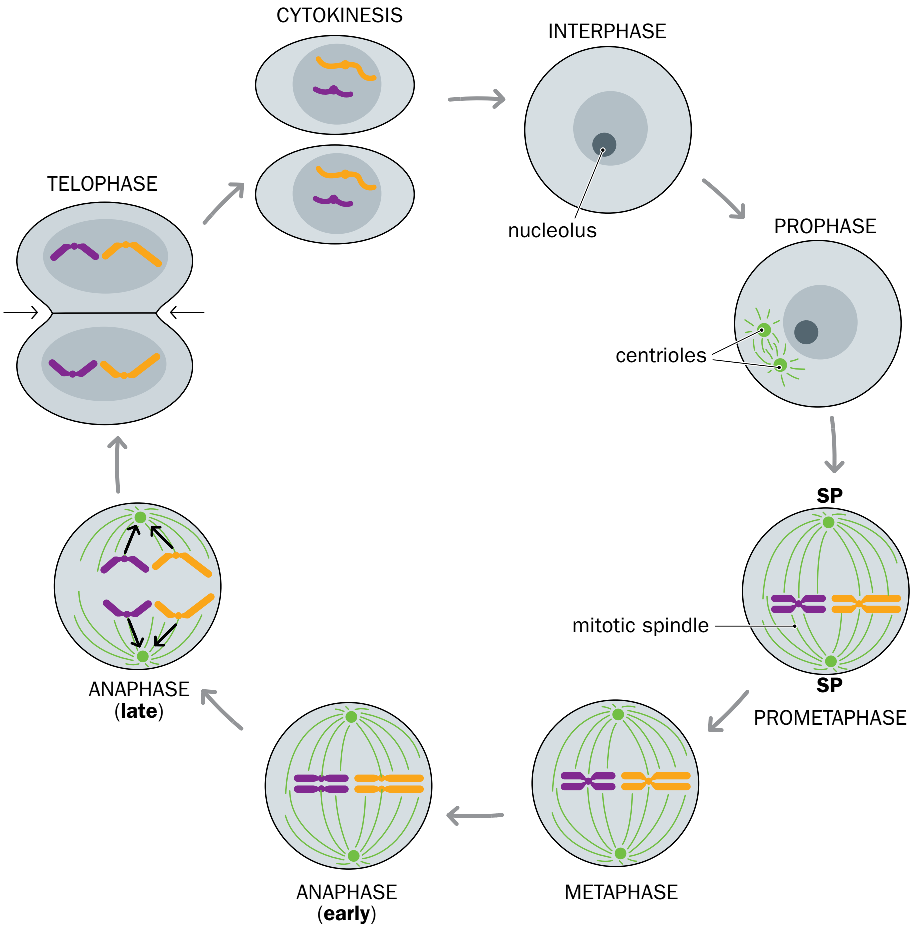


Figure 2.11 Mitosis (nuclear division) and cytokinesis (cell division). After S phase (in late interphase) each chromosome consists of two immensely long sister chromatids (not shown here) that are held together along their lengths by cohesin protein complexes. Then, in preparation for mitosis, the chromosomes begin to shorten and thicken. Early in prophase, centrioles (short, cylindrical structures composed of microtubules and associated proteins) begin to separate and migrate to opposite poles of the cell to form the spindle poles (SP). In prometaphase, the nuclear envelope breaks down, and the now highly-condensed chromosomes become attached at their centromeres to the array of microtubules extending towards the mitotic spindle. At metaphase, the chromosomes lie along the middle of the mitotic spindle, the equatorial plane, still with the sister chromatids bound together; at this stage most of the cohesin complexes have been removed, but residual cohesins at the centromere hold the duplicated DNA helices together. Removal of the residual cohesins at the centromere allows the onset of anaphase: the sister chromatids separate for the first time to form independent chromosomes, each with their own centromere. Later in anaphase, the centromeres are pulled by the microtubules of the spindle in the direction of opposing poles (arrows). The nuclear envelope forms again around the daughter nuclei during telophase, and the chromosomes decondense, completing mitosis. Before the final stages of mitosis, and most obviously at telophase, cytokinesis begins with constriction of the cell that will increase progressively to produce two daughter cells.

chromatids that are attached together at the centromere by the residual cohesin complexes that continue to bind the two DNA helices at this position.

Later still, at the start of anaphase, the residual cohesin complexes holding the sister chromatids together at the centromere are removed. The two sister chromatids can now disengage to become independent chromosomes that will be pulled to opposite poles of the cell and then distributed equally to the daughter cells (see **Figure 2.11**). Interaction between the mitotic spindle and the centromere is key to this process and we will consider this in detail in Section 2.4.

Note that we portray cell division here as being symmetrical, but some types of cell divisions, including many types of cell division in early development, and some stem cell divisions are asymmetric (see **Box 2.2** for a brief overview).

Meiosis is a specialized reductive cell division that gives rise to sperm and egg cells

Diploid primordial germ cells migrate into the embryonic gonad and engage in repeated rounds of mitosis, to generate spermatogonia in males and oogonia in females. Further growth and differentiation produce primary spermatocytes in the testis and primary oocytes in the ovary. This process requires many more mitotic divisions in males than in females, and likely contributes to sex differences in the mutation rate. The diploid spermatocytes and oocytes can then undergo **meiosis**, the cell division process that is designed to produce genetically unique haploid gametes. That is, each sperm cell produced by a man and each egg cell produced by a woman is designed to have a genome sequence that is unlike that of any other sperm or egg cell that they, or anybody else, produces.

Meiosis is a reductive division because it involves two successive cell divisions (meiosis I and II) but only one round of DNA replication. As a result, it gives rise to four haploid cells. In males, the two meiotic cell divisions are each symmetrical, producing four functionally equivalent spermatozoa. Female meiosis is different because, at each meiosis, asymmetric cell division results in unequal division of the cytoplasm. The products of female meiosis I (the first meiotic division) are a large secondary oocyte and a small cell (**polar body**), which is discarded. During meiosis II, the secondary oocyte then gives rise to the large mature egg cell and a second polar body, which again is discarded (Figure 2.12).

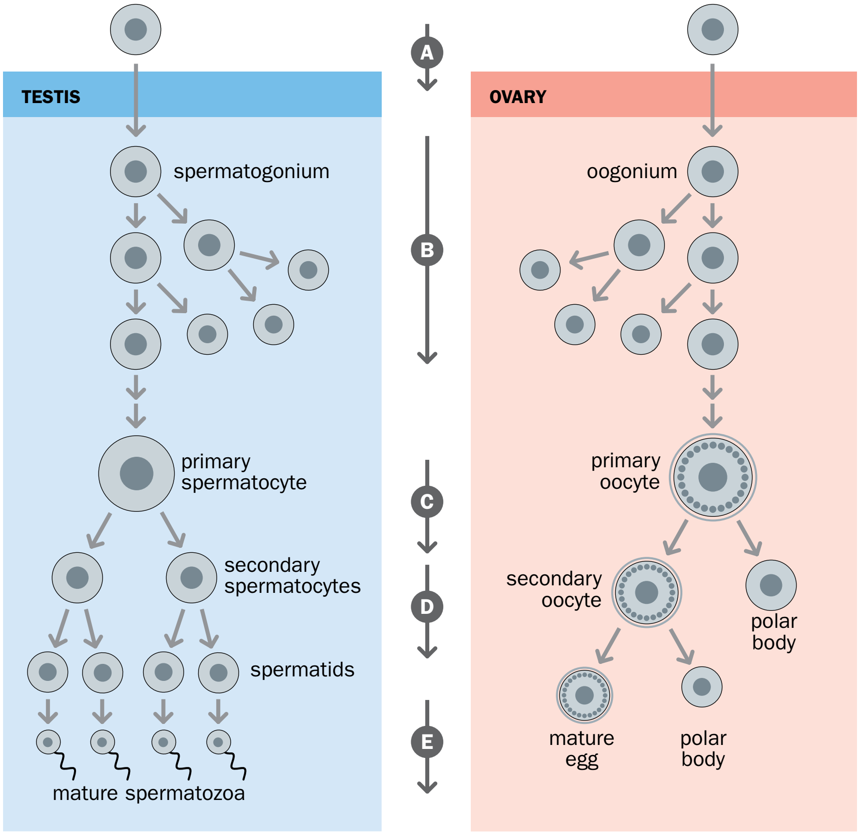


Figure 2.12 **Male and female germ-line development and gametogenesis.** (**A**) Diploid primordial germ cells migrate to the embryonic gonad (the male testis or the female ovary) and enter rounds of mitosis that establish spermatogonia (in males) and oogonia (in females). (**B**) These undergo further mitotic divisions, growth, and differentiation to produce diploid primary spermatocytes and diploid primary oocytes, which can enter meiosis. (**C**) Meiosis I. After DNA duplication, the cells become tetraploid but then divide to produce two diploid cells. In male gametogenesis, the cell division is symmetrical, generating identical, diploid secondary spermatocytes. In female meiosis I, by contrast, the division is asymmetric; the secondary oocyte is much larger than the first polar body, which is discarded. (**D**) Meiosis II. The diploid secondary spermatocyte and secondary oocyte divide without prior DNA synthesis to give haploid cell products. In male gametogenesis, this division is again symmetrical, producing two haploid spermatids from each secondary spermatocyte. In female meiosis II, the egg produced is much larger than the second (also discarded) polar body. (**E**) Maturation produces four spermatozoa and a single egg.

In humans, primary oocytes enter meiosis I during fetal development but are then all arrested at prophase until after the onset of puberty. After puberty in females, one primary oocyte completes meiosis with each menstrual cycle. Because ovulation can continue up to the fifth and sometimes sixth decades, this means that meiosis can be arrested for many decades in primary oocytes that are used in ovulation in later life. While arrested in prophase, the primary oocytes continue to grow to become large in size, acquiring an outer jelly coat and cortical granules, as well as reserves of ribosomes, mRNA, yolk, and other cytoplasmic resources that would sustain an early embryo. In males, huge numbers of sperm are produced continuously from puberty onward.

The second division of meiosis is identical to mitosis, but the first division has important differences. Its purpose is to generate genetic diversity, creating genetic differences between the daughter cells. This is done by two mechanisms: independent assortment of paternal and maternal homologs, and recombination.

Independent assortment

Every diploid cell contains two chromosome sets, and so has two copies (**homologs**) of each chromosome (except in the special case of the X and Y chromosomes in males). One homolog is paternally inherited and the other is maternally inherited.

During meiosis I the maternal and paternal homologs of each pair of replicated chromosomes undergo **synapsis** by pairing together to form a **bivalent**. (Although the X and Y chromosomes have very different sequences, they too can form a bivalent in male meiosis; see below.) Following DNA replication, the homologous chromosomes each comprise two sister chromatids, so each bivalent is a four-stranded structure at the metaphase plate. Spindle fibers then pull one complete chromosome (two chromatids) to either pole. In humans, for each of the 23 homologous pairs, the choice of which daughter cell each homolog enters is independent. This allows 2^{23}, or about 8.4×10^6, different possible combinations of parental chromosomes in the gametes that might arise from a single meiotic division (Figure 2.13).

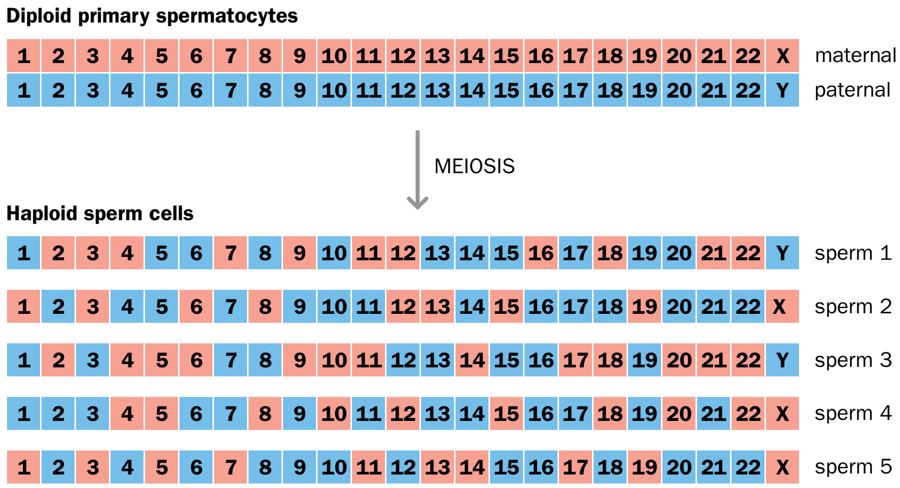

Diploid primary spermatocytes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X maternal
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 Y paternal
MEIOSIS
Haploid sperm cells
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 Y sperm 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X sperm 2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 Y sperm 3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X sperm 4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X sperm 5

Figure 2.13 Independent assortment of maternal and paternal homologs during meiosis. The figure shows a random selection of just 5 of the 8,388,608 (2^{23}) theoretically possible combinations of homologs that might occur in haploid human spermatozoa after meiosis in a diploid primary spermatocyte. Maternally derived homologs are represented by pink boxes, and paternally derived homologs by blue boxes. For simplicity, the diagram ignores recombination (but see **Figure 2.16**).

Recombination

The five stages of prophase of meiosis I (**Figure 2.14**) begin during fetal life and, in human females, can last for decades. During this extended process, the homologs within each bivalent normally exchange segments of DNA at randomly positioned but matching locations. At the zygotene stage (**Figure 2.14B**), a proteinaceous synaptonemal complex, consisting of proteins, forms between closely apposed homologous chromosomes. Completion of the synaptonemal complex marks the start of the pachytene stage (**Figure 2.14C**), during which recombination (crossover) occurs. Crossover involves physical breakage of the DNA in one paternal and one maternal chromatid, and the subsequent joining of maternal and paternal fragments.

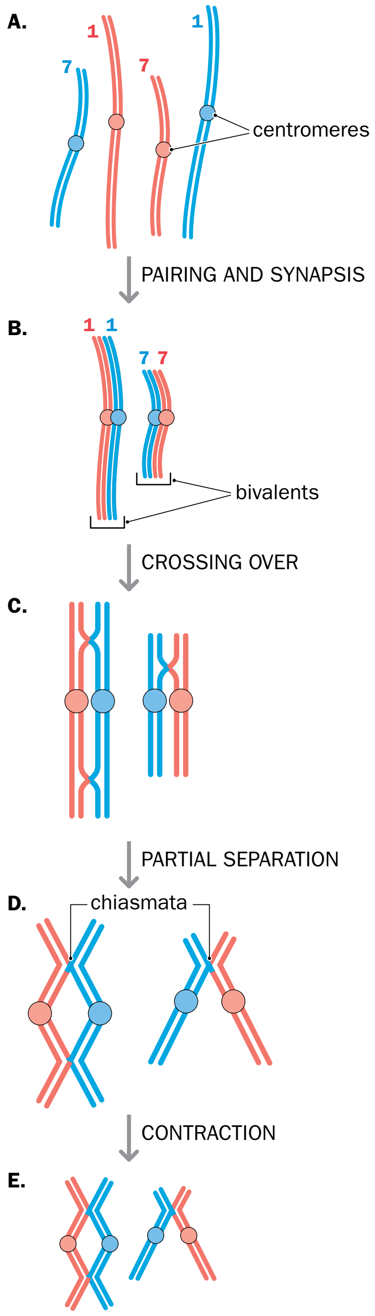


Figure 2.14 The five stages of prophase in meiosis I. (**A**) In leptotene, the duplicated chromosomes, each with a pair of sister chromatids, begin to condense but remain unpaired (shown here are representative maternal and paternal chromosome 1 homologs, and maternal and paternal chromosome 7 homologs). (**B**) In zygotene, duplicated maternal and paternal homologs pair, to form *bivalents* comprising four chromatids. Pairing of homologous chromosomes leads to fusion of the maternal and paternal homologs (*synapsis*). (**C**) In pachytene, recombination (crossing over) occurs via the physical breakage and subsequent rejoining of maternal and paternal chromosome fragments. There are two crossovers in the bivalent on the left and one in the bivalent on the right. For simplicity, both crossovers on the left involve the same two chromatids. In reality, more crossovers may occur, involving three or even all four chromatids in a bivalent. (**D**) During diplotene, the homologous chromosomes may separate slightly, except at the chiasmata. (**E**) Diakinesis is marked by contraction of the bivalents and is the transition to metaphase I.

The mechanism allowing alignment of the homologs (**Figure 2.14A** and **B**) is not known, although such close apposition is required for recombination. Located at intervals on the synaptonemal complex are very large multiprotein assemblies, called recombination nodules, that may mediate recombination events. Recombined homologs appear to be physically connected at specific points. Each such connection marks the point of a crossover and is known as a **chiasma** (plural chiasmata). There are an average of 55 chiasmata per cell in human male meiosis, and around 90 or so chiasmata per cell in female meiosis, and most of these occur at recombination hotspots.

In addition to their role in recombination, chiasmata are thought to be essential for correct chromosome segregation during meiosis I. By holding maternal and paternal homologs of each chromosome pair together on the spindle until anaphase I, they have a role analogous to that of the centromeres in mitosis and in meiosis II. Children with incorrect numbers of chromosomes have been shown genetically to be often the product of gametes where a bivalent lacked chiasmata.

Meiosis II resembles mitosis, except that there are only 23 chromosomes instead of 46. Each chromosome already consists of two chromatids that become separated at anaphase II. However, while the sister chromatids of a mitotic chromosome are genetically identical, the two chromatids of a chromosome entering meiosis II (**Figure 2.15**) are usually genetically different from each other, as a result of recombination events that took place during meiosis I.

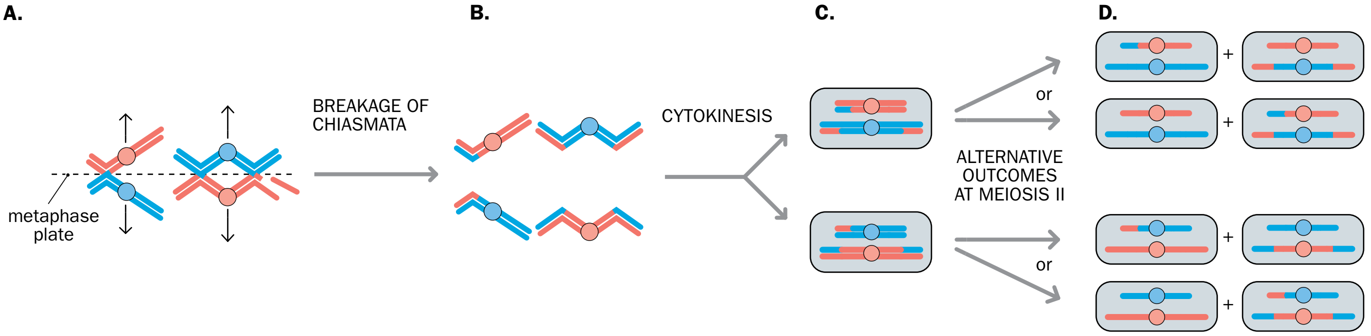


Figure 2.15 Metaphase I to production of gametes. (**A**) At metaphase I, the bivalents align on the metaphase plate, at the center of the spindle apparatus. Contraction of spindle fibers draws the chromosomes in the direction of the spindle poles (arrows). (**B**) The transition to anaphase I occurs at the consequent rupture of the chiasmata. (**C**) Cytokinesis segregates the two chromosome sets, each to a different primary spermatocyte in males. Note that following recombination during prophase I (**Figure 2.14C**), the chromatids share a single centromere but are no longer identical. (**D**) Meiosis II in each primary spermatocyte, which does not include DNA replication, generates unique genetic combinations in the haploid secondary spermatocytes. Only 2 of the possible 23 different human chromosomes are depicted, for clarity, so only 2^2 (= 4) of the possible 2^{23} (8,388,608) possible combinations are illustrated. Although oogenesis can produce only one functional haploid gamete per meiotic division (see **Figure 2.12**), the processes by which genetic diversity arises are the same as in spermatogenesis.

In **Figure 2.13**, we illustrated how independent assortment of homologs (during anaphase I) would have an effect on genetic variation by itself alone. But if we include the additional effects of recombination between homologs (during prophase I), and consider just chromosome 1, the combined effects could produce the additional variation seen in **Figure 2.16**. The combined effects of independent assortment of homologs and recombination ensure that each gamete is genetically unique. Each man can produce vast numbers of genetically distinct gametes, but only a limited number of eggs are produced by a woman. The genetic consequences of recombination are considered more fully in a later chapter.

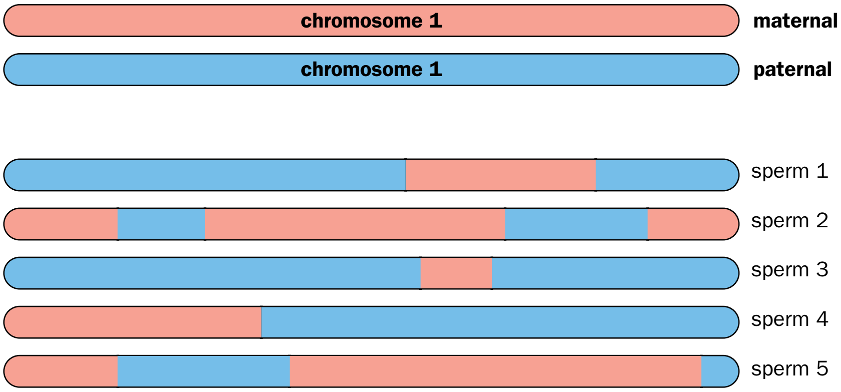


Figure 2.16 Recombination superimposes additional genetic variation at meiosis I. Figure 2.13 illustrates the contribution to genetic variation at meiosis I made by independent assortment of homologs, but for simplicity it ignores the contribution made by recombination. In reality, each transmitted chromosome is a mosaic of paternal and maternal DNA sequences, as shown here.

X–Y pairing

During meiosis I in a human primary oocyte, each chromosome has a fully homologous partner, and the two X chromosomes synapse and engage in crossover just like any other pair of homologs. In male meiosis there is a problem. The human X and Y sex chromosomes are very different from one another. Not only is the X very much larger than the Y, but it has a rather different DNA content and very many more genes than the Y. Nevertheless, the X and Y do pair during prophase I, thus ensuring that at anaphase I each daughter cell receives one sex chromosome, either an X or a Y.

Human X and Y chromosomes pair end-to-end rather than along the whole length, thanks to short regions of homology between the X and Y chromosomes at the very ends of the two chromosomes. Pairing is sustained by an obligatory crossover in a 2.6 Mb homology region at the tips of the short arms, but crossover also sometimes occurs in a second homology region, 0.32 Mb long, at the tips of the long arms. Genes in the terminal X–Y homology regions have some interesting properties:

- They are present as homologous copies on the X and Y chromosomes
- They are mostly not subject to the transcriptional inactivation that affects most X-linked genes as a result of the normal decondensation of one of the two X chromosomes in female mammalian somatic cells (**X-inactivation**)
- They display inheritance patterns like those of genes on autosomal chromosomes, rather than X-linked or Y-linked genes

As a result of their autosomal-like inheritance, the terminal X–Y homology regions are known as **pseudoautosomal regions**. We will describe them in more detail in Chapter 13 when we consider how sex chromosomes evolved in mammals.

Mitosis and meiosis: the key similarities and differences

Mitosis involves a single turn of the cell cycle. After the DNA is replicated during S phase, the two sister chromatids of each chromosome are divided equally between the daughter cells during M phase. Meiotic cell division also involves one round of DNA synthesis, but this is followed by two cell divisions without an intervening second round of DNA synthesis, allowing diploid cells to generate haploid products. While the second cell division of meiosis is identical to that of mitosis, the first meiotic division has distinct features that enable genetic diversity to arise. This relies on two mechanisms: independent assortment of paternal and maternal homologs, as well as recombination (**Table 2.2**).

TABLE 2.2 COMPARING MITOSIS AND MEIOSIS

Characteristic	Mitosis	Meiosis
Location	All tissues	Specialized germ-line cells in testis and ovary
Products	Diploid somatic cells	Haploid sperm and egg
DNA replication and cell division	Normally one round of replication per cell division	Only one round of replication per two cell divisions
Duration of prophase	Short (~30 min in human cells)	Can take decades to complete
Pairing of maternal and paternal homologs	No	Yes, during meiosis I
Recombination	Rare and abnormal	During each meiosis; normally occurs at least once in each chromosome arm after pairing of maternal and paternal homologs
Relationship between daughter cells	Genetically identical	Genetically different as a result of independent assortment of homologs and recombination

Mitochondrial DNA replication and segregation

In advance of cell division, mitochondria increase in mass, and mtDNA molecules replicate before being segregated into daughter mitochondria that then need to segregate into daughter cells. Whereas the replication of nuclear DNA molecules is tightly controlled, the replication of mtDNA molecules is not directly linked to the cell cycle.

Replication of mtDNA molecules simply involves increasing the number of DNA copies in the cell, without requiring equal replication of individual mtDNAs. That can mean that some individual mtDNAs might not be replicated and other mtDNA molecules might be replicated several times (**Figure 2.17**).

Whereas the segregation of nuclear DNA molecules into daughter cells needs to be equal and is tightly controlled, segregation of mtDNA molecules into daughter cells can be unequal. Even if the segregation of mtDNA molecules into daughter mitochondria is equal (as shown in **Figure 2.17**), the segregation of the mitochondria into daughter cells is thought to be stochastic.

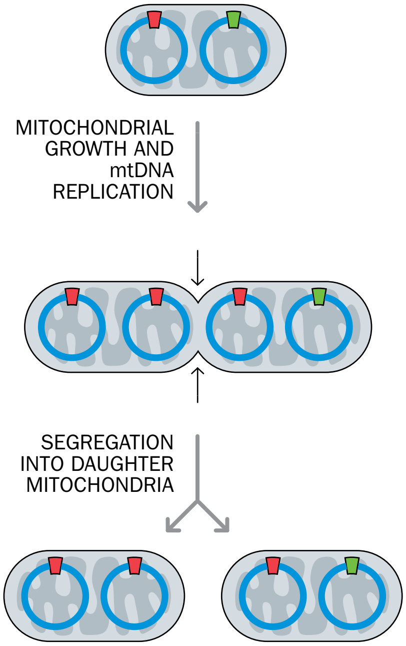

MITOCHONDRIAL GROWTH AND mtDNA REPLICATION
SEGREGATION INTO DAUGHTER MITOCHONDRIA

Figure 2.17 Unequal replication of individual mitochondrial DNAs. Unlike in the nucleus, where replication of each chromosomal DNA molecule normally produces two copies, replication of mitochondrial DNA (mtDNA) is not so tightly regulated. When a mitochondrion increases in mass in preparation for cell division, the overall amount of mitochondrial DNA increases in proportion, but individual mtDNAs replicate unequally. In this example, the mtDNA with the green tag fails to replicate and the one with the red tag replicates to give three copies. Variants of mtDNA can arise through mutation so that a person can inherit a mixed population of mtDNAs (heteroplasmy). Unequal replication of pathogenic and nonpathogenic mtDNA variants can have important consequences, as described in Chapter 16.

2.4 STRUCTURE AND FUNCTION OF CHROMOSOMES

Chromosomes have two fundamental roles: faithful transmission of genetic information and expression of the information. The processes of cell division are fascinating, and changes to the arrangement of chromosomes can have profound medical consequences. Knowledge of the detailed structure of chromosomes is crucial to understanding these vital processes.

Chromosome structure as generally illustrated in textbooks represents only the state that occurs during metaphase, while cells prepare to undergo the last stages of cell division. At this time the chromatids are still connected to each other at their centromeres and they are so condensed that they can be seen with a light microscope. But metaphase chromosomes are so tightly packed that their genes cannot be expressed. Chromosomes have a quite different structure during most of the cell cycle. Throughout interphase, most chromosome regions are comparatively very highly extended, allowing genes to be expressed.

For a chromosome to be copied and transmitted accurately to daughter cells, it requires just three types of structural elements, each of which is discussed in this section of the chapter:

- A centromere, which is most evident at metaphase—the narrowest part of the chromosome and the region at which spindle fibers attach;
- Replication origins—certain DNA sequences along each chromosome at which DNA replication can be initiated;
- Telomeres—the ends of linear chromosomes that have a specialized structure to prevent internal DNA being degraded by nucleases.

Artificial chromosomes that include large, introduced DNA fragments function normally in both yeast and mammalian cells if, and only if, they contain all three of the elements above.

Chromosomal DNA is compacted by coiling that begins with binding of histone proteins to form nucleosomes

In the eukaryotic cell, the structure of each chromosome is highly ordered. To achieve this, the large, negatively-charged nuclear DNA molecules are bound by various proteins, including both positively-charged, highly-conserved histone proteins and also non histone proteins. The DNA–protein complex is often described as **chromatin**, but certain noncoding RNAs can be intimately associated with chromosomal DNA too.

The greatest constraint on chromosome structure occurs when cells prepare to divide—the immensely long chromosomal DNA molecules must be very carefully packaged so that they do not get tangled during cell division. At metaphase, therefore, the chromosomes are extremely condensed: their linear size is about 0.01% of the length of the fully extended chromosomal DNA. Metaphase chromosomes have a protein scaffold that contains high amounts of certain non histone proteins, including topoisomerase II and protein complexes known as condensins. Condensins organize tight packaging of the chromatin, and they have been imagined to bring together distant regions of the DNA, possibly by enclosing them in ring structures, but the exact mechanism is presently unclear.

At interphase, the long part of the cell cycle that occurs between successive mitoses, the DNA is in a very highly-extended form. Nevertheless, the 2 nm thick DNA double helix is compacted to a small degree. A first level of DNA packaging involves periodic coiling of the double helix round a complex of histone proteins. The **nucleosome** has a core DNA region, uniformly 146 base pairs (bp) in length, that is wrapped around eight histone proteins (two molecules each of four core histones; Figure 2.18A and B). Adjacent nucleosomes are connected by a short stretch of linker DNA that can be as long as 114 bp (but varies between species) in transcriptionally active ("open") chromatin. A fifth type of histone, histone H1, binds to the linker DNA close to the nucleosome (see Figure 2.18A and B). Electron micrographs of suitable preparations show nucleosome filaments to have a "string-of-beads" appearance (Figure 2.18C).

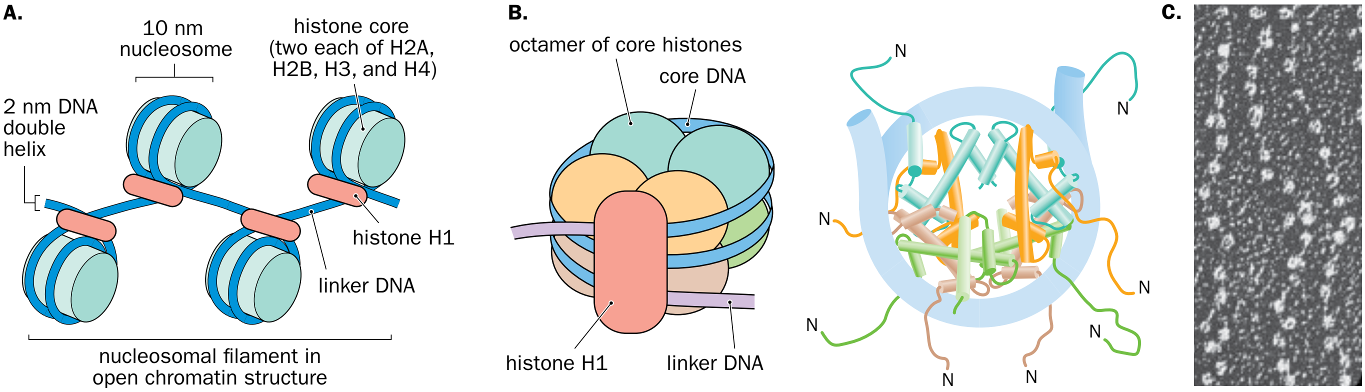


Figure 2.18 Nucleosome organization as a key step in compacting DNA in eukaryotic cells. (**A**) Binding of basic histone proteins causes the 2 nm thick DNA double helix to undergo a first level of compaction. The key structure is the nucleosome, a stretch of 146 bp of DNA wrapped in almost two turns around eight core histone proteins: two each of histones H2A, H2B, H3, and H4. A further histone, H1, is bound to linker DNA immediately outside the nucleosome; it seems to keep in place the DNA wrapped round the nucleosome. (**B**) Nucleosome detail. Left, a magnified view of a nucleosome. Right, the extensive α-helical structure of the core histones and their protruding N-terminal tails (note that many of the amino acids of the N-terminal tails are chemically modified, notably by methylation, acetylation, or phosphorylation, but are not shown here). (**C**) Electron micrograph of nucleosomal filaments showing the classic "beads-on-a-string" structure.

The N-terminal tails of the core histones protrude from the nucleosomes (Figure 2.18B). Specific amino acids in the histone tails can undergo various types of post-translational modification, notably acetylation, phosphorylation, and methylation, and so on. As a result, different proteins can be bound to the chromatin in a way that affects how the chromatin is packed and the local level of transcriptional activity. Additional histone genes encode variant forms of the core histones that may be associated with specialized functions and particular chromosomal regions, such as centromeres (see below).

Euchromatin, heterochromatin, and the variable degree of compaction of interphase chromatin

To differentiate different types of chromatin, cells can be stained with DNA-binding chemicals and analyzed by microscopy. Much of the chromatin in interphase cells (about 90% in the case of human cells) shows diffuse staining that can be seen to be dispersed through the nucleus. Chromatin like this, which stains poorly because it is in a comparatively extended state, is called **euchromatin** (Figure 2.19A). It is distinguished by relatively weak binding of histone H1 molecules and by extensive acetylation of core nucleosomal histones.

In unspecialized cells of the very early embryo, a large proportion of the euchromatin has the "open chromatin" structure shown in Figure 2.19A, the first level of DNA packaging, and the only one that will allow transcriptional activity in the cells of eukaryotes. But as cells differentiate and become specialized, the euchromatin is not so uniform: in many regions across chromosomes the euchromatin is significantly condensed, with reduced lengths of linker DNA.

The tight packing of many neighboring nucleosomes in condensed euchromatin means that RNA polymerases and transcription factors may not gain access to potential binding sites on the DNA (Figure 2.19B). It is the pattern of the open and condensed euchromatin regions across chromosomes that primarily determines which genes are expressed and which are switched off, thereby defining the identity of a cell, whether it be a lymphocyte, hepatocyte, or cardiomyocyte, and so on.

A minority of the chromatin, known as **heterochromatin**, is revealed as dark-staining regions in microscopy studies (see Figure 2.19A). It remains highly condensed (Figure 2.19C) throughout interphase, and is associated with tight binding of histone H1.

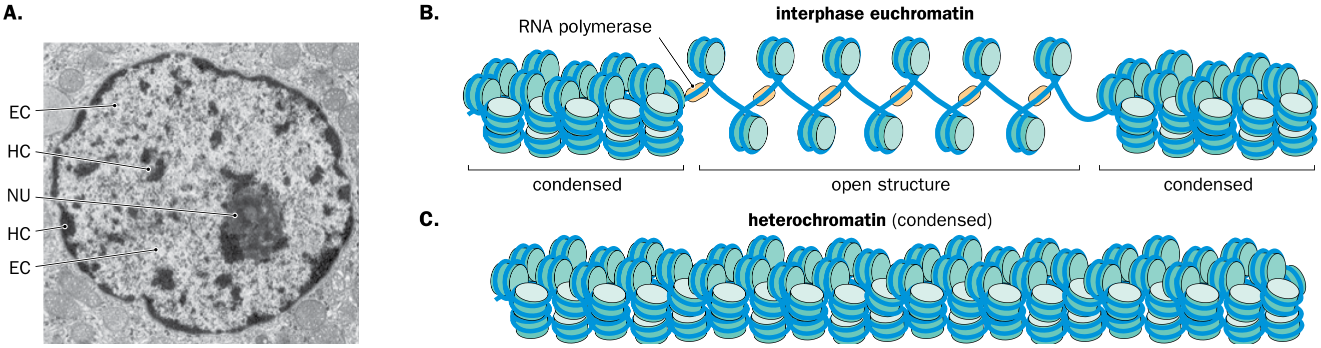


Figure 2.19 Euchromatin, heterochromatin, and higher-level packing of DNA. (**A**) Transmission electron microscopy of a typical cell nucleus clearly distinguishes the comparatively diffuse euchromatin (EC) from the electron-dense heterochromatin (HC). The euchromatin is dispersed within the interior of the nucleus. The heterochromatin is partly distributed at some interior locations, and includes nucleolus (NU)-associated heterochromatin, but just inside the nuclear envelope is a thin, electron-dense region containing the nuclear lamina and more heterochromatin. Magnification, ×26,000. (**B**) Alternating regions of open euchromatin and condensed euchromatin are typically found on chromosomes of somatic interphase cells. The open euchromatin can be accessed by RNA polymerase and the transcription machinery, but in the condensed chromatin multiple nucleosomes are tightly packed together and the lengths of linker DNA are significantly reduced. (**C**) Long regions of very highly-compacted DNA are typical of heterochromatin. (A, from Mescher AL [2018] *Junqueira's Basic Histology: Text and Atlas*, 15th edn. Republished with permission of McGraw-Hill Education; permission conveyed through Copyright Clearance Center, Inc.)

The two types of heterochromatin

Most of the heterochromatin in cells is described as **constitutive heterochromatin** because it is permanently, irreversibly condensed. The associated DNA is gene-poor and consists very largely of highly-repetitive DNA sequences, such as those found in and around the centromeres, telomeres, and over much of the Y chromosome in mammals. Constitutive heterochromatin is consistently genetically inactive in somatic cells, and if through some chromosome rearrangement an actively expressed gene is transposed from a euchromatic region to a heterochromatic region, it is silenced.

Unlike constitutive heterochromatin, **facultative heterochromatin** has a condensed structure that can be reversed (decondensed) and can be rich in genes. In each somatic cell of female mammals, for example, one of the two X chromosomes is highly condensed as a result of a process known as **X-inactivation**, and becomes a heterochromatic chromosome that migrates to the nuclear periphery. But in oogenesis, this chromosome is decondensed and reactivated. (Presumably, having both X chromosomes active is required to permit correct pairing and recombination in meiosis.) Also, both the X and the Y chromosomes become reversibly condensed for about 15 days during meiosis in spermatogenesis, forming the XY body that is segregated into a special nuclear compartment.

Each chromosome has its own territory in the interphase nucleus

The nucleus is highly organized, with many subnuclear compartments in addition to the nucleolus, where rRNA is transcribed and ribosomal subunits are assembled. The positioning of the chromosomes within the nucleus is also highly organized, as revealed by specialized techniques that analyze the movements of individual chromosomes during interphase within living cells.

The centromeres of different interphase chromosomes in human cells are less clearly aligned than they are in other organisms. They tend to cluster together at the periphery of the nucleus during the G_1 phase before becoming much more dispersed during S phase. Although the chromosomes are all in a highly-extended form, they are not extensively entwined. Instead, they appear to occupy relatively small, nonoverlapping territories (**Figure 2.20**).

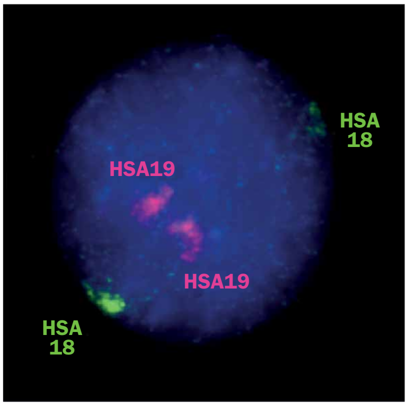


Figure 2.20 Individual chromosomes occupy distinct chromosome territories in the interphase nucleus. The nucleus of this human cell at interphase appears blue as a result of staining with DAPI, a fluorescent DNA-binding dye. DNA probes specific for gene-poor human chromosome 18 or gene-rich chromosome 19 were labeled with green or red fluorescent dyes, respectively. Within the nucleus both copies of chromosome 18 (HSA18; green signal) are seen to be located at the periphery of the nucleus, but the chromosome 19 copies (HSA19; red signal) are shown to be within the interior of the nucleus. (Courtesy of Wendy Bickmore, MRC Human Genetics Unit, Edinburgh.)

Although interphase chromosomes do not appear to have favorite nuclear locations, chromosome positioning is nevertheless nonrandom. The human chromosomes that have the highest gene density tend to concentrate at the center of the nucleus; gene-poor chromosomes are located toward the nuclear envelope (see **Figure 2.20**). Chromosome movements are probably restrained by telomere interaction with the nuclear envelope and also by internal nuclear structures (including the nucleolus in the case of chromosomes containing ribosomal RNA genes).

Centromeres play a pivotal role in chromosome movement but centromeric DNA is very different in different species

Chromosomes normally have a single **centromere**, the region where duplicated sister chromatids remain joined until anaphase. In metaphase chromosomes, the centromere is readily apparent as the primary constriction that separates the short and long arms. The centromere is essential for attaching chromosomes to the mitotic spindle and for chromosome segregation during cell division. Abnormal chromosome fragments that lack a centromere (**acentric** fragments) cannot attach to the spindle and fail to be correctly segregated to the nuclei of either daughter cell.

The centromere is effectively a chromatin structure that specifies where a large multiprotein complex, known as a **kinetochore**, will form on each sister chromatid at later prophase. The pair of kinetochores serve to tether the centromere to microtubules attached to the spindle poles (see **Box 2.3**). At anaphase, the kinetochore microtubules pull the

BOX 2.3 COMPONENTS OF THE MITOTIC SPINDLE

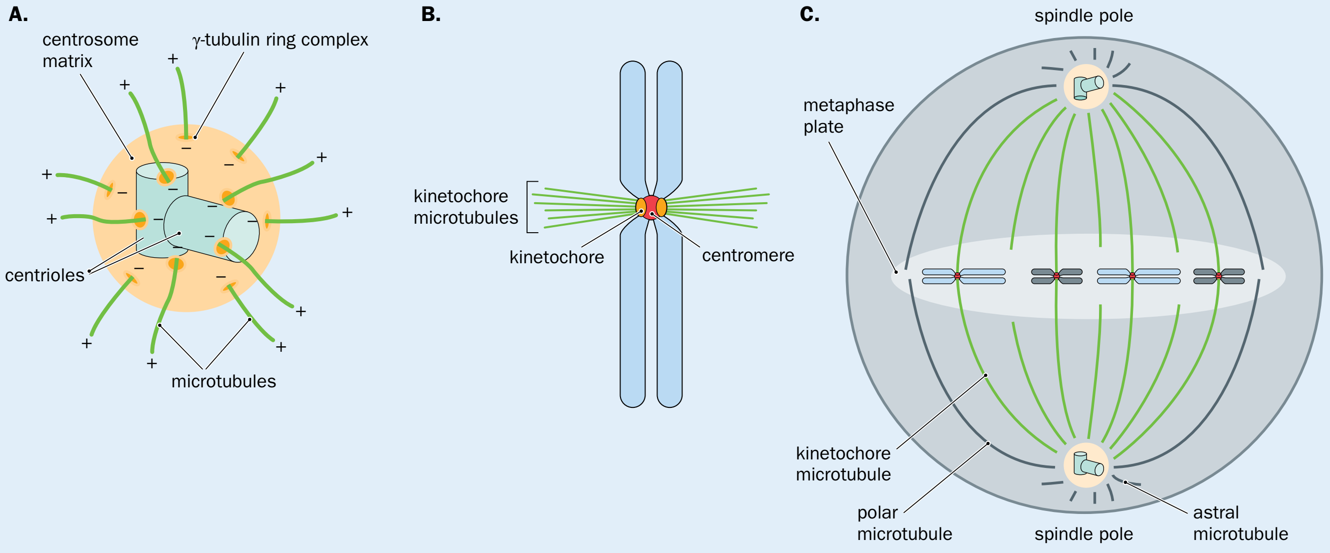
The **mitotic spindle** is formed from **microtubules** (polymers of a heterodimer of α-tubulin and β-tubulin) and microtubule-associated proteins. At each of the two spindle poles in a dividing cell is a **centrosome** that seeds the outward growth of microtubule fibers and is the major microtubule-organizing center of the cell. Because their constituent tubulins are synthesized in a particular direction, the microtubule fibers are polar, with a minus (–) end (the one next to the centromere) and a plus (+) end (the distal growing end).

Each centrosome is composed of a fibrous matrix containing a pair of **centrioles**—short, cylindrical structures composed of microtubules and associated proteins; the two centrioles are arranged at right angles to each other (**Figure 1A**). During G_1, the two centrioles in a pair separate, and during S phase, a daughter centriole begins to grow at the base of each mother centriole until it is fully formed during G_2.

The two centriole pairs remain close together in a single centrosomal complex until the beginning of M phase. At that point the centrosome complex splits in two and the two halves begin to separate. Each daughter centrosome develops its own array of microtubules and begins to migrate to one end of the cell, where it will form a **spindle pole** (**Figure 1C**).

Three different forms of microtubule fiber occur in the fully formed mitotic spindle:

- Polar fibers, which develop at prophase, extend from the two poles of the spindle toward the equator;
- Kinetochore fibers, which develop at prometaphase, connect the large multiprotein structure at the centromere of each chromatid (the **kinetochore**; **Figure 1B**) and the spindle poles;
- Astral fibers form around each centrosome and extend to the periphery of the cell.



Box 2.3 Figure 1 (**A**) Centrosome structure, (**B**) kinetochore–centromere association, and (**C**) mitotic spindle structure.

previously paired sister chromatids toward opposite poles of the spindle. The kinetochores control assembly and disassembly of the attached microtubules, which drives chromosome movement.

In the budding yeast *Saccharomyces cerevisiae*, the sequences that specify centromere function are very short, as are other functional chromosomal elements (**Figure 2.21**). The centromere element (CEN) is about 120–125 bp long and contains three principal sequence elements, of which the central one, CDE II, is particularly important for attaching microtubules to the kinetochore. A centromeric CEN fragment derived from one *S. cerevisiae* chromosome can replace the centromere of another *S. cerevisiae* chromosome with no apparent consequence.

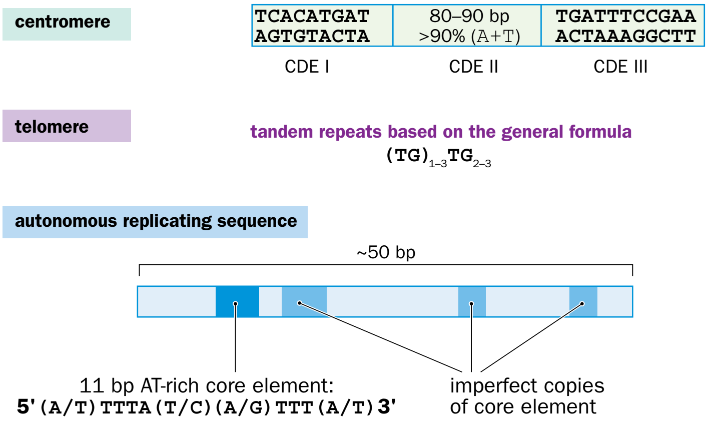


Figure 2.21 In *S. cerevisiae*, chromosome function is dependent on short, defined DNA sequence elements. *S. cerevisiae* centromeres are very short (often ~120 bp) and, unusually for eukaryotic centromeres, are composed of defined sequence elements. There are three contiguous centromere DNA elements (CDE) of which CDE II and CDE III are the most functionally important. Telomeres are composed of tandem TG-rich repeats. Autonomous replicating sequences are defined by short AT-rich sequences. The three types of short sequence shown here can be combined with foreign DNA to make an artificial chromosome in yeast cells.

S. cerevisiae centromeres are highly unusual because they are very small and the DNA sequences specify the sites of centromere assembly. They do not have counterparts in the centromeres of the fission yeast, *S. pombe*, or in those of multicellular animals. Centromere size has increased during eukaryote evolution and, in complex organisms, centromeric DNA is dominated by repeated sequences that evolve rapidly and are species-specific. The relatively rapid evolution of centromeric DNA and associated proteins might contribute to the reproductive isolation of emerging species.

Although centromeric DNA shows remarkable sequence heterogeneity across eukaryotes, centromeres are universally marked by the presence of a centromere-specific variant of histone H3, generically known as CenH3 (the human form of CenH3 is named CENP-A). At centromeres, CenH3/CENP-A replaces the normal histone H3 and is essential for attachment to spindle microtubules. Depending on centromere organization, different numbers of spindle microtubules can be attached (**Figure 2.22**).

Mammalian centromeres are particularly complex. The associated DNA often extends over several megabases and contains some chromosome-specific as well as repetitive DNA. A major component of human centromeric DNA is α-satellite DNA, whose structure is based on tandem repeats of a 171 bp monomer. Adjacent repeat units can show minor variations in sequence, and occasional tandem amplification of a sequence of several slightly different neighboring repeats results in a higher-order repeat organization. This type of α-satellite DNA is characteristic of centromeres and is marked by 17 bp recognition sites for the centromere-binding protein CENP-B.

Unlike the very small, discrete centromeres of *S. cerevisiae*, the much larger centromeres of other eukaryotes are not dependent just on sequence organization. Neither specific DNA sequences (for example, α-satellite DNA) nor the DNA binding protein CENP-B are essential or sufficient to dictate the assembly of a functional mammalian centromere. Poorly understood DNA characteristics specify a chromatin conformation that somehow, via epigenetic sequence-independent mechanisms, controls the formation and maintenance of a functional centromere.

There is little evidence for conserved sequence motifs in the replication origins of complex eukaryotes

In order for a chromosome to be copied, it needs an origin of replication, a *cis*-acting DNA sequence to which protein factors bind in preparation for initiating DNA replication. Eukaryotic origins of replication have been most comprehensively studied in yeast, where a genetic assay can be used to test whether fragments of yeast DNA can promote autonomous replication.

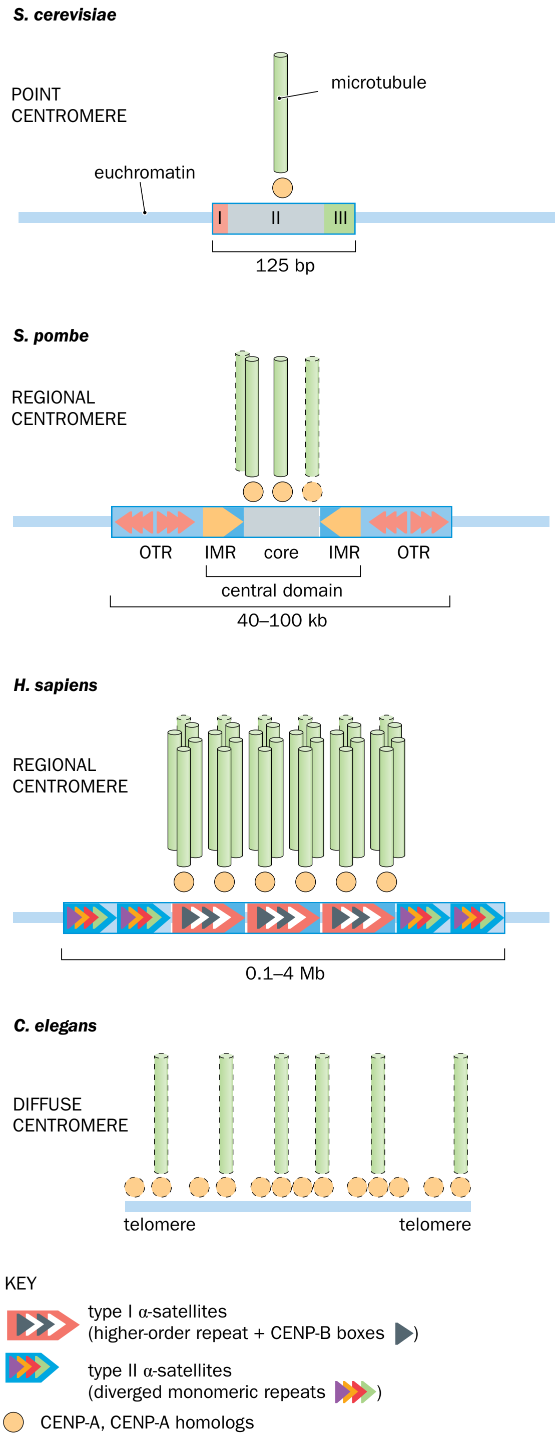


Figure 2.22 Differences in eukaryotic centromere organization. The budding yeast *S. cerevisiae* has the simplest form of centromere organization, a point centromere with just ~120 bp of DNA wrapped around a single nucleosome; each kinetochore makes only one stable microtubule attachment during metaphase. As in all centromeres, a centromere-specific variant of histone H3 (generically called CenH3), is implicated in microtubule binding. Other eukaryotes typically have defined regional centromeres that span from tens of kilobases up to a few megabases of DNA, and their kinetochores bind several microtubules. In the fission yeast *S. pombe*, microtubule attachment is centered on a nonrepetitve core sequence that is flanked by different types of repeat sequence (IMR, innermost; OTR, outermost). In humans, higher-order α-satellite DNA repeats are prominent at centromeres. In addition to binding to nucleosomes containing a CenH3 protein, CENP-A, they also have binding sites for the CENP-B protein. In some eukaryote species, such as the nematode *Caenorhabditis elegans*, the centromeres are diffuse: multiple kinetochores distributed across the length of the chromosome bind to microtubules of the metaphase spindle. (Chromosomes with diffuse centromeres are said to be *holocentric*.) Repeat direction is shown by left or right orientation of arrows and chevrons. Variable position/presence of some microtubules and CENP-A homologs is indicated by dashed outer lines. (Adapted from Vagnarelli P *et al*. [2008] *FEBS Lett* **582**:1950–1959; PMID 18435926. With permission from John Wiley & Sons, Inc. © 2008 Federation of European Biochemical Societies.)

In the yeast assay, test fragments are stitched into bacterial plasmids, together with a particular yeast gene that is essential for yeast cell growth. The hybrid plasmids are then used to transform a mutant yeast that lacks this essential gene. Transformants that can form colonies (and have therefore undergone DNA replication) are selected. Because the bacterial origin of replication on the plasmid does not function in yeast cells, the identified colonies must be cells in which the yeast DNA in the hybrid plasmid possesses an **autonomously replicating sequence** (ARS) element.

Yeast ARS elements are functionally equivalent to origins of replication and are thought to derive from authentic replication origins. They are only about 50 bp long and consist of an AT-rich region with a conserved 11 bp sequence plus some imperfect copies of this sequence (see **Figure 2.21**). An ARS encodes binding sites for both a transcription factor and a multiprotein origin of replication (ORC) complex.

In mammalian cells, the absence of a genetic assay has made it more difficult to define origins of DNA replication, but DNA is replicated from multiple initiation sites along each chromosome, with an average of roughly one initiation site per 40–80 kb of DNA. Unlike in yeast, mammalian artificial chromosomes do not require specific ARS elements. Structural motifs may be important: presumed replication origins often have guanine-rich DNA sequences with the potential to form G-quadruplexes, a four-stranded DNA structure that relies upon Hoogsteen bonding between guanines (as shown in **Figure 1.10**).

Telomeres have specialized structures to preserve the ends of linear chromosomes

Telomeres are specialized heterochromatic DNA–protein complexes at the ends of linear eukaryotic chromosomes. As in centromeres, the nucleosomes around which telomeric DNA is coiled contain modified histones that promote the formation of constitutive heterochromatin.

Telomere structure, function, and evolution

Telomeric DNA sequences are almost always composed of moderately long arrays of short tandem repeats that, unlike centromeric DNA, have generally been well conserved during evolution. In all vertebrates that have been examined, the repeating sequence is the hexanucleotide TTAGGG (**Table 2.3**). The repeats are G-rich on one of the DNA strands (the G-strand) and C-rich on the complementary strand. On the centromeric side of the human telomeric TTAGGG repeats are a further

TABLE 2.3 EVOLUTIONARY CONSERVATION OF TELOMERIC REPEAT SEQUENCES IN EUKARYOTIC CELLS

Organism(s)	Consensus telomere repeat sequence[a]
Saccharomyces cerevisiae	TG_{1-3}
Saccharomyces pombe	$TTACAG_{1-8}$
Neurospora crassa	TTAGGG
Paramecium	TTGGGG
Trypanosoma	TAGGGG
Arabidopsis	TTTAGGG
Nematodes	TTAGGC
Vertebrates	TTAGGG

[a] In the direction toward the end of the chromosome.

Note: although telomere function is conserved throughout eukaryotes and telomeric repeat sequences are generally strongly conserved, the telomeres of arthropods, such as *Drosophila*, are radically different in structure, being composed of long DNA repeats that are unrelated to the TG-rich oligonucleotide repeats found in other eukaryotes.

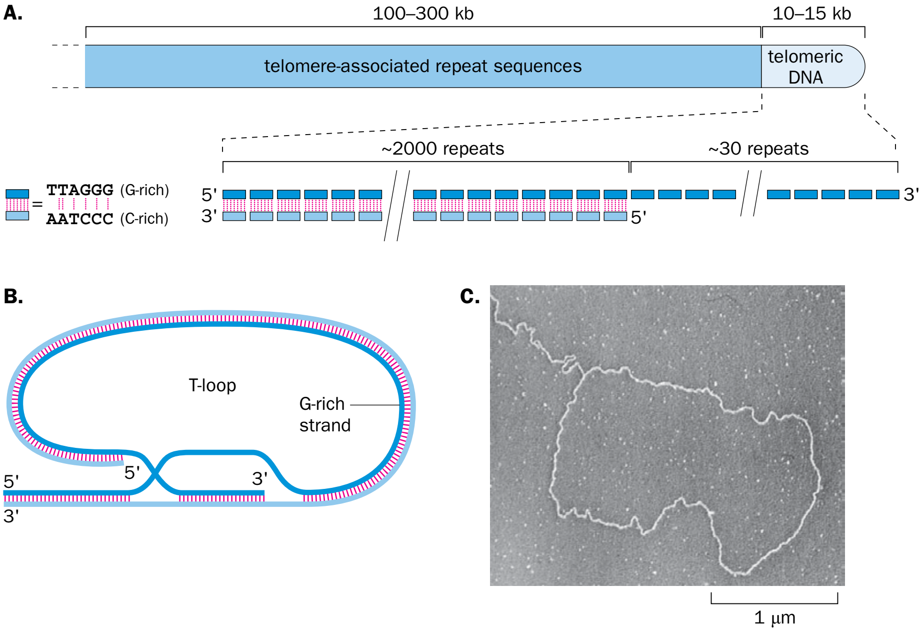


Figure 2.23 At telomeres, highly-conserved oligonucleotide repeats are bound by specialized proteins to form a protective loop. (**A**) Telomere structure. The DNA at the very ends of human chromosomes is defined by a tandem array of roughly 1700–2500 copies of the hexanucleotide TTAGGG (that is conserved in vertebrates, see **Table 2.3**). The G-rich strand, however, protrudes at the terminus to form a single-stranded region composed of ~30 TTAGGG repeats. The array of distinctive, conserved short repeats is bound by the shelterin (or telosome) complex (not shown for simplicity; two of its subunits, the **t**elomere **r**epeat binding **f**actors TRF1 and TRF2, directly bind to double-stranded regions, while POT1 can bind to the single-stranded repeats). Like centromeric DNA, telomeric DNA has modified histones that act as signals for forming constitutive heterochromatin. (**B**) T-loop formation. The single-stranded terminus of the G-rich strand can loop back and invade the double-stranded region by base-pairing with the complementary C-rich strand sequence. The resulting T-loop is thought to protect the telomere DNA from natural cellular mechanisms that repair double-strand DNA breaks. (**C**) Electron micrograph showing T-loop formation. The example shows formation of a ~15 kb T-loop at the end of an interphase human chromosome (after fixing, deproteination, and artificial thickening to assist viewing). (From Griffith JD *et al*. [1999] Cell **97**:503–514; PMID 10338214. With permission from Elsevier.)

100–300 kb of telomere-associated repeat sequences (**Figure 2.23A**). These have not been conserved during evolution, and their function is not yet understood.

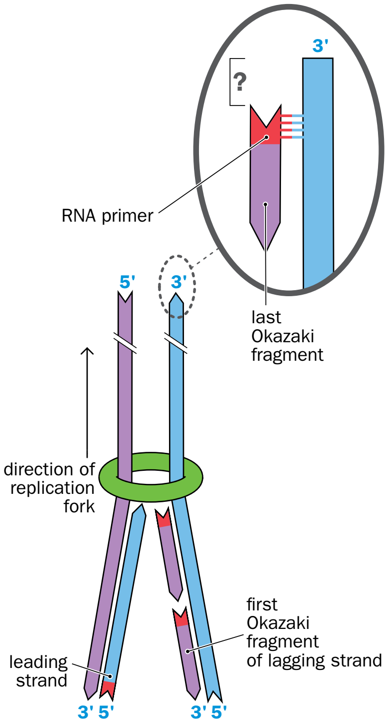
The $(TTAGGG)_n$ array of a human telomere often spans about 10–15 kb (see **Figure 2.23A**). A very large protein complex (called shelterin, or the **telosome**) contains several components that recognize and bind to telomeric DNA. Of these components, two telomere repeat binding factors (TRF1 and TRF2) bind to double-stranded TTAGGG sequences.

As a result of natural difficulty in replicating the lagging DNA strand at the extreme end of a telomere (discussed in the next section), the G-rich strand has a single-stranded overhang at its 3′ end that is typically 150–200 nucleotides long (see **Figure 2.23A**). This can fold back and form base pairs with the other, C-rich, strand to form a telomeric loop known as the T-loop (**Figure 2.23B** and **C**).

The T-loop probably represents a conserved mechanism for protecting chromosome ends. If a telomere is lost following chromosome breakage, the resulting chromosome end is unstable; it tends to fuse with the ends of other broken chromosomes, or to be involved in recombination events, or to be degraded. Telomere-binding proteins, notably the telosome component POT1, binds to single-stranded TTAGGG repeats and can protect the terminal DNA *in vitro* and perhaps also *in vivo*.

Telomerase and the chromosome end-replication problem

During DNA synthesis, the DNA polymerase extends the growing DNA chains in the 5′ → 3′ direction. One of the new DNA strands, the leading strand, grows in the 5′ → 3′ direction of DNA synthesis but the other strand, the lagging strand, is synthesized in pieces (Okazaki fragments) because it must grow in a direction opposite to that of the 5′ → 3′ direction of DNA synthesis. A succession of "backstitching" syntheses is required to produce a series of DNA fragments whose ends are then sealed by DNA ligase (**Figure 2.24**).



3'
?
RNA primer
5'
3'
last Okazaki fragment
direction of replication fork
first Okazaki fragment of lagging strand
leading strand
3' 5'
3' 5'

Figure 2.24 The problem with replicating the extreme ends of DNA in linear chromosomes. In normal DNA replication by DNA-dependent DNA polymerases, an existing DNA strand is used as a template for making a complementary new DNA strand. Here, as the replication fork advances in the upward direction, it can synthesize a continuous DNA strand, the **leading strand**, upward in the 5′ → 3′ direction from one original DNA strand (colored purple), but for the pale blue original strand, the 5′ → 3′ direction for DNA synthesis is in a direction opposite to the upward direction of the replication fork. Here, the **lagging strand** is synthesized in short pieces, called Okazaki fragments, starting from a position beyond the last fragment and moving backward toward it. (DNA-dependent DNA polymerases use short RNA primers to initiate the synthesis of DNA; the RNA primers are degraded, DNA synthesis fills in, and adjacent Okazaki fragments are ligated.) The question mark indicates a problem that is reached at the very end of the strand: how is synthesis to be completed when there can be no DNA template beyond the 3′ terminus?

Unlike RNA polymerases, DNA polymerases absolutely require a free 3′ hydroxyl group from a double-stranded nucleic acid from which to extend synthesis. This is achieved by employing an RNA polymerase to synthesize a complementary RNA primer that primes synthesis of each of the DNA fragments used to make the lagging strand. In these cases, the RNA primer requires the presence of some DNA ahead of the sequence to be copied that serves as its template. However, at the extreme end of a linear DNA molecule, there can never be a template ahead of the sequence to be copied, and a different mechanism is required to solve the problem of completing replication at the ends of a linear DNA molecule.

A solution to the end-replication problem is provided by a specialized **reverse transcriptase** (RNA-dependent DNA polymerase) that completes leading-strand synthesis. Telomerase is a ribonucleoprotein enzyme whose polymerase function is critically dependent on an RNA subunit, TERC (**te**lomerase **R**NA **c**omponent), and a protein subunit, TERT (**te**lomerase **r**everse **t**rancriptase). At the 5′ end of vertebrate TERC RNA is a hexanucleotide sequence that is complementary to the telomere repeat sequence (**Figure 2.25**). It will act as a template to prime extended DNA synthesis of telomeric DNA sequences on the leading strand. Further extension of the leading strand provides the necessary template for DNA polymerase to complete synthesis of the lagging strand.

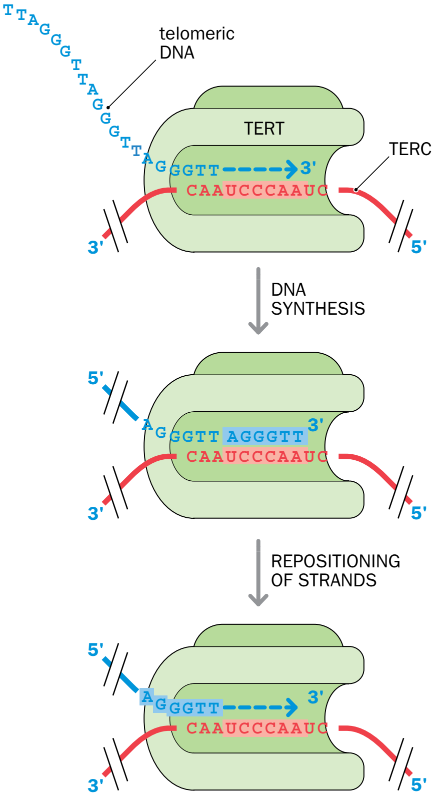


Figure 2.25 Telomerase uses a reverse transcriptase and a noncoding RNA template to make new telomere DNA repeats. The telomerase reverse transcriptase (TERT) is an RNA-dependent DNA polymerase: it uses an RNA template provided by its other subunit, TERC (telomerase RNA component). Only a small part of the RNA is used as a template—the hexanucleotide that is shaded—and so the telomeric DNA is extended by one hexanucleotide repeat (blue shading). Repositioning of the telomeric DNA relative to the RNA template allows the synthesis of tandem complementary copies of the hexanucleotide sequence in the RNA template.

In humans, telomere length is known to be highly variable and telomerase activity is largely absent from adult cells except for certain cells in highly-proliferative tissues such as the germ line, blood, skin, and intestine. In cells that lack telomerase, the extreme ends of telomeric DNA do not get replicated at S phase and their telomeres progressively shorten. Telomere shortening is effectively a way of counting cell divisions and has been related to cell senescence and aging. Cancer cells find ways of activating telomerase, leading to uncontrolled replication.

SUMMARY

- Based upon fundamental aspects of cell architecture, organisms can be divided into prokaryotes (unicellular organisms that have a simple structure) and eukaryotes (organisms with complex cells that have organelles, internal membranes, and a cytoskeleton).
- Prokaryotes can be further divided into two domains of life, bacteria and archaea, that are as different from each other as they are from the third domain of life, the eukaryotes.
- Archaea have similar information processing systems (DNA replication, transcription, recombination, and repair) to eukaryotes; bacteria resemble eukaryotes more in terms of operational functions (metabolism and so on).
- Eukaryotic cells evolved by endosymbiosis, a type of cell fusion in which an anaerobic archaeon engulfed an aerobic bacterium to produce a stable cell with two genomes and two sets of protein-synthesis machinery. The internalized cell eventually gave rise to mitochondria.
- Our cells show extraordinary diversity in size, form, and function. Histology recognizes over 200 different cell types in adult humans but this is a gross underestimate of cell diversity.
- All cells arise from other cells, almost always by cell division. In multicellular organisms, cell division is not just required for growth during development: it is still needed in the mature organism for replacing short-lived cells.
- Chromosomes have two fundamental roles: the faithful transmission of genetic information in the nucleus and the appropriate expression of that information.
- Eukaryotic chromosomes consist of linear, double-stranded DNA molecules bound to both histone and non-histone proteins that serve both structural roles and regulatory roles. The DNA–protein matrix is known as chromatin.
- Mitochondria and the vast majority of prokaryotic cells have circular, double-stranded DNA that is relatively protein-free.
- Chromosomes undergo major changes in the cell cycle, notably at S phase when they replicate and at M phase when the replicated chromosomes become separated and allocated to two daughter cells.
- DNA replication at S phase produces two double-stranded daughter DNA molecules that are held together at a specialized region, the centromere. When the daughter DNA molecules remain held together like this they are known as sister chromatids, but once they separate at M phase they become individual chromosomes.
- At the metaphase stage of M phase, the chromosomes are so highly condensed that gene expression is uniformly shut down. This is the optimal time for viewing them under the microscope.
- During interphase, the long period of the cell cycle that separates successive M phases, chromosomes have generally very long, extended conformations. Genes can be expressed efficiently but the chromosomes are too extended to be viewed by optical microscopy.
- Even during interphase some chromosomal regions always remain highly condensed and transcriptionally inactive (heterochromatin) while others are extended to allow gene expression (euchromatin).
- Sperm and egg cells have one copy of each chromosome (haploid cells), but most of our cells are diploid cells, having two sets of chromosomes.
- Fertilization of a haploid egg by a haploid sperm generates the diploid zygote from which all other body cells arise via cell division.
- In mitosis, a cell divides to give two daughter cells, each with the same number and types of chromosomes as the original cell.
- Meiosis is a specialized reductive cell division used to produce genetically unique haploid sperm and egg cells, and is confined to certain cells of the testes and ovaries. During meiosis, new genetic combinations are randomly created, partly by exchanging sequences between maternal and paternal chromosomes.
- Three types of functional element are needed for eukaryotic chromosomes to transmit DNA faithfully from mother cell to daughter cells: the centromere (ensures correct chromosome segregation at cell division); replication origins (initiate DNA replication); and telomeres (cap the chromosomes to stop the internal DNA being degraded by nucleases).

FURTHER READING

Cell architecture and evolution

Alberts B *et al.* (2014) *Molecular Biology of the Cell*, 6th edn. Garland Science.

Booth A & Doolittle WF (2015) Eukaryogenesis, how special really? *Proc Natl Acad Sci USA* **112**:10278–10285; PMID 25883267.

Eme L & Doolittle WF (2015) Archaea. *Curr Biol* **25**:R851–855; PMID 26439345.

Friedland JR & Nunnari J (2014) Mitochondrial form and function. *Nature* **505**:335–343; PMID 24429632.

Harold FM (2014) *In Search of Cell History: The Evolution of Life's Building Blocks*. University of Chicago Press.

Koonin EV & Yutin N (2014) The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb Perspect Biol* **6**:a016188; PMID 24691961.

Libby E *et al.* (2014) Geometry shapes evolution of early multicellularity. *PLoS Comput Biol* **10**:e1003803; PMID 25233196.

Chromosome structure, function, and recombination

Chan SR & Blackburn EH (2004) Telomeres and telomerase. *Philos Trans R Soc Lond B Biol Sci* **359**:109–121; PMID 15065663.

Cremer T & Cremer M (2010) Chromosome territories. *Cold Spring Harb Perspect Biol* **2**:a003889; PMID 20300217.

Fragkos M *et al*. (2015) DNA replication origin activation in space and time. *Nat Rev Mol Cell Biol* **16**:360–374; PMID 25999062.

Fukagawa T & Earnshaw WC (2014) The centromere: chromatin foundation for the kinetochore machinery. *Dev Cell* **30**:496–508; PMID 25203206.

Henikoff S *et al*. (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**:1098–1102; PMID 11498581.

Marks AB *et al*. (2016) Replication origins: determinants or consequences of nuclear organization? *Curr Opin Genet Dev* **37**:67–75; PMID 26845042.

Mézard C *et al*. (2015) Where to cross? New insights into the location of meiotic crossovers. *Trends Genet* **31**:393–401; PMID 25907025.

Ozer G *et al*. (2015) The chromatin fiber: multiscale problems and approaches. *Curr Opin Struct Biol* **31**:124–139; PMID 26057099.

Riethman H (2008) Human telomere structure and biology. *Annu Rev Genomics Hum Genet* **9**:1–19; PMID 18466090.

Scott KC & Sullivan BA (2014) Neocentromeres: a place for everything and everything in its place. *Trends Genet* **30**:66–74; PMID 24342629.

Fundamentals of cell–cell interactions and immune system biology

3

We introduced basic aspects of cells in Chapter 2 but we focused mostly on cells in isolation, how they are organized, how they have evolved, and the basics of the cell cycle and cell division. In this chapter we now examine select aspects of how cells interact with other cells in multicellular organisms. We continue the theme in Chapter 4 when we consider cell–cell interactions during early development.

In addition to responding to externally administered changes in their environment, the cells of a multicellular organism need to signal to each other so that they can interact in specific ways. Different types of cell signaling are used, as outlined in Section 3.1, but in molecular signaling pathways there is a common principle: a signal molecule produced by a transmitting cell is bound by a specific receptor on a responding cell, thereby activating a cell signaling pathway within the responding cell. The end result is a change in gene expression in the responding cell, causing it to change its behavior in some way.

Throughout development and life, cells die and new cells are formed. There needs to be tight control on cell proliferation to reduce the risk of tumors that threaten survival, and cell signaling pathways provide the necessary regulation. Our cells also need to receive signals from other cells, simply to survive. That is, the default state is an in-built cell suicide pathway that needs to be actively overcome by cell survival signals obtained from other cells. We consider cell signaling pathways that regulate cell proliferation and cell death in Section 3.2.

Cells also need to co-operate to form the tissues of the body, and we describe the key cell adhesion mechanisms and the importance of the extracellular matrix in Section 3.3. Additionally, cell–cell interactions are crucially important in mounting immune responses, and we introduce this topic against a general background of the biology of the immune system in Section 3.4. However, we cover the mechanisms responsible for extraordinary genetic variation in the immune system in Chapter 10 when we take a broad look at human genetic variation.

3.1 PRINCIPLES OF CELL SIGNALING

Single-celled organisms usually function independently, but in multicellular organisms the cells co-operate with each other for the benefit of the organism. To coordinate and regulate physiological and biochemical functions, the cells must communicate effectively, constantly sending and receiving signals. The roles of cell signaling in various types of cell–cell interaction will be covered in later sections; here we describe some of the underlying principles.

Signaling molecules bind to specific receptors in responding cells to trigger altered cell behavior

In a multicellular organism, virtually all aspects of cell behavior—metabolism, movement, proliferation, differentiation—are regulated by cell signaling. Transmitting cells produce signaling molecules that are recognized by responding cells, causing them to change their behavior. Intercellular signaling can take place over long distances where the transmitting cell and responding cell are far removed from each other, as when the signaling molecule is a hormone that is secreted and then must travel some distance to be received by responding cells.

Classes of cell signaling

In most cases, the transmitting and responding cells are in close proximity, and here three types of signaling mechanism can be distinguished, as listed below. In the first two cases, the signaling produces a change in gene expression in the responding cell, but in synaptic signaling the result is a change in electric potential of the cell membrane.

- ***Paracrine signaling.*** A cell sends a secreted signaling molecule that diffuses over a short distance to bind to receptors on responding cells in the local neighborhood.
- ***Juxtacrine signaling.*** The transmitting cell is in direct contact with the responding cell; the signaling molecule is tethered to the surface of the transmitting cell, and is bound by a receptor on the surface of the responding cell.
- ***Synaptic signaling.*** This specialized form of signaling occurs between adjacent neurons or between adjacent neuron and muscle cells, and produces changes in membrane potential, notably depolarization.

In paracrine and juxtacrine cell signaling, the transmitting cells and the responding cells are usually different cell types. However, an alternative is for the signaling molecule to bind to a receptor on the surface of the transmitting cell or identical neighbor cells. This type of autocrine signaling can be used to reinforce a signaling decision, or to co-ordinate decisions by groups of the same kind of cell.

Different types of action by signaling molecules and receptors

Some small, hydrophobic signaling molecules can pass directly through the plasma membrane of the responding cell and bind to intracellular receptors (**Figure 3.1B**). In many examples of cell signaling, however, the signaling molecule cannot cross the cell membrane and works by binding to a receptor on the surface of the responding cell. In those cases, the signal is often a soluble molecule that diffuses a short distance before binding to its receptor (**Figure 3.1A**). In juxtacrine signaling, however, the signal molecule is anchored in the plasma membrane of the transmitting cell (**Figure 3.1C**). **Table 3.1** provides some examples of different cell signaling systems in vertebrates; we will consider details of some mechanisms in the sections below.

How gene expression is altered

The endpoint of most cell signaling is altered gene expression producing behavioral changes in the responding cells. The usual key to this change, the final step in signal transduction, is activation of a specific transcription factor so that it selectively binds to the DNA of certain target genes to modulate gene expression. One part of a transcription factor protein is used to recognize and bind the target DNA sequence; another part is used to activate gene expression (**Box 3.1**).

Activation of the transcription factor that causes the change in gene expression is indirect in the case of signaling pathways that use a cell surface receptor. Binding of a signaling molecule to the receptor induces a change in the receptor's cytoplasmic domain. The alteration in the receptor activates a signal-transduction pathway that typically culminates in activation (or sometimes inhibition) of a transcription factor. However, as described in the next section, a signaling molecule that passes through the cell membrane and binds directly to an intracellular receptor causes a conformational change in the receptor that directly induces it to become an active transcription factor.

Some signaling molecules bind intracellular receptors that activate target genes directly

Small, hydrophobic signaling molecules such as steroid hormones are able to diffuse through the plasma membrane of the target cell and to bind intracellular receptors in the nucleus or cytoplasm. These receptors, often called nuclear hormone receptors, are therefore inducible transcription factors. Following ligand binding, the receptor protein is activated and associates with a specific DNA response element located in the promoter regions of perhaps 50–100 target genes and, with the help of suitable co-activator proteins, activates their transcription.

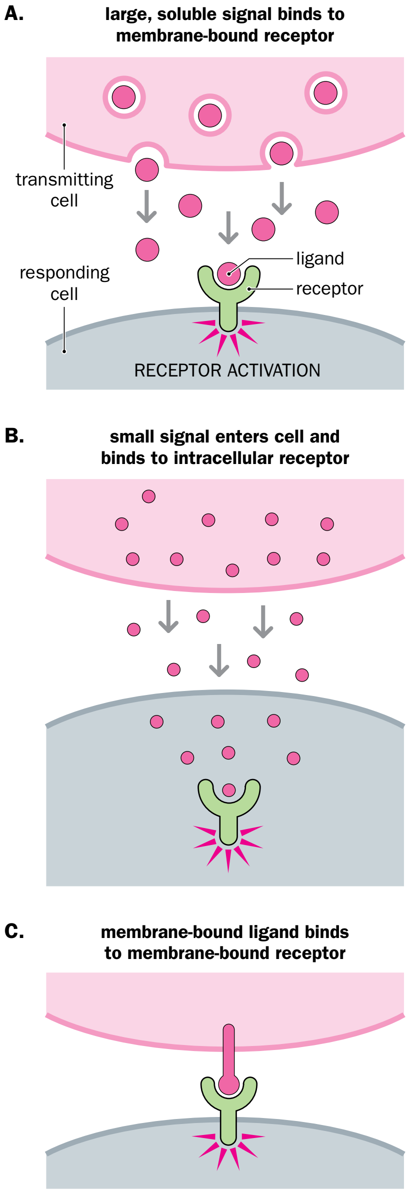

A.
large, soluble signal binds to membrane-bound receptor
transmitting cell
responding cell
ligand
receptor
RECEPTOR ACTIVATION
B.
small signal enters cell and binds to intracellular receptor
C.
membrane-bound ligand binds to membrane-bound receptor

Figure 3.1 Three types of relationship between ligand and receptor in cell signaling. (**A**) Soluble ligand and cell surface receptor. In this type of paracrine signaling, the ligand is often a protein that binds to a transmembrane protein receptor, activating its cytoplasmic tail. (**B**) Small, soluble ligand and intracellular receptor. In this type of paracrine signaling, the ligand may be a gas (such as nitric oxide) or a protein or steroid that is so small that it can freely pass through membranes. (**C**) Membrane-bound ligand and cell surface receptor on adjacent cells (juxtacrine signaling). See **Table 3.1** for examples.

TABLE 3.1 IMPORTANT CLASSES OF VERTEBRATE CELL SIGNALING MOLECULES AND THEIR RECEPTORS

Signaling molecule	Receptor	General figures and comments
SOLUBLE SIGNALING MOLECULES UNABLE TO CROSS CELL MEMBRANES	CELL SURFACE RECEPTORS	FIGURE 3.1A
Some hormones (e.g. insulin), some growth factors (e.g. FGFs, EGFs), ephrins	Receptors with intrinsic kinase activity	Many (FGF, EGF, and ephrin signaling widely used in embryonic development)
Cytokines (e.g. interleukins, interferons), some hormones and growth factors (e.g. growth hormone, prolactin, erythropoietin)	Receptors with associated tyrosine kinase activity. Internal domains have associated JAK protein (**Figure 3.4**)	Many
Various hormones (e.g. epinephrine, serotonin, glucagon, FSH), opioids, neurokinins, histamine	G-protein-coupled receptors (**Figure 3.5**). Internal domains have associated GTP-binding proteins	Signaling involving olfactory receptors, taste receptors, rhodopsin receptors, and so on
Neurotransmitters	Ion-channel-coupled receptors	Many
TGFβ family	Receptors with associated serine/threonine kinase activity. Internal domain associated with SMAD proteins	Many examples in signaling during embryonic development
Wnt and Hedgehog families	Frizzled (Wnt) and Patched (Hedgehog)	
SOLUBLE SIGNALING MOLECULES THAT CROSS THE CELL MEMBRANE	INTERNAL RECEPTORS	FIGURE 3.1B
Steroid hormones, retinoids	Receptors may be in cytoplasm or nucleus. They are converted to active transcription factors when bound by their ligand	Signaling using nuclear hormone receptors (**Figures 3.2** and **3.3**)
MEMBRANE-BOUND LIGANDS	CELL SURFACE RECEPTOR	FIGURE 3.1C
Death signals	Death receptors	Fas signaling in apoptosis (**Figure 3.10**)
Delta/Serrate family	Notch	Important in neural development

FGFs, fibroblast growth factors; EGFs, epidermal growth factors; FSH, follicle-stimulating hormone; TGF, transforming growth factor.

BOX 3.1 TRANSCRIPTION FACTOR STRUCTURE

Genes are regulated by a variety of protein transcription factors that recognize and bind a short nucleotide sequence in DNA. Eukaryotic transcription factors generally have two distinct functions located in different parts of the protein:

- A **DNA-binding domain** that allows the transcription factor to bind to a specific sequence element in a target gene;
- An **activation domain** that stimulates transcription of the target gene (probably by interacting with basal transcription factors in the transcription complex on the promoter).

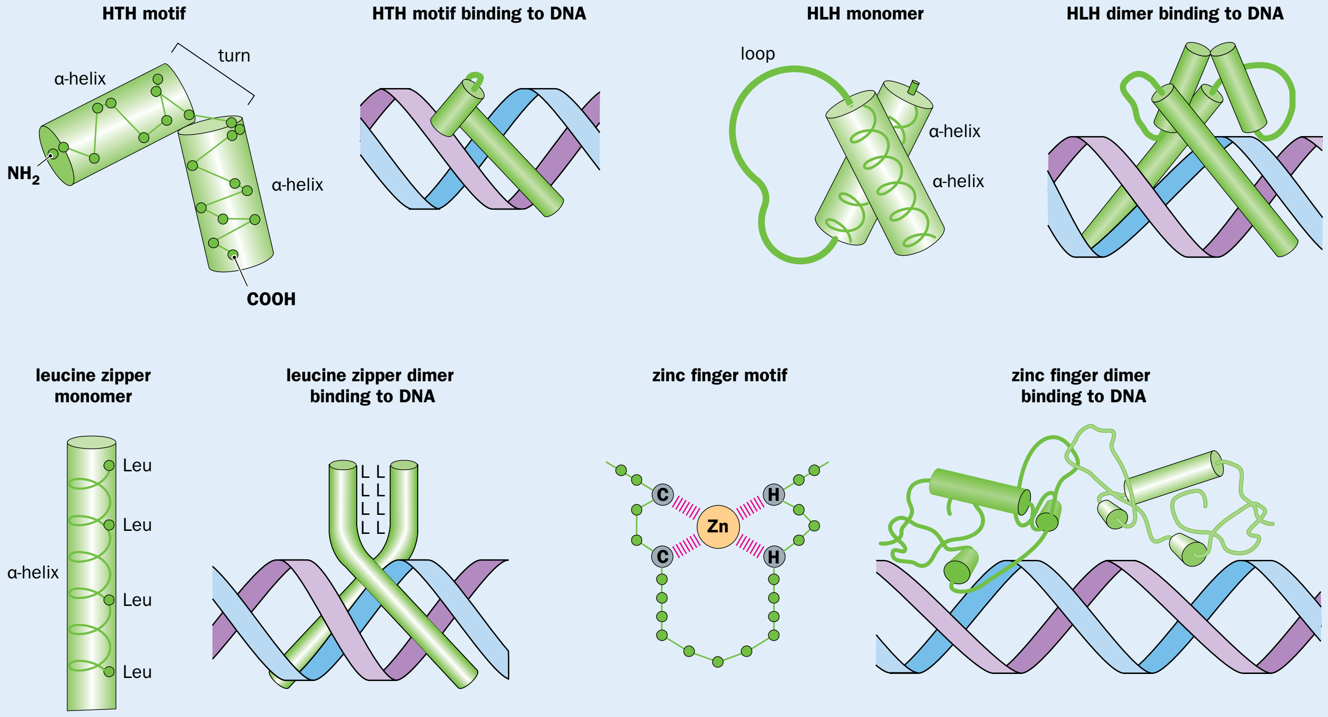
Some proteins, such as steroid hormone receptors, have only a DNA-binding domain but, after binding a ligand, they can co-operate with proteins called **co-activators** to carry out the activities of a transcription factor. See the example of the glucocorticoid receptor in Figure 3.3 in the main text.

COMMON DNA-BINDING MOTIFS IN TRANSCRIPTION FACTORS

Several common protein structural motifs have been identified, most of which use α-helices (or occasionally β-sheets) to bind to the major groove of DNA. While such structural motifs provide the basis for DNA binding, the precise sequence of the DNA-binding domain determines the DNA sequence-specific recognition. Most transcription factors bind to DNA as dimers (often homodimers), and the DNA-binding region is often distinct from the region specifying dimer formation.

The **helix-turn-helix** (**HTH**) motif (Figure 1) is a common motif found in transcription factors. It consists of two short α-helices separated by a short amino acid sequence that induces a turn, so that the two α-helices are oriented in different planes. Structural studies have suggested that the C-terminal helix (shown to the right in the image showing DNA binding) acts as a specific recognition helix because it fits into the major groove of the DNA, controlling the precise DNA sequence that is recognized.

The **helix-loop-helix** (**HLH**) motif also consists of two α-helices, but this time connected by a flexible loop that, unlike the short turn in the HTH motif, is flexible enough to permit folding back so that the two helices can pack against each other (that is, the two helices lie in planes that are parallel to each other). The HLH motif mediates both DNA binding and protein dimer formation. Heterodimers comprising a full-length HLH protein and a truncated HLH protein, which lacks the full length of the α-helix necessary to bind to the DNA, are unable to bind DNA tightly. As a result, HLH



Box 3.1 Figure 1 Structural motifs commonly found in transcription factors and DNA-binding proteins. HTH, helix-turn-helix motif; HLH, helix-loop-helix motif.

heterodimers are thought to act as a control mechanism, by enabling inactivation of specific gene regulatory proteins.

The **leucine zipper** is a helical stretch of amino acids, rich in hydrophobic leucine residues aligned on one side of the helix. These hydrophobic patches allow two individual α-helical monomers to join together over a short distance to form a coiled coil. Beyond this region, the two α-helices separate, so that the overall dimer is a Y-shaped structure. The dimer is thought to grip the double helix much like a clothes peg grips a clothes line. Leucine zipper proteins normally form homodimers but can occasionally form heterodimers. The latter provides an important combinatorial control mechanism in gene regulation.

The **zinc finger** motif involves binding of a zinc ion by four conserved amino acids (normally either histidine or cysteine) so as to form a loop (finger), which is often tandemly repeated. The so-called C2H2 (Cys_2/His_2) zinc finger typically comprises about 23 amino acids with neighboring fingers separated by a stretch of about seven or eight amino acids. The structure of a zinc finger may consist of an α-helix and a β-sheet held together by coordination with the Zn^{2+} ion, or of two α-helices, as shown in the image of the zinc finger binding to DNA (**Figure 1**). In either case, the primary contact with the DNA is made by an α-helix binding to the major groove.

The receptors for steroid hormones, and also those for the signaling molecules thyroxine and retinoic acid, belong to a common nuclear receptor superfamily. Each receptor in this superfamily contains a centrally located DNA-binding domain of about 68 amino acids, and a ligand-binding domain of about 240 amino acids located close to the C-terminus (**Figure 3.2A**). The DNA-binding domain contains structural motifs known as zinc fingers (see **Box 3.1**) and binds as a dimer, with each monomer recognizing one of two hexanucleotides in the response element. The two hexanucleotides are either inverted repeats or direct repeats, typically separated by three or five nucleotides (**Figure 3.2B**).

The nuclear hormone receptors are normally found in the cytoplasm. They are transcription factors that in the absence of bound ligand are maintained in an inactive state: either the ligand-binding domain directly represses the DNA-binding domain, or the receptor is bound to an inhibitory protein (as in the case of the glucocorticoid receptor; see **Figure 3.3**). Binding of the ligand to the receptor overcomes the inhibition and the activated ligand–receptor complex migrates to the nucleus, where it works as a specific transcription factor.

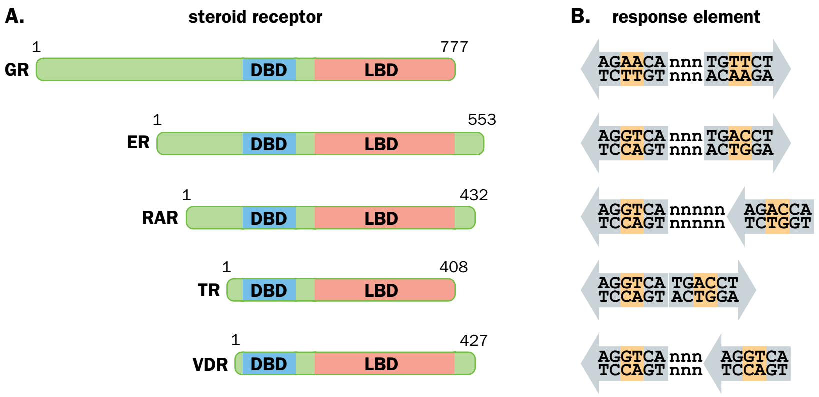


Figure 3.2 The nuclear receptor superfamily. (**A**) Members of the nuclear receptor superfamily all have a similar structure, with a central DNA-binding domain (DBD) and a C-terminal ligand-binding domain (LBD). Numbers refer to the protein size in amino acid residues. GR, glucocorticoid receptor; ER, estrogen receptor; RAR, retinoic acid receptor; TR, thyroxine receptor; VDR, vitamin D receptor. (**B**) The response elements recognized by the nuclear receptors also have a conserved structure, with two hexanucleotide recognition sequences typically separated by either three or five nucleotides (n). The hexanucleotide sequences have the general consensus of AGNNCA with the two central nucleotides (pale orange shading) conferring specificity and belonging to one of three classes: AA, AC, or GT.

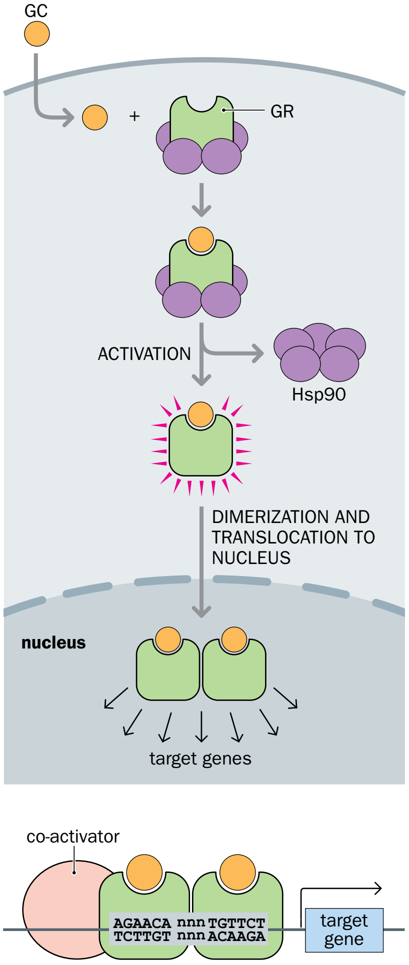


Figure 3.3 Cell signaling by ligand activation of an intracellular receptor. A glucocorticoid (GC), like other hydrophobic hormones, can pass through the plasma membrane and bind to a specific intracellular receptor. The glucocorticoid receptor (GR) is normally bound to an Hsp90 inhibitory protein complex and is found within the cytoplasm. After binding of the receptor to glucocorticoid, however, the inhibitory complex is released and the now-activated receptor forms dimers and translocates to the nucleus. Here, it works as a transcription factor by specifically binding to a particular response element sequence in target genes (shown in the lower panel; see also **Figure 3.2B**) and by activating the target genes with the co-operation of specific co-activator proteins.

Signaling through cell surface receptors often involves kinase cascades

The plasma membranes of animal cells positively bristle with transmembrane receptors for signaling molecules. The primary cilium, an extension of the membrane that protrudes into the external environment (see **Box 2.1 Figure 1**), is especially endowed with such receptors. This previously mysterious organelle is now thought to play a major role as a sensor, using the multiple different receptors to sense, and respond to, alterations in the extracellular environment.

When a receptor binds its signaling molecule on the external surface of the cell, a conformational change is induced in its internal (cytoplasmic) domain. This change alters the properties of the receptor, perhaps affecting its contact with another protein, or stimulating some latent enzyme activity. For many signaling receptors, either the receptor, or an associated protein, has integral kinase activity. The activated kinase often causes the receptor to phosphorylate itself and then allows it to phosphorylate other proteins inside the cell, thereby activating them. These target proteins are themselves often kinases. They, in turn, can phosphorylate and thereby activate proteins further down a signal-transduction pathway, resulting in a *kinase cascade*.

Eventually, a transcription factor is activated (or inhibited, as appropriate), resulting in a change in gene expression. The length of the signaling cascade can be short (as in the cytokine-regulated JAK-STAT pathway; see **Figure 3.4**) or have many steps, such as the MAP kinase pathway. Pathways where receptors do not have kinase activity often have several downstream components with either kinase or phosphorylase activity.

Small intermediate intracellular signaling molecules in signal transduction

Signal-transduction pathways initiated by binding of a signaling molecule to a cell surface receptor are often complex. In addition to the kinases and other enzymes mentioned above, the cascade of interacting molecules involved in communicating the signal within the cell can include various small, diffusible intracellular signaling molecules that act as intermediates in signal transduction. These are known as

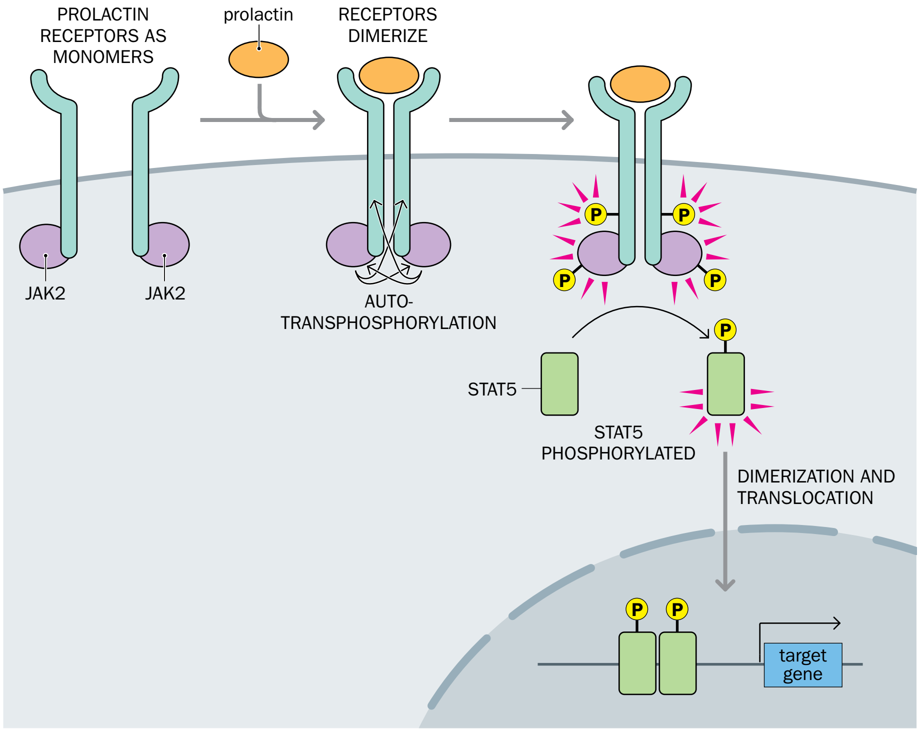

PROLACTIN RECEPTORS AS MONOMERS
prolactin
RECEPTORS DIMERIZE
JAK2
JAK2
AUTO-TRANSPHOSPHORYLATION
P
STAT5
STAT5 PHOSPHORYLATED
DIMERIZATION AND TRANSLOCATION
target gene

Figure 3.4 Ligand activation of a plasma membrane receptor. JAK-STAT signaling involves JAK kinases that are bound to the cytoplasmic domain of certain transmembrane receptors (notably cytokine receptors), and members of the STAT transcription factor family. For example, as shown here, JAK2 is bound to the prolactin receptor. Prolactin induces monomeric receptors to dimerize, bringing two JAK2 kinase molecules into close proximity. Each JAK2 molecule then cross-phosphorylates the other and the cytoplasmic domain of its attached receptor. Phosphorylation of the receptors activates a dormant receptor kinase activity that then targets a specific STAT protein, in this case STAT5. Once activated by phosphorylation, STAT5 dimerizes, translocates to the nucleus, and activates the transcription of various target genes.

TABLE 3.2 EXAMPLES OF SECOND MESSENGERS IN CELL SIGNALING

Class	Examples	Origin	Role
Hydrophilic (cytosolic)	Cyclic AMP (cAMP)	Produced from ATP by adenylate cyclase	Effects are usually mediated through protein kinase
	Cyclic GMP (cGMP)	Produced from GTP by guanylate cyclase	Best characterized role is in visual reception in the vertebrate eye
	Ca^{2+}	External sources or released from intracellular stores in the endoplasmic reticulum	Ca^{2+} channels can be voltage-gated and receptor-operated. Probably the most widely used intracellular messenger (see **Figure 3.5**)
Hydrophobic (membrane-associated)	PIP_2 (phosphatidylinositol 4,5-bisphosphate) DAG (diacylglycerol) IP_3 (inositol trisphosphate)	PIP_2 is a phospholipid that is enriched at the plasma membrane and can be cleaved to generate DAG on the inner layer of the plasma membrane while releasing IP_3 into the cytoplasm	IP_3 binds to receptors in the endoplasmic reticulum, thereby causing the release of Ca^{2+} into the cytosol. Ca^{2+}-dependent protein kinase C is thereby recruited to the plasma membrane (see **Figure 3.5**)

second messengers (the first messenger being the extracellular signaling molecule), see **Table 3.2** for some examples.

Second messengers are a feature of pathways that use G-protein-coupled receptors (GPCRs). This large family of membrane-bound receptors (encoded by over 1000 genes in mammals) includes many receptors for prostaglandins and related lipids, various neurotransmitters, neuropeptides, and peptide hormones; other GPCRs are responsible for relaying the sensations of sight, smell, and taste.

GPCRs are distinguished by having a transmembrane domain that passes through the plasma membrane seven times and a cytoplasmic domain that can bind a G-protein. G-proteins are membrane-bound proteins with three subunits, α, β, and γ, extending into the cytoplasm, and are so called because the α subunit can bind GDP (keeping the G-protein inactive) or GTP. In the absence of ligand, a GPCR can bind an inactive G-protein (with GDP bound to its α subunit), but binding of ligand to the GPCR stimulates the G-protein by causing the α subunit to release GDP and bind GTP instead, causing the α subunit to dissociate from the βγ dimer and leading to activation of both (**Figure 3.5A**).

Each of the activated α and βγ units can then interact with proteins downstream in the signal-transduction pathway, and according to the type of G-protein, different

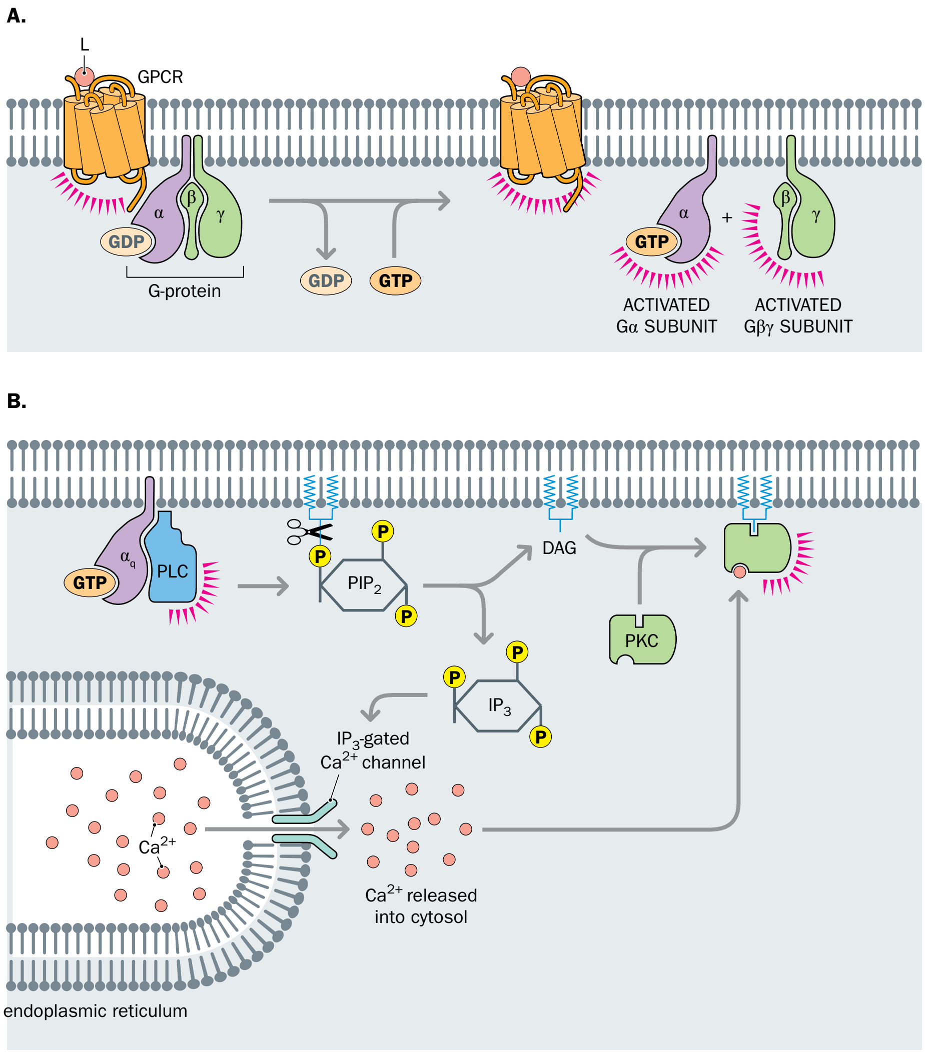


Figure 3.5 Principles of G-protein signaling. (**A**) G-protein-coupled receptors (GPCRs) have seven transmembrane helices and a short cytoplasmic region bound by an associated G-protein with three subunits: α, β, and γ. Binding of ligand (L) to the extracellular domain of a GPCR activates its cytoplasmic domain and causes exchange of GTP for GDP on the G-protein. As a result, the Gα subunit is activated and dissociates from the Gβγ dimer, which in turn becomes activated. (**B**) The G-protein subunit $G\alpha_q$ uses various lipids and Ca^{2+} as second messengers. Activation of a type of G-protein α subunit known as $G\alpha_q$ causes it to bind and activate phospholipase C (PLC). Activated PLC then migrates along the plasma membrane to bind and cleave membrane-bound phosphatidylinositol 4,5-bisphosphate (PIP_2). The reaction leaves a diacylglycerol (DAG) residue embedded in the membrane and liberates inositol 1,4,5-trisphosphate (IP_3). The released IP_3 diffuses to the endoplasmic reticulum, where it promotes the opening of an IP_3-gated calcium ion channel, causing an efflux of Ca^{2+} from stores in the endoplasmic reticulum. With the help of Ca^{2+}, the membrane-bound DAG activates protein kinase C (PKC), which is then recruited to the plasma membrane where it phosphorylates target proteins that differ according to cell type.

second messengers can be involved in signal transduction. Some G-proteins stimulate or inhibit the membrane-bound enzyme adenylate cyclase, causing a change in intracellular cyclic AMP (cAMP) levels. Other G-proteins stimulate the production of lipids (such as inositol 1,4,5-trisphosphate and diacylglycerol) and the release of calcium ions (**Figure 3.5B**). In turn, the second messengers activate downstream protein kinases such as cAMP-dependent protein kinase A and calcium-dependent protein kinase C, which go on to phosphorylate particular transcription factors, causing them to change their activity.

There is extensive crosstalk between different signaling pathways. At any moment, the response given by a particular cell depends on the sum of all signals that it receives and the nature of the receptors available to it.

Synaptic signaling does not require activation of transcription factors

Signaling between neurons needs to occur extremely rapidly and is achieved by synaptic signaling using chemical synapses (the most common form) or electrical synapses.

- Chemical synapses. Here, the axon termini of a transmitting neuron are closely apposed to dendrites of receiving neurons, but separated by a short gap, the synaptic cleft (**Figure 3.6**). A dendrite receives an incoming neurotransmitter signal (often glutamate or γ-aminobutyric acid [GABA], which are used extensively throughout the nervous system). In response, the local plasma membrane is depolarized to generate an action potential, an electrical impulse that spreads as a traveling wave along the plasma membrane of the axon. As a result of the change in electrical potential, a neurotransmitter is released from the axon terminus and diffuses across the synaptic cleft. The released neurotransmitter then binds to receptors on dendrites of all interconnected neurons, causing, in turn, local depolarization of their plasma membrane. In a similar fashion, neurons also transmit signals to muscles (at neuromuscular junctions) and glands.

- Electrical synapses. Here, the gap between the two connecting neurons is very small, only about 3.5 nm, and is known as a gap junction. Neurotransmitters are not involved; instead, gap junction channels allow ions to cross from the cytoplasm of one neuron into another, causing rapid depolarization of the membrane.

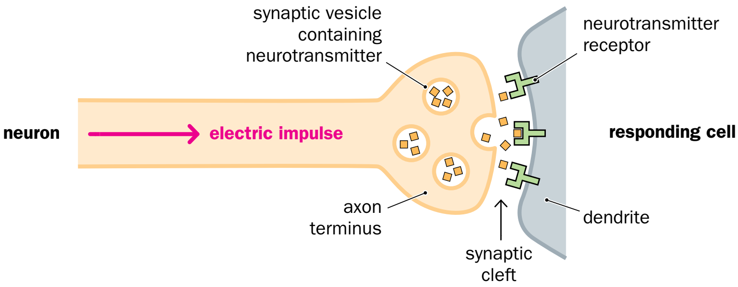


Figure 3.6 Synaptic signaling. At chemical synapses, a depolarized axon terminus of a transmitting (presynaptic) cell releases a neurotransmitter, such as glutamate or GABA, that is normally stored in vesicles. The release occurs by exocytosis: the vesicles containing the neurotransmitter fuse with the plasma membrane, releasing their contents into the narrow (about 20–40 nm) synaptic cleft. The neurotransmitter then binds to transmitter-gated ion channel receptors on the surface of the dendrites of a communicating neuron, causing local depolarization of the plasma membrane.

3.2 CELL PROLIFERATION AND PROGRAMMED CELL DEATH

The vast majority of cells are formed by cell division. Cell division may be common in certain tissues but especially so in early development, when rapid cell proliferation underlies the growth of multicellular organisms and the progression toward maturity. Throughout development, cell death is common, and by the time of maturity, an equilibrium is reached between cell proliferation and cell death.

While some cell loss is accidental—the result of injury or disease—planned or programmed cell death is very common and is functionally important. Like the organisms that contain them, cells age and the aging process, cell senescence, is related to various factors, including the frequency of cell division.

Most of the cells in mature animals are nondividing cells, but some tissues and cells turn over rapidly

The number of cells in a multicellular organism is determined by the balance between the rates of cell proliferation (which depend, in turn, on continued cell division) and of cell death. Tracking the birth and death of mammalian cells *in vivo* is problematic. Most of our knowledge about rates of mammalian cell proliferation has therefore come from cultured cells where, under normal circumstances, one turn of the mammalian **cell cycle** lasts approximately 20–30 hours.

M phase (mitosis and cytokinesis) lasts only about 1 hour and cells spend the great majority of their time (and do most of their work) in interphase. Interphase comprises three cell cycle phases: S phase (DNA synthesis) and two intervening, or gap, phases that separate it from M phase, G_1 phase (cell growth, centrosome duplication, and so on) and G_2 phase (preparation of factors needed for mitosis). Of the two gap phases, G_1 is particularly important and its length can vary greatly, under the control of several regulatory factors. Furthermore, when the supply of nutrients is poor, progress through the G_1 phase of the cycle may be delayed.

If the cells receive an antiproliferative stimulus they may exit the cell cycle altogether to enter a modified G_1 phase called the G_0 phase (**Figure 3.7A**). Cells in the G_0 phase are in a prolonged nondividing state. But they are not dormant. They can become terminally differentiated; that is, irreversibly committed to serve a specialized function. Most cells in the body are in this state, but they often actively synthesize and secrete proteins and may be highly motile.

G_0 cells can also continue to grow. For example, after withdrawing from the cell cycle, neurons become progressively larger as they project long axons that continue to lengthen until growth stops at maturity. For some neurons, the cytoplasm–nucleus ratio increases by more than 100,000 times during this period. Some G_0 cells do not become terminally differentiated but are quiescent. In response to certain external stimuli they can rejoin the cell cycle and start dividing again in order to replace cells lost through accidental cell death or tissue injury.

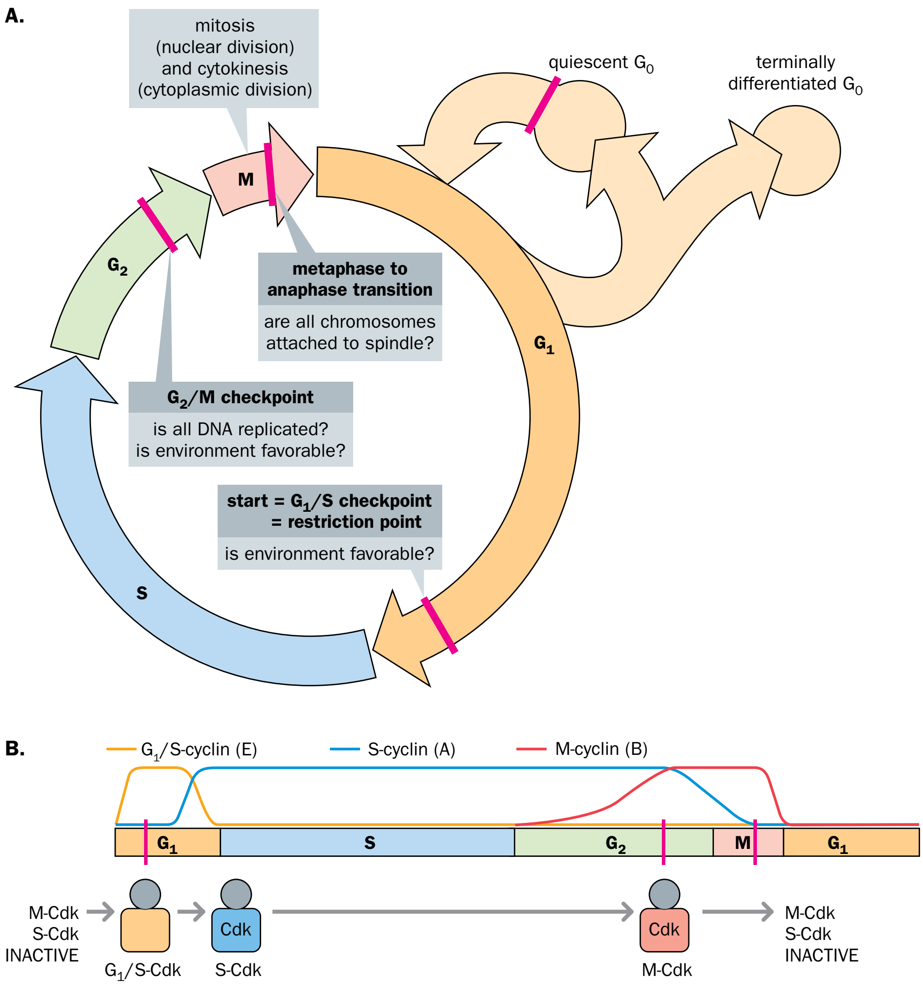

A.
mitosis (nuclear division) and cytokinesis (cytoplasmic division)
quiescent G_0
terminally differentiated G_0
M
G_2
metaphase to anaphase transition
are all chromosomes attached to spindle?
G_1
G_2/M checkpoint
is all DNA replicated? is environment favorable?
start = G_1/S checkpoint = restriction point
is environment favorable?
S
B.
G_1/S-cyclin (E)
S-cyclin (A)
M-cyclin (B)
G_1
S
G_2
M
G_1
M-Cdk S-Cdk INACTIVE
Cdk
Cdk
G_1/S-Cdk
S-Cdk
M-Cdk
M-Cdk S-Cdk INACTIVE

Figure 3.7 The cell cycle showing the three major checkpoints. (**A**) Passage through the cell cycle is controlled by checkpoints preceding transitions between phases. For example, cells can only leave G_2 phase to proceed with mitosis (M phase) if DNA has been replicated and conditions are conducive for cell division. Cells in G_1 phase may enter either a quiescent G_0 phase, from which they can re-enter the cell cycle, or a terminally differentiated state. (**B**) Each checkpoint is regulated by a specific cyclin-dependent kinase (Cdk) (shown here by a rectangle) bound to a cyclin protein (represented here by the gray circle). Cyclins include cyclin E (important for G1 phase), cyclin A (important for S phase), and cyclin B (important for M phase), and are synthesized and degraded at specific times within the cell cycle, limiting the availability of each cyclin–Cdk complex.

Mature multicellular animals do contain some dividing cells that are needed to replace cells that naturally undergo a high turnover. Sperm cells are continuously being manufactured in male mammals. There is also a high turnover of blood cells, and gut and skin epithelial cells are highly proliferative to compensate for the continuous shedding of cells from these organs. Even the adult mammalian brain, long believed to be unable to make new neurons, is now known to have three regions where new neurons are born.

Mitogens promote cell proliferation by overcoming braking mechanisms that restrain cell cycle progression in G_1

Cell proliferation is regulated by intrinsic (intracellular) factors and by extracellular signals. The intrinsic signals that regulate the cell cycle generally hold back (restrain) the cell cycle in response to sensors that indicate some fault, or unfavorable circumstance, at certain cell cycle checkpoints (Figure 3.7A).

The transition from one phase of the cell cycle to the next one is regulated by different cyclin-dependent kinases (Cdk). For example, Cdk1 and Cdk2 regulate entry into mitosis and S phase, respectively. Cdk concentrations are generally constant throughout the cell cycle but the Cdks are only active when they are bound by a cyclin protein. Different cyclins are synthesized and degraded at specific points in the cell cycle (Figure 3.7B). Thus, the amounts of individual cyclin–Cdk complexes, and ultimately of activated Cdks, parallels that of the amounts of the cyclins that they bind.

Unicellular organisms tend to grow and divide as rapidly as they can, but in multicellular organisms, cells divide only when the organism needs more cells. The start checkpoint at the G_1/S boundary is a major target for regulators that prevent cell division. For example, in mammalian cells, the Rb retinoblastoma protein and the p53 protein cause cells to arrest in G_1 if they contain damaged DNA.

In multicellular organisms, the cells that do divide must receive extracellular signals called **mitogens** that stimulate them to divide. Mitogens typically regulate cell division by overcoming intracellular braking mechanisms operating in the G_1 phase that restrain progress through the cell cycle.

The braking mechanisms in G_1 are naturally overcome by factors promoting S phase, such as E2F, a regulator that controls the synthesis of many proteins needed for S phase. During G_1, E2F is initially inhibited by being bound by the negative regulator Rb. As G_1 progresses, regulatory protein complexes (cyclin D–Cdk4 and cyclin E–Cdk2) accumulate, resulting in phosphorylation of Rb. Phosphorylated Rb has much lower affinity for E2F, freeing it to promote the synthesis of factors needed for S phase.

To ease the normal restraints on passage through G_1, mitogens must first bind to transmembrane receptor tyrosine kinases, stimulating a signal-transduction pathway that includes a small GTPase known as Ras and a MAP (mitogen-activated protein) kinase cascade. Ultimately transcription factors are activated that promote the transition to S phase (**Figure 3.8**). Downstream targets include proteins such as MYC, which stimulates production of both E2F and of cyclin–Cdk complexes that phosphorylate Rb and so liberate E2F.

As detailed in Chapter 16, cancer cells find ways of avoiding restrictions on the cell cycle, sometimes by mutating genes that code for checkpoint control proteins.

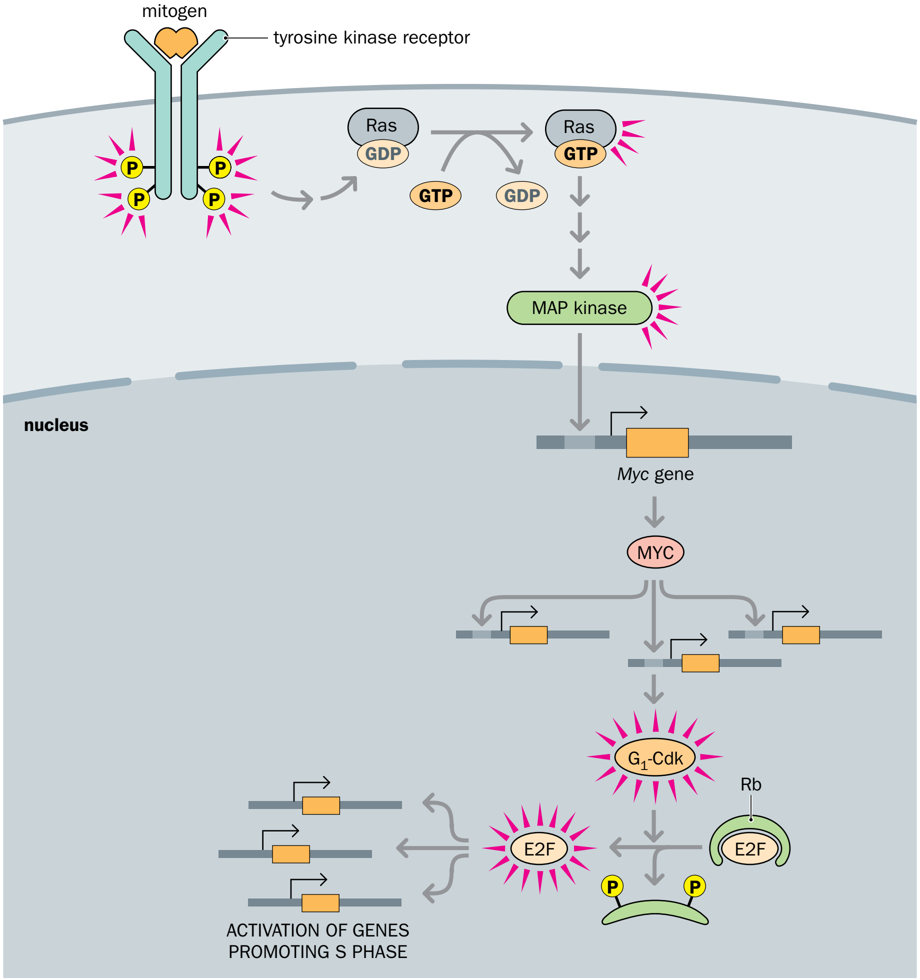


Figure 3.8 Mitogens promote cell proliferation through MAP kinase pathways. Mitogens bind to tyrosine receptor kinases causing the monomers to dimerize and cross-phosphorylate each other. The activated receptor relays the signal through an accessory protein leading ultimately to activation of a MAP kinase and subsequently the activation of transcription of target genes, such as the gene encoding the MYC transcription factor. MYC in turn activates various genes, including some that lead to increased G_1-Cdk activity, which in turn causes phosphorylation of the retinoblastoma protein Rb. Unphosphorylated Rb normally binds the transcription factor E2F and keeps it in an inactivate state, but phosphorylation of Rb causes a conformational change so that it releases E2F. The activated E2F transcription factor then activates the transcription of genes that promote S phase, notably the cyclin A gene. Red spikes on the phosphate groups of the tyrosine kinase receptor and on the Ras, MAP kinase, G1-Cdk and E2F proteins signify activation of the relevant protein; black arrows indicate transcriptional activation.

Cell proliferation limits and the concept of cell senescence

Intracellular mechanisms limit cell proliferation when it is not required. As organisms age (senesce), physiological deficits accumulate that undoubtedly have a cellular basis, including progressive oxidative damage to macromolecules. Cell senescence cannot easily be studied *in vivo*; instead, cell culture models have been used. Fibroblasts grown in culture from surgically removed human tissue typically achieve about only 30–50 population doublings in standard medium—the Hayflick limit. Proliferation rates are initially normal but gradually decline and then halt as the cells become arrested in G_1. They enter a terminal, nondividing state where they remain metabolically active for a while, but eventually die.

How this replicative cell senescence relates to cells *in vivo* and to organismal aging is not fully understood. Links between cellular and organismal senescence were postulated and for many years the Hayflick limit was viewed to be age-dependent. The modern consensus, however, is that there is no compelling evidence to support a relationship between replicative capacity *in vitro* and the age of the donors that provided the fibroblasts.

The phenomenon of cell senescence described above suggested the existence of a cellular "biological clock'' or, more accurately, replication counter. In cells undergoing senescence, the telomeres (chromosome ends) progressively shorten at each cell division (because of the problem of replicating chromosome ends). Eventually, the erosion of the telomeres destabilizes the telomeric T-loops (see **Figure 2.23**), leading to removal of the protective telomere cap. When this happens, the uncapped DNA at the end of the chromosomes is no different from the double-stranded DNA breaks resulting from DNA damage. As a result, cell surveillance systems monitoring DNA integrity likely induce the cells to enter a state of senescence. Cell senescence effectively appears to be a type of DNA damage response.

Not all cells are subject to cell senescence. Certain cells from embryos can be propagated for very long periods in culture and are effectively immortal, as are tumor cells, which subvert normal cell cycle controls. The idea that lack of telomere integrity is important in cell senescence is supported by studies of telomerase, an enzyme that counteracts telomere shortening by re-elongating telomeres. While most normal human somatic cells have negligible or tightly regulated telomerase activity, immortal cells have constitutively high telomerase levels. Fibroblasts and other somatic cells can be artificially immortalized by transfecting them with a gene encoding the catalytic subunit of telomerase.

Large numbers of our cells are naturally programmed to die

Throughout the existence of a multicellular organism, individual cells are born and die. Cell death occasionally occurs because of irreversible accidental damage to cells (necrosis). Causes include physical trauma, exposure to extreme temperatures, and oxygen starvation. Typically, large groups of neighboring cells are simultaneously affected. The process leads to leakiness of the plasma membrane, and water rushes in and causes the cells to swell up until the cellular membranes burst. Thereafter, the cells undergo autodigestion, producing local inflammation and attracting macrophages that ingest the cell debris.

In addition to accidental cell death, very large numbers of cells are also deliberately and naturally selected to die throughout the existence of a multicellular organism, even at very early stages of embryonic development. Such **programmed cell death** (**PCD**) can occur in response to signals sent, or withheld, by other cells (during development or immune surveillance, for example), or it can arise after a cell's internal monitoring systems sense major damage to vital cell components such as its DNA, mitochondria, and so on.

A variety of different types of programmed cell death are known. Of these, **apoptosis**, or type I PCD, has been extensively studied and is characterized by very specific changes in cell structure. Typically, individual cells (rather than groups of cells) undergo apoptosis and, as they die, the cells shrink rather than swell. A characteristic feature is that the chromatin condenses into compact patches that accumulate around the periphery of the nucleus. The nuclear DNA fragments and the nucleus breaks into discrete chromatin bodies. There is violent cellular movement (the cell appears to "boil"), and eventually the cell breaks apart into several membrane-lined vesicles called apoptotic bodies that are phagocytosed.

Another form of PCD is **autophagy**, a catabolic process where the cell's components are degraded as a way of coping with adverse conditions such as nutrient starvation or infection by certain intracellular pathogens. Intracellular double-membrane structures engulf large sections of the cytoplasm and fuse with lysosomes, thereby targeting the enclosed proteins and organelles for degradation. Taken to an extreme, autophagy can lead to cell death. Other forms of PCD are known but are poorly characterized.

The importance of programmed cell death

Programmed cell death is crucially important in many aspects of the development of a multicellular organism and is also vital for the normal functioning of mature organisms.

A quantitative illustration of its importance in development is provided by the nematode *Caenorhabditis elegans*, the only multicellular organism for which the lineage of all body cells is known. The adult worm has 959 somatic cells but is formed from a total of 1090 cells, 131 of which apoptose during embryonic development.

The selection of cells destined to die in *C. elegans* development is highly specific: the same 131 cells die in different individuals. Although 959 out of the 1090 cells survive, apoptosis is the default cell fate: all 1090 cells are programmed to die. The cells that survive do so by signaling to each other: they secrete proteins called survival factors that bind to cell surface receptors and override default apoptosis pathways. While our understanding of programmed cell death in early human development is incomplete, PCD is used to remove defective and excess or unwanted cells during the development of mammalian embryos and fetuses (see **Figure 3.9** for an example).

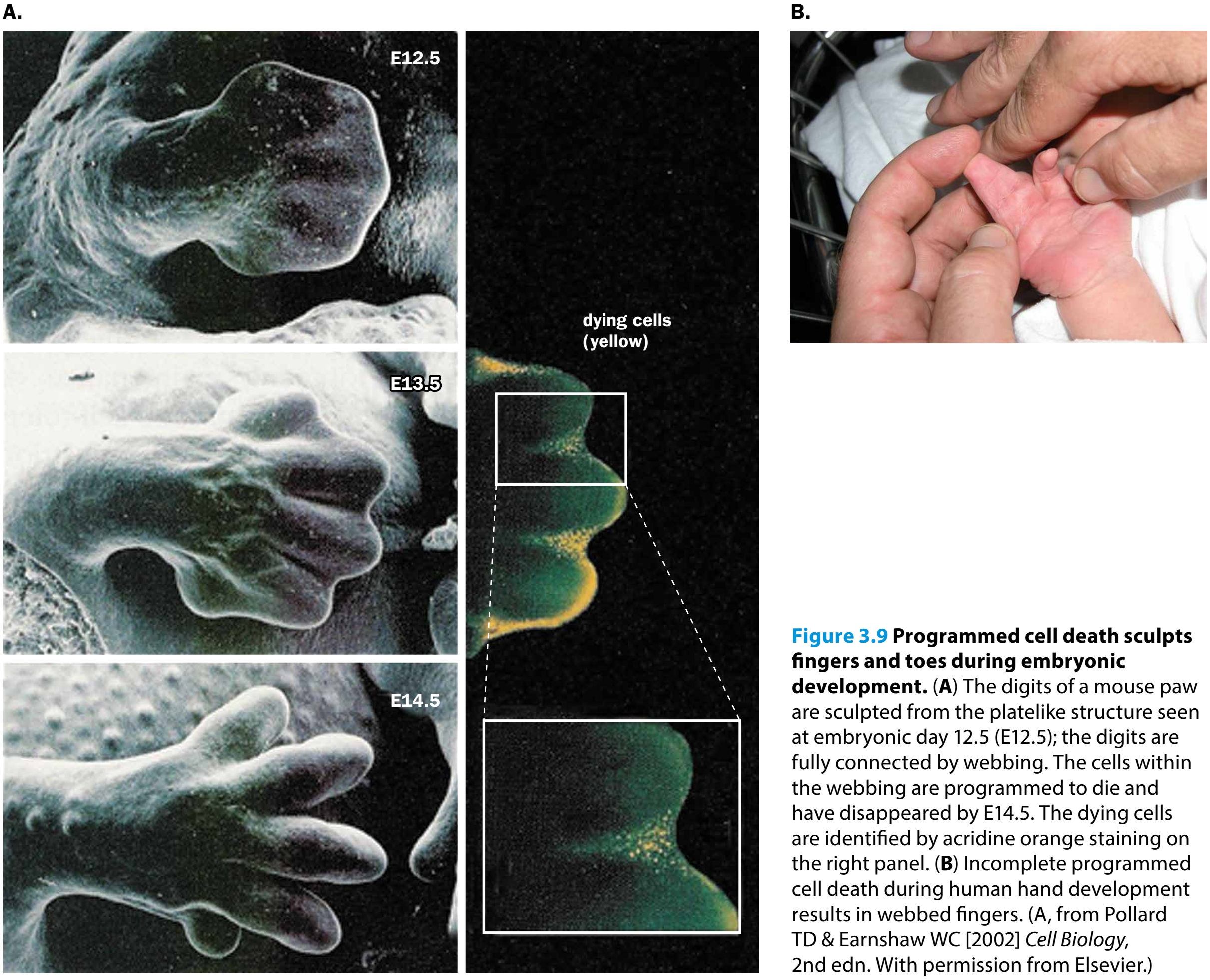


Figure 3.9 Programmed cell death sculpts fingers and toes during embryonic development. (**A**) The digits of a mouse paw are sculpted from the platelike structure seen at embryonic day 12.5 (E12.5); the digits are fully connected by webbing. The cells within the webbing are programmed to die and have disappeared by E14.5. The dying cells are identified by acridine orange staining on the right panel. (**B**) Incomplete programmed cell death during human hand development results in webbed fingers. (A, from Pollard TD & Earnshaw WC [2002] *Cell Biology*, 2nd edn. With permission from Elsevier.)

In complex multicellular organisms, cell death and cell proliferation need to be carefully balanced in the mature organism. There is a high cell turnover in some systems, such as for mammalian blood cells and epithelial cells in gut and skin. About 100,000 cells are programmed to die each second in adult humans but are replaced by mitosis. For some cells, such as B and T lymphocytes and neurons, special mechanisms are used to generate diversity, and here PCD is responsible for quality control, removing cells where the diversity-generating mechanisms have been unproductive (**Table 3.3**). As described in Section 3.4, PCD is also important in the immune system, where T cells and natural killer cells are employed to induce the death of body cells that are perceived to be harmful in some way, including virus-infected cells and tumor cells.

PCD has increasingly been recognized to be important in human disease. Aberrations in apoptosis play important parts in the etiology of autoimmune diseases, virally-induced diseases, and cancer. For example, helper T lymphocytes are key cells in immunosurveillance systems that we use to recognize and kill virally-infected cells.

Human immunodeficiency virus (HIV) proteins cause apoptosis of these key immune system cells, allowing disease progression toward acquired immunodeficiency syndrome (AIDS). Many cancer cells have devised strategies to oppose the immunosurveillance systems that normally induce apoptosis of cancer cells. Successful chemotherapy often relies on using chemicals that help induce cancer cells to apoptose. Other forms of PCD are important in neurodegenerative disease, such as in Huntington disease and Alzheimer's disease, as well as in myocardial infarction and in stroke, where secondary PCD caused by oxygen deprivation in the area surrounding the initially affected cells greatly increases the size of the affected area.

TABLE 3.3 SOME IMPORTANT FUNCTIONS OF PROGRAMMED CELL DEATH

Function	Examples	Remarks
Killing of defective cells	Removal of defective immature lymphocytes	Natural mistakes occur in DNA rearrangements that give T and B cells their specific receptors. Up to 95% of immature T cells have defective receptors and are eliminated by apoptosis
Killing of harmful cells	Removal of harmful T lymphocytes to allow self-tolerance	Some immature T cells contain receptors that recognize self-antigens instead of foreign antigens and need to be eliminated by apoptosis before they cause normal host cells to die
	Removal of virally-infected cells and tumor cells	A class of T lymphocytes known as killer T cells is responsible for inducing apoptosis in host cells that express viral antigens or altered antigens/neoantigens produced by tumor cells
	Removal of cells with damaged DNA	Cells with damaged DNA tend to accumulate harmful mutations and are potentially harmful. DNA damage often induces cell suicide pathways
Killing of excess, obsolete, or unnecessary cells	Removal of surplus neurons during development	In the embryo, many more neurons are produced than are needed. Those neurons that fail to make the right connections to other neurons or muscle cells are apoptosed
	Removal of interdigital cells	During development, fingers and toes are sculpted from spadelike handplates and footplates by apoptosis of the unnecessary interdigital cells (see **Figure 3.9**)

Apoptosis is carried out by caspases in response to death signals or sustained cell stress

The key molecules that execute apoptosis are the caspase family of proteases. These enzymes have cysteine at their active site and cleave their substrates on the C-terminal side of aspartate residues. Inactive caspase precursors (procaspases) are synthesized naturally by all cells and have an N-terminal prodomain that needs to be cleaved off to activate the caspase. There are two classes of procaspases:

- Initiator procaspases, such as caspase 8 and caspase 9, have long prodomains and can undergo autoactivation. Their job is to start off cell death pathways after they have been activated by signals transmitted through cell surface receptors or from internal sensors;
- Effector procaspases, such as caspases 3, 6, and 7, have short prodomains that are cleaved by initiator caspases. Effector caspases cleave about 100 different target proteins, including nuclear lamins (causing breakdown of the nuclear envelope) and cytoskeletal proteins (destroying cell architecture). In addition, cells naturally possess a cytoplasmic caspase-activated DNase. The DNase is normally kept inactive by being bound by an inhibitor protein. Effector caspases can cleave the inhibitor protein, however, thereby releasing the DNase, which migrates to the nucleus and fragments cellular DNA.

Protein modifications such as phosphorylation and ubiquitylation are reversible. By contrast, proteolysis is irreversible: once a peptide bond is cleaved, cells cannot re-ligate the cleavage products. Once apoptosis has been initiated, therefore, there is no going back.

As described below, extrinsic and intrinsic apoptosis pathways use different initiator caspases, but they use the same types of effector caspases to cleave target proteins.

Extrinsic apoptosis pathways

During development, neighboring cells exchange death signals and survival factors and have specific receptors for both types of signal. Cells that receive enough of a survival factor will live, because they initiate cell survival pathways that suppress programmed cell death pathways; those that don't receive enough survival factor will die. Competition between cells to receive enough survival factor is thought to control cell numbers both during development and in adulthood.

Many death receptors are members of the TNF (tumor necrosis factor) superfamily. Fas is a well-studied example. Its ligand (FasL) forms trimers and induces the Fas receptor to trimerize, causing clustering of death domains on the receptor's cytoplasmic tail. The clustered death domains attract binding of an adaptor protein, FADD, that then recruits procaspase 8 to initiate a series of caspase cleavages (**Figure 3.10**).

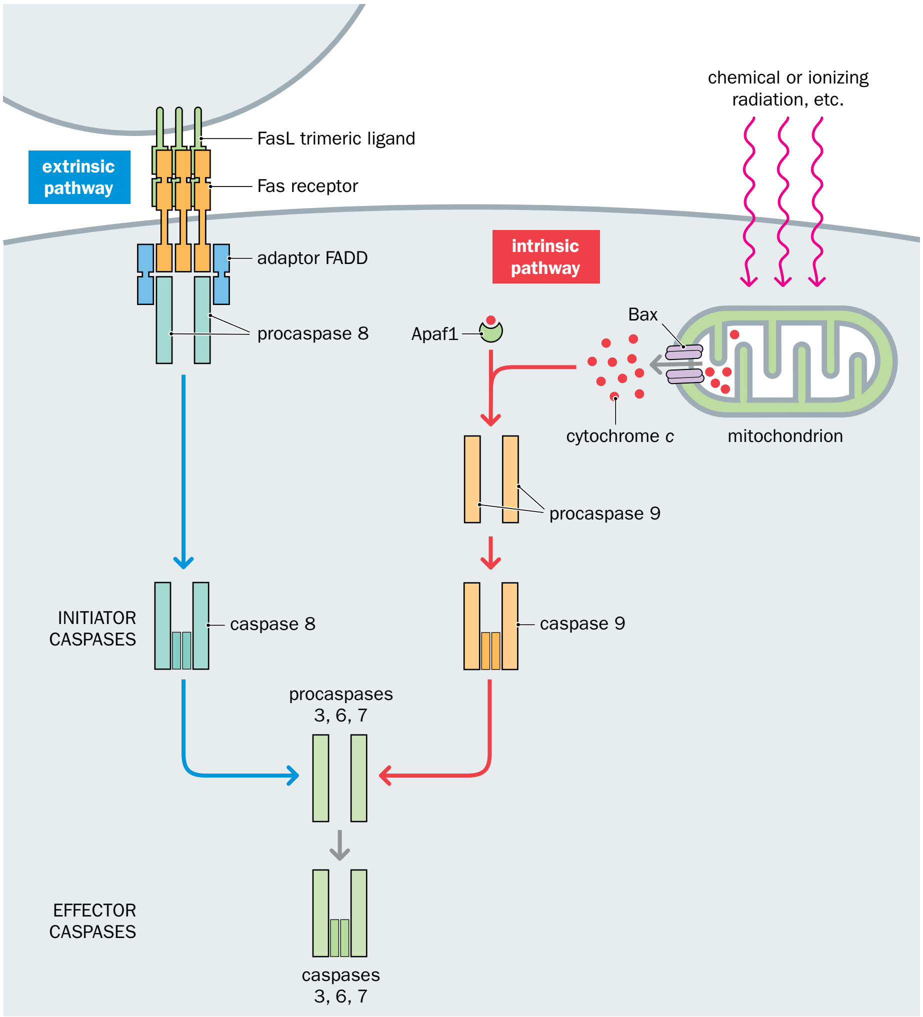


Figure 3.10 Apoptosis pathways. *Extrinsic apoptosis pathways* involve activation of a cell surface death receptor by a ligand on a neighboring cell. The Fas death receptor is normally found as a monomer but is induced to form a trimer by its trimeric ligand FasL. The resulting clustering of Fas receptors recruits the FADD adaptor protein. FADD acts as a scaffold to recruit procaspase 8, which undergoes autoactivation to initiate a caspase cascade. *Intrinsic apoptosis pathways* are activated when vital cell components are damaged or stressed, for example in response to harmful radiation, chemicals, hypoxia, and so on. These pathways are activated from within cells using mitochondrial and endoplasmic reticulum components. In the mitochondrial pathways, proapoptosis factors, such as Bax shown here, form oligomers in the mitochondrial outer membrane, forming pores that allow release of cytochrome *c*. In the cytosol, cytochrome *c* binds and activates the Apaf1 protein and induces the formation of an apoptosome that activates procaspase 9 and ultimately the same effector caspases (caspase 3, caspase 6, and caspase 7) as the death receptor pathways.

Intrinsic apoptosis pathways

Cells are induced to apoptose when they are sufficiently stressed that sensors indicate significant damage to certain key components. The integrity of genomic DNA needs to be protected, and mitochondria are vitally important energy producers. The endoplasmic reticulum is also crucial. As well as being required for protein and lipid synthesis, it is essential for correct protein folding and is the major intracellular depot for storing Ca^{2+}, the most widely used second messenger in cell signaling. Prolonged changes in Ca^{2+} concentration in the endoplasmic reticulum or the accumulation of unfolded or misfolded proteins can lead to apoptosis.

The mitochondrial pathway of apoptosis is initiated when proapoptosis cytoplasmic proteins such as Bax are activated. Bax then binds to the mitochondrial outer membrane and forms oligomers, permitting release of cytochrome *c*, which, in turn, activates the cytoplasmic Apaf1 protein causing activation of procaspase-9 (see **Figure 3.10**). Bax belongs to a large family of apoptosis regulators that includes antiapoptosis factors (such as Bcl-2) as well as proapoptosis factors.

3.3 CELL ADHESION AND TISSUE FORMATION

The cells of a multicellular organism need to be held together. In vertebrates and other complex organisms, cells are assembled to make tissues—collections of interconnected cells that perform a similar function—and organs. Various levels of interaction contribute to this process:

- As they move and assemble into tissues and organs, cells must be able to recognize and bind to each other, a process known as **cell adhesion**;
- Cells in animal tissues frequently form cell junctions with their neighbors that can have different functions;
- The cells of tissues are also bound by the extracellular matrix (ECM), the complex network of secreted macromolecules occupying the space between cells. Most human tissues contain ECM, but the proportion can vary widely.

Even where cells do not form tissues, as in the case of blood cells, cell adhesion is vitally important, permitting transient cell–cell interactions that are required for various cell functions.

During embryonic development, groups of similar cells are formed into tissues. For even simple tissues such as epithelium, the descendants of the progenitor cells must not be allowed to simply wander off. The requirement becomes more critical when the tissue is formed after some of the progenitor cells arrive from long and complicated cell migration routes in the developing embryo. Cells are kept in place by cell adhesion, and the architecture of the tissue is developed and maintained by the specificity of cell adhesion interactions.

Cell adhesion molecules work by having a receptor and a complementary ligand attached to the surfaces of adjacent cells. There may be hundreds of thousands of such molecules per cell and so binding is very strong. Cells may stick together directly and/or they may form associations with the ECM. During development, changes in the expression of adhesion molecules allow cells to make and break connections with each other, facilitating cell migration. In the mature organism, adhesion interactions between cells are generally strengthened by the formation of cell junctions (see below).

Cell adhesion molecules (CAMs) are typically transmembrane receptors with three domains: an intracellular domain that interacts with the cytoskeleton; a transmembrane domain; and an extracellular domain that interacts either with identical CAMs on the surface of other cells (homophilic binding) or with different CAMs (heterophilic binding), or the ECM. There are four major classes of cell adhesion molecule:

- Cadherins are the only class to participate in homophilic binding;
- Integrins are adhesion heterodimers. They usually mediate cell–ECM interactions but some leukocyte integrins are involved in cell–cell adhesion;
- Selectins mediate transient cell–cell interactions in the bloodstream. They are important in binding leukocytes to the endothelial cells lining blood vessels so that blood cells can migrate out of the bloodstream into a tissue (extravasation).
- Ig-CAMs (immunoglobulin superfamily cell adhesion molecules) possess immunoglobulin-like domains (see Section 3.4).

Different types of cell junction can regulate the contact between cells

As listed below, different types of cell junction can regulate contact between adjacent cells in vertebrate organisms, and between cells and the ECM. They can have different functions: helping to anchor cells, acting as barriers, or permitting direct intercellular passage of small molecules.

Anchoring cell junctions

Some cell junctions mechanically attach cells (and their cytoskeletons) to their neighbors or to the ECM using dedicated proteins, notably cadherins (cell–cell joining) and integrins (principally for cell–ECM joining). In each case, actin filaments or intermediate filaments are tethered to the cell junction: cell–cell adhesion also involves linking of cytoskeletons of the neighboring cells, and cell–matrix adhesion also involves the cytoskeleton of the anchored cell. The four major types of anchoring cell junction are listed below and illustrated in **Figure 3.11A**.

- Adherens junctions. Cadherins on one cell bind to cadherins on another. The cadherins are linked to actin filaments using anchor proteins such as catenins, vinculin, and α-actinin.

- Desmosomes. Desmocollins and desmogleins on one cell bind to the same molecule types on another. They are linked to intermediate filaments using anchor proteins such as desmoplakins and plakoglobin.
- Actin-linked focal adhesions. Integrins on a cell surface bind to ECM proteins. The integrins are connected internally to actin filaments using anchor proteins such as talin, vinculin, α-actinin, and filamin.
- Hemidesmosomes. Integrins on epithelial cell surfaces bind to a protein component, laminin, of the basal lamina. The integrins are connected internally to intermediate filaments using anchor proteins such as plectin.

Cell junctions acting as barriers

Tight junctions are primarily designed to act as barriers, and are especially prevalent in the epithelial cell sheets lining the free surfaces and all cavities of the body. Here they serve as selective permeability barriers by separating fluids with different chemical compositions on either side. By creating such tight seals between cells, they can prevent even small molecules from leaking from one side of the epithelial sheet to the other (**Figure 3.11**).

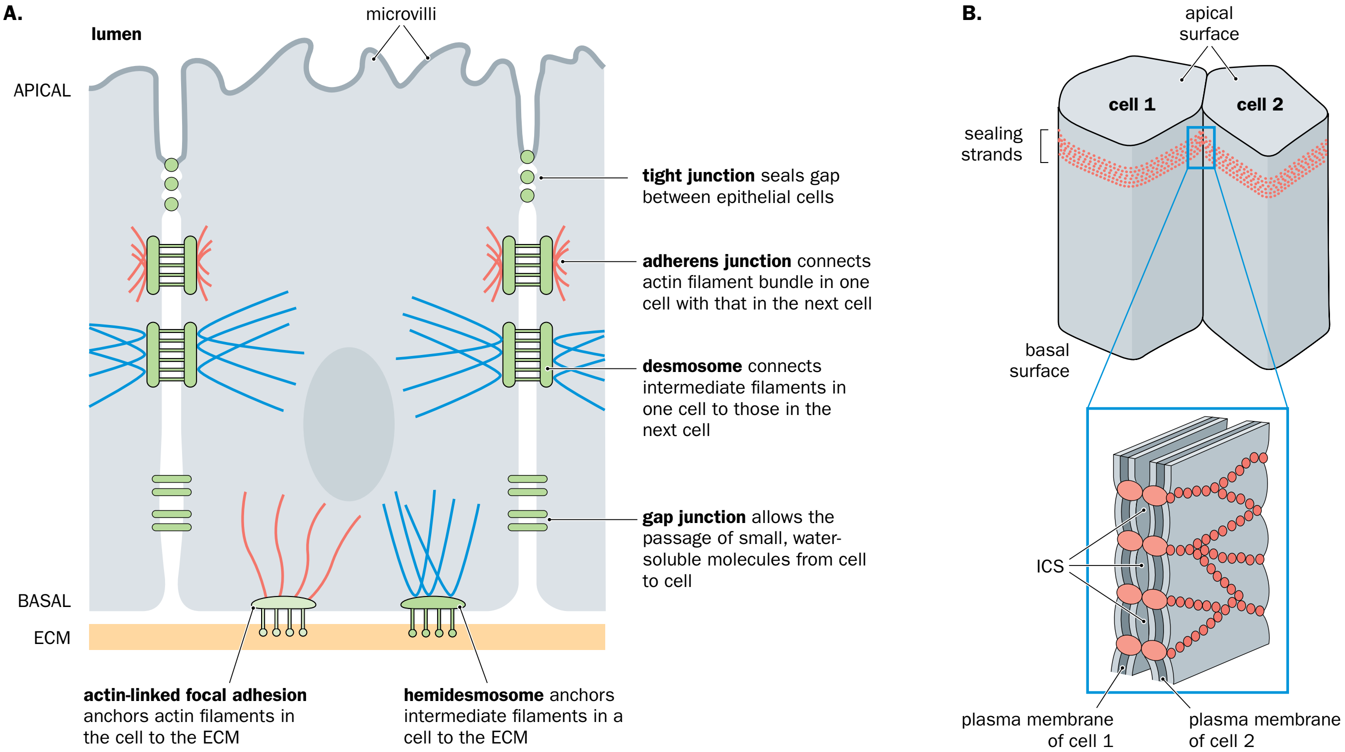

A.
lumen
microvilli
APICAL
tight junction seals gap between epithelial cells
adherens junction connects actin filament bundle in one cell with that in the next cell
desmosome connects intermediate filaments in one cell to those in the next cell
gap junction allows the passage of small, water-soluble molecules from cell to cell
BASAL
ECM
actin-linked focal adhesion anchors actin filaments in the cell to the ECM
hemidesmosome anchors intermediate filaments in a cell to the ECM
B.
apical surface
cell 1
cell 2
sealing strands
basal surface
ICS
plasma membrane of cell 1
plasma membrane of cell 2

Figure 3.11 The six principal classes of cell junctions found in vertebrate epithelial cells. (**A**) This example shows intestinal epithelial cells that are arranged in a sheet overlying a thin layer of extracellular matrix (ECM), known as the basal lamina. Individual cells are symmetrical along the axes that are parallel to the ECM layer but show polarity along the axis from the top (apical) end of the cell that faces the lumen to the bottom (basal) part of the cell. Actin filaments (red lines) and intermediate filaments (blue lines) are each important in two types of cell junctions. Thus, cells are anchored to the ECM via actin filaments (actin-linked cell matrix adhesion) or intermediate filaments (at hemidesomosomes), and are bound to their neighbors via actin filaments (at adherens junctions) or intermediate filaments (at desmosomes, located below the adherens junctions). Two additional cell junctions are used in cell-cell adhesion. Tight junctions act as barriers. They occupy the most apical position and divide the cell surface into an apical region (which is rich in intestinal microvilli) and the remaining basolateral cell surface. Gap junctions are communicating junctions and are located in more basal regions. (**B**) Three-dimensional structure of neighboring epithelial cells showing the apical sealing bands that encircle the cells at tight junctions (top) with an exploded view of the connecting plasma membranes at bottom. Sealing strands are shown in red; ICS, intercellular spaces. (A, adapted from Alberts B *et al*. [2014] *Molecular Biology of the Cell*, 6th edn. Garland Science. With permission from WW Norton.)

Tight junctions are made up of a network of sealing strands formed by direct joining of the extracellular domains of transmembrane proteins embedded in the two plasma membranes. The sealing strands completely encircle the apical (outward-facing) ends of each epithelial cell (**Figure 3.11B**).

Communicating cell junctions

Gap junctions permit inorganic ions and other small, hydrophilic molecules (<1 kDa) to pass directly from a cell to its neighbors (**Figure 3.11A**). The plasma membranes of participating cells come into close contact, establishing a uniform gap of about 2–4 nm.

The gap is bridged by contact between a radial assembly of six connexin molecules on each plasma membrane; when oriented in the correct register, they form an intercellular channel. Gap junctions allow electrical coupling of nerve cells and co-ordinate cell functions in a variety of other tissues.

The extracellular matrix regulates cell behavior as well as acting as a scaffold to support tissues

The ECM is a supportive matrix that occupies the space between cells of different types of tissue and accounts for a substantial amount of tissue volume, especially in connective tissues (which are the major component of cartilage and bone and provide the framework of the body). Its hallmark is a three-dimensional array of secreted protein fibers embedded in a gel of complex carbohydrates; these molecular components are mostly made locally by some cells embedded within the ECM.

The molecular composition of the ECM is variable and dictates the physical properties of connective tissue. It can be calcified to form very hard structures (bones, teeth), it can be transparent (cornea), and it can form strong ropelike structures (tendons). The ECM is not just a scaffold for supporting the physical structure of tissues, however. It also regulates the behavior of cells that come into contact with it. It can influence their shape and function, and affect their development and their capacity for proliferation, migration, and survival. Neighboring cells can, in turn, modify ECM structure by secreting enzymes, such as proteases.

In accordance with its diverse functions, the ECM contains a complex mixture of macromolecules (**Figure 3.12**). In connective tissue, for example, the matrix macromolecules are largely secreted by fibroblast-type cells (see also **Figures 3.13** and **3.14**, when we give the example of the structure of intestinal tissue). In addition to proteins, the ECM macromolecules include two other types of complex polymers, as listed below.

- Glycosaminoglycans are extremely long polysaccharide chains assembled from tandem repeats of particular disaccharides. Hyaluronic acid is the only example of a free glycosaminoglycan in the ECM, one that is not covalently attached to a protein. Its structure is based on repeats of the disaccharide *N*-acetyl-D-glucosamine–D-glucuronic acid; there can be as many as 25,000 repeats.
- Proteoglycans are a type of glycoprotein that has a protein core with sugar side chains, at least one of which is a glycosaminoglycan. They exist in various different forms in the ECM.

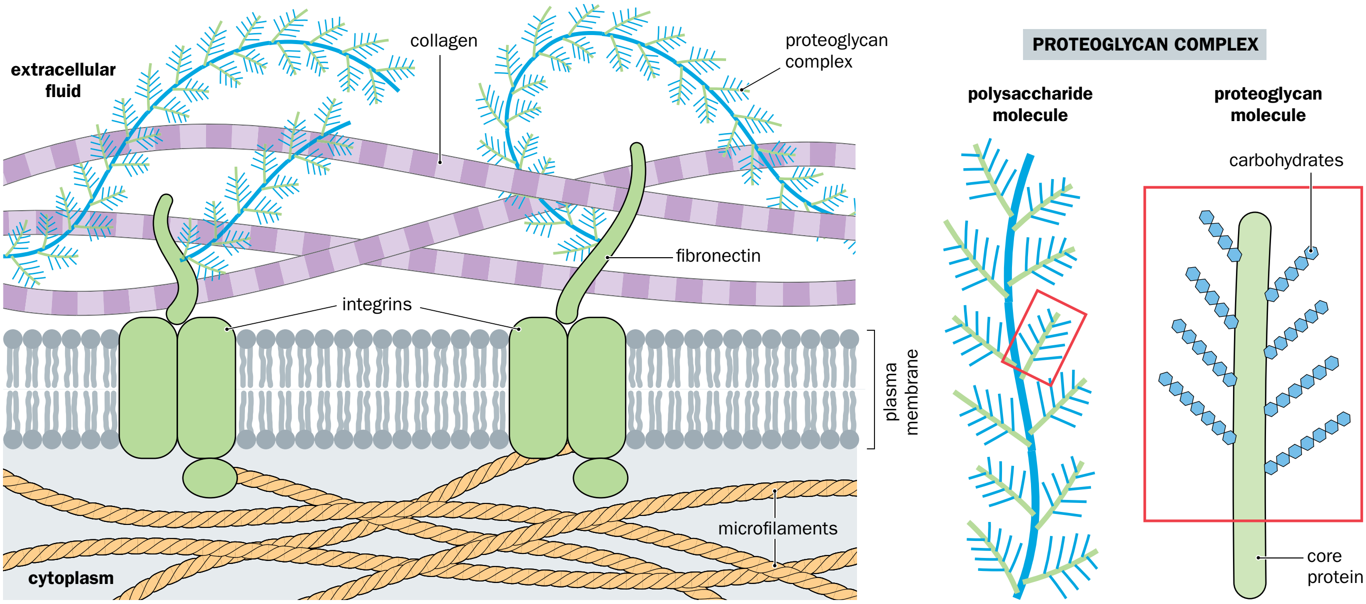


Figure 3.12 Molecular structure of the extracellular matrix (ECM). The principal molecular components of the ECM, glycoproteins and proteoglycans, are made within the cell and exported by exocytosis. The most prominent glycoproteins are long collagens that have a triple-helical structure (which gives the resulting fibers a high tensile strength and great elasticity) and fibronectins that help attach cells to the ECM via integrin receptor proteins in the plasma membrane. Not shown are additional glycoproteins such as elastin (which confers flexibility on the ECM) and laminin (which forms webs that help hold neighboring cells together). The proteoglycans are small glycoproteins bound to long polysaccharides; they regulate the movement of molecules through the matrix and also the binding of cations and water. (The consistency of the matrix as a whole depends on how much water can be trapped; the more interlinks, the more water can be trapped, making the consistency soft, such as that of cartilage.) Note that multiple cell types can be surrounded by a common ECM, as in the example of connective tissue shown in **Figure 3.14**. (Modified from Urry LA *et al*. [2016] *Biology*, 11th edn.)

Being extremely large and highly hydrophilic, glycosaminoglycans readily form hydrated gels that generally act as cushions to protect tissues against compression. Tissues such as cartilage, where the proteoglycan content of the ECM is particularly high, are highly resistant to compression. Proteoglycans can form complex superstructures in which individual proteoglycan molecules are arranged around a hyaluronic acid backbone. Such complexes can act as biological reservoirs by storing active molecules such as growth factors, and proteoglycans may be essential for the diffusion of certain signaling molecules.

The ECM macromolecules have different functional roles. The glycoproteins predominantly have structural roles (notably collagens; elastin also allows tissues to regain their shape after being deformed), or have roles in adhesion (fibronectin and vitrinectin in cell–matrix adhesion; laminins in adhesion of cells to the basal lamina of epithelial tissue (see below). Proteoglycans can bind growth factors and other bioactive molecules, and are important in regulating adhesion and some other processes. For example, hyaluronic acid is involved in regulating cell migration, particularly during development and tissue repair).

Specialized cell types are organized into tissues

There are many different types of cells in adult humans, but they are organized into just a few major types of tissue. Organs are typically composed of a small number of different tissue types; for example, the gut comprises a layer of epithelium, connective tissue, and smooth muscle (**Figure 3.13**). Of the common tissues, epithelial, muscle, nervous, and connective tissues are outlined below; lymphoid tissue is described in Section 3.4 when we consider immune system cells.

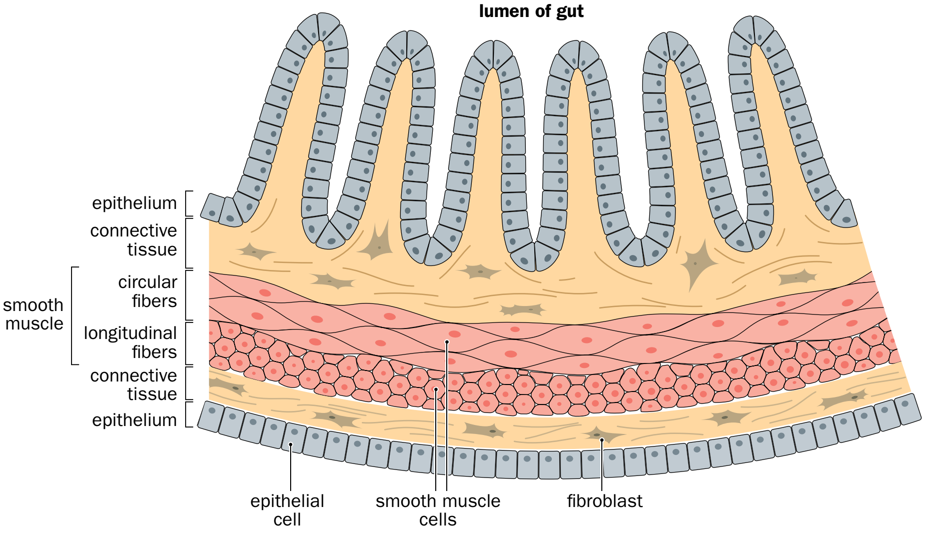


Figure 3.13 The gut as an example of the relationships between cells, tissues, and organs. The gut is a long, tube-shaped organ largely constructed from three tissues. Epithelial tissues form the inner and outer surfaces of the tube and are separated from internal layers of muscle tissue by connective tissue. The latter is mostly composed of extracellular matrix (extracellular fluid containing a complex network of secreted macromolecules; see **Figure 3.12**). The inner epithelial layer (top) is a semi-permeable barrier, keeping the gut contents within the gut cavity (the lumen) while transporting selected nutrients from the lumen through into the extracellular fluid of the adjacent connective tissue layer. (Adapted from Alberts B *et al.* [2014] *Molecular Biology* of the Cell, 6th edn. Garland Science.)

Epithelial tissue

Epithelial tissue has little ECM and is characterized by tight cell binding between adjacent cells, forming cell sheets on the surface of the tissue. The cells are bound to their neighbors by strong adhesive forces that permit the cells to bear most of the mechanical stress that the tissue is subjected to. Here, the ECM mostly consists of a thin layer, the **basal lamina**, that is secreted by the cells in the overlying epithelium layer (**Figure 3.14**).

The individual cells within a layer of epithelium show consistent internal asymmetry (**cell polarity**) in a plane that is at right angles to the cell sheet. The apex of an epithelial cell, the end facing the exterior (or the lumen of a cylindrical tube), is the one part of the cell not attached to its cell neighbors or to the basal lamina. On the apical surface are microvilli. These small, hairlike projections are composed of complex plasma membrane folds surrounding an actin microfilament core, and are not found at the basal end of the cell (the part attached to the basal lamina) or on the lateral regions, the sides attached to neighboring cells (see **Figure 3.14**).

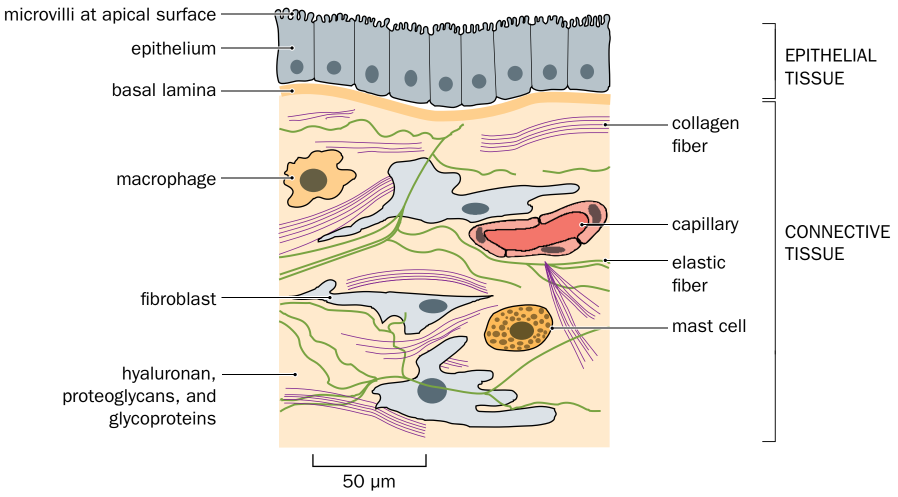


Figure 3.14 Connective tissue: cells and structure. The figure shows the example of connective tissue underlying epithelium. The epithelial tissue consists of a cellular layer plus an underlying thin layer (the basal lamina) consisting of extracellular matrix secreted by the cells above. Connective tissue is dominated by an extracellular matrix (ECM), a three-dimensional array of protein fibers embedded in a gel of complex carbohydrates (glycosaminoglycans), with cells sparsely distributed within the ECM. Some of the cells are indigenous, including fibroblasts (cells that synthesize and secrete most of the ECM macromolecules), fat cells, and mast cells that secrete histamine-containing granules in response to insect bites or exposure to allergens. In addition, there are various immigrant blood and immune system cells (such as monocytes, macrophages, T cells, plasma cells, and leukocytes). (Adapted from Alberts B *et al.* [2014] *Molecular Biology of the Cell*, 6th edn. Garland Science. With permission from WW Norton.)

Connective tissue

Connective tissue is largely composed of ECM that is rich in fibrous polymers, notably collagen. Sparsely distributed within this tissue is a remarkable variety of specialized cells. The indigenous cells—mesenchymal stem cells and the differentiated cells that they give rise to, notably fibroblasts—synthesize and secrete most of the ECM macromolecules (see **Figure 3.14**). There are also some immigrant cells, notably immune system and blood cells. Because cells are sparsely distributed in the supporting ECM, it is the ECM rather than the cells within it that bears most of the mechanical stress falling on connective tissue. Two types of connective tissue are recognized, as listed below.

- Loose connective tissue. This has fibroblasts surrounded by a flexible collagen-fiber matrix, and is found beneath the epithelium in skin and many internal organs. It also forms a protective layer over muscle, nerves, and blood vessels.
- Fibrous connective tissue. Here, the collagen fibers are densely packed, providing strength to tendons and ligaments. Cartilage and bone are rigid forms of connective tissue.

Muscle tissue

Muscle tissue is composed of contractile cells that have the special ability to shorten or contract in order to produce movement of the body parts. Skeletal muscle fibers are cylindrical, striated, under voluntary control, and multinucleated because they arise by fusion of precursor cells called myoblasts (see **Figure 2.8D**). Smooth muscle cells are spindle-shaped, have a single, centrally located nucleus, lack striations, and are under involuntary control (see **Figure 3.13**). Cardiac muscle has branching fibers, striations, and intercalated disks; the component cells, cardiomyocytes, each have a single nucleus and contraction is not under voluntary control.

Nervous tissue

Nervous tissue is limited to the brain, spinal cord, and nerves. Neurons are electrically excitable cells that process and transmit information via electrical signals (impulses) and secreted neurotransmitters. They have three principal parts: the cell body (the main part of the cell, performing general functions); a network of dendrites (extensions of the cytoplasm that carry incoming impulses to the cell body); and a single, long axon that carries impulses away from the cell body to the end of the axon (**Figure 3.15**).

Neurons account for less than 10% of cells in the nervous system; the other 90% are glial cells. Glial cells do not transmit impulses, but instead support the activities of the neurons in a variety of ways. The axons of neurons have an insulating sheath of a phospholipid, myelin, that is produced by certain glial cells: oligodendrocytes (in the central nervous system) and Schwann cells (in the peripheral nervous system). Astrocytes are small, star-shaped glial cells that ensheath synapses and regulate neuronal function. Microglial cells are phagocytic and protect against bacterial invasion. Other glial cells provide nutrients by binding blood vessels to the neurons.

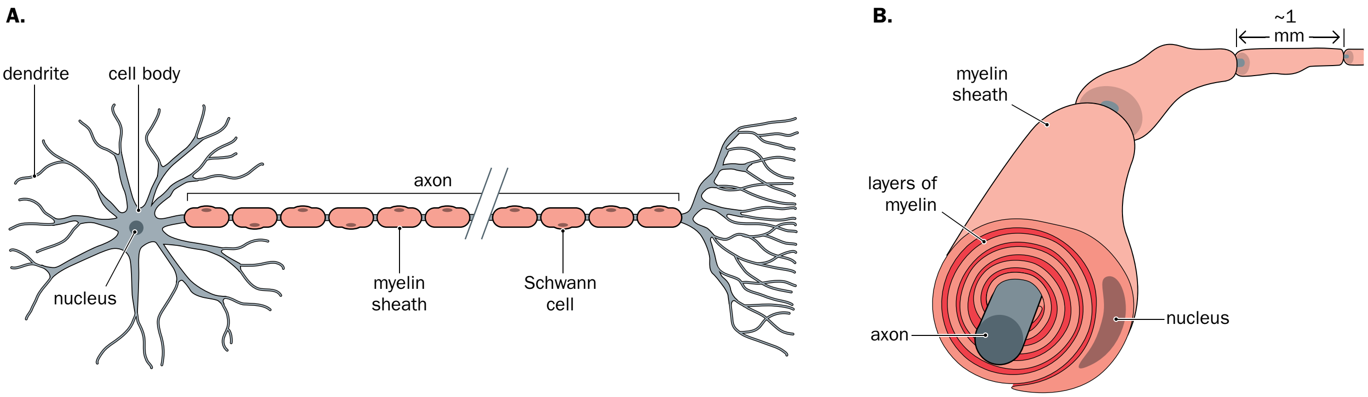


Figure 3.15 Neurons and myelination. (**A**) Neuron structure. Each neuron has a single, long axon with multiple axon termini (dendrites) that are connected to other neurons or to an effector cell such as a muscle cell. Neurons are insulated by certain glial cells such as Schwann cells that form a myelin sheath. (**B**) Myelination of an axon from a peripheral nerve. Each Schwann cell wraps its plasma membrane concentrically around the axon, forming a myelin sheath covering 1 mm of the axon. (Adapted from Alberts B *et al.* [2014] *Molecular Biology of the Cell*, 6th edn. Garland Science. With permission from WW Norton.)

3.4 IMMUNE SYSTEM BIOLOGY

The vertebrate immune system is an intricate network of cells, lymphoid organs, and proteins that protects the body against infection by a wide range of microbial pathogens (bacteria, viruses, fungi, parasitic worms, and protozoa, but archaea are not known to be pathogenic), and can also protect against tumor development. Its protective capabilities depend on the ability to distinguish between normal self (components of the host organism) and nonself (foreign intruders) or altered self (notably abnormal cancer cells). Powerful immune responses developed against **antigens** (any substance associated with nonself or altered self that the immune system perceives as being foreign or dangerous) are then unleashed in an attempt to eradicate an infection or kill harmful body cells.

Immune system cells include not just the familiar B and T lymphocytes that develop from immature lymphoblasts, but also many other types of blood cells and a range of different tissue cells, as detailed below. Immune system proteins include: signaling proteins and receptors; antibacterial antibodies; antibacterial peptides and enzymes; and many other soluble proteins, notably those of the plasma complement protein system.

Not all immune responses against antigens are beneficial. While fighting infection, substantial collateral damage can be inflicted on healthy tissues. Exaggerated or inappropriate immune responses against foreign substances (such as pollen) can cause allergies. Because the immune system occasionally makes mistakes in distinguishing self from nonself, it can sometimes attack healthy body cells, leading to a wide variety of autoimmune diseases. The immune system is also responsible for the rejection of transplanted organs and tissues.

In this section we explain the workings of the mammalian immune system, and how it comprises two types of immune system (**Figure 3.16**):

- The **innate immune system**, an evolutionarily ancient system (found in both vertebrates and invertebrates) that provides first-level and general protection against pathogens;
- The **adaptive immune system**, a vertebrate innovation that provides powerful backup protection through acquired responses to specific antigens.

The innate immune system is very valuable as a rapid first line of defense, providing protection in a general, nondiscriminating way against all pathogens. That is, no matter what the pathogen is, much the same type of basic protection cover is provided. It relies on various barriers designed to reduce the chance that active pathogens gain access to internal body tissues, plus internal defense systems carried out by certain protein classes and various immune system cells other than B and T lymphocytes. We describe the details below.

The adaptive immune system is more specific and more flexible than the innate immune system. And although it takes some time (several days) to mount an adaptive immune response, it is both stronger and more long-lasting than innate immune responses. The adaptive immune responses are ultimately dependent on two classes of lymphocytes: effector B cells (which secrete antibodies) and effector T cells (which have receptors that can recognize body cells harboring internal pathogens such as viruses and intracellular protozoa).

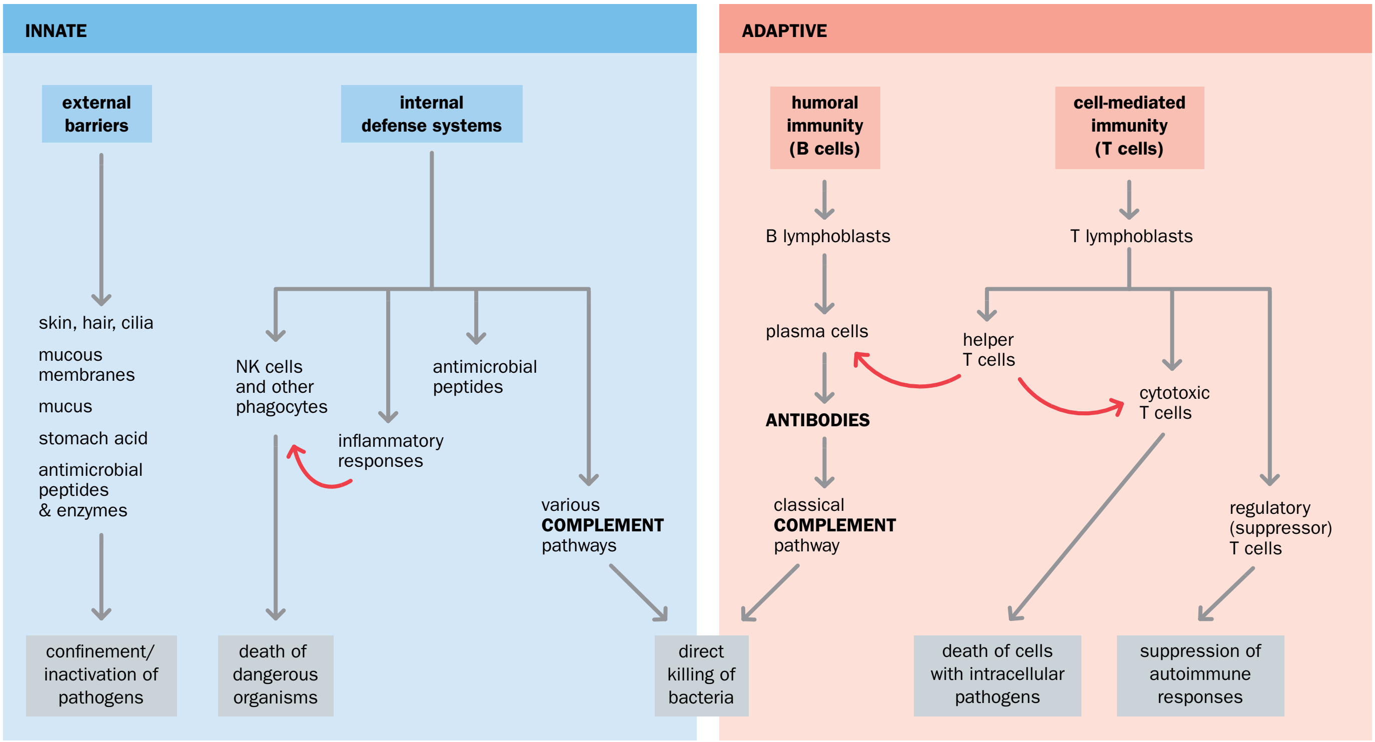

INNATE
external barriers
internal defense systems
skin, hair, cilia
mucous membranes
mucus
stomach acid
antimicrobial peptides & enzymes
NK cells and other phagocytes
antimicrobial peptides
inflammatory responses
various COMPLEMENT pathways
confinement/ inactivation of pathogens
death of dangerous organisms
direct killing of bacteria
ADAPTIVE
humoral immunity (B cells)
cell-mediated immunity (T cells)
B lymphoblasts
T lymphoblasts
plasma cells
helper T cells
cytotoxic T cells
ANTIBODIES
classical COMPLEMENT pathway
regulatory (suppressor) T cells
death of cells with intracellular pathogens
suppression of autoimmune responses

Figure 3.16 A simplified overview of components of the innate and adaptive immune systems. In the innate immune system, the external barriers can be physical (such as tightly locked epithelial cells of the skin and linings of the gastrointestinal tract), chemical (such as hydrochloric acid in the stomach), or biological. Various phagocytic cells and antimicrobial proteins and peptides are also deployed. The adaptive immune system is dependent on lymphocytes that develop from immature lymphoblasts: B lymphocytes that secrete antibodies and T lymphocytes that use membrane-bound receptors to recognize foreign antigens that are displayed on the surfaces of host cells. Not shown are B and T memory cells that are produced after initial exposure to foreign antigen and quickly deployed on repeated exposure to the same antigen. NK cells, natural killer cells.

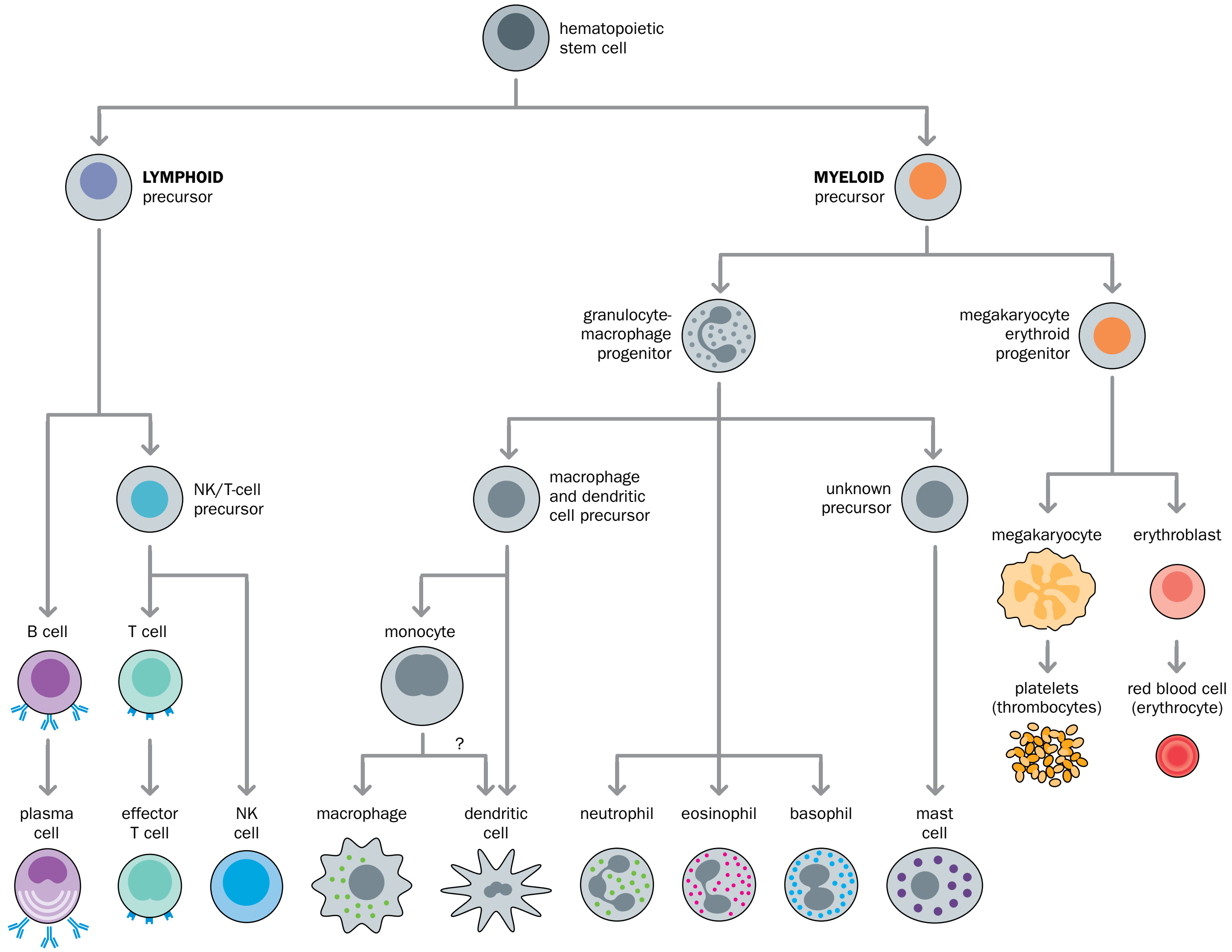
As individuals, we produce very large populations of different quiescent B and T lymphocytes, each one of which recognizes a specific foreign antigen. That provides each person with a massive repertoire of parallel antigen-detection systems, with the collective potential to recognize and deal with a huge number of possible pathogens. Exposure to specific antigens from one pathogen drives a process of selection, growth, and differentiation of only those B and T cells that can specifically recognize these antigens, producing a powerful immune response targeted against that pathogen.

For both the innate and adaptive immune systems, the immune response depends on whether the pathogen is detected outside of cells or inside cells. In the former case, certain soluble immune system proteins, such as secreted immunoglobulins (antibodies) and complement proteins, are important in destroying the pathogen. If the pathogen has infected a body cell and so escapes the attention of soluble proteins, the strategy is to kill the body cell, and seek to attack the released pathogens.

Close co-operation between the innate and adaptive immune systems is dependent on extensive cell–cell interactions and cell signaling. Immune system cells have numerous types of cell surface receptor that help them engage with other immune system cells and recognize foreign antigens or molecular patterns associated with microbes. And secreted signaling molecules called **cytokines**—notably, members of the interleukin and interferon protein families—are extensively employed to send messages between immune system cells and to co-ordinate the often complex immune responses needed to combat infection.

Origins of immune system cells and their characteristics

Active immune system cells originate from hematopoietic stem cells produced by the bone marrow, but are conveyed within blood to different tissues in the body. There are two major lineages. The lymphoid lineage leads to lymphocytes and the related natural killer (NK) cells. The myeloid lineage leads to the other types of immune system cell, including both blood cells (the granulocytes—neutrophils, eosinophils, and basophils—and monocytes) and different tissue cells (macrophages, dendritic cells, and mast cells).



	plasma cell	effector T cell	NK cell	macrophage	dendritic cell	neutrophil	eosinophil	basophil	mast cell
INNATE IMMUNE SYSTEM	–	–	+	+	+	+	+	+	+
ADAPTIVE IMMUNE SYSTEM	+	+	–	+	+	–	–	–	–

Figure 3.17 Hematopoietic stem cells in the bone marrow give rise to all blood cells and to different tissue cells with immune system functions. The multipotent hematopoietic stem cell shown at the top divides and differentiates to give more specialized progenitor cells. A lymphoid precursor gives rise to B and T lymphocytes of the adaptive immune system and natural killer (NK) cells of the innate immune system. The myeloid precursor gives rise to other classes of immune system cell, and also an erythroid lineage ultimately producing red blood cells and platelets. Macrophages and dendritic cells work in tissues and are important not just in the innate immune system, but also when they act as antigen-presenting cells in the adaptive immune system. Mast cells are another type of tissue cell and seem to be the tissue equivalent of the basophils in blood. Note: data from recent whole-genome sequencing studies are incompatible with the monophyletic theory, in which all blood cells were envisaged to arise from a common stem cell. Instead they suggest that blood cells are founded by multiple stem cells with polyphyletic ancestry. (Adapted from Parham P [2014] *The Immune System*, 4th edn. Garland Science. With permission from WW Norton.)

Another branch of the myeloid lineage gives rise to red blood cells, megakaryocytes, and platelets (**Figure 3.17**).

B and T lymphocytes

These lymphocytes, the central players of the adaptive immune system, originate in two primary lymphoid tissues: the bone marrow (where B cells complete their maturation before entering the circulation) and the thymus (immature T cells migrate from the bone marrow through the bloodstream to the thymus where they mature).

B and T cells were called lymphocytes because, unlike other blood cells, they can also circulate in the lymph, the extracellular fluid that bathes tissues. They become concentrated in multiple different secondary lymphoid tissues where they are stimulated to respond to invading pathogens. Among these peripheral lymphoid tissues are multiple lymph nodes and the spleen; following an infection, the spleen acts to filter foreign antigen from the blood, and lymph nodes filter foreign antigen from lymph. Both the spleen and lymph nodes are packed with mature immune cells, predominantly lymphocytes, but they also contain macrophages, dendritic cells, and other cells. See **Figure 3.18** for the other principal locations of secondary lymphoid tissue.

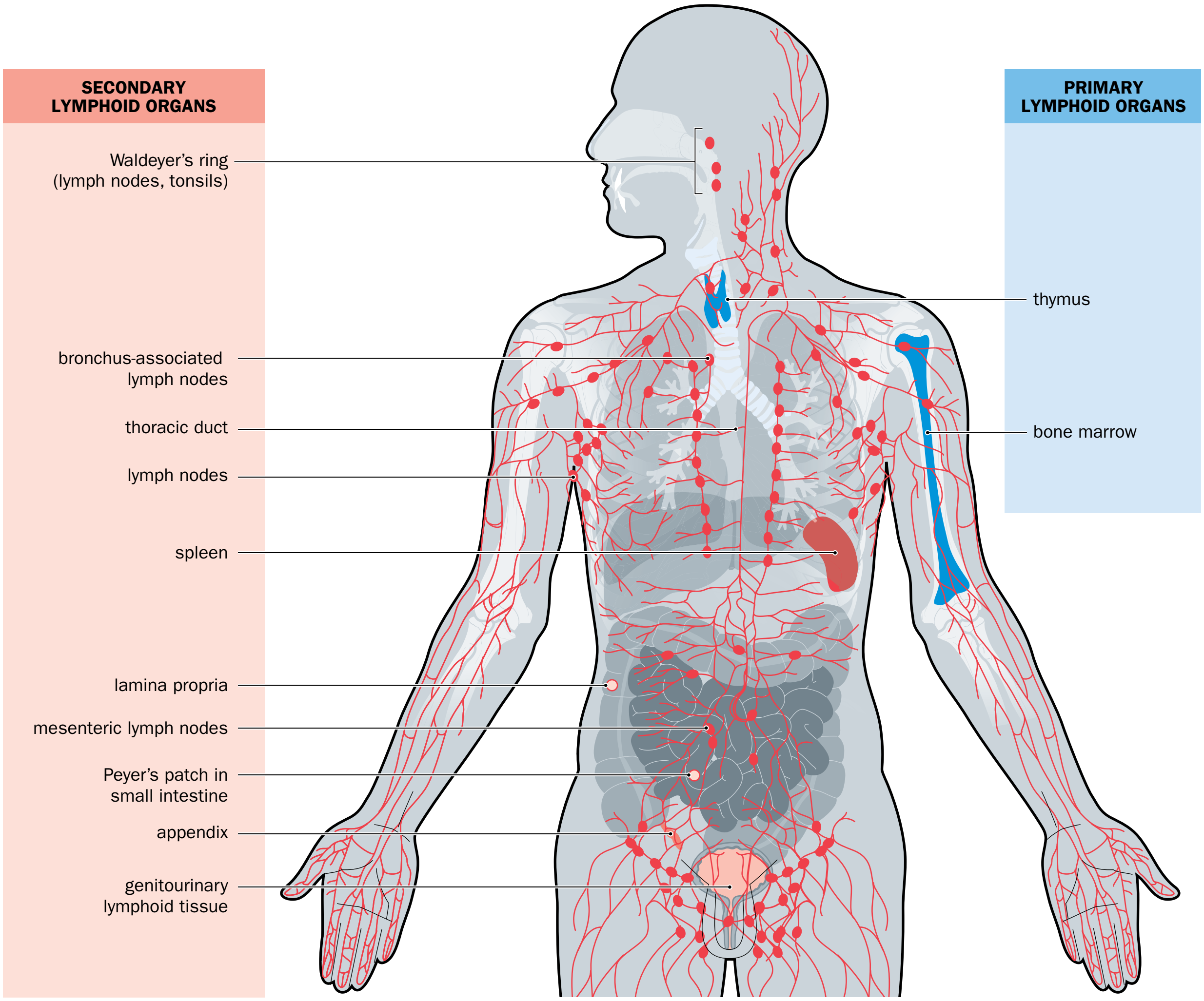

SECONDARY LYMPHOID ORGANS
Waldeyer's ring (lymph nodes, tonsils)
bronchus-associated lymph nodes
thoracic duct
lymph nodes
spleen
lamina propria
mesenteric lymph nodes
Peyer's patch in small intestine
appendix
genitourinary lymphoid tissue
PRIMARY LYMPHOID ORGANS
thymus
bone marrow

Figure 3.18 Locations of principal lymphoid tissues within the human body. The primary lymphoid organs are where B and T cells are produced: B cells complete their maturation in the bone marrow, but immature T cells migrate from the bone marrow in the blood to the thymus to complete their development. Secondary lymphoid tissues are where mature B and T cells encounter foreign antigen, which stimulates them to respond to invading pathogens and initiate adaptive immune responses. The secondary lymphoid tissues include lymph nodes (which filter antigens from lymph) and the spleen (which filters antigens from blood). The lymph nodes lie at junctions of a network of lymphatic vessels (also called *lymphatics*), with significant aggregates found close to the skin in the neck, armpit, and groin regions. (The lymphatics originate in the connective tissues of the body; they collect the plasma that is constantly leaking out of blood vessels to form the extracellular fluid, lymph, that bathes tissues, eventually returning the lymph to the blood.) Rather less organized mucosa-associated lymphoid tissue is found in various sites. Gut-associated lymphoid tissue (tonsils, adenoids, Peyer's patches in the small intestine, and lymphoid aggregates in the appendix and large intestine) collectively constitutes the largest lymphoid organ, consistent with its need to interact with a huge load of antigens from food and commensal bacteria. Epithelial-associated lymphoid tissue is also found in the skin and in the mucous membranes lining the upper airways, bronchi, and genitourinary tract. (Adapted from Parham P [2014] *The Immune System*, 4th edn. Garland Science. With permission from WW Norton.)

B cells are distinguished by the making of immunoglobulins (Ig). Early ("naive") B cells (B lymphoblasts) make IgM or IgD immunoglobulins that are incorporated into the cell membrane (a transmembrane B-cell receptor). After exposure to antigens, however, the B lymphoblasts are stimulated primarily in the lymph nodes to make effector B cells known as plasma cells. Instead of making a membrane-bound B-cell receptor, plasma cells secrete soluble immunoglobulins (IgM, IgG, IgA, or IgE classes) as **antibodies** that can recognize a specific antigen and combat bacterial infection in ways that we describe below.

T cells are distinguished by the making of a transmembrane receptor known as a **T-cell receptor**. The job of effector T cells is to recognize and deal with sick or damaged host cells that express foreign antigen on their cell surface, including, notably, virus-infected cells. To do that, they are assisted by certain antigen-presenting cells that present the foreign antigen on the cell surface as a complex with a major histocompatibility complex (MHC) protein. As described below, cytotoxic T lymphocytes induce the death of the harmful body cells, but various other types of effector T cells help in the process.

Natural killer (NK) cells

NK cells are large, lymphocyte-like effector cells of the innate immune system and are important in the defense against viral infections. Their job is to enter infected tissues and limit the spread of the infection in two ways: by killing body cells infected by the virus, and by secreting certain types of cytokine to impede viral replication in host cells.

Granulocytes (polymorphonuclear leukocytes)

As the name suggests, these white blood cells have prominent cytoplasmic granules and irregularly shaped nuclei (usually with two to five lobes). The cytoplasmic granules contain assorted antimicrobial reagents—defensins (peptides that insert into and then disrupt the cell membranes of pathogens), lysozyme, myeloperoxidase, and so on—that can be secreted as required. There are three types of granulocyte, as listed below.

- Neutrophils are the most common type of white blood cell, and an important effector cell of the innate immune system. They are also the most abundant and most lethal type of **phagocyte** (a cell that specializes in engulfing and killing a microbial pathogen; see Figure 3.19). Large reserves of neutrophils are stored in the bone marrow and are mobilized when needed to fight infection. They travel in the bloodstream and from there to infected tissues where they engulf and kill bacteria, but they are short-lived and die at the infection site, forming pus.
- Eosinophils are comparatively rare (1–6% of white blood cells) and protect against helminth worms and other intestinal parasites.
- Basophils are very rare and also protect against parasites, being recruited into tissues at sites of infection.

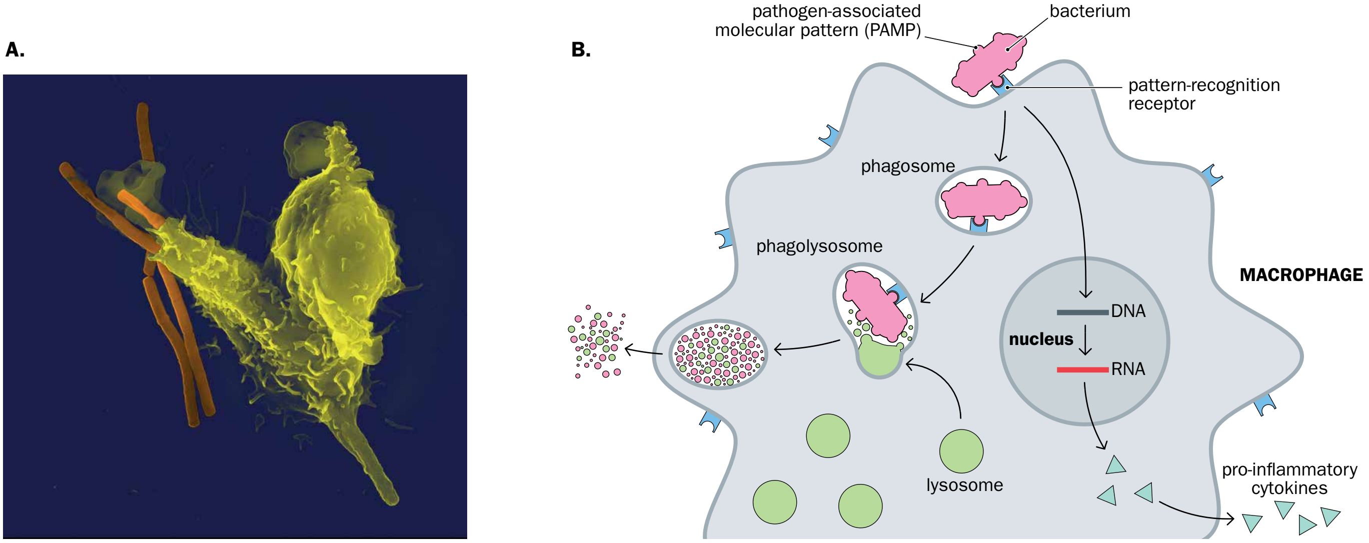

A.
B.
pathogen-associated molecular pattern (PAMP)
bacterium
pattern-recognition receptor
phagosome
phagolysosome
MACROPHAGE
DNA
nucleus
RNA
lysosome
pro-inflammatory cytokines

Figure 3.19 How immune system phagocytes kill microbial pathogens. (**A**) A single neutrophil (yellow) engulfing rod-shaped anthrax bacilli (orange). Like other phagocytes, neutrophils engulf microbial pathogens, which are then destroyed within the phagocyte. (**B**) Immune system phagocytes, such as macrophages, have cell surface receptors that can recognize certain types of pattern on microbes (such as components of bacterial cell walls), identifying them as foreign cells. The process of phagocytosis begins with binding of a bacterium (or other microbe) by cell surface receptors, followed by internalization of the microbe within a vacuole called a phagosome. Lysosomes fuse with phagosomes to form phagolysosomes and then discharge their hydrolytic enzymes and dangerous chemicals to degrade the microbe. A signal-transduction pathway is also triggered when the pattern-recognition receptor binds the pathogen, resulting in activated transcription of genes that make inflammatory cytokines. The secreted cytokines bind to surface receptors on other immune system cells, recruiting them to participate in the immune response. (A, original image by Volker Brinkmann [2005] *PLoS Pathog* **1**(3):cover page; B, used with permission from Dr Victoria J Drake and Linus Pauling Institute Micronutrient Information Center at Oregon State University.)

Monocytes and macrophages

Monocytes account for about 2–10% of circulating white blood cells and are also abundant in the spleen. They are bigger than granulocytes and have a more consistent appearance with a distinctive indented nucleus. They are a class of free-roaming phagocyte and they also give rise to long-lived, specialized tissue phagocytes known as **macrophages**, the general scavenger cells of the body that are particularly active in phagocytosing dead cells and debris as well as invading microorganisms.

Tissue macrophages arise after monocytes respond to inflammation signals: the monocytes migrate rapidly in the blood and enter tissues at infection sites, whereupon they differentiate. According to the tissue that they inhabit, tissue macrophages can be known by other names, such as Kupffer cells (liver), microglia (brain and spinal cord), and osteoclasts (bone).

Dendritic cells and mast cells

Dendritic cells are star-shaped tissue immune cells with many of the properties of macrophages, but their main purpose is to act as messenger cells: when required, they are sent to summon up an adaptive immune response. If the initial innate immune response to infection seems to be inadequate, dendritic cells within the infected tissue migrate to one of the secondary lymphoid tissues that specialize in making adaptive immune responses.

Mast cells are granulated cells that seem to be the tissue equivalent of basophils, being present in all connective tissues. When activated at sites of infection they release their cytoplasmic granules (degranulation). They are important in inflammation and allergic responses.

The innate immune system: countering pathogens using barriers and a rapid response based on general pattern recognition

The innate immune system provides defense against, and an immediate response to, all types of pathogen. It works at two levels in the body. An external defense system is designed to prevent active pathogens gaining access to internal tissues and body fluids (by internal tissues, we mean those beyond the surface epithelium of the skin, digestive tract, respiratory tract, urinary tract, and so on). The external defense system comprises some barriers that simply act as passive defensive shields plus active defense systems that seek to inactivate pathogens before they get to internal tissues.

The second level of defense is internal. It occurs when the external barriers have been breached, and pathogens have gained access to internal tissues and body fluids. Then a rapid immune response (which is much the same from one normal individual to another) is mounted to combat infection. (If the defense is unsuccessful after about four or so days, the adaptive immune system is pressed into action as a last resort.)

Innate immune responses depend on two components. First, there must be highly accurate pattern-recognition mechanisms that identify microbial pathogens as being foreign, for example by identifying protein or carbohydrate components of the cell walls or cell membranes of the pathogen. Second, effector mechanisms are then activated to kill the recognized invaders, requiring diverse immune system cells and many different soluble proteins and cell surface proteins.

Defensive barriers

To protect us against dangerous pathogens, defense is a priority: surface layers in contact with the environment must be securely protected from pathogens, notably those of the skin and the gastrointestinal and respiratory tracts. In the case of skin, the epidermis (the top, outermost layers) is made of stratified squamous epithelium sheets that are renewed by stem cells in the lower (basal) layers. The layers lying above the basal layers progressively contain differentiated cells, culminating in an outer layer of very tough keratinocytes that lack nuclei and other organelles, providing a resistant outer coat. If the skin is damaged by cuts, the resulting bleeding and blood clotting process can provide a temporary repair while the skin effects proper repairs.

Epithelial cells, such as those of the intestinal epithelium, can also benefit by a type of cell junction known as a tight junction that provides the closest contact between adjacent cells in nature. The tight junctions act like bands that encircle each epithelial cell close to the surface and also attach the cell tightly to its neighbors, preventing molecules from diffusing across the epithelial sheet between adjacent cells (see **Figure 3.11**).

Mucus, a viscous, slippery substance secreted by mucous membranes, acts as a protective lubricant coating the cells and glands of some mucous membranes, such as the nasal mucosa and mucosal tissues lining the airways. The thick mucus can trap bacteria; thereafter, coordinated beating of cilia lining the surrounding epithelia drives the trapped bacteria upward to the throat, to be expelled by coughing or sneezing, or to be swallowed and destroyed in the stomach.

Chemical and biological barriers include hydrochloric acid in the stomach, antimicrobial enzymes (lysozyme) in tears and saliva, antimicrobial peptides (notably defensins), and the nonpathogenic bacteria of our microbiome. Defensins, peptides with about 35–40 amino acids, are constitutively secreted at mucosal surfaces and can enter the cell membranes of microbes and disrupt them by forming pores. The nonpathogenic microbes on our skins, and notably in the gut, discourage colonization by pathogenic microbes simply by providing competition for growth resources.

Complement activation

In addition to phagocytosis (by macrophages, monocytes, and neutrophils), one of the very first innate immune responses is provided by the complement system, often just called **complement**. This group of more than 20 interacting soluble complement proteins is found in blood and lymph and works with complement receptors, found notably on the plasma membranes of macrophages. Complement proteins are mostly made by the liver, but blood monocytes, tissue macrophages, and epithelial cells of the gastrointestinal and genitourinary tracts can also make significant amounts.

Constitutively produced complement proteins circulate in blood and lymph but remain inactive until the complement system is activated by some trigger. Then the complement proteins work in pathways that begin with a protease cascade (an activated complement protease specifically cleaves another complement protein, converting it in turn to an active protease that cleaves the next complement protein in order to activate it, and so on).

The major purpose of complement is to kill or inactivate microbial pathogens and induce inflammatory responses. In the former case, pathogenic bacteria can be killed directly when complement proteins form a membrane attack complex that punches holes in the pathogen's cell membrane (**Figure 3.20A**). An additional option is indirect killing of pathogenic microbial cells (and inactivation of viruses) by assisting macrophages: a complement protein, C3b, is deposited on the surface of microbial pathogens to make them more readily recognized and destroyed by phagocytes such as macrophages (a process called **opsonization**; see **Figure 3.20B**).

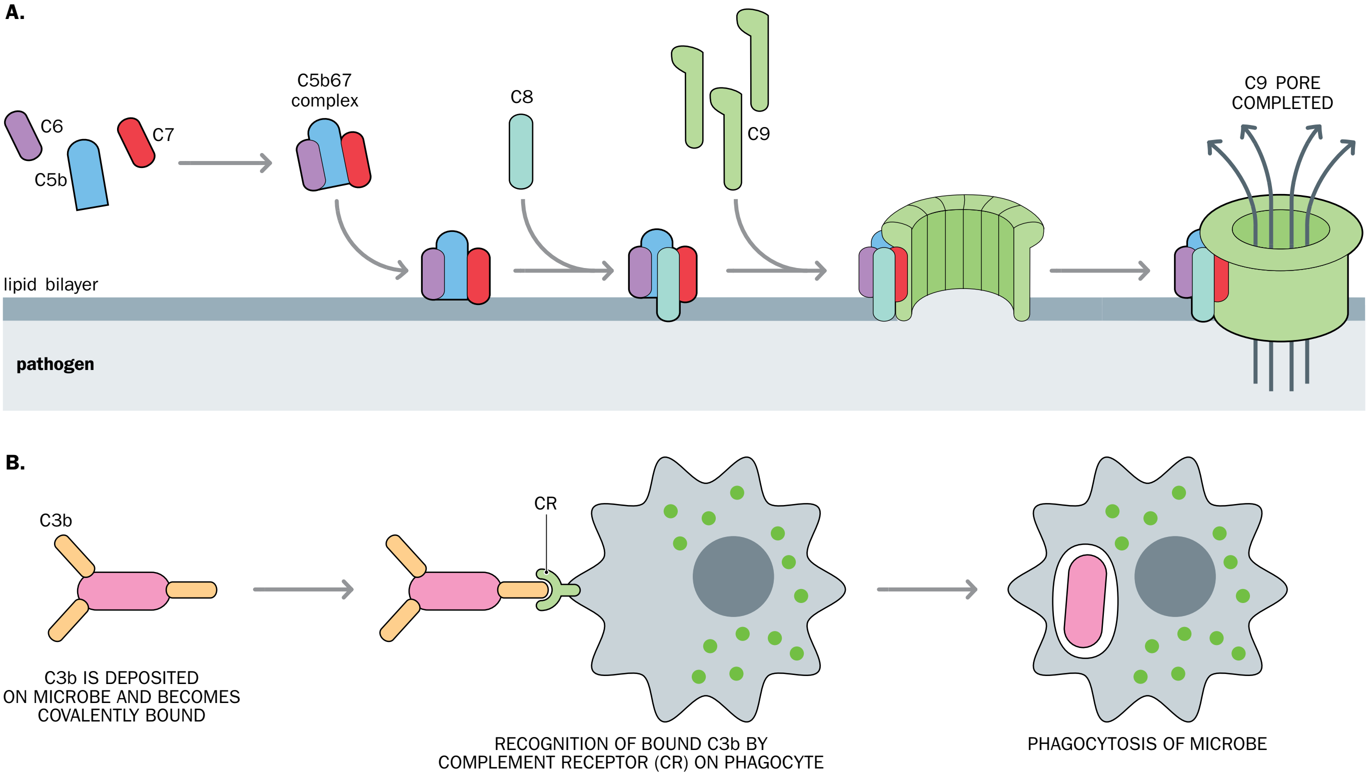


Figure 3.20 Two principal effector mechanisms of the complement system. (**A**) Killing of microbial pathogens by assembling a membrane attack complex to perforate the cell membrane. A complex of complement proteins C5b, C6, C7, and C8 assembles at the cell membrane of the pathogen and is able to recruit multiple copies of the complement C9 that integrates into the cell membrane and forms a large pore. Dark gray arrows indicate the resulting efflux of essential material from the cell, causing it to be destroyed. (**B**) Assisting macrophages by opsonization. Complement C3b selectively coats the surfaces of microbial pathogens (by covalently binding to amino or hydroxyl groups of cell surface molecules) and targets them for destruction by phagocytes, notably macrophages. The latter have cell surface complement receptors (CR) plus pattern-recognition receptors to identify microbial cells (and viruses). After binding the C3b-coated pathogen they internalize it and destroy it by phagocytosis (as shown in **Figure 3.19B**). (Adapted from Parham P [2014] *The Immune System*, 4th edn. Garland Science. With permission from WW Norton.)

According to the trigger, three variant complement pathways are sequentially employed in innate immune responses (Table 3.4). Although the initiating steps are different in the three pathways, in each case production of the key complement C3 protein is amplified to very large quantities. Small amounts of C3 spontaneously hydrolyze to give two fragments, C3a and C3b, but when a complement pathway is initiated, C3 convertase enzymes are produced that cleave C3 to produce large amounts of C3a and C3b and they become important effector molecules (Figure 3.21).

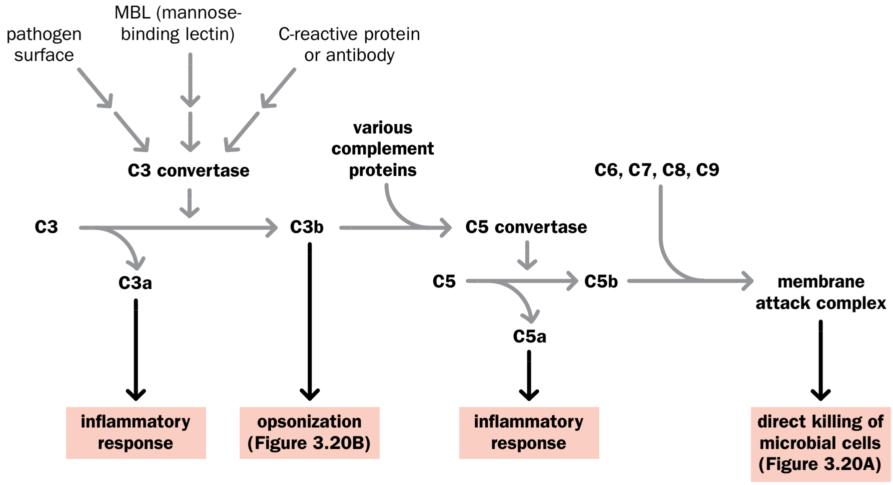

pathogen surface
MBL (mannose-binding lectin)
C-reactive protein or antibody
C3 convertase
various complement proteins
C6, C7, C8, C9
C3
C3b
C5 convertase
C3a
C5
C5b
membrane attack complex
C5a
inflammatory response
opsonization (Figure 3.20B)
inflammatory response
direct killing of microbial cells (Figure 3.20A)

Figure 3.21 Complement proteins and effector functions. Different upstream initiation events are used in the three different complement pathways that respond to different triggers (see **Table 3.4**). In each case, complement activation results in cleavage of the complement C3 protein (which is produced in large amounts) to give two fragments. The larger fragment, C3b, is used in two ways. First, it works in opsonization, coating the surface of microbial pathogens so that they can be recognized and killed by phagocytes (such as macrophages, which have complement receptors). Second, it binds to various other complement proteins to produce a complex that cleaves the complement C5 protein to give a small fragment C5a plus a large C5b protein. C5b binds to other complement proteins to form a membrane attack complex that directly kills pathogens (see **Figure 3.20**). The small peptides C3a and C5a act on nearby blood vessels to augment local inflammatory responses.

When C3 is cleaved, a thioester bond that had been hidden in the hydrophobic interior of C3 is suddenly exposed, as part of C3b, to the hydrophilic environment. Although the thioester bonds of most C3b fragments are spontaneously hydrolyzed by water, some react with hydroxyl and amino groups on molecules on the surface of a pathogenic microbial cell or virus and C3b becomes covalently bound to the pathogen (complement fixation). The C3b tags on the surface of the pathogen allow it to be recognized and bound by macrophages (which have complement receptors—see **Figure 3.20B**—plus other receptors that can recognize cell surface components of pathogens, as described in the next section). C3b also gives rise to C5b, which is important in forming membrane attack complexes, and C3a and another complement peptide that are important in inflammatory responses (see **Figures 3.20A** and **3.21**).

TABLE 3.4 CHARACTERISTICS OF THE THREE PATHWAYS OF COMPLEMENT ACTIVATION

Pathway	Trigger	Where and when used
Alternative pathway	The pathogen surface is by itself sufficient to create a local environment conducive to complement activation	Innate immune response only. Initiated at the start of infection
Lectin pathway	The plasma MBL (mannose-binding lectin) protein recognizes carbohydrate patterns found on the surface of a large number of pathogenic microorganisms (including bacteria, viruses, protozoa, and fungi). MBL is one of a group of acute-phase proteins whose plasma concentrations increase in response to cytokines secreted by neutrophils and macrophages at inflammation sites	Innate immune response only. Induced by infection, but takes some time to act
Classical pathway	The trigger in innate immune responses is the plasma C-reactive protein (CRP), an acute-phase protein. In adaptive immune system responses, antigen–antibody binding is the trigger	Innate and adaptive immune responses. The last of the three pathways to be activated

Pattern-recognition mechanisms

In the adaptive immune system, special mechanisms exist to diversify antibodies and T-cell receptors so that we can each recognize huge numbers of foreign antigens. Given the extensive variety of pathogenic organisms and viruses, how does the innate immune response distinguish pathogens? The answer is pattern-recognition mechanisms used by a wide variety of receptors in various cell types, including epithelial cells as well as immune system cells such as macrophages. The receptors scan for particular types of molecular patterns that are unusual for body cells but are instead associated with pathogens (for example, components of cell walls of bacteria) or with danger.

As an illustration of the diversity of pattern-recognition receptors, we provide some characteristics of three of the more important receptor families in **Table 3.5**. Classification of receptor families is primarily based on shared structural motifs, but there can be wide variation in the types of ligand recognized by members of a receptor family. Thus, whereas members of the C-lectin receptor family recognize various carbohydrate patterns on the surface of fungal and bacterial cells, members of the Toll-like receptor family detect a very wide range of ligands (**Table 3.5**).

Scanning for pathogen-associated molecular patterns occurs at different levels. Plasma membrane receptors carry out extracellular scans, and receptors located on internal endosomal membranes detect their ligands in intracellular vesicles. Some pattern-recognition receptors act as true intracellular sentinels, including cytoplasmic and occasionally nuclear receptors, and are able to detect unusual molecular patterns within these intracellular compartments. For example, the cytoplasmic receptors NOD1 and NOD2 detect bacteria or bacterial components entering the cytoplasm, and in response initiate an NF-κB signaling pathway that produces various cytokines that trigger inflammatory responses.

Note that the pattern-specific receptors are also important in recruiting the adaptive immune system. For example, once stimulated, Toll-like receptors induce the surface expression of co-stimulatory molecules that are essential for initiating adaptive immune responses (see below) and stimulate the secretion of pharmacologically active molecules (mostly prostaglandins and cytokines) that both initiate an inflammatory response and also help induce an adaptive immune response.

TABLE 3.5 SOME CHARACTERISTICS OF THREE IMPORTANT FAMILIES OF INNATE RESPONSE PATTERN RECEPTORS

Receptor family	Human family members	Cellular locations	Examples of ligands for individual receptors	
Toll-like receptors (TLR)	Ten members (TLR1–10)	Transmembrane (TM) receptors that line the plasma and endosomal membranes	TLR2	Peptidoglycan and lipoteichoic acid in bacterial cell walls; zymosan from yeast cell walls; lipoarabinomannan from mycoplasma
			TLR3	Double-stranded RNA (a pattern associated with viral infection)
			TLR4	Lipopolysaccharide in the outer wall of all gram-negative bacteria
			TLR5	Flagellin, a protein in the flagellum of gram-positive and gram-negative bacteria
			TLR9	Specific unmethylated CpG motifs present in microbial DNA but absent in vertebrate DNA
C-lectin type receptors	Dectin-1, dectin-2; mannose receptor, mannose-binding lectin (MBL), and many more	Mostly TM receptors on plasma membrane and endosomal membranes	Dectin-1	β-1,3-glucans in many fungi and mycobacteria
			Dectin-2	α-mannans in fungi
			MBL	Various carbohydrates, notably mannans in fungi and bacteria
NOD-like receptors	NOD1, NOD2, plus 20 others	Some in cytoplasm, others in plasma membrane and endosomal membranes	NOD1	Meso-diaminopimelic acid, a peptidoglycan component from the cell wall of gram-negative bacteria
			NOD2	Muramyl dipeptide, a peptidoglycan component from the cell walls of both gram-negative and gram-positive bacteria

Natural killer (NK) cells

NK cells, a class of giant, granular, cytotoxic lymphocyte (but lacking antigen-specific receptors), have an effector function that is broadly similar to that of cytotoxic T lymphocytes (CTLs) in the adaptive immune system: to induce apoptosis in virus-infected cells and other damaged or abnormal body cells, such as tumor cells. In the former case, recall that extracellular viruses can be coated by complement C3b and then targeted for destruction by macrophages, but viruses remaining inside cells are not visible to the complement system. However, once viruses have been detected inside body cells, NK cells are recruited to induce the virus-infected cells to undergo apoptosis.

Both intrinsic and extrinsic apoptosis pathways are used by NK cells. In the former case, NK cells bind to diseased cells, and via exocytosis release the contents of their secretory granules (perforins and granzymes) into the intercellular space (**Figure 3.22**). The perforins insert into the membrane of the target cell in a way that creates pores in the membrane to allow the pro-apototic granzymes to enter the target. For the extrinsic pathway, Fas ligands on NK cell membranes activate Fas receptors on the surface of target cells to initiate apoptosis (see **Figure 3.10** for the mechanism).

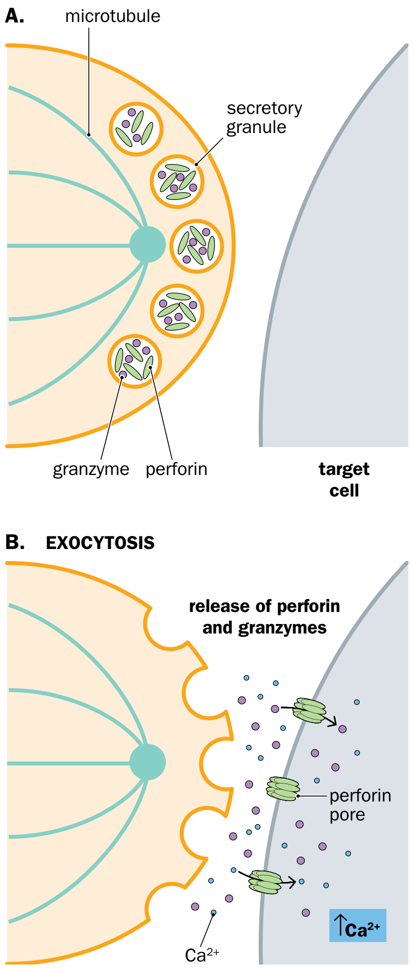


Figure 3.22 NK cells can induce apoptosis by secreting perforins and granzymes close to the surface of target cells. (**A**) Cytoplasmic secretory granules, containing perforin and granzyme molecules, are transported via the microtubule network toward the NK cell membrane at a point close to the target cell. (**B**) The secretory granules fuse with the NK cell plasma membrane (exocytosis) and released perforins form large transmembrane pores in the target cell membrane, enabling the diffusion of granzymes into the cytosol of the target cell. The granzymes then initiate intrinsic apoptosis pathways by cleaving procaspases or by cleaving the proapoptotic BID protein to activate the mitochondrial apoptosis pathway (see **Figure 3.10**). (Adapted from Voskoboinik I *et al*. [2015] *Nat Rev Immunol* **15**:388–400; PMID 25998963. With permission from Springer Nature. Copyright © 2015.)

The adaptive immune system mounts highly-specific immune responses that are enhanced by memory cells

The adaptive immune system arose in early vertebrates, providing a powerful additional defense against pathogens. It is mobilized by components of the innate immune system when this fails to provide adequate protection against an invading pathogen. It has three key characteristics:

- Exquisite specificity: it can discriminate between tiny differences in molecular structure;
- Extraordinary adaptability: it can respond to an unlimited number of molecules;
- Memory: it can remember a previous encounter with a foreign molecule and respond more rapidly and more effectively on a second occasion.

Whereas innate immune responses are much the same in all healthy members of a species, adaptive immune responses vary between individuals: one individual may mount a strong reaction to a particular antigen that another individual may never respond to. There are two major arms to the adaptive immune response:

- Humoral (antibody) immunity is mediated by B lymphocytes (also called **B cells**)
- Cell-mediated immunity is effected by T lymphocytes (**T cells**) and is an important antiviral defense system

Both B and T cells have dedicated cell surface receptors that can specifically recognize individual antigens.

What makes the adaptive immune system so proficient at recognizing antigens is that during its development in the primary lymphoid organs, each naive B or T cell acquires a cell surface antigen receptor of a unique specificity. Binding of this receptor to its specific antigen activates the cell, causing it to proliferate to give a clone of cells with the same immunological specificity as the parent cell (**clonal selection**, see **Figure 3.23**). As a result, the number of lymphocytes that can recognize the specific antigen can be rapidly expanded.

Immunological memory is another important feature of the adaptive immune system. During the massive clonal expansion of antigen-specific lymphocytes that occurs following an initial encounter with antigen, some of the expanding daughter cells

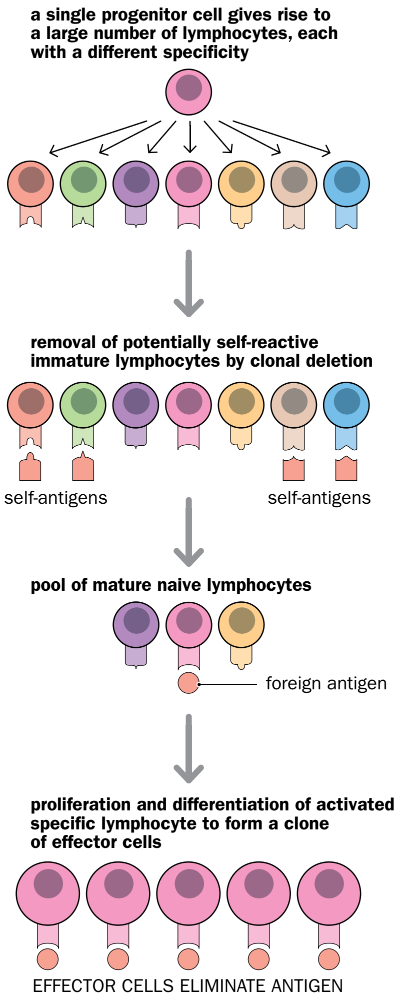

a single progenitor cell gives rise to a large number of lymphocytes, each with a different specificity
removal of potentially self-reactive immature lymphocytes by clonal deletion
self-antigens
self-antigens
pool of mature naive lymphocytes
foreign antigen
proliferation and differentiation of activated specific lymphocyte to form a clone of effector cells
EFFECTOR CELLS ELIMINATE ANTIGEN

Figure 3.23 Clonal selection of lymphocytes is the central principle of adaptive immunity. Each lymphocyte progenitor gives rise to many lymphocytes, each bearing a distinct antigen receptor. Lymphocytes with receptors that bind ubiquitous self-antigens are eliminated before they become fully mature, ensuring tolerance to such self-antigens. When antigen interacts with the receptor on a mature naive lymphocyte, that cell is activated and starts to divide. It gives rise to a clone of identical progeny, all of whose receptors bind the same antigen. Antigen specificity is thus maintained as the progeny proliferate and differentiate into effector cells. Once antigen has been eliminated by these effector cells, the immune response ceases. (Adapted from Parham P [2014] *The Immune System*, 4th edn. Garland Science. With permission from WW Norton.)

differentiate into memory cells that are able to respond to the antigen more rapidly or more effectively. Memory cells are endowed with higher-affinity antigen receptors, and also combinations of adhesion molecules, homing receptors, and cytokine receptors that direct the lymphocytes to migrate efficiently through the walls of blood vessels into specific tissues (extravasation).

Humoral immunity depends on the activities of soluble antibodies

B cells produced in the bone marrow have cell surface immunoglobulins (Igs) as their antigen receptors (B-cell receptors). When B cells are activated, however, they can differentiate into plasma cells that secrete their Ig receptors as soluble antibodies.

Immunoglobulin structure

Igs are composed of two identical heavy chains and two identical light chains that are held together by disulfide bonding (**Figure 3.24**). The light chains can be one of two varieties (κ or λ) that are functionally equivalent. The heavy chains can be one of five functionally distinct types that define five immunoglobulin classes (**Table 3.6**). Each Ig chain contains two distinct regions, as listed below.

- An N-terminal **variable region** is involved in antigen binding and has a variable sequence that accounts for the unique specificity of each B-cell receptor and antibody. The variability is not distributed evenly throughout the domain but is concentrated in hypervariable regions (also called **complementarity determining regions**), which are the regions that directly interact with antigen.
- A C-terminal **constant region** is invariant within each class of immunoglobulin but differs significantly between the different heavy-chain classes, accounting for the different functionality of each isotype.

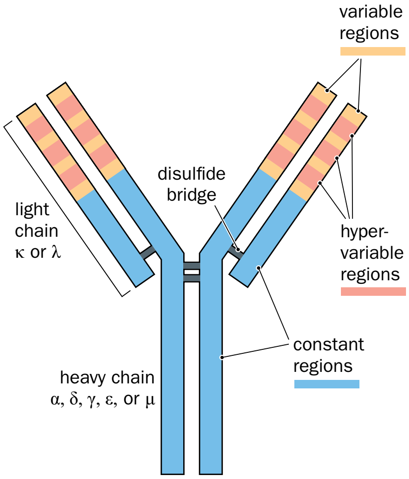


Figure 3.24 Antibody (immunoglobulin) structure. Antibodies are soluble immunoglobulins (Igs) that consist of two identical heavy chains (one of five classes) and two identical light chains (one of two classes). The two heavy chains are held together by disulfide bonds and each light chain is linked to one heavy chain by disulfide bridges. Each chain is composed of globular domains that are maintained by intrachain disulfide bridges. The N-terminal regions are known as *variable regions* as their sequence varies significantly from one antibody to another. Most of the sequence variation is concentrated in *hypervariable regions* (also called *complementarity determining regions*), which are the sequences involved in antigen binding. The constant region of each heavy chain determines the class of the heavy chain and the response to bound antigen (see **Table 3.6**). Sequences at the bottom of each heavy chain can be recognized by specific receptors (Fc receptors) on the surface of effector cells.

Light and heavy chains are structurally closely related, being made up of Ig domains, each ~100 amino acids long and held together by an internal disulfide bond. The light chain has two Ig domains, one each for the variable and constant regions; the heavy chain has one variable Ig domain and three constant Ig domains—see the B-cell receptor/Ig structure in **Figure 3.25**

Immunoglobulin function

Antibodies can bind to specific receptors at multiple regions of the surface of microbes, forming a surface coat. Once a microbe has been coated with IgM or IgG antibodies the classical pathway of complement activation can be activated, causing lysis of the microbes. Complement can also promote phagocytosis via complement receptors on

TABLE 3.6 IMMUNOGLOBULIN ISOTYPES

Isotype	Heavy (H) chain	Structure	Working form	Location/functions
IgM	μ	Pentamer/monomer	B-cell receptor (monomer) and secreted antibody (pentamer)	First Ig to be produced. Activates complement
IgD	δ	Monomer	B-cell receptor only	Second Ig to be produced. Function is unclear
IgA	α	Often dimer or tetramer	B-cell receptor, but predominantly as antibodies secreted by plasma cells	Predominant Ig in external secretions such as breast milk, saliva, tears, and mucus of the bronchial, genitourinary, and digestive tracts
IgG	γ	Monomer		Principal serum Ig. Binds to Fc receptors on phagocytic cells. Activates complement
IgE	ε	Monomer		Binds to Fc receptors of mast cells and blood basophils. Mediates immediate hypersensitivity reactions responsible for symptoms of allergic conditions such as hay fever and asthma

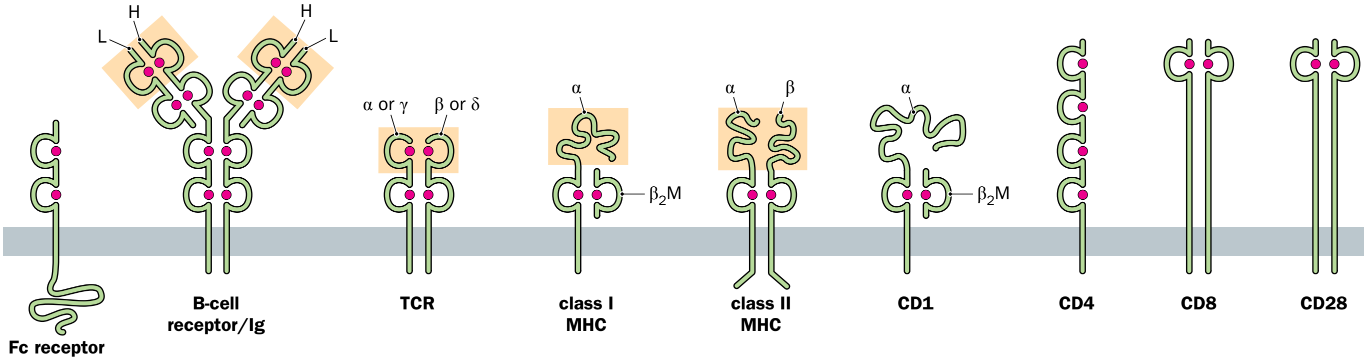


Figure 3.25 The immunoglobulin superfamily. The immunoglobulin (Ig) domain is a barrel-like structure held together by a disulfide bond (red dot). As shown here, multiple copies of the Ig domain are found in many proteins with important immune system functions—the pale orange boxes show variable domains. Note that immunoglobulins have essentially the same structure as the B-cell receptor (they differ by lacking membrane-binding sequences near the end of the constant region). TCR, T-cell receptor; β_2M, β_2-microglobulin.

macrophages and neutrophils. In addition, antibodies have several other important functions as listed below.

- Blocking pathogen entry into cells. Viruses and certain microorganisms that can exist within animal cells enter into the cells by binding to certain preferred receptors on the cell surface. Antibodies can physically block this process (**Figure 3.26A**).
- Neutralizing toxins. Antibodies can directly bind and neutralize toxins released by bacteria, inhibiting their enzymatic activity or their ability to bind to cell surface receptors (**Figure 3.26B**).
- Activating effector cells. Many immune system cells express receptors that recognize the invariant domains of Ig molecules. Antibodies bound to the surface of microbes can therefore not only activate complement but also directly activate those immune system cells that carry appropriate Fc receptors. IgG antibodies, for example, promote phagocytic uptake by neutrophils and macrophages and antibody-dependent cell-mediated cytotoxicity by NK cells (**Figure 3.26C**). IgE bound to IgE-specific Fc receptors on eosinophils, basophils, and mast cells can trigger the release of powerful pharmacological mediators and activate these cells to kill antibody-coated parasites.

In cell-mediated immunity, T cells recognize cells containing fragments of foreign proteins

Microorganisms and viruses that penetrate cells and multiply within them are out of reach of antibodies. T cells are equipped to deal with this need. On their cell surfaces are dedicated T-cell receptors (TCRs) that, like antibodies, show very high specificity in recognizing and binding sequences from foreign antigens. The TCRs are structurally and evolutionarily related to Igs (see **Figure 3.25**).

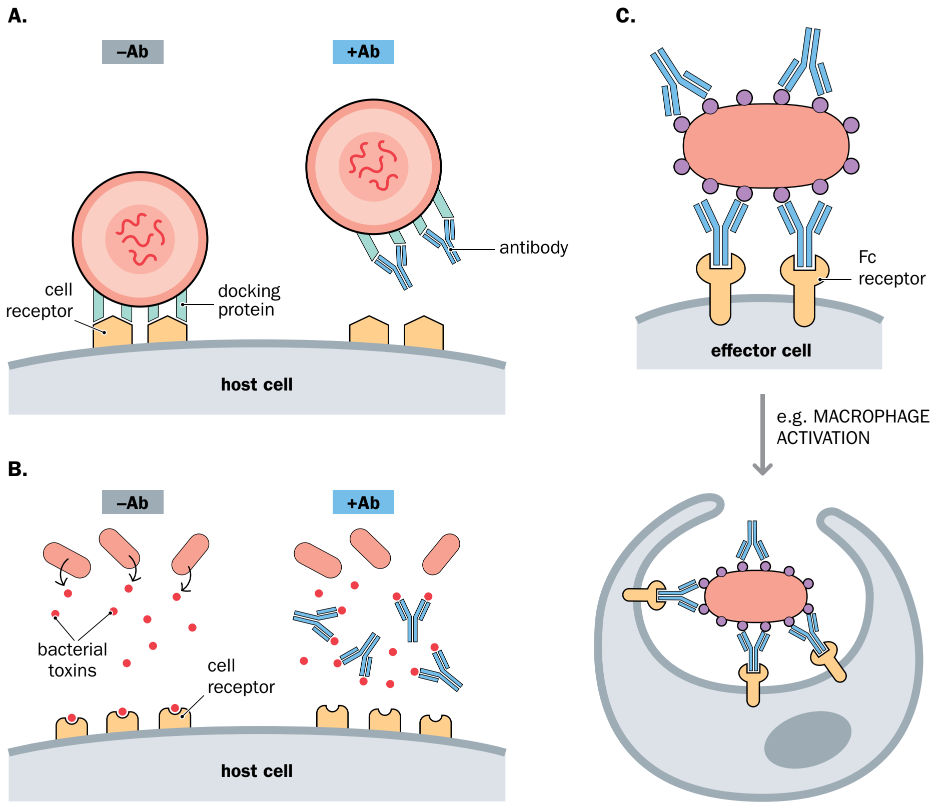


Figure 3.26 Aspects of antibody function. (**A**) *Inhibiting viral infection*. Viruses infect cells by first using docking proteins to bind to certain receptors on the plasma membrane of host cells. Antibodies can bind to the viral docking proteins to prevent them binding to host-cell receptors. (**B**) *Neutralizing toxins*. Antibodies bind to toxins released from invading microbes and so stop them binding to cell receptors. (**C**) *Activating effector cells*. Antibodies can bind and coat the surfaces of microbes and large target cells. Various immune system effector cells (notably macrophages, natural killer cells, neutrophils, eosinophils, and mast cells) carry Fc receptors that enable them to bind to the Fc region on IgA, IgG, or IgE antibodies. Antibody binding to an Fc receptor activates the effector cell and can lead to cell killing by phagocytosis, or release of lytic enzymes, death signals, and so on. (Adapted from Alberts B *et al*. [2014] *Molecular Biology of the Cell*, 6th edn. Garland Science. With permission from WW Norton.)

TCRs are heterodimers and come in two classes. The vast majority have α and β chains (αβ TCRs); a few have γ and δ chains (γδ TCRs). Unlike the immunoglobulins of B cells, which can collectively recognize foreign antigens on a wide range of different molecular classes, TCRs are especially focused on protein antigens (the predominant αβ TCRs are limited to recognizing protein components).

There are three major classes of T cells with αβ TCRs—killer (cytotoxic) T cells, helper T cells, and regulatory T cells—and individual αβ T cells that have previously encountered foreign antigen can be stimulated to undergo clonal expansion (Table 3.7).

TABLE 3.7 CLASSES OF T CELLS WITH αβ T-CELL RECEPTORS THAT BIND MHC–PEPTIDE ANTIGENS

T cell class	Features	Roles
Killer (cytotoxic) T cells	Bind predominantly to class I MHC–peptide signals on antigen-presenting cells. A CD8 receptor binds to the non-polymorphic part of the MHC protein while the T-cell receptor binds to the peptide and variable regions of the MHC protein	Kill virally-infected host cells and tumor cells. Like NK cells, they induce apoptosis, either by the Fas pathway (see **Figure 3.10**) or by delivering granules containing perforin and granzymes (**Figure 3.22**)
Helper T cells (T_h cells)	Bind predominantly to class II MHC–peptide signals on antigen-presenting cells. Usually have a CD4 receptor that binds to the non-polymorphic part of the MHC antigen while the T-cell receptor binds to the variable regions and peptide (see **Figure 3.27**). There are three major subclasses: T_h1, T_h2, and T_h17 cells	Signal to other immune system cells, stimulating them to proliferate and be activated: T_h1 cells activate macrophages and killer T cells T_h2 cells activate eosinophils and promote antibody production by B cells T_h17 cells promote autoimmunity and inflammation
Regulatory (suppressor) T cells	Like helper T cells, they express CD4 but can be distinguished by constitutive cell surface expression of CD25 (one of the chains of the interleukin-2 receptor) and intracellular expression of the transcription factor FOXP3	Suppress autoimmune responses
Memory T cells	Originate by clonal expansion of an αβ T cell that has previously encountered a foreign antigen	Stimulated to expand rapidly in secondary immune responses
Additional T-cell classes include NKT cells (having T-cell receptors that interact with CD1 proteins instead of MHC proteins) and γδ T cells. These T-cell classes are important regulatory cells and include cells that recognize glycolipid antigens.		

Although αβ TCRs recognize protein components, they do so only after the proteins have been degraded inside cells, whereupon the resulting peptides are individually bound by a newly made **major histocompatibility complex** (MHC) protein. Holding the peptide in a cleft, an MHC protein transports it to the surface of the cell (see Box 3.2 for a background on MHC genes and proteins, and protein degradation within cells).

BOX 3.2 THE MAJOR HISTOCOMPATIBILITY COMPLEX AND MHC PROTEIN STRUCTURE AND FUNCTION

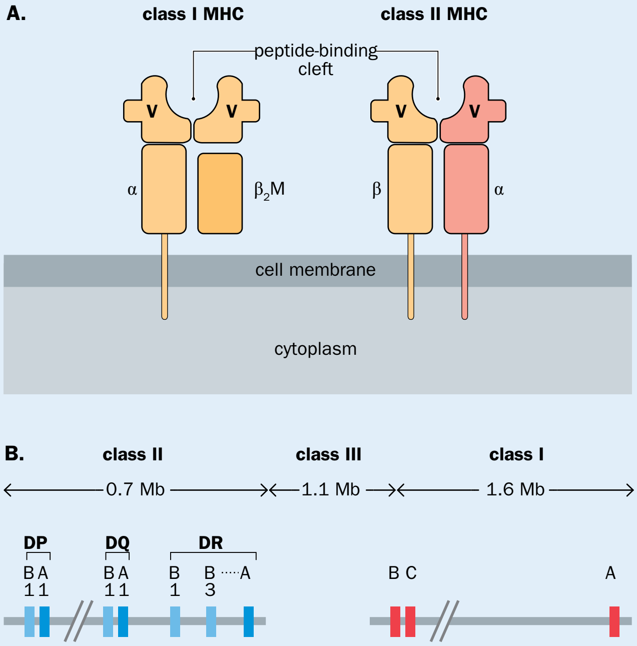
The major histocompatibility complex (MHC) is a gene cluster that contains classical MHC genes (which are extremely polymorphic and are the primary determinants of transplant rejection), nonclassical MHC genes (which make proteins that are highly related to classical MHC proteins but show limited polymorphism), and various other genes.

MHC genes make transmembrane protein components of heterodimeric cell surface proteins that belong to two classes. Class I MHC proteins are formed by association of a heavy chain, consisting of an MHC protein with three extracellular domains, and a nonpolymorphic, non-MHC light chain, β_2-microglobulin (which does not span the cell membrane). Class II MHC proteins are composed of two transmembrane proteins, each encoded by an MHC gene (Figure 1A).

Class Ia MHC genes, a subset of class I MHC genes, are considered to be classical MHC genes because they produce highly-polymorphic heavy chains. They are expressed by almost all nucleated cells of the body. The other classical MHC genes are polymorphic class II MHC genes that are expressed by certain immune system cells, notably dendritic cells, macrophages, and B cells. The basic organization of the human MHC, the **HLA complex** (originally, human leukocyte antigen complex), is shown in Figure 1B.

Because the classical class I MHC α genes and both the class II MHC α and β genes are highly polymorphic, individuals can be expected to be heterozygous at several MHC loci, and both alleles are co-expressed at the cell surface. As a result, transplanting an organ from a randomly selected donor to an unrelated recipient will mean that the grafted tissue will very likely have quite a different profile of MHC proteins to that of the cells of the recipient, and will be rejected. (As the transplanted cells carry antigens not present on host cells, the graft will be recognized to be foreign and attacked by the immune system.) To increase transplant success rates, *tissue typing* is carried out to identify alleles at MHC loci, so that potential donors can be sought with an MHC profile that closely matches the intended recipient.

The extreme polymorphism of MHC genes did not evolve to frustrate transplant surgeons! The job of MHC proteins is to help T cells to identify and kill potentially dangerous cells, and the extreme polymorphism of MHC loci is the result of selection pressure imposed by pathogens. Over long periods of human history when medical intervention was



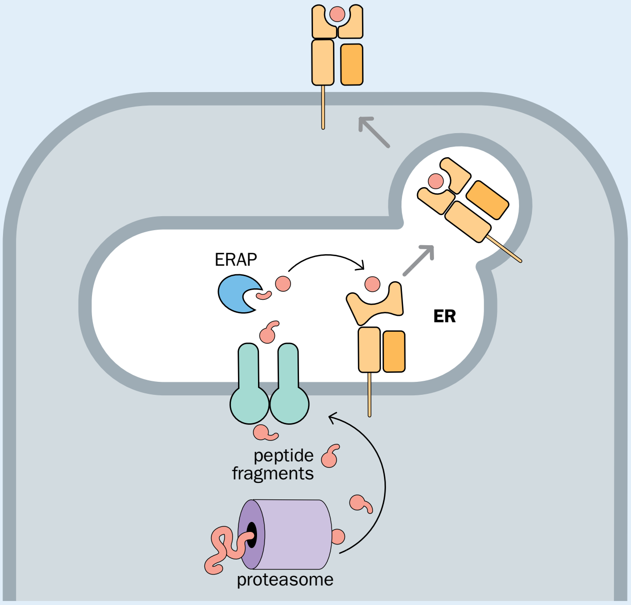
Box 3.2 Figure 1 MHC protein structure and organization of the HLA (human MHC) gene cluster. (**A**) Class I MHC proteins have a transmembrane α chain (encoded by an MHC gene) in association with β_2-microglobulin. The α chain has three extracellular domains: two that are distal to the membrane and can be highly variable, and a constant domain next to the membrane. Class II MHC proteins have two transmembrane proteins encoded by the MHC, each with two extracellular domains: one constant domain and one variable domain. The variable (V) domains are used to bind a peptide (see text). (**B**) Genes in the class II HLA region encode α chains (dark shading) and β chains (pale shading) that pair up to form heterodimers within specific classes (indicated by horizontal bars above, with gene names *HLA-DPB1*, *DPA1*, and so on). The class I HLA genes include the highly-polymorphic class Ia subfamily (*HLA-B*, *HLA-C*, and *HLA-A*) and class Ib genes (not shown), which have very limited polymorphism. Additional genes with an immune system function are found in the intervening class III region, notably several complement genes.

nonexistent or extremely limited, individuals heterozygous at multiple MHC loci had much higher survival rates than those with limited MHC heterozygosity (as having multiple MHC alleles extends the ability to recognize foreign antigens, providing increased protection against various infectious diseases associated with high mortality and morbidity).

FUNCTION OF MHC PROTEINS

At the cellular level, the function of classical MHC proteins is to bind and transport peptide antigens to the cell surface and present them to be recognized by T cells carrying αβ TCRs. However, the choice of a class I or class II MHC protein for peptide binding depends on the cellular origin of the proteins from which the peptides originated, as outlined below.

- Class I MHC proteins predominantly bind peptides derived from endogenous proteins (that is, proteins that have been synthesized and then degraded inside the same cell as the one that makes the MHC protein) and present them to killer T cells. As well as normal cellular proteins, endogenous proteins include abnormal tumor proteins and foreign proteins that have been synthesized within host cells as a result of infection (by viruses and other intracellular pathogens). All endogenous proteins are degraded within the cytosol using proteasomes, large complexes containing enzymes that cleave peptide bonds, converting proteins to peptides (**Figure 2**).
- Class II MHC proteins bind peptides predominantly derived from exogenous proteins (that were not synthesized in the same cell as the MHC proteins) and present them to helper T cells. An exogenous protein may have been synthesized in a cell that was then phagocytosed by a macrophage or a neutrophil, for example.



Box 3.2 Figure 2 MHC–peptide binding and antigen presentation. Class I MHC proteins (called class I HLA proteins in human cells) serve to bind peptides that predominantly originate from endogenous proteins and display them on the cell surface. The peptides are produced by the degradation within the proteasome (bottom) of *any* protein synthesized within that cell (either a host-cell protein or one made by a virus or other intracellular pathogen). Peptide fragments produced within the proteasome are transported into the endoplasmic reticulum (ER). Here they are snipped by an endoplasmic reticulum aminopeptidase (ERAP) to the proper size (~8–9 amino acids long) needed for loading on to a partly unfolded class I MHC protein. Once the peptide has been bound, the MHC protein completes its folding and is transported to the plasma membrane with the bound peptide displayed on the outside. In a similar process, class II MHC proteins bind somewhat larger peptides derived from exogenous proteins.

Antigen presentation

After a peptide has been bound by an MHC molecule and then transported with it to the cell surface, the MHC protein–peptide complex acts as a signal for a T cell with an αβ TCR that can specifically bind it. That is, an MHC protein is required to present a peptide on the cell surface so that it can be recognized by a T cell with the appropriate αβ TCR (**antigen presentation**). T-cell receptors on killer (cytotoxic) T cells recognize class I MHC–peptide signals, and those on helper T cells recognize class II MHC–peptide signals, but additional types of receptor–signal interactions are also required (**Figure 3.27A** and **B**).

Because class I MHC–peptide signals are expressed on almost all nucleated cells, almost any cell has the potential to present antigen to a killer T cell. However, class II MHC proteins are expressed on a very limited set of cells, notably dendritic cells, macrophages, and B cells. These cells are very active in presenting antigens to helper T cells and are often described as “professional” antigen-presenting cells.

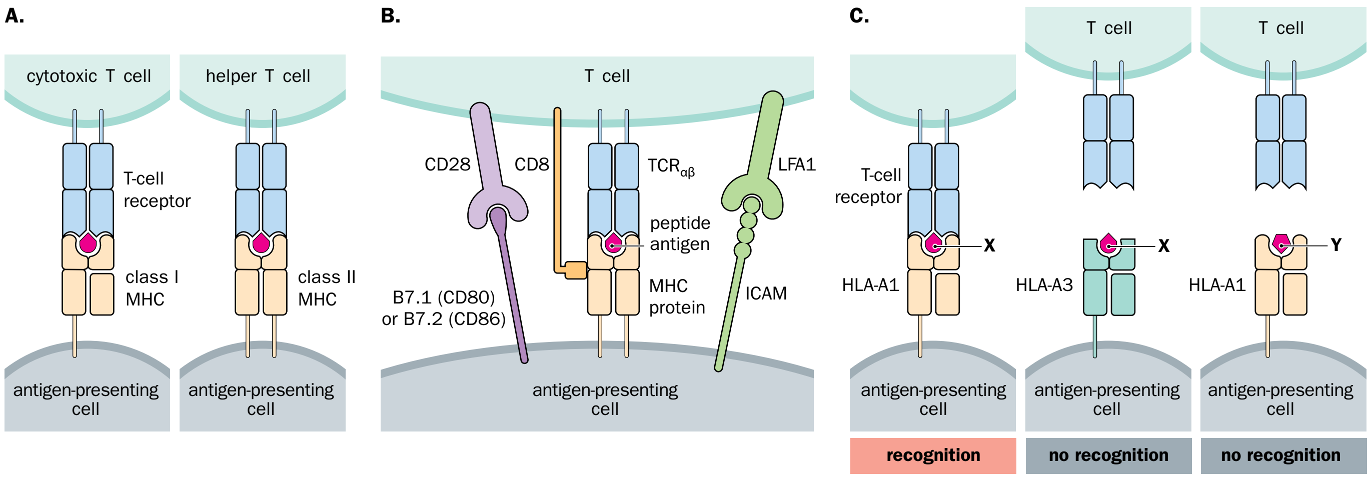


Figure 3.27 Antigen presentation. (**A**) Differential recognition of class I and class II MHC–peptide signals. For killer (cytotoxic) T cells, the T-cell receptor recognizes a class I MHC–peptide signal, but on helper T cells it recognizes a class II MHC protein–peptide signal. (**B**) Different types of receptors on the T cell are required to recognize signals on the antigen-presenting cell. Here we illustrate the example of a professional antigen-presenting cell presenting a class II MHC–peptide signal to be recognized by a T-cell receptor on a helper T cell. In addition, a CD8 protein on the helper T cell is required to bind a nonpolymorphic component of the class II MHC protein, and a co-stimulatory signal molecule, such as a member of the B7 family, needs to be recognized by CD28 receptors on the T-cell surface. Cell adhesion is promoted by, for example, using an LFA1 receptor on the T cell to recognize an ICAM signal on the presenting cell. (**C**) MHC restriction. T cells have cell-specific receptors that recognize a combination of a specific peptide and a specific MHC protein. In this case, we imagine a human T cell whose receptor is specific for a combination of peptide X and a class I MHC allele, HLA-A1. (Adapted from Murphy K & Weaver C [2016] *Janeway's Immunobiology*, 9th edn. Garland Science. With permission from WW Norton.)

From antigen presentation to T-cell activation

In addition to MHC–peptide signals, **co-stimulatory molecules** on the surface of an antigen-presenting cell interact with specific receptors on T cells (see Figure 3.27B). The co-stimulatory signal delivered to helper T cells is crucially important in inducing them to synthesize interleukin-2 (IL-2; a T-cell growth factor) and to express high-affinity receptors for IL-2. The secreted IL-2 stimulates proliferation of T cells expressing the CD4 or CD8 cell surface receptors. This complex way of inducing T-cell responses is presumably necessary to avoid inadvertent (and potentially damaging) activation of T cells, and to carefully regulate T-cell proliferation and differentiation.

Co-stimulatory molecules are so critically important for initiating and regulating immune responses that they are not constitutively expressed, even by professional antigen-presenting cells. Their expression can be induced via the TCR-triggered transient expression of CD40 ligand on T cells and the subsequent interaction with, and cross-linking of this ligand to, CD40 on professional antigen-presenting cells. Expression of co-stimulatory molecules is also triggered by signaling arising from the recognition of the conserved features of pathogens that is part of the innate immune system, most importantly the signaling through Toll-like receptors expressed on dendritic cells, macrophages, and certain other cells.

MHC restriction

The process whereby αβ T cells recognize protein antigens only after they have been degraded to form peptides that become associated with MHC molecules is described as **MHC restriction**. Note that it is the combination of a specific MHC protein and a specific peptide that an αβ T-cell receptor recognizes (Figure 3.27C). The MHC protein must be a self-MHC protein, one that is expressed naturally by host cells (which is why organ transplantation usually provokes strong immune responses in the recipient—MHC molecules on the transplanted tissue are treated as being foreign proteins).

Because all proteins in a cell routinely undergo degradation in proteasomes, MHC restriction allows T cells to survey a peptide library derived from the entire set of proteins contained within a presenting cell, but which is presented on the surface of the cell by MHC molecules. As a result, MHC restriction provides a remarkable evolutionary solution to the problem of how to detect intracellular pathogens. At the same time, it restricts T cells to recognizing only those antigens that are associated with host-cell MHC molecules and that are derived from intracellular spaces. T cells therefore complement B cells and antibodies, which can only recognize extracellular antigens and pathogens.

The need for self-tolerance

An important aspect of antigen presentation is that MHC proteins cannot distinguish self- from nonself-antigens, and therefore, except on the rare occasions when a cell does indeed become infected or captures a microbial protein, the MHC proteins on its surface are presenting peptides derived from the degradation of self-proteins. To avoid widespread attack of body cells by T cells, a **self-tolerance** mechanism is established during the development of αβ T cells in the thymus: only those T cells that have receptors that potentially recognize foreign peptides in association with self-MHC molecules are allowed to mature and are released into the periphery (positive selection). Those that recognize self-peptides are induced to commit suicide by apoptosis (negative selection).

SUMMARY

- Cells show extraordinary diversity in size, form, and function. Histology recognizes over 200 different cell types in adult humans but this is a massive underestimate—the number of different neurons alone is likely to be in the thousands.
- Many cell functions require that cells signal to each other over both short and long distances. Signals transmitted by one cell change the behavior of responding cells by altering the activity of proteins, notably transcription factors that bring about a change in gene expression.
- Transmitting cells often secrete a ligand molecule that binds to a receptor on the surface of responding cells to initiate a downstream signal-transduction pathway.
- Other small signaling molecules pass through the plasma membrane of responding cells to bind to intracellular receptors, or are anchored in the membrane of the transmitting cell and interact with receptors on the surfaces of adjacent cells.
- Cell proliferation is regulated at different stages in the cell cycle. During development, cell proliferation causes rapid growth of an organism. At maturity, cell proliferation is limited to certain cell types that need to be renewed and there is an equilibrium between cell birth and cell death.
- Programmed cell death is functionally important in development and throughout life. There are different pathways to get rid of cells that are unwanted, unnecessary, or potentially dangerous.
- Cells make connections with each other (cell adhesion); transient connections allow cells to perform various functions, while stable connections allow functionally similar cells to form tissues.
- Tissues are composed of cells of one or more types plus an extracellular matrix (ECM), a complex network of secreted macromolecules that supports cells and interacts with them to regulate many aspects of their behavior.
- Cell adhesion is regulated by a limited number of different types of cell adhesion molecule that link cells to neighboring cells or to the ECM. Neighboring cells also make contact through different types of cell junctions.
- Immune system cells are diverse and highly specialized. Those of the innate immune system work in nonspecific recognition of foreign or altered host molecules known as antigens.
- B and T lymphocytes are the core of the adaptive immune system that mounts strong, highly-specific immune responses.

FURTHER READING

Cell biology

Alberts B *et al*. (2014) *Molecular Biology of the Cell*, 6th edn. Garland Science.

Cell signaling

Brivanlou AH & Darnell JE (2002) Signal transduction and the control of gene expression. *Science* **295**:813–818; PMID 11823631.

Christensen ST & Ott CM (2007) Cell signaling: a ciliary signaling switch. *Science* **317**:330–331; PMID 17641189.

Database of Cell Signaling. Science STKE (Signal Transduction Knowledge Environment). http://stke.sciencemag.org/cm/ (Permits browsing of details of signal transduction pathways.)

Gavi S *et al*. (2006) G-protein-coupled receptors and tyrosine kinases: crossroads in cell signaling and regulation. *Trends Endocrinol Metab* **17**:48–54; PMID 16460957.

Cell proliferation and programmed cell death

Adams JM (2003) Ways of dying: multiple pathways to apoptosis. *Genes Dev* **17**:2481–2495; PMID 14561771.

Berry D (2007) Molecular animation of cell death mediated by the Fas pathway. *Science STKE* 2007(380):tr1. http://stke.sciencemag.org/content/2007/380/tr1

Elmore S (2007) Apoptosis: a review of programmed cell death. *Toxicol Pathol* **35**:495–516; PMID 17562483.

Morgan DO (2007) *The Cell Cycle: Principles of Control.* New Science Press Ltd.

Raff M (1998) Cell suicide for beginners. *Nature* **396**:119–122; PMID 9823889.

Cell adhesion and tissue formation

Beckerle M (ed.) (2002) *Cell Adhesion (Frontiers in Molecular Biology)*. Oxford University Press.

Extravasation Animation. http://multimedia.mcb.harvard.edu/media.html (This video shows how regulating cell adhesion permits white blood cells to migrate into tissues.)

Gumbiner BM (1996) Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell* **84**:345–357; PMID 8608588.

Heller E & Fuchs E (2015) Tissue patterning and cellular mechanics. *J Cell Biol* **211**:219–231; PMID 26504164.

Hynes RO (2009) The extracellular matrix: not just pretty fibrils. *Science* **326**:1216–1219; PMID 19965464.

Immune system biology

Blum JS *et al.* (2013) Pathways of antigen processing. *Annu Rev Immunol* **31**:443–473; PMID 23298205.

Brubaker SW *et al*. (2015) Innate immune pattern recognition: a cell biological perspective. *Annu Rev Immunol* **33**: 257–290; PMID 25581309.

Delves PJ *et al*. (2011) *Roitt's Essential Immunology*, 12th edn. Wiley-Blackwell.

Immune-Cell Lineage Commitment (2007) *Immunity* **26**:669–750. (*Immunity* special issue on immune system cell lineages.)

Parham P (2015) *The Immune System*, 4th edn. Garland Science.

Punt J *et al*. (2019) *Kuby Immunology*, 8th edn. WH Freeman and Company.

Aspects of early mammalian development, cell differentiation, and stem cells

4

We covered some aspects of cell–cell interactions in Chapter 3. Now we go on to consider aspects of early development, when signaling between cells is important in dictating key developmental processes.

In Section 4.1, we consider how the mammalian zygote gives rise to early lineages of more specialized cells, charting the initial steps of tissue differentiation and the formation of the three primary germ layers and germ cells against a background of mammalian embryonic development. As well as dictating how different cell types and tissues are formed, cell–cell signaling is required to position cells and organize them within a body plan for the organism and to mold the overall morphology of the organism.

In Section 4.2, we introduce various aspects of stem cells and cell differentiation. We consider what is known about natural stem cells in the body. Then we explain how artificial pluripotent stem cells are made, and cover the science of cell reprogramming. The ways in which stem cells and cell reprogramming can be applied toward treating disease will be covered in Chapter 22.

4.1 CELL LINEAGES AND TISSUE DIFFERENTIATION IN EARLY MAMMALIAN DEVELOPMENT

Mammalian development is unusual in that the embryo and fetus depend on the mother for a supply of nutrients. The fertilized mammalian egg cell gives rise to both embryonic cell lineages (which will form the organism) and extra-embryonic cell lineages. The latter will form extra-embryonic membranes and a component of the placenta, both of which are needed to support the developing embryo and fetus.

The principal aims of this section are to explain how a single fertilized mammalian egg cell, the zygote, gives rise to different embryonic and extra-embryonic cell lineages during early development, the molecular basis of this early tissue differentiation, and the development of the three somatic germ layers and of germ-line cells. For ethical and practical reasons, knowledge of human development has been limited; much of our understanding of mammalian development has been gleaned from animal models, principally the mouse.

The background to this section, early mammalian development, is a very broad field that we do not cover in detail; for interested readers, we recommend at the end of the chapter several excellent textbooks that provide comprehensive coverage of this topic. In Section 4.2, we follow up by considering stem cells, further aspects of cell differentiation, and cell reprogramming.

An overview of mammalian development

Traditionally, animal development has been divided into an embryonic stage (during which all the major organ systems are established) and a postembryonic stage (which in mammals consists predominantly of growth and refinement).

Once the basic body plan has been established, it is not clear when development stops. In humans, postembryonic growth begins at the start of the fetal period, in the fetus, at about 9 weeks after fertilization but continues for up to two decades after birth, with some organs reaching maturity before others. It can be argued that human development ceases when the individual becomes sexually mature, but many tissues and cells—intestinal epithelium, skin, blood, and sperm cells, for example—need to be actively replenished throughout life. In such cases, development never really stops at all; instead it reaches an equilibrium. Even aging, a natural part of the life cycle, can be regarded as a part of development.

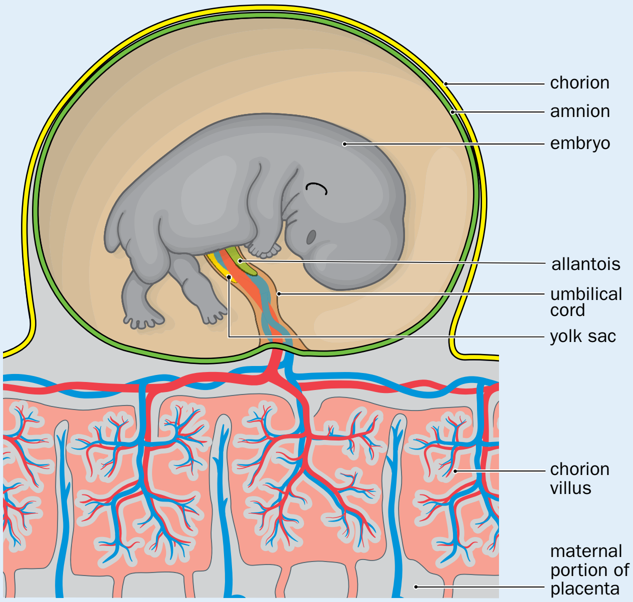
Development is a gradual process. The fertilized egg initially gives rise to a simple embryo. Different cell lineages are formed, becoming cells of the embryo proper or cells of the four extra-embryonic membranes and placenta (that support the developing embryo and fetus; see **Box 4.1**).

BOX 4.1 MAMMALIAN EXTRA-EMBRYONIC MEMBRANES AND THE PLACENTA

There are four **extra-embryonic membranes**: amnion, yolk sac, allantois, and chorion (**Figure 1**). After the embryo implants into the uterine wall, the chorion combines with maternal tissue to form the **placenta**. As well as protecting the embryo and fetus, these life support systems are required to provide for its nutrition, respiration, and excretion.

AMNION

The innermost of the extra-embryonic membranes, the amnion, remains attached to and immediately surrounds the embryo. It contains amniotic fluid that bathes the embryo, preventing it from drying out, helping it to float (reducing the effects of gravity), and protecting it from mechanical jolting (by acting as a hydraulic cushion).



Box 4.1 Figure 1 Human extra-embryonic membranes and placenta. (From Mader SS [2013] *Human Biology*, 13th edn. Copyright © 2013, with permission from McGraw-Hill Education.)

YOLK SAC

In many mammals, including humans and mice, the yolk sac does not contain any yolk. It is, however, the source of the first blood cells of the conceptus and most of the first blood vessels. The **primordial germ cells** (progenitor cells that ultimately give rise to spermatocytes and oocytes) pass through the yolk sac on their migration from the epiblast to the gonadal ridge.

ALLANTOIS

The allantois acts as a waste (urine) storage system in most amniotes but not in placental mammals. In some mammals, including humans, it is vestigial, lacking any function except that its blood vessels give rise to the umbilical cord vessels.

CHORION AND PLACENTA

The outermost extra-embryonic membrane, the chorion, has an outer layer of **trophoblast** cells and an inner layer of mesoderm. It serves as a source of hormones that influence the uterus, a surface for respiratory exchange, and as the fetal component of the **placenta** in placental mammals. The remainder of the placenta arises from the decidua basalis, a component of the maternal uterine wall (endometrium).

The placenta develops after the implanted blastocyst induces the neighboring maternal endometrium to become a nutrient-packed highly-vascular tissue, the decidua. Thereafter, the trophoblast tissue develops vacuoles that connect to nearby maternal capillaries, rapidly filling with blood. As the chorion forms, it projects outgrowths known as **chorionic villi** into the vacuoles, bringing the maternal and embryonic blood supplies into close contact. At the end of the third week of development, the chorion has differentiated fully and contains a vascular system connected to the embryo. Exchange of nutrients and waste products occurs over the chorionic villi.

As development proceeds, the number of cell types increases and the organization of these cell types becomes more intricate. Complexity is achieved progressively. Development incorporates several different molecular processes that affect cell behavior. The processes listed below are inter-related, occurring separately in different parts of the embryo or in combination.

- Cell proliferation: repeated cell division, leading to an increase in cell number. In the mature organism, this is balanced by regular cell loss.
- Growth: an increase in the overall size of the organism and of biomass.
- Differentiation: the process by which cells become specialized, both structurally and functionally.
- Pattern formation: the process by which cells become organized, initially to form the fundamental body plan of the organism. First, the three axes need to be specified in the early embryo: the anterior–posterior (head-to-tail) axis,

the dorso–ventral (back-to-front) axis, and the left–right axis. Subsequently, the detailed structures of different organs and tissues are formed.

- Morphogenesis: the changes in the overall shape or form of the developing organism. Underlying mechanisms include differential cell proliferation, selective cell–cell adhesion or cell–matrix adhesion, changes in cell shape and size, the selective use of programmed cell death, and control over the symmetry and plane of cell division.

Each of the processes above is controlled by regulatory genes specifying when and where in the embryo particular gene products that direct the behavior of individual cells will be synthesized. That is, these changes occur by changes in gene expression.

Epigenetic control of development and the need for symmetry-breaking

During development, a single unspecialized cell—the fertilized egg cell—gives rise to cells that are progressively more specialized (this process is known as **cell differentiation**). But if all nucleated cells in an individual originate from the zygote and have the same set of DNA molecules, how do uncommitted cells in the very early embryo give rise to increasingly specialized lineages of cells? That is, how does differentiation ever arise in the first place? The answer is that **epigenetic** factors are involved that do not involve changes to the DNA sequence.

Intrinsic and extrinsic asymmetry

Differentiation occurs at an early stage in embryonic development after some asymmetry arises. The cells may have the same DNA molecules but they may nevertheless be different because they have inherited different amounts of fate-determining protein factors (intrinsic asymmetry). Alternatively, they may be exposed to slightly different microenvironments and so the extracellular chemical signals received may differ between neighboring cells (extrinsic asymmetry).

In some nonmammalian organisms, intrinsic asymmetry develops in the egg: gradients of certain proteins that are important in early gene regulation can be established in the egg cell (which is a large, sometimes huge, cell). As a result, when the egg undergoes cleavage divisions without cell growth (to give progressively smaller descendent cells), cell division can be asymmetric: one daughter cell receives more of an asymmetrically distributed regulatory protein than the other daughter cell (Figure 4.1A). One might also consider the potential for asymmetry at fertilization: the sperm entry point (the point at which the sperm makes contact with the egg to inject its nuclear package) might, conceivably, define an axis.

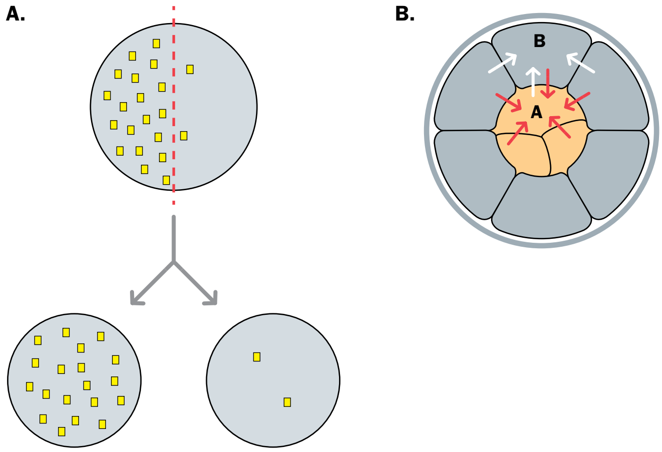


Figure 4.1 Examples of intrinsic and extrinsic asymmetry during early embryonic development. (**A**) Intrinsic asymmetry. In some organisms the egg is formed with asymmetric localization of certain fate determinants (which may be signaling proteins, specific transcription factors, and so on). In the zygote the asymmetric distribution of the fate determinants (shown here as yellow boxes) will be maintained (being concentrated toward one pole of the zygote). Following cell division (without growth) the daughter cells can have significantly different amounts of the fate determinants. (**B**) Extrinsic asymmetry. Inner cell A receives chemical signals from neighboring cells on all sides (red arrows). Outer cell B can receive chemical signals from neighboring cells next to it and from beneath it (white arrows) but not from above it.

Extrinsic asymmetry can arise at later stages. As the embryo grows, cells will have different locations. Simply on the basis of their positions in the embryo, cells may be exposed to different local environments, and might differ in the signal inputs they receive from neighboring cells. For example, cells on the outer surface of an embryo can receive signals from cells beneath them, but not above them, whereas cells in internal locations may receive signals on all sides from many different neighbors (Figure 4.1B). The choice of whether a cell is on the outside or inside of an embryo may largely be due to chance and so stochastic factors can be important as well. As explained below, cell position is known to be important in establishing different cell lineages in the early mammalian embryo.

In mammals, the first overt evidence for different cell lineages is apparent at very early stages of embryonic development. We consider this below when we describe the sequential stages of early development in mammals, from zygote to the implanted blastocyst. Figure 4.2 provides a road map of the major differentiation events in early mammalian development, and the changes from mammalian zygote to blastocyst are charted below.

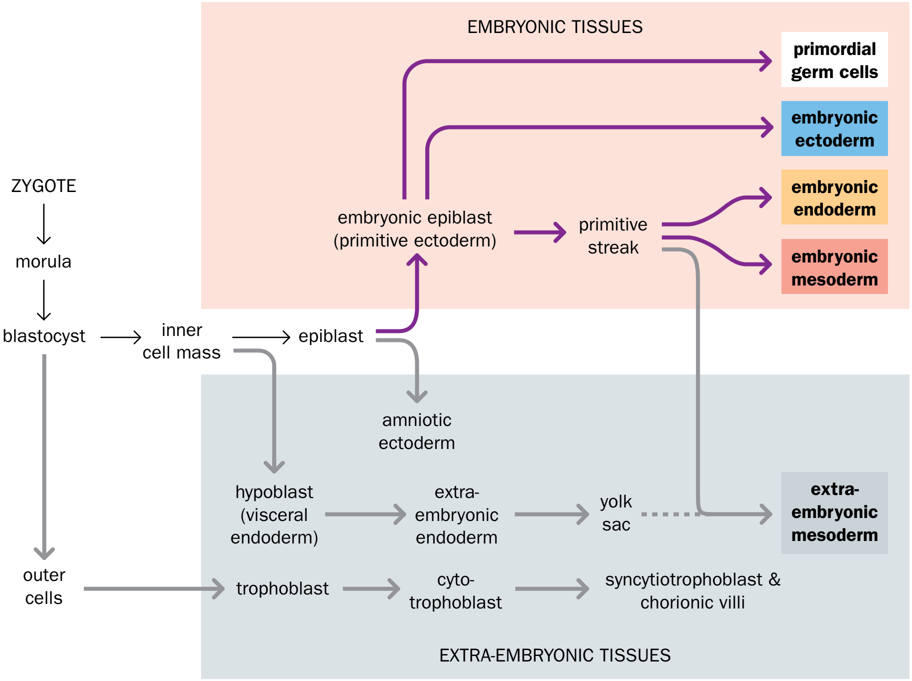


Figure 4.2 A road map for early differentiation events in human embryos. At, or shortly after, the late 8-cell stage, the embryo undergoes compaction and cell polarity develops (**Box 4.2**). Overt signs of resulting tissue differentiation first become apparent at the mammalian blastocyst stage, where there are two clearly different cell layers (as shown In **Figure 4.4B**). The outer trophoblast cells of the blastocyst (trophectoderm) give rise to cytotrophoblast that will form chorionic villi, and syncytiotrophoblast, which will ingress into uterine tissue. Tissues of the embryo will be formed exclusively by the epiblast cells of the inner cell mass that eventually give rise to three embryonic germ layers, ectoderm, endoderm, and mesoderm (see **Figure 4.9** for their derivatives). Some embryonic epiblast cells are induced by extra-embryonic ectoderm to form primordial germ cells, the precursors of germ cells, that will later migrate to the gonads. Other cells in the inner cell mass, those of the hypoblast and also some from the epiblast, give rise to other extra-embryonic membranes. The dashed line indicates a possible dual origin of the extra-embryonic mesoderm. (Adapted from Gilbert SF [2006] *Developmental Biology*, 8th edn. Fig. 11.32, p. 352. By permission of Oxford University Press.)

Fertilization: the beginning of a new life

Fertilization involves fusion of a unique haploid sperm cell and a unique haploid egg (oocyte) to create a diploid zygote and a new individual. The small sperm cell has a greatly reduced cytoplasm and a nucleus with highly-condensed, transcriptionally inactive chromatin (the normal histones are replaced by a special class of packaging proteins known as protamines). A long flagellum at the posterior end provides propulsion (Figure 4.3A).

The mammalian egg, although small compared to that of many other vertebrates, is still a large cell. During growth in the mouse ovary, for example, the volume of the developing oocyte increases by a factor of around 500. The egg provides the material necessary for the beginning of growth and development (the sperm loses most of its non-nuclear organelles during spermatogenesis, and so the zygote genome is dependent on maternal factors to activate it).

The cytoplasm of the egg is extremely well endowed. It has very large numbers of mitochondria and ribosomes, and large amounts of protein, including DNA and RNA polymerases. There are also very considerable amounts of RNAs, protective chemicals, and morphogenetic factors (but yolk is not required in mammalian eggs as the embryo will be nourished by the placental blood supply). Outside the plasma membrane of the mammalian egg is a proteinaceous extracellular matrix, the zona pellucida, surrounded by a thick extracellular matrix containing cumulus cells that nurture the egg prior to, and just after, ovulation (Figure 4.3B).

Human sperm cells have to migrate very considerable distances. Out of the 280 million or so ejaculated into the vagina, only about 200 reach the required part of the oviduct where fertilization takes place. Fertilization begins with attachment of a sperm to the zona pellucida, followed by release of enzymes from the acrosomal vesicle, causing local digestion of the zona pellucida. The sperm head then fuses with the plasma membrane of the oocyte and the sperm nucleus passes into the cytoplasm. Within the oocyte, the haploid sets of sperm and egg chromosomes are initially separated from each other; these male and female pronuclei (Figure 4.3C) subsequently fuse to form a diploid nucleus.

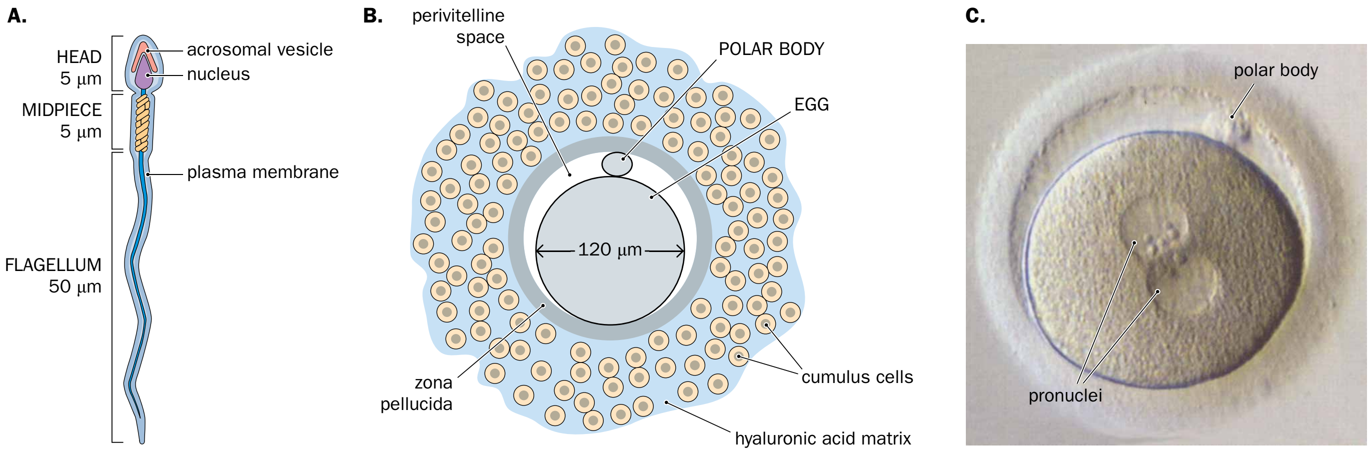


Figure 4.3 Mammalian egg and sperm cells. (**A**) A human sperm cell has a 5 µm head containing highly-compacted DNA, a 5 µm cylindrical body (the midpiece, which contains many mitochondria), and a 50 µm tail (flagellum). At the front, the acrosome contains enzymes that help the sperm to make a hole in the zona pellucida, allowing it to access and fertilize the egg. (**B**) The mammalian egg is a large cell, 120 µm in diameter. The surrounding zona pellucida contains three or four glycoproteins that polymerize to form a gel, and is enclosed by an extracellular matrix that is made up mostly of hyaluronic acid and contains the supporting cumulus cells. The first polar body (the product of meiosis I) lies under the zona pellucida within the perivitelline space. At ovulation, oocytes are in metaphase II. Meiosis II is not completed until after fertilization. (**C**) The fertilized oocyte (zygote) initially shows two *pronuclei*: a male pronucleus from the sperm and a female pronucleus from the egg cell. Fertilization triggers secretion of cortical granules by the egg that effectively inhibit further sperm from passing through the zona pellucida. (B, from Okabe M [2013] *Development* **140**:4471–4479; PMID 24194470. Adapted with permission from The Company of Biologists Ltd.; C, from Veek L & Zaninovic N [eds.] [2003] *An Atlas of Human Blastocysts*. With permission from CRC Press/Taylor & Francis.)

From the human zygote to the blastocyst stage and the development of three cell lineages

The journey from zygote to blastocyst involves a progressive increase in cell number with increasing cell specialization (**Figure 4.4**); the physiological context for that journey is shown in **Figure 4.5**.

Cleavage is the developmental stage where the zygote undergoes several cell divisions to form a number of progressively smaller cells, called **blastomeres**, that will initially form a solid ball of cells (**Figure 4.4A**). The cell divisions have a predominant S phase, but G phases are short and there is no net growth. (Note that the mammalian embryo is unusual in several ways. First, blastomeres do not all divide at the same time and the increase in cell number is not exponential—from 2 to 4 to 8 cells, and so on. The embryos frequently contain odd numbers of cells, and the blastomeres of one embryo can be of different sizes.)

Maternal factors in the egg cytoplasm are essential for initiating development. However, mammalian cleavage is exceptional in that the zygotic genome is activated early on, as early as the two-cell stage in the mouse, and at this stage maternal RNA transcripts get degraded. As a result, the cleavage divisions are largely controlled by the zygotic genome rather than by maternally-inherited gene products (which happens in many other animals).

Right up to the eight-cell stage, the blastomeres of a mammalian embryo are **totipotent**, like the zygote. That is, they can give rise to not just all possible cells of the organism but also to the cells of the supporting extra-embryonic membranes. As explained below, in very early mammalian embryos, cell fate has not yet been determined.

Compaction and developing cell polarity

When it forms, the eight-cell mammalian embryo appears to have symmetrical, roughly spherical cells that are loosely packed. Uniformly distributed over the cell surface are microvilli, small hairlike projections composed of complex plasma membrane folds surrounding an actin microfilament core. Shortly after reaching the eight cell stage, the embryo undergoes a key transformation known as **compaction** that is driven by cell–cell interactions. The hallmark of this process is a change in the shape of the embryo and of individual cells, but the most significant aspect is that individual cells become obviously asymmetric (see **Box 4.2**); the foundation is built for the first overt specification of two different cell lineages.

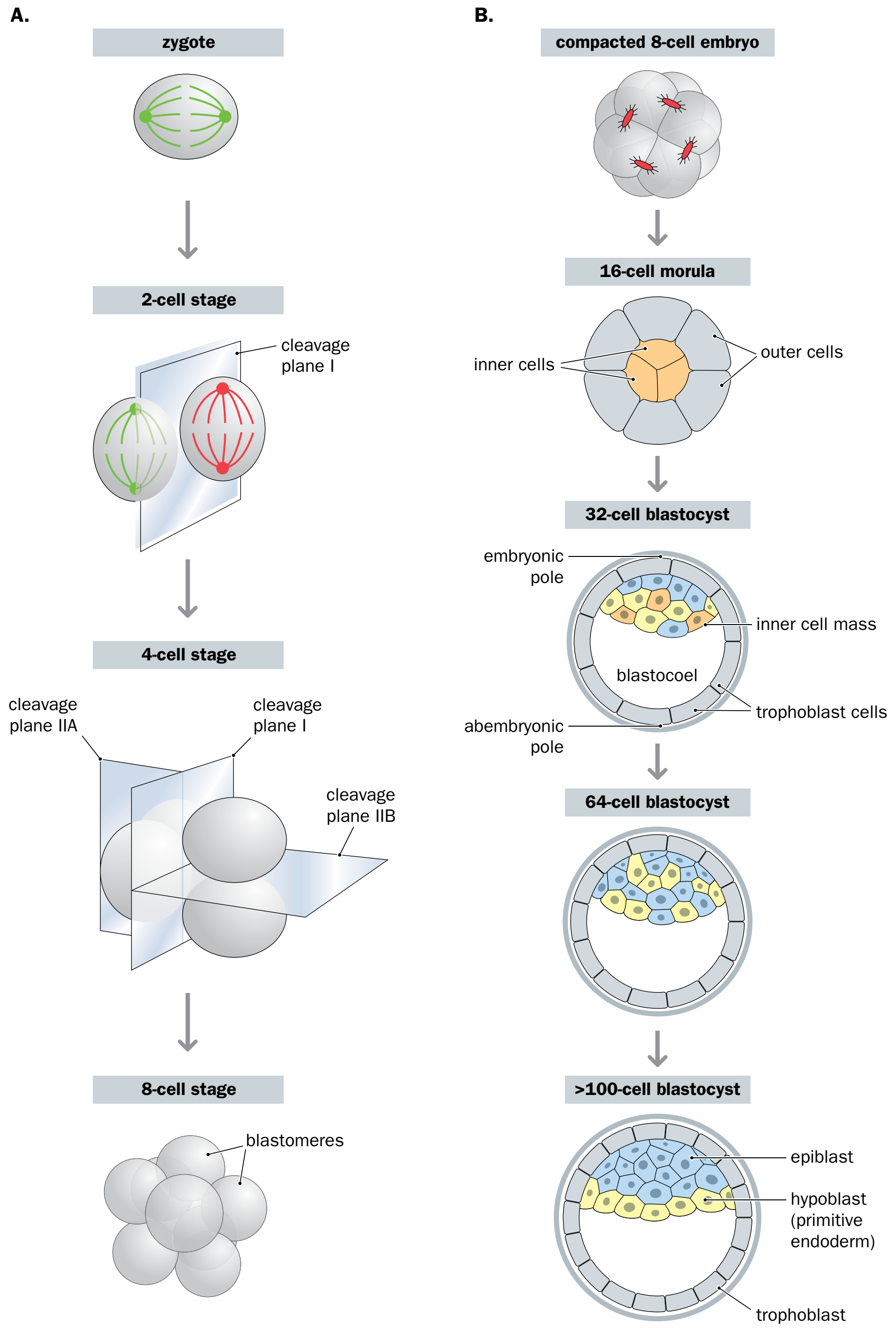


Figure 4.4 Early development of the mammalian embryo, from zygote to blastocyst. (**A**) Cleavage divisions. In mammals, the first cleavage is a normal meridional division (along the vertical axis), but in the second cleavage one of the two cells (blastomeres) divides meridionally while the other divides at right angles, equatorially (rotational cleavage). At the early eight-cell stage, the blastomeres are symmetrical and loosely packed with little contact between the cells. (**B**) From compacted embryo to blastocyst. Later, at about the eight-cell stage, the embryo undergoes compaction: the blastomeres flatten against each other to maximize cell–cell contacts, and tight junctions begin to be formed between the cells (see **Box 4.2**). By the 16-cell morula stage there are two populations of cells. Outer cells are held together by tight junctions, sealing off the inside of the sphere, which has inner cells that have formed gap junctions, enabling small molecules and ions to pass between them. The morula does not have an internal cavity, but afterward fluid is secreted from the outer cells into the interior, producing a blastocyst—a hollow ball of cells with a fluid-filled internal cavity, the blastocoel. The blastocyst has an outer layer of cells, the trophoblast that will contribute to the chorion, plus an inner cell mass (ICM) located at one end of the embryo, the embryonic pole. Differentiation of ICM cells (buff-colored cells) continues until by the mid-blastocyst stage (64-cell stage) the ICM has fully differentiated into two cell populations that initially appear in a mosaic form (blue and yellow cells). By the later blastocyst stage (>100 cells) the two ICM cell populations have sorted out into two distinct layers: an outer epiblast and an inner hypoblast. The epiblast will give rise to all the cells of the organism plus components of the extra-embryonic membranes; the hypoblast gives rise to the yolk sac. For convenience, the zona pellucida surrounding the early embryo (see **Figure 4.3B**) is not shown in part (**A**), and polar bodies are also omitted.

Symmetry is broken again at about the 12-cell stage, the time when the first cell leaves the surface of the embryo and moves to the interior. By the 16-cell **morula** stage (when the embryo resembles a mulberry, the Latin word for which is “morula”) there are two cell populations: outer polarized cells (which show apicobasal polarity) and inner apolar (symmetrical) cells (see **Figure 4.4B**). As the population of apolar cells increases, the cells begin to communicate with each other through gap junctions. This is the first overt demonstration of two different cell types; the distinction between them is a fundamental one.

Thereafter, at around the 32-cell stage, cavitation occurs as the embryo develops a cavity that becomes filled with fluid. Outer cells actively pump Na^+ ions into the extracellular space and, in response to the osmotic gradient, water diffuses out of the cells and accumulates in the interior. The morula is transformed into a hollow ball of cells, a **blastocyst** (the fluid-filled central cavity is known as the blastocoel and forms at about

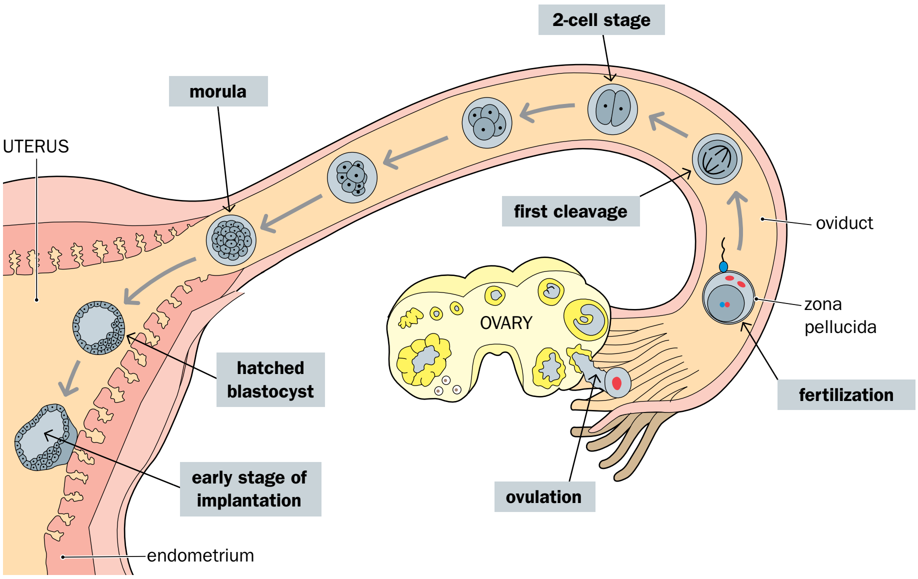


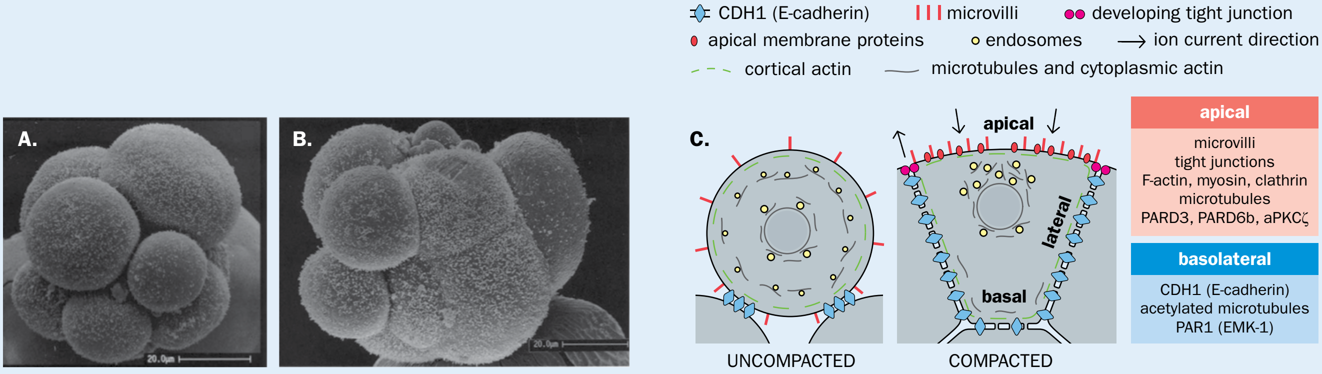
Figure 4.5 The physiological context of early embryonic development. Sperm deposited in the seminal fluid swim up into the uterus and then into the oviducts (the Fallopian tubes). During ovulation an egg is released from an ovary into the oviduct where it may be fertilized by a sperm. The fertilized egg is slowly propelled along the oviduct by cilia on the inner lining of the oviduct. During its journey, the zygote goes through various cleavage divisions but the zona pellucida (see **Figure 4.3B**) usually prevents it from adhering to the oviduct walls (this occasionally happens in humans, causing a dangerous *ectopic pregnancy*). Once in the uterus, the zona pellucida is partially degraded, releasing the blastocyst so that it can embed into the wall of the uterus (see **Figure 4.6**).

BOX 4.2 COMPACTION OF MAMMALIAN EMBRYOS: EARLY CELL POLARITY AND A FOUNDATION FOR DEVELOPING TWO CELL LINEAGES

At the early eight-cell stage, mammalian embryos are symmetrical and the individual blastomeres appear morphologically identical. The cells are loosely packed with quite small areas of contact between the cells. Shortly afterward, at about the eight- or ten-cell stage, compaction occurs. The first morphogenetic process in mammalian development involves a change in the shape of both the embryo and of its constituent blastomeres. The cells flatten their contacts with each other to maximize cell contact, and the embryo now becomes a tightly packed ball of cells (Figure 1A and B). Compaction is dependent on expression of E-cadherin (CDH1) but seems to be primarily driven by the ***cell cortex*** (or actomyosin cortex, the actin-rich layer on the inner face of the plasma membrane that is responsible for cell surface movements). It gives rise to pulsed contractions starting at the eight-cell stage in the mouse embryo (see Maître JL *et al.* [2015]; PMID 26075357).

Compaction is essential for the mammalian embryo to develop, being required for the first type of epithelial organization that will shortly manifest in the single-cell trophoblast layer. Epithelial cells are polarized, with an apical surface that has microvilli not present elsewhere on the cell surface, and compaction is the stage when cell–cell adhesion and **cell polarity** first become evident. The previously symmetrical blastomeres, with microvilli uniformly distributed on their cell surfaces, become asymmetric and, like in epithelial cells, microvilli become restricted to the apical portion of the cell surface, the outward-facing part not attached to other cells.

For each of the blastomeres, apicobasal polarity is also evident in the distribution of many plasma membrane

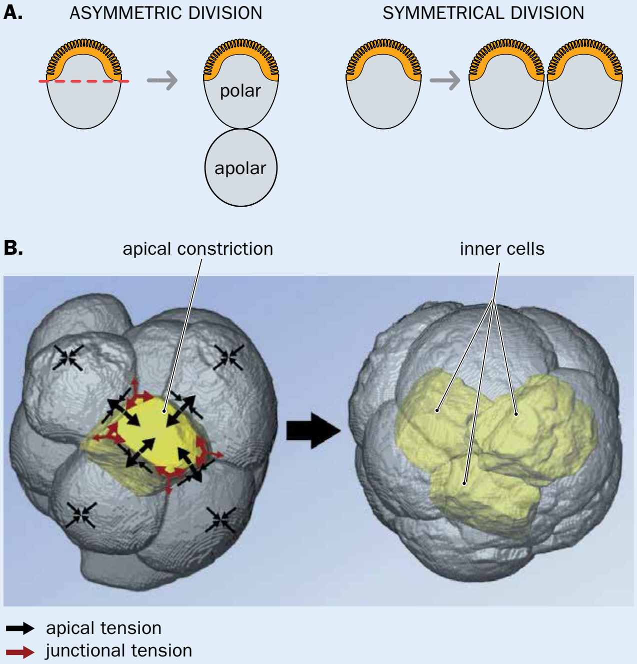


Box 4.2 Figure 1 Compaction of the mammalian embryo and developing cell polarity. (**A**) A 10-cell human embryo. Note the unequal size of blastomeres and presence of pronounced intercellular clefts. (**B**) A compacted 10-day human embryo. Note complete absence of clefts between several blastomeres. Not all of the blastomeres are compacted to the same degree, and they show differences in size and microvillus density. (**C**). During the transition from uncompacted embryo to compacted embryo the cells develop apicobasal polarity. Their apical (outer) surfaces retain microvilli and are enriched in membrane glycoproteins compared to the remaining basolateral part of the cell surface, while the underlying apical cytoplasm is enriched in various proteins and cytoskeleton components. (A and B, from Nikas G *et al.* [1996] *Biol Reprod* **55**:32–37; PMID 8793055. With permission from Oxford University Press; C, adapted from Fleming TP (ed.) [1992] *Epithelial Organization and Development*. With permission from Springer International Publishing AG. Copyright © 1992.)

proteins and components of the cell cortex, including polarity regulators such as PARD3, PARD6b, and aPKCζ and some cytoskeleton components (**Figure 1C**). Tight junctions begin to be formed between the cells and their neighbors, and the cytoskeletal elements are re-organized to form an apical band.

After the polar cells of the compacted embryo divide, the embryo has outer cells with apicobasal polarity (giving rise to trophectoderm) and inner apolar cells that begin to form the inner cell mass. This can happen in two ways. Occasionally, a polar cell undergoes asymmetric cell division where the cleavage plane is parallel to the surface and so at right angles to the apicobasal axis. One daughter cell is a polar cell that remains on the outside of the embryo; because of the angle of cleavage, the other daughter cell is now positioned in the interior of the embryo surrounded by cells on all sides and is apolar (**Figure 2A**).

In the mouse embryo, however, it is more common for early interior cells to be originally located on the surface and then be propelled into the interior by cortical tension: the apical surface of a cell undergoes shrinkage driven by cortical actomyosin contraction (**Figure 2B**).



Box 4.2 Figure 2 Formation of inner cell mass (ICM) cells by asymmetric division or cortical tension. (**A**) Polarized cells at the surface of the recently compacted embryo can undergo asymmetric cell division by cleavage parallel to the outer surface (red dashes). One daughter cell remains on the surface of the embryo and will retain the key fate determinants in the apical domain and remain polarized. The other daughter cell is born in an internal position and will contribute to the ICM; it is apolar because it lacks the apical domain. If cleavage occurs along the apicobasal axis, however, symmetrical cell division results in two identical polar outer cells, each with an apical domain. (**B**) An alternative origin for an early ICM cell is a cell originally located on the surface of the embryo but propelled into the interior by shrinkage of the apical surface (driven by actomyosin contractile forces). (B, from Samarage CR *et al*. [2015] *Dev Cell* **34**:435–447; PMID 26279486. With permission from Elsevier.)

the 32-day stage in humans). The inner cells within the blastocyst congregate at one end, the embryonic pole, to form an off-center **inner cell mass** (ICM). This asymmetry defines the first overt axis in the mammalian embryo: the embryonic–abembryonic axis. The ICM defines the embryonic pole and the distal end defines the abembryonic pole (see **Figure 4.4B**).

The early blastocyst has two distinct lineages of cells and, in both cases, the cells have lost some differentiation potential. The outer cells of the blastocyst, known as the **trophoblast** (or **trophectoderm**), are now quite specialized: they will give rise to the outer layer of the chorion, the outermost extra-embryonic membrane. The cells of the ICM will give rise to all the cells of the mammalian fetus plus the other three extra-embryonic membranes. The ICM cells have traditionally been considered to be **pluripotent**: they can give rise to all of the cells of the embryo but, unlike totipotent cells, they do not normally give rise to extra-embryonic structures derived from trophoblast.

The division into two cell lineages in the early blastocyst is a transient one. Before implantation, the ICM begins to differentiate into two distinct tissue types: the outer epiblast (= primitive ectoderm) and the inner hypoblast (= primitive endoderm or visceral endoderm). Initially, the two types of ICM cells are mixed in a mosaic pattern but then they sort themselves into two separate layers; see **Figure 4.4B**). The epiblast gives rise to all cells of the embryo proper plus some extra-embryonic membrane components, but the hypoblast gives rise to extra-embryonic cell lineages only, as follows:

- Epiblast: embryonic ectoderm, mesoderm, and endoderm; primordial germ cells; amniotic ectoderm; extra-embryonic mesoderm;
- Hypoblast: extra-embryonic endoderm lining the primary yolk sac and the blastocoel.

Implantation

Just prior to implantation, at around day 5 of human development, a protease is released that bores a hole through the zona pellucida, the proteinaceous extracellular matrix surrounding the egg and early embryo, and the blastocyst hatches (see Figure 4.5). The blastocyst is now free to interact directly with the lining of the uterus, the endometrium. Very soon after arriving in the uterus (day 6 of human development), the blastocyst attaches tightly to the epithelium of the uterine wall (**implantation**).

Trophoblast cells proliferate rapidly and differentiate into an inner layer of cytotrophoblast and an outer, multinucleated cell layer, the syncytiotrophoblast, that starts to invade the connective tissue of the uterus. The syncytiotrophoblast provides protection against maternal immune system cells; it effectively acts as a single giant cell (see Figure 4.6), leaving no gaps for maternal immune cells to migrate through (if they were to do so, an immune reaction would result when they reached the fetal side of the placenta and encountered paternal antigens).

A fluid-filled cavity, the amniotic cavity, forms within the inner cell mass, enclosed by the amnion. The embryo now consists of distinct epiblast and hypoblast layers and is known as the bilaminar germ disk. It is located between two fluid-filled cavities, the amniotic cavity on one side and the yolk sac on the other (see Figure 4.6B).

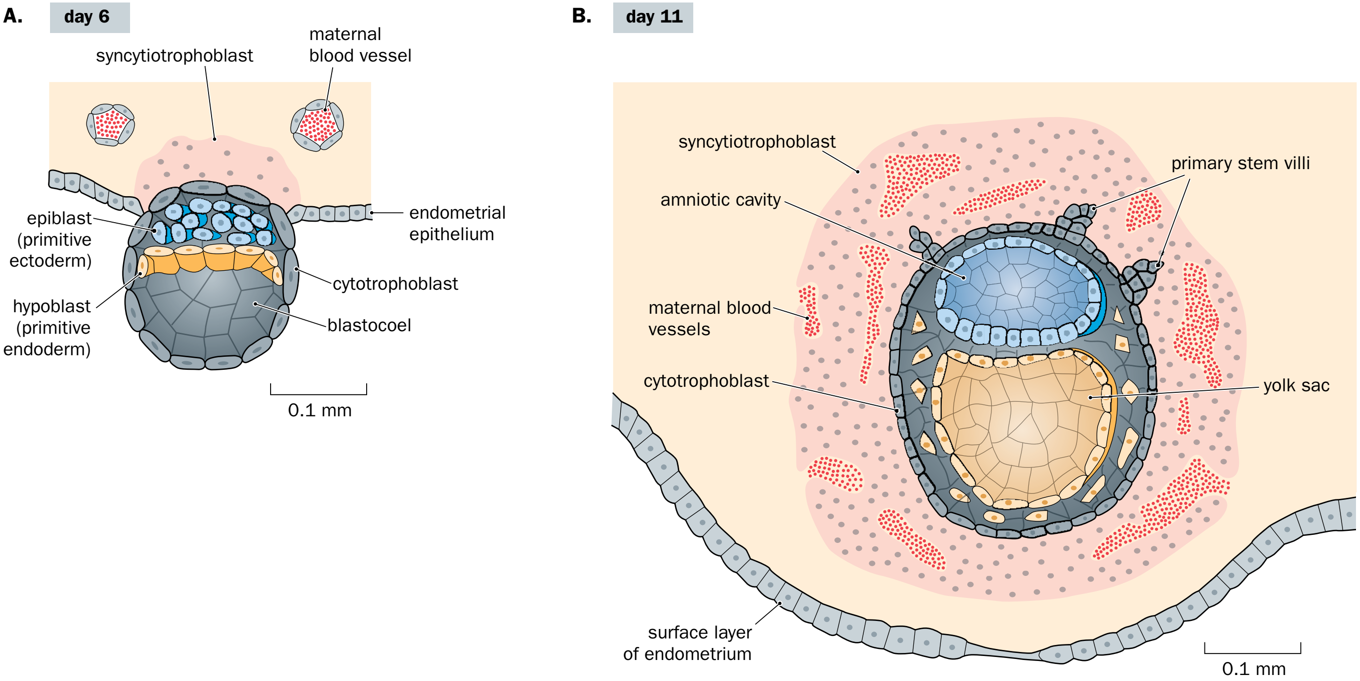


Figure 4.6 Detail of human embryo implantation. (**A**) Human embryo implantation begins at about 6 days after fertilization, when the hatched blastocyst adheres to the wall of the uterus (endometrium). Trophoblast cells differentiate into an inner layer of *cytotrophoblast* and an outer, multinucleated cell layer, the *syncytiotrophoblast*, that invades the connective tissue of the uterus. The inner cell mass of the blastocyst has given rise to two distinct cell layers: the epiblast and the hypoblast. (**B**) By about 11 days after fertilization, the primary yolk sac detaches from the surrounding cytotrophoblast. Loose endodermal cells are scattered round the yolk sac. Villi composed of cytotrophoblast cells begin to extend into the syncytiotrophoblast. (Reproduced with permission from McLachlan JC [1994] *Medical Embryology*. Addison Wesley.)

Gastrulation and the formation of the three somatic germ layers

Gastrulation, the first major morphogenetic process in development, takes place during the third week of human development. In this process, the orientation of the body is laid down, and the bilaminar germ disk is converted into a trilaminar disk with three fundamental germ layers: ectoderm, endoderm, and mesoderm. (Germ-line cells are separated out at an early stage, as described later.)

The primitive streak, an early but transient marker of the anterior–posterior axis, is first evident at about day 15 in human development, and marks a period when gastrulation has just begun. It appears as a faint linear groove along the longitudinal midline of the dorsal surface of the now oval-shaped bilaminar germ disk (Figure 4.7). Over the course of the next day it deepens and elongates to occupy about half the length of the embryo. By day 16, a deep depression surrounded by a slight mound of epiblast (the primitive node) is evident at the end of the groove, near the center of the germ disk.

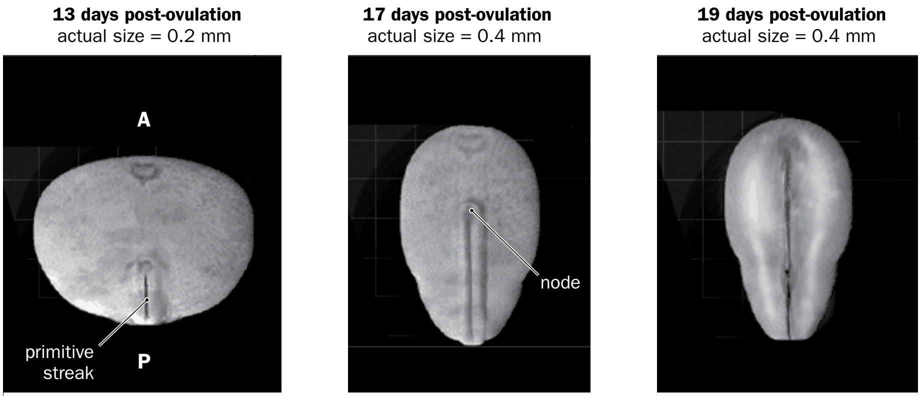


Figure 4.7 Progression of the primitive streak in human embryos. The view is of the dorsal surface of the illustrated embryos. At 13 days post-ovulation, a narrow line of cells appears on the dorsal surface of the formerly two-layered embryonic disk. This *primitive streak* marks bilateral symmetry in the embryo and indicates gastrulation has begun as cells migrate from the outer edges of the disk into the primitive streak and downward to create a new layer (as illustrated in **Figure 4.8**). At 17 days, gastrulation is continuing with the formation of a new cell layer, the ectoderm, changing the bilaminar embryonic disk into a trilaminar disk. By 19 days, the ectoderm has thickened to form the neural plate. The edges of the plate rise and form a concave area, the neural groove, which is the precursor of the nervous system. (Images reproduced from The Visible Embryo at http://www.visembryo.com.) A, anterior end; P, posterior end.

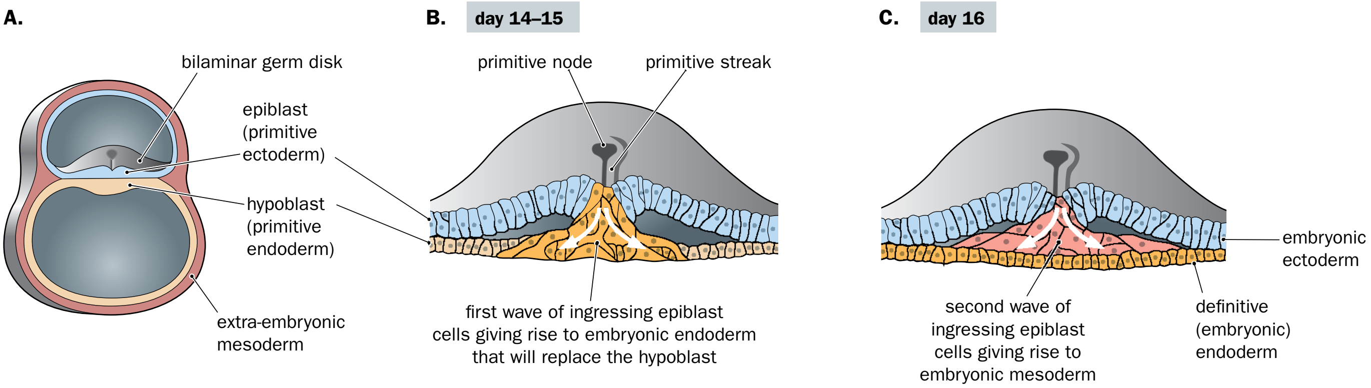

A.
bilaminar germ disk
epiblast (primitive ectoderm)
hypoblast (primitive endoderm)
extra-embryonic mesoderm
B.
day 14–15
primitive node
primitive streak
first wave of ingressing epiblast cells giving rise to embryonic endoderm that will replace the hypoblast
C.
day 16
embryonic ectoderm
second wave of ingressing epiblast cells giving rise to embryonic mesoderm
definitive (embryonic) endoderm

Figure 4.8 During human gastrulation, a flat bilaminar germ disk transforms into a trilaminar embryo. The outcome of gastrulation is similar in all mammals, but major differences may occur in the details of morphogenesis, particularly in how extra-embryonic structures are formed and used. (**A**) In humans, the epiblast and hypoblast come into contact to form a flat bilaminar germ disk. The epiblast cells within this disk are described as the primitive ectoderm but will give rise to all three germ layers—ectoderm, endoderm, and mesoderm—as described in panels (**B**) and (**C**). The epiblast cells that are not in contact with the hypoblast will give rise to the ectoderm of the amnion. The hypoblast will give rise to the extra-embryonic endoderm that lines the yolk sac. (**B**) The bilaminar germ disk at 14–15 days of human development. Along the length of the primitive streak, epiblast cells migrate downward to invade the hypoblast, and in so doing they become converted to embryonic endoderm and displace the cells of the hypoblast. (**C**) The bilaminar germ disk at 16 days of human development. A second wave of ingressing epiblast cells diverges into the space between the epiblast and the newly formed embryonic endoderm to form embryonic mesoderm. The remaining epiblast cells are now known as the embryonic ectoderm. (From Schoenwolf GC *et al.* [2014] *Larsen's Human Embryology*, 5th edn. Churchill Livingstone. With permission from Elsevier.)

The process of gastrulation is extremely dynamic, involving very rapid cell movements (**Figure 4.8**). At day 14–15 in human development, the epiblast cells near the primitive streak begin to proliferate, flatten, and lose their connections with one another. These flattened cells develop pseudopodia that allow them to migrate through the primitive streak into the space between the epiblast and the hypoblast (**Figure 4.8B**). Some of the ingressing epiblast cells invade the hypoblast and displace its cells, leading eventually to complete replacement of the hypoblast by a new layer of cells, the definitive endoderm. Starting on day 16, some of the migrating epiblast cells diverge into the space between the epiblast and the nascent definitive endoderm to form a third layer, the intra-embryonic mesoderm (**Figure 4.8C**). When the intra-embryonic mesoderm and definitive endoderm have formed, the residual epiblast is now described as the ectoderm and the new three-layered structure is referred to as the trilaminar germ disk (see **Figure 4.8C**).

At any time up until the late blastocyst stage, the pluripotency of the embryonic cells is demonstrated by the ability of the embryo to form twins, as described in **Box 4.3**. But by the time the three germ layers are formed, the constituent cells are already fairly restricted in their differentiation potential. They can give rise to just a few different cell

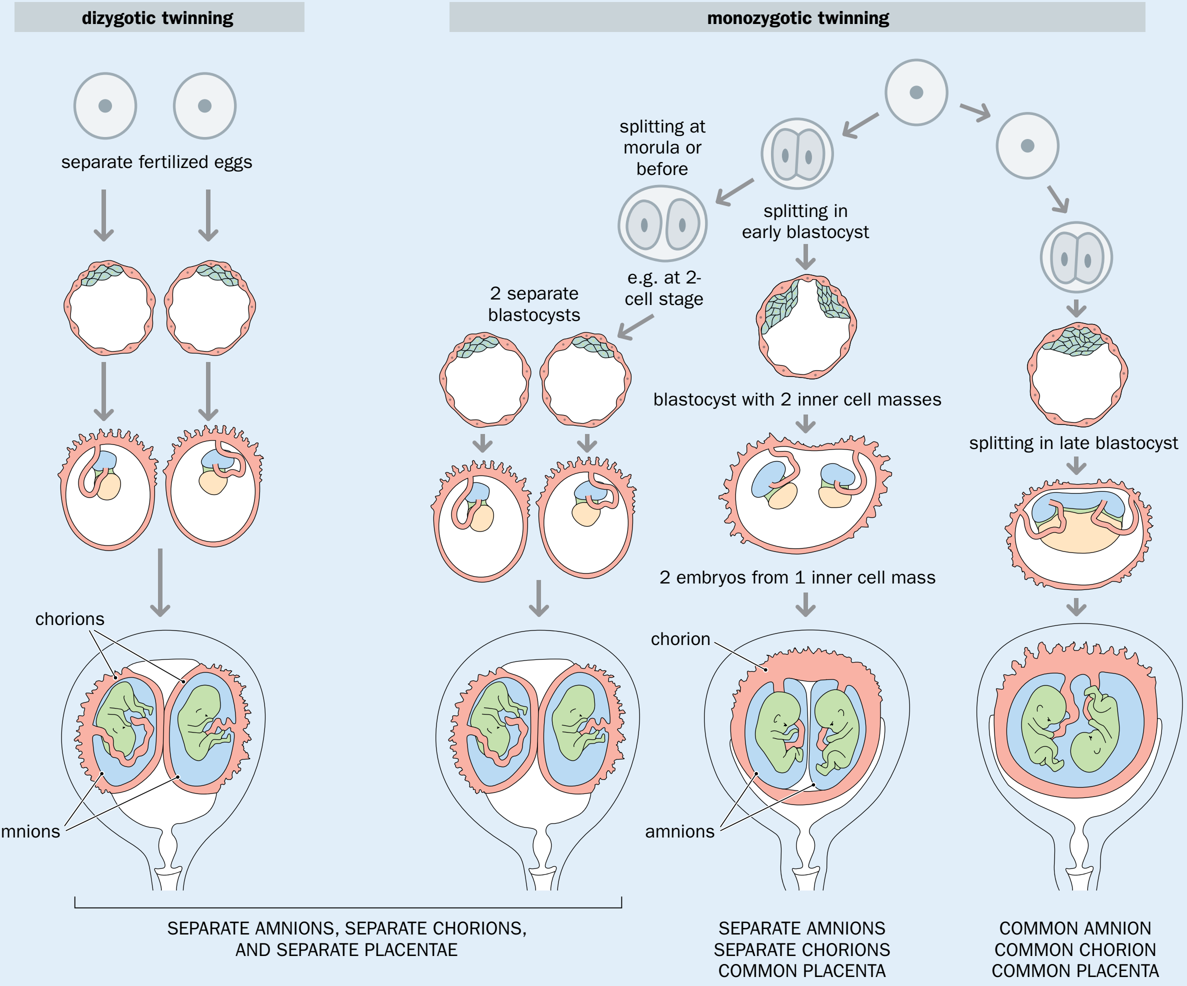
BOX 4.3 HUMAN TWINNING

Approximately one in every 200 human pregnancies gives rise to twins. There are two distinct types, as illustrated in Figure 1.

Fraternal (**dizygotic**) twins result from the independent fertilization of two eggs and are no more closely related than any other siblings. Although developing in the same womb, the embryos have separate and independent sets of extra-embryonic membranes.

Identical (**monozygotic**) twins arise from the same fertilization event, and are produced by the division of the embryo while the cells are still totipotent or pluripotent. About one-third of monozygotic twins are produced by an early division of the embryo, occurring during or prior to the morula stage. As a result, two separate blastocysts are formed, giving rise to embryos shrouded by independent sets of extra-embryonic membranes. In the remaining two-thirds of cases, twinning occurs at the blastocyst stage, and involves division of the inner cell mass.

The nature of monozygotic twinning reflects the exact stage at which the division occurs and how complete the division is. In most cases, the division occurs before day 9 of gestation, which is when the amnion is formed. Such twins share a common chorionic cavity, but are surrounded by individual amnions. In a very small proportion of births, the division occurs after day 9 and the developing embryos are enclosed within a common amnion. Either through incomplete separation or subsequent fusion, these twins are occasionally conjoined.



Box 4.3 Figure 1 **Human twinning.** (Adapted from Schoenwolf GC *et al.* [2014] *Larsen's Human Embryology*, 5th edn. Churchill Livingstone. With permission from Elsevier.)

types and are said to be multipotent. The ectoderm cells of the embryo, for example, give rise to epidermis, neural tissue, and neural crest (Figure 4.9), but they cannot normally give rise to kidney cells (mesoderm-derived) or liver cells (endoderm-derived). Cells from each of the three germ layers undergo a series of sequential differentiation steps. Eventually, unipotent progenitor cells give rise to terminally differentiated cells that have specialized functions.

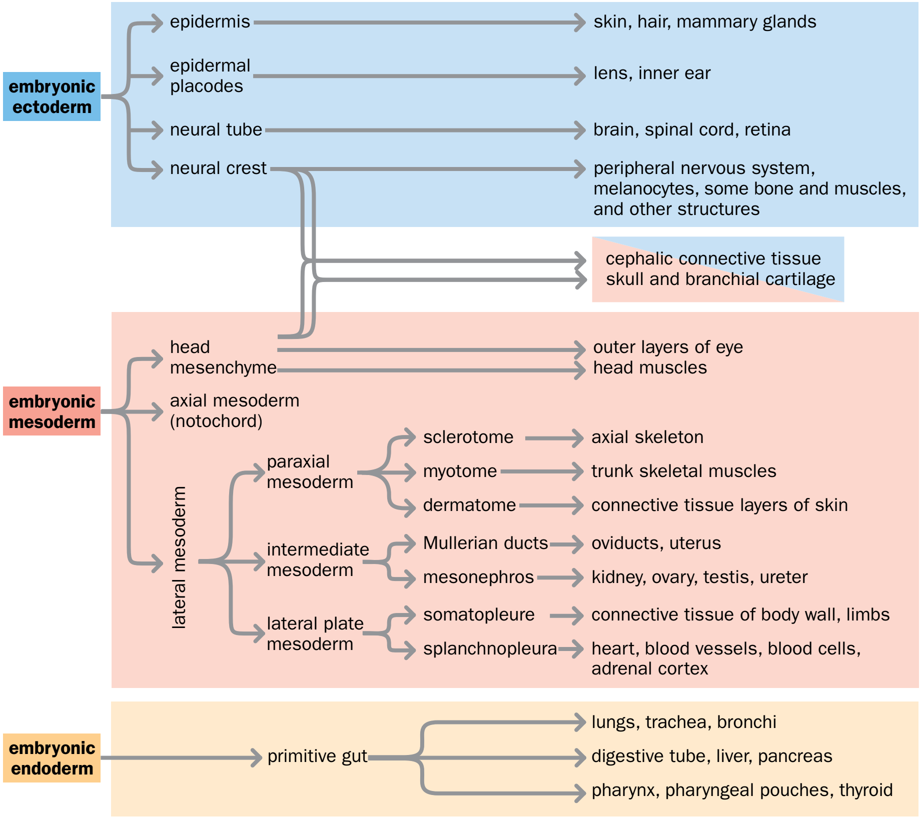

embryonic ectoderm
epidermis
skin, hair, mammary glands
epidermal placodes
lens, inner ear
neural tube
brain, spinal cord, retina
neural crest
peripheral nervous system, melanocytes, some bone and muscles, and other structures
cephalic connective tissue
skull and branchial cartilage
embryonic mesoderm
head mesenchyme
outer layers of eye
head muscles
axial mesoderm (notochord)
lateral mesoderm
paraxial mesoderm
sclerotome
axial skeleton
myotome
trunk skeletal muscles
dermatome
connective tissue layers of skin
intermediate mesoderm
Mullerian ducts
oviducts, uterus
mesonephros
kidney, ovary, testis, ureter
lateral plate mesoderm
somatopleure
connective tissue of body wall, limbs
splanchnopleura
heart, blood vessels, blood cells, adrenal cortex
embryonic endoderm
primitive gut
lungs, trachea, bronchi
digestive tube, liver, pancreas
pharynx, pharyngeal pouches, thyroid

Figure 4.9 Principal derivatives of the three germ layers. The three germ layers formed during gastrulation will eventually form all tissues of the embryo. The connective tissue of the head and the cartilage of the skull and of structures derived from branchial arches are a mixture of ectodermal and mesodermal tissue, as shown. Although the notochord persists in adults in some primitive vertebrates, in mammals and other higher vertebrates it becomes ossified in regions of forming vertebrae and contributes to the center of the intervertebral disks. Note that some of the embryonic mesoderm cells go on to form extra-embryonic mesoderm.

Cell fate decisions are generated by inductive signals from surrounding tissues or through asymmetric cell divisions

In development, the **fate** of a cell means the range of cell types that the cell will normally give rise to. A cell's fate affects all aspects of its behavior, defining its morphology, proliferative and migratory status, and its ability to execute a range of specific functions associated with its differentiated state. Muscle cells are specialized for contraction, neurons for electrical activity, and white blood cells for immunity, for example; each of these functions requires specific cellular properties (cell shape, size, and so on) and choice of neighbors.

Most cells gain their identity during development, and the molecular basis of the decisions that shape cell fate involves various cell signaling pathways and competition between suites of transcription factors that direct alternative cell fates at key decision-making points in development. Cell fate is specified in one of two ways:

- ***Conditional specification.*** The cells must receive some kind of inductive signal from neighboring cells. Usually, groups of cells have their fates specified after encountering suitable extracellular signals;
- ***Autonomous specification.*** Some previously established feature within a cell lineage favors asymmetric cell divisions that produce daughter cells with different amounts of a fate-determining factor. Typically, individual cells have their fates specified in this way.

Cell fate determination by inductive signals

During mammalian development, cell fate is often determined by cell–cell signaling. Cells may migrate into a region where some chemical signal, such as a morphogen, has been released, for example. Or close physical contact between two cell populations allows one of the cell populations to influence the development of neighboring cells via close-range interactions (juxtacrine or paracrine signaling) that will determine the fate of the neighboring cells (*induction*). In that case, cells in one tissue type (the inducer) typically send signals to cells in an immediately adjacent tissue (the responder), causing the latter to change its behavior in some way.

Clear evidence for induction comes from surgical transplantation experiments that are carried out easily in the very large *Xenopus* embryo. A good example is the formation of the neural plate that gives rise to the neural tube and then to the central nervous system (brain plus spinal cord). The neural plate arises from ectoderm cells positioned along the dorsal midline surface of the embryo; ectoderm cells on either side of the midline give rise to epidermis. Initially, however, all of the surface ectoderm is uncommitted (or *naive*): it is *competent* to give rise to either epidermis or neural plate. The ectodermal cells are therefore flexible. If positioned on the ventral surface of the embryo they will normally give rise to epidermis, but if ventral ectoderm cells are grafted to the dorsal midline surface they give rise to neural plate (**Figure 4.10**). Similarly, if dorsal midline ectoderm cells are grafted to the ventral or lateral regions of the embryo they will form epidermis. The dorsal ectoderm cells are induced to form the neural plate along the midline because they receive specialized signals from underlying mesoderm cells.

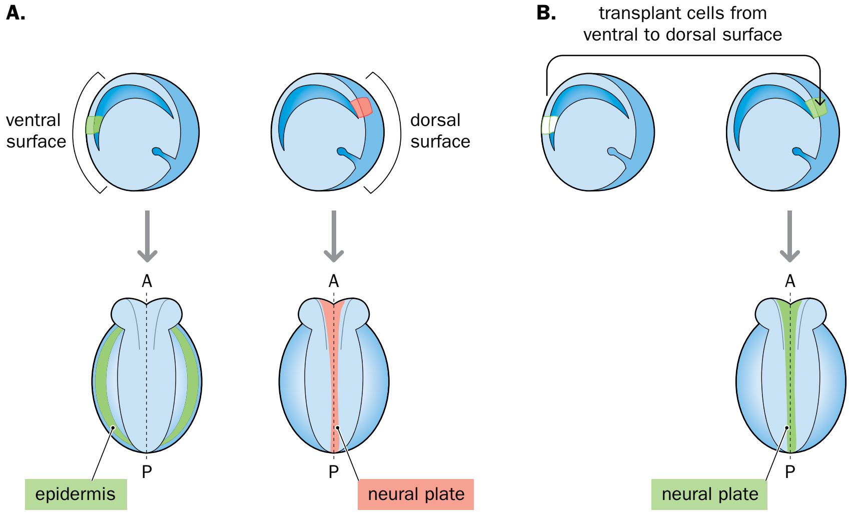


Figure 4.10 According to their position, ectoderm cells in the early *Xenopus* embryo become committed to epidermal or neural cell fates. (**A**) Ventral ectoderm cells (green) normally give rise to epidermis but cells in the dorsal midline ectoderm (red) give rise to the neural plate that is formed along the anterior–posterior (A-P) axis. (**B**) If a part of the ventral ectoderm is grafted onto the dorsal side of the embryo, it is re-specified and now forms the neural plate instead of epidermis. This shows that the fate of early ectoderm cells is flexible and does not depend on their lineage but on their position. The fate of dorsal midline ectoderm cells is specified by signals from cells of an underlying mesoderm structure, the *notochord*, that forms along the anterior–posterior axis.

In this case, the fate of the ectoderm—epidermal or neural—depends on the position of the cells not their lineage, and is initially reversible. At this point, cell fate is said to be specified, which means it can still be altered by changing the environment of the cell. Later on, the fate of the ectoderm becomes fixed. It can no longer be altered by moving the cells through grafting and now the cells are said to be **determined**, irreversibly committed to their fate. (That happens because some molecular process has been initiated that inevitably leads to differentiation: a new transcription factor may be synthesized that cannot be inactivated, or a particular gene expression pattern is locked in place through chromatin modifications. There may also be a loss of competence for induction. For example, ectoderm cells that are committed to becoming epidermis may stop synthesizing the receptor that responds to the signal emanating from the underlying mesoderm).

Cell fate determined by cell lineage

There are fewer examples where cell fate is specified by lineage in vertebrate embryos. However, sometimes stem cells divide by a form of asymmetric cell division that produces inherently different daughter cells: one with the same type of properties as the parent stem cell, ensuring stem cell renewal, and one with altered properties, becoming committed to producing a lineage of differentiated cells. In cases where the fate of the committed daughter cell is not influenced by its position, or by signals from other cells, the decision is intrinsic to the stem cell lineage, and the specification of cell fate is said to be autonomous (nonconditional). We consider stem cells in detail in Section 4.2.

The molecular basis of early lineage specification and alternative models of how different lineages develop

In Section 4.2, when we consider stem cells and cell reprogramming, we examine the molecular basis of differentiation from pluripotent cells to ectoderm, mesoderm, and endoderm. Here, we examine the molecular basis of early lineage specification to give the three cell types in the blastocyst, and we consider alternative models of lineage segregation at this early stage of development. Two of the cell lineages giving rise to blastocyst cells—trophoblast and primitive endoderm—are exclusively extra-embryonic cell lineages. The epiblast lineage gives rise to all the cells of the embryo proper (somatic cells arising from ectoderm, mesoderm, and endoderm, plus germ cells), and also amniotic ectoderm and mesoderm cells that will become components of extra-embryonic membranes.

Cellular differentiation depends on the actions of specific transcription factors; by regulating the expression of certain target genes, they promote a specific type of differentiation. Like all transcription factors, they work by binding to specific DNA sequences in, or near, the genes they regulate. Certain of the transcription factors act as master regulators, controlling the expression of large numbers of downstream genes.

Cell fate decision #1: ICM or trophectoderm?

Blastomeres of the early embryo express master transcription factors associated both with pluripotency (OCT4) and with the trophoblast state (CDX2) that are mutually antagonistic: OCT4 represses the gene encoding CDX2, and CDX2 represses the gene encoding OCT4. When progressing to the blastocyst, a decision is made to tip cells into one state or the other; either pluripotency-associated genes need to be switched off (to give trophoblast) or the trophoblast-promoting genes are switched off (to give ICM).

Once cells are tipped toward the ICM state, many pluripotency-promoting genes are directed by three master transcription factors: OCT4, SOX2, and NANOG. They are auto-stimulatory and mutually stimulatory (**Figure 4.11A**). In the blastocyst, CDX2 is expressed by trophoblast cells only; it down-regulates genes encoding both OCT4 and NANOG, thereby repressing the pluripotency-promoting pathway. CDX2 production is regulated by the TEAD4 transcription factor, and it is regulated by components of the Hippo signaling pathway, notably the transcriptional co-activator YAP, or its closely related homolog TAZ. In turn, the YAP and TAZ proteins are regulated by allowing or denying them access to the nucleus (**Figure 4.11B**).

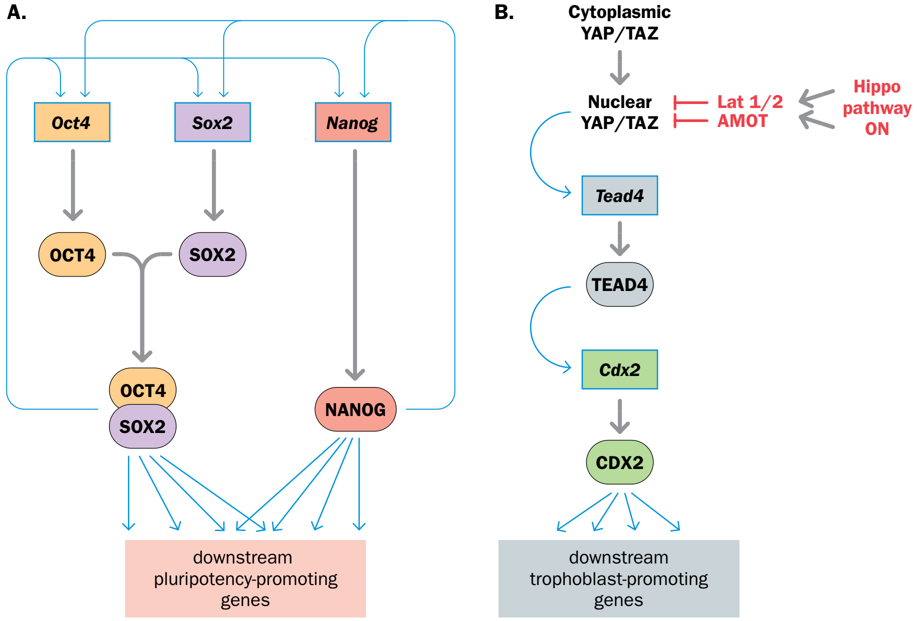


Figure 4.11 Regulation of pluripotency-promoting genes and trophoblast-promoting genes. Boxes signify genes, ovals signify protein transcription factors. Blue arrows indicate regulation of genes by the indicated transcription factor. (**A**) Pluripotency gene network. The *Oct4*, *Sox2*, and *Nanog* genes are master genes regulating pluripotency. They encode transcription factors that regulate hundreds of downstream regulatory genes (the OCT4 and SOX2 proteins work together as a heterodimer). These proteins also bind enhancers in their own genes and those of the other master genes in order to promote their transcription. (**B**) Regulation of trophoblast-promoting genes. The master regulator CDX2 controls the transcription of many downstream trophoblast-promoting genes. CDX2 production is controlled by upstream transcription factors: TEAD4 (which binds to an enhancer in the *Cdx2* gene to promote transcription) and the YAP and (closely related) TAZ proteins that act as co-activators of *Tead4* transcription. This pathway can be negatively regulated by the Hippo signaling pathway (shown in red), which prevents YAP and TAZ entering the nucleus. In that case, YAP and TAZ are bound by AMOT (angiomotin) cell junction proteins and recruited to tight junctions or the actin cytoskeleton; after being sequestered in this way, YAP and TAZ show reduced nuclear localization and activity. Additionally, AMOT proteins activate Lat1 and Lat2 kinases, which phosphorylate YAP and TAZ, and so target them for destruction (see also **Figure 4.12**).

As described in **Box 4.2**, compaction is a key stage that induces cell polarity. In the eight-cell embryo all eight cells are symmetrical, and are symmetrically positioned. But after compaction the cells are polarized. Then, after the cells have divided, the 16-cell morula has outer polar cells, and symmetrical inner cells. According to their position, therefore, the cells show differences in polarity. In the outer cells, upstream Hippo kinase components known as angiomotin (AMOT) proteins are sequestered in the apical domain and, as a result, YAP and TAZ can enter the nucleus to drive a pathway leading to CDX2 expression (**Figure 4.12**). These cells will become trophoblast cells. However, in the inner cells, YAP and TAZ are prevented from entering the nucleus and CDX2 production is inhibited; because OCT4 and NANOG are free to maintain pluripotency, these cells will give rise to the ICM.

Cell fate decision #2: epiblast or hypoblast (primitive endoderm)?

Two key master transcription factors are central to this decision: NANOG is an epiblast-promoting regulator, and GATA6 promotes primitive endoderm formation. Fibroblast growth factor (FGF) signaling is also crucially important, specifically the interaction of the FGF4 ligand and its receptor FGFR2. However, the fates of ICM cells are not irreversibly determined after some of the early cell divisions: when exposed to a different environment, they can switch between epiblast and primitive endoderm identities.

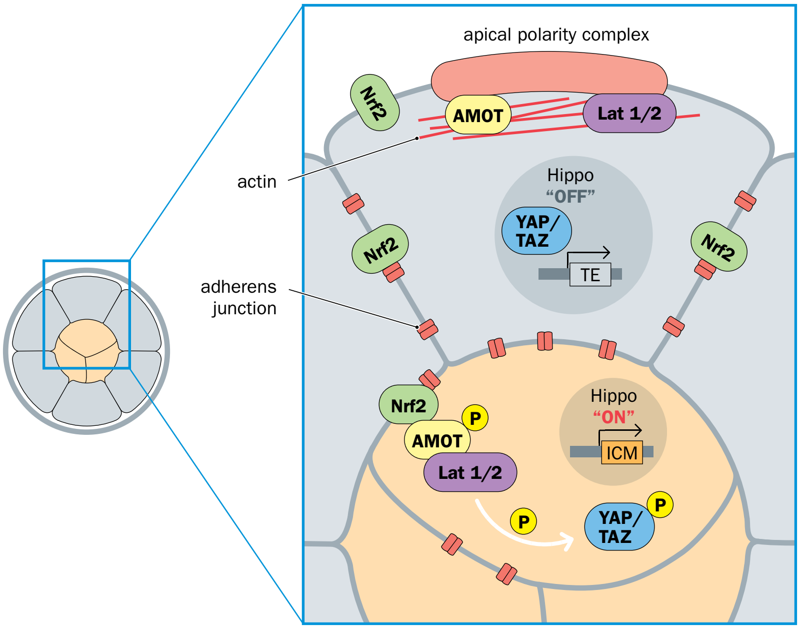


Figure 4.12 Molecular pathways leading to trophectoderm (TE) and inner cell mass (ICM) specification. In blastomeres of the eight-cell mouse embryo, OCT4 (ICM-promoting) and CDX2 (TE-promoting) are co-expressed, and are mutually antagonistic. However, further cell divisions result in two cell populations: inner apolar cells (pale brown color) and outer polar cells (gray color). In the inner cells, the Hippo pathway is activated after angiomotin (AMOT) proteins are localized to adherens junctions, where they bind and activate Lat1 and Lat2 kinases. YAP and TAZ are phosphorylated by active Lat1/Lat2 kinases, and excluded from the nucleus. As a result, CDX2 production is inhibited (see **Figure 4.11B**), and ICM-promoting genes are transcribed. In the outer cells, AMOT is sequestered at the apical domain (possibly by binding to cell polarity proteins such as PARD6b and aPKCζ) and does not activate Hippo signaling. Unphosphorylated YAP/TAZ enter the nucleus to activate *Cdx2*, and so activate transcription of TE genes. At the blastocyst stage, CDX2 inhibits *Oct4* expression in outer cells. OCT4 may contribute indirectly to *Cdx2* repression in inner cells. (From White MD & Plachta N [2015] *Curr Top Dev Biol* **112**:1–17; PMID 25733136. With permission from Elsevier.)

Prior to the 32-cell stage in mouse embryos, NANOG and GATA6 are co-expressed, but from the 32-cell stage onward, epiblast precursor cells in which NANOG inhibits expression of the gene encoding GATA6 begin to be segregated from primitive endoderm precursor cells in which FGFR2 is strongly expressed and FGF signaling inhibits production of NANOG. Initially, the precursors segregate in a mosaic fashion, but eventually cell–cell adhesion allows the formation of separate layers of epiblast and primitive endoderm (see **Figure 4.4B**).

Models of early cell lineage segregation

As detailed by Wennekamp (2013) (PMID 23778971; see Further Reading), different classical models seek to explain early lineage segregation in mammalian embryos. Most attention has focused on models that deal with post-compaction breaking of symmetry: either the symmetry of cells (cell polarity), or of their positions in the embryo (cell surface location versus interior location), or both. In the mouse, cell polarity is evident at the late eight-cell stage, followed by inner cell–outer cell asymmetry developing at the 16-cell stage (when the embryo is called a *blastula*, but is effectively a counterpart of the early human blastocyst). One linkage between cell polarity and inner–outer cell position is the Par family of cell polarity proteins. They are known to be restricted to the apical domain of the cells of the compacted eight-cell mouse embryo, and Par family members PARD6b and aPKCζ seem to be able to bind to angiomotin (AMOT). They may be responsible for sequestering angiomotin (see **Figure 4.12**) so that Hippo signaling is switched off in what will be outer cells of the 16-cell morula.

There is no evidence of pre-patterning of the mouse egg and the fate of blastomeres would not be expected to be fixed in cleavage embryos because of the high capacity of early mammalian embryos for *regulation*, the ability to restore normal development even when portions of the embryo are removed or rearranged. Cleavage embryos can be combined, and several pre-implantation mouse embryos can be aggregated together, for example, without affecting development. Conversely, after the majority of cells in the mouse blastocyst are destroyed, the embryo can still adapt and form a normal conceptus. (Human embryos also regulate, and a blastomere from an *in vitro* fertilization eight-cell embryo can be removed and analyzed in invasive pre-implantation diagnosis; the remaining seven-cell embryo undergoes normal development.) Conventional thinking, therefore, views cells of the early mammalian embryo as having a certain degree of plasticity.

Epigenetic factors are important in early development. Blastomeres of the four-cell mouse embryo, for example, show significant differences in arginine methylation of histone H3: those with maximal levels seem to direct the descendent cells to contribute to the inner cell mass; those with low levels appear to be biased to contribute to trophectoderm. While blastomere fates are not fixed, individual mouse blastomeres harbor intrinsic biases regarding which lineage they will adopt as early as the four-cell stage.

That bias seems to depend in part on positional cues, but stochastic factors can also be expected to have a role.

Germ cell development and sex determination in mammals

All mammals have male and female sexes. The decision between male and female development is made at conception, when the sperm delivers either an X chromosome or a Y chromosome to the egg, which always contains an X chromosome. (Exceptions can occur, however, notably when errors in meiosis produce gametes with missing or extra sex chromosomes, resulting in individuals with sex-chromosome aneuploidies.)

Sperm and egg cells are the differentiated cells that mark the endpoint of development from germ cell progenitors within individuals (but unlike somatic cell lineages that perish when individuals die, the germ line is a potentially immortal cell lineage). In mammals, germ cells are induced by cell–cell interactions in the early embryo, unlike in many animal models (such as *Xenopus*, zebrafish, *Drosophila*, and *Caenorhabditis elegans*) where the germ cells are determined by material within the cytoplasm of the egg (pre-formation). As in other vertebrates, there is a clear separation of mammalian germ cells from somatic cells in the early embryo. The earliest committed germ cell progenitor cells are known as **primordial germ cells**. Much of our knowledge of mammalian germ cell development comes from studies on mice, but, as described below, there are significant human–mouse differences in how germ cells develop.

Primordial germ cell development and migration

The primordial germ cells (PGCs) do not form in the developing gonads, but must migrate there from their site of origin elsewhere in the early embryo. The reason for the physical separation of the site of origin and final destination is unclear. Possibly, the aim is to protect the crucially important germ-line cells by excluding them from the significant upheavals required to lay down the body plan. Or maybe it is a way of selecting for the healthiest germ cells, the ones that survive the extensive migration required.

In the mouse, BMP4/BMP8 (bone morphogenetic protein 4/8) signals transmitted from neighboring extra-embryonic ectoderm cells induce expression of *Fragilis* in posterior epiblast cells, and also of *Blimp1* in a small subset of such cells, about 6–8 cells lying immediately proximal to the ectoderm cells. The latter are the earliest PGCs that can be detected in the mouse (appearing first at E6.25 = embryonic day 6.25; **Figure 4.13**). They express the BLIMP1 transcriptional repressor protein to repress genes required for establishing the somatic development program. By escaping from a somatic cell fate, these early PGCs retain the potential to be totipotent; OCT4, the most important of the pluripotency master transcription factors, is expressed, and there is widespread demethylation of the genome.

The PGCs proliferate and move to the primitive streak where, by E7.25, they number around 40 cells. They then migrate away from the posterior region of the primitive

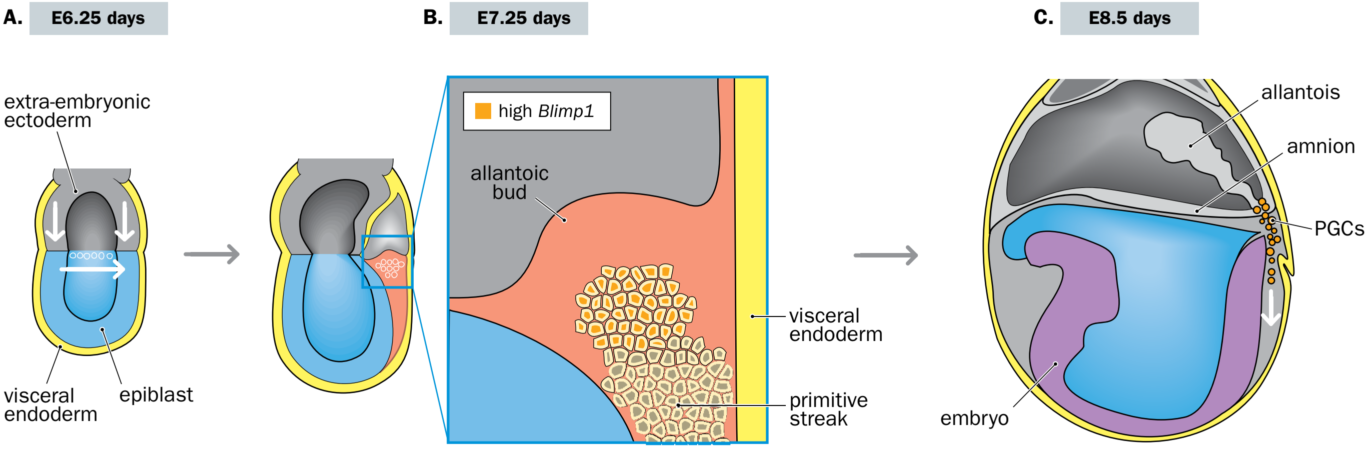


Figure 4.13 Primordial germ cell origins and migration in the mouse. (**A**) A small number of primordial germ cells (PGCs; shown in white) expressing *Blimp1* are first detectable in the proximal epiblast at E6.25 (embryonic day 6.25, or 6.25 days after fertilization). (**B**) During gastrulation, these cells and the surrounding prospective embryonic mesoderm move to the posterior end of the embryo above the primitive streak, where the PGCs start also to express the germ cell lineage-specific gene *stella*. By E7.25 about 40 PGCs (shown in orange) are present in the primitive streak. (**C**) At around E8.5, the PGCs start to migrate toward the gonads (via the hindgut); by the time they reach the genital ridge, their numbers will have increased to about 8000. (Adapted from Hogan B [2002] *Nature* **418**:282–283; PMID 12124605. With permission from Springer Nature. Copyright © 2002.)

streak into the endoderm and enter the developing hindgut for a short period before they migrate into the gonadal (genital) ridge. At this stage, the gonad is bipotential—capable of developing into either testis or ovary.

The migration path of the PGCs is regulated by their environment: they continuously receive chemical signals from cells in the tissues through which they must migrate, notably chemoattractant signals, but also survival and proliferation signals. Eventually, about 8000 PGCs arrive at the genital ridge. Later they will differentiate in the developing gonad, giving rise ultimately to sperm or egg cells.

Human primordial germ cell development is similar to that in mouse but there are differences in transcription factor regulation. In mouse, BLIMP1 and two other transcription factors, PRDM14 and TFAP2C (= AP2γ), are the key regulators, but in humans the SOX17 transcription factor is the key determinant with BLIMP1 acting in tandem.

Sex determination

Although the sex of the human embryo is established at conception, sexual differentiation does not begin to occur until the embryo is about 5 weeks old. Primary sexual characteristics (the development of the gonad and the choice between sperm and egg development) are intrinsic, being dependent on the genotype of the embryo; however, secondary sex characteristics (the sex-specific structures of the urogenital system and the external genitalia) are dependent on signals from their environment, mediated by hormonal signaling.

Male development normally depends on the presence or absence of the Y chromosome. A critical male-determining gene called *SRY* (sex-determining region of the Y chromosome) encodes a transcription factor that activates downstream genes required for testis development. The testis then produces sex hormones required for the development of male secondary sex characteristics.

Female gonad development was previously considered a “default state.” The *SRY* gene was thought sufficient to switch the bipotential embryonic gonad from female to male differentiation. Consistent with this, rare XX males often have a small fragment of the Y chromosome, including *SRY*, translocated onto the tip of one of their X chromosomes, and genetically female mice transgenic for the mouse *Sry* gene develop as males. More recent studies suggest, however, that genes on the X chromosome and autosomes are also involved in positive regulation of ovarian development: over-expression of genes such as *DAX* and *WNT4A* can feminize XY individuals even if they possess a functional *SRY* gene.

Early gamete development appears to be controlled more by the environment than the genotype of the germ cells. Female PGCs introduced into the testis will begin to differentiate into sperm, and male PGCs introduced into the ovary will begin to differentiate into oocytes. This may reflect the regulation of the cell cycle, since PGCs entering the testis arrest prior to meiosis; those entering the ovary commence meiosis immediately. Therefore, PGCs of either sex that colonize somatic tissue outside the gonad begin to differentiate into oocytes since there is no signal to arrest the cell cycle. In all of these unusual situations, however, functional gametes are not produced. Differentiation aborts at a relatively late stage, presumably because the genotype of the germ cells themselves also plays a critical role in gamete development.

Unlike primary sex characteristics, the default state is female for secondary sex characteristics. One of the genes regulated by SRY makes the SF1 transcription factor, which activates genes required for the production of male sex hormones, including *HSD17B3* (encoding hydroxysteroid-17-β-dehydrogenase 3, required for testosterone synthesis) and *AMH* (encoding AMH the anti-Mullerian hormone). Both hormones play important roles in the differentiation of the male urogenital system. AMH, for example, causes breakdown of the Mullerian ducts (which would normally become the Fallopian tubes and uterus in females).

Mutations inhibiting the production, distribution, elimination, or perception of such hormones produce feminized XY individuals. For example, androgen insensitivity syndrome results from defects in the testosterone receptor that prevent the body responding to the hormone even if it is produced at normal levels. XY individuals with this disease appear outwardly as normal females but, due to the effects of SRY and AMH, they possess undescended testes instead of ovaries, and they lack a uterus and Fallopian tubes.

Mutations that lead to the overproduction of male sex hormones in females have the opposite effect: XX individuals are virilized. Occasionally, this occurs in developing male/female fraternal twins, when the female twin is exposed to male hormones from her brother. The CYP19 enzyme converts androgens to estrogens, so mutations that increase its activity can result in the feminization of males; those decreasing or abolishing its activity can lead to the virilization of females.

4.2 STEM CELLS AND CELL DIFFERENTIATION

All our cells originate from the fertilized egg, the ultimate progenitor cell. As detailed in Section 4.1, the zygote gives rise to a series of increasingly more specialized progenitor cells with reduced differentiation potential (potency). That is, progenitor cells produced early in development have wide differentiation potential; those produced later have more limited potency. The endpoint of differentiation is a wide variety of specialized cell types. Some, such as hepatocytes, lymphocytes, and cardiomyocytes, are adapted to very specific functions; others, such as fibroblasts, may have more general functions and be present in a wide range of tissues and organs.

During early development growth is rapid but increasingly as we approach maturity the proportion of dividing cells falls. In adults, the majority of cells are terminally differentiated *post-mitotic cells* (which now do not divide). Just a small minority of the cells are capable of cell division (some are actively dividing; others may remain in a quiescent state and divide only occasionally). Within this grouping are progenitor cells that ultimately give rise to terminally differentiated somatic cells or gametes.

Progenitor cells in the body give rise to more specialized cells, but a subgroup of progenitor cells called **stem cells** also have the ability to renew themselves by cell division (and are capable of unlimited division). As well as the stem cells that occur naturally in the body (where their main job is to replace short-lived cells), a wide variety of artificial stem cell lines have been developed in the laboratory. As listed in **Table 4.1**, stem cells can be classified into two broad divisions, according to whether they can occur naturally or are entirely artificial constructions.

TABLE 4.1 CHARACTERISTICS OF MAMMALIAN TISSUE STEM CELLS AND PLURIPOTENT STEM CELLS

Stem cell class	Provenance	Subclasses	Differentiation potential (potency)	Examples
Tissue stem cells (adult and fetal stem cells)	Occur naturally in certain tissues of the body, notably those with high turnover, such as skin, intestine, and blood; but have also been grown in culture with variable degrees of success	Multipotent stem cells	Can give rise to multiple cell lineages	Hematopoietic stem cell (**Figure 3.17**) Mesenchymal stem cells
		Oligopotent stem cells	Can give rise to a few types of differentiated cell only	Neural stem cell that can create a subset of neurons in brain
		Unipotent stem cells	Can form one type of differentiated cell only	Spermatogonial stem cell
Pluripotent stem cells (PSC)	Do not occur naturally. Artificially produced, beginning with short-lived pluripotent cells from the early embryo or somatic cells	Pluripotent cells of the early embryo that have been cultured and selected	Can potentially give rise to all of the cells of the organism (but not to the extra-embryonic membranes)	Embryonic stem cells Epiblast stem cells Embryonic germ cells
		Differentiated cells that have been epigenetically reprogrammed into PSCs		SCNT stem cells Induced pluripotent stem cells

SCNT, somatic cell nuclear transfer.

Why stem cells are important for both scientific and medical research

Analysing stem cells *in vivo* (primarily in laboratory animals) and *in vitro* (using cultured stem cell lines) provides fundamental knowledge of basic stem cell biology. In addition, some stem cells, notably embryonic stem cell lines, have been especially valuable in allowing the functions of mammalian genes to be dissected, as described in Chapter 8. In addition, as described in Chapters 19, 21, and 22, stem cells have attracted huge interest in the field of medicine for four reasons. First, because of their capacity to keep dividing and producing defined cell types in the body, they are important target cells for delivering gene constructs in gene therapy. Secondly, because of their general potential to replenish specific cell populations, they have propelled new types of cell therapy and a developing field of regenerative medicine. Thirdly, they provide important routes for making animal and cellular models of disease. Finally, it has become clear that cancers may often be the result of aberrant stem cells that have subverted normal constraints on cell proliferation.

Stem cell division and the balance between stem cell proliferation and differentiation

The defining property of stem cells is their ability to generate progeny (offspring) with different cell fates: daughter stem cells (to maintain stem cell number) and daughter progenitor cells committed to differentiation. And there needs to be a balance between stem cell proliferation and the generation of differentiated progeny.

In the classical stem cell view, individual stem cells divide to give the different daughter cells as a result of intrinsic or extrinsic asymmetry. In the former case, the daughter cells acquire different cell fates as a result of asymmetric division of the parent stem cell causing unequal distribution in the daughter cells of a protein that regulates cell fate (Figure 4.14A). The alternative is extrinsic asymmetry: here stem cell division is symmetrical, but asymmetry arises when the daughter cells are placed in different microenvironments where they receive different external chemical signals, causing one of them to remain a stem cell and the other to become a progenitor cell committred to differentiation (Figure 4.14B); we cover the role of stem cell microenvironments in more detail in Section 4.3.

Symmetrical stem cell divisions are now believed to be very common and recent evidence suggests that the balance between stem cell renewal and differentiation can also be obtained at a population level. That is, some stem cells may divide to give identical daughter stem cells, while other stem cells divide to generate two progenitor cells committed to differentiation (Figure 4.14C).

Whichever way the stem cell generates a progenitor cell committed to differentiation, asymmetry is required to create downstream differentiated cells. The progenitor cell first produces **transit amplifying cells** that go through a finite number of symmetrical cell divisions to rapidly expand their numbers before generating the different types of differentiated cells found in the tissue (Figure 4.14D). Transit amplifying cells normally account for the majority of dividing cells in an adult tissue. They can be multipotent, but they are not stem cells: they are very short-lived, and more dispensable, than adult stem cells (which can continue to divide over very long periods, up to a whole lifetime).

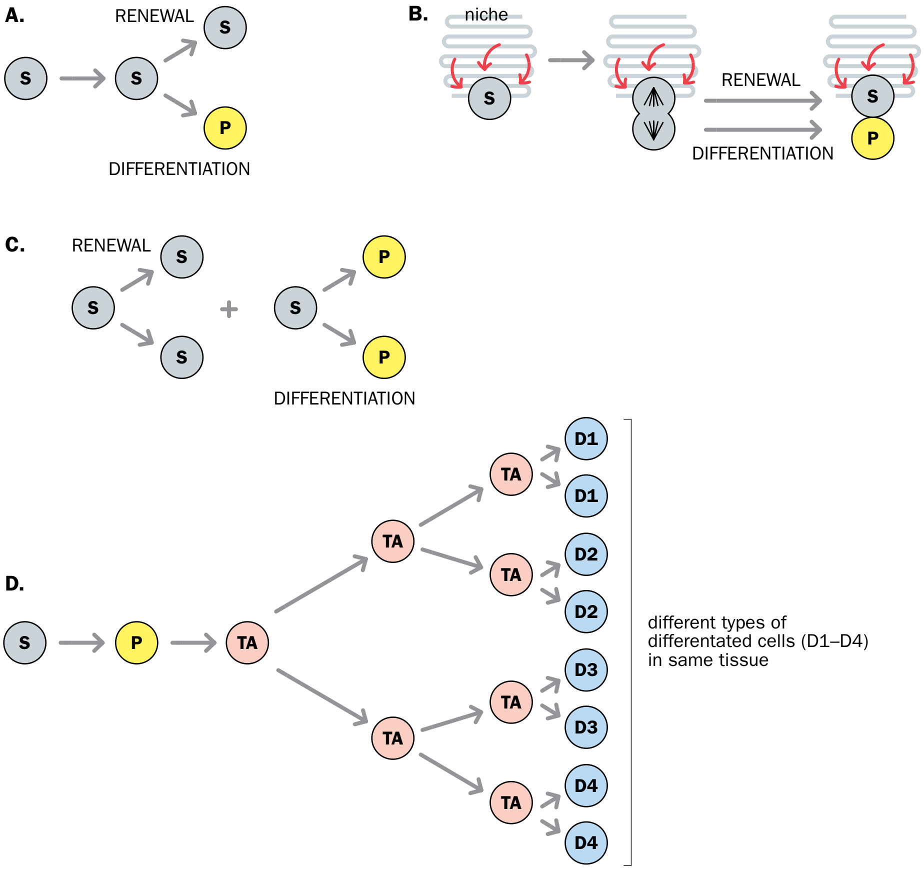


Figure 4.14 Asymmetry arising at or from cell division can explain how stem cells generate both new stem cells and differentiated cell progeny. (**A**) Intrinsic (autonomous) asymmetry can arise as a result of asymmetric division of a stem cell, S, to give a daughter stem cell and a daughter progenitor cell, P, committed to differentiation. Before cell division a cell fate regulator, such as a polarity protein, accumulates at one pole of the cell and the plane of division is adjusted by re-orienting the mitotic spindle so that one daughter cell receives the great majority of the cell fate regulator. (**B**) Alternatively, extrinsic asymmetry can be involved: cell division is symmetrical but the two daughter cells receive different external signals from their microenvironments. Tissue-specific stem cells occur in specialized microenvironments (**stem cell niches**) where stem cell differentiation is suppressed by signals (red curly arrows) received from immediately neighboring cells (not shown) within the niche. When the physical site occupied by a stem cell in its niche is extremely limited and confined to the niche boundary, just one daughter cell can remain at the original stem cell location (and becomes a stem cell); the other daughter cell exits from the niche, escapes the differentiation-suppressing signal, and becomes a progenitor cell, P, that gives rise to differentiated cells. (**C**) Population-based origin of asymmetry. Asymmetry can also arise at the population level: some stem cells produce identical daughter stem cells, while others divide to produce daughter progenitor cells committed to differentiation. (**D**) Downstream asymmetry. When a tissue-specific stem cell gives rise to progenitors committed to a differentiation pathway, it first produces transit amplifying cells, TA, that divide quickly by symmetrical divisions to rapidly expand the population of cells committed to differentiate (with multiple TA generations rather than the two generations shown here for simplicity). Asymmetry at subsequent cell divisions can produce different types of differentiated cells.

Tissue stem cells are important in tissue renewal and are maintained in specialized microenvironments

Naturally occurring stem cells in the body are not pluripotent (the pluripotent cells of the early embryo are transiently existing cells, not stem cells). To distinguish them from embryonic stem cells, our natural tissue stem cells are often loosely called adult stem cells. Usually multipotent, they typically give rise to the various types of differentiated cell within the tissue they reside in, and their job is to ensure production of suitable differentiated cells as replacements for cells lost from the tissue (normally through natural turnover). That is, they are principally tissue-specific stem cells; a stem cell in intestinal epithelium does not make hepatocytes or muscle cells, for example.

Stem cells are conspicuously lacking in tissues and organs where there is a very low rate of cell turnover, such as in the adult brain (where neurons are very long-lived). Conversely, they are prominent in tissues with a high cell turnover, notably the epithelium lining the small intestine (where cells are renewed roughly every four days), the epidermis of the skin (cells are renewed about once a fortnight), and blood cells. As detailed below, the cells in these tissues are each capable of being maintained by a single type of stem cell.

Tissue-specific stem cells are maintained in special supportive microenvironments, called **stem cell niches**, where chemical signals are conveyed from neighboring cells and extracellular matrix to receptors on the stem cell to support stem cell activity and renewal (and to suppress stem cell differentiation). In tissues that are subject to significant mechanical, chemical, and environmental assault the stem cell niches are located in deep-lying regions for protection, and may be protected by neighboring cells from microbial attack. We give some examples below of some of the more commonly studied tissue stem cells. (Note that much of our information on tissue stem cells comes from studies on mice where it is possible to label specific cells and trace their progeny).

Bone marrow stem cells

Human blood cells are initially made in certain embryonic structures before the fetal liver takes over production. From fetal week 20 onward, blood cells originate from the bone marrow, where two types of stem cell have long been known.

- Hematopoietic stem cells (HSCs) are multipotent stem cells anchored to fibroblast-like osteoblasts of the "spongy" inner marrow of the long bones, such as the femur. Studies in mice indicate that a single type of HSC is capable of making all the different blood cells: after a mouse's bone marrow cells have been destroyed by radiation, a graft of purified HSCs from another mouse allows the incoming cells to differentiate and re-populate the blood. Only about 1 in 10,000 to 15,000 bone marrow cells is an HSC; the frequency in blood is ~1 in 100,000. HSCs are naturally multipotent, ultimately producing all of the terminally differentiated blood cells plus some tissue cells that work in the immune system (see **Figure 3.17**).
- Mesenchymal stem cells (MSCs) are stromal cells found not just in bone marrow but in organs throughout the body. Bone marrow MSCs (also known as bone marrow stromal cells) are poorly defined and heterogeneous, and unlike HSCs they do not self-renew quite so regularly, but are relatively long-lived. In artificial culture conditions, MSCs derived from the bone marrow can give rise to a variety of cell types, including cartilage, fat, and fibrous connective tissue, as well as bone, but *in vivo* their normal function is likely to be focused on the gradual turnover of bone. Umbilical cord MSCs have a particularly broad potency *in vitro*.

Epidermal and intestinal stem cells

In mammals, epidermal stem cells are known to occur in three locations, of which those in the bulge region of the hair follicles give rise to both the hair follicle and to epidermis (**Figure 4.15A**). The stem cells that will form epidermis give rise to keratinocytes that progressively differentiate as they move toward upper layers. In the epithelium of the small intestine, stem cells are located near the base of intestinal crypts, small pits that are interspersed between the fingerlike villi that protrude into the lumen (**Figure 4.15B** and **C**).

In order to be able to self-renew, the stem cells rely on receiving chemical signals, often members of the Wnt protein family, from neighboring cells in their niche. These signals may be transmitted by different types of neighboring cells (as in the case of intestinal stem cells) or by neighboring stem cells (as in the case of stem cells in the epidermal basal layer); see **Figure 4.16**.

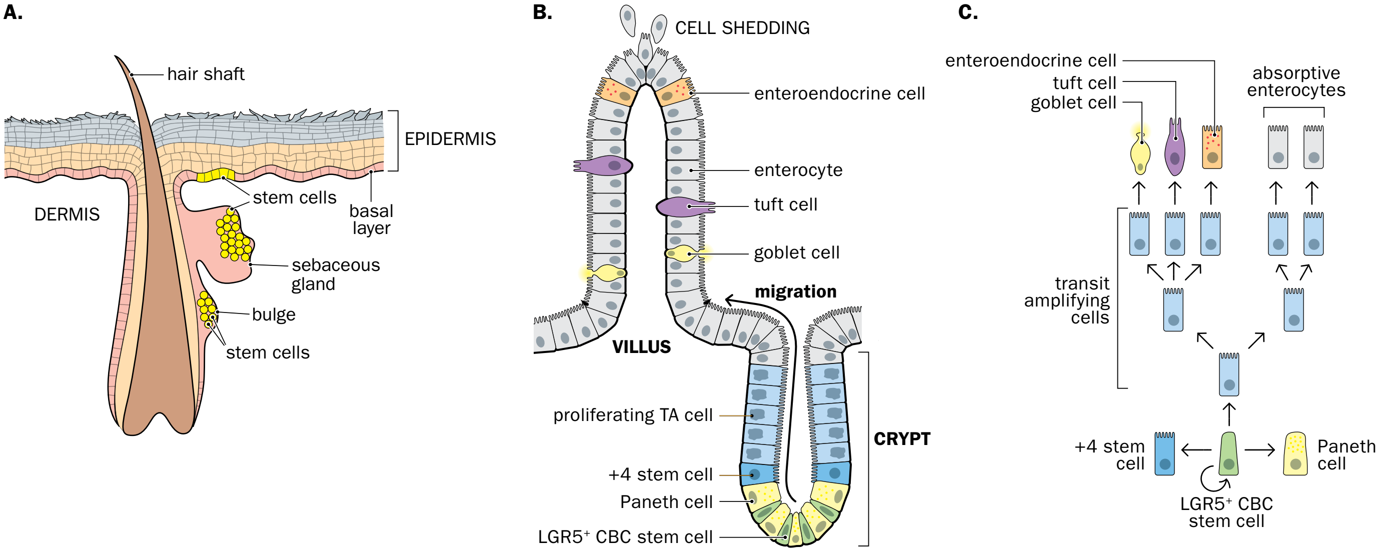


Figure 4.15 Epidermal and intestinal stem cell locations. (**A**) Epidermal stem cells can be found in three types of niche: in the bulge region of the hair follicle, the sebaceous gland, and the basal layer of the epidermis (where they produce proliferating transit amplifying cells that, in turn, give rise to increasingly differentiated upper cell layers). (**B**) Intestinal stem cells (called crypt base columnar or CBC stem cells) are protected by being located at the base of pits (crypts) positioned next to fingerlike projections (villi) that extend into the lumen. They are identified by testing positive for LGR5, a type of G-protein-coupled receptor. They alternate with Paneth cells (which defend the crypt base against microbes by secreting antimicrobial peptides [defensins] and proteins [including lysozyme]). (**C**) CBC stem cells generate all the cells at the base of the crypt, plus rapidly proliferating, but short-lived, transit amplifying (TA) cells, which occupy the remainder of the crypt. TA cells differentiate into the various functional cells on the villi (enterocytes, tuft cells, goblet cells, and enteroendocrine cells) to replace the epithelial cells that are continuously shed at the villus tip. The +4 "reserve" stem cells (which occupy the fourth position from the crypt base) can restore the LGR5-positive (LGR5⁺) CBC stem cell compartment following injury. Epithelial turnover occurs about every 4 days. New Paneth cells are supplied every 3–6 weeks. Note that TA cells migrate upwards out of the crypt and up into the villi to produce desired differentiated cells to replace the cells constantly being shed from villi. (A, courtesy of The Graduate School of Biomedical Science, University of Medicine and Dentistry of New Jersey; C, from Barker N [2014] *Nat Rev Mol Cell Biol* **15**:19–33; PMID 24326621. With permission from Springer Nature. Copyright © 2014.)

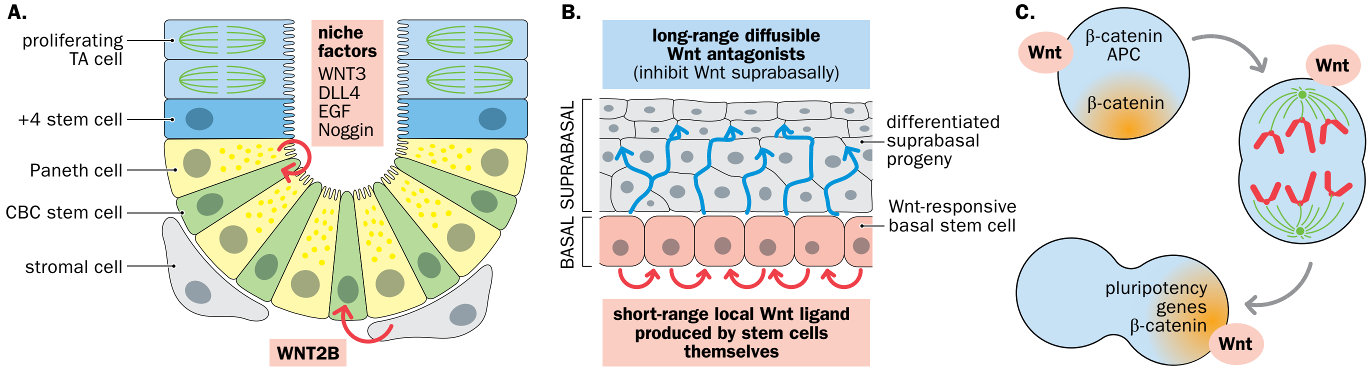

A.
proliferating TA cell
+4 stem cell
Paneth cell
CBC stem cell
stromal cell
niche factors
WNT3
DLL4
EGF
Noggin
WNT2B
B.
long-range diffusible Wnt antagonists
(inhibit Wnt suprabasally)
SUPRABASAL
BASAL
differentiated suprabasal progeny
Wnt-responsive basal stem cell
short-range local Wnt ligand produced by stem cells themselves
C.
Wnt
β-catenin
APC
β-catenin
Wnt
pluripotency genes
β-catenin
Wnt

Figure 4.16 Stem cell renewal is often regulated by Wnt signaling within stem cell niches. (**A**) Intestinal stem cell niche. Components of the CBC (crypt base columnar) stem cell niche at the crypt base include neighboring Paneth cells and pericryptal stromal cells, which supply factors—Wnt proteins, the Notch ligand Delta-like 1 (DLL4), epidermal growth factor (EGF), and Noggin—to regulate the survival and function of the CBC stem cells *in vivo*. TA cell, transit amplifying cell. (**B**) Epidermal basal layer stem cell niche. Within the interfollicular epidermis, basal layer stem cells act as their own niche: by expressing ligands for Wnt proteins they continuously induce their own self-renewal. Basal stem cells also express long-range antagonists of Wnt proteins that diffuse to suprabasal layers, basally limiting the Wnt signaling field and "self-organizing" the stratified epidermal architecture. (**C**) How a local Wnt protein signal induces asymmetric cell division. A cell exposed to a local, extracellular Wnt signal distributes Wnt signaling pathway components to the side of the cell receiving the Wnt protein signal. This orients the mitotic spindle and centrosomes during cell division. The daughter cell close to the Wnt source maintains nuclear β-catenin and stem cell gene expression; the daughter cell remote from the Wnt protein signal loses expression of such genes. APC, adenomatous polyposis coli protein. (A, adapted from Barker N [2014] *Nat Rev Mol Cell Biol* **15**:19–33; PMID 24326621. With permission from Springer Nature. Copyright © 2014; B and C, from Clevers H *et al*. [2014] *Science* **346**:1248012; PMID 25278615. Reprinted with permission from the AAAS.)

Manipulating cells obtained from early embryos to establish immortal pluripotent stem cell lines

A variety of mammalian pluripotent stem cell lines have been artificially created and are used for different purposes. In addition to enabling studies on stem cell properties, such as stem cell renewal and differentiation, they have been applied in many other ways, notably as tools for studying mammalian gene function, modeling disease, and drug screening. There is also the potential for therapeutic applications (by generating appropriate cells to replace cells lost through injury or disease).

A major way to produce pluripotent stem cell lines involves obtaining suitable cells from the early mammalian embryo and manipulating them *in vitro*. As detailed in Section 4.1, certain cells have very high differentiation potential. The zygote and the blastomeres of cleavage embryos are totipotent, but are not numerous and not so amenable to culture. Later on, pluripotent cells are found in the undifferentiated inner cell mass (ICM) and in the epiblast of later-stage blastocysts (**Figure 4.4B**). These naturally pluripotent cells are not stem cells that renew themselves while producing more specialized cells: the ICM and epiblast are transient structures with short-lived founder cells that quickly give rise to differentiated tissues. However, the pluripotency of ICM/epiblast cells can be captured by culturing them under conditions that favor cell renewal and suppress differentiation. Success in this endeavour yields a pluripotent stem cell line that can undergo unlimited cell division.

Germ cell progenitors are another special case. Although they express germ-line-specific genes, primordial germ cells and other germ cell progenitors have genomes with very similar epigenetic settings to those of totipotent cells. They can readily be induced toward higher potency; when cultured under certain conditions they convert to pluripotency.

Many different types of pluripotent stem cell lines have been made by culturing cells from early mouse embryos, but according to the source of embryonic cells and the state of pluripotency, three major classes of pluripotent mouse cell line have been distinguished, as listed below.

- ***Embryonic stem cell (ESC).*** Formed by manipulating pre-implantation blastocyst cells in culture, the cells demonstrate a state of *naive* pluripotency resembling that of the early epiblast.
- ***Epiblast stem cell (EpiSC).*** Formed by manipulating later-stage, post-implantation epiblasts in culture, the cells are said to be in a *primed* pluripotency state: they express certain lineage-specific factors that make them more predisposed to differentiate than ESCs.
- ***Embryonic germ cell (EGC).*** Formed by culturing germ-line cells, such as primordial germ cells, that convert to pluripotency *in vitro*.

Origins of embryonic stem cells

ESCs have been particularly well studied and intensively used. They were first reported in 1981 after successful culturing of cells from the ICM of blastocysts from the 129 mouse strain. This mouse strain is unusual because males sporadically develop testicular **teratocarcinomas**, malignant germ cell tumors that can be maintained continuously by serial transplantation. Teratocarcinomas and related benign teratomas contain multiple tissue types that can represent the three germ layers, and fully differentiated structures can form, such as teeth and hair (**Figure 4.17**).

Teratocarcinomas are associated with the presence of embryonal carcinoma cells. These cells are not germ cells (which do not normally differentiate into other lineages); instead, they are proliferative pluripotent cells that closely resemble cells from the ICM

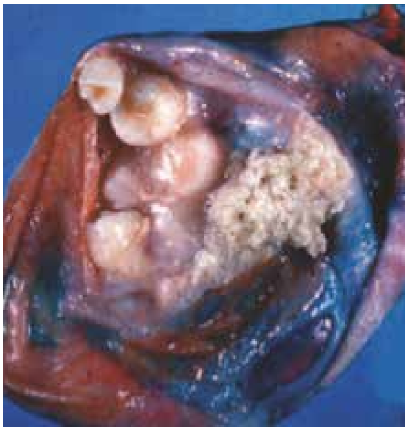


Figure 4.17 Teratomas, like teratocarcinomas, are tumors of germ cells converted to a pluripotent state that can differentiate into diverse somatic tissues. Teratomas and teratocarcinomas are, respectively, benign and malignant germ cell tumors. They have a disorganized collection of multiple different tissue types, and can have fully differentiated structures, such as teeth and hair, giving a bizarre appearance, as in this example of a teratoma. Both types of tumor arise from changes in germ cell progenitors. While expressing germ-line-specific genes, primordial and embryonic germ cells can be induced to a pluripotent state by genetic and epigenetic changes responsible for tumor formation and then differentiate to give diverse somatic tissues.

in morphology, in ultrastructure, and in molecular markers. When injected into blastocysts, to test if they could behave like cells of the early embryo, cells from some embryonal carcinoma cell lines were able to colonize the host embryo and produce live-born **chimeras** (with cells originating from two zygotes). Even in the best cases, however, the efficiency was low. Embryonal carcinoma cells had another major drawback: they were genetically abnormal (having been derived from tumors).

Because of the difficulties with embryonal carcinoma cells, attention focused on deriving pluripotent stem cell cultures using blastocyst explants of the 129 mouse strain. Because pluripotent cells from the early embryo have a natural tendency to differentiate, the cell culture system was required to maintain the pluripotency of isolated ICM cells and suppress cell differentiation, while stimulating cell growth. That was possible by co-culturing with a layer of irradiated fibroblasts in the presence of medium containing fetal calf serum. The fibroblasts act as feeder cells: they are stimulated to produce matrix and growth factor support for the ESCs (but after having been irradiated they cannot divide). When grafted into adult mice, the cultured ESCs give rise to teratocarcinomas. The final proof of pluripotency was successful germ-line transmission following injection of ESCs into isolated blastocysts that were then implanted in a foster mother (Figure 4.18).

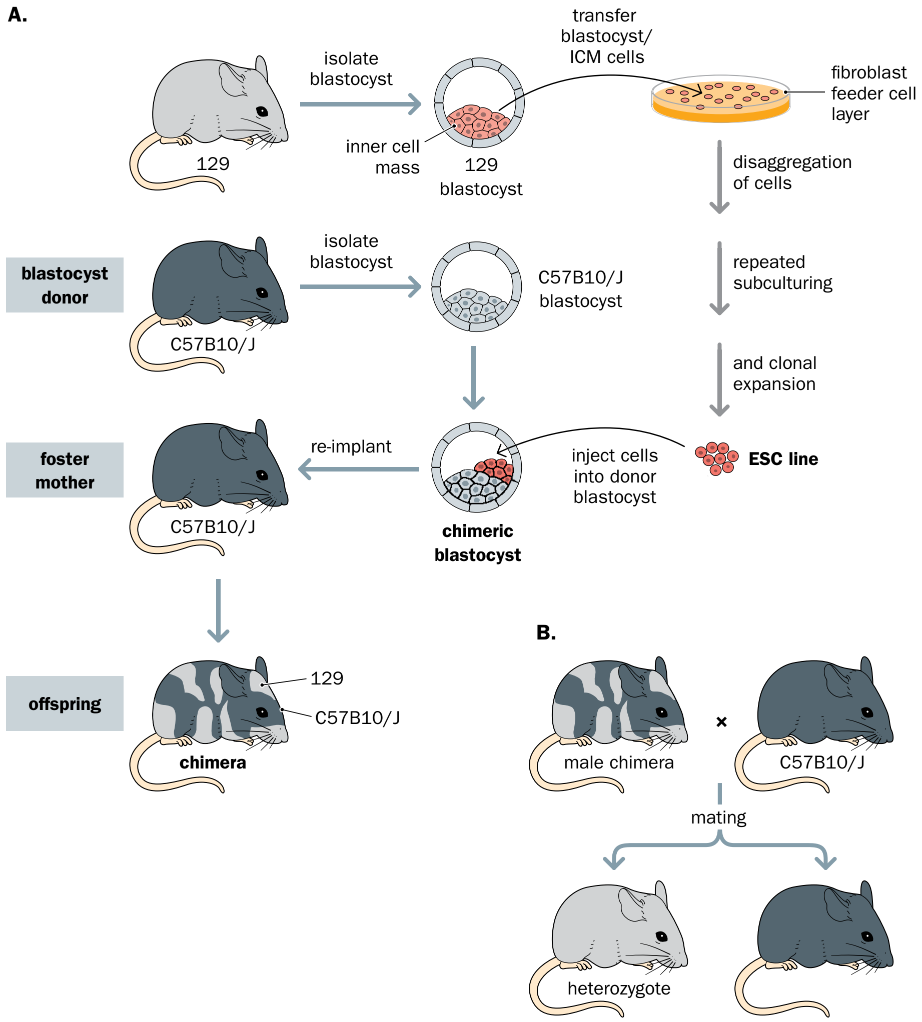


Figure 4.18 Development of an embryonic stem cell (ESC) line from the 129 mouse strain and how germ-line transmission can be demonstrated. (**A**) ESC isolation and chimera formation. To construct the ESC lines from the 129 mouse strain, blastocysts were excised from the oviducts of the ICM (inner cell mass) donor mouse and ICM cells were layered on top of a feeder cell layer of mouse embryonic fibroblasts in a culture dish with media supplemented by fetal calf serum. After various cell culture steps, a stable pluripotent ESC line was produced. To demonstrate germ-line transmission, ESCs can be injected into isolated blastocysts obtained from a mouse strain with a different coat color (such as C57B10/J, which has a black coat color that is recessive to the agouti color of the 129 strain). The resulting blastocysts can then be implanted into a pseudopregnant foster mother of the same strain as the donated blastocyst. Subsequent development of the introduced chimeric blastocyst results in chimeras with two populations of cells deriving from different zygotes (129 and C57B10/J in this case). The chimeras are readily identified because their coats have patches with different colors. (**B**) Germ-line transmission from chimeras. Breeding of male chimeras to C57B10/J mice can result in offspring with an agouti coat color, signifying a heterozygote: a sperm with a haploid strain 129 genome fertilized an egg with a haploid C57B10/J genome (the agouti coat color is dominant in the heterozygote).

Expanding the range of ESCs

The effectiveness of the cell culture system using fibroblast feeder cells and serum was heavily dependent on the genetic background of the inbred mouse strain; in practice, almost all stably pluripotent ESCs derived in this way are from the 129 mouse strain, or hybrids thereof. To isolate ESCs from other mouse strains and other mammals, there was

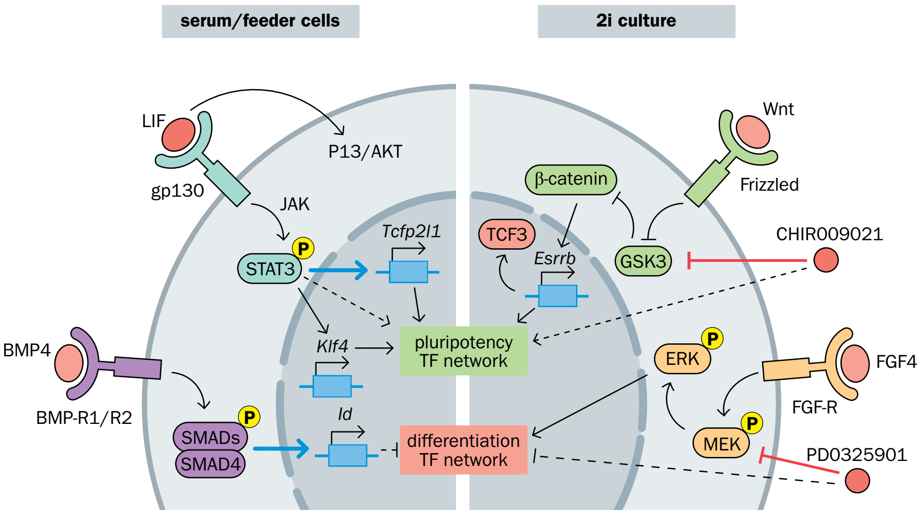


Figure 4.19 Extrinsic signaling pathways that feed into reinforcing or antagonizing naive pluripotency. Simplified schematic of various signaling cascades that affect self-renewal. Blue arrows indicate activation of indicated target, whereas red T-bars show inhibition or blockade. Solid lines indicate a direct or known downstream target; dashed lines indicate indirect/inferred effects. Small, filled red circles indicate small-molecule inhibitors. BMP4 is present in serum and functions via SMADs to activate *Id* genes that repress differentiation-promoting transcription factors. Leukemia inhibitory factor (LIF) signaling affects many pathways but primarily acts via JAK-mediated phosphorylation of STAT3, which activates *Tcfp2l1* and *Klf4*. Canonical Wnt signaling blocks GSK3 (glycogen synthase kinase-3) activity leading to stabilization of β-catenin, which in turn abrogates TCF3-mediated repression of pluripotency genes including *Esrrb*. CHIR009021 closely mimics Wnt signaling by inhibiting GSK3. FGF signaling activates the MAPK pathway leading to phosphorylation of MEK kinases, which in turn phosphorylate and activate ERK. Activated ERK promotes transition to a "primed" state of pluripotency that is blocked by the MEK inhibitor PD0325901. TF, transcription factor; -R, receptor. (Adapted from Hackett JA & Surani MA [2014] *Cell Stem Cell* **15**:416–430; PMID 25280218. With permission from Elsevier.)

a need to identify the extracellular signals (transmitted by feeder cells and serum) that promoted ESC self-renewal and suppressed differentiation. Leukemia inhibitory factor (LIF), which signals through the transcription factor STAT3, was quickly found to be an important factor, as was BMP4 (bone morphogenetic protein 4), which signals through SMAD transcription factors (**Figure 4.19**).

Subsequently, activation of the Wnt/β-catenin signaling pathway and inhibition of the FGF (fibroblast growth factor)/MAPK (mitogen-activated protein kinase) pathway were also found to be very important in maintaining renewal of ESCs while suppressing differentiation. A screen then identified chemical ways of manipulating these two pathways using two small-molecule inhibitors: PD0325901 inhibits MEK1/MEK2 (kinases that phosphorylate MAPK) and as a result inhibits FGF/MAPK signaling; CHIR009021 inhibits glycogen synthase kinase-3 (GSK3) to promote Wnt signaling through β-catenin (see **Figure 4.19**). The use of these two small-molecule inhibitors in culture conditions (known as 2i culture) does away with the need for feeder cells and serum by increasing the efficiency of keeping ESCs in a "ground state" of **naive pluripotency**. By using these conditions (sometimes with the addition of LIF), ESCs could be isolated from diverse mouse strains, and for the first time it was possible to develop rat ESCs, which were first reported in 2008, a full 27 years after the first mouse ESC lines.

Mouse EpiSC and EGC lines

Mouse epiblast stem cell (EpiSC) lines were established by culturing later-stage egg cylinder epiblasts. The culture conditions do not use LIF or 2i but instead use the fibroblast growth factor FGF2 and activin-A. Compared to ESCs, the cells are more heterogeneous and are associated with a state of pluripotency known as **primed pluripotency** that has certain disadvantages (**Table 4.2**).

Primordial germ cells that will normally develop into mature gametes can be isolated from the gonadal ridge of E7.5 mouse embryos and cultured *in vitro,* leading to pluripotent embryonic germ cell (EGC) lines that are virtually identical to ESCs.

Human pluripotent stem cell lines

Human EGCs, derived from primordial germ cells of embryos and fetuses from 5 to 10 weeks old, were first cultured in the late 1990s. To isolate human embryonic stem cell lines, surplus embryos arising from *in vitro* fertilization (which were donated with consent) were cultured to the blastocyst stage and the ICM cells were cultured.

The first human pluripotent stem cell lines derived from ICM cells of the blastocyst were reported in 1998, but as well as being difficult to manipulate, they had some very different properties from mouse ESCs. Now, it is widely accepted that they are the human counterparts of mouse EpiSCs (which were not reported until 2007). Like the mouse EpiSCs, the human "embryonic stem cell" lines exhibit primed pluripotency and associated features (see **Table 4.2**). Presumably, during the explant procedure, developmental progression continued further than anticipated.

In order to isolate genuine human counterparts of mouse ESCs, attempts were made to convert the human "embryonic stem cells" from primed pluripotency to naive pluripotency. However, culturing them in 2i medium alone (using inhibitors of MEK and GSK3;

TABLE 4.2 PROPERTIES OF TWO MAJOR PLURIPOTENCY STATES IN MAMMALIAN PLURIPOTENT STEM CELL LINES

Property	Naive pluripotency	Primed pluripotency[a]
Cell colony morphology	Dome shaped	Flattened
X-inactivation status in females[b]	XaXa	XaXi
DNA methylation	Very low levels	Low levels
Respiration	Oxidative phosphorylation plus glycolysis	Almost entirely glycolysis
Key cell signaling dependence	JAK-STAT signaling	TGFβ-activin A signaling
Enhancer used to regulate expression of the *OCT4* gene	Proximal enhancer	Distal enhancer
Cloning efficiency	High	Usually low
Ability to form chimeric blastocysts	Good	Little or none

[a] This pluripotency state is usually found in mouse epiblast stem cells, induced pluripotent stem cells, and in human "embryonic stem cells" (which really are the counterpart of mouse epiblast stem cells), but can be converted to naive pluripotency by culturing the cells under certain conditions (see text).
[b] Xa, active X chromosome; Xi, inactivated X chromosome. TGF, transforming growth factor.

see **Figure 4.19**) causes the cells to differentiate into cells resembling neural stem cells. Additional chemical compounds and/or growth factors were needed to supplement the medium (see Theunissen TW *et al*. [2014]; PMID 25090446, for an example).

Making pluripotent stem cell lines by epigenetic reprogramming of the genomes of differentiated cells

The previous section mostly dealt with making pluripotent stem cell lines by capturing pluripotency, harvesting pluripotent cells from the early embryo and finding ways of getting them to renew in culture without differentiating. Here we describe two alternative ways of making pluripotent stem cell lines that involve creating pluripotency by epigenetic reprogramming of the genomes of differentiated cells. That is, the pattern of DNA methylation and histone modifications in the genome of a differentiated cell is reset by artificial intervention; as a result, the chromatin structure is reset so that it resembles that found in pluripotent cells of the early embryo.

Somatic cell nuclear transfer (SCNT)

Although egg and sperm are highly-differentiated cells, they fuse to form a totipotent zygote. The cytoplasm of the egg contains factors that somehow naturally reprogram the DNA in the zygote: epigenetic marks such as repressive methylation signals are removed over large regions of the genome. **Somatic cell nuclear transfer** means artificially removing the nucleus of a differentiated somatic cell and placing it in an enucleated egg cell. Like the original nucleus of the egg cell, the introduced nucleus can be reprogrammed by factors in the egg cytoplasm. The egg cell with a somatic cell nucleus behaves like a zygote and can give rise to an adult organism.

When John Gurdon transferred nuclei of adult frog cells into enucleated eggs in the 1960s, he obtained a series of cloned adult frogs that were genetically identical to the original adult frog donor. But for decades afterward it was not possible to repeat this success with mammals. That is, until 1996, when a sheep called Dolly was born and became a world celebrity. The problem was that SCNT was highly inefficient in mammals. Refinements to the method were gradually introduced and a variety of other types of animal cloning followed. As detailed in Chapter 20, they have been used to introduce genetic modifications into animals to make disease models, such as a sheep model of cystic fibrosis, but they have also permitted the construction of human pluripotent stem cell lines (**Figure 4.20**).

Induced pluripotent stem cells (iPSC)

During development, transcription factors play key roles in determining the transition between different states of cell potency. As progenitor cells give rise to more differentiated cell lineages, the cells can simultaneously express transcription factors promoting differentiation to different cell lineages; they initially oppose each other, until one side

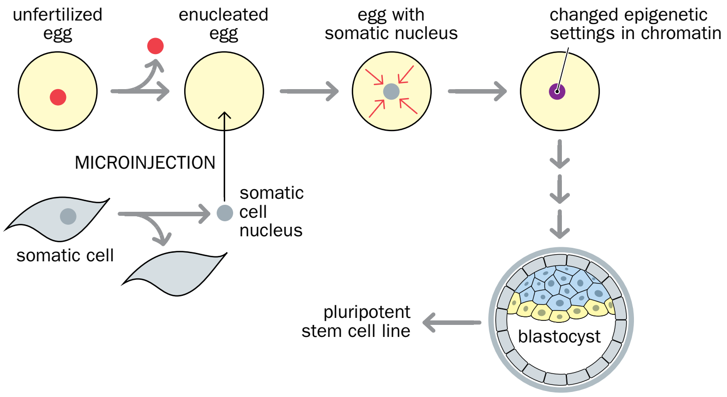


Figure 4.20 Constructing personalized pluripotent stem cell lines from a cloned blastocyst obtained by somatic cell nuclear transfer (SCNT). In SCNT, microsurgical techniques are used to remove a nucleus from an unfertilized egg and replace it with a nucleus taken from a differentiated somatic cell, such as a fibroblast. The introduced somatic cell nucleus has comparatively condensed chromatin but is epigenetically reprogrammed by cytoplasmic factors in the egg (red arrows) so that its pattern of chromatin conformation resembles that of a zygote. Continued development in culture allows production of a blastocyst that is genetically identical to the donor of the somatic cell. Inner cell mass cells from the blastocyst can be used to produce a pluripotent stem cell line.

triumphs in the tug of war. For example, cells of the undifferentiated ICM simultaneously express NANOG (promoting differentiation to epiblast) and the opposing GATA6 (promoting differentiation to the hypoblast, or primitive endoderm). Cells in which NANOG triumphs go on to become epiblast cells; those in which GATA6 triumphs become hypoblast cells.

If transcription factors drive pathways toward increasing differentiation, a question arises: Can they drive pathways in the opposite direction, toward reduced differentiation (**dedifferentiation**)? That question became increasingly relevant to mammalian cells after the birth of Dolly the cloned sheep: at last there was proof that mammalian cells could be epigenetically reprogrammed toward dedifferentiation, all the way to totipotency. But SCNT is technically challenging and time consuming. Could somatic differentiated cells be reprogrammed to pluripotency simply by exposing cultured cells to appropriate transcription factors? The answer was yes, and perhaps surprisingly, only four transcription factors were found to be needed (**Figure 4.21**). They included OCT4 and SOX2, two of the three master transcription factors that regulate pluripotency *in vivo* in the ICM, but not the third such factor, NANOG (expression of OCT4 and SOX2 is sufficient because they work together to up-regulate NANOG; see **Figure 4.21A**). KLF4 up-regulates OCT4 expression, and, like MYC, it is important in self-renewal of embryonic stem cells.

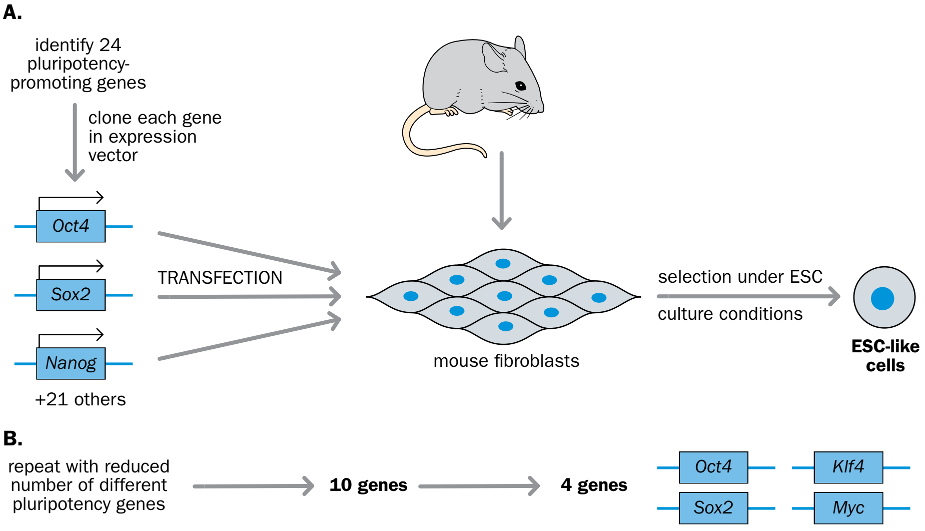

A.
identify 24 pluripotency-promoting genes
clone each gene in expression vector
Oct4
Sox2
Nanog
+21 others
TRANSFECTION
mouse fibroblasts
selection under ESC culture conditions
ESC-like cells
B.
repeat with reduced number of different pluripotency genes
10 genes
4 genes
Oct4
Klf4
Sox2
Myc

Figure 4.21 Differentiated cells can be converted to pluripotency by artificially expressing just four transcription factors. Shinya Yamanaka and colleagues reported an astounding breakthrough in 2006 (Takahashi K & Yamanaka S [2006]; PMID 16904174) when they were able to regress cultured mouse fibroblasts to a pluripotent state by transfecting genes encoding just four types of transcription factor. (**A**) The starting point was 24 genes known to be important in pluripotency. They were individually cloned into retroviral expression vectors, transfected into mouse fibroblasts, and the recombinants were then cultured under conditions used to support embryonic stem cell (ESC) cultures. Pluripotent ESC-like cells could be observed because of their distinctive morphology (round, large nucleus containing large nucleoli, a thin rim of cytoplasm). (**B**) Not all of the 24 genes were necessary for this effect: by withdrawing some and testing different combinations, 10 genes were identified as being more important, and from this smaller set just four genes were found to be required to induce pluripotency. The four transcription factors, OCT4 (historically called Oct-3/4), SOX2, KLF4, and MYC, are sometimes known as Yamanaka factors or OSKM (from their initials).

This new type of pluripotent stem cell line came to be known as **induced pluripotent stem cells** (**iPSCs**). In 2009, viable fertile mice were reported that originated exclusively from mouse iPSCs: when introduced into blastocysts, the iPSCs can colonize the embryo to produce chimeras, and contribute to the germ line as well as to all of the somatic tissues. Although the mouse iPSCs resembled ESCs, they did exhibit some differences, notably showing the primed pluripotency reminiscent of epiblast stem cells (see **Table 4.2**). Subsequently, iPSCs have been made from more than 20 mammalian species, including humans, and different ways have been used to induce pluripotency of differentiated cells, including using purified transcription factors (instead of genes that express them), miRNAs, and small synthetic hydrocarbon molecules (selected after screening for their ability to interfere with signaling pathways).

Personalized human pluripotent stem cells

Human pluripotent stem cell lines made by epigenetic reprogramming methods have one major advantage over those made by culturing cells from the early embryo: they can be made from any individual who wishes to donate easily accessible cells, such as skin fibroblasts. Pluripotent stem cell lines can therefore be made from patients and directed to differentiate to provide cells that are not readily accessible (such as neurons) for modeling disease and for drug screening. There is also the possibility of correcting genetic defects in iPSCs prepared from a patient with a genetic disorder, deriving suitable progenitor cells, and re-introducing them into the patient so that they can be induced to differentiate into the desired cells. We consider the potential in this area in Chapter 22.

Artificially directing differentiation and engineering transdifferentiation of human cells

When human embryonic stem cell lines were first analyzed it became clear that they were naturally prone to differentiate in certain ways. Because they are capable of giving rise to any type of body cell, there was the possibility of artificially directing cellular differentiation *in vitro*. To direct differentiation toward a desired cell type, genes encoding suitable lineage-specific transcription factors could be transfected into the cells and overexpressed. A new era of regenerative medicine was envisaged where cells could be instructed to change into other cells, both *in vitro* and *in vivo*.

The idea of forcibly directing differentiation by overexpressing an appropriate transcription factor dates back to pioneering work by Harold Weintraub and colleagues in the late 1980s. They transfected mouse cultured fibroblasts with a cDNA encoding MYOD, and by overexpressing just this one transcription factor, were able to convert the fibroblasts into myoblasts. That result, although attracting considerable attention, still appeared unusual: there was considerable resistance to the idea that the epigenetic settings that determined the identity of a cell could easily be reset. Even by 2006, 10 years after the birth of Dolly the cloned sheep, the discovery that induced pluripotent stem cells could be created by overexpressing just four transcription factors was met with widespread amazement.

Since then there has been a paradigm shift: the idea that irreversible epigenetic marks are laid down as cells travel along the pathways of cell specialization is one whose time has come and gone. In hindsight, we should not have been too surprised: there are extraordinary examples of tissue regeneration in some species, and even in humans natural epigenetic reprogramming occasionally causes cells to change identity (Figure 4.22).

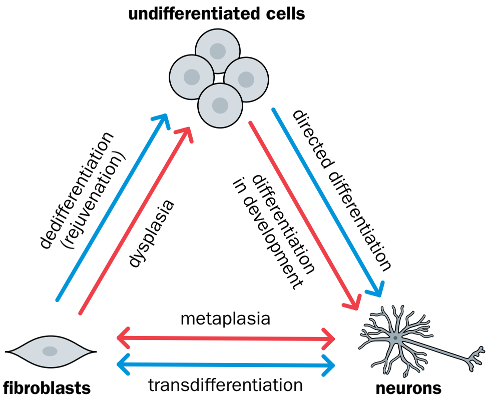


Figure 4.22 Modes of epigenetic reprogramming to change cell identity. Red arrows signify naturally occurring modes of epigenetic reprogramming; blue arrows indicate artificial epigenetic reprogramming. Metaplasia—conversion of a differentiated cell to another of a different type—is common in some organisms but also occurs naturally in humans when cells are subject to extended physiological or pathological stress. Prolonged exposure to cigarette smoke, for example, can convert pseudostratified columnar epithelial cells of the airways into squamous epithelial cells, and gastroesophageal reflux causes squamous epithelial cells of the esophagus to convert to columnar epithelial cells. Artificial transdifferentiation can make larger changes in mammals, for example convertibility of fibroblasts and neurons. Dysplasia entails expansion of immature cells at the expense of more differentiated cells and is common in cancer where epigenetic changes cause cells to revert to undifferentiated states. (Adapted from Cherry A & Daley G [2012] *Cell* **148**:1110–1122; PMID 22424223. With permission from Elsevier.)

Directed transdifferentiation

If artificial epigenetic reprogramming were to be used for therapeutic purposes, then inducing dedifferentiation to form a pluripotent stem cell followed by differentiation of the pluripotent cells to a suitable tissue progenitor cell might not seem to be the most efficient route. To replace insulin-producing pancreatic β cells, for example, it might be simpler to convert other pancreatic cells *in vivo*, or one might try to convert patient fibroblasts *in vitro* to some tissue progenitor cells that might simply then be directed toward the desired cell type *in vivo*.

Significant effort has therefore gone into changing the identity of a differentiated cell toward another desired cell type, a process known as **transdifferentiation**. The example of converting fibroblasts to myoblasts given above required a single transcription factor, but a variety of different transdifferentiations have been carried out and often require two or more transcription factors. And transdifferentiation has been possible between the three germ layers—ectoderm, endoderm, and mesoderm—as well as between cell types belonging to one germ layer (see **Figure 4.23** for some examples). We consider the potential therapeutic applications in Chapter 22.

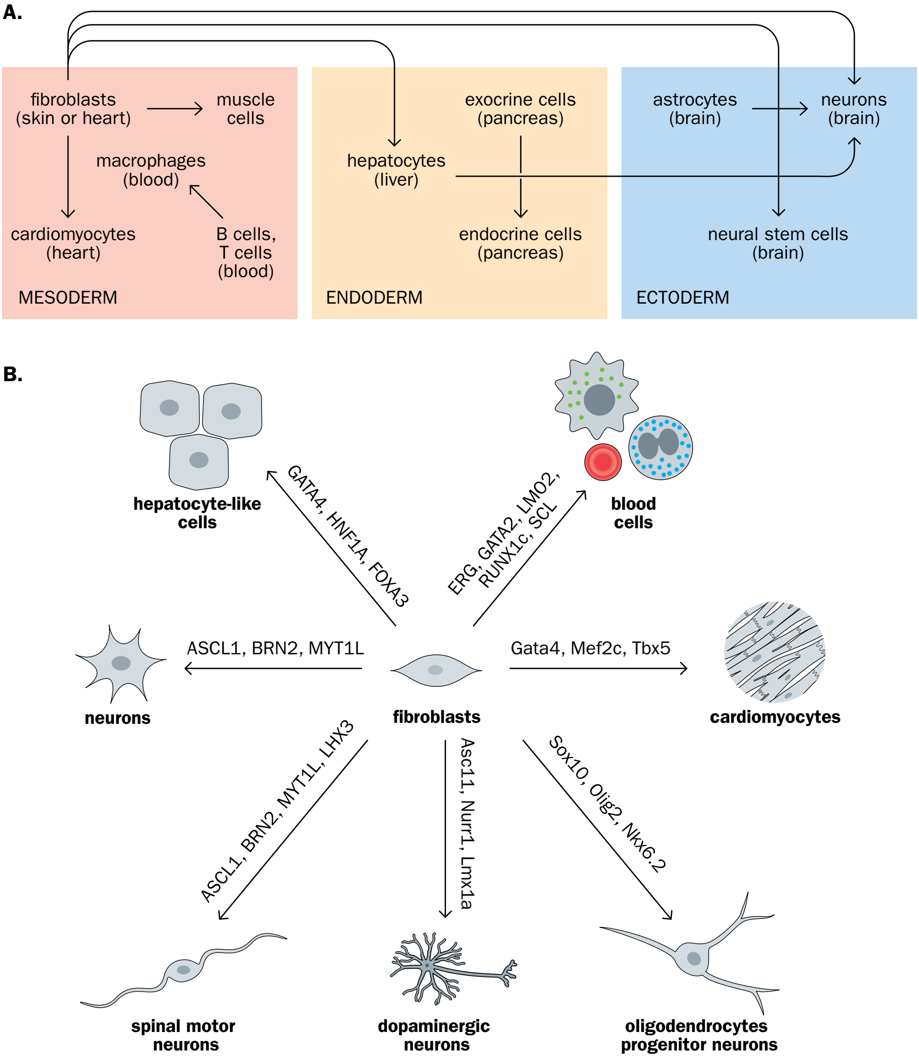


Figure 4.23 Transdifferentiation by overexpression of transcription factors. (**A**) Examples of successful transdifferentiation of human and mouse cells, including conversion of cells belonging to different germ layers and to cells belonging to the same germ layer. (**B**) Examples of how mammalian fibroblasts can be programmed by combinations of transcription factors to give different cell types. As described in the following article, various miRNAs have also been overexpressed to direct transdifferentiation. (A, From Ladewig J *et al.* [2013] *Nat Rev Mol Cell Biol* **14**:225–236; PMID 23847783. With permission from Springer Nature. Copyright © 2013; B, Adapted from Wang H *et al.* [2015] *Differentiation* **90**:69–76; PMID 26525508. With permission from Elsevier.)

SUMMARY

- The fertilized egg (zygote) and each cell in very early stage mammalian embryos (up to the 16-cell stage in mouse) are totipotent; they can each give rise to every type of adult cell and to all the different extra-embryonic cells need to support embryonic development.
- As development proceeds, the choice of cell fate narrows, and cells become progressively more specialized (differentiated). At maturity most cells are terminally differentiated.
- All cells in an animal embryo have the same DNA molecules, but epigenetic events occurring very early in development cause differences between cells leading to different cell lineages, then formation of different tissues and further cell differentiation.
- Epigenetic influences on animal embryonic development include symmetry-breaking events and stochastic factors. Intrinsic asymmetry can arise in some animal embryos because of asymmetric positioning of fate-determining proteins in the maternal egg (the daughter cells inherit different amounts of the protein).
- In the early vertebrate embryo, choice between alternative cell fates primarily depends on the position of a cell and its

interactions with other cells, rather than cell lineage. Outer cells in an early embryo may, for example, receive different external chemical signals than those located in the interior.

- In mammalian development, the early embryo gives rise not only to every type of adult cell, but also to the cells of four extra-embryonic membranes that support the developing embryo and fetus, and to the fetal component of the placenta.
- Overt asymmetry is apparent in the early mammalian embryo. In the late eight-cell stage, mouse embryo cells become polarized and certain key regulatory proteins become concentrated at the apical ends of cells. When the cells divide the daughter cells can inherit different amounts of the regulatory proteins.
- In the 16-cell human embryo, distinctly different outer and inner cell layers are evident. Afterward, a cavity forms and fluid is secreted from outer cells into the interior to form a fluid-filled hollow ball of cells (the blastocyst, first evident at the 32 cell-stage in humans).
- The inner cells of the blastocyst (inner cell mass) are pluripotent: they give rise to every type of adult cell plus some cells of the extraembryonic membranes. The outer (trophoblast) cells give rise to many cells of the extraembryonic membranes and the fetal component of the placenta.
- Gastrulation is a key morphogenetic process in early vertebrate development. Rapid cell migrations cause drastic restructuring of the embryo to form three types of somatic cell lineage (germ layers)—ectoderm, mesoderm, and endoderm—that will be precursors of defined tissues of the body.
- The germ cell lineage (germ line) ultimately gives rise to sperm and cells, and is potentially immortal. The earliest committed progenitors, primordial germ cells, are physically segregated from somatic cell lineages and need to migrate as the embryo develops, ultimately taking up residence in the developing gonads.
- By escaping from a somatic cell fate, early primordial germ cells retain the potential to be totipotent.
- Primary sexual characteristics (gonad development and the choice of sperm or egg development) are determined by the genotype with the male-determining Y-linked *SRY* gene usually having a dominant role.
- Stem cells are comparatively unspecialized cells that can divide to generate daughter stem cells (stem cell renewal) and daughter progenitor cells (which are committed to producing more differentiated descendant cells).
- Tissue-specific stem cells are often multipotent. They produce various types of differentiated cells to replace equivalent cells lost from the tissue through natural cell turnover. They are located in protective microenvironments (stem cell niches) where they receive signals from nearby cells and extracellular matrix that support stem cell renewal.
- Immortal pluripotent embryonic stem cell (ESC) lines can be constructed by artificial manipulation in culture of naturally pluripotent cells obtained from the inner cell mass of blastocysts. Suitably manipulated ESCs can be directed to differentiate to give desired cell types found in the mature organism.
- Cell reprogramming means changing the epigenetic settings of cells. Cultured cells are manipulated to change their potency or cell type, often by exposing the cells to appropriate transcription factors.
- Terminally differentiated somatic cells can be reprogrammed so that they become pluripotent (dedifferentiation). The resulting induced pluripotent stem cells (iPSCs) resemble pluripotent stem cells and can be directed to differentiate to desired cell types. Unlike ESCs, iPSCs can be prepared readily from any individual, allowing studies of pathogenesis in all individuals with genetic disorders.
- Somatic cells can also be artificially reprogrammed to become a different type of somatic cell; fibroblasts can be directed to become neurons, for example (transdifferentiation).

FURTHER READING

General developmental biology

Gilbert SF & Barresi (2016) *Developmental Biology*, 11th edn. Sinauer Associates.

Slack JMW (2012) *Essential Developmental Biology*, 3rd edn. Wiley-Blackwell.

Wolpert L & Tickle C (2015) *Principles of Development*, 5th edn. Oxford University Press.

Human embryology and early development

embryology.ch at http://www.embryology.ch/. Online educational resource covering human embryology and organogenesis. Available in English, Dutch, French, and German.

Larsen WJ (2001) *Human Embryology*, 3rd edn. Churchill Livingstone.

The Virtual Human Embryo at http://virtualhumanembryo.lsuhsc.edu/.

UNSW Embryology. University of New South Wales. http://embryology.med.unsw.edu.au (Online educational resource for teaching concepts in embryology.)

Compaction, cell adhesion, and cell polarity in the early mammalian embryo

Nance J (2014) Getting to know your neighbor: cell polarization in early embryos. *J Cell Biol* **206**:823–832; PMID 25267293.

White MD & Plachta N (2015) How adhesion forms the early mammalian embryo. *Curr Top Dev Biol* **112**:1–17; PMID 25733136.

Ziomek CA & Johnson MH (1980) Cell surface interaction induces polarization of mouse 8-cell blastomeres at compaction. *Cell* **21**:935–942; PMID 7438209.

Cell fate specification in the early mouse embryo

Boroviak T & Nichols J (2014) The birth of embryonic pluripotency. *Philos Trans R Soc Lond B Biol Sci* **369**:20130541; PMID 25349450.

Frum T & Ralston A (2015) Cell signaling and transcription factors regulating cell fate during formation of the mouse blastocyst. *Trends Genet* **31**:402–410; PMID 25999217.

Leung CY & Zernicka-Goetz M (2015) Mapping the journey from totipotency to lineage specification in the mouse embryo. *Curr Opin Genet Dev* **34**:71–76; PMID 26343010.

Sasaki H (2015) Position- and polarity-dependent Hippo signaling regulates cell fates in preimplantation mouse embryos. *Semin Cell Dev Biol* **47**-48:80–87; PMID 25986053.

Wennekamp S *et al*. (2013) A self-organization framework for symmetry breaking in the mammalian embryo. *Nat Rev Mol Cell Biol* **14**:452–459; PMID 23778971. (Includes an account of the three classical models for explaining lineage segregation in the early embryo.)

Germ cells and sex determination in mammals

Günesdogan U *et al*. (2014) Primordial germ cell specification: a context-dependent cellular differentiation event. *Philos Trans R Soc Lond B Biol Sci* **369**:20130543; PMID 25349452.

Hayashi K *et al.* (2007) Germ cell specification in mice. *Science* **316**:394–396; PMID 17446386.

Saitou M & Yamaji M (2012) Primordial germ cells in mice. *Cold Spring Harb Perspect Biol* **4**:a008375; PMID 23125014.

Wilhelm D *et al*. (2007) Sex determination and gonadal development in mammals. *Physiol Rev* **87**:1–28; PMID 17237341.

Stem cells (general)

NIH Stem Cell Information at https://stemcells.nih.gov/

Adult (tissue-specific) stem cells

Barker N *et al*. (2010) Tissue-resident adult stem cell populations of rapidly self-renewing organs. *Cell Stem Cell* **7**:656–670; PMID 21112561.

Blanpain C & Fuchs E (2014) Stem cell plasticity. Plasticity of epithelial stem cells in tissue regeneration. *Science* **344**:1242281; PMID 24926024.

Clevers H *The Intestinal Crypt: a clonal conveyor belt*. Available at https://www.hubrecht.eu/onderzoekers/clevers-group/ (accessed 02/07/18).

Clevers H & Watt FM (2018) Defining adult stem cells by function, not by phenotype. *Annu Rev Biochem* **87**:1015–1027; PMID 29494240.

Clevers H *et al*. (2014) An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control. *Science* **346**:1248012; PMID 25278615.

Lane SW *et al*. (2014) Modulating the stem cell niche for tissue regeneration. *Nat Biotechnol* **32**:795–803; PMID 25093887.

Simons BD & Clevers H (2011) Strategies for homeostatic stem cell self-renewal in adult tissues. *Cell* **145**: 851–862; PMID 21663791.

Embryonic stem cells and pluripotency states

De Los Angeles A *et al*. (2015) Hallmarks of pluripotency. *Nature* **525**:469–478; PMID 26399828.

Hackett JA & Surani MA (2014) Regulatory principles of pluripotency: from the ground state up. *Cell Stem Cell* **15**:416–430; PMID 25280218.

Martello G & Smith A (2014) The nature of embryonic stem cells. *Annu Rev Cell Dev Biol* **30**:647–675; PMID 25288119.

Nichols J & Smith A (2011) The origin and identity of embryonic stem cells. *Development* **138**:3–8; PMID 21138972.

Wu J & Izpisua Belmonte JC (2015) Dynamic pluripotent stem cell states and their applications. *Cell Stem Cell* **17**:509–525; PMID 26544113.

Induced pluripotent stem cells and epigenetic reprogramming

Ladewig J *et al*. (2013) Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nat Rev Mol Cell Biol* **14**:225–236; PMID 23847783.

Sánchez Alvarado A & Yamanaka S (2014) Rethinking differentiation: stem cells, regeneration, and plasticity. *Cell* **157**:110–119; PMID 24679530.

Takahashi K & Yamanaka S (2015) A developmental framework for induced pluripotency. *Development* **142**:3274–3285; PMID 26443632.

Wilmut I *et al*. (2015) Somatic cell nuclear transfer: origins, the present position and future opportunities. *Philos Trans R Soc Lond B Biol Sci* **370**:20140366; PMID 26416677.

5 Patterns of inheritance

Genetics as a science started with Gregor Mendel's experiments in the 1860s. His original paper, *"Versuche über Pflanzen-Hybriden"* (*"Experiments in plant hybridization"*), can be read, in the original German and in English translation, at http://www.mendelweb.org/Mendel.html. Mendel bred and crossed pea plants, counted the numbers of plants in each generation that manifested certain traits, and deduced mathematical rules to explain his results. One should not fall for a romantic picture of the simple monk in his monastery garden stumbling on the laws of genetics. Mendel's monastery, now in the Czech Republic, was an intellectual powerhouse, and Mendel was recruited to work on a preexisting program of genetics research, albeit focused on agricultural improvement. Writing in 1865, Mendel knew nothing of chromosomes, still less of DNA, but we now understand that the patterns he identified are the consequence of the way chromosomes segregate in meiosis. They apply to any character that is determined by a DNA sequence at a single, fixed chromosomal location in a sexually reproducing diploid organism.

Our first task in this chapter is to define a number of terms, none of which was used by Mendel.

A **locus** (plural **loci**) is a unique chromosomal location defining the position of an individual gene or DNA sequence. Thus, we can speak of the ABO blood group locus, the Rhesus blood group locus, and so on.

Alleles are alternative versions of a gene. For example, A, B, and O are alternative alleles at the ABO locus.

The **genotype** is a list of the alleles present at one or several loci.

Phenotypes, **characters**, or **traits** are the observable properties of an organism. The means of observation may range from simple inspection to sophisticated laboratory investigations.

A person is **homozygous** at a locus if both alleles at that locus are the same, and **heterozygous** if they are different. For different purposes we may check more or less carefully whether the two alleles at a locus really are the same. For looking at patterns of inheritance in this chapter, we are concerned only with the phenotypic consequences of a person's genotype. We will describe a person as homozygous if both alleles at the locus in question have the same phenotypic effect (wild-type or mutant), even though inspection of the DNA sequences might reveal differences between them. For various aspects of population and evolutionary genetics covered in later chapters one would use a definition based on DNA sequence, and only describe a person as homozygous if the two alleles had identical sequences.

We have left to the end the most difficult term to define. A **gene** might be defined in either of two ways:

- A gene is a determinant, or a co-determinant, of a character that is inherited;
- A gene is a functional unit of DNA.

Human genes have uppercase italicized names such as *CFTR* or *CDKN2A*. The process of discovery often involved several competing research groups, so that the same gene may initially be referred to by several different names. Eventually an official name is assigned by the HUGO (Human Genome Organisation) Nomenclature Committee, and this is the name that should be used henceforth. Official names can be found on the nomenclature website (http://www.genenames.org) or from entries in OMIM (described below). The formal nomenclature for alleles uses the gene name followed by an asterisk and then the allele name or number, for example, *PGM*1*, *HLA-A*31*. However, the rules for nomenclature of alleles are less rigorously adhered to in publications than those for genes. For pedigree interpretation, alleles are usually simply denoted by upper- and lowercase versions of the same letter, so that the genotypes at a locus would be written *AA*, *Aa*, or *aa*. By convention the uppercase letter is used for the allele that determines the dominant character (see below).

A major aim of human molecular genetics is to understand how genes as functional DNA sequences determine the observable characters of a person. This is the subject of later chapters, especially Chapter 16. In this chapter we are concerned with the patterns of inheritance: how the segregation of alleles at meiosis determines the transmission of characters in pedigrees.

5.1 MONOGENIC VERSUS MULTIFACTORIAL INHERITANCE

The simplest genetic characters are those whose presence or absence depends on the genotype at a single locus. That is not to say that the character itself is programmed by only one pair of genes: expression of any human character is likely to depend on the action of a large number of genes and environmental factors. However, sometimes a particular genotype at one locus is both necessary and sufficient for the character to be expressed, given the normal range of human genetic and environmental backgrounds. Such monogenic characters are called **Mendelian** because their pattern of inheritance follows that established by Gregor Mendel.

Mendelian characters can be recognized by the characteristic pedigree patterns they give, as described in the next section. The best starting point for acquiring information on any such character, whether pathological or nonpathological, is the Online Mendelian Inheritance in Man (OMIM) database, https://www.ncbi.nlm.nih.gov/omim. OMIM contains about 24,000 entries, which may be sequenced genes, characters or diseases associated with known sequenced genes, or characters that are inherited in a Mendelian way but for which no gene has yet been identified. Some entries describe Mendelian subsets of characters that are not in aggregate Mendelian—breast cancer, for example. Users need to be aware that in those cases the OMIM entry will therefore not give a balanced picture of the overall etiology. Each entry is a detailed review, usually historically ordered, of the genetics of a character or the history and function of a gene, with subsidiary clinical and other information, and a very useful list of references. Entries have accumulated text over many years with only patchy rewrites, so that the early part of an entry may not reflect current understanding. Where appropriate, throughout this book the OMIM reference number is quoted for each human character or gene when it is first described.

Most human genetic or partly genetic characters are not Mendelian. They are governed by genes at more than one locus. The more complex the pathway between a DNA sequence and an observable trait, the less likely it is that the trait will show a simple Mendelian pedigree pattern. Thus, DNA sequence variants are almost always inherited in a cleanly Mendelian manner. Disease states or other traits that reflect the biochemical action of a protein within a cellular or organismal context are less often entirely determined by alleles at a single genetic locus. The failure or malfunction of a developmental pathway that results in a birth defect is likely to involve a complex balance of factors. Thus, the common birth defects such as cleft palate, spina bifida, or congenital heart disease are rarely overall Mendelian, though there may be Mendelian subsets where major malfunction at a single locus derails an entire pathway. Behavioral traits such as IQ test performance or schizophrenia are still less likely to be Mendelian—but they may still be genetically determined to a greater or lesser extent.

Non-Mendelian characters may depend on two, three, or many genetic loci. We use **multifactorial** here as a catch-all term covering all these possibilities. More specifically, the genetic determination may involve a small number of loci (**oligogenic**) or many loci each of individually small effect (**polygenic**); or there may be a single major locus with a polygenic background—that is, the genotype at one locus has a major effect on the phenotype, but this effect is modified by the cumulative minor effects of genes at many other loci. In reality, characters form a continuous spectrum, from perfectly Mendelian through to truly polygenic (**Figure 5.1A**). Superimposed on this there may be a greater or smaller effect of environmental factors. The overall etiology of a character could be represented by a point somewhere within the triangle of **Figure 5.1B**.

Mendelian characters are necessarily **dichotomous**: characters such as cystic fibrosis or extra fingers that you either have or do not have. Most human characteristics are not dichotomous. Think of the way you would describe a person so as to enable a friend to recognize them. Apart from sex (in most cases) and maybe hair and eye color, almost none of the characters you would describe are dichotomous. Most are **continuous** or **quantitative characters** such as height or weight—characters we all have, but to different degrees.

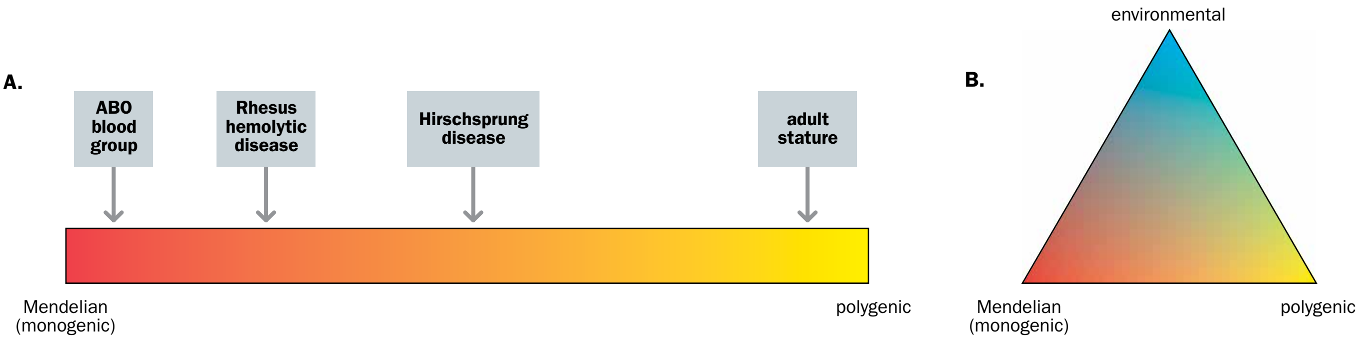


Figure 5.1 Determinants of a phenotype. (**A**) ABO blood group depends (with rare exceptions) on the genotype at just one locus, the *ABO* locus at chromosome 9q34 (see **Box 15.1** for an explanation of this nomenclature). Rhesus hemolytic disease of the newborn depends on the genotypes of mother and baby at the *RHD* locus at chromosome 1p36, but also on mother and baby being ABO compatible. Hirschsprung disease depends on the interaction of several genetic loci. Adult stature is determined by the cumulative small effects of many loci. (**B**) Nongenetic factors, collectively termed environmental, are important for many phenotypes, for example Hirschsprung disease and adult stature. The overall etiology of a character could be represented as a point somewhere within the triangle.

Such characters are necessarily non-Mendelian, but that is not to say that genes can have no role in determining them. The underlying loci are described as **quantitative trait loci** (QTLs). Dichotomous characters can also be non-Mendelian but wholly or partly genetically determined: they may tend to run in families, but the pedigrees do not fit any standard Mendelian pattern. The genetic factors may be described as **susceptibility genes**. QTLs and susceptibility genes are not different at the molecular level from Mendelian genes, just the characters concerned are determined in a more complex way. Variants in the same gene may be a Mendelian determinant of one phenotype and a QTL for another.

In a further layer of complication, a common human condition such as diabetes is likely to be very heterogeneous in its causation. Some cases may have a simple Mendelian cause, some might be entirely the result of environmental factors, while the majority of cases may be multifactorial. Such conditions are called **complex**.

5.2 MENDELIAN PEDIGREE PATTERNS

When the presence, absence, or specific nature of a character is normally determined by variation at a single chromosomal location, the character is called monogenic or Mendelian. If a monogenic character is such that it can be transmitted through a family (that is, it does not pose a major obstacle to successful reproduction), it may show a characteristic pedigree pattern. The patterns are the result of the way chromosomes segregate during meiosis, and any heritable character that is determined by the DNA or chromatin at one fixed chromosomal location can follow them. Mendel had no idea what his “determinants” might actually be physically, and it remains the case that determinants of Mendelian characters do not need to be genes in the traditional sense of protein-coding sequences—all that matters is that they should occupy a single, fixed chromosomal location. Facioscapulohumeral muscular dystrophy, an autosomal dominant Mendelian character, is a good example of that (see OMIM #158900). It is nevertheless true that the determinants of most Mendelian characters are variants in, or affecting, protein-coding sequences.

Figure 5.2 shows the symbols commonly used for pedigree drawing. Ideally, each generation is drawn on a separate horizontal line (see examples in **Figures 5.3–5.10**). Generations are usually labeled with Roman numerals, and individuals within each generation with Arabic numerals. Thus, III-7 or III_7 is the seventh person from the left (unless explicitly numbered otherwise) in generation III. An arrow can be used to indicate the **proband** or **propositus** (female: proposita) through whom the family was ascertained.

There are four basic Mendelian pedigree patterns illustrated in **Figures 5.3–5.6** and described in **Box 5.1**. A character can be autosomal or X-linked, depending on the chromosomal location of the relevant gene, and it can be dominant or recessive. A character is **dominant** if it is evident in a heterozygous person, **recessive** if not. Note that dominance and recessiveness are properties of characters, not genes or alleles; the tendency to speak of a “dominant gene,” and so on, should be resisted as far as possible. The reason why some monogenic characters are dominant and others recessive is discussed by Wilkie (1994) (PMID 8182727; see Further Reading); see also **Figure 16.17**. Y-linked characters would form a fifth class (see **Figure 5.7**), but the Y chromosome carries very few genes, and the main pathogenic effect of variants is to cause male infertility. Thus, the only Y-linked character that commonly gives an extended pedigree pattern is maleness itself, due to the *SRY*

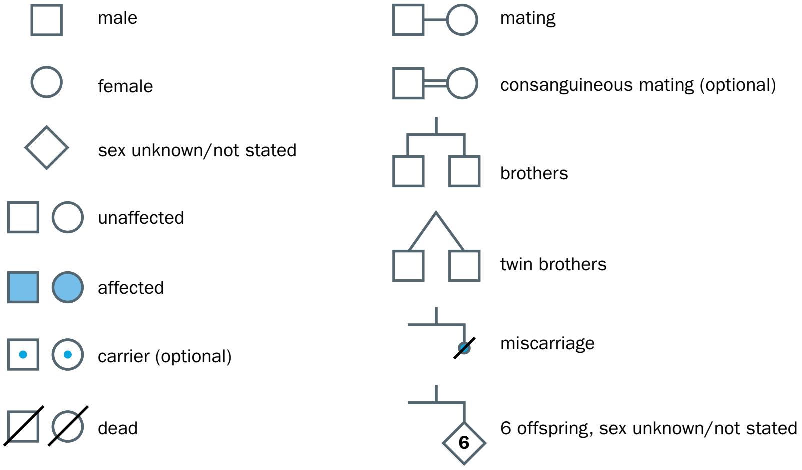


Figure 5.2 Main symbols used in pedigrees. The dot symbol for a carrier, and the double marriage lines for consanguineous matings, are optional. They are used to draw attention to those features, but their absence does not necessarily mean that a person is not a carrier, or that a union is not consanguineous.

gene on the Y chromosome. Various characters, such as hairy ears, have been unconvincingly described as Y-linked; possibly the only genuine example, *DFNY1* (OMIM #400043), is shown in **Figure 5.7**. Characters determined by variations in the mitochondrial DNA (see Chapter 9) show yet another pattern (**Figure 5.8**). These basic patterns are subject to various complications that are discussed below and illustrated in **Figures 5.9–5.15**.

BOX 5.1 SUMMARY OF BASIC PATTERNS OF INHERITANCE

AUTOSOMAL DOMINANT INHERITANCE (FIGURE 5.3)

- An affected person usually has at least one affected parent (but exceptions are due to new mutations or non-penetrance, see **Figures 5.11** and **5.17**).
- It affects either sex.
- It is transmitted by either sex.
- A child with one affected and one unaffected parent has a 50% chance of being affected (this assumes that the affected person is heterozygous, which is usually true for rare conditions).

AUTOSOMAL RECESSIVE INHERITANCE (FIGURE 5.4)

- Affected people are usually born to unaffected parents.
- Parents of affected people are usually asymptomatic carriers.
- There is an increased incidence of parental consanguinity (see Section 12.4).
- It affects either sex.
- After the birth of an affected child, each subsequent child has a 25% chance of being affected (assuming that both parents are heterozygous carriers).

X-LINKED RECESSIVE INHERITANCE (FIGURE 5.5)

- It affects mainly males.
- Affected males are usually born to unaffected parents; the mother is normally an asymptomatic carrier but may have affected male relatives.
- Females may be affected if the father is affected and the mother is a carrier, or occasionally as a result of non-random X-inactivation (see Section 10.4).
- There is no male-to-male transmission in the pedigree (but matings of an affected male and carrier female can give the *appearance* of male-to-male transmission; see **Figure 5.14**).

X-LINKED DOMINANT INHERITANCE (FIGURE 5.6)

- It affects either sex, but more females than males.
- Usually at least one parent is affected.
- Females are often more mildly and more variably affected than males (because of X-inactivation; see Section 10.4).
- The child of an affected female, regardless of its sex, has a 50% chance of being affected.
- For an affected male, all his daughters but none of his sons are affected.

Y-LINKED INHERITANCE (FIGURE 5.7)

- It affects only males.
- Affected males always have an affected father (unless there is a new mutation).
- All sons of an affected man are affected.

MITOCHONDRIAL INHERITANCE (FIGURE 5.8)

- It affects both sexes.
- It is usually inherited from an affected mother (but is often caused by *de novo* mutations, with the mother unaffected).
- It is not transmitted by a father to any of his children.
- Clinical manifestations are often highly variable.

Autosomal characters in both sexes, and X-linked characters in females, can be classified as dominant or recessive by observing the phenotype of a heterozygote. Males are **hemizygous** for loci on the X and Y chromosomes; that is, they have only a single copy of the DNA at each locus. Thus, chromosomally normal men are never heterozygous for any X-linked or Y-linked character and the concepts of dominance and recessiveness do not apply. In the rare XYY males (see Chapter 15), the two Y chromosomes are duplicates, although the two X chromosomes in most XXY men are not.

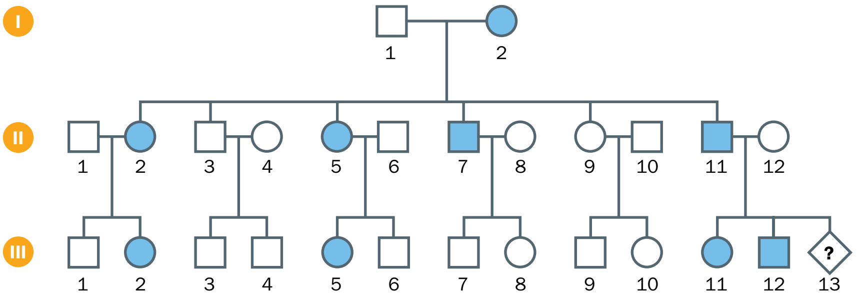


Figure 5.3 An ideal autosomal dominant pedigree. Affected people are heterozygotes. The risk of being affected for the person marked with a query is 1 in 2. Real pedigrees of human dominant characters often show irregularities, as shown in **Figures 5.9** and **5.10**.

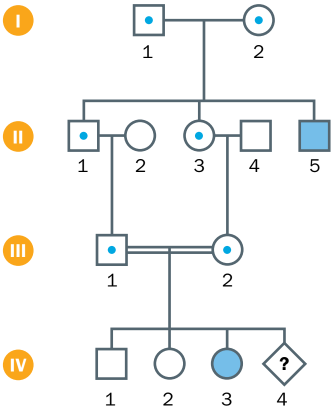


Figure 5.4 Pedigree of an autosomal recessive character. People who must be carriers are indicated with dots; IV$_1$ and/or IV$_2$ might also be carriers, but we do not know. Note the double marriage line, drawing attention to the consanguineous mating. The risk for the individual marked with a query is 1 in 4.

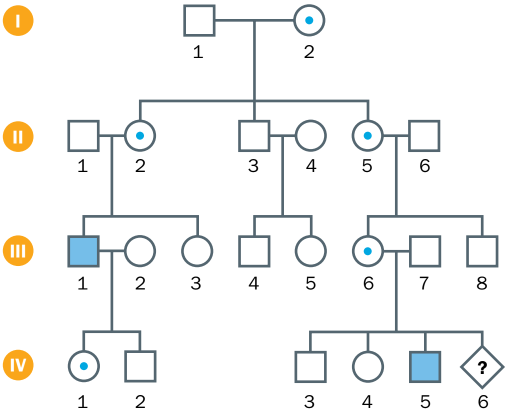


Figure 5.5 Pedigree pattern of an X-linked recessive condition. The females marked with dots are definite (obligate) carriers; individuals III$_3$ and/or IV$_4$ may also be carriers, but we do not know. The risk that the individual marked with a query would be affected is 1 in 2 (if male) or 1 in 4 of all offspring, regardless of sex.

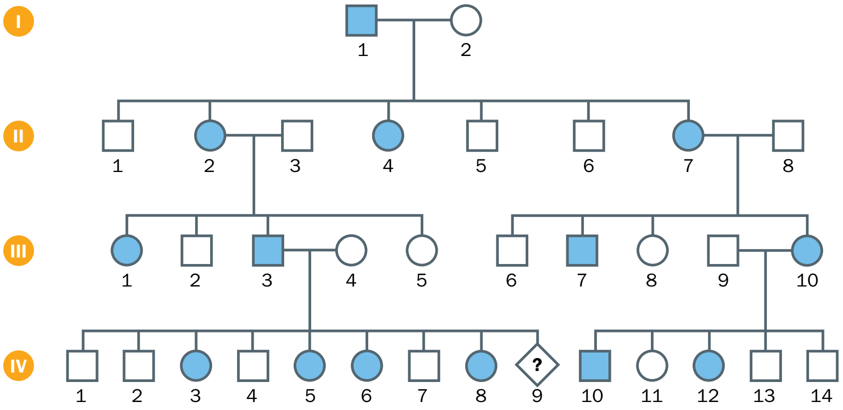


Figure 5.6 Pedigree pattern of an X-linked dominant condition. The risk for the individual marked with a query is negligibly low if male, but 100% if female.

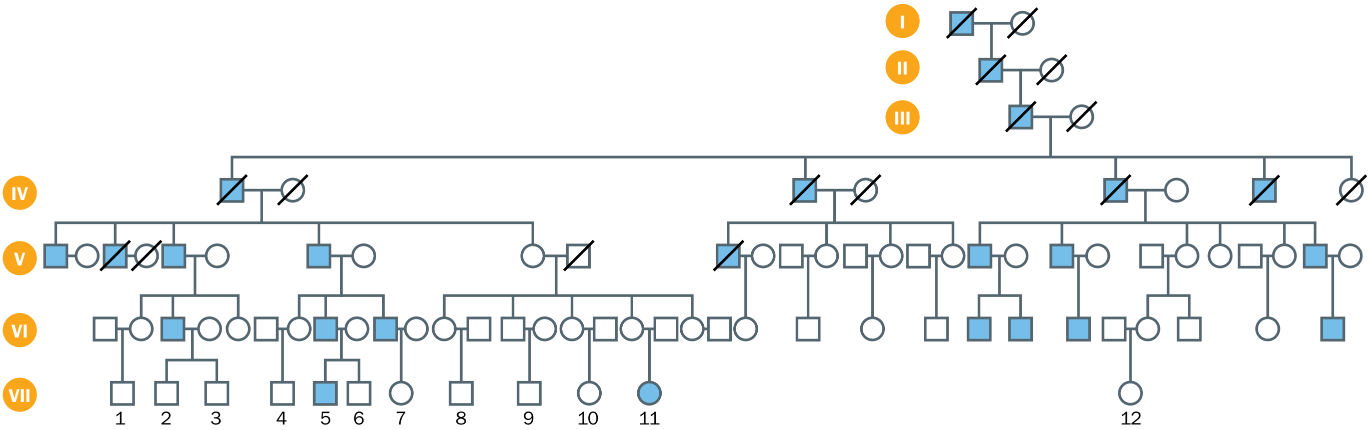


Figure 5.7 Part of the pedigree of a Chinese family in which deafness segregates as a Y-linked character (*DFNY1*, OMIM #400043). The unaffected males VII_2, VII_3, and VII_6 were aged <1, 2, and 4 years, respectively, when examined and were too young to manifest the condition; in affected males, the age of onset ranged from 7 to 27 years. The affected female VII_{11} had been given the antibiotic gentamycin, which is known to cause hearing loss in susceptible people, thus she is probably a phenocopy. The causative gene is probably not normally located on the Y chromosome. It turned out that male family members had a Y-chromosome structural abnormality in which 160 kb of sequence from chromosome 1, including the known hearing-loss gene *DFNA49*, had been inserted into the Y chromosome. (Adapted from Wang QJ *et al*. [2004] *J Med Genet* **41**:e80; PMID 15173246. With permission from the BMJ Publishing Group Ltd.)

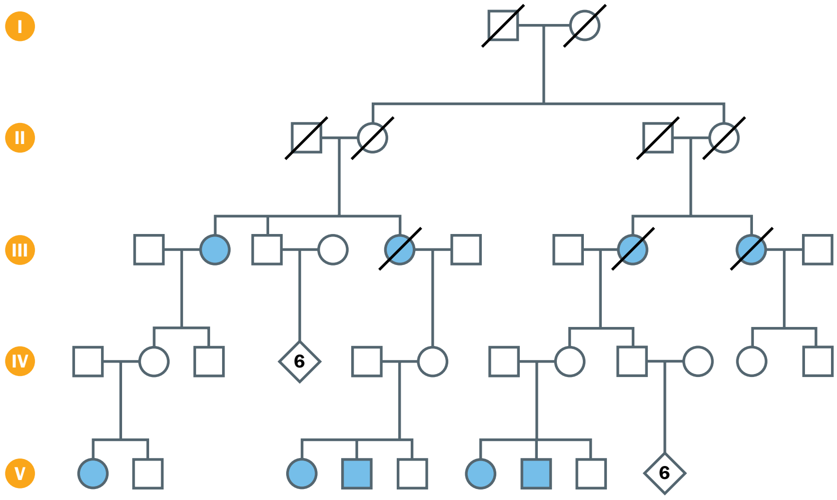


Figure 5.8 Pedigree of a mitochondrially-determined condition. Affected individuals in this Chinese family suffered hearing loss after taking streptomycin. Susceptibility is caused by a variant in the mitochondrial DNA, m.1555A>G (see **Table 10.4** for a guide to the nomenclature of variants; the prefix m. labels it as a variant in the mitochondrial DNA). As described in Section 9.1, mitochondrial DNA is inherited exclusively from the mother. All the sons and daughters of a susceptible woman inherit her m.1555G variant, but only those who were exposed to the antibiotic suffer hearing loss. (Family C reported in Prezant TR *et al*. [1993] *Nat Genet* **4**:289–294; PMID 7689389. With permission from Springer Nature. Copyright © 1993.)

Identifying the mode of inheritance

In a laboratory breeding experiment, one would identify whether a character is dominant or recessive by counting the proportions of the different classes of offspring from a suitable mating. One would do a chi-squared test to check for a 1 in 2, or 1 in 4, and so on, ratio. This is not the way it is done in humans. For a start, families are usually far too small to generate reliable statistics. Additionally, there is a systematic **bias of ascertainment** if one attempts to show that a condition is recessive by showing that in affected families 1 in 4 children are affected. Figure 5.9 illustrates the problem. Relevant families are identified through affected children—but this systematically omits families where, by good luck, none of the children was affected. Segregation analysis offers a range of sophisticated statistical techniques to correct such biases, but requires families to have been collected according to rigid predefined protocols. In general nowadays one looks to see if the pedigree seems to fit one of the patterns set out in Box 5.1, and hopes for molecular testing to resolve uncertainties.

Most people with a dominant character, and especially people with a dominant disease, are heterozygotes. For most rare dominant conditions, homozygotes have never been reported. In some cases, such as achondroplastic dwarfism (OMIM #100800), there is assortative mating (like marrying like) and we do see homozygotes. Babies with homozygous achondroplasia have extreme features of the condition; their rib cage is so small that they cannot breathe, and so they die at birth. It could be argued, since the

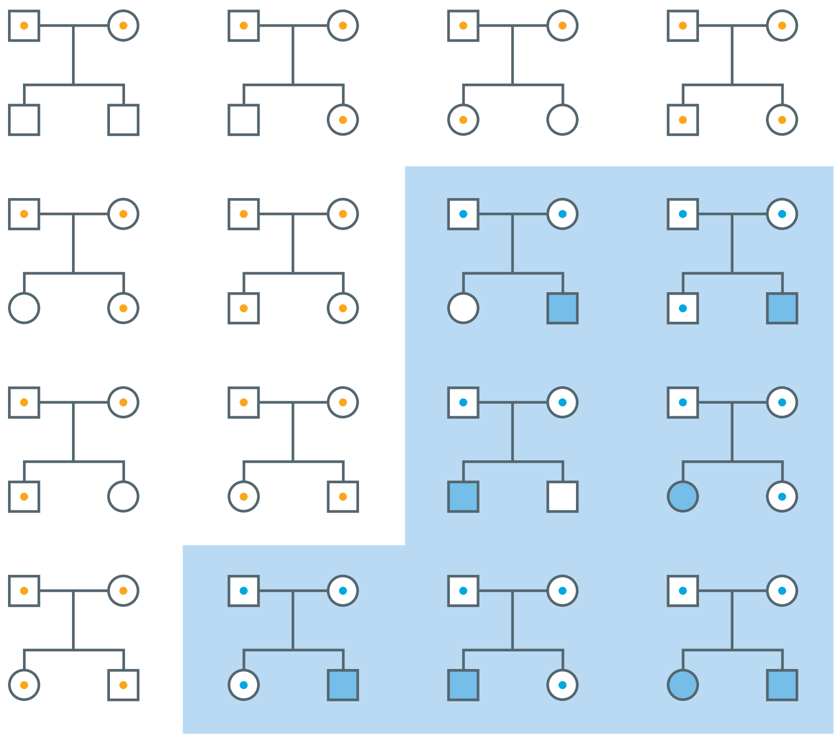


Figure 5.9 Biased ascertainment. In each of these 16 pedigrees, both parents are carriers of an autosomal recessive condition. Overall, their 32 children show the expected 1:2:1 distribution of genotypes (*AA*, 8/32; *Aa*, 16/32; *aa*, 8/32). However, if families can be recognized only through affected children, only the families shown in the shaded area will be picked up, and the proportion of the children in that sample who are affected will be 8/14, not 1/4. Statistical methods are available for correcting such biased ascertainment and recovering the true ratio.

phenotype of heterozygous achondroplastics is intermediate between normal and the full homozygotes, that achondroplasia should be described as a co-dominant or semi-dominant condition, not simply dominant. That would be in keeping with the way characters in experimental organisms are described. But recall that dominance is a property of phenotypes, not alleles or genes. Achondroplasia is a well-characterized phenotype: a person of normal intelligence and fertility with a characteristic build and appearance, including very short arms and legs so that they stand no more than four feet high, and has various problems consequent on the skeletal dysplasia. That is the phenotype that we label achondroplasia, and that phenotype is dominant. A similar argument applies to most of the other human dominant conditions where homozygotes have been described. Huntington disease (progressive neurodegeneration; OMIM #143100) is a rare example of a dominant condition where homozygotes are known and are indistinguishable from the usual heterozygotes.

Complications to the basic patterns

In real life, various complications often disguise a basic Mendelian pattern. Figures 5.10–5.15 illustrate several common complications.

Many conditions show variable expression

Variable expression describes the frequent observation that affected individuals within a pedigree may show different degrees of severity or different features of the condition. Variable expression is especially a feature of dominant conditions. Figure 5.10 shows an example from a family with the autosomal dominant condition Waardenburg syndrome (OMIM #193500). As a general rule, recessive conditions are less variable than dominant ones, probably because the phenotype of a heterozygote involves a balance between the effects of the two alleles, so that the outcome is likely to be more sensitive to outside influence than the phenotype of a homozygote. Careful examination of people with a recessive condition will nevertheless often also show some degree of variability.

These complications are much more conspicuous in humans than in experimental organisms. Laboratory animals and crop plants are far more genetically uniform than humans, and live in much more constant environments. What we see in human genetics is typical of a natural mammalian population. Nevertheless, mouse geneticists are familiar with the way in which the expression of a mutant gene can change when it is bred onto a different genetic background—an important consideration when studying mouse models of human diseases.

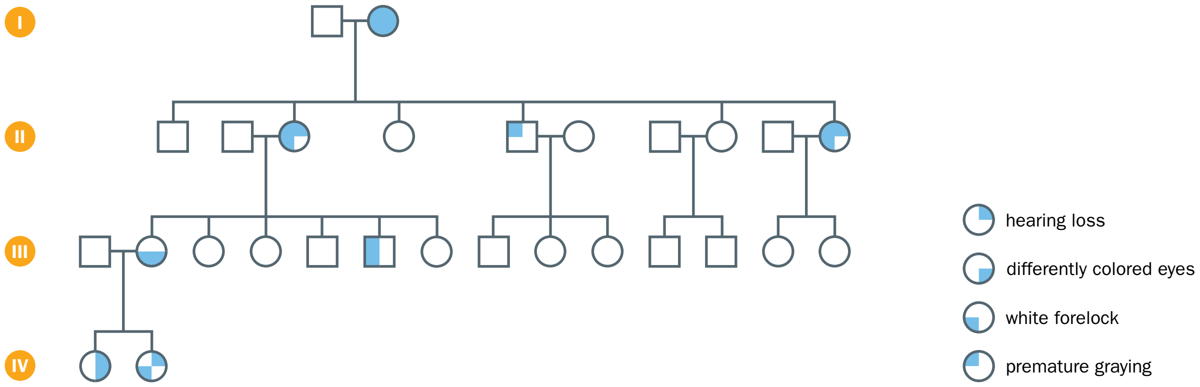


Figure 5.10 Complications to the basic Mendelian patterns (1): variable expression. Different affected family members show different features of type 1 Waardenburg syndrome, an autosomal dominant trait (OMIM #193500), although they all have the same mutation in the *PAX3* gene.

Nonpenetrance: a dominant condition may fail to manifest itself

Nonpenetrance is the extreme of variable expression. The **penetrance** of a character, for a given genotype, is the probability that a person who has the genotype will manifest the character. By definition, a dominant character is manifested in a heterozygous person, and so should show 100% penetrance. Nevertheless, many human characters, although generally showing dominant inheritance, occasionally skip a generation. In **Figure 5.11**, II$_2$ has an affected parent and an affected child, and almost certainly carries the mutant gene, but is phenotypically normal. This would be described as a case of nonpenetrance.

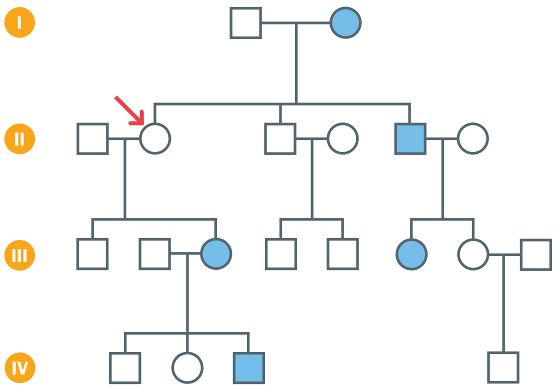


Figure 5.11 Complications to the basic Mendelian patterns (2): nonpenetrance. The pedigree shows transmission of an autosomal dominant condition. Individual II$_2$ (arrowed) evidently carries the gene for the condition but does not show symptoms. Other unaffected family members, such as II$_3$, III$_1$, III$_7$, IV$_1$, or IV$_2$, might also be nonpenetrant gene carriers.

There is no mystery about nonpenetrance—indeed, 100% penetrance is the more surprising phenomenon. Very often the presence or absence of a character depends, in the main and in normal circumstances, on the genotype at one locus, but an unusual genetic background, a particular lifestyle, or maybe just chance means that the occasional person may fail to manifest the character. Nonpenetrance is a major pitfall in genetic counseling. It would be an unwise counselor who, knowing that the condition in **Figure 5.11** was dominant and seeing that III$_7$ was free of signs, told her that she had no risk of having affected children. One of the jobs of genetic counselors is to know the usual degree of penetrance of each dominant condition.

Age-related penetrance in late-onset diseases

A particularly important case of reduced penetrance is seen with late-onset diseases. Genetic conditions are not necessarily **congenital** (present at birth). The genotype is fixed at conception, but the phenotype may not manifest until adult life. In such cases the penetrance is age related. Huntington disease is a well-known example (**Figure 5.12**).

Delayed onset might be caused by the slow accumulation of a noxious substance, by incremental tissue death, or by an inability to repair some form of environmental damage. Hereditary cancers are caused by a chance second mutation affecting a cell of a person who already carries one mutation in a tumor suppressor gene in every cell (see Chapter 19). That second mutation could occur at any time, and so the risk of having acquired it is cumulative and increases through life. Depending on the disease, the penetrance may become 100% if the person lives long enough, or there may be people who carry the gene but who will never develop symptoms no matter how long they live. Age-of-onset curves such as those in **Figure 5.12** are important tools in genetic counseling, because they enable the geneticist to estimate the chance that an at-risk but asymptomatic person will subsequently develop the disease.

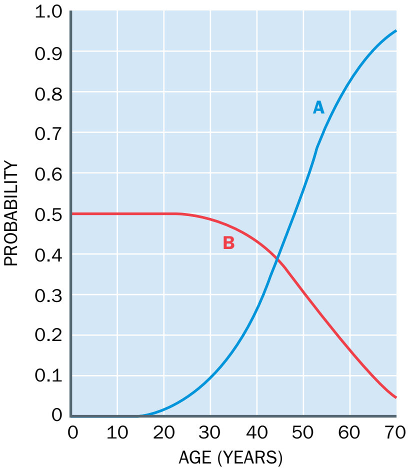


Figure 5.12 Age-of-onset curves for Huntington disease. Curve A shows the probability that an individual carrying the disease allele will have developed symptoms by a given age. Curve B shows the risk at a given age that an asymptomatic person who has an affected parent nevertheless carries the disease allele. (From Harper PS [2010] *Practical Genetic Counselling*, 7th edn. With permission from CRC Press.)

Multigeneration pedigrees often give the appearance of anticipation

Anticipation describes the tendency of some conditions to become more severe, or have earlier onset, in successive generations. As described in Section 16.3, true anticipation is a hallmark of conditions caused by a very special genetic mechanism, dynamic mutation. But when a dominant condition shows random variations in severity, this can easily produce a false impression of anticipation. Mildly affected parents who have a severely affected child will bring it to the clinic. On the other hand, severely affected people may never become parents, or if they do and have a mildly affected child, they might not see any reason to bring it to clinical attention. Thus the clinician's experience is usually of mildly affected parents having severely affected children, and not the reverse. There is a systematic bias of ascertainment that mimics true anticipation. Claims of anticipation without evidence of a dynamic mutation should be treated with great caution. To be credible, a claim of anticipation requires careful statistical backing or direct molecular evidence, not just clinical impression.

Male lethality may complicate X-linked pedigrees

For some X-linked dominant conditions, absence of the normal allele is lethal before birth. Thus affected males are not born, and we see a condition that affects only females, who pass it on to half their daughters but none of their sons. If the family were large enough, one might notice that there are only half as many boys as girls, and a history of miscarriages (because the 50% of males who inherited the mutant allele miscarry before birth). An example is incontinentia pigmenti (**Figure 5.13**; linear skin defects following defined patterns known as Blaschko's lines, often accompanied by neurological or skeletal problems; OMIM #308300). Rett syndrome (OMIM #312750) is another case (see Section 10.3). Affected girls are normal at birth and develop normally for the first year or two, but then stop developing, and eventually regress, losing speech and other abilities that they acquired in early life. In males, Rett syndrome is usually lethal before birth, but rare survivors have a severe neonatal encephalopathy. Until the causative gene was cloned, it was not recognized that these males had the same gene defect as females with classical Rett syndrome.

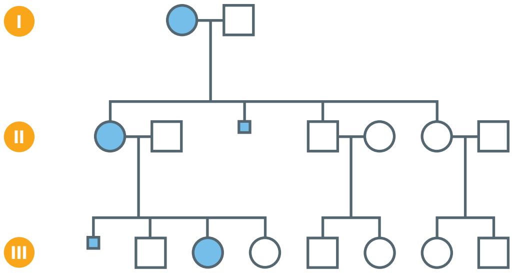


Figure 5.13 Complications to the basic Mendelian patterns (3): a male-lethal X-linked condition. In this family with X-linked dominant incontinentia pigmenti (OMIM #308300), affected males abort spontaneously (small squares).

Inbreeding can complicate pedigree interpretation

The absence of male-to-male transmission is a hallmark of X-linked pedigree patterns—but if an affected man marries a carrier woman, he may have an affected son. Naturally this is most likely to happen as a result of inbreeding in a family in which the condition is segregating. Such matings can also produce homozygous affected females. **Figure 5.14** shows an example.

Metabolic interference could result in heterozygotes for a condition being affected while both homozygotes are unaffected

Metabolic interference was suggested by WG Johnson (1980) (PMID 6770678) as a hypothetical mechanism by which two alleles at a locus, each in itself fully functional, could conflict so as to produce a phenotype in heterozygotes, while both homozygotes would be unaffected. Craniofrontonasal syndrome (OMIM #304110) has often been cited as a possible example. It is an X-linked condition in which males carrying the mutant gene are very mildly affected compared to heterozygous females. The causative mutation is in the *EFNB1* (Ephrin B1) gene at Xp13. Ephrin B1 is involved in defining tissue boundaries. It appears that it is largely dispensable, because males with null mutations have minimal disease signs. Heterozygous females have problems because of X-inactivation. As discussed in Section 10.4, because of X-inactivation a heterozygous female has clones of ephrin-expressing cells mingled with clones of cells expressing no ephrin. The problems arise when cells from positive and negative clones try to form a boundary. Thus, this is not a true example of metabolic interference as conceived by Johnson; it is, however, an example of cellular interference—a conclusion strengthened by the observation that males mosaic for loss-of-function mutations are more severely affected than males with constitutional mutations. It is not clear that any good example of simple metabolic interference as proposed by Johnson has been identified in humans.

The classic Mendelian patterns are best seen with rare conditions

If a recessive trait is common in a population, there is a good chance that it may be brought into the pedigree independently by two or more people. A common recessive character such as blood group O may be seen in successive generations because of repeated matings of group O people with heterozygotes. This produces a pattern resembling dominant inheritance (**Figure 5.15**). The classic Mendelian pedigree patterns are best seen with rare conditions, where there is little chance that somebody who marries into the family might coincidentally also carry the disease mutation that is segregating in the family.

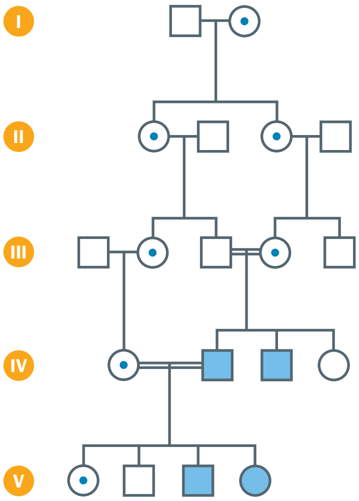


Figure 5.14 Complications to the basic Mendelian patterns (4): an X-linked recessive pedigree with inbreeding. There is an affected female and apparent male-to-male transmission. The pedigree could easily be misinterpreted as showing an autosomal recessive condition.

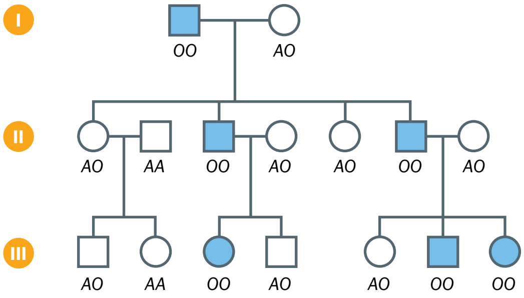


Figure 5.15 Complications to the basic Mendelian patterns (5): a common recessive condition giving an apparently dominant pedigree. If a recessive trait is sufficiently common that unrelated people marrying into the family often carry it, the pedigree may misleadingly resemble that of a dominant trait. The condition in the figure is blood group O.

All these complications to the basic Mendelian patterns reinforce the fact that only a very small fraction of all variants do determine a phenotype directly and with high penetrance. It is a failing in the way genetics is taught that students all too often imagine that clean Mendelian inheritance is the norm for any genetic character, and that any greater complexity in the mode of inheritance is somehow exceptional, abnormal, and best ignored as long as possible. Genes are always Mendelian, but phenotypes are not. Reduced penetrance and variable expression show the effect of the genotypes at other loci (“modifier genes”), plus nongenetic factors and maybe simple chance. As the role of these other factors increases, there comes a point where it is no longer useful to describe a condition as Mendelian. As shown in **Figure 5.1**, all genetic determination lies along a spectrum, ranging from fully penetrant monogenic characters through to polygenic, where the phenotype is the result of the combined effects of variants at many loci, no one of which has a major effect by itself.

Identifying the mode of inheritance and estimating recurrence risks for Mendelian conditions is as much an art as a science. Increasingly nowadays molecular testing identifies causative mutations and removes the necessity of interpreting the pedigree, but where this still has to be done, the answer is often not completely clear. Families are often too small to make the pattern unambiguous, variable expression leads to uncertainty whether an individual is affected or not, and nonpenetrance obscures the transmission of a disease allele. This is not a problem for students sitting exams—the examiner will have made sure that there is one correct answer—but in real life it would be wise not to attempt amateur genetic counseling. Leave it to trained (and insured) professionals!

5.3 MOSAICISM AND NEW MUTATIONS

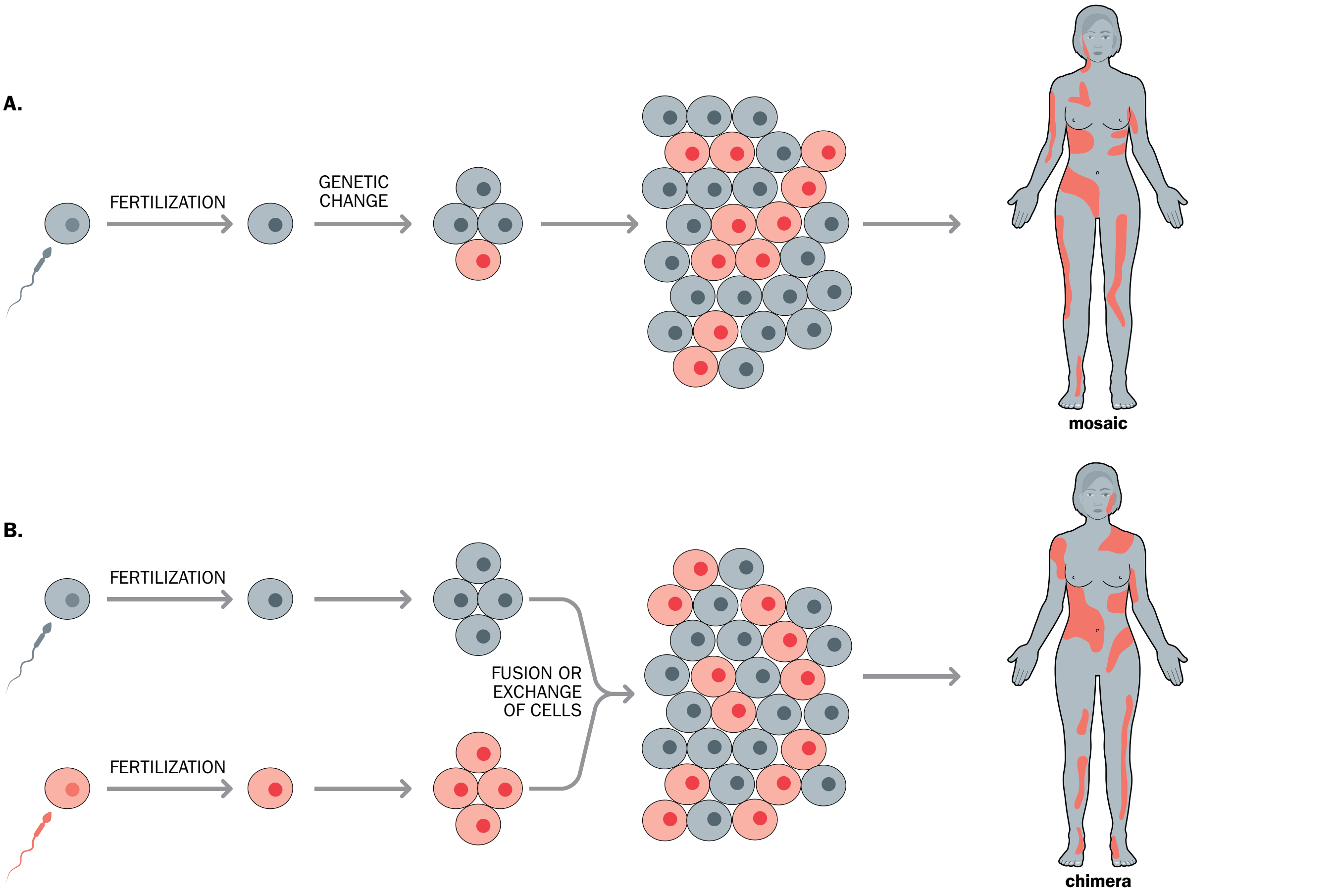
Genetic abnormalities can occur in constitutional or mosaic form

All genetic abnormalities—everything from a gross chromosomal abnormality to a single nucleotide change—can be constitutional or mosaic.

A **constitutional abnormality** is present in all cells of the body. It was inherited in the egg or sperm from a parent who carried the abnormality (or just possibly it could have arisen very early in embryonic development, so that the one abnormal cell gave rise to the whole person).

Mosaicism is when an individual has two or more genetically different cell lines, all derived from one original zygote (**Figure 5.16A**). It is the result of a post-zygotic genetic change, probably in just a single cell of a developing embryo. The event might have occurred early in embryonic development, resulting in large-scale multitissue mosaicism, or much later, producing limited tissue-restricted mosaicism. Conditions that would be lethal in constitutional form may be seen in mosaic form in patients. Intriguingly, when an individual has a constitutional mutation that impairs the proliferation of a certain type of cell, one sometimes sees **revertant mosaicism**, where by chance a cell back-mutates to normal and so acquires a growth advantage over the mutant cells.

Figure 5.16 Mosaics and chimeras. (**A**) Mosaics have two or more genetically different cell lines derived from a single zygote. The genetic change indicated may be a gene mutation, a numerical or structural chromosomal change, or the special case of X-inactivation (see Section 10.4). (**B**) A chimera is derived from two zygotes, which are usually both normal but genetically distinct.



Mosaicism is **somatic** if it involves only somatic cells, **gonadal** or germinal if it is in the germ line, and **gonosomal** if both somatic and germ-line cells are involved. Somatic mosaicism has rather different phenotypic consequences depending whether the gene product involved is diffusible or cell-autonomous. Mosaicism for lack of a circulating protein would show as a reduced level of the protein. Depending on the protein, this might or might not cause clinical symptoms. It might cause a mild version of the phenotype produced by the same mutation when it is in constitutional form. Mosaicism for a cell-autonomous product would create patches of tissue with the mutant phenotype. This would be particularly noticeable in skin, where various nevi and marks are the result of genetic mosaicism. Somatic mosaicism should be suspected in any condition that shows a patchy or variegated phenotype. Gonadal mosaicism, where a phenotypically normal person has a clone of mutant cells in his or her gonads, is a serious problem when estimating recurrence risks, as discussed below.

Note that mosaicism is different from **chimerism** (**Figure 5.16B**). Mosaics start life as a single fertilized egg. Chimeras, in contrast, are the result of fusion of two zygotes to form a single embryo (the reverse of twinning), or chimerism can be the result of intrauterine transfusion between dizygotic twins that share a placenta. Chimerism is rare. It is proved by the presence of too many parental alleles at several loci in a sample that is prepared from a large number of cells. If just one locus were involved, one would suspect mosaicism for a single mutation, rather than the much rarer phenomenon of chimerism.

Blood-grouping centers occasionally discover chimeras among normal donors, and some intersex patients turn out to be XX/XY chimeras. A fascinating example was described by Strain *et al*. (1998) (PMID 9428825; see Further Reading). They showed that a 46,XY/46,XX boy was the result of two embryos amalgamating after an *in vitro* fertilization in which three embryos had been transferred into the mother's uterus.

Most mosaicism goes unnoticed

If we look carefully enough we are all mosaic. Most mutations arise through errors in DNA replication or cell division, or mistakes in repairing DNA damage. Given the number of mitoses involved as a fertilized egg develops into an adult human, and given the finite risk of error whenever a cell divides or repairs damage, it must follow that each one of us is mosaic many times over for a great variety of abnormalities. Additionally, all females are mosaic by virtue of X-inactivation (see Section 10.4). However, mosaicism is only noticed and commented on when a person has a relevant phenotypic abnormality or if they have multiple children affected by a dominant condition that occurs *de novo* in the family. This can happen in three ways:

- If the mutation occurred in an early embryo, affecting a cell that was the progenitor of a significant fraction of the whole person, they might show phenotypic signs of the mutation;
- If the abnormality conferred a growth advantage on cells they might multiply disproportionately. Most obviously this happens in cancer, but there are also a number of congenital syndromes such as Proteus syndrome (OMIM #176920) where there is overgrowth of some part of the body caused by mosaicism;
- If the mutation occurred in a germ-line cell early in development, it could result in a phenotypically normal person harboring a clone of mutant germ-line cells. As a result, a normal couple with no previous family history may produce one or more children with the same serious disease. The possibility of germ-line mosaicism must be considered whenever there is a new mutant case of a condition. As described below, with serious dominant or X-linked conditions such cases are frequent and cause difficulties with pedigree interpretation and estimation of recurrence risks.

New mutations are often originally present in mosaic form

When an individual carries a new mutation, a common assumption is that an entirely normal parent produced a single mutant gamete. However, this is not necessarily what happened. Unless there is something special about the mutational process, such that it can happen only during gametogenesis, a transmitted mutation could have arisen in the parent at any time during post-zygotic life.

New mutations are frequent with serious dominant or X-linked recessive conditions

New mutations are individually rare. In Chapter 11 we describe the various ways they can arise and ways of estimating their frequency. However, in the context of serious dominant or X-linked diseases, they may appear to be very far from rare: they may account for a significant proportion of all cases. If a serious dominant or X-linked condition persists in a population over many generations despite selection removing disease alleles, there must be a compensating production of new mutant alleles.

- A fully penetrant, lethal dominant condition would necessarily always occur by fresh mutation, because the parents could never be affected—an example is thanatophoric dysplasia (severe shortening of long bones and abnormal fusion of cranial sutures; OMIM #187600).
- For a nonlethal but deleterious dominant condition a similar argument applies, but to a lesser degree. Achondroplastic dwarfism is an example, as described in Chapter 12.
- Serious X-linked recessive diseases also show a significant proportion of fresh mutations, because the disease allele is exposed to natural selection whenever it is in a male.
- Autosomal recessive pedigrees, by contrast, are not significantly affected. Ultimately there must have been a mutational event, but the mutant allele can propagate for many generations in asymptomatic carriers, and so it is reasonable to assume that the parents of an affected child are both carriers.

The relation between intensity of selection, mutation rate, and population frequency of a condition is explored further in **Box 12.3**.

New mutations complicate pedigree interpretation

When a normal couple with no relevant family history have a child with severe abnormalities (**Figure 5.17**), deciding the mode of inheritance can be very difficult—the problem might be autosomal recessive, autosomal dominant with a new mutation, X-linked recessive (if the child is male), or nongenetic. In the absence of a direct molecular test, this leads to uncertainties in interpreting the pedigree and still greater uncertainties in estimating recurrence risks.

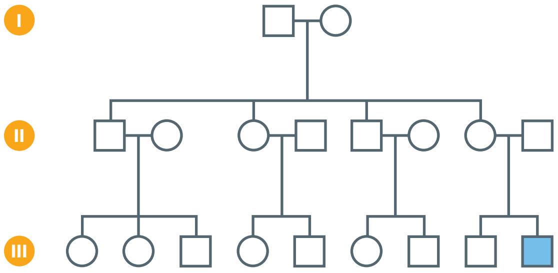


Figure 5.17 A problem in pedigree interpretation. Is this an inherited autosomal or X-linked recessive condition, a new autosomal dominant or X-linked mutation, or maybe a nongenetic condition?

In every pedigree with a new mutation the possibility of germ-line mosaicism must be considered. This complicates estimation of the recurrence risk. For the pedigree in **Figure 5.17**, even if it is proven that the condition affecting the child is caused by a new dominant mutation, it is very difficult to calculate a recurrence risk to use in counseling the parents. **Figure 5.18** shows an example of this uncertainty with an X-linked disease. The problem is discussed by van der Meulen *et al.* (1995) (PMID 7760316; see Further Reading). Usually an empiric risk (see below) is quoted.

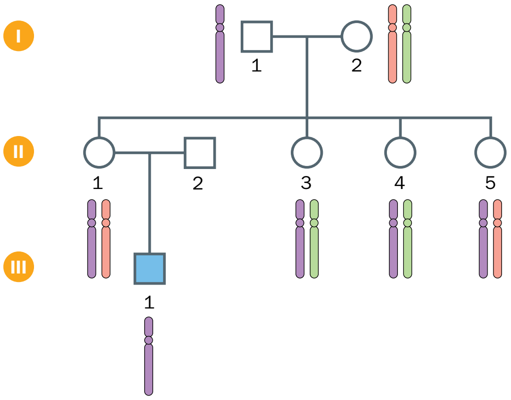


Figure 5.18 An X-linked pedigree with a new mutation. III$_1$ has a serious X-linked disease. The three grandparental X chromosomes were distinguished by using genetic markers; here we distinguish them with three different colors (ignoring recombination). III$_1$ has the grandpaternal X, which must have acquired a mutation at some point in the pedigree. See text for discussion.

In the family shown in **Figure 5.18** there has been a new mutation. It could have occurred at any one of four possible points in the pedigree—each with very different implications for genetic counseling:

- If III$_1$ carries a new mutation that was not present in any form in II$_1$, the recurrence risk for all family members is very low;
- If the mutation happened post-zygotically in II$_1$ so that she is a germinal mosaic, there is a significant (but hard to quantify) risk for her future children, but not for those of her three sisters;
- If II$_1$ was the result of a single mutant sperm, her own future offspring have the standard recurrence risk for an X-linked recessive trait, but her sisters are free of risk;
- If I$_1$ was a germinal mosaic, all four sisters in generation II have a significant (but hard to quantify) risk of being carriers of the condition.

Detecting mosaicism

Molecular studies can be a great help where there is a suspicion of mosaicism. Sometimes it is possible to demonstrate directly that a normal father is producing a proportion of mutant sperm, in which case the proportion can be used to give an accurate recurrence risk. Direct testing of the germ line is not feasible in women, but other accessible tissues such as fibroblasts or hair roots can be examined for evidence of mosaicism. A negative result on somatic tissues does not rule out germ-line mosaicism, but a positive result, in conjunction with an affected child, proves it.

If individual cells are being examined for the presence of a mosaic variant—for example, if karyotyping or fluorescence *in situ* hybridization (Chapter 15) is being used to check for a chromosomal variant—the ability to detect mosaicism depends simply on the number of cells examined. Mosaicism can never be totally excluded, but an upper limit can be placed on its possible extent. If the actual proportion of mutant cells is x, then the chance of failing to see any mutant cell when a sample of N cells is tested is $(1 - x)^N$. So, for example, if 300 cells are checked and no variant found, might the variant actually be present in 1% of the general cell population? The chance of this can be calculated to be 0.05 since $(0.99)^{300} = 0.05$.

A similar argument holds for next-generation sequencing. A very high read depth will allow detection of low-level mosaicism. By contrast, Sanger sequencing is unlikely to detect mosaicism present in less than around 20% of molecules, and other techniques that produce a single answer from bulk DNA will similarly be poor at picking up low-level mosaicism.

Often the question is whether a particular known sequence variant is present in low-level mosaic form. Provided PCR primers can be designed that amplify the variant but not the wild-type sequence (following the principles described in Section 6.2), the variant sequences can be detected and quantitative real-time PCR (Section 6.2) can be used to estimate their frequency. Often one is interested in very-low-level mosaicism—for example, an oncologist might want to check for the low-level presence of a pathogenic variant to attack it before it can progress, or a virologist might want to check whether a treatment has eliminated a patient's HIV infection. For detecting and quantifying a variant present in a few copies per million cells, droplet digital PCR (**Figure 5.19**) is a useful technique. Target sequences can be directly counted without the need for calibration with standard samples, and for very-low-level targets the results are more reliable than those obtained with real-time quantitative PCR.

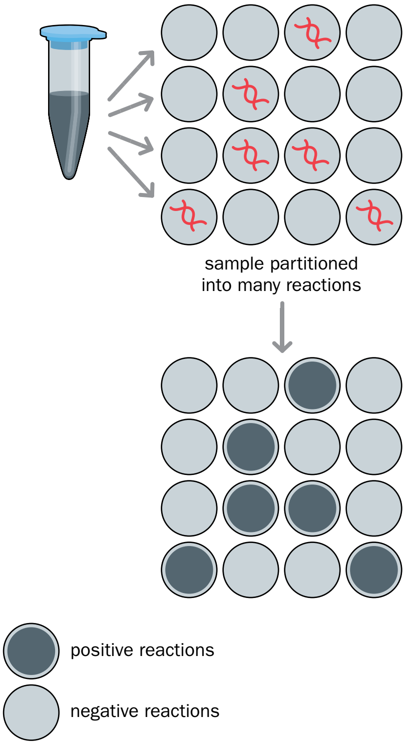


Figure 5.19 Droplet digital polymerase chain reaction (PCR). To detect a very few mutant sequences among a vast preponderance of wild-type sequences, the test sample is mixed with a reaction mix that will specifically PCR-amplify the mutant sequence, then emulsified. The PCR is run on the emulsion, which is then partitioned into individual droplets and the proportion of positive droplets measured. Ideally the initial concentration should be such that most droplets contain either 0 or 1 mutant molecule. Knowing the total concentration of the wild-type sequence, the proportion of mutant sequences can be directly calculated.

5.4 NON-MENDELIAN CHARACTERS

A major thrust of genetic research is toward identifying the genetic variants that underlie common differences between people, including health-related conditions like diabetes, obesity, heart disease, and mental health problems. Such differences almost never behave as Mendelian characters. Hopefully, knowledge of the underlying variants will shed light on the pathogenesis of these conditions and suggest approaches for prevention or treatment. Genotyping a person for a set of variants might sometimes allow a prediction of individual risk. In this section we will survey the theoretical tools that help us understand non-Mendelian inheritance. Work on identifying the genetic factors involved is described in Chapter 18.

In the early twentieth century there was controversy between proponents of Mendelian and quantitative models of inheritance

Mendel sent copies of his paper to the leading plant breeders of his day, but its significance was not appreciated. If you read the paper without hindsight you will perhaps appreciate the problem; in any case, all awareness of it was lost. By the time that it was rediscovered in 1900, a rival school of genetics was well established in the UK and elsewhere. Francis Galton, the remarkable and eccentric cousin of Charles Darwin, devoted much of his vast talent to systematizing the study of human variation. Starting with an article, "*Hereditary Talent and Character*," published in the same year, 1865, as Mendel's paper (and expanded in 1869 to a book, "*Hereditary Genius*"), he spent many years investigating family resemblances. Galton was devoted to quantifying observations and applying statistical analysis. His Anthropometric Laboratory, established in London in 1884, recorded from his subjects (who paid him three pence for the privilege) their weight, sitting and standing height, arm span, breathing capacity, strength of pull and of squeeze, force of blow, reaction time, keenness of sight and hearing, color discrimination, and judgments of length. In one of the first applications of statistics, he compared physical attributes of parents and children, and established the degree of correlation between relatives. By 1900 he had built up a large body of knowledge about the inheritance of such attributes, and a tradition (**biometrics**) of their investigation.

When Mendel's work was rediscovered, a controversy arose. Biometricians accepted that a few rare abnormalities or curious quirks might be inherited as Mendel described, but they pointed out that most of the characters likely to be important in evolution (fertility, body size, strength, and skill in catching prey or finding food) were continuous or quantitative characters and not amenable to Mendelian analysis. We all have these characters, only to different degrees, so you cannot define their inheritance by drawing pedigrees and marking in the people who have them. Mendelian analysis requires dichotomous characters that you either have or do not have.

Polygenic theory explains how quantitative traits can be genetically determined

A controversy, heated at times, ran on between Mendelians and biometricians until 1918. That year saw a seminal paper by RA Fisher in which he demonstrated that characters governed by a large number of independent Mendelian factors (polygenic characters) could display precisely the continuous nature, quantitative variation, and family correlations described by the biometricians. His paper can be read at https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15097/1/9.pdf. It could not be described as easy reading. Later, DS Falconer extended this model to cover dichotomous non-Mendelian characters like birth defects. Fisher's and Falconer's analyses created a unified theoretical basis for human genetics. Here we set out their ideas, in a nonmathematical form. A more rigorous treatment can be found in textbooks of quantitative or population genetics.

Any variable quantitative character that depends on the additive action of a large number of small independent causes (whether genetic or not) will show a Normal (Gaussian) distribution in the population. **Figure 5.20** gives a highly-simplified illustration of this for a genetic character.

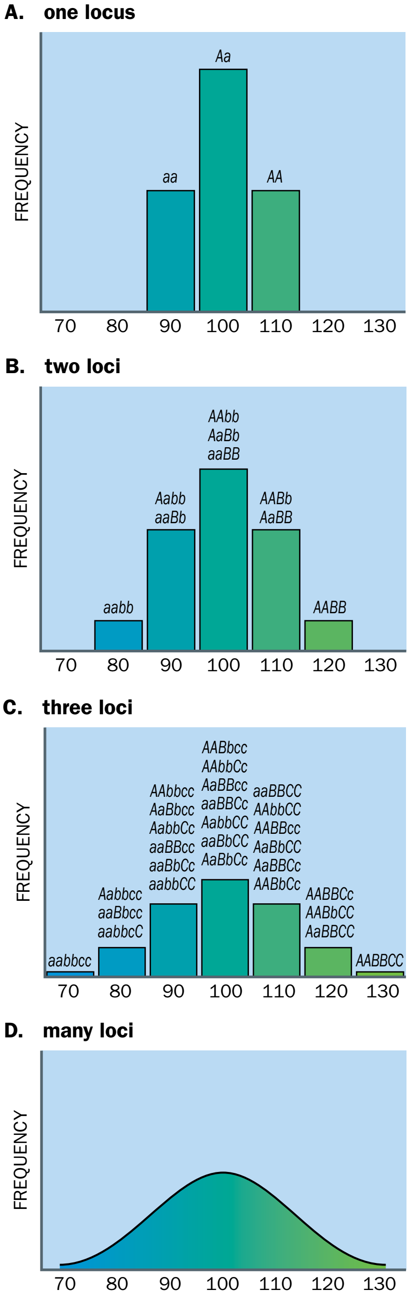


Figure 5.20 Successive approximations to a Gaussian distribution. The charts show the distribution in the population of a hypothetical continuous character that has a mean value of 100 units. The character is determined by the additive (co-dominant) effects of alleles. Each uppercase allele adds 5 units to the value, and each lowercase allele subtracts 5 units. All allele frequencies are 0.5. The character is determined by (**A**) a single locus, (**B**) two loci, and (**C**) three loci. (**D**) The addition of a minor amount of random (environmental or polygenic) variation produces a Gaussian curve.

In **Figure 5.20** we suppose a character to depend on alleles at a single locus, then at two loci, then at three. As more loci are included, we see two consequences:

- The simple one-to-one relationship between genotype and phenotype disappears. Except for the extreme phenotypes, it is not possible to infer the genotype from the phenotype;
- As the number of loci increases, the distribution looks increasingly like a Gaussian curve. The addition of a little environmental variation would smooth out the three-locus distribution into a good Gaussian curve.

A more sophisticated treatment, allowing dominance and varying allele frequencies, leads to the same conclusions. Because relatives share genes, their phenotypes are correlated, and Fisher's 1918 paper predicted the size of the correlation for different relationships.

Regression to the mean

A much-misunderstood feature, both of biometric data and of polygenic theory, is **regression to the mean**. Imagine, for the sake of example only, that IQ were a meaningful quantitative character in which all variation was entirely genetically determined. **Figure 5.21** shows that in our simplified two-locus model, for each class of mothers, the average IQ of their children is halfway between the mother's value and the population mean. This is regression to the mean—but its implications are often misinterpreted. Two common misconceptions are:

- After a few generations everybody will be exactly the same;
- If a character shows regression to the mean, it must be genetic.

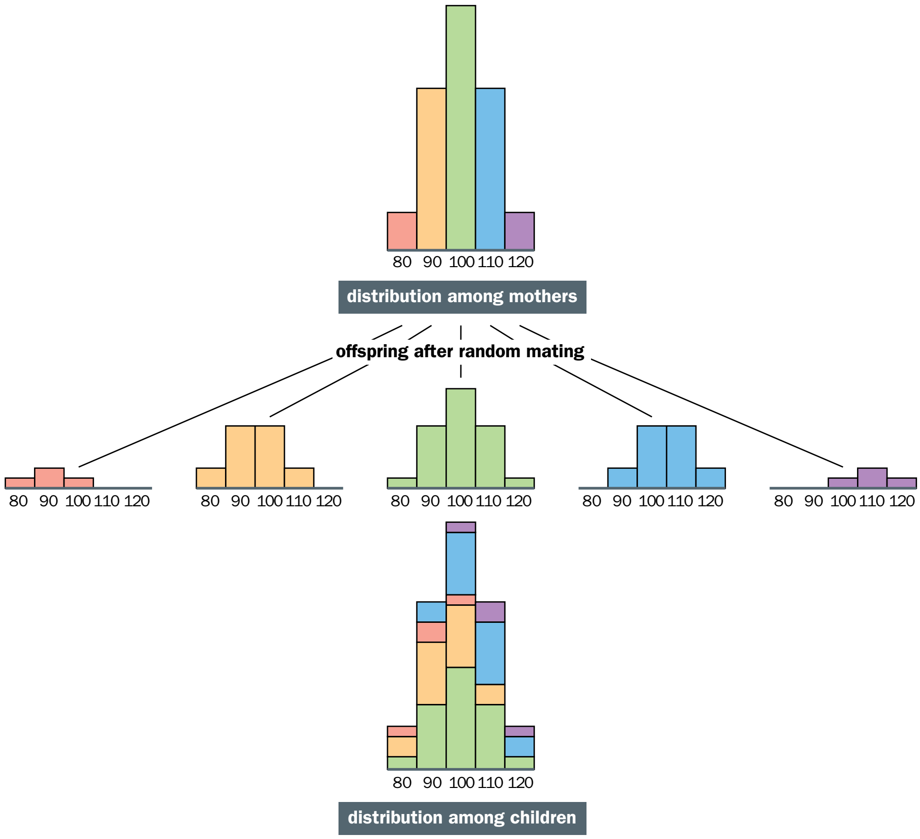


Figure 5.21 Regression to the mean. The same character as in **Figure 5.20B**: mean 100, determined by co-dominant alleles *A*, *a*, *B*, and *b* at two loci, all allele frequencies = 0.5. Top: distribution in a series of mothers. Middle: distributions in children of each class of mothers, assuming random mating. Bottom: summed distribution in the children. Note that: (1) the distribution in the children is the same as the distribution in the mothers; (2) for each class of mothers, the mean for their children is halfway between the mothers' value and the population mean (100); and (3) for each class of *children* (bottom), the mean for their *mothers* is halfway between the children's value and the population mean.

Figure 5.21 shows that the first of these beliefs is wrong. In this simple genetic model:

- The overall distribution is the same in each generation;
- Regression works both ways: for each class of children, the average for their mothers is halfway between the children's value and the population mean. This may sound paradoxical, but it can be confirmed by inspecting, for example, the right-hand column of the bottom histogram in the figure (children of IQ 120). One-quarter of their mothers have IQ 120, half 110, and one-quarter 100, making an average of 110.

Regarding the second of these beliefs, regression to the mean is not a genetic mechanism but a purely statistical phenomenon. Whether the determinants of IQ are genetic, environmental, or any mix of the two, if we take an exceptional group of mothers (for example, those with an IQ of 120), then these mothers must have had an exceptional set of determinants. If we take a second group who share half those determinants (their children, their sibs, or either of their parents), the average phenotype in this second group will deviate from the population mean by half as much. Genetics provides the figure of one-half—it is because each child inherits one-half of his or her genes from their mother that the average IQ of the children (in this simple model) is halfway between the mother's IQ and the population mean—but genetics does not supply the principle of regression.

The simplified model has hidden assumptions

In the simple model of Figure 5.21 there is a hidden assumption: that there is random mating. For each class of mothers, the average IQ of their husbands is assumed to be 100. Thus, the average IQ of the children is actually the mid-parental IQ, as common sense would suggest. In the real world, highly-intelligent women tend to marry men of above average intelligence (**assortative mating**). The regression would therefore be less than halfway to the population mean, if IQ were a purely genetic character—which, of course, it isn't.

A second assumption of our simplified model is that there is no dominance. Each person's phenotype is assumed to be the simple sum of the contribution of each allele at each relevant locus. If we allow dominance, the effect of some of a parent's genes will be masked by dominant alleles and invisible in their phenotype, but they can still be passed on and can affect the child's phenotype. Given dominance, the expectation for the child is no longer the mid-parental value. Our best guess about the likely phenotypic effect of the masked recessive alleles is obtained by looking at the rest of the population. Therefore, the child's expected phenotype will be displaced from the mid-parental value toward the population mean. How far it will be displaced depends on how important dominance is in determining the phenotype.

Heritability measures the proportion of the overall variance of a character that is due to genetic differences

Gaussian curves are specified by just two parameters, the mean and the variance (or the standard deviation, which is the square root of the variance). Variances have the useful property of being additive when they are due to independent causes. Thus, the overall variance of the phenotype V_P is the sum of the variances due to the individual causes of variation—the genetic variance V_G and the environmental variance V_E:

$$V_P = V_G + V_E$$

V_G can in turn be broken down to a variance V_A, which is due to simply additive genetic effects, and two extra terms. V_D accounts for dominance effects: because of dominance, the effect of a certain combination of alleles at a locus may not be simply the sum of their individual effects. V_I is an interaction variance: the overall effect of genes at several loci may not be simply the sum of the effects that each would have if present alone:

$$V_G = V_A + V_D + V_I$$

Therefore:

$$V_P = V_A + V_D + V_I + V_E$$

The **heritability** (h^2) of a trait is the proportion of the total variance that is genetic (the h^2 notation reflects the fact that heritability is a correlation coefficient). The **broad heritability** is defined as:

$$h^2 = V_G / V_P$$

The **narrow heritability**, V_A/V_P, governs the response to selection. For animal breeders interested in breeding cows with higher milk yields, this is an important measure of how far a breeding program can create a herd in which the average animal resembles today's best.

For humans, heritability figures require very careful interpretation. For many human traits, especially behavioral traits, the simple partitioning of variance into genetic and environmental components is not applicable. We give our children both their genes and their environment. Genetic disadvantage and social disadvantage often go together, so genetic and environmental factors are not independent. Moreover, different genotypes are likely to respond differently to different environments. A much-quoted study of the effects of poor upbringing on a person's behavior in adult life (Caspi *et al.* [2002]; PMID 12161658; see Further Reading) illustrated this. In a sample of 1037 individuals in Dunedin, New Zealand, followed to age 26, these authors showed that, as expected, those who had suffered maltreatment in childhood were more likely to commit antisocial acts later on. However, the tendency was much stronger among those who had a certain common variant in the *MAOA* gene that encodes monoamine oxidase. A variable tandem repeat polymorphism present in 35% of the group caused low expression of the enzyme, which is important in regulating turnover of the neurotransmitter serotonin. Individuals with low expression of MAOA were much more prone to respond to childhood abuse by developing antisocial behavior.

If genetic and environmental factors are not independent, we need to introduce additional variances to account for the correlations or interactions between specific genotypes and specific environments. A proliferation of variances can rapidly reduce the explanatory power of the models, and in general this has been a difficult area in which to work.

Misunderstanding heritability

The term "heritability" is often misunderstood. Heritability is quite different from the mode of inheritance. The mode of inheritance (autosomal dominant, polygenic, and so on) is a fixed property of a trait, but heritability is not. Heritability of IQ is shorthand for heritability of variations in IQ. Contrast the following two questions:

- To what extent is IQ genetic? This is a meaningless question. Somebody's ability and willingness to sit down and fill in an IQ test paper depend on innumerable factors in both their genetics and their upbringing;

- How much of the differences in IQ between people in a particular country at a particular time is caused by their genetic differences, and how much by their different environments and life histories? This is a meaningful question (in so far as IQ is a meaningful concept), even if difficult to answer.

In different social circumstances, the heritability of IQ will differ. The more equal a society is, the higher the heritability of IQ should be. If everybody has equal opportunities, several of the environmental differences between people have been removed. Therefore more of the remaining differences in IQ will be due to the genetic differences between people.

The threshold model extended polygenic theory to cover dichotomous characters

Although some continuously variable characters such as blood pressure or body mass index are of great importance in public health, medical geneticists are more concerned with dichotomous characters: the innumerable diseases and malformations that tend to run in families but do not show Mendelian pedigree patterns. DS Falconer provided a major conceptual tool in non-Mendelian genetics by extending polygenic theory to dichotomous or discontinuous characters (those that you either have or do not have).

The key concept is that even for a dichotomous character, there is an underlying continuously variable **susceptibility**. You may or may not have a cleft palate, but every embryo has a certain susceptibility to cleft palate. The susceptibility may be low or high; it is polygenic and follows a Gaussian distribution in the population. Together with the polygenic susceptibility, we postulate the existence of a threshold. Embryos whose susceptibility exceeds a critical threshold value develop cleft palate; those whose susceptibility is below the threshold, even if only just below, develop a normal palate. Stripped of mathematical subtlety, the model can be represented as in **Figure 5.22**. The threshold can be imagined as the neutral point of the balance. Changing the balance of factors tips the phenotype one way or the other.

For cleft palate, a polygenic threshold model seems intuitively reasonable. All embryos start with a cleft palate. During early development the palatal shelves must become horizontal and fuse together. They must do this within a specific developmental window of time. Many different genetic and environmental factors influence embryonic development, so it seems reasonable that the genetic part of the susceptibility should be polygenic. Whether the palatal shelves meet and fuse with time to spare, or whether they only just manage to fuse in time, is unimportant—if they fuse, a normal palate forms; if they do not fuse, a cleft palate results. There is therefore a natural threshold superimposed on a continuously variable process.

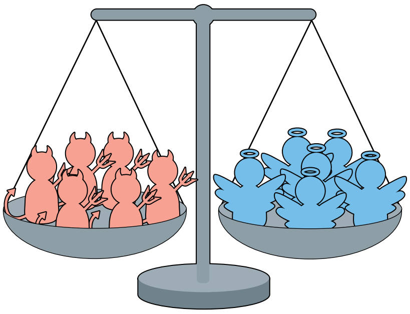


Figure 5.22 Multifactorial determination of a disease or malformation. The angels and devils can represent any combination of genetic and environmental factors. Adding an extra devil or removing an angel can tip the balance, without that particular factor being the cause of the disease in any general sense. (From an idea by the late Professor RSW Smithells.)

Threshold theory helps us understand recurrence risks

Threshold theory helps explain how recurrence risks for non-Mendelian conditions vary in families. Affected people must have an unfortunate combination of high-susceptibility alleles. Their relatives who share genes with them will also, on average, have an increased susceptibility, with the divergence from the population mean depending on the proportion of shared genes. Thus, polygenic threshold characters tend to run in families (**Figure 5.23**). Moreover, in complete contrast to Mendelian conditions, the recurrence risk for polygenic conditions depends on the previous history. Parents who have had several affected children may have just been unlucky, but on average they will have more high-risk alleles than parents with only one affected child. The threshold is fixed, but the average susceptibility, and hence the recurrence risk, increases with an increasing number of previous affected children.

Thresholds may be sex-specific. **Table 5.1** shows some old data that illustrate an example. Congenital pyloric stenosis is five times more common in boys than girls. It has a tendency to run in families. For parents who have had an affected baby, **Table 5.1** shows that the recurrence risk is higher if the affected baby was a girl. Applying polygenic threshold theory, we can understand this. The threshold must be higher for girls than for boys. To be affected, a girl must on average have a higher liability than a boy. Relatives of an affected girl therefore have a higher average liability than relatives of an affected boy (**Figure 5.24**). The recurrence risk is correspondingly higher, although in each case a baby's risk of being affected is five times higher if it is a boy because a less extreme liability is sufficient to cause a boy to be affected.

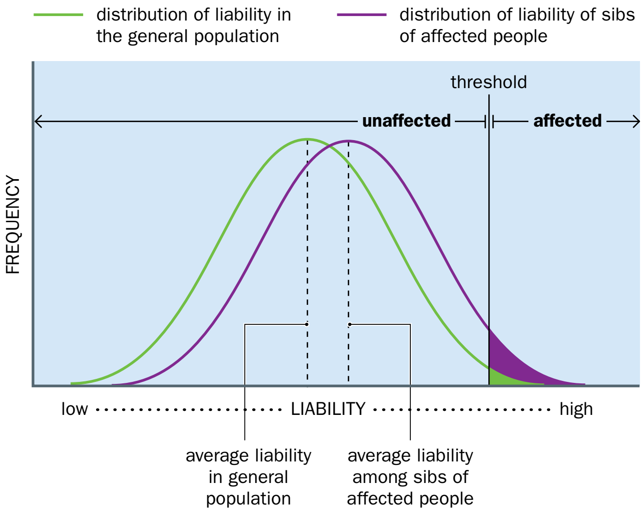


Figure 5.23 A polygenic threshold model for dichotomous non-Mendelian characters. Liability to the condition is polygenic and Normally distributed (green curve). People whose liability is above a certain threshold value (the balance point in **Figure 5.22**) are affected. The distribution of liability among sibs of an affected person (purple curve) is shifted toward higher liability because they share genes with their affected sib. A greater proportion of them have liability exceeding the fixed threshold. As a result, the condition tends to run in families.

TABLE 5.1 RECURRENCE RISKS FOR PYLORIC STENOSIS

Relatives of	Sons	Daughters	Brothers	Sisters
Male proband	19/296 (6.42%)	7/274 (2.55%)	5/230 (2.17%)	5/242 (2.07%)
Female proband	14/61 (22.95%)	7/62 (11.48%)	11/101 (10.89%)	9/101 (8.91%)

More boys than girls are affected, but the recurrence risk is higher for relatives of an affected girl. (Data from Fuhrmann W & Vogel F [1976] *Genetic Counselling*, 2nd edn. Springer-Verlag.)

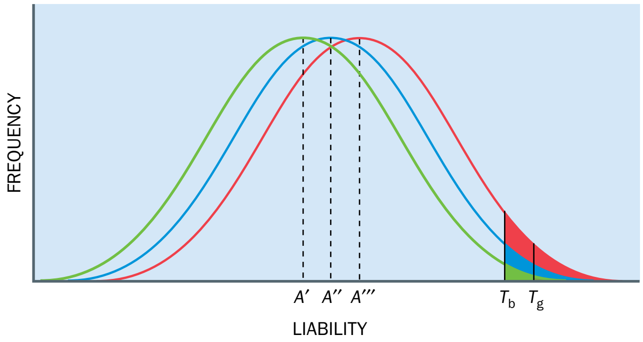


Figure 5.24 A polygenic dichotomous character with sex-specific thresholds. The figure shows a model that explains data such as those in **Table 5.1**. As in **Figure 5.23**, the general population displays a liability to this polygenic disease that is Normally distributed, with an average liability of *A′* (green curve). Boys with a liability above the threshold value T_b manifest the condition; for girls to be affected, the liability must be above the female-specific threshold value T_g. Among siblings of affected boys, the liability (blue curve) is higher, with average *A″*, and a greater proportion of these brothers and sisters have a liability that exceeds the respective threshold levels. Among siblings of affected girls, the liability is still higher (red curve, average liability *A‴*), and an even greater proportion of these brothers and sisters will be affected because they have a liability that exceeds their sex-specific threshold levels.

All of this theory is not used by counselors to predict risks for people who consult them. Those predictions are based on **empirical risks**—risks defined by population surveys, like those in Table 5.1. For such purposes it is important to use data that are recent (unlike those in Table 5.1) and from the same population as the consultand. Different populations can have differing spectra of susceptibility factors, and environmental factors vary both between populations and over time. The value of the models discussed in this section is not to provide actual risk figures but to provide a mental framework that makes sense of the way non-Mendelian characters run in families, and how differing family histories and structures affect recurrence risks. Eventually geneticists want to know the specific genetic variants that contribute to liability. The ways of doing this are discussed in Chapter 18.

SUMMARY

- This chapter defines some of the terminology used to describe the pattern of inheritance of genetic or part-genetic phenotypes. These form an unbroken spectrum, ranging from purely Mendelian to polygenic.
- The pattern of inheritance of a character follows Mendel's rules if its presence, absence, or specific nature is normally determined by the genotype at a single locus.
- Mendelian characters can be dominant or recessive. A character is dominant if it is evident in a heterozygous person, recessive if not. A character is co-dominant if the heterozygote shows features due to both alleles, like blood groups A and B in an AB person.
- Human Mendelian characters give autosomal dominant, autosomal recessive, or X-linked pedigree patterns that are often recognizable, although seldom as unambiguous as the results of laboratory breeding experiments because of the limited size and nonideal structure of most human families.
- The ideal pedigree patterns are often complicated by variable expression, nonpenetrance, and new mutations. Pure monogenic phenotypes are an ideal; in reality there is almost always some influence of other genes and/or environmental factors. As the extent of these other influences increases, there is a spectrum of phenotypes ranging from monogenic through to polygenic or multifactorial.
- Mosaicism describes the situation where a person has two or more different cell populations derived from a single zygote and resulting from a post-zygotic change originally affecting a single cell of an embryo or person.
- Chimeras are individuals who have two cell populations derived from different zygotes.
- Somatic mosaicism can result in a person having a mild or patchy form of a condition; germ-line mosaicism can result in a phenotypically normal person having one or more children with a dominant or X-linked condition that was not present in either parent or the rest of the family.
- New mutations are frequent among people with serious dominant or X-linked conditions. Unless a molecular test is available, they complicate pedigree interpretation and make defining recurrence risks difficult.
- Non-Mendelian characters can be dichotomous (present or absent) or continuous (quantitative). The genes controlling quantitative characters are called quantitative trait loci.
- Most characters depend on more than one genetic locus, and often also on environmental factors. In such cases, no one genetic variant is either necessary or sufficient to determine the character. A variety of environmental and genetic factors each reduce or increase the value of a quantitative character or the likelihood of manifesting a dichotomous character.
- Common diseases are usually complex, having many different possible causes.
- Polygenic theory provides a mathematical framework explaining why many quantitative characters show a Normal (Gaussian) distribution in a population. The theory can also explain some of the features of dichotomous conditions using the concepts of an underlying susceptibility with a threshold value for being affected. It provides a conceptual framework but it is not a tool for predicting the characteristics of individuals.
- The heritability of a character is the proportion of the variance in the character that is due to the genetic differences between people. For a dichotomous character it is the proportion of the variance in susceptibility.
- The heritability of a character is not a fixed property of that character. It is specific to a given population at a given time, and it is sensitive to the range of differences between the environments and lifestyles of different people.

FURTHER READING

Single gene conditions

Albers CA *et al*. (2012) Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat Genet* **44**:435–439; PMID 22366785. (An antidote to simplistic thinking about single-gene conditions.)

Johnson WG (1980) Metabolic interference and the +- heterozygote. A hypothetical form of simple inheritance which is neither dominant nor recessive. *Am J Hum Genet* **32**:374–386; PMID 6770678.

Wang QJ *et al*. (2004) Y-linked inheritance of non-syndromic hearing impairment in a large Chinese family. *J Med Genet* **41**:e80; PMID 15173246.

Wilkie AO (1994) The molecular basis of dominance. *J Med Genet* **31**:89–98. PMID 8182727. (An excellent review of why some characters are dominant and others recessive, with the emphasis on human clinical conditions.)

Zschocke J (2008) Dominant versus recessive: molecular mechanisms in metabolic disease. *J Inherit Metab Dis* **31**:599–618; PMID 18932014. (A detailed discussion, with many examples, of the limitations of the simple division of Mendelian characters into dominant or recessive.)

Mosaicism

Strain L *et al*. (1998) A true hermaphrodite chimera resulting from embryo amalgamation after in vitro fertilization. *N Engl J Med* **338**:166–169; PMID 9428825.

Van der Meulen MA *et al*. (1995) Recurrence risk for germinal mosaics revisited. *J Med Genet* **32**:102–104; PMID 7760316. (Mathematical modeling of the risks.)

Genetics of multifactorial characters

Caspi A *et al*. (2002) Role of genotype in the cycle of violence in maltreated children. *Science* **297**:851–854; PMID 12161658. (There has been some controversy about the claims in this paper, but it illustrates a point.)

Falconer DS & Mackay TFC (1996) *Introduction to Quantitative Genetics*, 4th edn. Longmans Green. (An approachable text covering all aspects of quantitative genetics; not specifically focused on human genetics.)

Francis Galton. Biographical detail, commentaries, and a large range of facsimile documents on the life and achievements of Francis Galton. http://galton.org/

Harper PS (2010) *Practical Genetic Counselling*, 7th edn. CRC Press.

UNDERSTANDING GENOMES

PART TWO

Core DNA technologies: amplifying DNA, nucleic acid hybridization, and DNA sequencing

6

The vast majority of our genetic material is organized as immensely long DNA molecules, but changing just a single nucleotide out of the more than 6 billion nucleotides in our diploid genome can cause disease. Sophisticated technologies have been developed to purify and manipulate genes and other DNA sequences of interest, enabling studies on how they work, how they differ between individuals and across species, how they are changed in disease, and how they can be used therapeutically. We will outline how these approaches are used in later chapters. We describe here just the principal core technologies for purifying and analyzing DNA sequences. In Chapter 7 we describe how these methods have been applied to investigate the structure and expression of genes and genomes. We then go on to describe, in Chapter 8, additional techniques that allow precise changes to be targeted to the genomes of intact cells (genome editing) and the selective inactivation or suppression of a chosen gene (gene knock-outs and gene silencing).

Imagine we wish to analyze or manipulate a single human exon or gene. We can readily isolate DNA from human cells but a single exon, averaging just 230 bp or so, is a tiny fraction of the DNA, just 1/12,000,000 of the haploid genome. Many full-length genes are also extremely small components of the genome, and that poses difficulties for studying them. To overcome this difficulty, two quite different approaches are used, as listed below.

- **Selective amplification of desired DNA sequences.** The idea is to selectively replicate only small pieces of DNA that we are interested in. By making many copies of just these DNA sequences, we can effectively purify them, and get enough of them to work with and analyze. To do that, a DNA polymerase is used to make multiple copies of the desired sequences. The DNA can be amplified by replicating it within intact cells (DNA cloning, described in Section 6.1). Alternatively, the DNA is replicated using some *in vitro* method, notably the polymerase chain reaction (PCR), as described in Section 6.2. Note, however, that both cell-based and *in vitro* DNA cloning are also used to *nonselectively* amplify DNA sequences in a complex starting DNA population, to create collections of clones (DNA libraries) that represent that DNA population. The DNA libraries can then be screened to identify DNA sequences of interest.
- **Specific recognition of desired DNA sequences.** No attempt is made to purify the DNA. Instead, the object is to track the desired DNA sequence using nucleic acid hybridization (Section 6.3). In the first such applications specific short nucleic acid sequences that were complementary in sequence to the desired target DNA sequences were prepared, labeled in some way, and then allowed to bind to the target DNA and track it.

The ultimate way of understanding the structure of genes and DNA sequences, and to track small-scale genetic variation, is to obtain the base sequence of a DNA sample. DNA sequencing used to be expensive, time-consuming, and restricted in scope. All that has changed. Recent rapid technological advances now allow whole genomes to be sequenced quite quickly and cheaply. We cover the principles involved in DNA sequencing and the basis of routine Sanger dideoxy DNA sequencing in Section 6.4. And in Section 6.5 we go on to describe the principles and some details of massively parallel DNA sequencing technologies, often called next-generation DNA sequencing.

6.1 CLONING DNA IN BACTERIAL CELLS

DNA cloning means making identical copies (clones) of a DNA molecule using a DNA polymerase to replicate the DNA. Although it can be carried out in cell-free systems as well as in intact cells, the term is widely used to mean cloning DNA within intact cells, using a DNA polymerase naturally resident within the cells. The DNA sequences to be cloned must first be transferred into some suitable cells that can proliferate rapidly in culture. Cells that have correctly taken up the DNA of interest must be selected in some way, and allowed to proliferate in culture, resulting in a large increase in DNA copy number. After lysing the cells, the desired DNA sequences of interest are purified in some way.

Fractionating and purifying DNA by transforming bacterial cells with recombinant DNAs

The cells used in DNA cloning are typically well-studied bacterial cells, notably strains of *Escherichia coli*. They grow well in culture (and can quickly be expanded to very large numbers in large-volume containers), and they are amenable to genetic modification that maximizes the efficiency of the cloning process. Yeast cells are also used to clone very large DNA fragments, as described in Section 7.1.

The procedure initially involves treating the cells in some way so as to allow transfer of the DNA molecules that we wish to clone into the cells, a process known as **transformation**. In each case, the DNA to be cloned is covalently joined (ligated) to some **vector DNA** sequence that will help it replicate within the host cells, as detailed below.

The joining of DNA fragments to vector molecules results in the formation of an artificial **recombinant DNA** that may be linear in specialized cases (as in the case of cloning very large fragments in yeast). Usually, however, DNA cloning is carried out in bacteria where recombinant DNA molecules are circular (bacteria are more readily transformed by circular DNA).

The transformation process is selective: when foreign DNA does get into a cell, just a single DNA molecule is usually taken up by the cell. If a cell population is presented with a mixture of different foreign DNA fragments, therefore, different DNA fragments will be randomly allocated to different cells during transformation. That is, the population of cells serves as a kind of postal sorting office that can efficiently fractionate a complex mixture of DNA fragments (**Figure 6.1**).

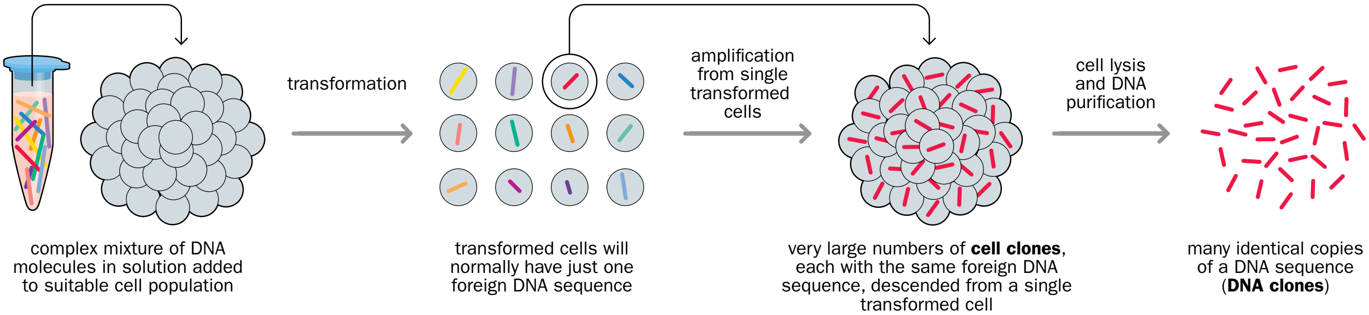

transformation
amplification from single transformed cells
cell lysis and DNA purification
complex mixture of DNA molecules in solution added to suitable cell population
transformed cells will normally have just one foreign DNA sequence
very large numbers of **cell clones**, each with the same foreign DNA sequence, descended from a single transformed cell
many identical copies of a DNA sequence (**DNA clones**)

Figure 6.1 Transformation as a way of fractionating a complex sample of DNA fragments. The key point is that transformation is selective: when a cell is transformed it usually picks up a single DNA molecule from the environment, and so different fragments are taken up by different cells. Cell clones can form by repeated cell division from a single transformed cell and be propagated to produce a large number of cells with an identical foreign DNA sequence that can be purified after breaking the cells open. (Note: for clarity, the figure shows only the DNA sequences that are to be cloned—in practice they would be joined to a vector molecule.)

Amplification

Bacterial cloning systems offer the chance of making large quantities of a cloned DNA. That is, the inserted DNA is amplified to very high copy numbers as indicated in images at the right in **Figure 6.1**. That is possible for two reasons. First, a single bacterium containing a cloned DNA can rapidly divide, leading eventually to a huge number of identical bacterial cell clones, each with the same foreign DNA sequence. Second, some vector molecules can replicate within a bacterial cell to reach quite high copy numbers; if they have a foreign DNA sequence covalently linked to them, that too will be amplified within the cell (**Figure 6.2**). We consider some of the details below.

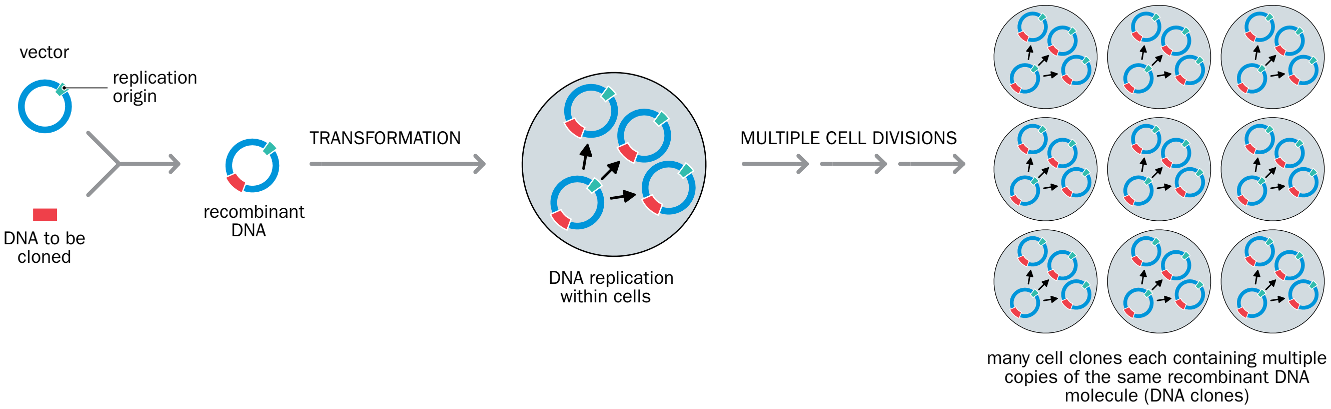

vector
replication origin
DNA to be cloned
recombinant DNA
TRANSFORMATION
DNA replication within cells
MULTIPLE CELL DIVISIONS
many cell clones each containing multiple copies of the same recombinant DNA molecule (DNA clones)

Figure 6.2 Recombinant DNA may be amplified to high copy number within individual cells. Vectors have their own replication origin and can replicate within a bacterial cell independently of the host chromosome, often replicating much more frequently than the host-cell chromosome. For simplicity, the illustration here shows a very modest 3× amplification of the recombinant DNA, but some plasmids allow amplification to 100 copies or more in bacterial cells. A transformed bacterial cell can divide in culture to produce huge numbers of descendants that may each contain multiple copies of a recombinant DNA, resulting in a huge amplification of the starting recombinant DNA.

Vector molecules

Fragments of human DNA would not normally be able to replicate if transferred into bacterial cells or yeast cells. To replicate within cells, the DNA molecules need a suitable origin of replication, a DNA sequence that will initiate DNA replication in that cell type (molecules like this are known as replicons). A convenient solution is to take advantage of replication origins in DNA molecules that naturally replicate within the host cells.

For cloning in bacteria, extrachromosomal replicons are typically used that replicate independently of the bacterial chromosome. Two useful sources are **plasmids** (small, circular, double-stranded DNAs that can replicate to high copy numbers in some cases) and bacteriophages (bacterial viruses).

Plasmid vectors are popular because they are easy to work with, and they are versatile. Different plasmid vectors are suited to cloning fragments of different sizes (**Table 6.1**). Often, the object is simply to clone DNA fragments, but specialized plasmid vectors allow the insert DNA sequences to be expressed to produce RNA transcripts and proteins.

To be useful as a cloning vector, the original plasmid, bacteriophage, or other replicon needs to be genetically modified so that we can efficiently join a foreign DNA to it (as described below) and so that transformed cells can easily be recognized. When cloning in bacteria, for example, the vector will have been genetically engineered to contain a gene that confers resistance to some antibiotic that the host cells are sensitive to. After transformation, the cells are grown on agar containing the antibiotic and untransformed cells die but transformed cells survive. Because some cells are transformed by naked

TABLE 6.1 DIFFERENT PLASMID CLONING VECTORS TO ALLOW CLONING OF DNA FRAGMENTS OF DIFFERENT SIZES

Vector class	Host cell	Insert sizes	Comments
Standard high-yield plasmid vector	*E. coli*	Often small, but up to 5–10 kb	Replicates independently of host chromosome and can reach high copy numbers; widely used for cloning small DNA fragments to achieve high yields of recombinant DNA
Cosmid vector	*E. coli*	30–44 kb	A plasmid vector that is modified to contain *cos* sequences from bacteriophage λ. The short terminal *cos* sequences naturally regulate how the λ genome is packaged into a λ phage particle. When inserted into a plasmid vector, they impose a constraint on the size of the inserts that can be stably accepted when packaging the recombinant DNA into a λ phage particle (which will infect the *E. coli* host cells)
BAC (bacterial artificial chromosome) vector	*E. coli*	Up to 300 kb	A plasmid vector containing a gene that confers a tight constraint on copy number, allowing large DNA fragments to be stably propagated
YAC (yeast artificial chromosome) vector	*Saccharomyces cerevisiae*	0.2–2.0 Mb	A plasmid vector containing short sequence elements needed for chromosome propagation in yeast cells. Recombinants are effectively small linear chromosomes that mostly consist of human (or other foreign) DNA sequence

vector DNA (lacking other DNA), screening systems are often also devised to ensure that cells with recombinant DNA can be identified.

Physical separation of clones

How can cells that have taken up different DNA fragments be separated from each other? The answer is to allow physically separated cell colonies to form. After transformation of bacterial cells, for example, aliquots of the cell mixture are spread over the surface of antibiotic-containing agar in Petri dishes (plating out); successfully transformed cells should grow and multiply and, if the plating density is optimal, they form well-separated cell colonies (**Figure 6.3**). Each colony consists of identical descendant cells (cell clones) that originate from a single transformed cell, and so the cell clones each contain the same single foreign DNA molecule.

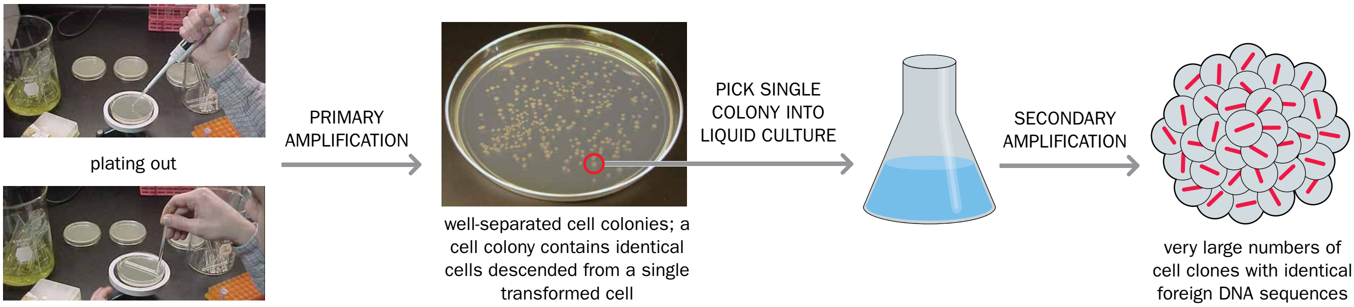

plating out
PRIMARY AMPLIFICATION
well-separated cell colonies; a cell colony contains identical cells descended from a single transformed cell
PICK SINGLE COLONY INTO LIQUID CULTURE
SECONDARY AMPLIFICATION
very large numbers of cell clones with identical foreign DNA sequences

Figure 6.3 Picking well-separated bacterial colonies from a culture dish allows purification of cells containing a single type of recombinant DNA. Irrespective of whether a vector is capable of high-copy number amplification or not, any recombinant DNA can be amplified simply as a result of repeated division of the host cell. Growth occurs originally in a solid medium after the transformed cells are plated out; that is, spread out on a plate of agar containing nutrients and antibiotics (the vector is designed to contain a gene that confers resistance to the antibiotic; an additional selection system is often applied to ensure that the cells contain recombinant DNA, as described in the main text). During plating out, the cells will be physically dispersed to different parts of the agar surface in the culture dish, and individual surviving cells containing the antibiotic-resistance gene go through several rounds of cell division *in situ* to form visible colonies that may be well separated. An individual colony, consisting of identical cells with the same recombinant DNA molecule, can be picked and allowed to go through a second round of amplification by growth in liquid culture. For simplicity, the cloned DNA fragments are shown in the absence of the vector molecule.

An individual, well-separated cell colony can then be physically picked and used to start growth of a large culture of identical cells all containing the same foreign DNA molecule, resulting in very large amplification of a single DNA sequence of interest (**Figure 6.3**). Thereafter, the cloned foreign DNA can be purified from the bacterial cells.

Making recombinant DNA

To make recombinant DNA, each DNA fragment of interest needs to be covalently joined (ligated) by a DNA ligase to a vector DNA molecule. The resulting recombinant DNA molecules will subsequently be transported into suitable host cells, often bacterial or yeast cells. Before that is done, there is a need to prepare the DNA of interest and the vector DNA so that they can be joined efficiently, and there is a need to ensure that the recombinant DNAs are of optimal size.

To clone DNA in bacterial cells, we normally need to use relatively small DNA fragments. When DNA is isolated from the cells of complex organisms, the extremely long nuclear DNA molecules are fragmented by physical shearing forces to give a complex collection of still rather long fragments with heterogeneous ends. The long fragments need to be reduced to pieces of a much smaller, manageable size with uniform end sequences to facilitate ligation.

Recombinant DNA technology was first developed in the 1970s. The crucial breakthrough was to exploit the ability of restriction endonucleases to cut the DNA at *defined* places. As a result, the DNA could be reduced to small, well-defined fragments with uniform end sequences that could be easily joined by a DNA ligase to similarly cut vector molecules (see **Box 6.1**). Note that while most recombinant DNAs are circular, sometimes very large pieces of DNA are cloned in yeast cells and here the recombinant DNA is a linear DNA molecule called a yeast artificial chromosome (YAC), because it resembles a small yeast chromosome.

Plasmid copy number varies significantly: high-copy number plasmids may reach over 100 copies per cell, but other plasmids may be restricted to just 1–2 copies per cell.

BOX 6.1 RESTRICTION ENDONUCLEASES: FROM BACTERIAL GUARDIANS TO GENETIC TOOLS

THE NATURAL ROLE OF RESTRICTION ENDONUCLEASES: HOST-CELL DEFENSE

Restriction endonucleases (also called restriction nucleases) are a class of bacterial enzymes that recognize specific short sequence elements within a double-stranded DNA molecule and then cleave the DNA on both strands, either within the recognition sequences or near to them.

The natural purpose of these enzymes is to protect bacteria from pathogens, notably bacteriophages (viruses that kill bacteria). They can disable the invading pathogen by selectively cutting the pathogen's DNA into small pieces. To ensure that its own genome is unaffected, the host cell produces a matching DNA methyltransferase that methylates the host's own DNA so that it is protected from subsequent cleavage by the restriction nuclease.

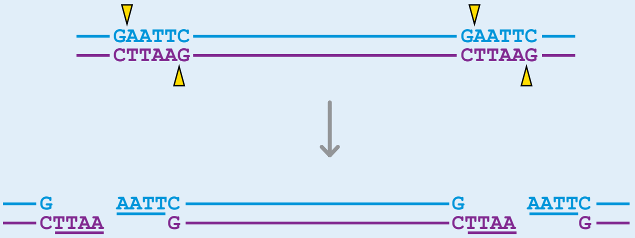
For example, restriction nuclease *Eco*RI from the *Escherichia coli* strain RY13 specifically recognizes the sequence GAATTC and cleaves DNA strands within this recognition sequence (called a **restriction site**). The same bacterial strain also initially produces an *Eco*RI methyltransferase that is used to modify its own genome: it recognizes the sequence GAATTC and methylates the central adenosine on both DNA strands. The *Eco*RI restriction nuclease cannot cleave at previously methylated GAATTC sequences within the bacterial genome but will cleave at unmethylated GAATTC sequences in the DNA of invading pathogens.

RESTRICTION ENDONUCLEASES AS MOLECULAR GENETIC TOOLS

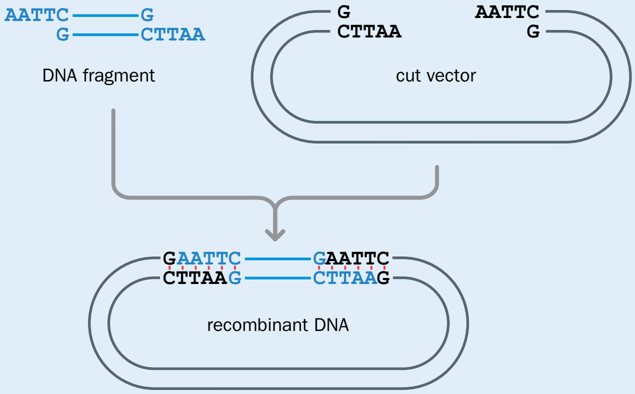
There are different classes of restriction nuclease but type II restriction nucleases are widely used in manipulating and analyzing DNA. They recognize short sequence elements that are typically palindromes (the 5′ → 3′ sequence is the same on both strands, as in the sequence GAATTC); they then cleave the DNA either within or very close to the recognition sequence. Cleavage often occurs at asymmetric positions within the two strands to produce fragments with overhanging 5′ ends (Figure 1) or overhanging 3′ ends.

Under appropriate conditions, it is possible to use a restriction nuclease to cut complex genomic DNA into thousands or millions of fragments that can then be individually joined (*ligated*) using a DNA ligase to a similarly cut vector molecule to produce recombinant DNA molecules (Figure 2).

For cloning DNA in bacterial cells, vector molecules are often based on circular plasmids that have been artificially engineered so that they contain unique restriction sites for certain restriction nucleases. The recombinant DNA molecules can then be transferred into suitable host cells and amplified (see main text).



Box 6.1 Figure 1 Asymmetric cutting of double-stranded DNA by the restriction nuclease *Eco*RI. Note that the underlined AATT sequence is an example of an overhanging 5′ end.



Box 6.1 Figure 2 Formation of recombinant DNA. In this example, the vector has been cut at a unique *Eco*RI site to produce 5′ ends with an overhanging AATT sequence and the DNA fragment to be cloned has the same 5′ AATT overhangs, having also been produced by cutting with *Eco*RI. The AATT overhangs are examples of *sticky ends* because they can hydrogen-bond to other fragments with the same overhang (as shown in the recombinant DNA; hydrogen bonds are shown as vertical red lines) and so facilitate intermolecular interactions.

Different plasmids can coexist in a cell. A typical *E. coli* isolate, for example, might have three different small plasmids present in multiple copies and one large single-copy plasmid. Natural examples of bacterial plasmids include plasmids that carry the sex factor (F) and those that carry drug-resistance genes. Some plasmids sometimes insert their DNA into the bacterial chromosome (integration). Such plasmids, which can exist in two forms, extrachromosomal replicons or integrated plasmids, are known as **episomes**.

An example of standard DNA cloning in bacterial cells using a plasmid vector and a genetically modified host cell

In order to use natural plasmids (and phages) as vector molecules they need to be genetically modified in different ways. First, it is important to design the vector so that restriction fragments produced by cutting the sample DNA are inserted into a unique location (the cloning site) in the vector molecule. To allow cloning of different types of restriction fragments, the vector is genetically engineered to contain a 20–60 bp **polylinker** sequence with many different restriction sites. (As required, naturally occurring restriction sites in the vector are mutated and inactivated to ensure that the introduced restriction sites in the polylinker are unique—see top of **Figure 6.4A** for an example.)

Vector selection

Another genetic modification is to have a marker gene in the vector for selecting only host cells that have been transformed by the vector. The plasmid vector pUC19, for example, contains a gene, *Amp*^{R}, that confers resistance to the antibiotic ampicillin (**Figure 6.4A**), and the host *E. coli* cells used for cloning are ampicillin-sensitive. To screen for cells transformed by the vector, the transformed cells are plated out on an agar surface that contains the relevant antibiotic; the resulting colonies should be descendants of originally antibiotic-sensitive cells that have been transformed by the vector to become antibiotic-resistant.

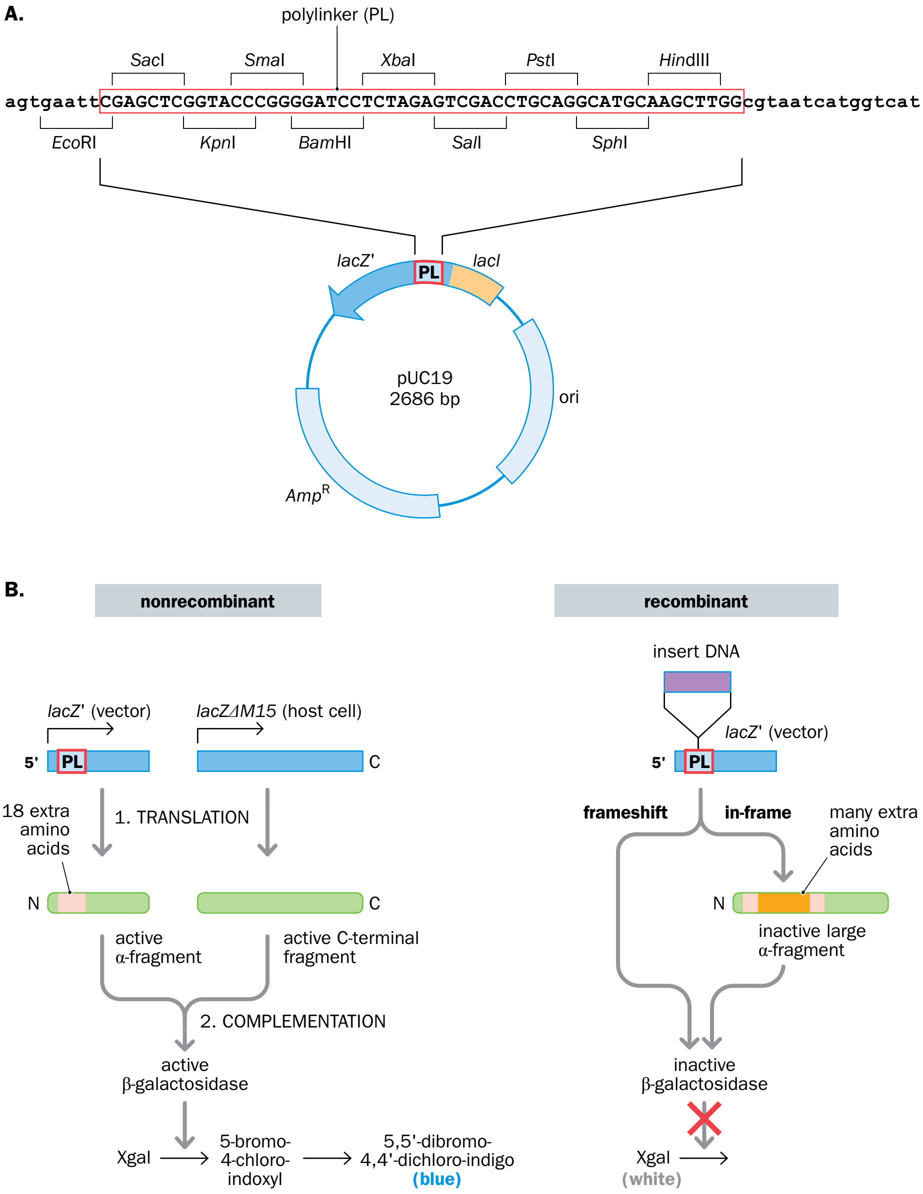


Figure 6.4 The high-copy number plasmid vector pUC19 and the basis of the *lacZ* color screen for recombinants. (**A**) Map of pUC19. The origin of replication (ori) enables more than 100 copies of pUC19 per host cell. The ampicillin-resistance gene (*Amp*^{R}) permits selection for cells containing the vector molecule. A 54 bp polylinker (PL; uppercase letters outlined in red) has been inserted into *lacZ′*, a 5′ fragment of the *lacZ* (β-galactosidase) gene, and provides multiple unique cloning sites. *lacI* encodes a lac repressor protein (which represses transcription of *lacZ′* until the inducer IPTG is added to the culture medium to bind to the repressor protein and inactivate it). (**B**) Basis of the *lacZ* complementation system and the blue/white color selection to identify recombinants. The host cell has been genetically modified to have a 5′ deletion in its *lacZ* gene so that it produces a protein lacking an N-terminal component of β-galactosidase. The vector's 5′ *lacZ′* sequence makes a polypeptide with the first 146 amino acids at the N-terminal end (α-fragment), but it is interrupted by 18 foreign amino acids as a result of translation of the inserted polylinker. The insertion is small and does not affect the activity of the α-fragment, and it can still complement the large C-terminal fragment of β-galactosidase (produced by the host cell's *lacZΔM15* gene) to make a functional β-galactosidase. In the β-galactosidase assay, a colorless substrate (Xgal, 5-bromo-4-chloro-indolyl-β-D-galactopyranoside) is converted to 5-bromo-4-chloroindoxyl, which spontaneously dimerizes to give an insoluble, deep-blue pigment. Cloning of a large DNA fragment into the polylinker inactivates the α-fragment (either by introducing a frameshift or producing a much larger protein). Although the host cell continues to produce the large C-terminal fragment of β-galactosidase (not shown), it can no longer be complemented by an active α-fragment. Because no β-galactosidase can be produced, the cells will be colorless in the assay.

Recombinant selection

Another type of genetic modification is to have a second marker gene in the vector for selecting only host cells that have been transformed by recombinant DNA. Sometimes the short polylinker is inserted into the coding sequence of the marker gene without changing the reading frame or disrupting the function of the marker gene. However, recombinants that have long inserts within the polylinker will disrupt the function of the marker gene (even if it does not change the reading frame, a large insert in the coding sequence will disrupt the function of the marker gene).

A widely used recombinant selection system derives from the *E. coli lacZ* gene and involves a color assay for β-galactosidase, the *lacZ* product. This enzyme can cleave the disaccharide lactose to give glucose plus galactose, but transcription of *lacZ* is normally repressed by a lac repressor protein encoded by the *lacI* sequence. Transcription of *lacZ* can, however, be induced by lactose, or a lactose analog such as isopropylthiogalactoside (IPTG), which binds to the repressor protein and inactivates it. IPTG is generally used as the inducer because it cannot be metabolized, and therefore its concentration does not change as the cells grow.

The *lacZ* marker assay is also designed to depend on complementation between different β-galactosidase fragments produced by the host cell and the vector. The *E. coli* host is a mutant in which the 5′ end of the *lacZ* gene has been deleted; it can produce a large C-terminal fragment of β-galactosidase, which by itself is nonfunctional. Vectors such as pUC19 (**Figure 6.4A**) have, however, been modified to contain a short 5′ component of the *lacZ* gene known as *lacZ′* that when transcribed can produce a short N-terminal fragment of β-galactosidase (the α-fragment). The N-terminal α-fragment complements the host's C-terminal β-galactosidase fragment to produce an active β-galactosidase that can be assayed by a reaction that gives a blue product (**Figure 6.4B**). However, because the cloning site is within the *lacZ′* sequence of the vector, the *lacZ′* sequence of recombinants is interrupted by a large insert and no β-galactosidase can be made.

The principle of DNA clone libraries and the applications and limitations of DNA cloning

Once DNA cloning was established it was soon used as a way of amplifying all DNA sequences from a starting source of DNA to make **DNA libraries**; that is, collections of DNA clones representing all types of DNA sequence in the starting DNA. DNA isolated from white blood cells, for example, provides a complex genomic DNA that can be cut into many pieces and attached to vector DNA molecules. The resulting mix of different recombinant DNA molecules is used to transform bacteria to produce very many different clones—a genomic DNA library. A good genomic DNA library would have so many different DNA clones that there was a good chance that the library included just about all the different DNA sequences in the genome.

An alternative was to make gene-centered DNA libraries. Until recently, it was imagined that the vast majority of human genes made proteins and an obvious starting point was mRNA. RNA cannot be cloned, however. Instead, DNA copies needed to be made using a specialized DNA polymerase, a reverse transcriptase that naturally copies a single-stranded RNA template to make a **complementary DNA** (**cDNA**) copy. Once the cDNA strand is made, the original RNA is destroyed by treatment with ribonuclease and the copied DNA strand is copied in turn to make a complementary DNA, thereby making double-stranded cDNA. Total double-stranded cDNA isolated from cells could then be used to make a cDNA library. But because different genes are expressed in different cell types, the range of DNA clones in a cDNA library could vary according to whether the cDNA originated from white blood cells or brain cells, and so on.

Once genomic and cDNA libraries were made from human cells (and the cells of model organisms) they could be screened using previously isolated DNA clones (or synthetic oligonucleotides) as probes to identify related DNA clones. We go through the details of how this is done in Chapter 7, where we also describe specialized large-insert DNA cloning systems that were first widely used in genome projects.

DNA cloning started a revolution in genetics. It prepared the way for obtaining panels of DNA clones representing all the sequences in the genome of organisms, making genome projects possible, and comprehensive collections of expressed sequences in different types of cell. To allow expression of coding DNAs, cDNA clones were subcloned into specialized plasmid vectors to produce large amounts of purified proteins that could be used for various purposes, including therapeutic purposes or for raising antibodies to track expression of that protein in cells and tissues. By cutting and pasting sequence components derived from different genes it became possible to create hybrid genes that have had different uses, including the production of different types of artificial antibody. By optimizing techniques for transferring gene/cDNA clones into human cells and expressing them within cells, a new field of gene therapy became possible.

There is a downside: cloning DNA in cells is laborious and time-consuming. It is therefore not suited to diagnostic testing and screening work (where parallel amplifications of one or more DNA sequences need to be carried out rapidly in multiple different DNA samples). That required a new technology, the polymerase chain reaction (PCR), as described in Section 6.2.

Expression cloning in bacterial cells as a way of making large amounts of a desired protein

The cloning systems described so far in this section are designed simply to amplify the target DNA to obtain sufficient quantities for structural and functional studies. However, in many circumstances it is also useful to be able to express the introduced gene in some way. In **expression cloning**, appropriate signals need to be provided alongside the introduced gene to enable the gene to be expressed in the transformed host cell.

Depending on the type of expression product required, and the purpose of the expression, many different expression cloning systems can be used. Sometimes an RNA product is sufficient, but in many cases the object is to produce a protein. In some cases, all that is required is to analyze gene expression, for which low expression levels are acceptable; in others, high expression is needed to retrieve, for example, large quantities of a specific protein. It is common to express the product in well-studied cells, but sometimes it may be sufficient to express the product *in vitro*.

Cloning of eukaryotic cDNA in an expression vector is often required for the production of proteins in large quantities for research purposes such as structural studies, or for biotechnology purposes, or as medically relevant compounds such as therapeutic proteins. Usually, a cDNA providing the genetic information specifying the protein sequences is inserted into a vector along with separate expression signals such as suitably strong promoters and other regulatory elements. Because the expression system is based on recombinant DNA, the resulting proteins are sometimes described, rather inaccurately, as *recombinant proteins*.

Bacterial cells, notably the well-studied *Escherichia coli*, have been widely used in expression cloning: they grow rapidly and can be expanded easily in culture to very large culture volumes. But the production of very large amounts of a nonendogenous protein can, however, be detrimental to the growth of the host cell, and can sometimes be toxic. An inducible promoter is often used, therefore, so that expression can be delayed until the transformed cells have been identified and grown in bulk. For example, the pET series of vectors contain a bacteriophage T7 promoter that is not recognized by the endogenous *E. coli* RNA polymerase (**Figure 6.5**). Such vectors are used to transform a genetically modified *E. coli* strain that contains a T7 RNA polymerase gene regulated by a *lac* promoter. The transformed cells can be selected and grown up in large quantities without expression of the foreign gene. Addition of the β-galactosidase inducer IPTG activates the *lac* promoter and expression of the adjacent foreign gene, and the cells can be harvested shortly afterward.

Bacteria have many advantages for expressing heterologous (foreign) proteins, but they also have limitations. Many eukaryotic proteins are modified by the addition of phosphate, lipid, or sugar groups after translation, and these modifications are often essential for the biological function of the protein. Bacterial cells lack the enzymes needed for eukaryotic post-translational processing, however, and so eukaryotic proteins artificially produced in bacterial cells often become unstable, or show limited or no biological activity. That has prompted the use of animal and mammalian cell systems for expressing many human and mammalian genes. We cover aspects of this when we consider genetic manipulation of human and animal cells in Chapter 8.

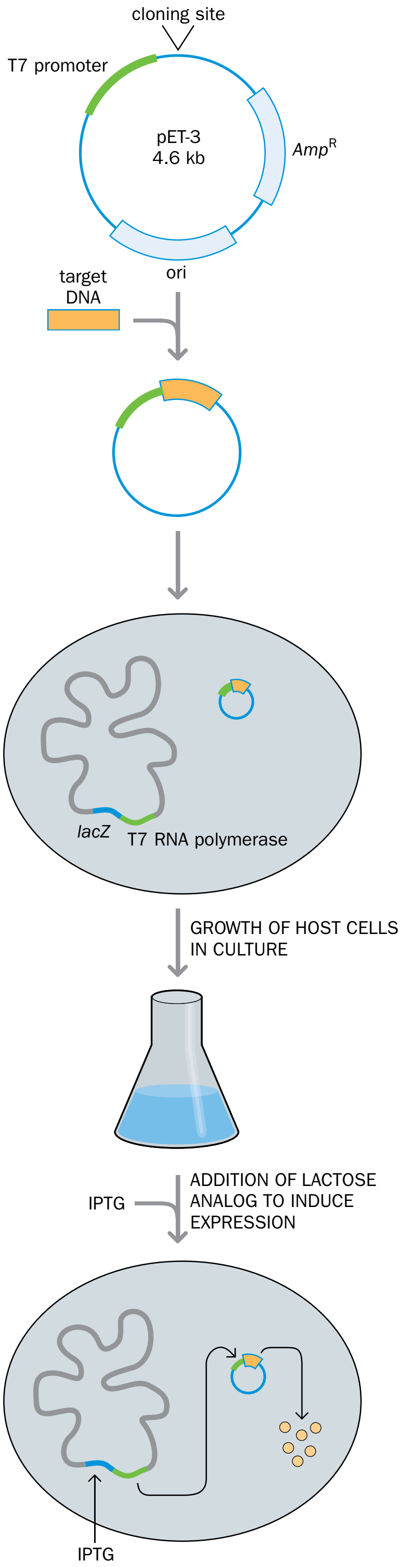


Figure 6.5 Inducible bacterial expression vectors. The pET-3 series of plasmid vectors contain a bacteriophage T7 promoter. They are used with a strain of *E. coli* that has been genetically modified to contain a gene for the phage T7 RNA polymerase under control of an inducible *lac* promoter. After transformation by a pET-3 recombinant DNA, the *E. coli* cells are allowed to grow to give large numbers of cells in culture. At a desired stage, the lactose analog isopropylthiogalactoside (IPTG) is added to induce expression of the host's T7 RNA polymerase gene. The induced T7 RNA polymerase specifically binds to the T7 promoter on the recombinant DNA to give high-level expression of the insert. *Amp*^{R}, ampicillin-resistance gene; ori, origin of replication.

The need for fusion proteins and affinity tags

The large size of many eukaryotic proteins, notably mammalian proteins, also poses difficulties for synthesizing them in *E. coli* (where average protein sizes are only about 300 amino acids). Overexpression of large proteins leads to the production of insoluble aggregates of misfolded protein (inclusion bodies). The inclusion bodies can easily be purified, but it is difficult to then solubilize the protein and achieve efficient refolding *in vitro*.

Efforts to increase yield and solubility have often involved the production of **fusion proteins**. For example, the vector can be modified so that immediately adjacent to the cloning site it contains a cDNA sequence for all or part of an endogenous protein. Recombinants will therefore express the desired protein fused to an endogenous protein sequence. Many modern protein expression vectors are modified to contain a coding DNA sequence for a specific peptide or protein that is easily purified by affinity chromatography. In such cases the recombinant DNAs are expressed to give the desired protein but with a short peptide or protein tag attached that is known as an **affinity tag** because it is attached to assist purification of the recombinant protein by affinity chromatography.

Two favorite systems that allow affinity purification of expressed proteins are based on GST–glutathione affinity and polyhistidine–nickel ion affinity. Glutathione-S-transferase (GST) is a small protein with a very high affinity for its substrate glutathione. The expression cloning vector positions the target DNA just after a gene encoding GST so that a GST-fusion protein is produced in the transformed host cell (**Figure 6.6**). This fusion protein can be purified by selective binding to a column containing glutathione. Alternatively, an affinity tag of six consecutive histidine residues can be attached to a protein. The side chains of the $(His)_6$ tag bind selectively and strongly to nickel ions, assisting purification by affinity chromatography using a nickel–nitrilotriacetic acid matrix.

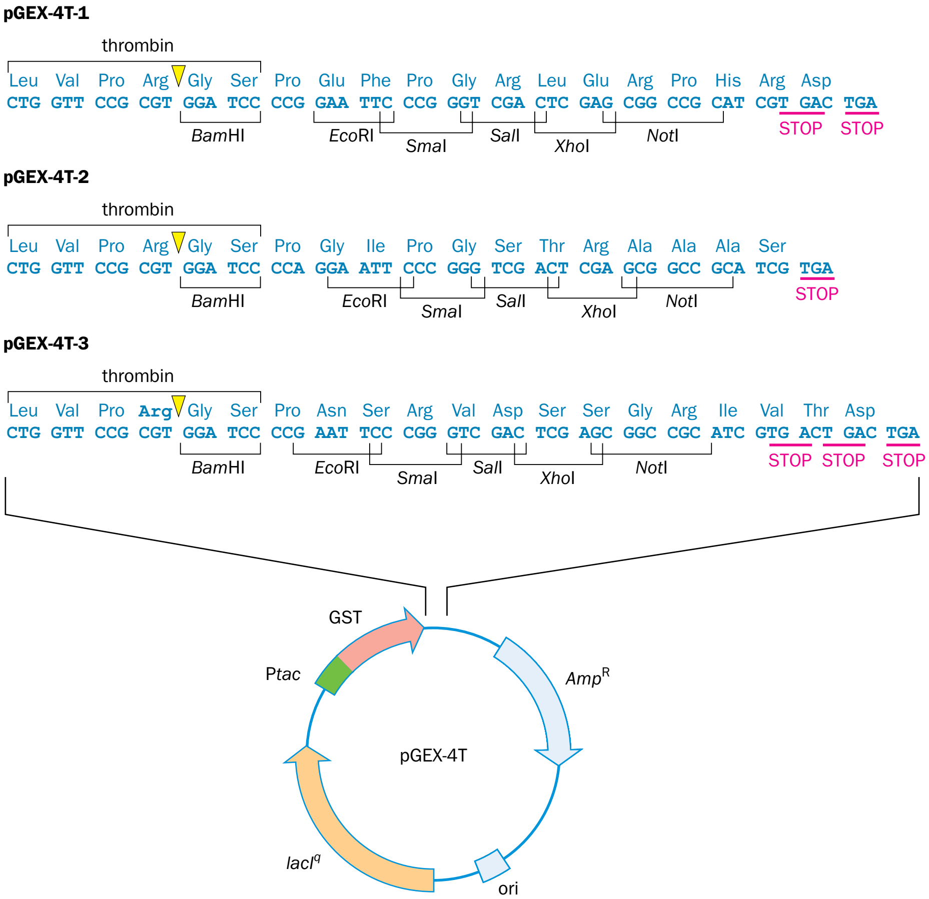

pGEX-4T-1
thrombin
Leu Val Pro Arg Gly Ser Pro Glu Phe Pro Gly Arg Leu Glu Arg Pro His Arg Asp
CTG GTT CCG CGT GGA TCC CCG GAA TTC CCG GGT CGA CTC GAG CGG CCG CAT CGT GAC TGA
BamHI
EcoRI
SmaI
SalI
XhoI
NotI
STOP
STOP
pGEX-4T-2
thrombin
Leu Val Pro Arg Gly Ser Pro Gly Ile Pro Gly Ser Thr Arg Ala Ala Ala Ser
CTG GTT CCG CGT GGA TCC CCA GGA ATT CCC GGG TCG ACT CGA GCG GCC GCA TCG TGA
BamHI
EcoRI
SmaI
SalI
XhoI
NotI
STOP
pGEX-4T-3
thrombin
Leu Val Pro Arg Gly Ser Pro Asn Ser Arg Val Asp Ser Ser Gly Arg Ile Val Thr Asp
CTG GTT CCG CGT GGA TCC CCG AAT TCC CGG GTC GAC TCG AGC GGC CGC ATC GTG ACT GAC TGA
BamHI
EcoRI
SmaI
SalI
XhoI
NotI
STOP
STOP
STOP
GST
Ptac
AmpR
pGEX-4T
lacIq
ori

Figure 6.6 Fusion protein vectors. The pGEX-4T series of vectors have a *tac* promoter (P*tac*, a hybrid promoter with elements of the *trp* and *lac* promoters). Downstream gene expression is normally repressed by the repressor protein encoded by the *lacI*^{q} gene but is inducible by the lactose analog IPTG. Immediately downstream of the *tac* promoter is a gene for the affinity tag glutathione-S-transferase (GST) followed by a multiple cloning site (MCS). The object is to clone a target coding sequence for a protein into the MCS so that a fusion protein is produced, with an N-terminal GST sequence fused to the protein encoded by the target cDNA. Three alternative vectors, pGEX-4T-1, pGEX-4T-2, and pGEX-4T-3, have slightly different MCS sequences that differ in the translational reading frame (see from proline codon onwards in the sequences shown in the top half of figure). By using all three alternative vectors, cloned inserts can be expressed in each of the three amino acid reading frames so that in at least one of the three cases, the target DNA should be in frame with the GST sequence. The expressed fusion protein can be purified easily on a glutathione affinity purification column such as glutathione sepharose 4B. Because the MCS is engineered to contain a thrombin cleavage site, the desired protein can be purified after cleavage at the thrombin cleavage site. *Amp*^{R}, ampicillin-resistance gene; ori, origin of replication; STOP, termination codon.

Phage display

Phage display involves inserting a coding DNA into a bacteriophage vector to produce a recombinant DNA that is transferred into bacteria, leading to the production of recombinant phage. The recombinants that express the foreign DNA give a protein that is displayed on the surface of a phage particle.

Because the cloning sites in the phage vector are designed to lie within a gene encoding a phage coat protein, expression results in a fusion protein that is incorporated into the phage's protein coat, so that it is displayed on the surface of the phage (but it does not affect the phage's ability to infect cells). If an antibody is available for the expressed protein, phage displaying the protein can be selected by preferential binding to the antibody: affinity purification of virus particles bearing such a protein can be achieved from a 10^8-fold excess of phage not displaying the protein, using even minute quantities of the relevant antibody (**Figure 6.7**).

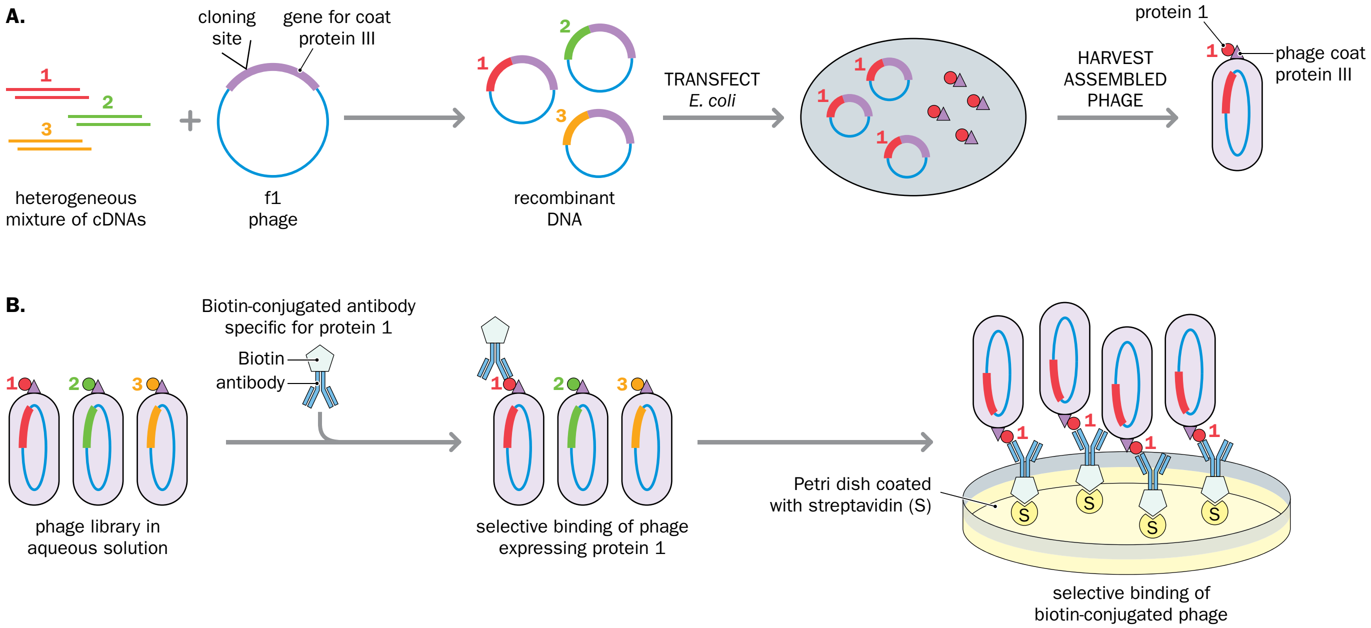

A.
cloning site
gene for coat protein III
1
2
3
+
heterogeneous mixture of cDNAs
f1 phage
1
2
3
recombinant DNA
TRANSFECT
E. coli
1
1
1
HARVEST ASSEMBLED PHAGE
protein 1
1
phage coat protein III
B.
Biotin-conjugated antibody specific for protein 1
Biotin
antibody
1
2
3
phage library in aqueous solution
1
2
3
selective binding of phage expressing protein 1
Petri dish coated with streptavidin (S)
1
1
1
1
S
S
S
S
selective binding of biotin-conjugated phage

Figure 6.7 Phage display. (**A**) A heterogeneous mixture of target cDNAs is cloned into a bacteriophage vector, such as one based on phage M13 or f1, in order to express foreign proteins on the phage surface. Here, DNA is inserted into a cloning site at the extreme N-terminal sequence of the gene for protein III from phage f1 that makes one of the phage coat proteins. The recombinants are allowed to transform host *E. coli* cells, whereupon phage DNA replicates, and phage particles are assembled, extruded from the host cell, and harvested. (**B**) The mixture of recombinant phage is known as a phage expression library. Recombinants with in-frame inserts may often be expressed to give a fusion protein in which the N-terminal component consists of a protein sequence encoded by the inserted DNA. An antibody specific for one of the inserted proteins will bind specifically to just the phage that displays that protein. If the antibody bears an affinity tag such as biotin, it will selectively bind to its partner streptavidin, allowing purification of the labeled phage.

Phage display is a very versatile system. It is used in protein engineering to select for desired variants from a library of mutants. It has also proved a powerful alternative source of constructing antibodies, bypassing normal immunization techniques and even hybridoma technology. Phage libraries can also be used to identify proteins that interact with a specific protein. In the same way in which antibodies can be used in affinity screening, a known protein (or any other molecule to which a protein can bind) can be used as a bait to select phages that display any other proteins that bind to the bait protein.

6.2 AMPLIFYING DNA BY *IN VITRO* DNA REPLICATION

The **polymerase chain reaction** (**PCR**), a cell-free method for amplifying DNA, was first developed in the mid-1980s. It was both very fast and readily allowed parallel amplifications of DNA sequences from multiple starting DNA samples. If you wanted to amplify each exon of the β-globin gene from blood DNA samples from 100 different individuals with β-thalassemia, a single person could now do that in a very short time. Because of its simplicity, rapidity, and versatility, PCR has revolutionized genetics. PCR uses reaction cycles consisting of consecutive steps that require different temperatures, but alternate methods use constant temperatures; we describe isothermal amplification approaches to amplifying DNA *in vitro* at the end of this section.

The polymerase chain reaction (PCR): basic features and quantitation

For most purposes, PCR is used for selective amplification. It relies on using a heat-stable DNA polymerase to synthesize copies of a small, pre-determined DNA segment of interest within a complex starting DNA (such as total genomic DNA from easily accessed blood or skin cells). To initiate the synthesis of a new DNA strand, a DNA polymerase needs a single-stranded oligonucleotide primer that is designed to bind to a specific complementary sequence within the starting DNA.

For the primer to bind preferentially at just one desired location in a complex genome, the oligonucleotide often needs to be about 20 nucleotides long or more and is designed to be able to base-pair perfectly to its intended target sequence (the strength of binding depends on the number of base pairs formed and the degree of base matching).

To allow the primer to bind, the DNA needs to be heated. At a high-enough temperature, the hydrogen bonds holding complementary DNA strands together are broken, causing the DNA to become single-stranded. Subsequent cooling allows the oligonucleotide primer to bind to its perfect complementary sequence in the DNA sample (annealing or **hybridization**). Once bound, the primer can be used by a suitably heat-stable DNA polymerase for it to synthesize a complementary DNA strand.

In PCR, two primers are designed to bind to complementary target sequences that are close to each other on the same DNA molecule but on opposing DNA strands. The primer binding sites are chosen to flank the region of interest, and the primers are normally designed to be convergent so that each new strand is synthesized in the direction toward the sequence bound by the other primer. In further cycles of DNA denaturation, primer binding, and DNA synthesis, previously synthesized DNA strands become targets for binding by the other primer, causing a chain reaction (**Figure 6.8**).

The end result is that millions of DNA copies (**amplicons**) can be made of just the desired DNA sequence of interest within the complex starting DNA. By amplifying the desired sequence, we can now study it in different ways—by directly sequencing the amplified DNA, for example. PCR can also be used to analyze RNA transcripts. In that case the RNA transcripts are first converted into cDNA using reverse transcriptase (the process is called reverse transcription-PCR or **RT-PCR**).

Quantitative and real-time PCR

In routine PCR, all that is required is to generate a detectable or usable amount of product. But for some purposes, there is a need to quantify the amount of product. There are different types of quantitative PCR. Some are variants of routine PCR and use standard PCR machines to give a relative quantification of a sequence of interest within test samples and controls.

Real-time PCR is a form of quantitative PCR carried out in specialized PCR machines. It provides both absolute quantification (the absolute number of copies) and also relative quantification. Instead of waiting for the end of the reaction, the measurements are performed within the PCR machine while the reaction is still progressing. That is, the amplified DNA is tracked by detecting and quantifying in real time the fluorescence from a reporter molecule included within the reaction mixture. Fluorescently-labeled PCR products from the *exponential phase* of the reaction (see **Figure 6.9** for the different phases of a polymerase chain reaction) are removed and analyzed to measure the ratio of the fluorescence exhibited by the PCR product from a test sample (for example, one that is associated with disease or that is suspected of being abnormal) to the fluorescence exhibited by the PCR product from a control sample. The basis of the quantitation is that during the exponential phase, the amount of PCR product is proportional to the amount of target DNA sequence in the input DNA. We expand on this method when we consider important applications in profiling gene expression (as detailed in Chapter 7), and in later chapters we also describe its use in assays for altered nucleotides in DNA.

Advantages and disadvantages of PCR

Because of its simplicity and speed PCR is widely used for myriad research purposes and diagnostic applications. It is also exquisitely sensitive and robust. Thus, it can successfully amplify DNA fragments from tiny amounts of tissue samples that may have been badly degraded, and even from single cells. As a result, there have been numerous applications in forensic and archaeological studies, and in research and diagnostic applications that involve single cells. PCR is robust enough to allow analysis of tissue samples that have been fixed in formalin.

Because PCR involves reaction cycles with steps at very high temperatures, prokaryotic heat-stable DNA polymerases are used (isolated from thermophilic bacteria or archaea that naturally live in hot springs or hydrothermal vents in oceans), and they can have comparatively high error rates. Taq DNA polymerase isolated from the bacterium *Thermus aquaticus* is popularly used, but it lacks a 3′-to-5′ exonuclease proofreading function and so has a significant base-misincorporation rate (about 1 in 10^5). More recently, heat-stable DNA polymerases with 3′-to-5′ exonuclease activity and lower error rates have become popular, such as ones isolated from archaea belonging to the *Pyrococcus* genus (including Pfu DNA polymerase and the Vent™ and Deep Vent™ DNA polymerases).

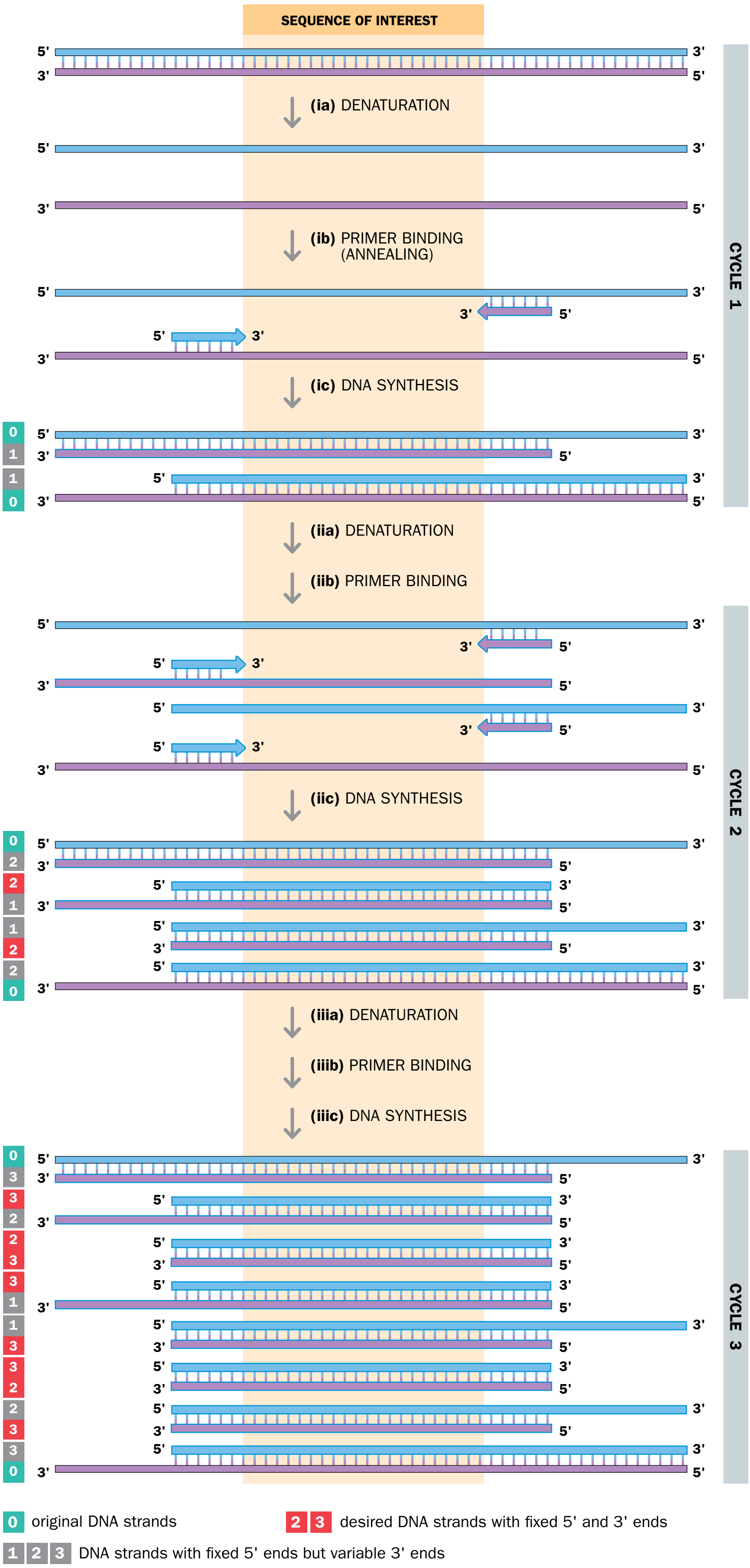


Figure 6.8 The polymerase chain reaction (PCR). The reaction usually consists of about 30 cycles of (1) DNA denaturation, (2) binding of oligonucleotide primers flanking the desired sequence, and (3) new DNA synthesis in which the desired DNA sequence is copied and primers are incorporated into the newly synthesized DNA strands. Numbers in the vertical strips to the left indicate the origin of the DNA strands, with original DNA strands represented by 0 and PCR products by 1 (made during the first cycle), 2 (second cycle), or 3 (third cycle). The first cycle will result in new types of DNA product with a fixed 5′ end (determined by the primer) and variable 3′ ends (extending past the other primer). After the second cycle, there will be two more products with variable 3′ ends but also two desired products of fixed length (shown at left by filled red squares) with both 5′ and 3′ ends defined by the primer sequences. Whereas the products with variable 3′ ends increase arithmetically (amount = $2n$ where n is the number of cycles), the desired products initially increase exponentially (amount = 2^{n-1}) until the reaction reaches a stationary phase as the amount of reactants becomes depleted (see **Figure 6.9**). After 25 or so cycles, the desired product accounts for the vast majority of the DNA strands.

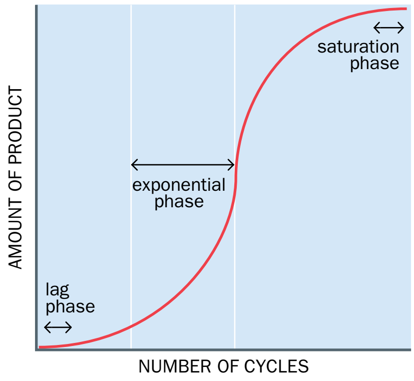


Figure 6.9 Different phases in a polymerase chain reaction (PCR). After a lag phase, the amount of PCR product increases gradually at first. In the exponential phase, beginning after about 16–18 cycles and continuing to approximately the 25th cycle, the amount of PCR product is taken to be proportional to the amount of input DNA; quantitative PCR measurements are made on this basis. With further cycles, the amount of product increases at first but then tails off as the saturation phase approaches, when the reaction efficiency diminishes as reaction products increasingly compete with the remaining primer molecules for template DNA.

PCR has two major disadvantages compared to cell-based DNA cloning. First, unlike cell-based DNA amplification, scaling up to give very large quantities of amplified DNA is not practical. Second, PCR is unsuited to amplifying large DNA sequences: the vast majority of applications produce amplicons less than 10 kb in length (whereas sequences up to a few hundred kilobases in length can be cloned in bacteria and DNA sequences longer than 1 Mb can be cloned in yeast cells).

Isothermal amplification is an alternative to PCR for amplifying DNA sequences *in vitro*

PCR is not the only type of method to allow amplification of DNA *in vitro*. As the name indicates, **isothermal amplification** means that the *in vitro* DNA amplification is carried out at a constant temperature. Isothermal amplification has the advantage that there is no need for specialized equipment to carry out complicated thermocycling; a simple water bath at a constant temperature will suffice. The isothermal amplification methods are, however, not so versatile as PCR, and are primarily used as diagnostic and detection techniques rather than for DNA cloning purposes.

Most of the isothermal methods are designed to allow amplification of specific sequences (and some examples, including the popular LAMP method, are listed in **Table 6.2**). In addition, some isothermal amplification methods allow indiscriminate amplification of all sequences in a complex starting nucleic acid, as described below.

TABLE 6.2 EXAMPLES OF SOME ISOTHERMAL AMPLIFICATION METHODS PERMITTING EXPONENTIAL AMPLIFICATION OF SPECIFIC SEQUENCES

Amplification method	Temperature °C	Characteristics
Exponential strand displacement amplification	37	Efficient amplification, but specificity of amplification is less than for methods using higher temperatures. Needs a nicking endonuclease plus a DNA polymerase
Helicase-dependent amplification	37–65	Requires a helicase as well as a DNA polymerase. Efficient amplification, but significantly less amplification than with the LAMP method
Loop-mediated isothermal amplification (LAMP)	60–65	Highly efficient amplification and high specificity of amplification. A heat-tolerant DNA polymerase with a high strand displacement activity is used, and the specificity is further increased by using at least four primers, a set of outer forward and backward primers, and an inner set of forward and backward primers

The review by Zhao Y *et al*. (2015) (PMID 26551336; see Further Reading) gives a comprehensive overview; diagnostic devices using isothermal amplification are reviewed by Chang C-C, Chen CC, Wei SC *et al*. (2012) *Sensors* **12**:8319–8337; PMID 22969402. For a video description of the popular LAMP method, see https://www.youtube.com/watch?v=L5zi2P4lggw.

Nonselective DNA amplification methods to provide sufficient material for study from samples with limited starting nucleic acid

Up until now we have considered PCR as a tool for selective amplification of a desired DNA sequence so that it can be purified, then studied or put to use in some way, or as a way of detecting a DNA sequence in some clinical or other diagnostic test. But both PCR and isothermal amplification methods can also permit *indiscriminate* amplification of all DNA fragments from a starting DNA. The object is simply to increase the amount of DNA for study.

This type of application is especially required when the source of DNA is in very limiting quantities, such as genomic DNA retrieved from bone samples at archaeological and historical sites, or from tiny human residues left at crime scenes, or when analyzing single cells, whether in pre-implantation diagnosis or for research purposes. As detailed in Section 6.5 it has also become essential for preparing DNA libraries for many high-throughput ("next-generation sequencing") methods where millions of different DNA fragments in the starting DNA need to be amplified (in order to increase the signals from incorporating individual nucleotides so that they can be detected during the sequencing reaction).

Various types of PCR can be used to amplify all sequences in a complex starting population. One popular approach is to attach a common DNA sequence onto the ends of all the fragments, and then use primers that are specific for the common flanking sequence. The first step is to prepare a double-stranded **adaptor oligonucleotide** (sometimes also called a linker oligonucleotide) by designing complementary synthetic oligonucleotides and allowing them to hybridize. The common adaptor oligonucleotide is ligated to all of the DNA fragments. Then, primers specific for the universal adaptor sequence are used to amplify any fragments that have the adaptor sequence at their ends (Figure 6.10).

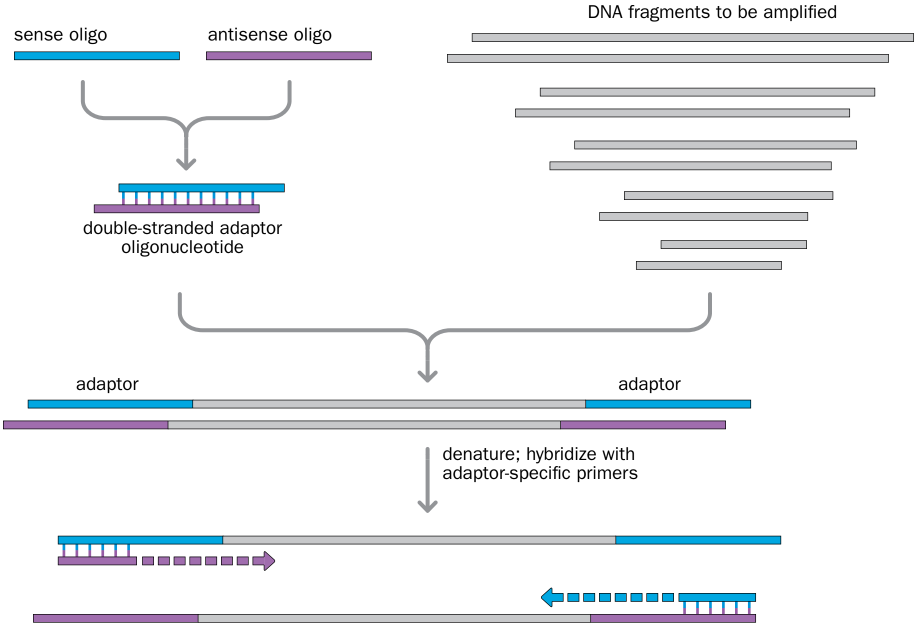

DNA fragments to be amplified
sense oligo
antisense oligo
double-stranded adaptor oligonucleotide
adaptor
adaptor
denature; hybridize with adaptor-specific primers

Figure 6.10 Nonselective DNA amplification using double-stranded adaptor oligonucleotides. Here the idea is to ligate a universal sequence to the ends of a heterogeneous collection of DNA fragments so that all DNA fragments can be amplified using a single set of primers. Two synthetic oligonucleotides are designed to have complementary sequences and allowed to hybridize to form the universal sequence. After the resulting double-stranded adaptor oligonucleotide is ligated to the DNA fragments, amplification of all of the DNA fragments is possible using primers that are specific for the adaptor sequence.

For some applications the nonselective amplification method requires that the amplified DNA fragments be flanked by two different adaptor sequences. For example, in high-throughput DNA sequencing methods that require amplified DNA templates, DNA fragments to be sequenced are often manipulated so that different adaptor sequences are bolted on at the two ends—we give details in Section 6.5.

Isothermal amplification methods have also been applied to allow nondiscriminate amplification. Whole-genome amplification is possible from single cells using, for example, the multiple displacement amplification (MDA) method. It relies on using the highly processive bacteriophage Φ29 DNA polymerase that can produce amplicons greater than 70 kb in length. Because of its importance in whole-genome amplification from single cells, we will detail the method when we consider single-cell technologies in Chapter 7.

6.3 NUCLEIC ACID HYBRIDIZATION: PRINCIPLES AND USES

In a double-stranded DNA molecule, the hydrogen bonds between paired bases act as a glue that holds the two complementary DNA strands together. Two hydrogen bonds form between A and T in each A-T base pair, and three hydrogen bonds hold G and C together in each G-C base pair (see Figure 1.7A). A region of double-stranded DNA that is GC-rich (having a high proportion of G-C base pairs) is therefore more stable than a region that is AT-rich.

Each hydrogen bond is individually weak but when base matching extends over many base pairs the cumulative strength of the hydrogen bonds becomes quite strong (think of Velcro$^{\text{TM}}$; a single Velcro hook-and-loop attachment is very weak, but thousands of them make for a strong fastening system).

Double-stranded DNA can be subjected to different treatments that result in breaking of the hydrogen bonds so that the two DNA strands are separated (**denaturation**). For example, if we heat the DNA to a high-enough temperature (or expose it to high concentrations of a strongly polar molecule such as formamide or urea), the hydrogen bonds break, causing the complementary DNA strands to separate. Subsequent gradual cooling allows the separated DNA strands to come together again, re-forming the base pairs in the correct order to restore the original double-stranded DNA (Figure 6.11A).

Artificial heteroduplexes

The association of any two complementary nucleic acid strands to form a double-stranded nucleic acid is known as nucleic acid **hybridization** (or **annealing**). Under experimental conditions, two single nucleic acid strands with a high degree of base complementarity

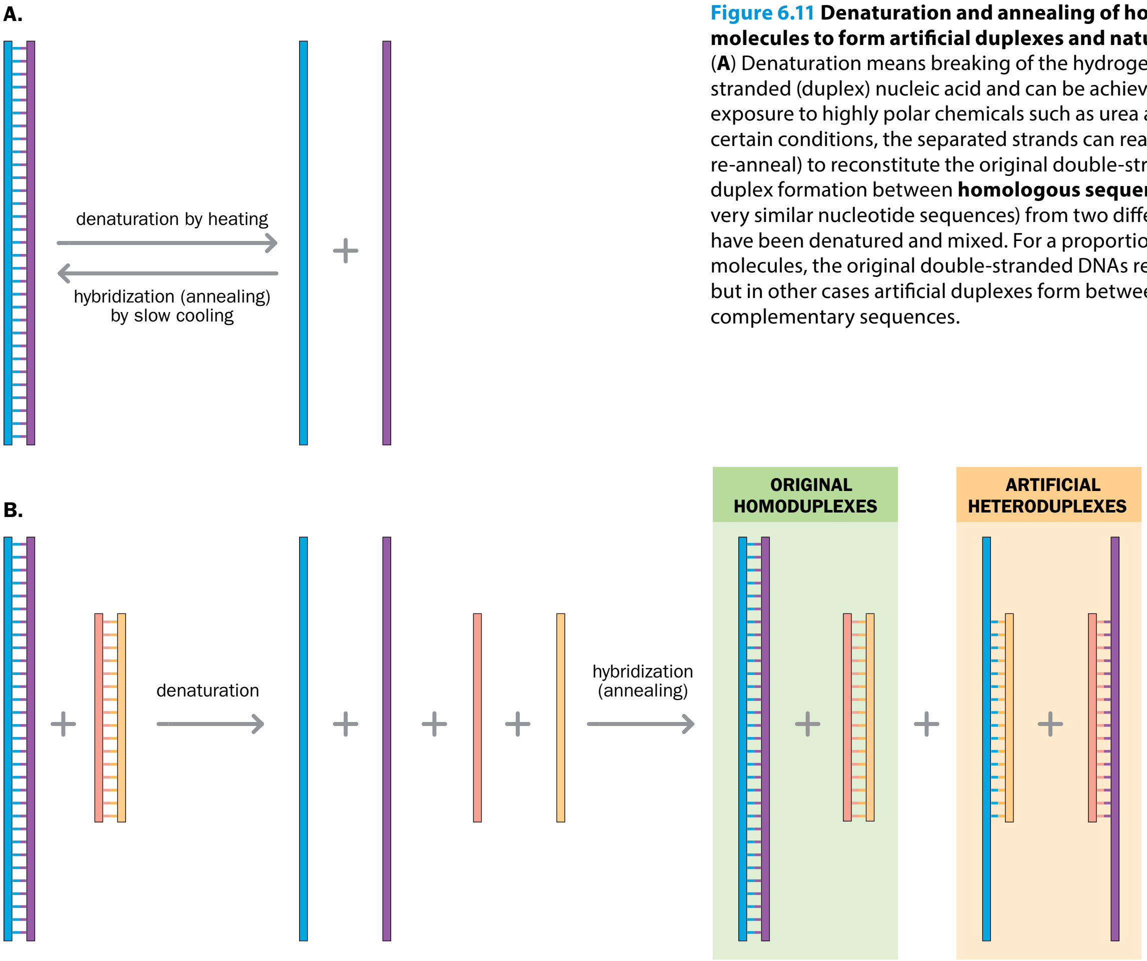

A.
denaturation by heating
hybridization (annealing) by slow cooling
+
B.
+
denaturation
+
+
+
hybridization (annealing)
ORIGINAL HOMODUPLEXES
+
+
ARTIFICIAL HETERODUPLEXES
+

Figure 6.11 Denaturation and annealing of homologous DNA molecules to form artificial duplexes and natural homoduplexes. (**A**) Denaturation means breaking of the hydrogen bonds in a double-stranded (duplex) nucleic acid and can be achieved by heating (or by exposure to highly polar chemicals such as urea and formamide). Under certain conditions, the separated strands can reassociate (hybridize or re-anneal) to reconstitute the original double-stranded DNA. (**B**) Artificial duplex formation between **homologous sequences** (ones that have very similar nucleotide sequences) from two different DNA sources that have been denatured and mixed. For a proportion of the denatured DNA molecules, the original double-stranded DNAs re-form (homoduplexes), but in other cases artificial duplexes form between the partially complementary sequences.

can be allowed to hybridize to form an artificial duplex. For example, if we mix cloned double-stranded DNA fragments that come from two different sources but have high levels of sequence identity, then heat the mixture to disrupt all hydrogen bonding, all the millions of molecules of double-stranded DNA in the DNA from each of the two sources will be made single-stranded (Figure 6.11B).

Now imagine allowing the mixture to cool slowly; two different types of DNA duplex can form. First, a proportion of the single-stranded DNA molecules will base-pair to their original partner to reconstitute the original DNA strands (homoduplexes). But, in addition, sometimes a single-stranded DNA molecule will base-pair to a complementary DNA strand in the DNA from the other source to form an artificial **heteroduplex** (Figure 6.11B).

(Note that we use the term heteroduplex to cover all artificial duplexes in which base pairing is not perfect across the full lengths of the two complementary strands. There may be obvious base mismatches. But even where there are not, as in the example in Figure 6.11B where each base of the smaller DNA strands is perfectly base paired, the term heteroduplex can be applied when the two participating strands are of unequal length, because many bases in the long strands of the heteroduplexes remain unpaired. Very rarely, complementary DNA strands from two different sources might be generated that have both identical lengths and perfect base matching—if so, they could be said to form artificial homoduplexes.)

Two different uses of nucleic acid hybridization

The specificity of base pairing to form stable nucleic acid duplexes is what makes nucleic acid hybridization such a useful tool. For most purposes, the object is to perform a hybridization assay that is used to track some sequence of interest. For convenience we have illustrated cloned double-stranded DNAs in Figure 6.11B. But as we will see below, the starting sources of nucleic acids often include RNA or synthetic oligonucleotides, as well as DNA. Often, too, one or both starting nucleic acids are complex mixtures of fragments, such as total RNA from cells or fragments of total genomic DNA. Like cloned DNA, the starting nucleic acids are usually isolated from millions of cells and so individual sequences are normally present in many copies, often millions of copies.

A second, and very different, use of nucleic acid hybridization, one that was developed fairly recently, is to enable selective DNA purification. Here the object is to capture some DNA sequences of interest from a complex test DNA sample, such as total genomic DNA, so that they can be further studied in isolation, for example by DNA. We describe this type of application at the end of this section.

Hybridization assays: using known nucleic acids or oligonucleotides to find related sequences in a test nucleic acid population

Nucleic acid hybridization assays exploit the specificity of hybridization. A stable double-stranded hybrid (duplex) forms only when there is a significant amount of base pairing between the two sequences (which can be DNA, RNA, or oligonucleotide sequences). Because the stability of the duplex depends on the extent of base matching, assay conditions can be chosen to allow perfectly matched duplexes only or to allow degrees of base mismatching.

Hybridization assays can be carried out in many different ways, with multiple applications in both research and diagnostics. But there is a common underlying principle: a known, well-characterized population of nucleic acid molecules or synthetic oligonucleotides (the **probe** population) is used to interrogate an imperfectly understood population of nucleic acids (the test sample). To do that, as required, both nucleic acid populations must be separated into single strands and then mixed so that single probe strands can form artificial duplexes with complementary strands in the test sample.

Because the object of a hybridization assay is to use the probe to identify complementary or partially complementary test-sample strands, the probe–test-sample duplexes need to be labeled in some way so that they can be identified. To do that, either the probe or the test sample needs to be labeled at the outset (see Figure 6.12 for one approach, where it is the probe that is labeled). As described in Box 6.2, different systems can be used to label nucleic acids and oligonucleotides, but in the former case the methods normally involve incorporating labeled nucleotides during DNA or RNA synthesis.

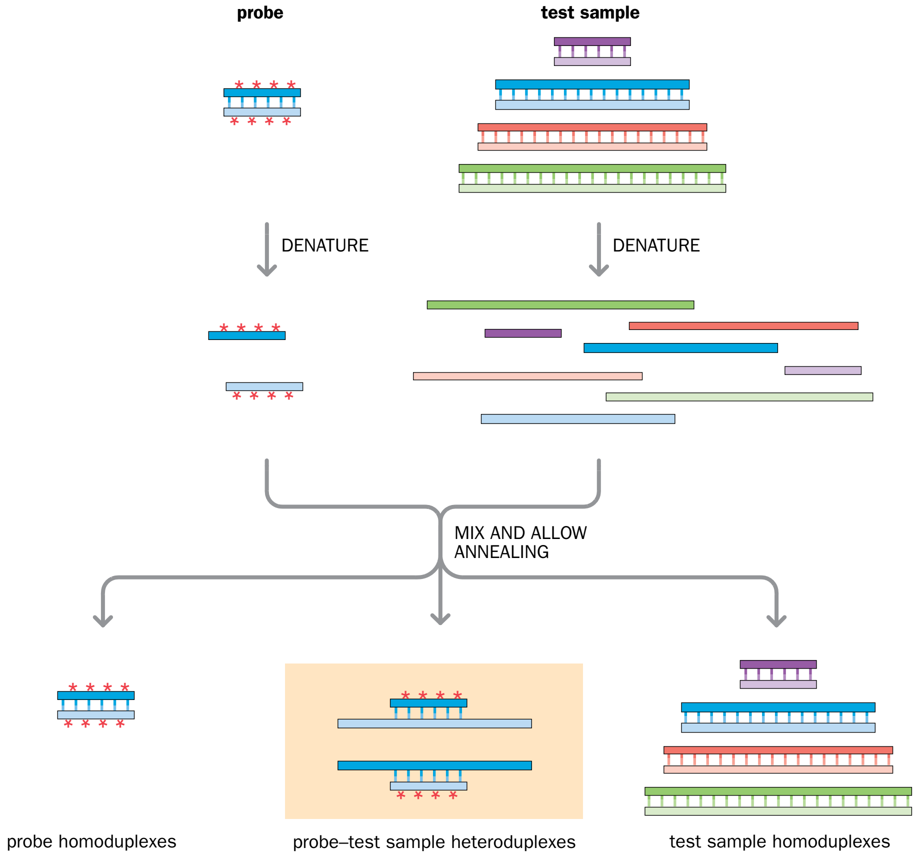


Figure 6.12 Heteroduplex formation in a nucleic acid hybridization assay. For ease of illustration we consider here a homogeneous probe population, consisting of a single type of defined nucleic acid that is labeled (shown by red asterisks), and an unlabeled test sample made up of many different nucleic acids. For the assay to work any double-stranded molecules need to be denatured to give single strands. Thereafter, single-stranded probe nucleic acids are mixed with single-stranded test sample nucleic acids. Strands with complementary sequences are then allowed to anneal. Many of the fragments that had previously been base paired in the two populations will re-anneal to reconstitute original homoduplexes (bottom left and bottom right). In addition, new artificial duplexes will be formed between (usually) partially complementary probe and test-sample sequences (bottom center). The hybridization conditions can be adjusted to favor formation of the novel duplexes. In this way, probes can selectively bind to and identify closely related nucleic acids within a complex nucleic acid population.

BOX 6.2 LABELING OF NUCLEIC ACIDS

Hybridization assays involve labeling of nucleic acids (or oligonucleotides) in either the probe or test-sample population. Often this involves synthesizing DNA or RNA in the presence of a nucleotide precursor carrying a distinctive group or *label* that is incorporated into the growing nucleic acid chain and that can be specifically detected in some way.

Labeled DNA copies can be made from a starting DNA or RNA using a suitable DNA polymerase in the presence of the four precursor deoxynucleotides (dATP, dCTP, dGTP, dTTP), at least one of which is labeled. In the case of a starting RNA, a reverse transcriptase uses the RNA as a template for making a labeled cDNA copy.

For some purposes, labeled RNA copies are made from a starting DNA using an RNA polymerase and the four precursor ribonucleotides (ATP, CTP, GTP, UTP), at least one of which is labeled. For example, labeled antisense RNA probes (riboprobes) are often used to track gene expression by hybridization to RNA transcripts in tissues (**Figure 7.10A** gives an example).

Unlike DNA or RNA, oligonucleotides are chemically synthesized by sequential addition of nucleotide residues to a starting nucleotide that will be the 3′ terminal nucleotide. Amine or thiol groups can be incorporated into the oligonucleotide and can then be conjugated with amine-reactive or thiol-reactive labels.

NONISOTOPIC LABELING

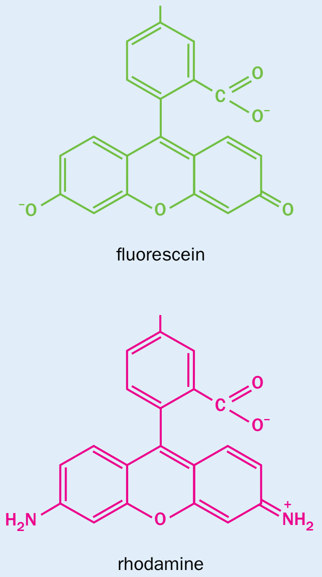
Different types of labeling system can be used. Radioisotope labeling offers high sensitivity of detection. Nonisotopic labeling is a convenient and safer alternative, and is now widely used. It involves incorporating nucleotides that contain a chemical group that may be directly or indirectly detected in some assay.

In a direct assay, the incorporated group directly serves as a label that is measured by the assay. Often a **fluorophore** is used, a chemical group that can readily be detected because it will absorb energy of a specific wavelength (excitation wavelength) and re-emit the energy at a longer, but equally specific wavelength (emission wavelength). Different fluorophores can be used, but various derivatives of fluorescein and rhodamine are especially popular—see **Figure 1** and **Table 1**.

In an indirect assay, the incorporated chemical group serves as a *reporter group* that is specifically recognized and bound by some affinity molecule, such as a dedicated antibody. The affinity molecule has a *marker group* bound to it, a chemical group or molecule that can be assayed in some way (see **Figure 2**).

The biotin–streptavidin system is widely used in indirect labeling of nucleic acids. Biotin, a vitamin, acts as the reporter group and gets incorporated into the nucleic acid by using a biotinylated nucleotide in the labeling reaction. (**Figure 3**, top). The bacterial protein streptavidin acts as the affinity molecule and binds to the biotin reporter (the affinity constant for bonding of streptavidin to biotin is 10^{-14}, one of the strongest known in biology). Another widely used reporter is digoxigenin, a steroid obtained from the *Digitalis* plant (**Figure 3**, bottom). A specific antibody raised against digoxigenin acts as the affinity molecule.

A variety of different marker groups or molecules can be conjugated to affinity molecules such as streptavidin and the digoxigenin antibody. They include various fluorophores that can be detected by fluorescence microscopy, or enzymes such as alkaline phosphatase and peroxidase that can permit detection via colorimetric assays or chemical luminescence assays.

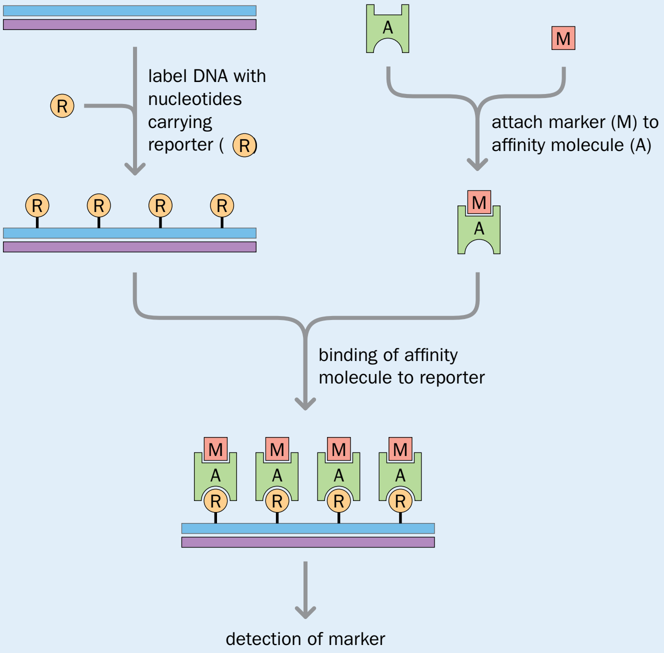


Box 6.2 Figure 1 Structure of two common fluorophores. TRITC and a variety of other fluorophores have been derived from rhodamine.

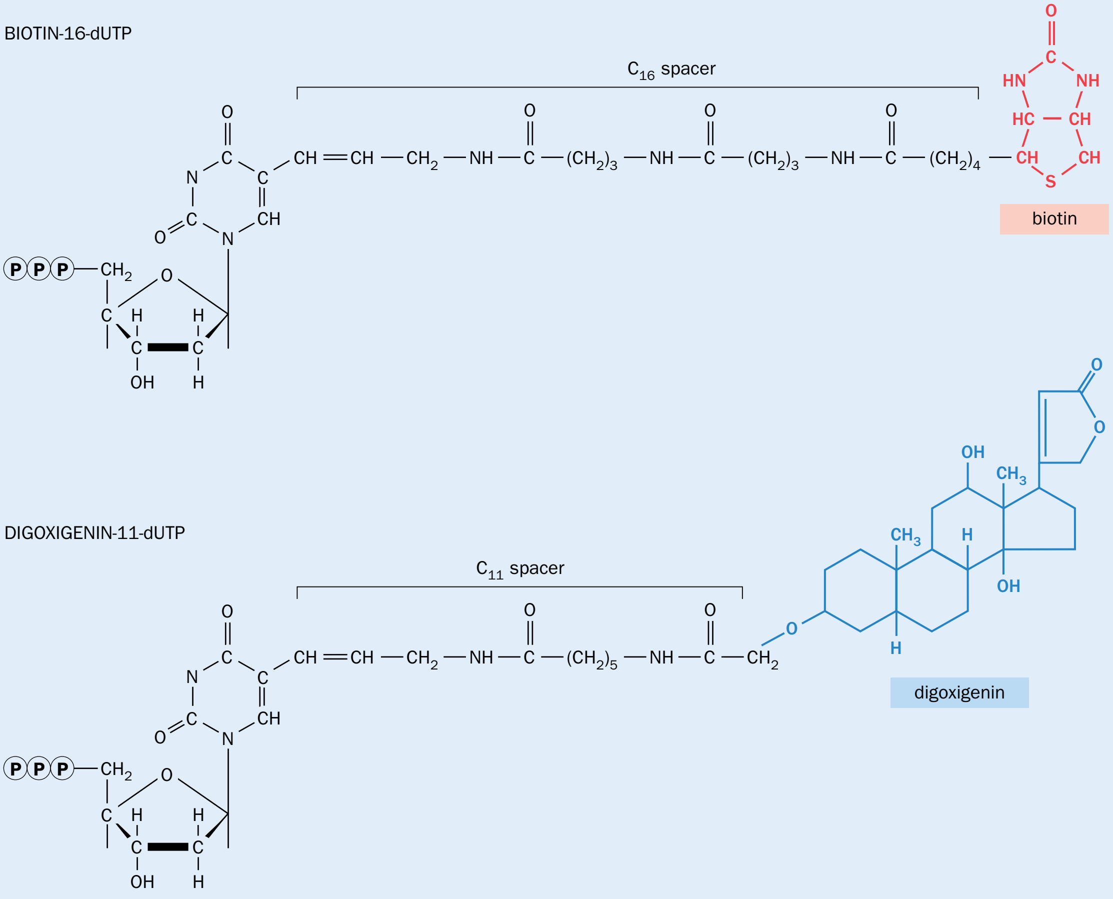
BOX 6.2 TABLE 1 FLUOROPHORES COMMONLY USED FOR LABELING NUCLEIC ACIDS

Fluorophore	Excitation maximum (nm)	Emission maximum (nm)
BLUE		
AMCA	350	450
DAPI	358	461
GREEN		
FITC	492	520
Fluorescein (see **Figure 1**)	494	523
RED		
CY3	550	570
TRITC	554	575
Rhodamine (see **Figure 1**)	570	590
Texas Red	596	620
CY5	650	670

AMCA, aminomethylcoumarin; DAPI, 4′,6-diamidino-2-phenylindole; FITC, fluorescein isothiocyanate; CY3, indocarbocyanine; TRITC, tetramethylrhodamine isothiocyanate; CY5, indodicarbocyanine.


A
M
R
label DNA with nucleotides carrying reporter (R)
attach marker (M) to affinity molecule (A)
R R R R
M
A
binding of affinity molecule to reporter
M M M M
A A A A
R R R R
detection of marker

Box 6.2 Figure 2 Indirect detection of labeled groups in nucleic acids. Incorporated reporter (R) groups are bound with high specificity by an affinity molecule (A) carrying a detectable marker (M). The marker can be detected in various ways. If it carries a specific fluorescent dye, it can be detected by fluorescence microscopy. A common alternative involves using an enzyme such as alkaline phosphatase to convert a substrate to give a colored product that is measured colorimetrically.



Box 6.2 Figure 3 Structure of biotin- and digoxigenin-modified nucleotides. The biotin and digoxigenin reporter groups shown here are linked to the 5′ carbon atom of the uridine of dUTP by spacer groups consisting, respectively, of a total of 16 carbon atoms (biotin-16-dUTP) or 11 carbon atoms (digoxigenin-11-dUTP). The spacer groups are needed to ensure physical separation of the reporter group from the nucleic acid backbone, so that the reporter group protrudes sufficiently far to allow the affinity molecule to bind to it.

Using high- and low-hybridization stringency

A hybridization assay can be used to identify nucleic acid sequences that are distantly related from a given nucleic acid probe. We might want to start with a DNA clone from a human gene and use that to identify the corresponding mouse gene. The human and mouse genes might be significantly different in sequence, but if we choose a long DNA probe and reduce the stringency of hybridization, stable heteroduplexes can be allowed to form even though there might be significant base mismatches (Figure 6.13A).

Conversely, we can choose more stringent hybridization conditions to accept only perfect base matching. If we choose an oligonucleotide probe, we can use a high-hybridization stringency so that the only probe–test duplexes that can form are ones that contain exactly the same sequence as the probe (Figure 6.13B). That can happen because a single mismatch out of, say, 18 base pairs can make the duplex thermodynamically unstable. Oligonucleotides can therefore be used to identify alleles that differ by a single nucleotide (allele-specific oligonucleotides).

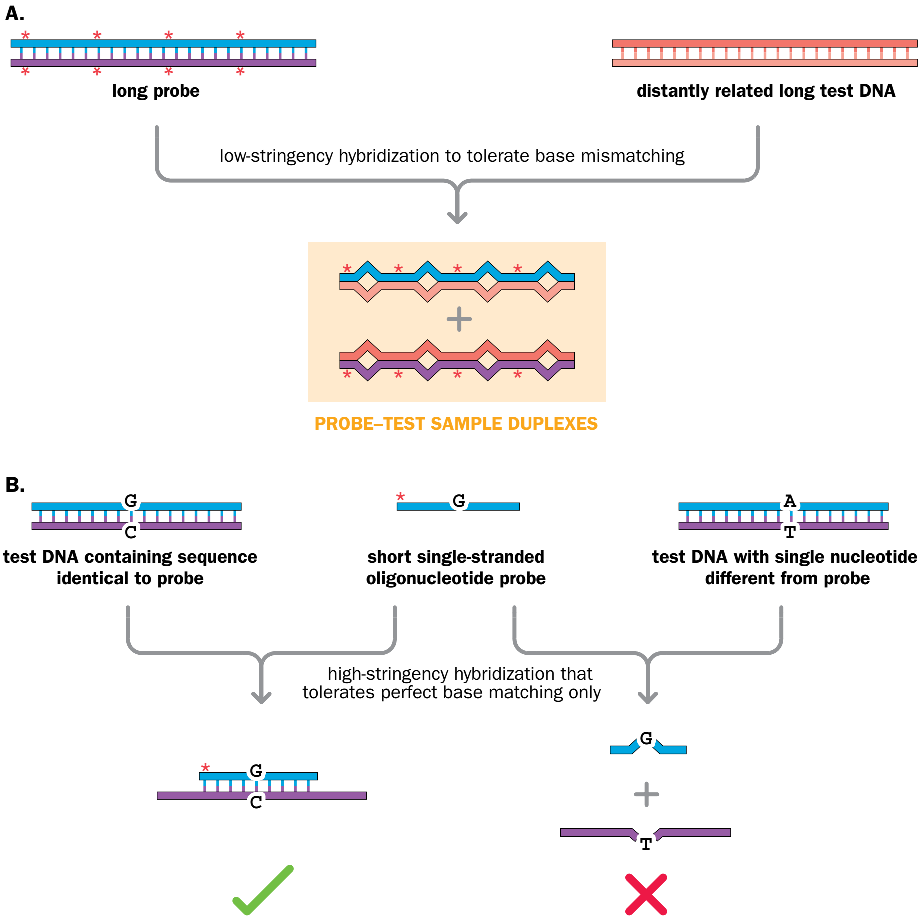

A.
long probe
distantly related long test DNA
low-stringency hybridization to tolerate base mismatching
+
PROBE–TEST SAMPLE DUPLEXES
B.
G
C
test DNA containing sequence identical to probe
G
short single-stranded oligonucleotide probe
A
T
test DNA with single nucleotide different from probe
high-stringency hybridization that tolerates perfect base matching only
G
C
G
+
T

Figure 6.13 Using low- or high-hybridization stringency to detect nucleic acid sequences that are distantly related or show perfect base matching with a given probe. In any hybridization assay we can control the degree of base matching between complementary strands in the probe and test sample. (**A**) If, for example, we increase salt concentrations and/or reduce the temperature, we lower hybridization stringency. In some circumstances a long probe strand can form a thermodynamically stable duplex with a comparable but distantly related strand within the test DNA (or RNA), even though there might be significant base mismatching. (**B**) Alternatively, we can use high temperatures and low salt concentrations to achieve high-hybridization stringency that might allow only perfect base matching. That is most easily achieved using a short oligonucleotide probe and it allows assays to discriminate between alleles that differ at a single nucleotide position. Labeling of nucleic acids and oligonucleotides is indicated by red asterisks.

Two classes of hybridization assay

There are many types of hybridization assay but they all fall into two broad classes, as listed below.

- Labeled probe–unlabeled test sample (left image in Figure 6.14). This has been the traditional type of hybridization. The probe is often a single type of cloned DNA, such as represented in Figure 6.12, and it is usually labeled by using a polymerase to synthesize complementary DNA or RNA strands in the presence of one or more labeled nucleotides.
- Unlabeled probe–labeled test sample (right image in Figure 6.14). This type of hybridization is sometimes called reverse hybridization because it is the reverse of the traditional approach. It has become popular since the development of microarrays, where unlabeled complex probe populations bound to a surface are used to interrogate a labeled test sample, as detailed below.

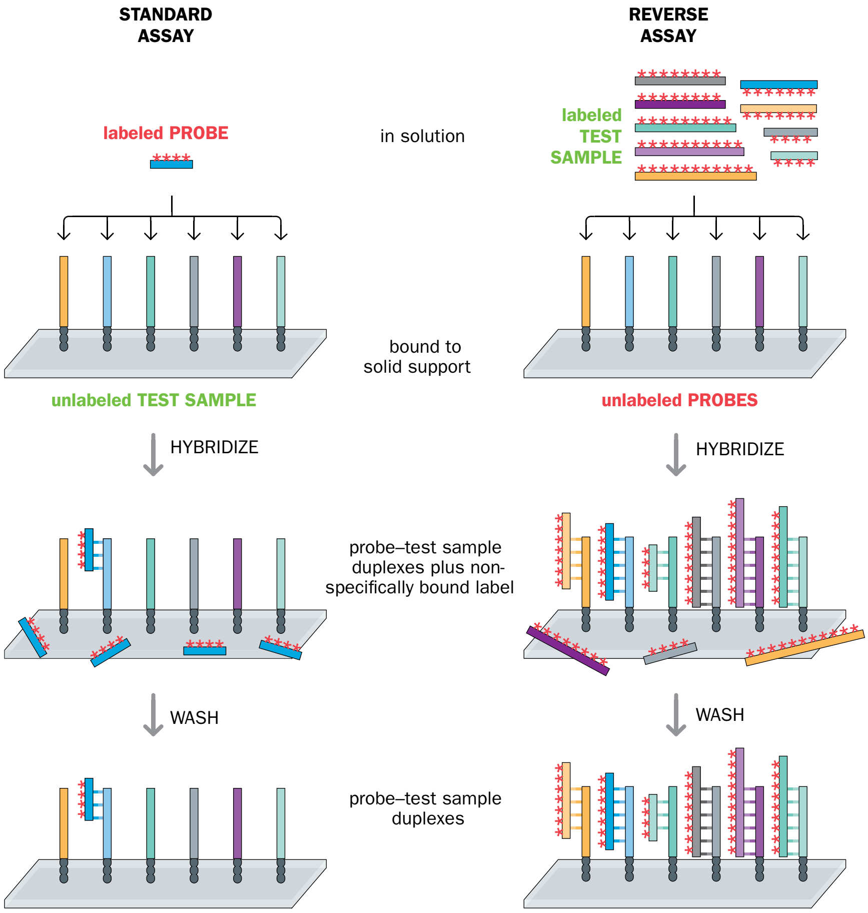


Figure 6.14 Standard and reverse hybridization assays and the use of solid supports to capture labeled probe–test sample duplexes. In both assays the unlabeled nucleic acid (or oligonucleotide) population is bound to a solid support and denatured and is exposed to an aqueous solution of the labeled nucleic acid (or oligonucleotide) population that has also, as required, been denatured. Single-stranded molecules in the labeled population can hybridize to complementary sequences in the unlabeled population and so become bound to the solid support. Other labeled sequences that have not bound, or have bound nonspecifically at incorrect locations on the support, can be washed off. The bound labeled nucleic acids can then be studied and are sometimes retrieved by washing at higher temperatures to break the hydrogen bonds connecting them to their unlabeled partner strands on the supports. In the past, most hybridization assays used the labeled probe/immobilized test sample scheme shown on the left (see **Table 6.3**), but microarray hybridization assays use the labeled test sample/immobilized probe scheme shown on the right (see **Figure 6.16**).

The point of using labeled nucleic acids in a hybridization assay is to allow probe–test sample heteroduplexes to be identified. But how can we distinguish between the label in these duplexes and the label in the original labeled probe or labeled test-sample DNA? The answer is to immobilize the unlabeled nucleic acid or oligonucleotide population on a solid support (often plastic, glass, or quartz) and expose it to an aqueous solution of the labeled nucleic acid population. When labeled nucleic acid strands hybridize to complementary sequences on the solid support, they will be physically bound to the support. However, labeled molecules that do not find a partner on the support (or that stick nonspecifically) can be washed off. That leaves behind the complementary partners that the assay is designed to find (see **Figure 6.14**). Hybridization assays are used for different purposes, as described later; see **Table 6.3** for some examples.

TABLE 6.3 POPULARLY USED HYBRIDIZATION ASSAYS

Hybridization class	Method	Applications	Examples
STANDARD ASSAY: labeled probe–unlabeled test sample	Southern blot	Looking for medium-size changes in genes/DNA (hundreds of base pairs to several kilobases) in test sample	**Figure 6.15**
	Tissue *in situ*	Tracking RNA transcripts in tissues and embryos	**Figure 7.10A,B**
	Chromosome *in situ*	Studying large-scale changes using fixed chromosomes on a slide as the test sample	**Figures 15.3** and **19.4**
REVERSE ASSAY: unlabeled probe–labeled test sample	Microarray CGH*	Genome-wide search for large (often megabase-sized) changes in DNA samples	**Figures 15.6** and **15.7**
	Microarray-based expression profiling	Simultaneously tracking expression of very many genes; used widely in cancer profiling	**Figure 7.11**

* Also called array CGH or aCGH, where CGH = comparative genome hybridization.

Standard hybridization assays using labeled probes to detect immobilized nucleic acids

Nucleic acid hybridization has long been popularly used to detect nucleic acids that have been immobilized in some way. In some cases, the nucleic acid samples have been isolated from cells and immobilized by binding them to a solid support such as a nylon membrane. That can be done very simply by injecting individual nucleic acid samples onto the membrane so that they are bound at regular intervals (dot blot hybridization). In other cases, the isolated nucleic acids may have been size-fractionated by gel electrophoresis before being transferred to the solid support. Alternatively, the hybridization assay is designed to detect nucleic acids *in situ* within native structures such as cells or chromosomes.

Southern and northern blot hybridization

In Southern hybridization assays, a sample population of purified DNA is digested with one or more restriction endonucleases, generating fragments that are several hundred to many thousands of base pairs in length. The restriction fragments are separated according to size by agarose gel electrophoresis, denatured, and transferred to a nitrocellulose or nylon membrane. Labeled probes are hybridized to the membrane-bound target DNA and the positions of the labeled heteroduplexes are revealed by autoradiography (**Figure 6.15**). Although this method has largely been superseded by PCR assays, it is still used in diagnostic assays that seek to identify large DNA changes (which are difficult to detect by PCR).

Northern blot hybridization is a variant of Southern blotting in which the samples contain undigested size-fractionated RNA instead of DNA. In the past, this method was regularly used to obtain information on which tissues genes were expressed in, and to identify tissue-specific isoforms. However, it has been superseded by RT-PCR and sequencing assays.

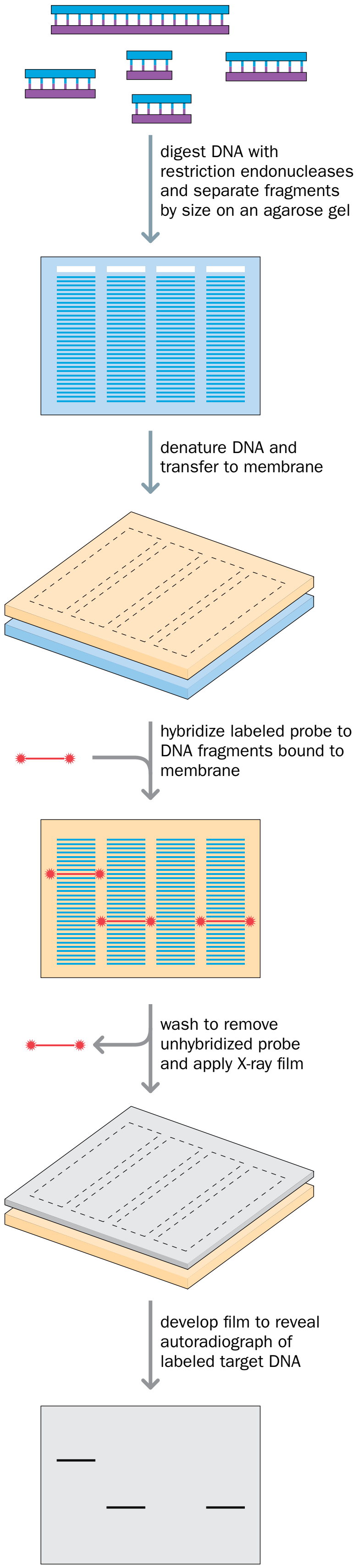

digest DNA with restriction endonucleases and separate fragments by size on an agarose gel
denature DNA and transfer to membrane
hybridize labeled probe to DNA fragments bound to membrane
wash to remove unhybridized probe and apply X-ray film
develop film to reveal autoradiograph of labeled target DNA

Figure 6.15 Southern blot hybridization. A complex DNA sample is digested with restriction endonucleases. The resulting fragments are applied to an agarose gel and separated by size using electrophoresis. The gel is treated with alkali to denature the DNA fragments and is then placed against a nitrocellulose or nylon membrane. DNA will be transferred from the gel to the membrane, which is then soaked in a solution containing a radiolabeled, single-stranded DNA probe. After hybridization, the membrane is washed to remove excess probe and then dried. The membrane is then placed against an X-ray film and the position of the labeled probe will cause a latent image on the film that can be revealed by development of the autoradiograph as a hybridization band.

Chromosome *in situ* hybridization

In **chromosome *in situ* hybridization**, a suitable labeled DNA probe is hybridized against chromosomal DNA that has been denatured *in situ*. First, an air-dried, microscope-slide chromosome preparation is made, typically using metaphase or prometaphase chromosomes from peripheral blood lymphocytes or lymphoblastoid cell lines. RNA and protein are then removed from the sample by treatment with RNase and proteinase K, and the remaining chromosomal DNA is denatured by exposure to formamide. The denatured DNA is exposed to an added solution containing a labeled nucleic acid probe and overlaid with a coverslip.

Depending on the particular technique, chromosome banding of the chromosomes can be arranged either before or after the hybridization step. As a result, the signal obtained after removal of excess probe can be correlated with the chromosome band pattern in order to identify a map location for the DNA sequences recognized by the probe. Chromosome *in situ* hybridization has been revolutionized by the use of fluorescently labeled probes in fluorescence *in situ* hybridization (FISH) techniques; see **Table 6.3** for cross-references to some figures containing example images.

Tissue *in situ* hybridization

In **tissue *in situ* hybridization**, labeled probes are hybridized against RNA in tissue sections. Very thin tissue sections are cut, either from paraffin wax-embedded tissue blocks or from frozen tissue using a cryostat, and then mounted onto glass slides. A hybridization mix including the probe is applied to the section on the slide and covered with a glass coverslip. Single-stranded antisense RNA probes (riboprobes; see **Box 6.2**) are preferred.

The riboprobes can be labeled with a radioisotope such as ^{33}P or ^{35}S, and the hybridized probes visualized using autoradiographic procedures. The localization of the silver grains is often visualized using only dark-field microscopy. Here, direct light is not allowed to reach the objective; instead, the illuminating rays of light are directed from the side so that only scattered light enters the microscopic lenses and the signal appears as an illuminated object against a black background. However, better signal detection is possible using bright-field microscopy, in which the image is obtained by direct transmission of light through the sample. Alternatively, probes are subjected to fluorescence labeling and detection is accomplished by fluorescence microscopy.

Microarray hybridization: large-scale parallel hybridization of labeled test-sample nucleic acids to immobilized probes

Innovative and powerful hybridization technologies developed in the early 1990s permit numerous hybridization assays to be simultaneously conducted on a common sample under the same conditions. A DNA or oligonucleotide microarray consists of many thousands or millions of different unlabeled DNA or oligonucleotide probe populations that have been fixed to glass or another suitable surface within a high-density grid format. Within each grid square are large numbers of identical copies of just one probe (a grid square with its probe population is called a **feature**). For example, oligonucleotide microarrays often have a 1.28 cm × 1.28 cm surface that contains very many different features that each occupy about 5 or 10 µm^2 (**Figure 6.16**).

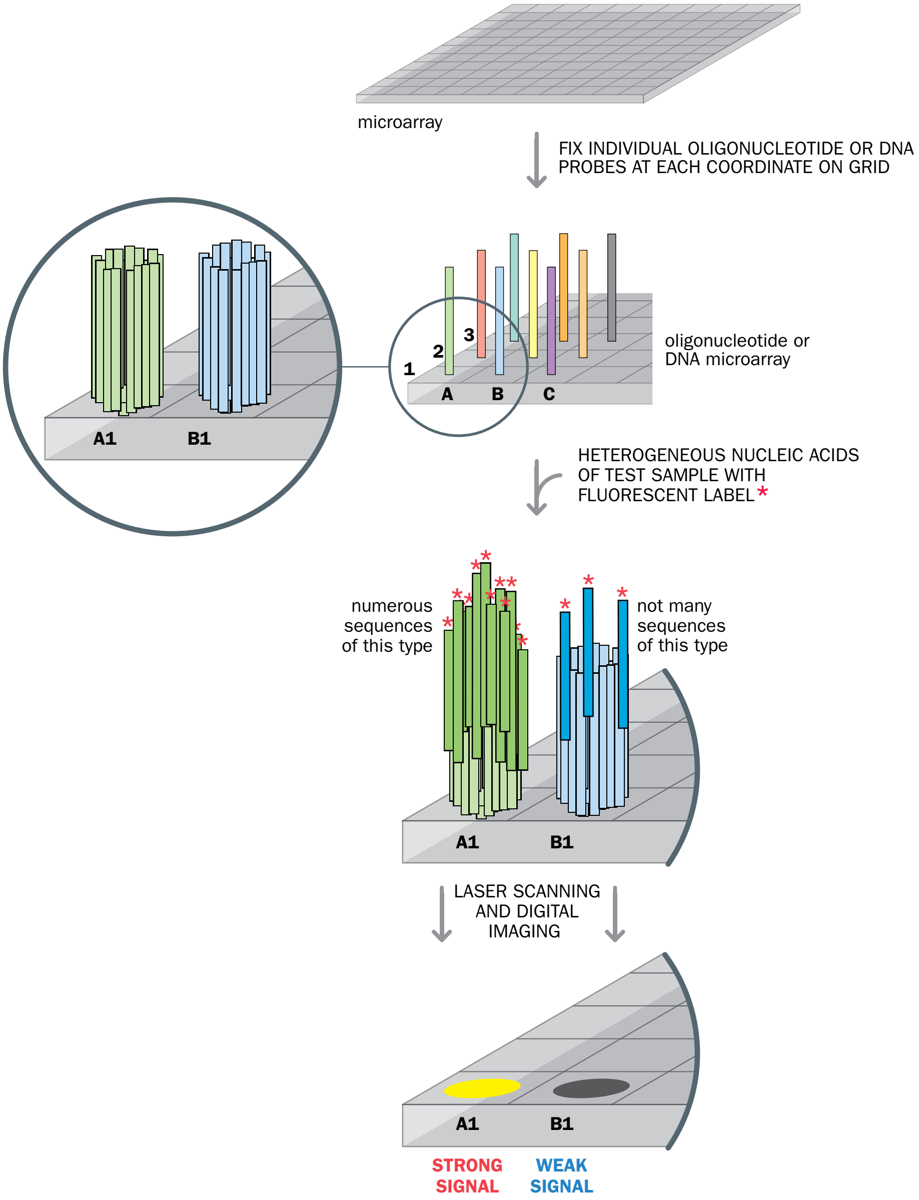

microarray
FIX INDIVIDUAL OLIGONUCLEOTIDE OR DNA PROBES AT EACH COORDINATE ON GRID
A1
B1
1
2
3
A
B
C
oligonucleotide or DNA microarray
HETEROGENEOUS NUCLEIC ACIDS OF TEST SAMPLE WITH FLUORESCENT LABEL *
numerous sequences of this type
not many sequences of this type
A1
B1
LASER SCANNING AND DIGITAL IMAGING
A1
B1
STRONG SIGNAL
WEAK SIGNAL

Figure 6.16 Principle of microarray hybridization. A microarray is a solid surface on which molecules can be fixed at specific co-ordinates in a high-density grid format. Oligonucleotide or DNA microarrays have thousands to millions of different synthetic, single-stranded oligonucleotide or DNA probes fixed at specific pre-determined positions in the grid. As shown by the expanded item enclosed within the circle (left), each grid square will have many identical copies of a single type of oligonucleotide or DNA probe (a *feature*). An aqueous test sample containing a heterogeneous collection of labeled DNA fragments or RNA transcripts is denatured and allowed to hybridize with the probes on the array. Some probes (e.g., the A1 feature) may find numerous complementary sequences in the test population, resulting in a strong hybridization signal; for other probes (e.g., the B1 feature) there may be few complementary sequences in the test sample, resulting in a weak hybridization signal. After washing and drying of the microarray, the hybridization signals for the numerous different probes are detected by laser scanning, giving huge amounts of data from a single experiment. (For ease of illustration, we show test-sample nucleic acids with end labels, but normally they would contain labels on internal nucleotides.)

A test sample—an aqueous solution containing a complex population of fluorescently labeled denatured DNA or RNA—is hybridized to the different probe populations on the microarray. After a washing step to remove nonspecific binding of labeled test-sample molecules to the array, the remaining bound fluorescent label is detected using a high-resolution laser scanner. The signal emitted from each feature on the array is analyzed using digital imaging software that converts the fluorescent hybridization signal into one of a palette of colors according to its intensity (see **Figure 6.16**).

Because the intensity of each hybridization signal reflects the amount of labeled molecules that have bound to a feature, microarray hybridization is used to quantitate different sequences in complex test-sample populations such as different samples of genomic DNA or total cellular RNA (or cDNA). As described in later chapters, frequent applications include quantifying different transcripts (expression profiling) and also scanning genomes to look for large-scale deletions and duplications.

Nucleic acid hybridization as a method to selectively purify desired nucleic acid sequences

Up until this point we have considered how nucleic acid hybridization is commonly used as an assay. That is, a well-known nucleic acid or oligonucleotide probe population is used to interrogate some other, less understood nucleic acid population, and the object is simply to get some information about the latter population (we might want to know the sizes of the hybridizing fragments, or their copy number, or where they are expressed in cells and tissues, and so on). But we can also use nucleic acid hybridization for a different purpose: to selectively purify a desired type of nucleic acid sequence. That is made possible by covalently attaching to the probe molecules some other molecule that can bind with very high specificity to "capture molecules" immobilized on a surface.

The most popular approach is to covalently attach biotin to the probe. Biotin, a naturally occurring vitamin (known as vitamin B7 or vitamin H), just happens to have an extraordinarily high affinity for streptavidin, a protein that originates from a *Streptomyces* bacterium. (A homotetramer of streptavidin can bind biotin with a dissociation constant of about 10^{-14} mol/L, one of the strongest noncovalent interactions between natural biological molecules; the strength of the interaction depends on the formation of numerous hydrogen bonds and van der Waals interactions between biotin and streptavidin.)

After biotin-linked probe molecules have been allowed to hybridize to complementary sequences in a test sample, the probe–test-sample heteroduplexes can be captured using magnetized beads coated with streptavidin. The magnetized beads, with attached probe–test-sample heteroduplexes, can then be selectively removed using a magnet and the desired nucleic acids can be eluted (**Figure 6.17**).

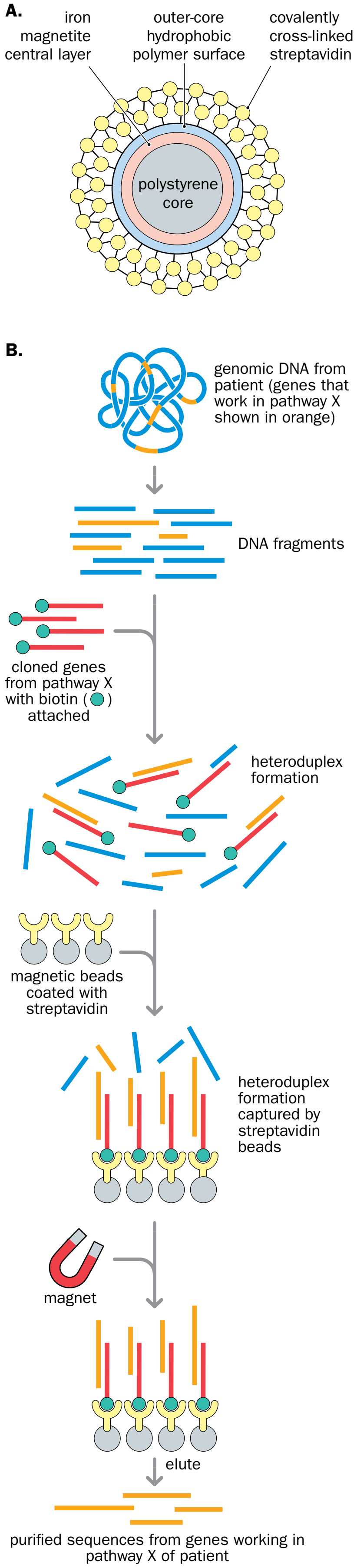

A.
iron magnetite central layer
outer-core hydrophobic polymer surface
covalently cross-linked streptavidin
polystyrene core
B.
genomic DNA from patient (genes that work in pathway X shown in orange)
DNA fragments
cloned genes from pathway X with biotin (●) attached
heteroduplex formation
magnetic beads coated with streptavidin
heteroduplex formation captured by streptavidin beads
magnet
elute
purified sequences from genes working in pathway X of patient

Figure 6.17 Selective purification of desired nucleic acid sequences using nucleic acid hybridization. (**A**) Structure of streptavidin-coated magnetic beads. (**B**) Example of purification. Here we imagine a situation where we wish to purify sequences from a family of many genes that work in some common cell signaling pathway, X, and are considered as possible candidates for having a disease-causing mutation in a group of patients. The idea is to take genomic DNA from individual patients, fragment it, and then mix the fragments with a combination of cloned DNA sequences representing all the normal genes that work in pathway X. The DNA sequences in the mixture are then denatured and allowed to re-anneal, allowing heteroduplexes to form between biotinylated pathway X sequences and complementary sequences from the patient. After streptavidin-coated magnetic beads are added, heteroduplexes with an end-biotin group are bound via the streptavidin to the beads, and can be selectively removed using a magnet. The desired X pathway gene sequences in the patient can be eluted by heating the sample so as to break the hydrogen bonds of the heteroduplex. The resulting purified sequences from each patient can then be investigated by DNA sequencing.

6.4 DNA SEQUENCING PRINCIPLES AND SANGER DIDEOXY SEQUENCING

DNA sequencing means working out the linear sequence of the four bases (A, C, G, and T) in fragments of DNA, usually fragments that have been amplified by PCR or DNA cloning. By working out the base sequence of many short DNA fragments with overlapping DNA sequences, one can get complete sequences for whole genes and whole genomes. The standard method does not distinguish between modified bases and the unmodified version of the base (such as between 5-methylcytosine and cytosine, or between N^6-methyladenine and adenine)—more specialized methods can do that job, as described for 5-methylcytosine/cytosine in **Box 10.3**. Note that DNA sequencing is also used to infer the linear sequence of bases in RNA molecules (the RNA must first be converted into a complementary DNA sequence using a reverse transcriptase).

Until quite recently, dideoxy DNA sequencing was essentially the only DNA sequencing method that was used (it was first published in 1997 by Fred Sanger, earning him his second Nobel Prize). It relies on amplifying individual DNA sequences that have been purified by DNA cloning or PCR. For each amplified DNA, nested sets of labeled DNA copies are made and then separated according to size using gel electrophoresis.

In the last few years, completely different technologies have allowed massively-parallel DNA sequencing. No attempt is made to obtain the sequence of just a purified DNA component; instead, millions of DNA fragments present in a complex DNA sample are simultaneously sequenced without the need for gel electrophoresis.

Dideoxy DNA sequencing provides highly accurate DNA sequences and is still widely used for investigating specific DNA sequences, such as testing whether individuals have mutations in a particular gene or confirming a suspected mutation. What the newer DNA sequencing technologies offer is a dramatic increase in sequencing capacity and the ability to sequence complex DNA populations, such as genomic DNA sequences, very rapidly. As a result of fast-developing technology, the running costs of DNA sequencing are plummeting and very rapid sequencing of whole exomes and even whole genomes is becoming routine.

The basics of Sanger dideoxy DNA sequencing

Like PCR, dideoxy DNA sequencing uses primers and a DNA polymerase to make DNA copies of specific DNA sequences of interest. To obtain enough DNA for sequencing, the DNA sequences are amplified using PCR (or sometimes by cloning in cells). The resulting purified DNAs are then sequenced, one after another, in individual reactions. Each reaction begins by denaturing a selected purified DNA. A single oligonucleotide primer is then allowed to bind and used to make labeled DNA copies of the desired sequence (using a provided DNA polymerase and the four dNTPs).

Instead of making full-length copies of the sequence, the DNA synthesis reactions are designed to produce a population of DNA fragments sharing a common 5′ end sequence (defined by the primer sequence) but with variable 3′ ends. This is achieved by simultaneously having present the standard dNTP precursors of DNA plus low concentrations of ddNTPs, dideoxynucleotide analogs that differ from a standard deoxynucleotide only in that they lack an OH group at the 3′ carbon of the sugar as well as at the 2′ carbon (**Figure 6.18A**).

DNA synthesis continues smoothly when dNTPs are used, but once a dideoxynucleotide is incorporated into a growing DNA chain, chain synthesis is immediately terminated (the dideoxynucleotide lacks a 3′-OH group to form a phosphodiester bond). To keep the balance tilted toward chain elongation, the ratio of each ddNTP to the corresponding dNTP is set to be about 1:100, so that a dideoxynucleotide is incorporated at only about 1% of the available nucleotide positions.

If we consider competition between ddATP and dATP in the example in **Figure 6.18B**, there are four available positions for nucleotide insertion: opposite the T at nucleotide positions 2, 5, 13, and 16 in the starting DNA. Because the DNA synthesis reaction results in numerous DNA copies, then, by chance, some copies will have a dideoxyA incorporated opposite the T at position 2, some will have a dideoxyA opposite the T at position 5, and so on. Effectively, chain elongation is randomly inhibited, producing sets of DNA strands that have a common 5′ end but variable 3′ ends.

Fluorescent dyes are used to label the DNA. One convenient way of doing this, as shown in **Figure 6.18B**, is to arrange that the four different ddNTPs are labeled with different fluorescent dyes. The reaction products will therefore consist of DNA strands that have a labeled dideoxynucleotide at the 3′ end, carrying a distinctive fluorophore according to the type of base incorporated.

All that remains is to separate the DNA fragments according to size using electrophoresis (see **Box 6.3**), and then to detect the fluorescence signals. The latter is usually achieved during gel electrophoresis: as the migrating DNA fragments reach a certain point in the gel, they pass a laser that excites the fluorophores, causing them to emit fluorescence at distinct wavelengths. The fluorescence signals are recorded, and an output is provided in the form of intensity profiles for the differently colored fluorophores (see **Figure 6.18C** for an example).

Dideoxy DNA sequencing is disadvantaged by relying on gel electrophoresis. It is not amenable, therefore, to full automation (slab polyacrylamide gels were used initially;

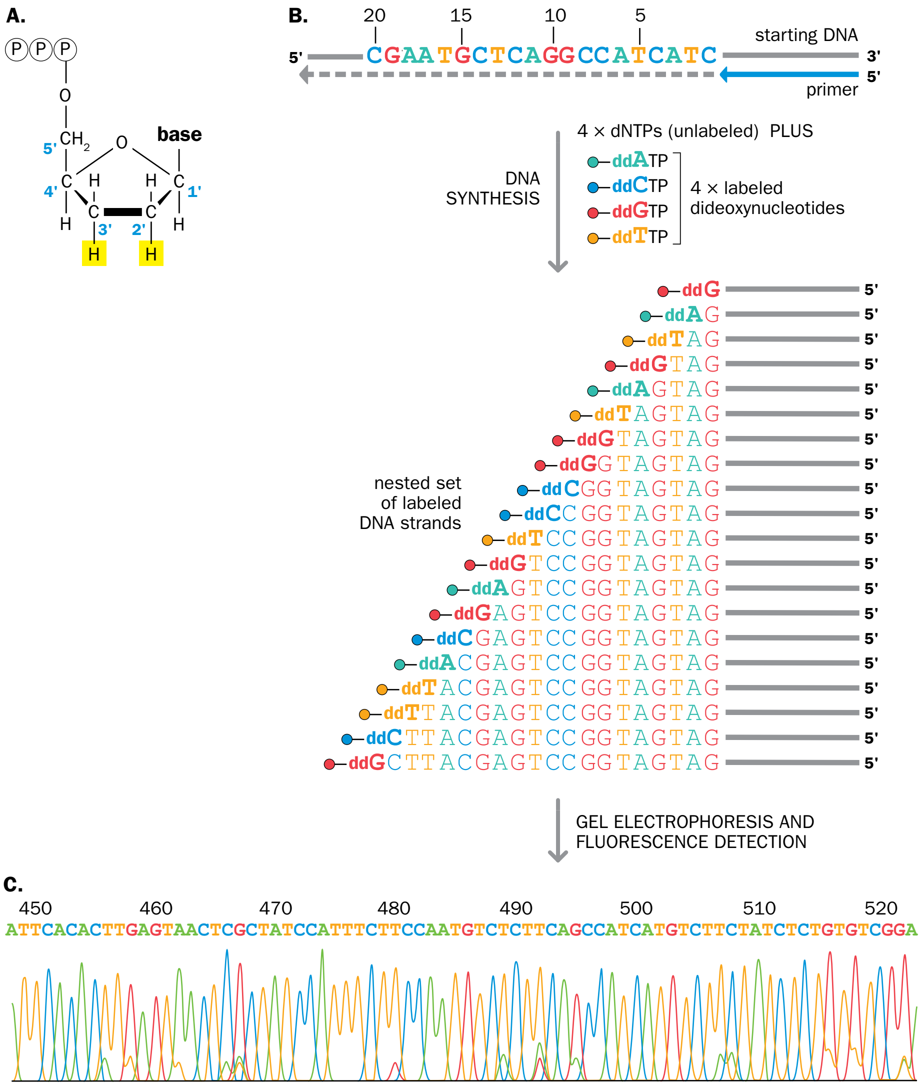


Figure 6.18 Principle of dideoxy sequencing. (**A**) Generalized structure of a 2′,3′ ddNTP. The sugar is *dideoxy*ribose because the hydroxyl groups attached to both carbons 2′ and 3′ of ribose are each replaced by a hydrogen atom (shown by yellow shading). (**B**) In dideoxy sequencing reactions, a DNA polymerase uses an oligonucleotide primer to make complementary sequences from a purified, single-stranded starting DNA. The sequencing reactions include ddNTPs that compete with the standard dNTPs for insertion of a nucleoside monophosphate into the DNA. Different labeling systems can be used but it is convenient to use labeled ddNTPs that have different fluorescent groups according to the type of base, as shown here by colored circles. The DNA copies will have a common 5′ end (defined by the sequencing primer) but variable 3′ ends, depending on where a labeled dideoxynucleotide has been inserted, producing a nested set of DNA fragments that differ by a single nucleotide in length. A series of nested fragments that incrementally differ by one nucleotide from their common 5′ end are fractionated according to size by gel electrophoresis, and the fluorescence signals are recorded and interpreted to produce a linear base sequence. (**C**) Example of DNA sequence output, showing a succession of dye-specific (and therefore base-specific) intensity profiles. This example shows a cDNA sequence from the *PHC3* polyhomeotic gene, provided by E. Tonkin, Newcastle University.

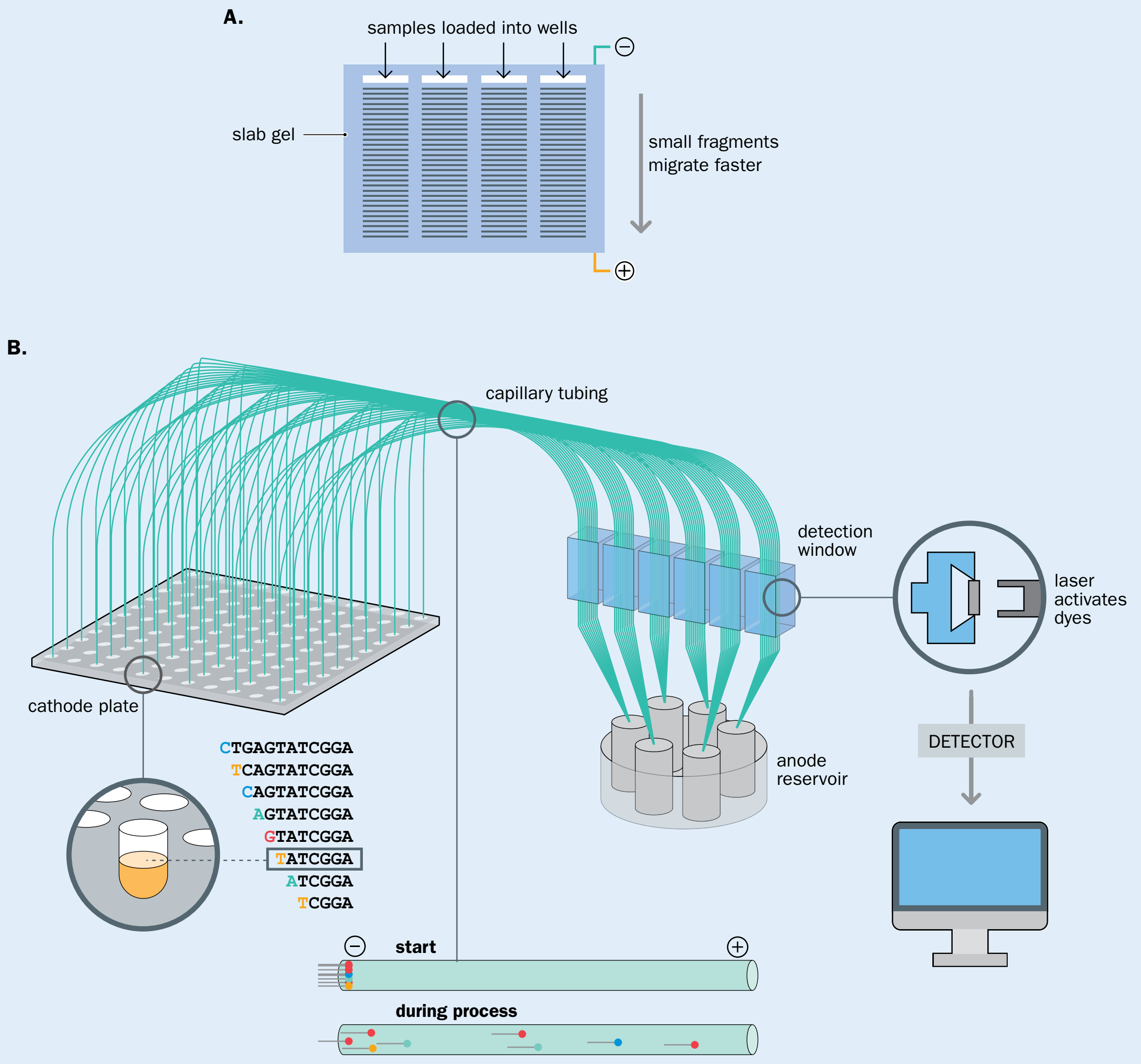
BOX 6.3 SLAB GEL AND CAPILLARY ELECTROPHORESIS FOR SEPARATING NUCLEIC ACIDS ACCORDING TO SIZE

Nucleic acids carry numerous negatively-charged phosphate groups and will migrate toward the positive electrode when placed in an electric field. By arranging that they migrate through a porous gel during electrophoresis, nucleic acid molecules can be fractionated according to size. The porous gel acts as a sieve: small molecules pass easily through the pores of the gel, but larger fragments are more impeded by frictional forces.

Standard gel electrophoresis using agarose gels allows fractionation of moderately large DNA fragments (usually from about 0.1 kb to 20 kb). Pulsed-field gel electrophoresis can be used to separate much larger DNA fragments (up to megabases long). It uses specialized equipment in which the electrical polarity is regularly changed, forcing the DNA molecules to periodically alter their conformation in preparation for migrating in a different direction. Polyacrylamide gel electrophoresis allows superior resolution of smaller nucleic acids (it is usually used to separate fragments in size ranges up to 1 kb), and is used in dideoxy DNA sequencing to separate fragments that differ in length by just a single nucleotide.

In slab gel electrophoresis, individual samples are loaded into cut-out wells at one end of a solid slab of agarose or polyacrylamide gel. They migrate in parallel lanes toward the positive electrode (**Figure 1A**). The separated nucleic acids can be detected in different ways. For example, after the end of an electrophoresis run, the gels can be stained with chemicals such as ethidium bromide or SYBR green that bind to nucleic acids and fluoresce when exposed to ultraviolet light. Sometimes the nucleic acids are labeled with fluorophores prior to electrophoresis, and during electrophoresis a recorder detects fluorescence of individual labeled nucleic acid fragments as they sequentially pass a recorder placed opposite a fixed position in the gel.

The disadvantage of slab gel electrophoresis is that it is labor-intensive. The modern trend is to use capillary gel electrophoresis, which is largely automated. Fluorescently-labeled DNA samples migrate through individual long and very thin tubes containing polyacrylamide gel, and a recorder detects fluorescence emissions as samples pass through a fixed point (**Figure 1B**). Modern Sanger DNA sequencing uses capillary electrophoresis, as do many different types of diagnostic DNA screening methods that we outline in Chapter 20.



Box 6.3 Figure 1 **(A) Slab gel electrophoresis and (B) capillary gel electrophoresis.** (B, courtesy of the Life Sciences Foundation.)

more modern machines use semi-automated capillary electrophoresis—see **Box 6.3 Figure 1B**). Because gel electrophoresis is not suitable for handling large numbers of samples at a time, dideoxy sequencing can generate a limited amount of sequence data only. It is, therefore, not well suited to genome sequencing. It does, however, provide highly accurate sequences over several hundred bases. In modern times it is often used for analyzing variation over small DNA regions, such as regions encompassing individual exons, and in validating some sequences obtained by newer, high-throughput DNA sequencing methods.

6.5 MASSIVELY-PARALLEL DNA SEQUENCING (NEXT-GENERATION SEQUENCING)

For thirty years, Sanger's dideoxy technique was the standard method of DNA sequencing. Over the years, incremental technical improvements saw radiolabeling replaced by fluorescence labeling, and the development of automated capillary sequencers with increasing numbers of capillaries per machine, but the basic principle was unchanged. Sanger sequencing is a highly reliable way of sequencing a pre-determined stretch of DNA up to around 800 bp in length. Although it was used for obtaining the sequence of the human genome and that of some other animal genomes, in the "post-genome era" it is now widely used for re-sequencing of amplified exons or other small DNA sequences (for example, to confirm a suspected DNA variant or pathogenic mutation).

Starting in 2005, a number of revolutionary new techniques, often collectively termed next-generation sequencing, burst upon the scene to transform both the scale and applications of DNA sequencing. They differ in many ways from Sanger DNA sequencing (**Table 6.4**), but the most significant characteristic is a vast increase in the amount of sequence data per run (the sequence throughput), with dramatic reductions in the cost of sequencing per megabase. As a result, the scale of sequencing could move from individual exons and genes (which usually requires sequencing of just a few different amplified DNA fragments) to whole exomes, whole genomes, and whole transcriptomes (where huge numbers of different DNA fragments are sequenced simultaneously).

TABLE 6.4 PRINCIPAL DIFFERENCES BETWEEN SANGER DIDEOXY SEQUENCING AND MASSIVELY-PARALLEL DNA SEQUENCING

Property	Sanger dideoxy sequencing	Massively-parallel DNA sequencing
Gel electrophoresis	Integral part of process	Not used
Automation of sequencing process	At most, semi-automated (when using capillary gel electrophoresis—see **Box 6.3 Figure 1B**)	Fully automated
Recording of sequence during sequencing reaction	Not possible. The nested set of fragments synthesized during each sequencing reaction must be size-separated by subsequent gel electrophoresis	The sequence is recorded during the sequencing reaction
DNA template preparation	DNA fragments are cloned in cells or amplified by PCR	DNA fragments may be ligated to adaptor oligonucleotides and amplified by PCR (see **Figure 6.10**); or single, unamplified DNA fragments are sequenced
Read lengths	Up to 800 nucleotides	From 35 up to 14,000 nucleotides (see **Figure 6.19**)
Number of DNA templates sequenced per reaction	One	Tens of thousands to billions of DNA fragments are sequenced simultaneously
Sequencing throughput	Advanced machines (ABI Prism 3200) allow 96 sequencing reactions per electrophoresis run; that is, a throughput of about 80 Mb of DNA per run	From 80 Mb to 1000 Gb per run, according to the sequencing platform (see **Figure 6.19**)
Sequencing error rates	Very low	Significantly high for individual fragments, but comparing multiple reads can give consensus sequence with acceptably low error rates (see **Figure 6.20**)
Applications	Routine, small-scale DNA sequencing. Limited by the small number of different DNA templates that can be sequenced at a time	Especially suited to global analyses, such as whole genomes, exomes, transcriptomes, and so on

Sanger sequencing is limited in its applications, but the new high-throughput DNA sequencing methods are extremely versatile. The applications include: whole genome sequencing; targeted DNA sequencing (defined subsets of the genome such as the exome or all genes that work in a disease-associated pathway); whole transcriptome sequencing (after total RNA is converted into cDNA); targeted transcriptome sequencing (defined subsets, such as transcripts from all genes known to regulate cell division); Methyl-Seq (to identify sites of DNA methylation across the genome); and so on—see Table 1 of Shendure & Aiden (2012) (PMID 23138308; see Further Reading) for a more comprehensive list.

Within the categories listed in the previous paragraph, one can compare different sources of DNA in different ways. In whole genome sequencing, for example, the genomes of different human individuals can be compared across different existing populations (population genomics, clinical genomics) and across human history and pre-history (human evolutionary genomics). Transcriptome sequencing may be targeted at different cell types, and even from single cells, or from the same cell type at different stages of disease and normality. In short, the potential of DNA sequencing to answer scientific and clinical questions has been revolutionized, and we provide many examples in later chapters.

As described below, different competing companies have developed different technologies, but all the next-generation methods share one key feature, namely that they are massively parallel—that is, they sequence millions of DNA fragments simultaneously (as opposed to the "one-DNA-fragment-at-a-time" limitation of Sanger sequencing). As a result, the sequence throughput is vastly greater than can be achieved using Sanger dideoxy sequencing. Apart from cost and throughput, important differences between the available technologies center on the average read length (the length of an average sequence read) and error rate. In the sections that follow, we describe the more popular massively-parallel DNA sequencing platforms. They can be classified into two broad groups:

- Methods that sequence PCR-amplified DNA templates. They provide from very short (35 nucleotides) to medium-length sequences (up to 800 nucleotides). Many of them involve sequencing-by-synthesis approaches where new DNA strands are synthesized using, as templates, amplified DNA sequences from the DNA sample. They exhibit high to very high sequence throughput (from close to 1000 Mb up to nearly 1,000,000 Mb of DNA—see **Figure 6.19**);
- Methods that sequence unamplified DNA templates ("single-molecule sequencing"). They enable sequences that are many thousands of nucleotides long to be obtained, but sequence throughput is more modest (100–1000 Mb of DNA). Because single molecules are sequenced, potential problems relating to amplification bias are avoided (some of the original DNA sequences in a starting DNA sample may be underrepresented or overrepresented after PCR amplification).

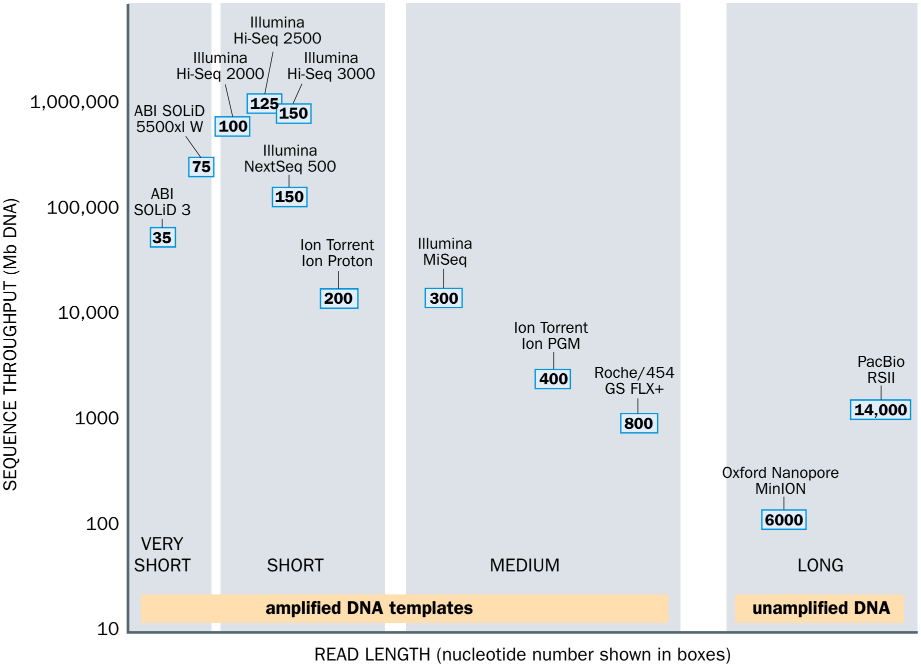


Figure 6.19 Sequence throughput and read lengths for some common massively-parallel DNA sequencing platforms. For some methods the accuracy of sequencing is low for individual sequence reads but can be compensated for by having high sequence coverage (when each region of the DNA is represented by many individual sequence reads). The maximum sequence throughput of close to 1,000,000 Mb of DNA corresponds to 1000 Gb or about 300 haploid human genomes. (Data from Reuter JA *et al*. [2015] *Mol Cell* **58**:586–597; PMID 26000844.)

For many of the methods that sequence PCR-amplified DNA, the sequencing reads are much shorter than in Sanger dideoxy sequencing (but companies are putting great efforts into increasing them). That is an important consideration because of the need to assemble finalized consensus sequences from individual sequence reads for millions of random fragments. The shorter the sequencing reads are, the greater the problems of assembly (ordering individual sequences according to perceived overlaps between different neighboring sequences), which can be a particular concern where the starting DNA is particularly complex (such as a large genome with high-copy-number repetitive DNA sequences). Assembly can be either *de novo* or by alignment against a previously obtained reference sequence (re-sequencing).

In human molecular genetics, re-sequencing involves aligning sequence reads from individuals against the most recent Reference Human Genome Sequence, which at the time of writing is Genome Reference Consortium Human Build 38 patch 12 (see https://www.ncbi.nlm.nih.gov/assembly/GCA_000001405.27). Compared to Sanger sequencing, massively-parallel sequencing methods have high error rates, and assembling short error-prone sequence reads into the final consensus sequence may not be

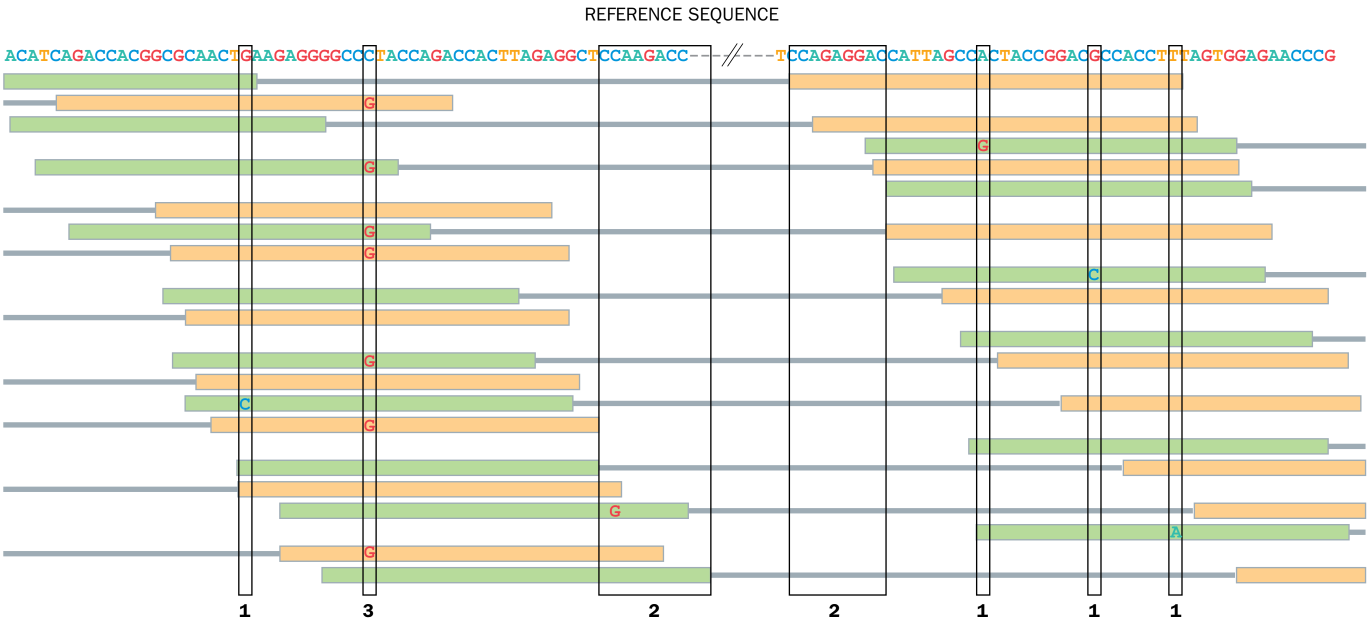

REFERENCE SEQUENCE
ACATCAGACCACGGCGCAACTGAAGAGGGGCCCTACCAGACCACTTAGAGGCTCCAAGACC---//---TCCAGAGGACCATTAGCCACTACCGGACGCCACCTTTAGTGGAGAACCCG
G
G
G
G
G
C
G
C
G
G
A
G
1
3
2
2
1
1
1

Figure 6.20 Aligning short sequence reads to a reference sequence. Some high-throughput sequencing platforms have quite short read lengths. This highly-schematic example illustrates sequence reads from paired-ends (shown in pale green and orange colored bars); that is, sequences obtained from the ends of individual DNA fragments (see **Figure 6.21** for more detail). They correspond to nucleotide sequences in the reference sequence shown at top, except for indicated base changes. (Note that the bars and connecting gray lines represent individual DNA fragments, with the gray lines representing the unsequenced central part of the fragments. DNA fragments at top left and bottom right are artificially truncated, because only part of their sequence corresponds to the reference sequence.) In this example, the bases at top are covered by at most 16 sequence reads. Vertical boxes indicate the following: (1) variants that are most likely sequencing errors; (2) regions of poor coverage—is the G in the first of these regions a sequencing error, or is the individual heterozygous with a T and a G? (3) Convincing evidence of heterozygosity (C/G). A real example would have much greater sequence coverage (read depth).

straightforward. Figure 6.20 shows a highly-schematic example to illustrate issues that may arise.

The sequencing reads represented by pale green and orange bars in Figure 6.20 are derived by sequencing both ends of each DNA fragment, so-called **paired-ends** (Figure 6.21). Paired-end sequencing has significant advantages over single-end sequencing: it provides double the amount of sequence data, and it aids sequence assembly, both generally (because there are two neighboring sequences) and also by permitting fragments to be anchored against the reference sequence where one end originates from a highly-repetitive DNA sequence (Figure 6.21B).

Paired-end sequences would normally be expected to be separated by the same distance as in the original chromosomal DNA molecule they arose from in the cells donated to provide the DNA sample. Comparing paired-end sequences from different DNA samples against the reference sequence can therefore be used to identify structural variants, including deletions, insertions, and inversions, extending over comparatively short distances (Figure 6.21C).

To identify structural variation extending over larger distances, mate-pair sequences are required. Unlike paired-end sequences, mate-pair sequences represent sequences that were originally well-separated on a chromosomal DNA molecule (Figure 6.21D) but were artificially brought together into a short DNA sequencing template after circularizing the intervening DNA (Figure 6.21E). Effectively, they can be thought of as being equivalent to paired-end sequences that have very long inserts, and so are also helpful in the de-novo assembly required for sequencing a genome for the first time (Figure 6.21D).

Driven by intense competition between the various companies involved, technical progress in high-throughput DNA sequencing is extremely rapid; readers are recommended to consult the most recent reviews available for updates.

Massively-parallel sequencing of amplified DNA: an overview

The first massively-parallel DNA sequencing systems to be developed used PCR-amplified DNA templates; the signals generated need to be amplified in order to enable their detection during the sequencing process. The methods used for the sequencing reactions give generally low signal-to-noise ratios (the chemistry of the sequencing reactions becomes

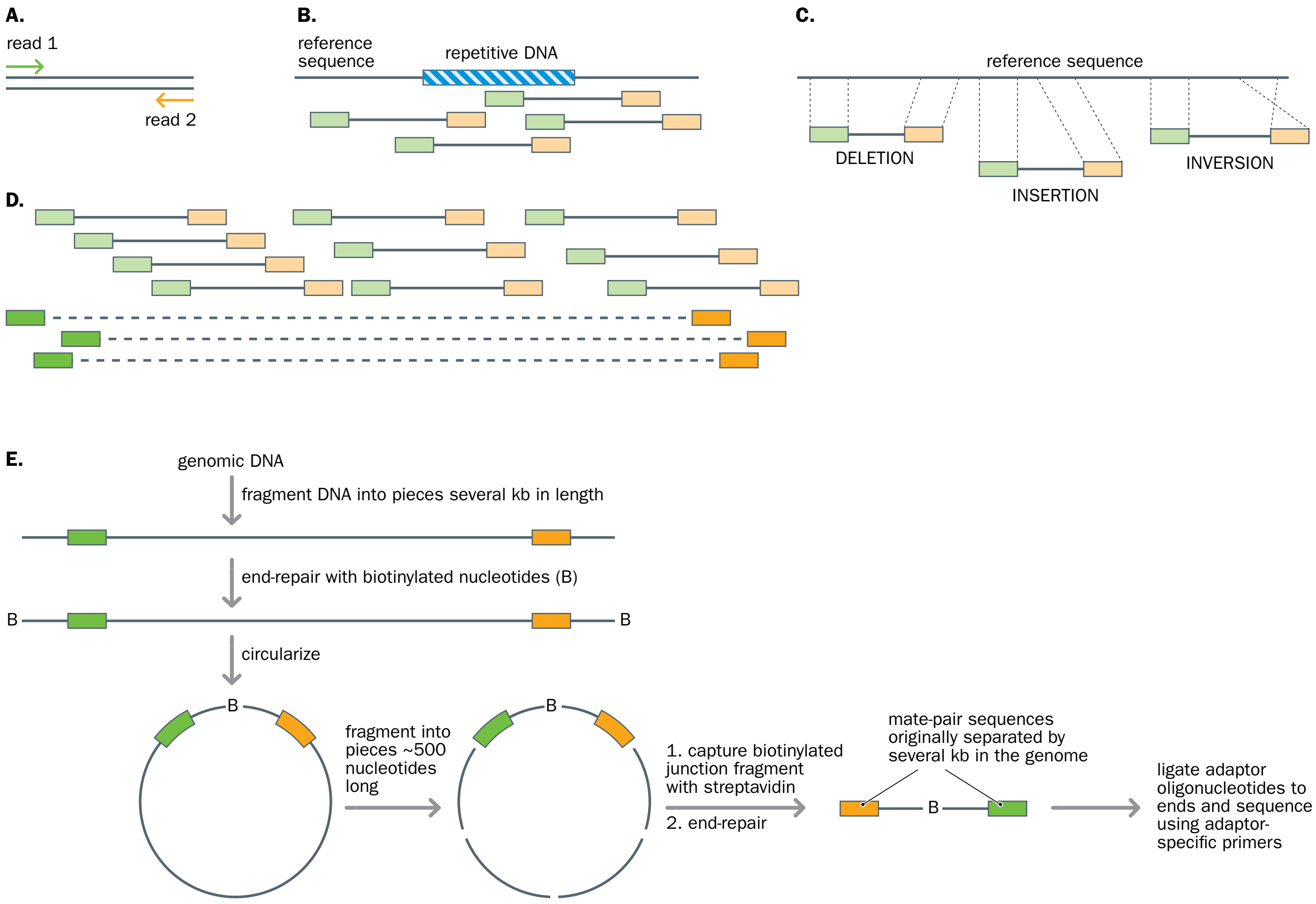

A.
read 1
read 2
B.
reference sequence
repetitive DNA
C.
reference sequence
DELETION
INSERTION
INVERSION
D.
E.
genomic DNA
fragment DNA into pieces several kb in length
end-repair with biotinylated nucleotides (B)
B
B
circularize
B
fragment into pieces ~500 nucleotides long
B
1. capture biotinylated junction fragment with streptavidin
2. end-repair
mate-pair sequences originally separated by several kb in the genome
B
ligate adaptor oligonucleotides to ends and sequence using adaptor-specific primers

Figure 6.21 Paired-end and mate-pair sequences. (**A**) Both paired-end and mate-pair sequences are obtained by sequencing the ends of individual DNA fragments. This requires ligating adaptor oligonucleotide sequences onto the ends, and sequencing involves using adaptor-specific primers to extend DNA synthesis inward to obtain sequences from each end of the individual DNA fragments (**Figure 6.22** gives the detail). (**B**) In re-sequencing, paired-end sequences have the advantage that they may anchor a DNA fragment to a position in the reference genome even if one end maps within a highly-repetitive DNA sequence in the reference genome. (**C**) Structural variants such as deletions and insertions can be identified because the size of the DNA fragment is smaller or greater than expected, respectively, when referred against the size predicted from the reference genome. Inversions can be identified by altered orientation of an end sequence. (**D**) A combination of paired-end sequences (pale green and orange bars) and mate-pair sequences (dark green and dark orange bars joined by dashed gray lines to indicate the original distance separating the two sequences in the genome) can be used in *de novo* genome assembly. (**E**) Mate-pair sequences (dark green and dark orange bars) are originally located several kilobases apart in the genome but can be artificially brought together by circularization of long fragments, fragmentation of the circular DNAs, capture of junction fragments (where the original ends of linear sequences were brought together), and then sequencing of the ends of junction fragments.

progressively more inefficient during the reactions, which is why the read lengths are generally low). In the ABI SOLiD sequencing platforms, for example, the relatively low signal-to-noise ratio means read lengths of only 35 to 75 bp, but in the Roche/454 GS FLX+ sequencing platform a significantly higher signal-to-noise ratio allows read lengths approaching that of Sanger sequencing (see **Figure 6.19**).

These methods offer a trade-off: generally high, and sometimes very high, sequencing throughput, but a quite high rate of sequencing errors in individual sequence reads. By sequencing very many PCR-amplified fragments, however, many sequence reads can be obtained for a given region of DNA. This "deep sequencing" allows sequencing errors to be identified in individual reads (see **Figure 6.20**) so that a consensus sequence can be established with acceptably low error rates.

DNA library construction

The preliminary objective is to fragment the DNA to give suitably sized DNA fragments, and then attach to them common sequences that can be recognized by universal primers for amplification and DNA sequencing. Fragmentation of the DNA is usually achieved using a mechanical method. Sonication uses sound energy to agitate the DNA molecules in solution; hydrodynamic shearing forces break the DNA at random locations.

Alternatively, nebulization randomly shears DNA by using pressurized gas to force the DNA through a small hole in a nebulizer unit. Conditions are chosen to prepare small DNA fragments within a desired size range.

The resulting DNA fragments have different lengths of overhanging 5′ and/or 3′ ends but can be converted to blunt-ended DNA using a DNA polymerase and 3′ exonuclease (**Figure 6.22A**). Blunt-ended DNA fragments are phosphorylated at the 5′ end, and then are usually A-tailed at the 3′ end to facilitate ligation to a double-stranded **adaptor oligonucleotide** designed to have an overhanging T.

The adaptor oligonucleotide provides defined target sequences to allow binding of complementary primers to enable clonal amplification of individual DNA fragments, and can be used to allow sequencing of both ends of each DNA fragment. Forked adaptors are typically used. These Y-shaped adaptors are designed to have complementary sequences at one end (which form a double-stranded stem) and two unrelated and non-complementary sequences at the other end (which form single-stranded arms). The amplification process allows individual DNA fragments to be flanked by two different sequences to which sequencing primers can subsequently bind for forward and reverse strands (**Figure 6.22B** and **C**).

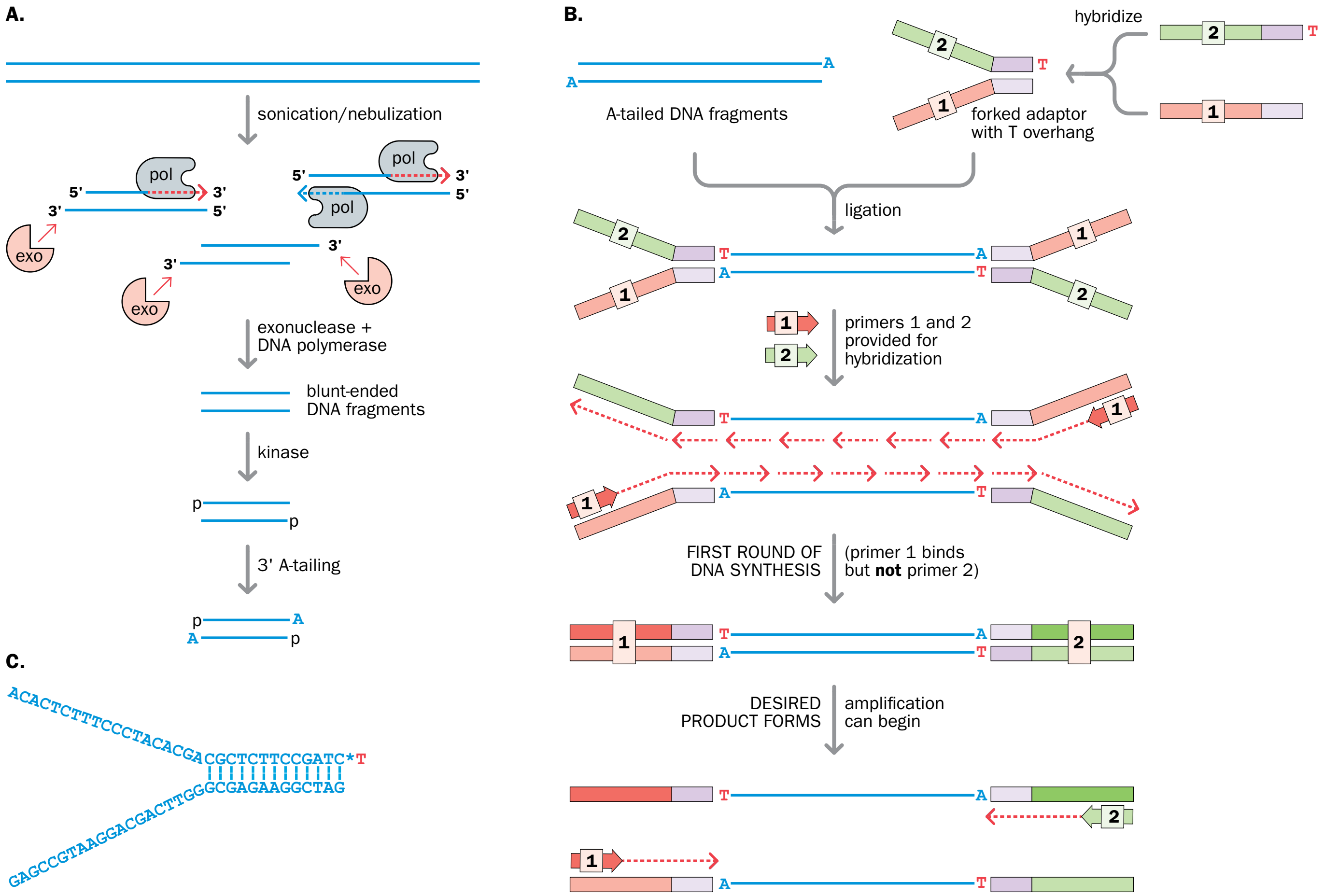


Figure 6.22 Preparation and amplification of DNA libraries for high-throughput DNA sequencing. (**A**) DNA fragment preparation. The DNA is randomly fragmented. The resulting fragments are made blunt-ended by digesting 3′ overhangs with a 3′ exonuclease and by “filling in” at overhanging 5′ ends (extending the complementary strand using DNA polymerase). Blunt-ended fragments are 5′ phosphorylated using polynucleotide kinase and are often then “A-tailed” using Taq polymerase to add a nontemplated adenine to the 3′ end. (**B**) Amplification using a forked adaptor oligonucleotide. A forked adaptor oligonucleotide is designed to be partially double-stranded (open boxes) but has distinctive sequences (labeled 1 and 2 here) that will provide target sequences for forward and backward primers in the sequencing reaction. To facilitate ligation, one of the two oligonucleotides is often designed to have an additional T (resulting in a 3′ T overhang). Thereafter, primers for adaptor sequences 1 and 2 are used to amplify the DNA fragments. The primers, however, are asymmetric. One primer (primer 1 here) has a sequence complementary to one of the unique adaptor sequences, and base-pairs immediately with its target. However, primer 2 has the same sequence (*not* the complementary sequence), as part of the other adaptor sequence (and so cannot bind initially). After one round of DNA synthesis (by primer 1 only), a desired product is formed with the two different, flanking double-stranded adaptor sequences, and can now be amplified after both primers 1 and 2 bind. (**C**) An example of an Illumina paired-end forked adaptor (the asterisk signifies a phosphorothioate bond that is resistant to 3′ exonuclease).

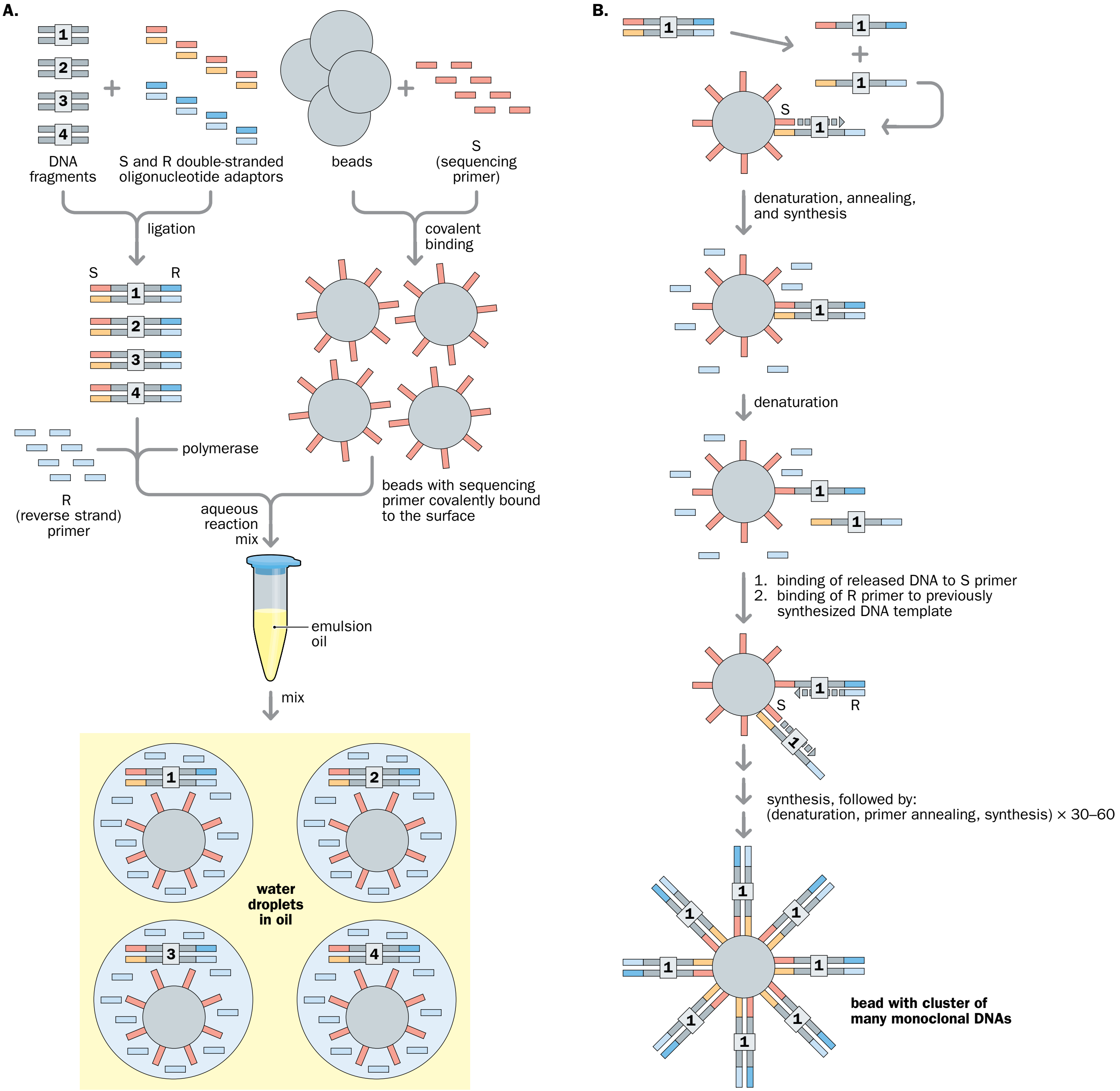

A.
1
2
3
4
+
DNA fragments
S and R double-stranded oligonucleotide adaptors
beads
+
S (sequencing primer)
ligation
covalent binding
S
R
polymerase
R (reverse strand) primer
beads with sequencing primer covalently bound to the surface
aqueous reaction mix
emulsion oil
mix
water droplets in oil
B.
denaturation, annealing, and synthesis
denaturation
1. binding of released DNA to S primer
2. binding of R primer to previously synthesized DNA template
S
R
synthesis, followed by:
(denaturation, primer annealing, synthesis) × 30–60
bead with cluster of many monoclonal DNAs

Figure 6.23 Principle of emulsion PCR. (**A**) Generating tiny water-droplet microreactors. DNA fragments (1, 2, 3, 4, and so on) prepared from the DNA source are ligated to double-stranded oligonucleotide adaptors so that each fragment has at one end an adaptor containing the same sequence as the sequencing primer (labeled S here) and at the other end a different adaptor (labeled R here). Multiple copies of the sequencing primer are tethered by a short linker sequence at the 5′ end to the surface of tiny beads (about 28 μm in diameter); each bead has very many copies of the S primer, but for the sake of clarity, only eight S primer copies are shown here. The adaptor-linked DNA fragments are mixed with the primer-linked beads plus reverse-strand primer and a heat-stable DNA polymerase, and the resulting aqueous reaction mix is inserted into the emulsion oil, and thoroughly mixed. Tiny water droplets are formed. By ensuring a low concentration of starting DNA, most bead-containing water droplets will have zero or one DNA template molecule. Productive droplets contain a single DNA fragment with all the necessary components for DNA amplification: a bead with bonded sequencing primers, free reverse primers, and DNA polymerase molecules (not shown), constituting a "microreactor." (**B**) DNA amplification in microreactors. After the DNA fragment (number 1 in this example) is denatured and cooled, the reverse strand can hybridize to a complementary sequencing primer (S) protruding from the bead, allowing DNA synthesis of a forward strand from the primer bound to the bead. Further denaturation and primer annealing allows a reverse DNA strand to bind to a sequencing primer attached to the bead, and also allows a reverse-strand primer to bind to the forward strand previously synthesized. After further cycles, clonal amplification results so that each bead will have a cluster of many copies of a single type of DNA fragment covalently bound to it.

Amplification of separated DNA fragments

After the DNA fragments have been flanked by adaptor sequences, individual DNA fragments must be physically separated in some way. They are then amplified independently to give separated clusters of monoclonal DNA that will provide the DNA templates for sequencing. (It is important that the amplified DNA within each cluster is composed of one type of DNA; if not, the recorded sequencing signals would be unreadable.) Two amplification methods have been particularly popular, as listed below.

- ***Emulsion PCR.*** This amplification method was pioneered in the Roche/454 sequencing approach, and variant methods have been used for the ABI SOLiD and Life Sciences/Ion Torrent DNA sequencing platforms. The object is to separate individual DNA fragments in individual tiny water droplets so that each fragment in the starting DNA can be amplified separately. This is done by mixing an aqueous phase containing a library of DNA fragments (plus primers and reagents for PCR amplification) with oil to create an emulsion: tiny water droplets become suspended in the oil and can entrap individual DNA fragments (**Figure 6.23**).
- ***Bridge amplification.*** Originally developed by the Solexa company, this amplification method is used by the Illumina sequencing platforms. It involves a type of two-dimensional PCR amplification of well-separated DNA fragments that are covalently bound to the surface of a glass slide within a flow cell (**Figure 6.24**). A related but different method of amplification known as Wildfire has more recently been adopted for use with ABI SOLiD sequencing, as described below.

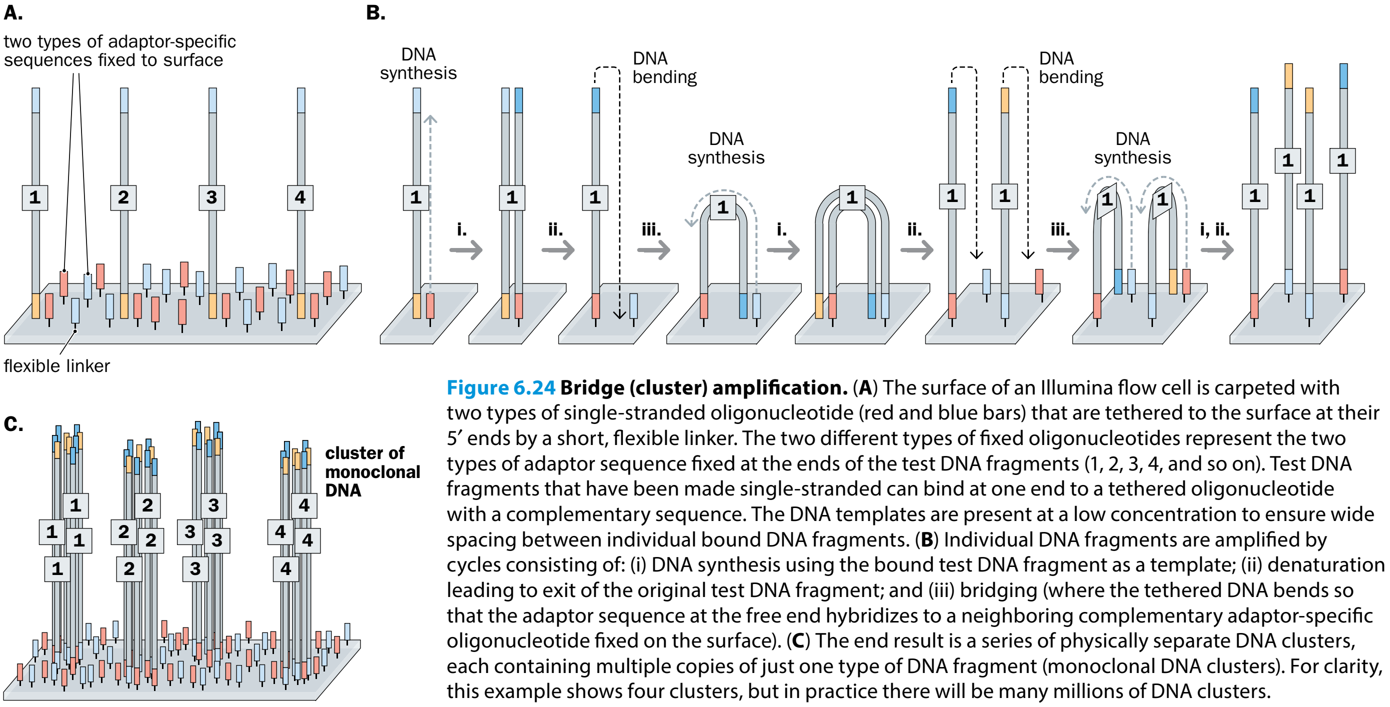


Figure 6.24 Bridge (cluster) amplification. (**A**) The surface of an Illumina flow cell is carpeted with two types of single-stranded oligonucleotide (red and blue bars) that are tethered to the surface at their 5′ ends by a short, flexible linker. The two different types of fixed oligonucleotides represent the two types of adaptor sequence fixed at the ends of the test DNA fragments (1, 2, 3, 4, and so on). Test DNA fragments that have been made single-stranded can bind at one end to a tethered oligonucleotide with a complementary sequence. The DNA templates are present at a low concentration to ensure wide spacing between individual bound DNA fragments. (**B**) Individual DNA fragments are amplified by cycles consisting of: (i) DNA synthesis using the bound test DNA fragment as a template; (ii) denaturation leading to exit of the original test DNA fragment; and (iii) bridging (where the tethered DNA bends so that the adaptor sequence at the free end hybridizes to a neighboring complementary adaptor-specific oligonucleotide fixed on the surface). (**C**) The end result is a series of physically separate DNA clusters, each containing multiple copies of just one type of DNA fragment (monoclonal DNA clusters). For clarity, this example shows four clusters, but in practice there will be many millions of DNA clusters.

DNA sequencing reactions

In the bridge amplification method, the amplified DNA, bound to a glass slide within a flow cell, is already in place for the sequencing reaction (the *flow cell* is a multichannel, sealed glass device, through which reagents required for DNA sequencing can be passed). In emulsion PCR, however, the amplified DNA is bound to tiny beads within water droplets, and the DNA–bead complexes must first be recovered, and then layered into individual wells in the receptacle that will be used for sequencing, usually a picotiter plate (PicoTiterPlate™).

The sequencing reactions often involve sequencing-by-synthesis. Like Sanger sequencing, this means using a single-stranded DNA template and carrying out DNA synthesis with a DNA polymerase and dNTPs. Unlike Sanger sequencing, the incorporation of nucleotides into the growing DNA chain is followed during the sequencing reaction, either directly or indirectly, and results in a base-specific light signal (recorded using a CCD camera) or an electrical signal. The sequencing-by-synthesis methods either use reversible chain-terminating nucleotides (Illumina) or single-nucleotide addition (Roche/454, Ion Torrent). ABI SOLiD sequencing platforms, however, use an alternative method: sequencing-by-ligation. That is, a series of base-specific ligation reactions to fluorescently labeled oligonucleotides is performed and the fluorescent label acquired in positive ligations is recorded, as described below.

Massively-parallel sequencing of amplified DNA: commonly used sequencing platforms

The sequence outputs and read lengths of the most commonly used sequencing platforms are shown in **Figure 6.19**. The run times vary from several hours to several days. At the time of writing, the Illumina platform was by some distance the market leader.

Roche/454 pyrosequencing

This, the first of the massively-parallel technologies, was brought to market in 2005. It is based on pyrosequencing (**Box 6.4**). The procedure is summarized in **Figure 6.25**, and described in detail by Margulies *et al*. (2005) (PMID 16056220; see Further Reading).

The strengths of the 454 technology are its relatively long reads and speed. It can produce 400–600 Mb of sequence in a 10-hour run. Disadvantages include its relatively low throughput, high reagent costs, and an inability to size homopolymer runs accurately (a run of eight adenines, for example, cannot be reliably distinguished from a run of nine adenines by the amount of light emitted). In the light of newer technologies, Roche discontinued support for 454 sequencing in 2016.

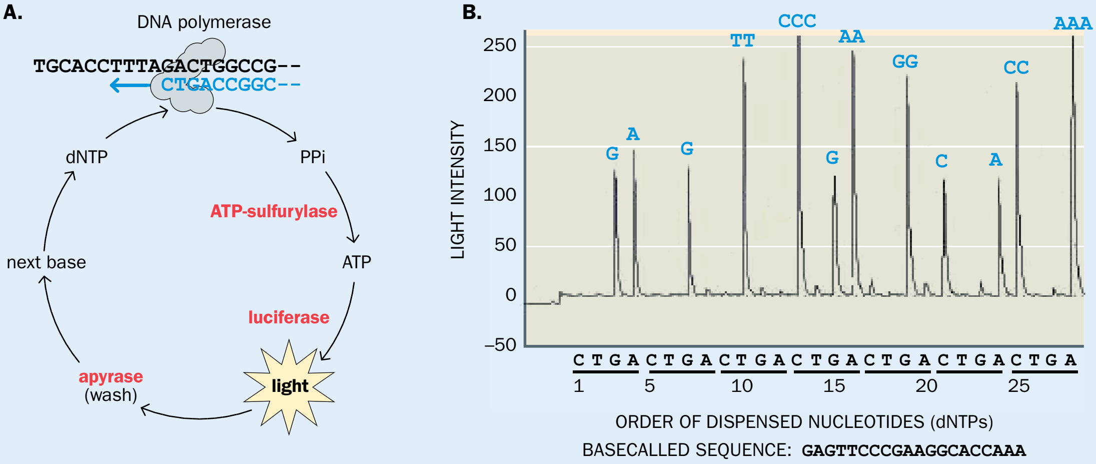
The ABI SOLiD technique

Applied Biosystems introduced the SOLiD (Sequencing by Oligonucleotide Ligation Detection) system in 2007. Its special feature is the use of an ingenious DNA ligation-based

BOX 6.4 PYROSEQUENCING

This alternative to Sanger sequencing was developed by Ronaghi and Nyrén at the Royal Institute of Technology in Stockholm in the 1990s (Ronaghi *et al*. [1996]; PMID 8923969; see Further Reading). Like Sanger sequencing, it works by synthesis, but detects the pyrophosphate produced when a deoxynucleoside triphosphate is incorporated into the growing chain (**Figure 1**). In the pyrosequencing procedure, each dNTP in turn is presented to the DNA polymerase reaction. Incorporation of the correct nucleotide produces a flash of light.

Between each addition, the remaining dNTP is broken down by an apyrase enzyme. A weakness of pyrosequencing is its inability to size homopolymer runs accurately, as, for example, a run of eight adenines cannot be reliably distinguished from a run of nine adenines by the amount of light emitted.



Box 6.4 Figure 1 The principle of pyrosequencing. (**A**) Incorporation of a dNTP into the growing chain releases a molecule of pyrophosphate (PPi). In the presence of adenosine 5′ phosphosulfate, the enzyme ATP-sulfurylase converts pyrophosphate into ATP, which luciferase uses to generate a flash of light. (**B**) A pyrosequencer trace. Individual dNTPs are dispensed as possible substrates for the pyrosequencing reaction in cycles with the order: C then T then G then A, and the intensity of any light emitted in response to each nucleotide dispensed is recorded to give the graph shown here. The interpreted sequence here would be: GAGTTCCCGAAGGCACCAAA.

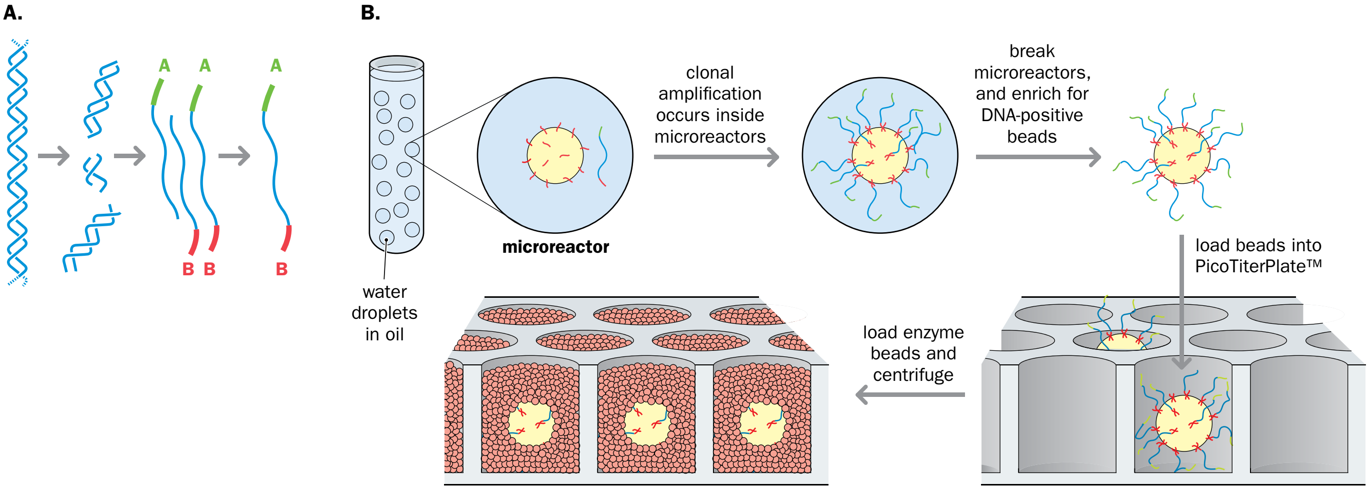


Figure 6.25 Roche/454 reiterative pyrosequencing. (**A**) The input DNA is nebulized to 300–800 bp fragments, made blunt-ended, and then ligated to double-stranded adaptor oligonucleotides, A and B, each of which contains primer sequences for PCR and then pyrosequencing. One strand of the B adaptor carries a biotinylated nucleotide, allowing tagged DNA fragments to be isolated by binding to streptavidin-labeled magnetic beads. Double-stranded fragments carrying A and B sequences are captured and then denatured. The single-stranded molecules released into the supernatant comprise the sequencing library. (**B**) A separate set of agarose beads with a covalently bound sequence complementary to the B adaptor is used to capture the single-stranded end-tagged fragments of the test DNA. The conditions are chosen so that most beads capture either zero or just one DNA fragment. PCR primers, polymerase, and the other components of a PCR mix are added to the beads, and the whole is broken into minute droplets ("microreactors") in an oil emulsion, so that on average a droplet includes only a single bead. The whole emulsion is then put through 40 cycles of PCR. At the conclusion of the PCR, the emulsion is broken down and the DNA anchored to the beads is made single-stranded by denaturation and washing. The beads, now each carrying several million identical, single-stranded copies of one particular fragment of the test DNA, are loaded into individual wells of a fiber-optic slide (a "picotiter plate" [PicoTiterPlate™]) that contains 1.6 million wells. Smaller beads carrying immobilized enzymes required for pyrosequencing are layered on top of each well. A CCD camera images the pattern of light emission across the wells as each successive dNTP is washed over the plate.

chemistry that checks each base independently twice for mismatches (two-base encoding). This results in a very low error rate for either miscalled or mismatched bases.

The test DNA is fragmented, ligated to adaptor oligonucleotides, bound to beads, and amplified by emulsion PCR. The beads, each carrying millions of clonally amplified single-stranded copies of one particular fragment, are immobilized on the surface of a glass slide and exposed to a cocktail of fluorescently labeled 8-mer oligonucleotides in the presence of DNA ligase. Ligation depends on correct base pairing of the two bases at the 5′ end of the oligonucleotide. After washing away unligated probes, a camera records the color of the ligated probe. The 3′ end of the probe, together with the fluorescent label, is then cleaved off and a further round of ligation started. Successive cycles of ligation and cleavage follow, until the maximum read length is reached. The resulting double-stranded DNA is then denatured, the ligation product washed away, and the whole process started over again, but using a sequencing primer displaced one nucleotide from the initial primer. This is done five times in total (**Figure 6.26**).

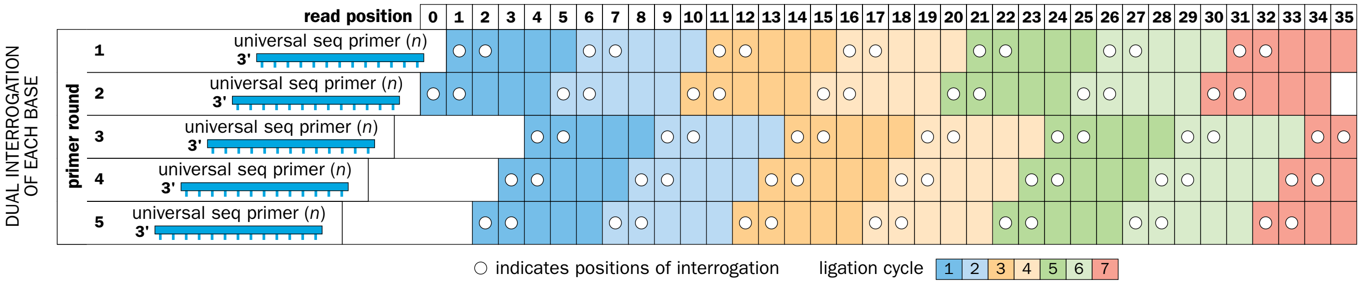


Figure 6.26 Principle of two-base encoding used in ABI SOLiD DNA sequencing. The top row shows how one primer is extended in successive cycles of ligation and cleavage (seven cycles are shown here, being color-coded as shown in the key at bottom). Ligation depends on correct base pairing (marked by dots) of nucleotides 1 and 2 of the probe with the template. Nucleotides 3–5 are degenerate. After ligation, the probe is cleaved after nucleotide 5, removing the fluorescent label and the degenerate downstream nucleotides. A second cycle of ligation and cleavage then takes place; this time, ligation depends on exact pairing of nucleotides 1 and 2 of the probe with nucleotides 6 and 7 of the template (marked by dots). After maybe a dozen such cycles, the whole newly synthesized strand is dissociated from the template and washed away, and the cycles of ligation and cleavage are started again using a different sequencing primer that anneals to a slightly different position on the template, as shown in the second row. Overall, five different sequencing primers are used. Each position in the template is interrogated twice, allowing very low mismatch or miscalling error rates.

In 2013, Applied Biosystems introduced a new and much simpler method ("Wildfire") for preparing the sequencing templates. This did away with beads and emulsion PCR, relying instead on a system of isothermal amplification of DNA fragments anchored to a glass slide, to produce compact clusters of clonal copies. Although the method is different, the effect is similar to the bridge PCR used by Illumina for its next-generation system, as described below. Full details of the Wildfire technology can be found in the paper by Ma *et al*. (2013) (PMID 23940326; see Further Reading).

The particular advantage of the SOLiD system is its high accuracy, including with homopolymer runs. Its main disadvantage is the short read length, up to 75 bp, which makes assembly more difficult. Each full sequencer run takes 7 days and generates around 4 terabytes of raw image data, which are analyzed to produce up to 200 Gb of finished sequence.

Illumina/Solexa sequencing

The market-leading Illumina technology was originally developed by the Solexa company (Bentley *et al*. [2008], PMID 18987734; see Further Reading). The bridge PCR technique (**Figure 6.24**) generates clonal clusters of amplified fragments bound to a glass slide, and these are sequenced by synthesis in a way similar to Sanger dideoxy sequencing, but using reversible chain terminators.

Illumina/Solexa sequencing uses dye-labeled chain-terminator dNTPs (**Figure 6.27A**), as in Sanger sequencing, but in this case no normal (unmodified) dNTPs are present. The reaction therefore stops after incorporation of a single nucleotide. This is imaged to record the color, and then both the blocking group and dye are removed, allowing a second nucleotide to be added. Thus sequencing goes in cycles of incorporation, imaging, and cleavage (**Figure 6.27B** and **C**).

The chemical cleavage step leaves an entirely normal 3′ hydroxyl group for addition of the next dNTP, but part of the dye linker remains attached to the base. A modified DNA polymerase is used, both to improve incorporation of the heavily modified dNTPs and to tolerate the "scar" left in the growing newly synthesized strand after cleavage of the dye. Accurate data depend on the incorporation and cleavage reactions being complete across all the millions of clusters on the flow cell, and this limits reads to around 100 nucleotides.

Illumina produces a range of machines to suit either large-scale genome projects or small clinical sequencing services, together with kits to simplify all aspects of library preparation and data acquisition.

Ion Torrent systems

The first Ion Torrent PGM (Personal Genome Machine) was released in 2010. The workflow of Ion Torrent systems has much in common with that of the 454 pyrosequencing system: both use fragmentation, adaptor ligation, capture by beads, emulsion PCR, and deposition of beads, each carrying millions of single-stranded copies of one particular DNA fragment, into wells of a plate. As in 454 technology, individual dNTPs are washed across the plate and sequencing proceeds by synthesis. However, Ion Torrent systems use a radically different, and much simpler, way to follow the synthesis.

Incorporation of a correctly paired dNTP into the growing strand releases not only the pyrophosphate detected by pyrosequencing but also a hydrogen ion, H^+. Ion Torrent systems detect the hydrogen ion directly as an electric signal. The Ion Torrent chip is a CMOS silicon chip, like those used in digital cameras. Below each well is an ion-sensitive field effect transistor that generates an electronic signal in response to release of a hydrogen ion. Unlike any of the systems detailed above, Ion Torrent systems do not require fluorescence and camera scanning, which allows higher speed, lower cost, and a smaller-sized machine.

Reads of 200 or 400 bp are delivered in 2.5 or 4 hours, respectively, and the throughput depends simply on the number of wells on a plate—for example, an Ion Torrent 540 chip allows 60–80 million reads of 200 bp, giving 10–15 Gb of sequence. Ion Torrent machines are thus marketed as relatively cheap and simple benchtop machines suitable for applications in microbiology or clinical resequencing. They suffer from the same problems with homopolymer runs as the 454 system.

Massively-parallel sequencing of unamplified DNA

DNA sequencing technologies that use single unamplified DNA templates—sometimes called single-molecule sequencing or third-generation sequencing—avoid the biases introduced by PCR, and have the potential for producing very long sequences at low cost. However, sequence accuracy can be an issue.

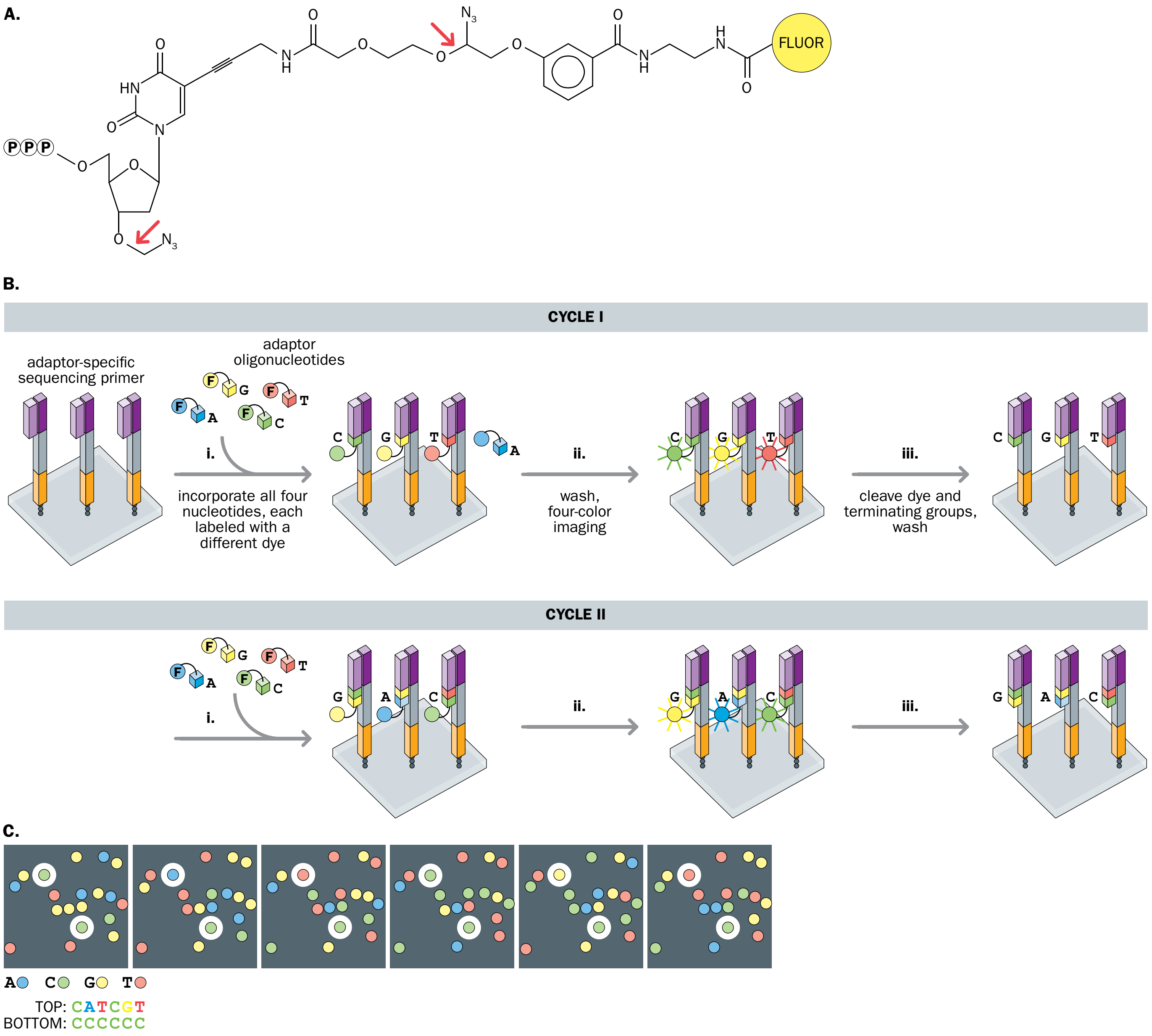


Figure 6.27 The Illumina/Solexa sequencing-by-synthesis method involves repeated cycles of incorporation of reversible terminator deoxynucleotides, imaging, and cleavage. (**A**) An example of a reversible terminator dNTP where the normal 3′ hydroxyl group is replaced by a 3′ O-azidomethyl group (imagine a 3′ O-methyl group but with the three hydrogen atoms replaced by nitrogen atoms). In addition, the base (thymine in this example) has a side chain containing an azidomethyl ($-N_3$) group and a fluor group. Once a reversible terminator has been incorporated into the growing DNA chain, the fluorescent signal from the fluor can be imaged, but then a single chemical treatment cleaves both azidomethyl groups (red arrows), leaving free hydroxyl groups. The result is that the fluor group is released, and the 3′ hydroxyl group is restored so that a new nucleotide can be incorporated. (**B**) The cycle of nucleotide incorporation, imaging, and cleavage. This shows synthesis of three out of the many millions of growing DNA strands synthesized using single-stranded DNA templates (gray bars) using an adaptor-specific primer. The process continues through many cycles, of which only the first two are shown here. (**C**) Example of four-color imaging. The 19 colored circles in each panel represent 19 of the millions of growing DNA strands. The color given in each circle represents the last nucleotide to be inserted in the growing DNA chain and the six individual panels show imaging after six cycles of nucleotide incorporation. For clarity, two of the growing DNA strands are shown enclosed in white rings so that successive color images (representing nucleotide incorporation) can be readily followed from one panel to the next. For the top one of the two encircled growing DNA strands, the successive color changes (moving between the images from left to right) indicate the sequence CATCGT; the bottom one remains green for six cycles, indicating the sequence CCCCCC. (B and C, adapted from Metzker ML [2010] *Nat Rev Genet* **11**:31–46; PMID 19997069. With permission from Springer Nature. Copyright © 2010.)

Pacific Biosciences systems

The PacBio RSII system, released in 2010, was heralded as the first DNA sequencing method to sequence single unamplified molecules in real time. The sequencing templates are double-stranded DNA molecules that have single-stranded hairpin oligonucleotides ligated to each end (**Figure 6.28A**). A single sequencing primer is annealed to one of the hairpins.

Sequencing takes place in a SMRT Cell, a fabrication containing 150,000 tiny wells (capacity per well = 10^{-21} liters) called zero-mode waveguides. A single DNA polymerase molecule is anchored to the bottom of each well. Dye-labeled dNTPs diffuse in and out of the wells (**Figure 6.28B**). The dye labels are on the terminal phosphates of the dNTPs, so that on incorporation the dye is lost and totally natural DNA is synthesized. A high-processivity, strand-displacement DNA polymerase is used, such as Φ29 or Bst polymerase. The synthesis point moves round and round the template, making long concatemers of the sequence and using both the sense and antisense strands as template in its journey round.

When the polymerase binds and then incorporates a dNTP, it increases the time the attached label spends at the bottom of the well, compared to the time spent by randomly diffusing dNTP molecules, as recorded by the laser imaging system (**Figure 6.28C**). The duration represents the real-time dynamics of the polymerase, and this may be different when template bases carry epigenetic modifications such as methylation. Thus the system has the unique ability to identify patterns of modification directly from the raw data. For the principles of the system and discussion of the templates, see Eid *et al*. (2009) (PMID 19023044) and Travers *et al*. (2010) (PMID 20571086) in Further Reading.

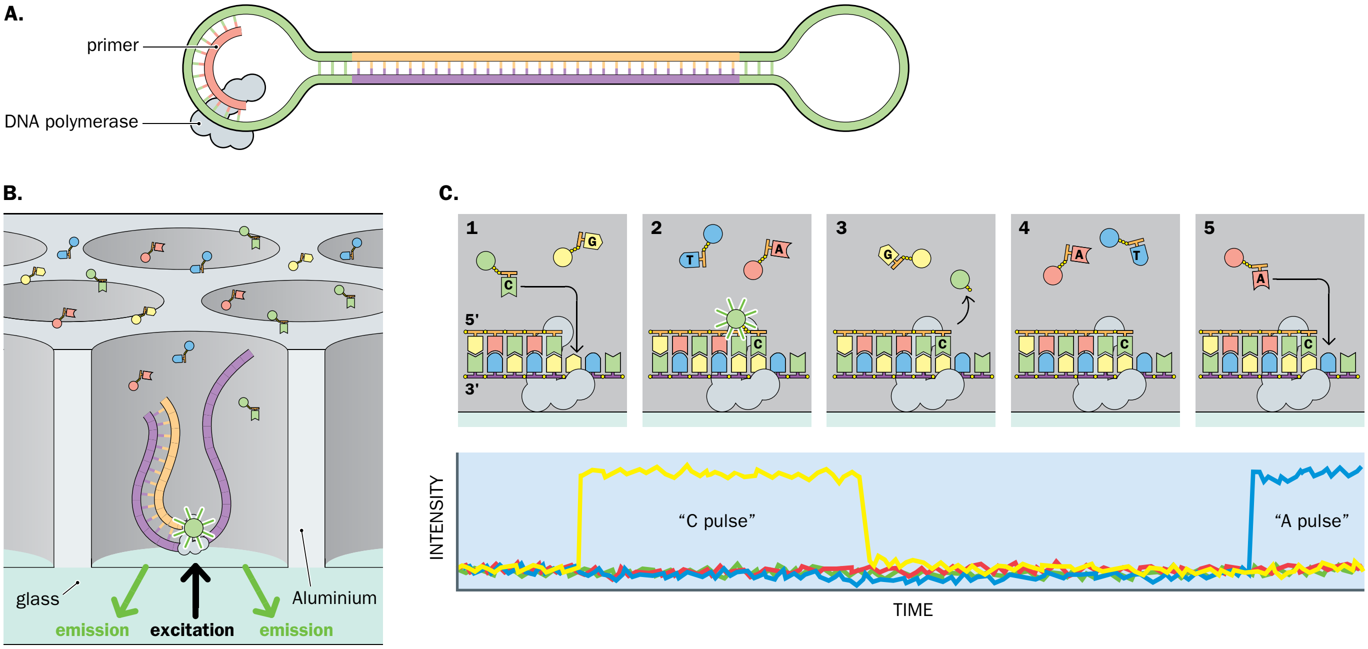


Figure 6.28 Single-molecule real-time sequencing using the PacBio system. (**A**) The template for PacBio sequencing is a double-stranded DNA with single-strand hairpins (green) ligated to either end. A sequencing primer (red) anneals to one of the hairpins. The strand-displacing polymerase (gray) moves the template continuously round, producing concatenated copies of the whole sequence, as detailed in Travers *et al*. (2010). (**B**) The polymerase is anchored at the bottom of a well; the four phospho-labeled dNTPs diffuse freely. (**C**) When a nucleotide is incorporated into the growing chain, its fluorescent label remains at the bottom of the well much longer than the freely diffusing dNTPs. (A, adapted from Travers KJ *et al*. [2010] *Nucl Acid Res* **38**:e159; PMID 20571086; B and C, adapted from Eid J *et al*. [2009] *Science* **323**:133–138; PMID 19023044. Reprinted with permission from the AAAS.)

The PacBio RSII allows extremely long reads (20 kb or more), and both sample preparation times and run times are short, allowing the whole procedure to be completed in a single day. The error rate per nucleotide, at around 11%, is much higher than in competing systems, but the errors are random and can be compensated by allowing the polymerase to run through the same circular template many times. The throughput is lower and the cost per base higher than for many competing systems. However, the long reads make it ideal for de-novo sequencing of small bacterial and viral genomes, and for sequencing low-complexity regions or structural variants in human and other genomes.

Oxford Nanopore Technologies system

Oxford Nanopore are developing a competing third-generation system. In the MinION device, released to early-access users in 2014, a flow cell contains maybe 500 wells, each of which is spanned by a synthetic, electrically resistant membrane in which a single nanopore is anchored. The nanopores are made of modified α-hemolysin protein. Single-stranded DNA feeds through the 10 μm long × 1 μm wide pore. As the different-sized nucleotides pass through, they block the ionic current flowing through the pore to different extents, potentially allowing each nucleotide to be recognized. In practice, the blocking effect depends on at least five contiguous nucleotides and must be deconvoluted to identify individual nucleotides. To get signals for analysis, the passage of the DNA through the pore must be slowed down by several orders of magnitude, and this is achieved by coupling it to a relatively slow-moving processive enzyme. The test DNA is double-stranded; the leader end has a single-strand extension coupled to the motor enzyme, while the far end can be ligated to a hairpin oligonucleotide carrying a second motor enzyme, thus allowing both strands to be sequenced (**Figure 6.29**).

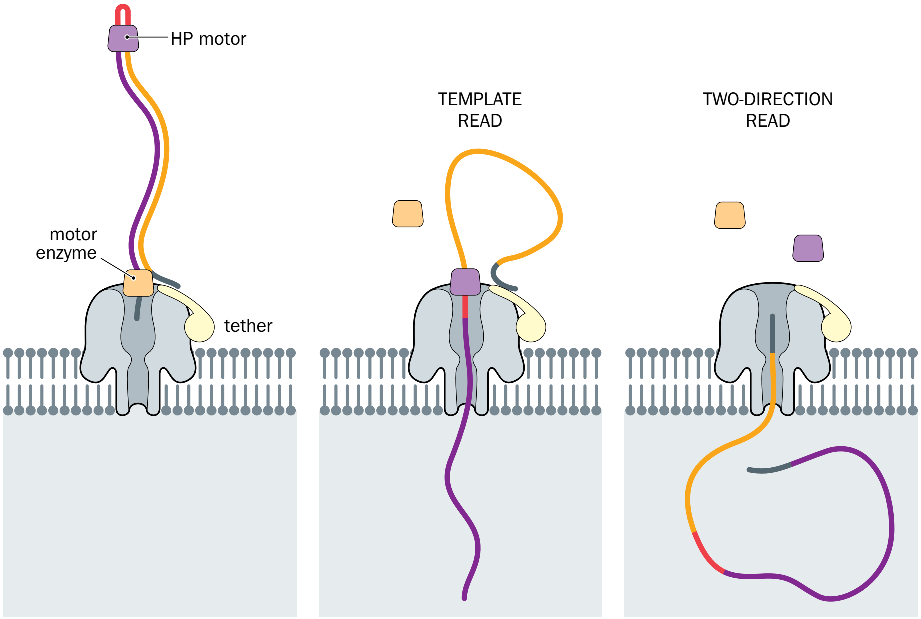


Figure 6.29 Feeding test DNA through a nanopore. Oxford Nanopore's sequencing strategy requires DNA templates to be ligated with two adaptors. The first adaptor is bound with a motor enzyme as well as a tether, whereas the second adaptor is a hairpin oligonucleotide that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the pore are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads). (From Reuter JA *et al.* [2015] *Mol Cell* **58**:586–597; PMID 26000844. With permission from Elsevier.)

Nanopore sequencing offers great promise. Potentially the read length is limited only by the length of the test DNA. Sample preparation is simple, the process is quick, and the device is small and simple enough to be portable. The big problem is the error rate, with insertion, deletion, and substitution rates of 4.9%, 7.8%, and 5.1%, respectively, reported by Jain *et al.* (2015) (PMID 25686389; see Further Reading). Reuter *et al.* (2015) (PMID 26000844; see Further Reading) also report a high run-failure rate. However, the device has already demonstrated its use in sequencing a previously unresolved, highly-repetitive region of human chromosome X (see Jain *et al.* 2015), and with increases in reliability and accuracy, it has many possible applications.

Other technologies

Many alternative massively-parallel sequencing technologies are available or under development by a variety of companies. The following list is intended to give a flavor of the current technical ferment, without in any way claiming to be comprehensive.

- Complete Genomics (http://www.completegenomics.com)
- Genapsys (http://genapsys.com)
- Genia (http://www.geniachip.com)
- Gnubio (http://gnubio.com)
- Lasergen (http://lasergen.com)
- Nabsys (http://nabsys.com)
- Stratos (http://stratosgenomics.com)
- ZSG (http://www.zsgenetics.com)

SUMMARY

- In complex genomes, a gene or exon or other sequence of interest (the target sequence) is often a tiny fraction of the genome. To study it, either we must first purify it (by artificially increasing its copy number using a DNA polymerase) or use some method that specifically tracks it.
- Amplifying DNA sequences to produce multiple identical copies is achieved using a DNA polymerase within cells (DNA cloning) or by using a purified DNA polymerase *in vitro*, as in the case of the polymerase chain reaction (PCR).
- Cell-based DNA cloning uses vector molecules that can readily replicate in the host cell. The DNA to be cloned is often fragmented by cutting with restriction nucleases so that the fragments can be joined efficiently to similarly cut vector molecules to produce recombinant DNAs.
- Recombinant DNAs can be induced to enter a suitable host cell (transformation). Transformation is selective: each transformed cell normally has taken up a single DNA molecule. A transformed bacterial cell can multiply many times, resulting in large numbers of identical copies (clones) of the recombinant DNA.
- A DNA library is a bank of DNA clones that collectively include many different DNA sequences representing a complex starting population of genomic DNA or reverse-transcribed RNA.
- PCR is often used to amplify a target DNA sequence from within a complex population, such as genomic DNA. Oligonucleotide primers are designed to bind to the starting DNA at positions flanking the target sequence, and a heat-stable DNA polymerase makes DNA copies of the target sequence that can themselves serve as templates for making further copies, causing exponential amplification.
- PCR can readily amplify short DNA sequences, but is poor at amplifyinf long DNA sequences. Unlike cloning DNA in cells, it is not suited to producing very large amounts of a purified DNA.
- Being a rapid, highly-sensitive, and robust reaction, PCR is also suited to diagnostic work to identify sequences of interest. It can quickly allow cloning of small amounts of DNA from tiny quantities of starting cells and degraded tissues, and from thousands of different starting samples at a time.
- Isothermal amplification means amplification of target DNA sequences *in vitro* using a purified DNA polymerase at a constant temperature, and according to the type of method, amplification may be linear or exponential. Isothermal amplification methods can be a valuable alternative to PCR in diagnostic assays.
- Both PCR and some isothermal amplification methods can also be used to amplify all sequences in a complex starting nucleic acid present in very limiting quantities, such as the genome of a single cell. That often involves ligating adaptor oligonucleotides to the ends of all DNA fragments and then nonspecifically amplifying all the fragments using adaptor-specific primers.
- PCR is sometimes used as a way of capturing subsets of a genome or transcriptome, such as all exons in a genome (the exome).
- Nucleic acid hybridization is the key method used to track a DNA or RNA sequence of interest. The method relies on the specificity of base pairing—if two different nucleic acids are related in sequence, they may be able to form an artificial duplex that is stable under selected experimental conditions.
- To carry out nucleic acid hybridization, a test nucleic acid population with some sequence of interest is made single-stranded (denatured) and mixed with a probe population of known denatured nucleic acids. The object is to identify heteroduplexes where a single-stranded sequence of interest in the test sample has formed a stable hybrid with a known sequence within the probe population.
- In many types of nucleic acid hybridization, a homogenous, labeled probe population is used to identify related sequences in an unlabeled test population that is typically bound to a solid surface.
- In microarray hybridization, many thousands of unlabeled oligonucleotide probes are attached to a solid surface in a regular grid formation and hybridized in parallel with a labeled test nucleic acid population provided in solution. According to the amount of labeled DNA bound to each type of oligonucleotide, it is possible to quantify specific sequences that are complementary to each of the different probes.
- In many types of DNA sequencing, DNA samples are made single-stranded and a DNA polymerase is used to synthesize a complementary DNA (cDNA) in a way that provides a readout of the base sequence.
- In standard dideoxy DNA sequencing, selected, individual DNA samples are sequenced. The cDNA synthesis step uses a mix of normal and chain-terminating nucleotides, producing a nested set of fragments that differ incrementally by one nucleotide and that can be separated by gel electrophoresis.
- In massively-parallel DNA sequencing (next-generation sequencing, or NGS), a complex population of very many DNA templates is sequenced simultaneously and indiscriminately, generating huge amounts of sequence data (high-sequence throughput). There is no gel electrophoresis. Many of the methods involve DNA synthesis and monitor which of the four nucleotides is being incorporated at each step in DNA synthesis.
- Because of the chemistries involved, many NGS methods have low signal:noise ratios so that short sequences can be obtained only, and error rates are comparatively high for individual sequence reads. However, by sequencing many fragments with overlapping sequences it is possible to get consensus sequences with acceptably low error rates.
- Some of the NGS methods use unamplified DNA templates. These single-molecule sequencing methods can produce very long sequences, but sequence throughput may not be so high as for many of the NGS methods that use amplified DNA templates.

FURTHER READING

DNA cloning, amplification, and hybridization

Brown TA (2010) *Gene Cloning and DNA Analyses. An Introduction*, 6th edn. Wiley-Blackwell.

Geschwind DH (2003) DNA microarrays: translation of the genome from laboratory to clinic. *Lancet Neurol* **2**:275–282; PMID 12849181.

McPherson M & Møller S (2006) *PCR*, 2nd edn. Taylor and Francis.

Zhao Y *et al*. (2015) Isothermal amplification of nucleic acids. *Chem Rev* **115**:12491–12545; PMID 26551336. (A comprehensive review.)

Massively-parallel DNA sequencing: general overviews

Goodwin S *et al*. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**:333–351; PMID 27184599.

Jain M *et al*. (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12(4)**:351–356, PMID 25686389.

Levy SE & Myers RM (2016) Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet* **17**:95–115; PMID 27362342.

Mardis ER (2013) Next-generation sequencing platforms. *Annu Rev Anal Chem* **6**:287–303; PMID 23560931.

Reuter JA *et al*. (2015) High-throughput sequencing technologies. *Mol Cell* **58**:586–597; PMID 26000844.

Shendure J & Lieberman Aiden E (2012) The expanding scope of DNA sequencing. *Nat Biotechnol* **30**:1084–1094; PMID 23138308.

Vierstraete A (2012) Next Generation Sequencing for Dummies. http://users.ugent.be/~avierstr/nextgen/Next_generation_sequencing_web.pdf (A graphics-rich overview of technologies.)

Massively-parallel sequencing of amplified DNA: platforms and technology

Bentley DR *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**:53–59; PMID 18987734. (Illumina/Solexa sequencing.)

Ma Z *et al*. (2013) Isothermal amplification method for next-generation sequencing. *Proc Natl Acad Sci USA* **110**:14320–14323; PMID 23940326. (Wildfire technology.)

Margulies M *et al*. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376–380; PMID 16056220. (Roche/454 sequencing.)

Merriman B *et al*. (2012) Progress in Ion Torrent semiconductor chip based sequencing. *Electrophoresis* **33**:3397–3417; PMID 23208921.

Ronaghi M *et al*. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**:84–89; PMID 8923969. (Principle behind pyrosequencing.)

Valouev A *et al*. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**:1051–1063; PMID 18477713. (SOLiD sequencing.)

Massively-parallel sequencing of unamplified DNA

Deamer D (2016) Three decades of nanopore sequencing. *Nat Biotechnol* **34**:518–524; PMID 27153285.

Eid J *et al*. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**:133–138; PMID 19023044. (Pacific Biosciences system.)

Lu H *et al*. (2016) Oxford Nanopore MinilON sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14**:265–279; PMID 27646134.

Travers KJ *et al*. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* **38**:e159; PMID 20571086.

Websites for some established DNA sequencing companies

http://www.454.com/

http://www.illumina.com/systems/sequencing-platform-comparison.html?sciid=2014107IBN2

https://www.nanoporetech.com (Oxford Nanopore)

http://www.pacificbiosciences.com/

http://www.thermofisher.com/uk/en/home/life-science/sequencing/sequencing-technology-solutions.html (for SOLiD and Ion Torrent machines)

Analyzing the structure and expression of genes and genomes

7

DNA cloning and sequencing technologies that developed in the 1970s made it possible for the first time to systematically characterize genes. Bacterial recombinant DNA clones could be obtained that collectively represented almost all of the DNA in a nuclear genome (genomic DNA clones). In parallel, efforts were made to clone genome-wide transcripts by preparing RNA fractions from cells, converting the RNA into double-stranded DNA, and then cloning the resulting complementary DNA (cDNA clones).

By identifying and analyzing individual genomic and cDNA clones for a gene of interest, the exon–intron organization could be established, and the sequences of the exons could then be inspected to predict any likely protein product. That then allowed researchers to track the expression of a gene of interest at the RNA level (using gene-specific DNA or antisense RNA probes) and at the protein level (using specific antibodies developed against any protein products).

During the 1970s and 1980s much of the research on genes and how they are expressed was focused on individual genes of interest to specific research groups. This gene-centered focus was recognized to be inefficient. As DNA technologies improved, researchers began to make plans for sequencing all the different DNA molecules in our genome. In 1990 a large-scale project was launched to determine the complete sequence of the human genome and that of several model organisms. The resulting Human Genome Project, a triumphant international collaboration, made it possible to get information on all the genes in our genome, and to relate them to genes in other species. (But, in retrospect, determining the genome sequence was the easy part: working out what it means, and what it does, is the hard part; we cover that in Chapter 9.)

In Section 7.1 we describe how the genome projects were established and carried out, how genomic DNA clones were ordered and sequenced, and how sequences can be assembled. This section also covers bioinformatic approaches used to analyze the sequence, and to predict genes and aspects of their function, plus efforts to annotate genes within genome sequences. We describe in this chapter, too, how human gene expression is studied: we outline low-throughput gene expression studies, including high-resolution studies, in Section 7.2, then high-throughput expression analyses in Section 7.3, covering highly-parallel studies of both RNA transcripts and of proteins. We conclude in Section 7.4 by looking at how genomic technologies are increasingly being applied to single cells for different purposes. We defer description of global analyses of the human transcriptome until Chapter 8 (when we follow up a description of the architecture of the human genome by describing the ENCODE Project and the quest to determine the functional elements of our genome).

7.1 GENOME STRUCTURE ANALYSIS AND GENOME PROJECTS

To begin to understand the structure and functions of complex organisms such as humans, a necessary starting point is to obtain the complete genome sequence. The genomes of complex metazoans (multicellular animals) consist of a small number of different DNA molecules: several types of very large nuclear DNA molecules, corresponding to the different chromosomes, plus one type of mitochondrial DNA (mtDNA) molecule. Mitochondrial DNA molecules are easy to purify: they are confined to the cytoplasm, and there is only one type of mtDNA molecule. And because they are tiny, mitochondrial DNA molecules were comparatively easy to sequence. When we talk about genome projects, therefore, we mean the nuclear genome, and sequencing of all the different large chromosomal DNAs.

Unlike bacterial genomes, which often consist of a single type of comparatively small DNA molecule, the nuclear genomes of metazoans consist of very large DNA molecules. In the case of the human genome, for example, the average size of a chromosomal DNA molecule is 130 Mb. Sequencing technologies allow only comparatively short DNA sequences to be obtained and so the problem reduces to breaking up the DNA in cell nuclei into small, manageable-sized pieces (DNA fragmentation), sequencing fragments, and then trying to piece together all the little sequences to reconstruct the sequences of the chromosomal DNAs. But the sheer problem of **genome assembly**, arranging the sequences to correspond to the original linear order on the chromosome, had seemed a daunting prospect, and up until the early 1980s the idea of sequencing the human genome had seemed impossibly remote.

DNA libraries had offered, however, the possibility of “shotgun sequencing” of large genomes, whereby randomly selected clones in a genomic DNA library are sequenced until the full genome is covered. Clones with overlapping inserts are normally generated during construction of the libraries as a result of the random fragmentation of the DNA. Identical copies of the same chromosomal DNA molecules will be cleaved at different locations on the DNA, so any unique short sequence will be represented by a series of overlapping DNA fragments of different sizes (**Figure 7.1**).

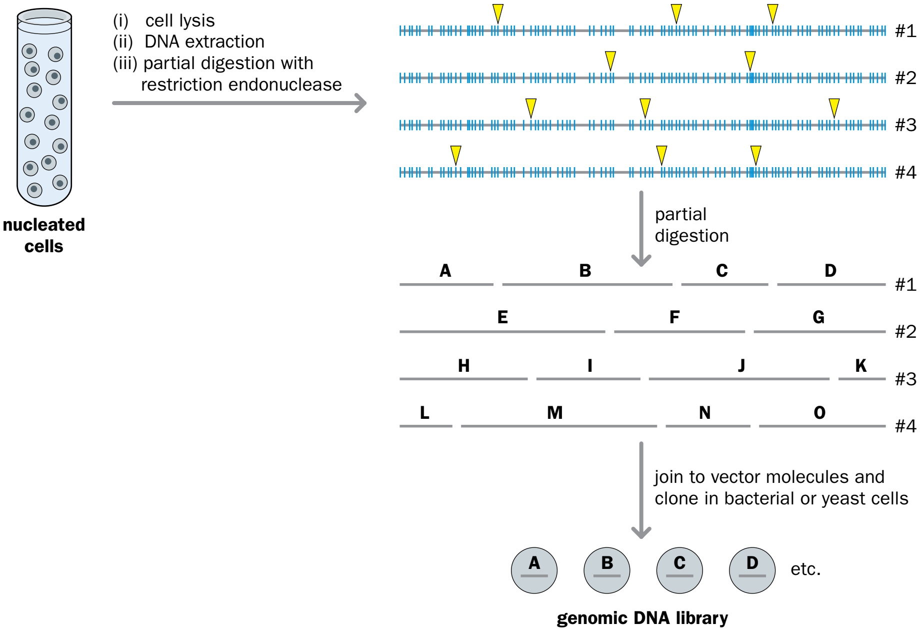

(i) cell lysis
(ii) DNA extraction
(iii) partial digestion with restriction endonuclease
nucleated cells
#1
#2
#3
#4
partial digestion
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
join to vector molecules and clone in bacterial or yeast cells
A
B
C
D
etc.
genomic DNA library

Figure 7.1 Generating clones with overlapping DNA inserts when constructing a genomic DNA library. Because all nucleated cells of an individual have essentially the same genomic DNA content, easily accessible cells (e.g., white blood cells) can be used as source material. Given that the cells contain the same sets of DNA molecules, the starting DNA will contain very many copies of each type of DNA molecule. Here, we imagine zooming in on a short region present on four copies of the same chromosomal DNA molecule (for example, paternal chromosome number 1 contributed by each of four different cells). The four copies of this sequence (#1 to #4) will have identical restriction sites for a specific restriction endonuclease (short vertical blue bars). However, because partial digestion is used, the DNA will be cleaved at only a small subset of the available restriction sites (indicated by yellow darts). Because the choice of which restriction site is cleaved is essentially random, the enzyme cuts the different copies of the same DNA sequence at different places, so that restriction fragments with overlapping sequences are produced (for example, fragment F from copy #2 partially overlaps fragments B and C from copy #1, fragments I and J from copy #3, and fragments M and N from copy #4).

Whole-genome shotgun sequencing (**Figure 7.2A**) is most successfully applied to small genomes; there are major difficulties in applying it to sequence large metazoan genomes for the first time. The human genome, for example, has huge numbers of repetitive DNA sequences, and because members of a repetitive DNA family can be very similar in sequence, it is difficult to map the individual sequences (and be certain of clone order).

To circumvent the problem of repetitive DNA, a different strategy was required for sequencing complex metazoan genomes. What was needed was some kind of initial scaffold, a series of **framework maps**, to anchor sequences of individual clones to defined subchromosomal regions, as a prelude to sequencing (**Figure 7.2B**). The burning question was this: what kind of framework maps could be established to aid sequencing of the human genome?

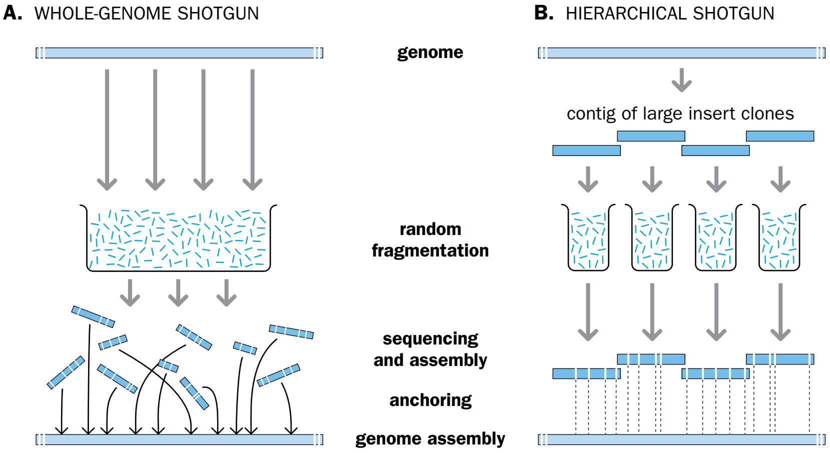


Figure 7.2 Two strategies for sequencing a genome. (**A**) Whole-genome shotgun sequencing involves indiscriminate fragmentation of the genome into small pieces of DNA that are readily sequenced. It quickly generates large amounts of sequence data but anchoring sequences to specific locations in the genome may be problematic. Large amounts of repetitive DNA in complex genomes make it difficult to unambiguously locate sequences to specific subchromosomal regions. (**B**) For first-time sequencing of complex genomes, it is more efficient to assemble contigs of large insert clones for each chromosome and then to fragment individual clones into pieces that are sequenced to reconstruct the sequence of the parent clone. (Adapted from Waterston RH *et al.* [2002] *Proc Natl Acad Sci USA* **99**:3712–3716; PMID 11880605. With permission from National Academy of Sciences. Copyright [2002] National Academy of Sciences, USA.)

Framework maps are needed in order to sequence complex genomes

For decades, human geneticists had envied geneticists working on model organisms where high-resolution classical genetic maps could be established readily. Such maps were based on mutant genes: by crossing mutants, the inheritance of individual phenotypes could be tracked through generations. If two mutant phenotypes showed a tendency to be co-inherited, the underlying genes could be expected to be reasonably closely linked on the same chromosome. Recombination between linked loci could provide a measure of the physical distance separating the two genes.

For ethical and practical reasons, classical genetic mapping could never be contemplated in humans. The breakthrough that paved the way to mapping the human genome came from a simple insight: genetic maps do not have to be based on gene mutations that affect the phenotype. Mutation is essentially a random process, and the great majority of mutations do not change the phenotype (only 1.2% of our DNA is coding DNA, and other highly-conserved functional sequences account for just a few percent of our genome). So, general DNA polymorphic markers can be used instead of gene variants. Once assays were developed to track this general type of DNA polymorphism, maps based on DNA markers could be established.

The initial genetic maps based on polymorphic DNA markers provided a skeleton for each chromosome, upon which more detailed framework maps could be built, containing a high density of DNA markers including polymorphic markers and numerous additional nonpolymorphic markers. The latter were simply chosen because they had a unique sequence that could be assayed by polymerase chain reaction (PCR) and mapped to specific subchromosomal locations using different methods, as described below.

Once suitably high-density marker–marker framework maps were developed, it was possible to build framework maps based on DNA clones. As described in the next section, DNA clone maps involve identifying and arranging DNA clones in a linear order that corresponds to the original linear chromosomal order of the inserts of the cloned DNA sequences. Comprehensive clone maps for each chromosome provided the final substrate for genome sequencing that delivered the ultimate physical map at 1 bp resolution.

The linear order of genomic DNA clones in a clone contig should mirror the original subchromosomal locations

The perfect substrate for sequencing of a complex genome would be to have each chromosomal DNA molecule represented by a set of DNA clones containing large inserts that have been linearly ordered according to the subchromosomal origins of their inserts, with neighboring clones having partially overlapping sequences, a **tiling path**. A series of such clones where the insert of each clone partially overlaps that of its neighbors with no gaps is known as a **clone contig**: the clones collectively represent a contiguous (continuous) DNA sequence from a subchromosomal region (**Figure 7.3**), or even a whole chromosome.

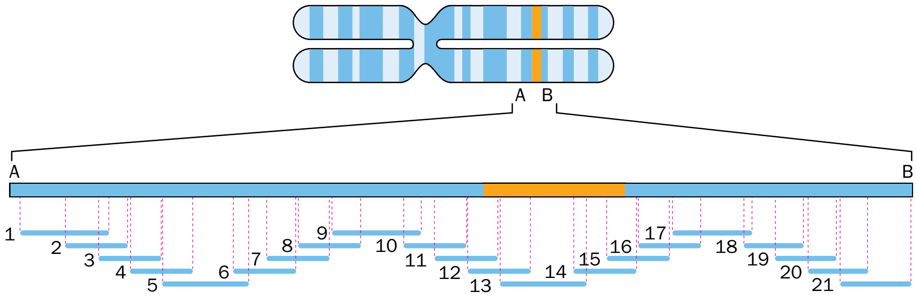


Figure 7.3 Assembling clones in a clone contig. Because partial digestion of genomic DNA is employed to construct a genomic DNA library, some DNA clones from a genome region will have inserts that partially overlap with each other (see **Figure 7.1**). Here, the chromosomal DNA sequence from positions A to B is represented by a linear series of overlapping DNA inserts, a tiling path. The collection of clones whose inserts produce a tiling path is known as a clone contig.

Recall from **Figure 7.1** that DNA library construction produces DNA fragments with overlapping sequences, but when they are cloned in cells the original order is lost; the problem for genome assembly is to put the clones back together with their inserts in the same order as on the chromosome of origin. To establish clone contigs, clones with overlapping inserts can be identified using **clone fingerprinting** methods. To do that, the standard procedure in genome projects is to establish a high-density DNA marker map for each chromosome, and then screen DNA clones for the presence or absence of DNA markers known to map to that chromosome. Clones are then identified that test positive for the same marker or markers. As we show below, the DNA markers needed to meet only two requirements: they should have a unique subchromosomal location, and be able to be conveniently assayed by PCR.

The Human Genome Project was an international endeavor and biology's first Big Project

The realization in the early 1980s that even large genomes, such as the human genome, could be sequenced sparked serious planning efforts to sequence the human genome. The official Human Genome Project (HGP) envisaged a 15-year timescale, commencing on October 1st, 1990. In addition to sequencing the human genome, the HGP had three other goals: to develop mapping and sequencing technologies; to carry out genome projects for five model organisms; and to investigate the ethical, legal, and societal implications.

The prioritized model organisms were the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the roundworm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the mouse *Mus musculus*. The first four were known to have substantially smaller genomes than the human genome and so were expected to be test beds to evaluate and then refine genome sequencing strategies and methodologies. The bulk of the sequencing for the more complex human and mouse genomes was expected to be carried out in the later stages, after learning from the smaller genome projects and extracting maximum benefit from technological improvements.

Because of the large scale involved, the HGP was biology's first Big Project. The ultimate aim was to achieve a periodic table for biology, based on genes rather than elements. The publically funded HGP was also a truly international endeavor. It came to be represented by the International Human Genome Sequencing Consortium of 20 different centers in the USA, UK, Japan, France, Germany, and China and was marked by extensive data sharing and collaboration on strategy, methodology, and data analysis.

Much of the genome sequencing technology was concentrated in a few very large genome mapping and sequencing centers with industrial-scale resources and massive data-analysis capabilities. Interacting with these centers was a worldwide network of small laboratories mostly attempting to map and identify disease genes and typically focusing on very specific subchromosomal regions. In addition to publically funded research efforts, privately funded research programs pursued similar but parallel sequencing projects.

The first framework maps for the human genome were genetic maps for individual chromosomes. The genetic maps were of low resolution but they provided the backbone on which to build a series of ever-more-detailed physical maps, culminating in clone contigs for each chromosome that were then used for the final sequencing stage.

The first human genetic maps were of low resolution and were constructed with mostly anonymous DNA markers

In genetic mapping, different polymorphic markers are assayed in all members of a variety of multigeneration families. The resulting genotypes are then analyzed using genetic linkage programs to work out how alleles of the different markers segregate at each meiotic event connecting parents to their offspring, and to identify markers where specific alleles co-segregate. The discovery of apparently random DNA polymorphisms in the human genome prompted the idea of constructing a comprehensive, nonclassical human genetic map, one that was predominantly based on anonymous DNA markers rather than on genes.

The first genetic linkage map of the human genome, published in 1987, was based on restriction fragment length polymorphisms (RFLPs), a type of DNA polymorphism that results in creation or destruction of one recognition sequence for a specific restriction nuclease (**Figure 7.4A**). Although an outstanding achievement, the map was of limited use: the markers were too few, just one marker per 9 Mb of DNA, and not very polymorphic (having just two alleles—either the restriction site is present or it is absent).

Second-generation human genetic maps were based on polymorphic microsatellite DNA markers, which are both quite common and often highly polymorphic (multiple alleles can often be distinguished; see **Figure 7.4B**). By 1994 an integrated genetic map (mostly based on microsatellites but containing some other markers) had a sufficiently high marker density (close to one marker per Mb), and from now on the major focus would be on developing and refining physical maps leading to the ultimate physical map, the complete DNA sequence of each chromosome. However, genetic mapping of the human genome continued outside the remit of the HGP: high-density single nucleotide polymorphism (SNP) maps were developed by the International HapMap (haplotype mapping) Consortium to help identify DNA variants contributing to common multifactorial diseases.

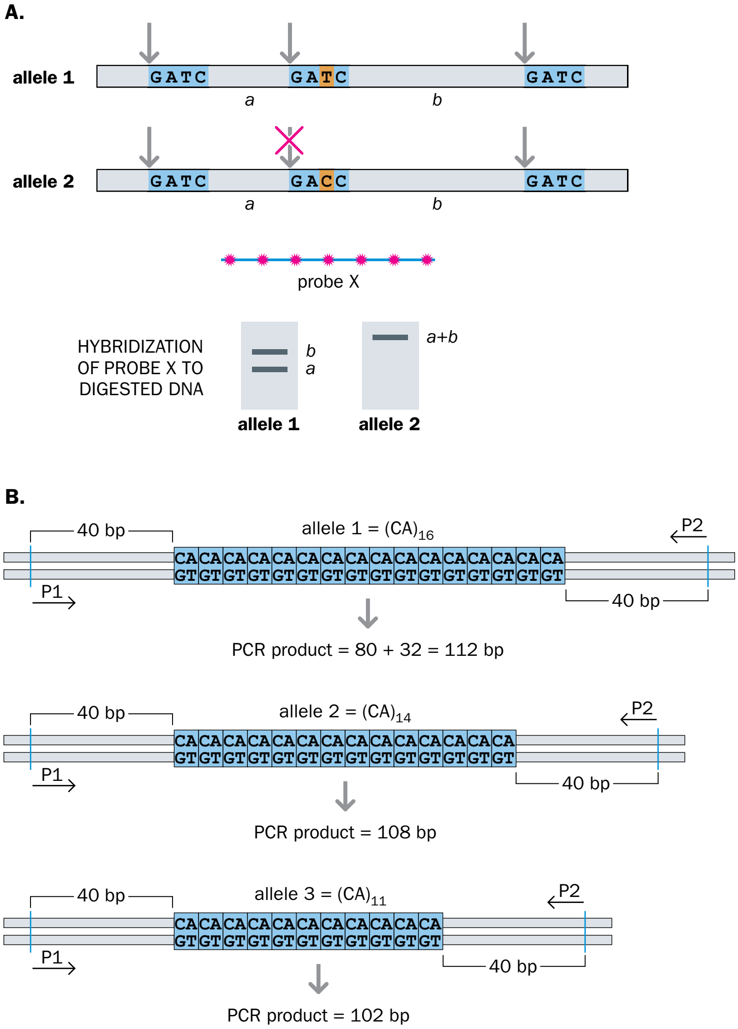


Figure 7.4 Restriction fragment length polymorphism (RFLP) and microsatellite DNA polymorphism. (**A**) RFLP. The example shows a polymorphic site for the enzyme *Mbo*I, which recognizes the sequence GATC. Allele 1 has a GATC sequence that is altered in allele 2 (GACC) so that *Mbo*I makes an extra cut in the DNA of allele 1 compared to allele 2 (at the middle GATC sequence shown). In the past, that difference would be detected by a Southern blot assay using a probe such as probe X, which would hybridize to the individual fragments *a* and *b* from allele 1 or a long fragment (*a* + *b*) from allele 2. The polymorphism is more conveniently detected by PCR using an upstream primer derived from the *a* sequence and a downstream primer from *b*, whereupon the amplified DNA can be digested with *Mbo*I and size-fractionated to distinguish between the alleles (not shown). (**B**) Microsatellite DNA polymorphism. Long runs of the (CA)/(TG) dinucleotide are prone to changes in copy number of the dinucleotide. Here the 5′ ends of upstream primer P1 and downstream primer P2 are located 40 bp from the microsatellite array, and so the length of amplified fragment will be 80 + the array length (32, 28, or 22 bp in this example). The different alleles can be separated by polyacrylamide gel electrophoresis.

Establishing high-density DNA marker maps and clone contigs for each human chromosome

Starting with the low-density human genetic map, the next task was to build a high-density marker map. That involved obtaining a large number of nonpolymorphic markers and mapping them to quite short, unique chromosomal regions. The key markers here were **sequence tagged site** (**STS**) markers, short (<1 kb) DNA sequences that occur at a unique location in the genome and that can readily be screened by a PCR assay. They are often nonpolymorphic (but some may contain a polymorphic site). Many STS markers were located outside genes, but some of them were located in genes and this subset of STS markers came to be known as *expressed sequence tag (EST)* markers.

STS markers were obtained by various routes, including from previously studied DNA clones that had been sequenced, and by randomly sequencing the ends of inserts of genomic DNA clones. They were mapped by different methods too. Sometimes **fluorescence *in situ* hybridization** (**FISH**) mapping was used: a genomic DNA clone containing the STS marker would be fluorescently labeled and hybridized to denatured DNA from a preparation of fixed metaphase or prometaphase chromosomes from human cells. In 1995 a human STS map was published with an average spacing of just less than one STS marker per 200 kb. The chromosomal locations of STS markers had been obtained by either using panels of human–rodent hybrid cells, or by STS content mapping as described below.

Mapping with somatic cell hybrid panels

Somatic cell hybrids are artificially constructed by fusing a human cell to a mouse or hamster cell. Initially unstable, the hybrid cells become stable after selectively jettisoning most of the human chromosomes. The loss of human chromosomes occurs randomly: different human chromosomes are retained in different hybrid cells (Figure 7.5A).

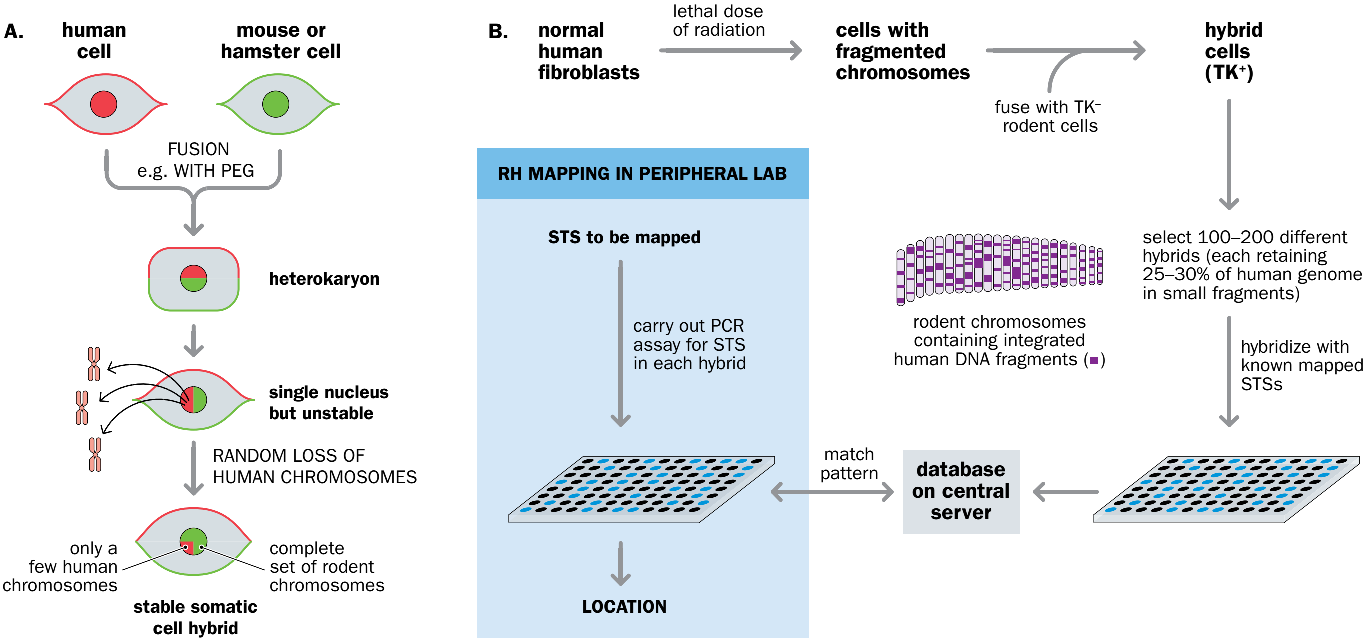

A.
human cell
mouse or hamster cell
FUSION e.g. WITH PEG
heterokaryon
single nucleus but unstable
RANDOM LOSS OF HUMAN CHROMOSOMES
only a few human chromosomes
complete set of rodent chromosomes
stable somatic cell hybrid
B.
normal human fibroblasts
lethal dose of radiation
cells with fragmented chromosomes
fuse with TK⁻ rodent cells
hybrid cells (TK⁺)
RH MAPPING IN PERIPHERAL LAB
STS to be mapped
carry out PCR assay for STS in each hybrid
LOCATION
rodent chromosomes containing integrated human DNA fragments (■)
select 100–200 different hybrids (each retaining 25–30% of human genome in small fragments)
hybridize with known mapped STSs
match pattern
database on central server

Figure 7.5 Somatic cell hybrids and the use of radiation hybrids to map human sequence tagged sites (STSs). (**A**) Principle of somatic cell hybrids. In the presence of certain agents, such as polyethylene glycol (PEG), cultured human and rodent cells can be induced to fuse, generating somatic cell hybrids. The initial fusion cells are heterokaryons, with both a human and a rodent nucleus. At mitosis the two nuclear envelopes dissolve, bringing human and rodent chromosomes together in a single nucleus. For unknown reasons, most human chromosomes fail to replicate in subsequent mitoses, and are lost. Eventually a variety of stable hybrid cell lines arise, each with the full set of rodent chromosomes plus a few types of human chromosome. (**B**) Mapping with radiation hybrids. Panels of radiation hybrids are produced by exposing cells with human chromosomes to radiation-induced chromosome breakage, prior to fusion to a rodent cell that is deficient in thymidine kinase (TK), followed by selection for TK⁺ cells. The radiation hybrids retain different sets of small human chromosome fragments; different human DNA markers will be present in some hybrid cells but absent in others. Although the pattern of fragment integration is mostly random, individual markers will give related typing patterns if they originate from closely spaced loci on a particular chromosome. The principle of a radiation hybrid map is therefore reminiscent of meiotic linkage analysis: the nearer together two DNA sequences are on a chromosome, the lower the probability that they will be separated by the chance occurrence of a breakpoint between them. Laboratories can map any unknown STS by assaying for it in a defined panel of radiation hybrids (RH) and comparing the pattern with patterns of previously mapped markers held on a central server.

Monochromosomal hybrids (with a single human chromosome) are particularly useful for mapping. To obtain them, donor human cells are first exposed to colcemid, causing the chromosome set to become partitioned into different discrete subnuclear packets (micronuclei). Subsequent centrifugation can produce **microcells** that have a single micronucleus containing just a few chromosomes and a thin rim of cytoplasm surrounded by an intact plasma membrane. When microcells are fused with recipient rodent cells, some hybrid cells retain just a single human chromosome. Monochromosomal hybrid cells were quickly developed to represent each of the human autosomes plus the X chromosome.

Subchromosomal mapping became possible after hybrids were designed to contain human chromosome fragments. The most popular approach involved exposing a cell containing one or more human chromosomes to a lethal dose of radiation (causing multiple chromosome breaks), and then fusing the irradiated cell with a rodent cell. The resulting **radiation hybrids** contained fragments of human chromosomes that integrated into rodent chromosomes.

Initially, the irradiated cells were monochromosomal hybrid cells, and fragments of a single human chromosome (plus fragments of many rodent chromosomes) randomly integrated into the genome of the rodent cell fusion partner. However, it was more efficient to irradiate a diploid human cell and then fuse it to a rodent cell. The resulting hybrids had multiple fragments from several different human chromosomes. The locations of the breakpoints on any one human chromosome varied from one irradiated cell to the next, and only a minority of the fragments successfully integrated, in a random way, into the rodent chromosomes (**Figure 7.5B**).

Constructing clone contigs by STS content mapping

Human genomic DNA libraries containing large insert DNAs were preferred for assembling clone contigs in preparation for DNA sequencing. The first such library was based on yeast artificial chromosomes (YACs) that can be constructed to have DNA inserts over a megabase in size (see **Box 7.1** for how YACs are made).

To build clone contigs efficiently, a quick and simple way of identifying clones with overlapping inserts was needed, and it was provided by constructing maps based on STS markers. By typing YAC clones with STS markers it was possible to assemble contigs of clones with partially overlapping inserts. But it soon became clear that YAC clones are unstable and YAC inserts were often not faithful representations of the original starting human DNA.

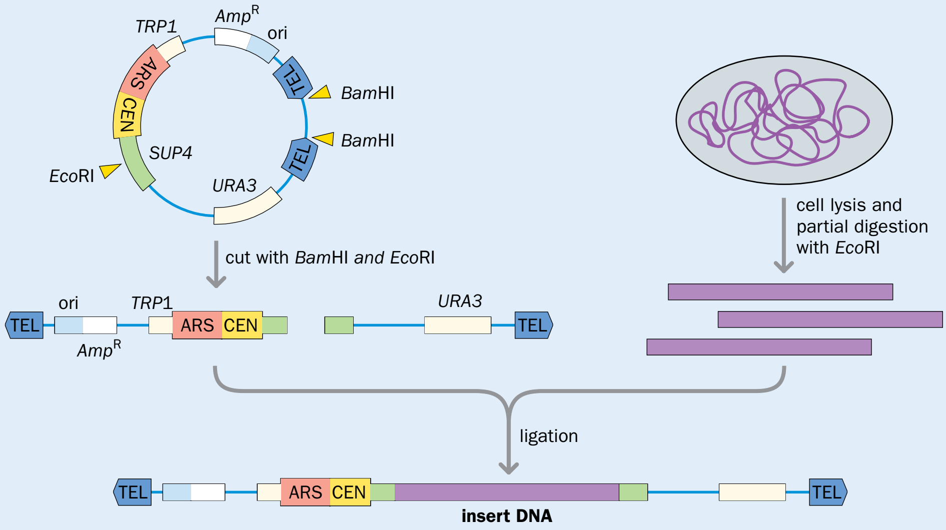
Because of their frequent instability, YAC clones could not be the template for the final genome sequencing effort, and alternative large-insert cloning systems were developed. Bacterial artificial chromosome (BAC) libraries (see **Box 7.1**) have smaller inserts than YACs but, crucially, human inserts are comparatively stable in BACs. Eventually, large BAC clone contigs were established for each human chromosome, paving the way for the final genome sequencing phase (see **Figure 7.6** for an overview of the different methodological approaches used in the Human Genome Project).

BOX 7.1 CLONING OF LARGE DNA FRAGMENTS USING YEAST AND BACTERIAL ARTIFICIAL CHROMOSOMES

YEAST ARTIFICIAL CHROMOSOME (YAC) CLONING

Very large DNA fragments can be cloned by making artificial chromosomes that are propagated in the budding yeast. The chromosomes of *Saccharomyces cerevisiae* are linear and vary in size from 200 kb to 1.5 Mb. As well as genes, they contain sequence elements essential for basic chromosomal functions (to protect the integrity of the underlying DNA, to replicate it, and to ensure faithful segregation into descendent cells). In *S. cerevisiae* the three types of sequence element needed for a DNA molecule to behave as a chromosome—centromere, telomere, and replication origin—are well defined and very short (see **Figure 2.21**).

To make a **YAC**, all that is needed, then, is to combine a suitably sized foreign DNA fragment with four short DNA sequences that can function in *S. cerevisiae* cells: two telomere sequences, one centromere sequence, and an autonomous replicating sequence that behaves as a replication origin. The resulting linear DNA molecule should have the telomere sequences correctly positioned at the termini (**Figure 1**). YACs cannot be transfected directly into yeast cells; instead, the external cell walls must first be removed. The resulting yeast spheroplasts can accept exogenous fragments but are osmotically unstable and need to be embedded in agar. The overall transformation efficiency is very low and the yield of cloned DNA is low (about one copy per cell). Nevertheless, the capacity to clone large DNA fragments (up to 2 Mb) allows functional studies on large regions of DNA and made YACs a vital tool in the physical mapping of some complex genomes, notably the human genome.



Box 7.1 Figure 1 Making yeast artificial chromosomes (YACs). A gene conferring ampicillin resistance (*Amp*^{R}) and a plasmid-derived origin of replication (ori) allow the YAC vector to be replicated to a high copy number in an *E. coli* host. In addition, the vector contains the three types of element required for DNA to behave as a chromosome in yeast cells: a centromere (CEN), an autonomous replicating sequence (ARS), and two telomeres (TEL). The vector also has three marker genes for growth in yeast cells (shown as white or green boxes). They comprise dominant alleles *TRP1* and *URA3*, which complement recessive alleles *trp1* and *ura3*, respectively, in the genome of the yeast host cell (a selection system for identifying transformed cells containing the YAC vector), plus the *SUP4* gene, which contains the cloning site. The yeast host cell has been engineered to have a gene mutation causing accumulation of red pigment (the host cells are normally red), but because the *SUP4* gene compensates for the mutation, cells transformed by the vector alone are colorless. Recombinants can be identified because the red color is restored after a foreign DNA fragment inserts into the *SUP4* gene, thereby inactivating it. To make recombinant YACs, the DNA to be cloned is partially digested with the restriction endonuclease *Eco*RI to give very large fragments (hundreds of kilobases) and mixed with linearized vector molecules cut with the restriction endonucleases *Bam*HI and *Eco*RI. Ligation occurs at the *Eco*RI ends to form linear recombinant DNA, with the insert DNA flanked by vector sequences that terminate in a TEL sequence.

BACTERIAL ARTIFICIAL CHROMOSOME (BAC) CLONING

A disadvantage of YACs is that they have a tendency to undergo rearrangements or deletions during the cloning process and during clone propagation. Using recombination-deficient yeast host strains helped reduce the frequency of transformation-associated alterations and mitotic instability, but because of the general problem with insert stability, attention turned to other cloning systems. One popular alternative was to use modified plasmid vectors that could propagate insert DNAs that were a few hundred kilobases in length.

When cloning DNA in plasmids, the aim is often to produce large amounts of cloned DNA for study. Plasmid vectors, such as pUC19, can be propagated at copy numbers of more than 100 per cell, allowing high yields of recombinant DNA. The downside of high-copy-number plasmid vectors is that insert sizes are low, and it is difficult to clone fragments of human DNA larger than 10 kb. However, if plasmid vectors are chosen to have an origin of replication that exerts stringent control of copy number (limiting the number of copies to one or two per cell), large inserts can be cloned into the vector and propagated stably.

The F-factor, an *E. coli* fertility plasmid, contains two genes, *parA* and *parB*, that maintain the copy number at 1–2 per *E. coli* cell. Plasmid vectors based on the F-factor system are able to accept large foreign DNA fragments (up to 300 kb), and the resulting recombinants can be transferred quite efficiently into bacterial cells using electroporation (an electrical field is applied to the cells, causing increased permeability of the cell membrane so that the large recombinant DNA molecules can enter the cell). The resulting **BACs** contain a low-copy-number replicon, and so only low yields of recombinant DNA can be recovered from the host cells. However, because of their great insert stability, BAC clones were the templates of choice for sequencing in the Human Genome Project.

The race to achieve a draft human genome sequence

The progress of the Human Genome Project was faster than expected (see **Box 7.2** for a timeline). Genetic maps were developed ahead of the original schedule, and the final stage of large-scale DNA sequencing was facilitated by developments in automated fluorescence-based DNA sequencing. Competition between publically and privately funded sequencing programs also drove a rapid final sequencing phase.

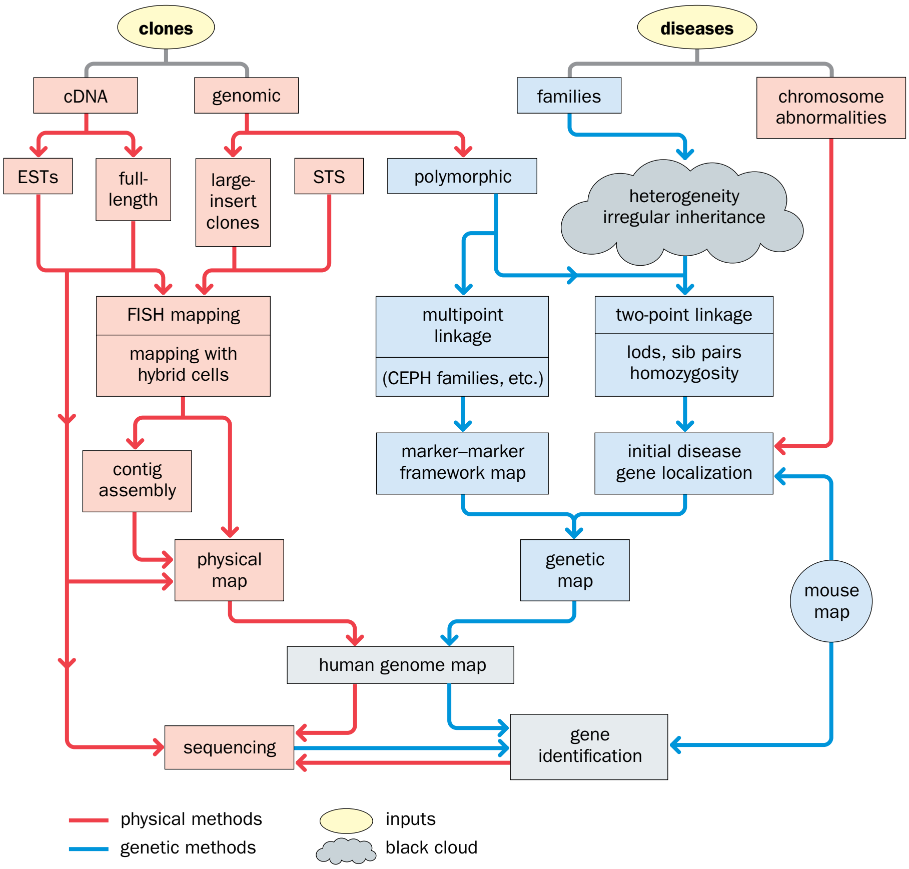


Figure 7.6 Major scientific strategies and approaches used in the Human Genome Project (HGP). The HGP required isolation of human genomic and cDNA clones. The clones were used to construct high-resolution genetic and physical maps that paved the way for genome sequencing. Inevitably, the HGP interacted with research on mapping and identifying human disease genes. The data produced were channeled into mapping and sequence databases permitting rapid electronic access and data analysis. Ancillary projects (not shown here) included studying genetic variation, genome projects for model organisms, and research on ethical, legal, and social implications. CEPH, Centre d'Etude du Polymorphisme Humain; EST, expressed sequence tag; FISH, fluorescence *in situ* hybridization; lods, logarithm of the odds scores; STS, sequence tagged site.

BOX 7.2 MAJOR MILESTONES IN MAPPING AND SEQUENCING THE HUMAN GENOME

1956: A first physical map of the human genome: light microscopy of stained tissue reveals that our cells usually contain 46 chromosomes, and that there are 24 different types. See the review by Gartler (2006) (PMID 16847465).

1977: Fred Sanger and colleagues publish the dideoxy DNA sequencing method (PMID 271968).

1980: David Botstein and colleagues propose that a human genetic map can be constructed using a set of random DNA markers such as RFLPs (PMID 6247908).

1981: Publication of the complete sequence of human mitochondrial DNA (PMID 7219534).

1984: A workshop, held in Alta, Utah, concludes that a human genome sequencing project is needed to permit high-efficiency detection of mutation.

1987: The USA Department of Energy publishes a report on a Human Genome Initiative.

1987: The first genetic linkage map of the human genome is published (PMID 3664638).

1987: The USA National Institutes of Health (NIH) sets up a dedicated Office of Human Genome Research (later renamed the National Center for Human Genome Research).

1987: The Human Genome Organization (HUGO) is established to co-ordinate international efforts.

1990: Official launch of the Human Genome Project (HGP) following implementation of a $3 billion 15-year project in the USA.

1992: The first comprehensive human genetic linkage map, based on microsatellite markers (PMID 1436057).

1993: A first-generation physical map of the human genome is reported, based on YAC clones (PMID 8259213).

1994: An improved human genetic map is published with one marker per centiMorgan (PMID 8091227).

1995: The first detailed physical map of the human genome, based on STS markers **(**PMID 8533086).

1996: A high-density human BAC library is published (PMID 8661051).

1997: An STS–radiation hybrid map of the human genome is published (PMID 9149939).

1998: Publication of the first comprehensive map of human gene-based markers (PMID 9784132).

1999: Publication of the essentially complete DNA sequence of chromosome 22 (PMID 10591208).

2001: Publication of draft human genome sequences by the International Human Genome Sequencing Consortium (IHGSC; PMID 11237011) and Celera (PMID 11181995).

2003: The essentially completed sequence of the human euchromatic genome is announced by the IHGSC. (For analysis, see PMID 15496913.)

2004: Over 21,000 human genes are validated by full-length cDNA clones (PMID 15103394).

2005/2007: The International HapMap Consortium reports increasingly detailed single nucleotide polymorphism (SNP) maps for the human genome. The 2007 map (PMID 17943122) has >3.1 million SNPs.

2007: The age of personal genome sequencing begins with euchromatic genome sequences for Craig Venter and James Watson. The latter sequence was obtained in just a few months using massively parallel sequencing (PMID 18421352).

An important rationale of the HGP—and a major motivation for privately funded genome sequencing—was to be able to study human genes. Starting in the early 1990s, attempts were made to obtain partial sequences from the 3′ untranslated regions of as many different human cDNA clones as possible, generating a huge number of ESTs. Introns are rarely found in the 3′ untranslated region of human genes and so a PCR assay based on the EST sequence could usually be used to type genomic DNA. Subsequently, systematic large-scale mapping of ESTs against panels of radiation hybrids produced the first comprehensive human gene maps. The resulting gene map was published in 1998 (see **Box 7.2**) and appeared to identify the positions of 30,000 human genes. However, the full extent of our genes could not be known with more precision until the genome sequence was delivered.

From an early stage it was clear that human genes were not uniformly distributed along or between chromosomes. Some chromosomes were rich in genes; others were gene-poor (**Figure 7.7**). The heterochromatic regions of the genome—including most of the Y chromosome, and substantial regions on chromosomes 1, 9, and 16—were known to be essentially devoid of genes and extraordinarily rich in repetitive DNA that would make mapping extremely difficult. As a result, the HGP was almost exclusively focused on the remaining euchromatic regions that collectively accounted for about 90% of the human genome.

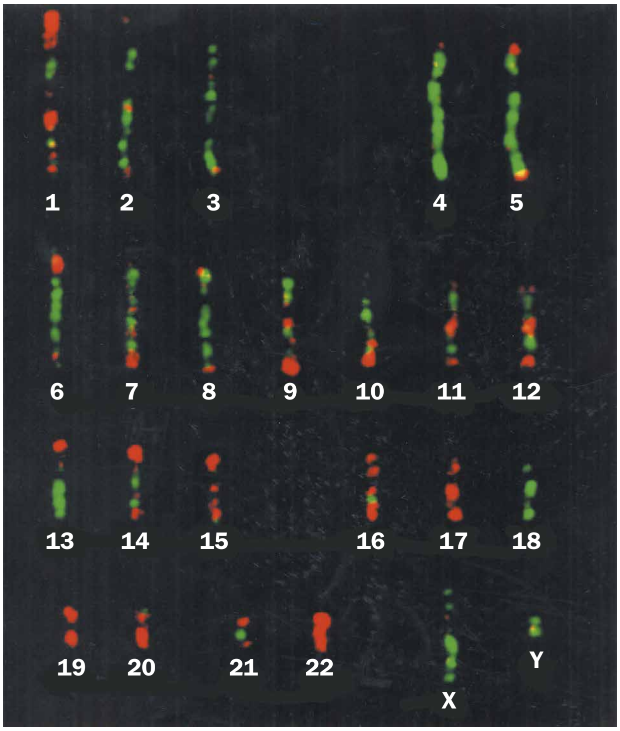


Figure 7.7 An early human gene map. Many human (and other vertebrate) genes have associated CpG islands (sequences about 1 kb long, often at the 5′ ends of genes, that differ from the bulk of the DNA in having many unmethylated CpG dinucleotides). The image shows the result of hybridizing a purified human CpG island fraction (labeled with a Texas Red stain) to human metaphase chromosomes. Late-replicating chromosomal regions (mostly transcriptionally inactive) are distinguished by incorporation of FITC-labeled bromodeoxyuridine (green signal). Yellow regions (overlap of red and green signals) denote late-replicating regions rich in genes (or strictly, CpG islands). Because CpG islands are gene markers, chromosomal regions that show a strong red signal have a high gene density (e.g., chromosome 22). Other chromosomes have very weak red signals and are gene-poor, such as chromosomes 4, 18, X, and Y. (Adapted from Craig JM & Bickmore WA [1994] *Nat Genet* **7**:376–382; PMID 7920655. With permission from Springer Nature. Copyright © 1994.)

For the publically funded International Human Genome Sequencing Consortium (IHGSC), most of the sequence was contributed by large genome centers. To ensure efficiency it was agreed that specific centers would take primary responsibility for the assembly of clone contigs and subsequent sequencing of individual chromosomes: for example, the Wellcome Trust Sanger Institute for chromosome 1, Washington University for chromosome 2, and so on.

A rival, commercial human genome sequencing effort was announced in 1999 when the Celera company declared its intention to produce a draft human genome sequence in 2 years: a whole-genome shotgun sequencing approach was to be used instead of the hierarchical shotgun sequencing approach of the IHGSC (which required BAC clones to be ordered into contigs before carrying out shotgun sequencing of the large inserts, which was more time-consuming). The ensuing race between the IHGSC and Celera accelerated the timetable. In 2001 both sides published a draft sequence of the human genome that covered about 90% of the euchromatic genome sequence. The euchromatic component is ~90% of the total genome, and so the draft sequences actually represented about 80% of the total genome, but 90% of the total gene sequence.

Although the race to achieve a draft genome sequence was perceived to have ended in a draw in 2001, it had not been a fair race, and the finishing line had not been reached—there was still some hard work to be done. The IHGSC had made their data freely available to all, posting sequence data updates on the Web every 24 hours. Celera took huge blocks of the IHGSC's sequence data, reprocessed it, and fed the data back into its own sequence compilation. The Celera sequence was, therefore, not an independently obtained human genome sequence. Unlike the IHGSC, Celera denied free external access to their sequence data (and continued to require expensive subscription charges to view their sequence data long after they had published their analyses).

The hard work toward finishing the euchromatic human genome sequence was carried out by the IHGSC, leading to publication of the virtually complete sequence in 2004. Even by the end of 2018, however, the human euchromatic genome sequence remained unfinished, largely because of genome assembly problems. Some of the problems arose because the source DNA was not a haploid genome (which would have been ideal for assembly purposes), and not even a diploid sequence of an individual person (for ethical reasons). Instead, a variety of different people contributed blood-cell samples for making the large-insert libraries used to provide the final DNA templates for sequencing (**Box 7.3**). Different regions of the human genome sequence therefore originated from different individuals, and structural variation between haplotypes impeded genome assembly. We report progress in finishing the human genome reference sequence in Chapter 9.

BOX 7.3 WHOSE GENOME IS IT ANYWAY?

BAC libraries provided the DNA templates for human genome sequencing in the publically funded Human Genome Project, and for ethical reasons it was decided that the cells used to provide the source DNA would come from multiple individuals who would be guaranteed anonymity. Rather than mixing the DNA samples from different individuals before making a library, individual libraries were made from cells contributed by a single individual. The initial plan envisaged making ten BAC libraries; each library was expected to contribute about 10% of the total DNA clones that would directly contribute to the genome sequence.

The libraries were constructed at the Roswell Park Cancer Institute (RPCI) in Buffalo, New York State, and at the California Institute of Technology, starting with anonymized blood-cell samples contributed by male and female volunteers who had responded to local advertising campaigns. Although 91.6% of the draft genome sequence originated from eight of the libraries prepared at the two centers, one library, RPCI-11, contributed to 74.3% of the draft genome sequence. (The RPCI-11 library was made at an early stage—but not reported until 2001 in PMID 11230172—and its large insert size proved popular, prompting early, widespread distribution of copies of the library to sequencing centers across the world; sequence variation analysis has since shown the anonymous male donor to have been primarily European in origin but with a 30% African Yoruban contribution.)

It subsequently became clear that allelic structural variation—large-scale deletions, duplications, inversions, and so on—were much more frequent than previously thought. As a result, in regions of the genome with high structural variation, the human genome reference sequence has been extended to include additional sequences so that multiple haplotypes can be represented. That has meant that DNA from several other individuals has also contributed to small proportions of the human genome reference sequence.

Genome projects for a diversity of model organisms

From the outset, the goals of the HGP included sequencing the genomes of five model organisms, partly as technology test beds. The genome sequences for four of the five model organisms prioritized by the HGP—*E. coli*, *S. cerevisiae*, *C. elegans,* and *D. melanogaster*—were obtained at an early stage in the project and were helpful in guiding the mapping and sequencing strategies that would be applied to the more complex human genome.

Draft mouse genome sequences were obtained in 2001/2002, and as detailed below, comparison of human and mouse sequences was to prove extremely important in identifying human genes and in establishing their exon–intron organizations. Genome projects were also launched for other organisms outside the remit of the IHGSC, and by late 2016 over 28,600 genome sequence projects had been completed and an additional 17,550 incomplete genome projects had been recorded.

Most of the sequenced genomes have come from prokaryotes, notably bacteria. As well as clarifying evolutionary relationships, the principal motivation for sequencing bacterial genomes has been to understand their involvement in pathogenesis or for applications in biotechnology. The several thousand genome projects for eukaryotes have been motivated by the need to understand general research models, models of disease and development, models for evolutionary and comparative genomic studies, farm animals and crops, and pathogenic protozoa and nematodes. Relevant genome data can be accessed at certain Websites that provide compilations of genome databases (see next section).

Powerful genome databases and browsers help to store and analyze genome data

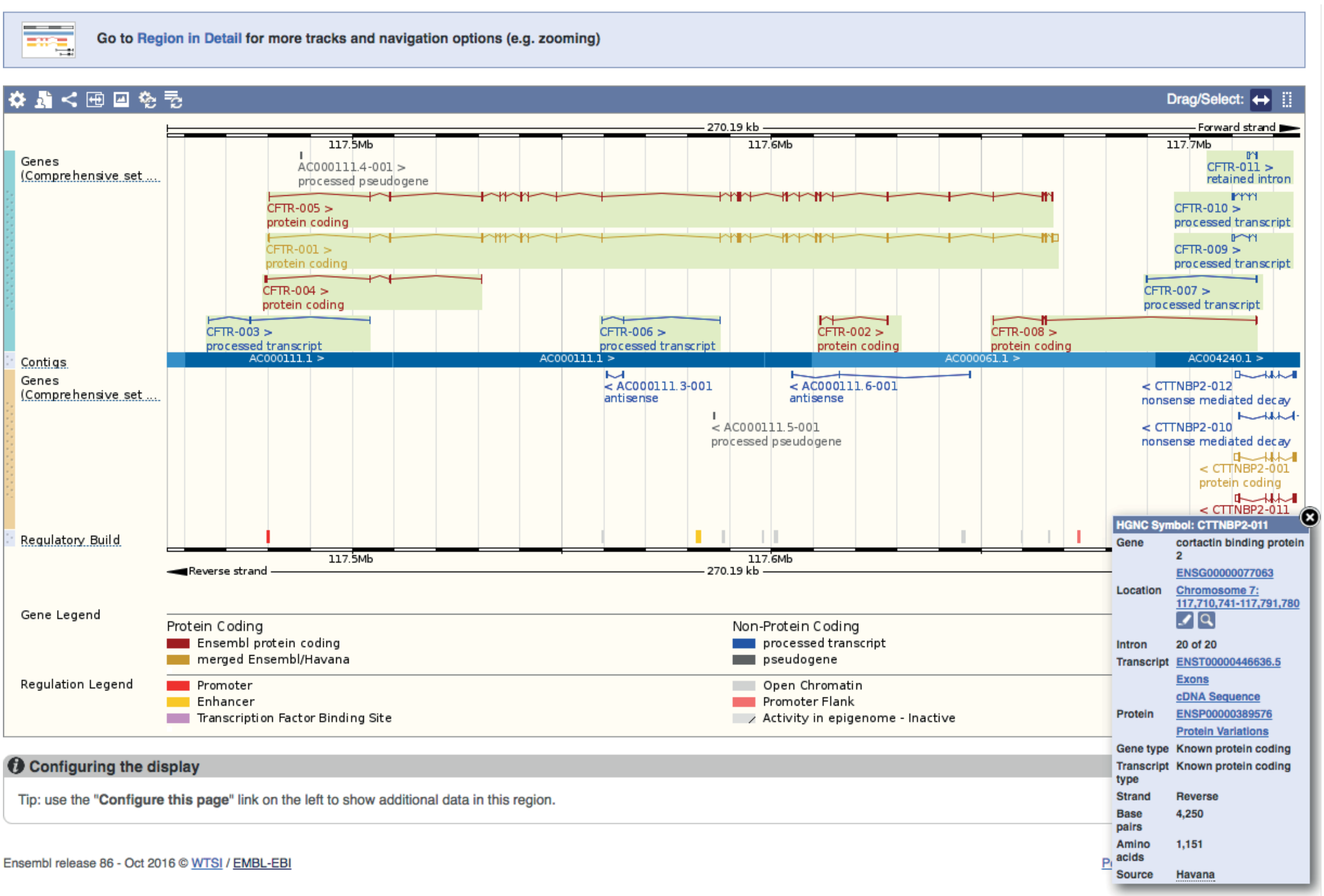
From the earliest stages of genome and gene analysis, central computer repositories were established for storing mapping data and sequence data produced in laboratories throughout the world. After major genome mapping and sequencing centers developed, a parallel data storage effort began when the individual genome centers developed dedicated in-house databases to store mapping and sequencing data produced in their own laboratories. The data from the publically funded genome projects were made freely available through the Web.

As genome data began to be produced in very large quantities, strenuous efforts were devoted to developing new genome databases (**Table 7.1**) and designing new software that would permit the huge amounts of mapping and sequence data and associated information to be searched in a systematic and user-friendly way. A major new focus on in-silico (computer-based) analyses made vital contributions to our understanding of the structure of genes and genomes.

An important advance was the development of **genome browsers** with graphical user interfaces to portray genome information for individual chromosomes and subchromosomal regions. Users of genome browsers can quickly navigate the sequence of a selected human chromosome moving from large scale to nucleotide scale, identifying genes and associated RNA transcripts in regions of interest, with exon-intron organization revealed as the user zooms in (see **Figure 7.8** for an example). Thereafter, the user can click on features of interest to allow numerous connections to other databases and programs (permitting, for example, amino acid sequences to be obtained for selected transcripts, or evolutionary conservation of selected sequences by comparison with homologs in other genomes). As more and more information is obtained for genes and other functional units, more informative and precise gene annotation will be available in frequent, periodic updates of the genome browsers and databases.

TABLE 7.1 SOME OF THE MAJOR EUKARYOTIC GENOME BROWSERS AND GENOME DATABASES

Resource	Originator/host	Website URL
GENOME BROWSERS		
Ensembl	Wellcome Trust Sanger Institute/ European Bioinformatics Institute (EBI)	http://www.ensembl.org
NCBI Genome Data Viewer	USA National Center for Biotechnology Information (NCBI)	https://www.ncbi.nlm.nih.gov/genome/gdv
UCSC Genome Browser	University of California at Santa Cruz	http://genome.ucsc.edu
COMPILATIONS OF GENOME SEQUENCES		
EBI Genomes	EBI	http://www.ebi.ac.uk/genomes
Genomes Online Database (gold)	US Department of Energy	https://gold.jgi.doe.gov/
ORGANISM-SPECIFIC GENOME DATABASES		
FlyBase (*Drosophila*)	FlyBase Consortium	http://flybase.org/
MGI (Mouse Genome Informatics)	Jackson Laboratory	http://www.informatics.jax.org/
NCBI Human Genome Resources	NCBI	http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/
SGD (*Saccharomyces* Genome Database)	Stanford University	http://www.yeastgenome.org/
WormBase (*C. elegans*)	WormBase Consortium	http://www.wormbase.org/
ZFIN (Zebrafish Information Network)	University of Oregon	http://zfin.org/



Tip: use the "**Configure this page**" link on the left to show additional data in this region.

HGNC Symbol: CTTNBP2-011

Gene	cortactin binding protein 2 ENSG00000077063
Location	Chromosome 7: 117,710,741-117,791,780
Intron	20 of 20
Transcript	ENST00000446636.5 Exons cDNA Sequence
Protein	ENSP00000389576 Protein Variations
Gene type	Known protein coding
Transcript type	Known protein coding
Strand	Reverse
Base pairs	4,250
Amino acids	1,151
Source	Havana

Figure 7.8 An example to illustrate using the Ensembl genome browser. Here the October 2016 version of Ensembl was queried with the human *CFTR* gene (the cystic fibrosis transmembrane regulator gene spans nucleotides 117,465,784–117,715,971 on chromosome 7). The two frames with the title "Genes" show exon–intron organizations of the different transcripts from the two DNA strands: the upper frame shows the *CFTR* sense transcripts, and the frame below the "Contigs" bar (which gives GenBank accession numbers for indicated DNA sequence contigs) shows *CFTR* antisense transcripts, plus transcripts from a neighboring, partially overlapping, protein-coding gene, *CTTNBP2* (transcribed from the opposite DNA strand). In each case, exons are represented by short vertical bars that are connected by introns (flattened chevrons); protein-coding transcripts are represented by red- and gold-colored lines, noncoding transcripts by blue lines (according to the gene legend at bottom). There are five *CFTR* protein-coding transcripts (two full-length isoforms and three smaller isoforms), six noncoding sense transcripts, and two antisense transcripts. The "Regulatory Build" shows colored vertical bars representing the positions of indicated regulatory sequences. Clicking over individual items brings up additional information, as shown here by clicking on the *CTTNBP2-011* transcript at bottom right (clicking on underlined items in blue font allows access to further information, often presented in further graphical frames).

Bioinformatic approaches to predicting genes and their functions, and gene annotation

Most human genes had not been uncovered before the Human Genome Project. As the DNA sequences of chromosomes were being spewed out of the big sequencing centers they were scanned by powerful bioinformatics programs to predict genes that could subsequently be validated. Some programs sought to identify genes *ab initio* by using hidden Markov models to analyze the human sequence by itself. However, comparisons with genome sequence data from other organisms and from previously studied genes provided extremely valuable support in identifying human genes, and also often provided important functional clues as to what kind of product a predicted human gene might make. As functions became mapped to genes, collaborative efforts sought to annotate genes in a comprehensive, systematic way and the Gene Ontology project was developed to address the need for consistent descriptions of gene products across species and databases, as described below.

Predicting genes and their functions *in silico*

Genome sequence data can be interrogated by a suite of computer programs that seek novel genes by screening for general gene characteristics. Genes are transcribed into RNA, for example, and so confidence in candidate gene sequences is high when sequence homology searching (**Box 7.4**) reveals many highly-related expressed sequence tags (ESTs) and/or cDNA sequences in sequence databases. Vertebrate genes also often show altered base composition: in addition to having a significantly higher percentage of GC (%GC) than the genome average, they are frequently associated with CpG islands, regions around 1 kb in length that are both GC-rich and also have a significantly higher frequency of the dinucleotide CpG than the bulk of the genome. (We consider CpG islands in more detail in Chapter 9.) For protein-coding genes, three gene-associated facets have been especially exploited to identify novel genes, as listed below.

- ***Open reading frames (ORFs)***. Open reading frames are needed in long coding DNA sequences to make a protein. There are three possible reading frames for each of the two DNA strands. If we make the simplifying assumption that the three types of termination codon occur with equal frequencies, then in human DNA, with an average base composition of 41% GC, one might expect that a stop codon would occur by chance roughly once every 50 nucleotides or so in each of the six possible reading frames (3/64 codons are termination codons, but termination codons are AT-rich: seven out of the nine nucleotides in the three stop codons are A or T). Coding DNA has a significantly higher %GC, and so statistically longer ORFs can generally be expected in coding DNA (making the same assumptions as above, a stop codon might be expected to occur by chance once every 80 nucleotides or so in DNA regions with 50% GC). Added to that is the frequent occurrence of intervening introns: the coding DNA is usually split, and an average internal exon in a human protein-coding gene is about 150 nucleotides in length. Long ORFs (>300 nucleotides) become prioritized for follow-up investigations.
- ***Exon prediction***. Programs such as GENSCAN exploit the observation that there are short conserved sequences at splice junctions and assign high probability to a predicted exon if there is also a large ORF.
- ***Evolutionary conservation***. Homology searching—comparing sequences at the nucleotide level, or a predicted protein against protein sequence databases or against translated nucleotide sequences—is an especially powerful tool for gene identification (see **Box 7.4**). That is so because functionally important sequences, such as proteins and coding DNA sequences, have been highly conserved during evolution. A recently discovered human gene would often be found to have previously characterized homologs in model organisms (including other vertebrates, invertebrates such as *Drosophila* and *C. elegans*, or even in microbial cells). Because protein sequences are more conserved than the corresponding DNA sequences, predicted translation products of a candidate gene are typically used to search for homology against all known protein sequences using BLASTP, and against the predicted translation products of all known nucleotide sequences using TBLASTN. Identifying related homologs in other species may provide valuable clues to the function of the human gene. Additionally, even if a related homolog is not found, homology searching may identify small subcomponents of a candidate gene that are suggestive of a function—for example, sequence motifs associated with DNA-binding properties, such as zinc fingers.

Integrated gene-finding software packages have been developed that combine programs designed to identify exons and gene-associated motifs with general sequence-homology-based database-searching programs. Although the bioinformatics approaches have been of great help in predicting genes (and some aspects of their function), they have also had their limitations. Gene prediction *in silico* has been especially valuable in the case of protein-coding genes, but programs like GENSCAN do have a tendency toward overprediction.

The Gene Ontology (GO) Project

Searching across databases (and the broad scientific literature) has been hampered by the sometimes wide variation in terminologies used in different species. The Gene Ontology (GO) Consortium was formed in 1998 to institute a standardized system of **gene ontology** to represent gene function across genomes and species, and the resulting GO project began to develop a controlled vocabulary to describe the attributes of genes and gene products

BOX 7.4 SEQUENCE HOMOLOGY SEARCHING

Powerful computer programs have been devised to permit rapid searching of nucleic acid and protein sequence databases (and of dedicated genome sequence databases) for significant sequence matching (homology) with a test sequence. BLAST, BLAT, and FASTA programs are popular (see **Table 1**).

Programs such as BLAST and FASTA use algorithms to identify optimal sequence alignments and typically display the output as a series of pairwise comparisons between the test sequence (*query sequence*) and each related sequence that the program identifies in the database (*subject sequences*).

In pairwise sequence alignments, two major approaches can be taken to calculate the optimal sequence alignment between a query sequence and a subject sequence.

- ***Global alignment***. The object is to get the best alignment that can be made over the entire length. This is appropriate when the two sequences are of similar length and have a significant degree of similarity throughout (**Figure 1A**).
- ***Local alignment***. The object is to get a number of short sequence alignments. It is used when comparing substantially different sequences that may also differ significantly in length (**Figure 1B**).

Popular algorithms used in nucleotide sequence alignments include the Needleman–Wunsch algorithm (which seeks to maximize the number of matched nucleotides) and the Waterman–Eggert algorithm (which seeks to minimize the number of mismatches); see Durbin *et al*. (1998) under Further Reading for background information.

For coding sequences, nucleotide sequence alignments can be aided by parallel amino acid sequence alignments using the assumed translational reading frame for the coding sequence. Pairwise alignments of amino acid sequences are aided by the fact that there are 20 different amino acids (compared to just four different nucleotides), and by the importance of chemical subclasses of amino acids (conservative nucleotide substitutions replace an amino acid with one that is chemically related to it, typically belonging to the same subclass). As a result, algorithms used to compare amino acid sequences typically use a scoring matrix in which pairs of scores are arranged in a 20 × 20 matrix, where higher scores are accorded to identical amino acids and to ones that belong to the same chemical subclass (such as lysine and arginine, or isoleucine and leucine) and lower scores are given to amino acids that are chemically rather different (such as arginine and leucine).

The typical output gives two overall results for percent sequence relatedness. The first is concerned with sequence identity (matching of identical amino acids only) and the second also takes into account amino acids that are chemically related; see **Figure 2**).

BOX 7.4 TABLE 1 COMMONLY USED PROGRAMS FOR BASIC SEQUENCE HOMOLOGY SEARCHING

Program	Features
BLASTN	Compares a nucleotide sequence against a nucleotide sequence database
BLASTX	Compares a nucleotide sequence translated in all reading frames against a protein sequence database
BLASTP	Compares an amino acid sequence against a protein sequence database
TBLASTN	Compares an amino acid sequence against a database of nucleotide sequences translated in all reading frames
BLAT	BLAST-like program that offers extremely rapid searching at nucleotide or protein levels against a selected genome
FASTA	Compares a nucleotide sequence or amino acid sequence against, respectively, a nucleotide or protein sequence database
TFASTA	Compares an amino acid sequence against a database of nucleotide sequences translated in all reading frames

BLAST and FASTA programs are widely available, such as from the European Bioinformatics Institute (http://www.ebi.ac.uk/Tools/sss/) and the USA NCBI (https://www.ncbi.nlm.nih.gov/BLAST/). BLAT is hosted at the University of California at Santa Cruz (http://genome.ucsc.edu/).

A.

```
CCGATAGAGGAC-GGTACTATAGCA-AGAGACCACGGAGACCATTGGGGACGAA-TGCA
** *** ***** ****** ***** *********** ******* **** *** ****
CC-ATACAGGACAGGTACTTTAGCAAAGAGACCACGGCGACCATTAGGGA-GAAATGCA
```

B.

```
CCGAAAGAGGT-----GCTTTAG-ACAC-----------AC-TATTTCG--GA--TACA
** * * ***       ** *** * *            ** *** * *  **  * **
CC-ATACAGGACAGGCACTATAGCAAAGAGACCACGGCGACCTATATGGGAGAAATGCA
```

Box 7.4 Figure 1 Examples of sequence pairs suited to global and local alignment. (**A**) The two sequences differ in length by just one nucleotide and are closely related; global alignment can be used to find an optimal alignment across the whole length. (**B**) The two sequences are significantly similar but the upper sequence is much shorter than the lower one. A local alignment strategy looks for matching of short sequence motifs. Asterisks signify identity, and dashes signify deletions.

```
score = 52.8 bits (125), expect = 9e-08
identities = 39/120 (32%), positives = 57/120 (47%), gaps = 9/120 (7%)

QUERY: 1    AKLLIKHDSNIGIPDVEGKIPLHWAANHKDPSAVHTVRCILDAAPTESLLNWQDYEGRTP 60
            A+LL++HD++       G  PLH A +H +   +  V+ +L    +     W  Y   TP
SBJCT: 548  AELLLEHDAHPNAAGKNGLTPLHVAVHHNN---LDIVKLLLPRGGSPHSPAWNGY---TP 601


QUERY: 61   LHFAVADGNLTVVDVLTSY-ESCNITSYDNLFRTPLHWAALLGHAQIVHLLLERNKSGTI 119
            LH A     + V   L  Y  S N  S   +  TPLH AA  GH ++V LLL +  +G +
SBJCT: 602  LHIAAKONOIEVARSLLOYGGSANAESVOGV--TPLHLAAOEGHTEMVALLLSKOANGNL 659
```

Box 7.4 Figure 2 Sequence identity and sequence similarity in aligned protein sequences. The BLASTP output shown here resulted from querying the Swiss-Prot protein database with a query sequence of amino acids 165–283 of the inversin protein. The subject sequence shown here is a mouse erythrocyte ankyrin sequence. The program considers not just sequence identity (39 of the 120 positions, or 32%, have identical residues in the two sequences, shown as red letters in the middle rows) but also sequence similarity (an additional 18 positions have chemically similar amino acids, represented by the + symbol), giving a total of 57 positive positions where there is sequence identity or sequence similarity.

in any organism. Three separate ontologies—biological process, cellular component, and molecular function—were developed to allow for the annotation of molecular characteristics across species, and they may be broad or more focused. For example, the biological process can be as broad as signal transduction or more restricted, such as alpha-glucoside transport. Each vocabulary is structured so that any term may have more than one parent as well as zero, one, or more children. This makes attempts to describe biology much richer than would be possible with a hierarchical graph. By 2014 the GO project had developed formal ontologies to represent more than 40,000 biological concepts.

Carrying out sequence assembly in complex genomes

Early approaches to sequencing a complex genome involved hierarchical shotgun sequencing (see **Figure 7.2B**). The aim then was to identify the minimum number of DNA clones that need to be sequenced to provide an acceptable "sequence depth" (ideally, a region of interest on the DNA molecule should be represented by a large number of sequence reads to maximize the chances of obtaining an unambiguous consensus sequence). The ideal template for genome assembly would be a continuous clone contig for each chromosomal DNA molecule. That is, for each chromosome there would be a continuous tiling path of clones with overlapping DNA inserts. (Visualize the contig shown in **Figure 7.3** extending across the whole chromosome.) For complex genomes, however, there are problems achieving that aim.

The long arrays of tandem, highly-repetitive DNA sequences associated with constitutive heterochromatin provide a major obstacle for genome assembly: very high levels of sequence identity between the repeats in a long array mean that identifying overlapping sequences is difficult. Some gene clusters provide obstacles to genome assembly, too, such as the long arrays of tandem repeats specifying the 18S, 5.8S, and 28S ribosomal RNAs.

Outside difficulties with long arrays of tandem repeats, different problems can affect genome assembly. Some genome regions may not be represented by sequence reads (certain sequences may not be propagated well in bacteria, for example, and so be poorly represented in libraries of DNA clones). Challenges can also be posed by other, closely related repeats. And structural variation between haplotypes can pose problems for establishing contigs. As a result of these difficulties, the sequences of even the euchromatic part of complex genomes, such as the human genome, are unfinished. A chromosome is typically represented by scaffolds that each contain multiple contigs, with individual contigs separated by gaps whose approximate lengths are known (**Figure 7.9**).

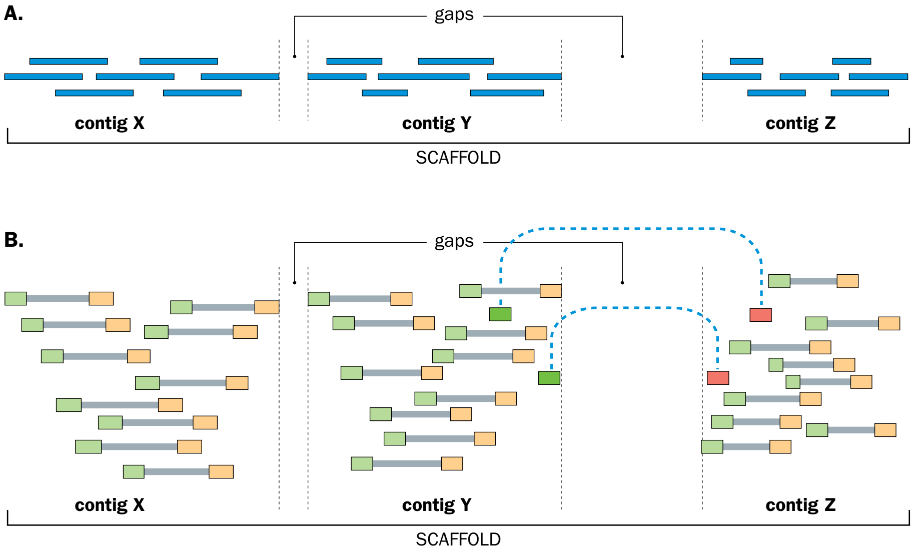


Figure 7.9 Scaffolds and contigs. Contigs are made up of overlapping DNA sequences without gaps. Scaffolds are made up of two or more contigs with gaps, where the gaps are of approximately known length and each contig has a known orientation within the scaffold. (**A**) An example of a scaffold with three clone contigs developed in a hierarchical shotgun sequencing strategy. The blue bars represent overlapping sequenced inserts of large-insert clones (whose sequences had been obtained after shotgun sequencing). (**B**) An example of contigs and scaffolds established using massively parallel sequencing of paired ends (pale green and orange boxes) and mate-pair sequences (dark green and red boxes). The approximate size of the large gap between contigs Y and Z may be known because of the linked mate-pair sequences (connected by dashed blue lines) that were originally derived from sequences that can be up to tens of kilobases apart in the genome (but have been artificially brought together onto short fragments for sequencing, as detailed in **Figure 6.21**.)

Assembly statistics

Different assembly statistics are used. N50, the most commonly used statistic, describes an average length of a set of sequences (contigs or scaffolds), but it is not the mean or median length. It is defined as the largest length *L* such that 50% of all nucleotides are contained in contigs of size at least *L*. To calculate a contig N50, for example, every contig is first ordered by length from longest to shortest. Then, starting from the longest contig, the lengths of each contig are sequentially summed until this running sum equals one half of the total length of all contigs in the assembly. At that point, the length of the shortest contig in the list is the contig N50. L50, another commonly used statistic, signifies the number of sequences whose sum length exceeds 50% of the total size of the sequence assembly. (Note: because of the confusing initial letters, some people have preferred to invert the meanings of N50 and L50, so that N50 becomes a number and L50 represents a length.) **Table 7.2** provides an example of statistics for scaffolds and contigs for the human genome reference sequence.

TABLE 7.2 SOME ASSEMBLY STATISTICS FOR THE HUMAN GENOME REFERENCE SEQUENCE GRCh38.p12

Statistic	Value
Total sequence assembly size (nucleotides)	3,257,319,537*
Total assembly gap length (nucleotides)	161,368,351
Number of gaps between scaffolds	349
Number of scaffolds	874
Scaffold N50 (nucleotides)	59,364,414
Scaffold L50	17
Number of contigs	1535
Contig N50 (nucleotides)	56,413,054
Contig L50	19

Here, the N50 statistic signifies the largest length *L* such that 50% of all nucleotides are contained in contigs or scaffolds of at least size *L*. L50 signifies the number of sequences when the sum length exceeds 50% of the total size of the sequence assembly. Data are from the Genome Reference Consortium's Human Build 38, patch release 12 (GRCh38.p12) published in December 2017 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405). The N50 value is a measure of assembly quality: higher numbers mean greater continuity of the sequence. For comparison, the contig N50 values for the IHGSC's draft human genome sequence and the "finished" human euchromatin genome sequence reported in 2004 were, respectively, 81 kb and 38.5 Mb.

* This value is higher than the total genome length (3.1 Gb) because it also includes additional sequences for alternative haplotypes, as detailed in Chapter 9.

7.2 BASIC GENE EXPRESSION ANALYSES

A first step in working out how a gene functions is to track its expression at the transcript or protein levels. Different starting sources can be used and different technologies can be employed with varying resolution and throughput (see **Table 7.3**).

Crude RNA/cDNA or protein extracts are frequently used as source material. Sometimes, however, expression is sampled in tissue sections or even whole embryos that have been fixed so as to preserve the original *in vivo* morphology. Gene expression can also be studied in live cells in tissue culture, and even in certain living experimental organisms that have optically transparent tissues at certain stages in development (allowing tracking of expression of genes tagged with a fluorescent chemical group).

Laser capture microdissection uses a laser to dissect out microscopic portions of a tissue to produce pure cell populations from sources such as tissue biopsies and stained tissue, and even single cells. As a result, gene expression analyses can be focused on single cells or on homogeneous cell populations that will be more representative of the *in vivo* state than cell lines.

Low-resolution expression patterns are initially sought from genes by tracking gross expression in RNA extracts or protein extracts. In addition to being able to sample expression in different tissues, these patterns may provide useful information on the level of expression and on expression-product variants that can differ in size (isoforms). Interesting expression patterns can be followed-up using methods to track expression within a cell, or within groups of cells and tissues that are spatially organized in a manner representative of the normal *in vivo* organization.

TABLE 7.3 DIFFERENT LEVELS OF EXPRESSION MAPPING

Study material	Resolution	Throughput*	Examples
RNA in cells or cell extracts, or purified cDNA	High	Low	Tissue and whole-embryo *in situ* hybridization Cellular *in situ* hybridization
	Low	Low to medium	Northern blot hybridization RNA dot-blot hybridization Ribonuclease protection assay RT-PCR/qPCR
	Low	High	Microarray hybridization RNA-Seq
Protein in cells or cell extracts	High	Low	Immunocytochemistry Fluorescence microscopy
	Low	Low	Immunoblotting (western blotting)
	Low	High	Mass spectrometry of separate peptides

* Number of genes or proteins studied at one time. RT-PCR, reverse transcriptase polymerase reaction; qPCR, quantitative polymerase reaction.

Low-throughput expression screening follows the expression of only one or a very small number of genes at a time. High-throughput methods can simultaneously track the expression of very many, often thousands of genes at a time, and can offer whole-genome expression screening.

Quantifying the expression of individual genes using real-time (quantitative) PCR

Various PCR-based methods can track gene expression in cell types or tissues that are not easy to access in great quantity. To do that, the starting RNA is first copied into a complementary DNA sequence. In **reverse transcriptase-PCR (RT-PCR)**, a cDNA copy is made of RNA using an oligo(dT) primer or random sequence oligonucleotide primers, and is then used to initiate a PCR. After the PCR is completed, an aliquot is submitted to standard gel electrophoresis. The size-fractionated DNA is exposed to a DNA-binding chemical such as ethidium bromide that allows the DNA to be visualized under ultraviolet light.

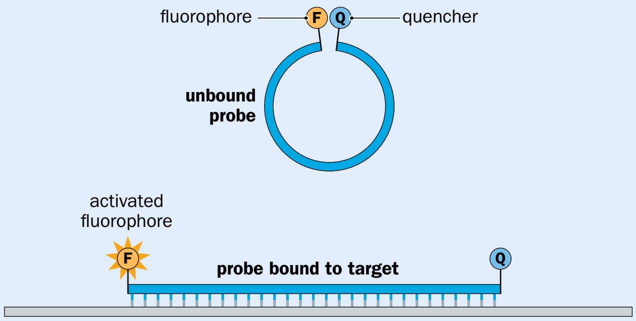
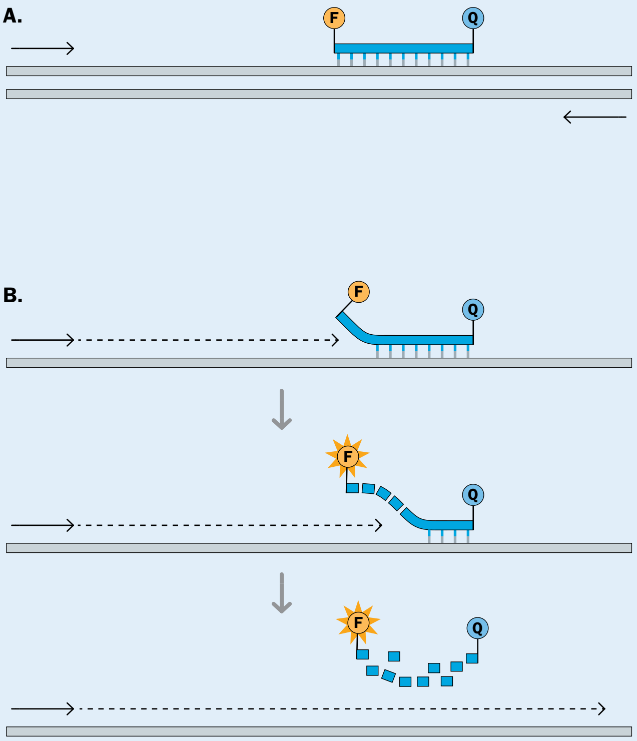
The basic RT-PCR method has been useful for identifying and studying different RNA isoforms of RNA transcript but is not suited to quantification (detection of DNA by ethidium bromide fluorescence is not very sensitive; amplicons have gone through a full 25–30 PCR cycles but, by this stage, the exponential stage of amplification has long since passed—see **Figure 6.9** for the stages of a PCR assay). To obtain accurate quantification of RNA, a kinetic reaction, **quantitative PCR (qPCR)**, is used: while the PCR is progressing, the amplification products are continuously quantified in a specialized PCR machine. To get the most accurate data, the measurements are made only during the early stages in the PCR assay, when the amplification is still exponential.

In order to detect amplification products during a qPCR assay, some measurable signal must be generated that is proportional to the amount of the amplified product. Current detection methods use fluorescence technologies, and the detection method may be nonspecific or specific (see **Box 7.5**).

BOX 7.5 DETECTION METHODS USED IN QUANTITATIVE (REAL-TIME) PCR

Quantification in qPCR depends on detecting a fluorescent signal that is generated after binding of some reagent to amplified product at the early stages of the PCR. Nonspecific detection methods, such as using the SYBR Green I dye, are limited to detecting only one type of target amplicon at a time. Specific detection methods, however, can distinguish between different targets and allow multiplex assays.

Specific detection is possible using single-stranded hybridization probes designed to bind to a specific type of amplicon during the annealing stage of the qPCR cycle. Popular specific detection methods often use a hybridization probe that carries a **fluorophore** at the 5′ end and a *quencher* group at the 3′ end that absorbs photons emitted by the fluorophore and then dissipates the energy absorbed either in the form of heat, or light of a different wavelength.

Method	Basis of method
SYBR Green I dye	When free in solution this dye displays relatively little fluorescence, but when it binds to double-stranded DNA its fluorescence increases by >1000-fold. It is, however, nonspecific in its binding of double-stranded DNA and so can bind to primer dimers or nonspecific amplification products (which is why it is not used in standard PCR). To control for this, a melting curve is carried out at the end of the run by increasing the temperature slowly from 60°C to 95°C while continuously monitoring the fluorescence. At a certain temperature, the whole amplified product will fully dissociate, resulting in a reduction in fluorescence as the dye dissociates from the DNA. The temperature of dissociation is dependent on the length and composition of the amplicon, allowing different DNA fragments to be distinguished (should there be nonspecific amplification or significant primer dimers)
Molecular Beacon probes	Molecular Beacon probes are designed to have a stem-loop structure that brings a fluorophore at the 5′ end in very close proximity to a quencher at the 3′ end, severely limiting the fluorescence emitted by the fluorophore. However, in the presence of a complementary target sequence, the probe unfolds and hybridizes to the target, causing the fluorophore to be displaced from the quencher (see **Figure 1**). As a result, the quencher no longer absorbs the photons emitted by the fluorophore and the probe starts to fluoresce  fluorophore F Q quencher unbound probe activated fluorophore F probe bound to target Q **Box 7.5 Figure 1 Molecular Beacon probes fluoresce only after they hybridize to a target DNA.**
TaqMan^{TM} double-dye probes	Here the probe is an oligonucleotide that has a fluorophore (F) such as FAM^{TM} at the 5′ end and a quencher (Q) group, such as TAMRA^{TM} or the Black Hole Quencher^{TM}, at the 3′ end. In this case the quenching mechanism is based on fluorescence resonance energy transfer (FRET) and can occur over a relatively long distance, as much as 10 nm or more. As above, the oligonucleotide probe can bind to a complementary sequence in the qPCR amplicon during an annealing step in the PCR cycle, but in this case the fluorophore of the intact hybridized probe does not fluoresce because it continues to pass on its energy to the quencher (see **Figure 2A**). After the probe has bound to an amplicon target, however, the advancing Taq polymerase displaces the 5′ end of the probe, which is then degraded by the 5′ → 3′ exonuclease activity of the Taq polymerase. While the polymerase continues to push aside the rest of the probe, cleavage continues and results in release of the fluorophore and quencher into solution (see **Figure 2B**). Now physically separated from each other, the fluorophore FAM exhibits strong fluorescence while the quencher emits much less energy (at a different wavelength to that of the fluorophore)  **Box 7.5 Figure 2 Using TaqMan double-dye probes.** (**A**) Hybridization of a TaqMan oligonucleotide probe containing fluorophore (F) and quencher (Q) dyes to a PCR amplicon. (**B**) The hybridized TaqMan probe is displaced by Taq polymerase and degraded by the associated exonuclease, thereby activating the probe's fluorophore.

High-resolution expression mapping by *in situ* hybridization and immunocytochemistry

High-resolution spatial expression patterns of RNA in tissues and groups of cells are normally obtained by tissue *in situ* hybridization or whole-embryo *in situ* hybridization (embryonic tissues are often used because their miniature size allows screening of many tissues in a single section). In tissue *in situ* hybridization, tissues are frozen or embedded in wax, then sliced using a microtome to give very thin sections (often 5 µm or less) that are mounted on a microscope slide. Hybridization of a suitable gene-specific probe to the tissue on the slide can then give detailed expression images representative of the distribution of the RNA in the tissue of origin (**Figure 7.10**).

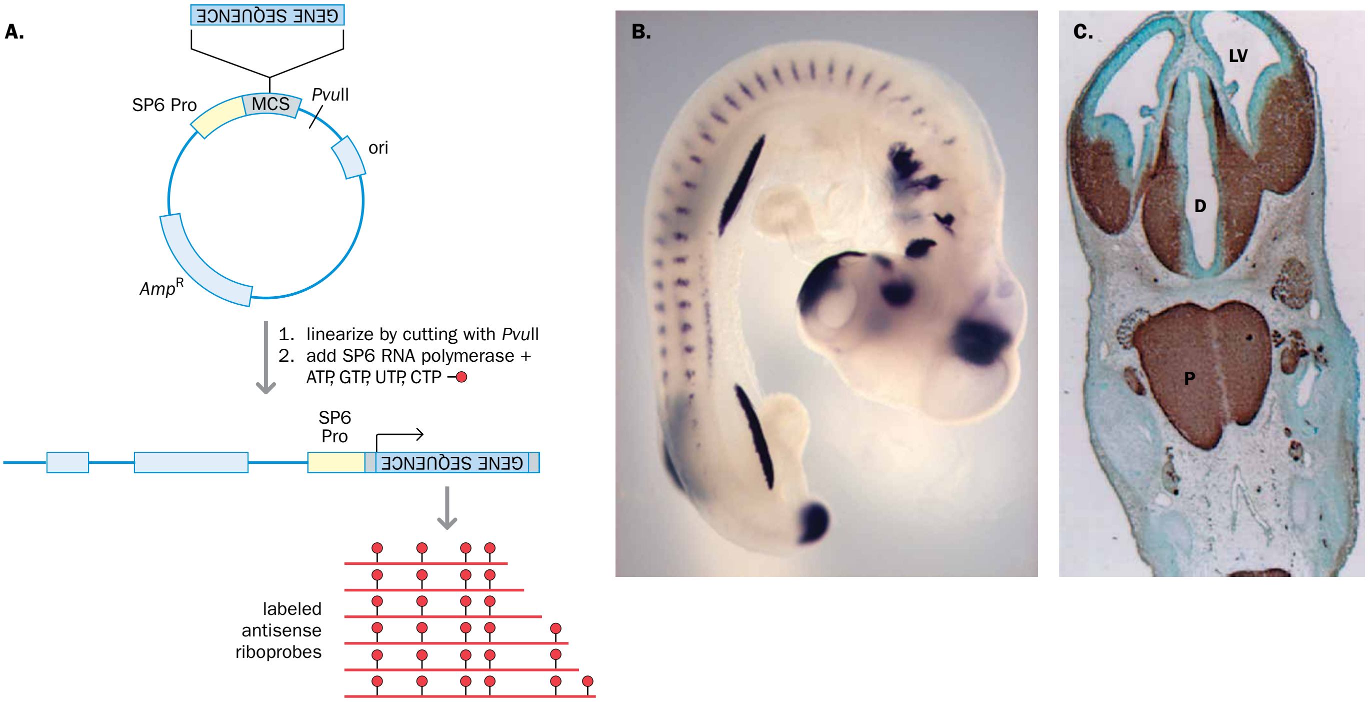

A.
GENE SEQUENCE
SP6 Pro
MCS
PvuII
ori
AmpR
1. linearize by cutting with PvuII
2. add SP6 RNA polymerase + ATP, GTP, UTP, CTP
SP6 Pro
GENE SEQUENCE
labeled antisense riboprobes
B.
C.
LV
D
P

Figure 7.10 High-resolution gene expression studies using *in situ* hybridization and immunocytochemistry. (**A**) Principle of generating antisense riboprobes. A coding DNA sequence from a gene of interest is cloned in reverse orientation into a multiple cloning site (MCS) of a plasmid expression vector that has the powerful SP6 phage promoter (SP6 Pro) adjacent to the MCS. The phage RNA polymerase is used to transcribe the inverted gene sequence in the presence of ribonucleotide precursors, one of which has a labeled group (filled red circle) that gets incorporated into the antisense RNA transcripts. (**B**) *In situ* hybridization. The example shows expression of the *Fgf8* fibroblast growth factor gene in a chick embryo at Hamburger Hamilton stage 20 (about 3 days of embryonic development after egg laying). Transcripts were labeled with an antisense digoxigenin-labeled *Fgf8* probe and were detected with antidigoxigenin antibodies coupled to alkaline phosphatase. The alkaline phosphatase assay used a combination of BCIP (5-bromo-4-chloro-3-indolyl-phosphate) and NBT (nitro blue tetrazolium), resulting in deep-blue expression signals. Expression was evident in the developing eye, isthmus, branchial arches, somites, limb buds, and tailbud. (**C**) Immunocytochemistry. In this example, β-tubulin expression was screened in a transverse section of the brain of a 12.5-day mouse embryo. The antibody detection system used identifies β-tubulin expression ultimately as a brown color reaction (based on horseradish peroxidase/3,3′ diaminobenzidine). The underlying histology was revealed by counterstaining with a toluidine blue stain. LV, lateral ventricle; D, diencephalon; P, pons. (B, image kindly provided by Dr Terence Gordon Smith, Newcastle University; C, courtesy of Steve Lisgo, Newcastle University.)

Using suitably labeled probes, specific RNA sequences can also be tracked within single cells to identify sites of RNA processing, transport, and cytoplasmic localization. By using quantitative fluorescence *in situ* hybridization (FISH) and digital imaging microscopy, it has even been possible to visualize single RNA transcripts *in situ*. A further refinement uses combinations of different types of oligonucleotide probe labeled with spectrally distinct fluorophores. This has allowed transcripts from multiple genes to be tracked simultaneously.

Immunocytochemistry

Because of their exquisite diversity, selectivity, and sensitivity in detecting proteins, antibodies are ideally placed to track gene expression at the protein level. Specific antibodies are prepared, labeled in some way, and then allowed to bind to proteins in cells and tissues, after which the bound label is detected.

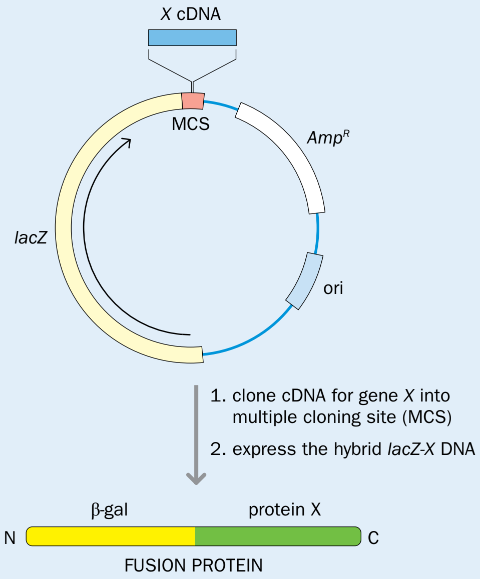
The traditional way to obtain an antibody is to repeatedly inject a suitable animal with a specific **immunogen**, a molecule that is detected as being foreign by the host immune system. To raise an antibody against a specific protein, the immunogen used was often a synthetic peptide in the past, but artificial *fusion proteins* have become popular alternatives (see **Box 7.6**).

BOX 7.6 OBTAINING ANTIBODIES

The major way of obtaining antibodies has been the traditional route of repeatedly injecting animals (rodents, rabbits, goats, and so on) with a suitable immunogen that represents the molecule (usually a protein) under investigation. (An immunogen is a molecule that must be able to provoke an immune response: as well as being viewed as a foreign molecule by the immune system of the injected animal, the immunogen needs to be of high molecular weight and to be chemically complex.)

For convenience, many immunogens are synthetic peptides (representing a favorable part of the protein sequence, often 20–50 amino acids long) conjugated to a carrier protein (which helps maximize the immunogenicity; keyhole limpet hemocyanin is particularly effective). The hope is that the peptide adopts a conformation resembling that of the native polypeptide sequence. Success is not guaranteed, and several different peptides may need to be designed.

An alternative type of immunogen is a **fusion protein** produced by expression cloning in suitable cells, often bacterial cells. The fusion protein has most, or all, of the target protein sequence joined to another protein that confers some advantages, notably in assisting purification of the protein. A cDNA for the desired target protein is cloned into a plasmid expression vector, downstream of a gene that can be expressed in the host cells, in a way such that hybrid transcripts are produced and translated (**Figure 1**). Because the fusion protein contains all or most of the desired polypeptide sequence, the probability of raising specific antibodies may be reasonably high.


X cDNA
MCS
Amp^R
lacZ
ori
1. clone cDNA for gene *X* into multiple cloning site (MCS)
2. express the hybrid *lacZ-X* DNA
β-gal
protein X
N
C
FUSION PROTEIN

Box 7.6 Figure 1 Producing a fusion protein in bacterial cells. The plasmid vector here is an expression vector designed to be expressed in bacteria and contains a bacterial gene such as the β-galactosidase (β-gal) gene, *lacZ*, immediately adjacent to the multiple cloning site (MCS). A cDNA for a protein of interest (such as human protein X) is inserted into the MCS so that transcripts produced from the *lacZ* promoter can continue through the *X* cDNA sequence and be translated in-frame to produce a fusion protein in the bacterial cells.

POLYCLONAL AND MONOCLONAL ANTIBODIES

If the animal's immune system has responded, specific antibodies should be secreted into the serum. The antibody-rich serum (antiserum) contains a heterogeneous mixture of antibodies, each produced by a different B lymphocyte. The different antibodies recognize different specific components (**epitopes**) of the immunogen molecule and a heterogeneous collection of antibodies like this is referred to as **polyclonal antibodies**.

A homogeneous preparation of antibodies with a defined specificity is often advantageous, and involves propagating a clone of cells originally derived from a single B lymphocyte. B cells have a limited life span in culture, however, and so an immortal cell line is made by fusing antibody-producing cells with cells derived from an immortal B-cell tumor. From the resulting heterogeneous mixture of hybrid cells, those hybrids that have both the ability to make a particular antibody and the ability to multiply indefinitely in culture are selected. Such *hybridomas* are propagated as individual clones, each of which can provide a permanent and stable source of a single type of **monoclonal antibody (mAb)**.

ANTIBODIES PRODUCED BY GENETIC ENGINEERING

After the cloning of human immunoglobulin genes, new antibodies have been developed that originate in part or in total from human gene sequences, as detailed in Section 22.2. They are more applicable for therapeutic purposes than the classical antibodies derived from animals.

Some approaches can bypass the need for hybridoma technology and immunization altogether. Phage display technology (described in **Figure 6.7**) permits the construction of a virtually limitless repertoire of human antibodies with specificities against both foreign and self-antigens. The essence of this approach is that the gene segments encoding antibody heavy- and light-chain variable sequences are cloned and expressed on the surface of a filamentous bacteriophage, and rare phage are selected from a complex population by binding to an antigen of interest.

Antibodies can be labeled in different ways. In direct detection methods, the purified antibody is labeled by attaching a reporter molecule, often a fluorophore or biotin, allowing the labeled antibody to bind directly to the target protein. Alternatively, the target protein is first bound by an unlabeled primary antibody that binds to the target protein, and this antibody is then specifically bound in turn by a suitably labeled secondary molecule that may be a secondary antibody, a specific antibody raised against the primary antibody. Sometimes a general secondary molecule is used, such as protein A, a protein found in the cell wall of *Staphylococcus aureus*. (Protein A happens to bind strongly to a common core region on antibody molecules—in the second and third constant regions of the Fc portion of immunoglobulin heavy chains.)

In immunocytochemistry (also referred to as immunohistochemistry) an antibody is used to obtain an overall expression pattern for a protein within a tissue or other multicellular structure. As in tissue *in situ* hybridization, the tissues are typically either frozen or embedded in wax and then cut into very thin sections with a microtome before being mounted on a slide. A suitably specific antibody is allowed to bind to the protein in the tissue section and can produce expression data that can be related to histological staining of neighboring tissue sections (see **Figure 7.10C**).

7.3 HIGH-THROUGHPUT GENE EXPRESSION ANALYSES

Highly-parallel analyses of gene expression allow simultaneous screening of the expression of many hundreds or thousands of genes, even the full set of different transcripts expressed by a cell (**transcriptome**) or all the different proteins (**proteome**). Unlike the genome, which is extremely stable, the transcriptome and proteome are highly variable between different cell types. And, because gene expression in a cell changes with time, they are dynamic: according to pre-determined developmental programs, genetic and epigenetic controls on gene expression are flicked on or off, and gene expression in cells is also controlled by variable environmental factors (including signals received from neighboring cells).

Analyzing the transcriptomes of complex cells, such as human cells, poses a special challenge. A human cell produces huge numbers of transcripts, alternative splicing and other differential processing events are common, and the transcriptome is dominated by noncoding transcripts whose functions largely remain to be elucidated. Only 1.2% or so of our DNA is coding DNA, and so proteomes are not so complex as transcriptomes. But, of course, there is wide cell-to-cell variation, too, in the amounts of different proteins, and there is the complication of post-translational modifications.

The high-throughput expression analyses may have different rationales. Transcripts may be analyzed for the purpose of gene annotation, for example; the object may be to catalog and physically map all the different types of transcript at each locus, including alternative isoforms. Alternatively, the analysis may be concerned with *comparative gene expression* between two cellular sources. Here, the motivation may be to understand aspects of cell or developmental biology or disease processes. One might want, for example, to compare the expression patterns of different subclasses of the same cell type, study how expression changes at different developmental stages, or compare the expression patterns of normal cells and diseased cells of the same cell type. Instead of whole-genome expression profiling, the analysis may be confined to subsets of transcripts or proteins of interest, such as all miRNAs or all transcripts of genes working in a specific cell signaling pathway (targeted expression studies).

DNA and oligonucleotide microarrays permit rapid global transcript profiling

The targets for transcript profiling are complex RNA populations from cellular sources of interest, often cultured cells, surgically excised tissues or tumors, or isolated portions of such. Typical microarrays use many hundreds or thousands of gene-specific probes. cDNA probes have been used but the modern trend has been to use oligonucleotide probes (see Section 6.3 for a general background). Two popular systems are Affymetrix GeneChip microarrays where oligonucleotides about 25 nucleotides long are synthesized *in situ* on an array, and Illumina microarrays where pre-synthesized oligonucleotides with a gene-specific component ~50 nucleotides long are attached to beads. More recently, high-density arrays of longer oligonucleotides have also become available.

Microarray-based expression analyses are often organized so as to compare two or more highly-related cellular or tissue sources that differ in an informative way. To carry out transcript profiling, the cellular RNA sample is normally reverse transcribed *en masse* to form a representative complex cDNA population. Labeling of the cDNA can be achieved during synthesis (by including a fluorophore-conjugated nucleotide in the reaction mix). Alternatively, a two-step procedure is used. First, unlabeled cDNA is made. Then the cDNA is converted into labeled complementary RNA (cRNA) by incorporating biotinylated nucleotides (the biotin labels will later be detected with fluorophore-conjugated streptavidin).

The labeled target cDNA or cRNA is applied to the array and allowed to hybridize. Each individual feature or spot on the array contains large numbers of copies of the same DNA sequence, and is therefore unlikely to be completely saturated in the hybridization reaction. Under these conditions, the intensity of the hybridizing signal at each feature on the array is proportional to the relative abundance of that particular cDNA or cRNA in the target population, which in turn reflects the abundance of the corresponding mRNA in the original source population. The relative abundance of thousands of different transcripts can therefore be monitored in one experiment. Multiple oligonucleotides are also used to help distinguish between closely related transcripts from individual genes. It is possible to monitor splice variants, for example, and to design oligonucleotides specific for every single known exon.

The huge amount of expression data generated by microarray-based hybridization analyses require careful statistical analyses (see **Box 7.7**). Stringent controls are also required to normalize expression data for cross-experiment variation. One way to avoid such problems is to hybridize cDNA populations labeled with different fluorophores to the same array simultaneously. Under nonsaturating conditions, the signal at each feature will represent the relative abundance of each transcript in the sample. If two samples are used, then the ratio of the signals from each fluorophore provides a direct comparison of expression levels between samples, fully normalized for variations in signal-to-noise ratio even within the array. The array is scanned at two emission wavelengths and a computer is used to combine the images and render them in false color. Usually, one fluorophore is represented as green and the other as red. Features representing differentially expressed genes show up as either green or red, while those representing equivalently expressed genes show up as yellow—see the example of using spotted cDNA arrays in **Figure 7.11A**.

BOX 7.7 ANALYZING MICROARRAY EXPRESSION DATA

Microarray expression data revolutionized biological and biomedical research but data interpretation posed a challenge. The raw expression data from microarray experiments are signal intensities that must be *normalized* (corrected for background effects and interexperimental variation) and checked for errors caused by contaminants and extreme outlying values.

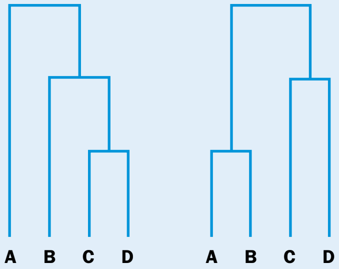
The data are summarized as a table of normalized signal intensities, where rows on the table represent individual genes and the columns represent different conditions under which gene expression has been measured. In the simplest cases, the table has two columns (for example, control and disease samples) and these may represent the signal intensities from two samples hybridized simultaneously to the array. However, there is no theoretical limit to the number of conditions that can be used.

Next, genes with similar expression profiles are grouped. Generally, the more conditions over which gene expression is tested, the more rigorous the analysis. Two types of algorithm are used to mine the gene expression data: one in which similar data are clustered in a hierarchy, and one in which the clusters are defined in a nonhierarchical manner.

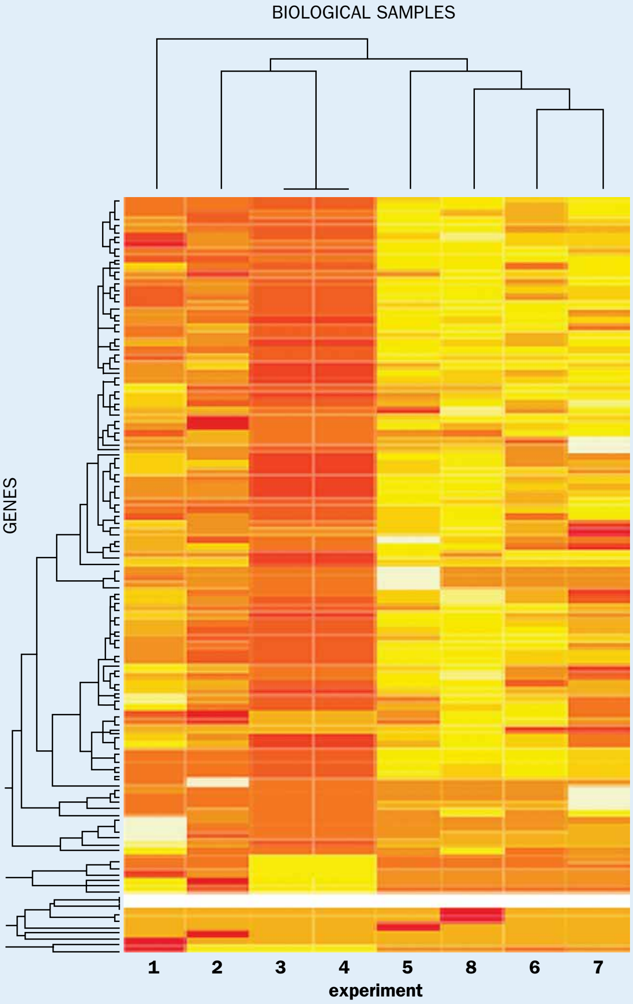
HIERARCHICAL CLUSTERING

The general approach in hierarchical clustering is to establish a distance matrix that lists the differences in expression levels between each pair of features on the array. Those showing the smallest differences, expressed as the distance function *d*, are then clustered in a progressive manner. Agglomerative clustering methods begin with the classification of each gene represented in the array as a singleton cluster (a cluster containing one gene). The distance matrix is searched and the two genes with the most similar expression levels (the smallest distance function) are defined as neighbors, and are then merged into a single cluster. The process is repeated until there is only one cluster left. There are variations on how the expression value of the merged cluster is calculated for the purpose of further comparisons.

In the nearest-neighbor (single linkage) method, the distance is minimized. That is, where two genes *i* and *j* are merged into a single cluster *ij*, the distance between *ij* and the next nearest gene *k* is defined as the lower of the two values *d*(*i*,*k*) and *d*(*j*,*k*). In the average linkage method, the average between *d*(*i*,*k*) and *d*(*j*,*k*) is used. In the farthest-neighbor (complete linkage) method, the distance is maximized. These methods generate dendrograms with different structures (see **Figure 1**). Less frequently, a divisive clustering algorithm may be used in which a single cluster representing all the genes on the array is progressively split into separate clusters. Hierarchical clustering of microarray data is often represented as a **heatmap** (**Figure 2**).



Box 7.7 Figure 1 Microarray data analysis using the hypothetical expression profiles of four genes, A–D. Hierarchical clustering methods produce branching diagrams (dendrograms) in which genes with the most similar expression profiles are grouped together, but alternative clustering methods produce dendrograms with different topologies. The pattern on the left is typical of the topology produced by nearest-neighbor (single linkage) clustering; the pattern on the right is typical of the topology produced by farthest-neighbor (complete linkage) clustering.



Box 7.7 Figure 2 Heatmaps as a tool for visualizing microarray analysis. Heatmaps give a quick overview of clusters of genes that show similar expression values. They consist of small cells, each consisting of a color, which represent relative expression values. Heatmaps are often generated from hierarchical cluster analyses of different biological samples (typically portrayed in columns, as here) and genes (usually in rows and grouped together according to similarity of expression). (Adapted from Allison DB *et al.* [2006] *Nat Rev Genet* **7**:55–65; PMID 16369572. With permission from Springer Nature. Copyright © 2006. See the same reference for alternative ways of visualizing microarray analysis.)

NONHIERARCHICAL CLUSTERING

A disadvantage of hierarchical clustering is that it is time-consuming and resource-hungry. As an alternative, nonhierarchical methods partition the expression data into a certain predefined number of clusters. As a result, the analysis is speeded-up considerably, especially when the dataset is very large. In the k-means clustering method, a number of points known as cluster centers are defined at the beginning of the analysis, and each gene is assigned to the most appropriate cluster center.

Based on the membership of each cluster, the means are recalculated (the cluster centers are repositioned). The analysis is then repeated so that all the genes are assigned to the new cluster centers. This process is reiterated until the membership of the various clusters no longer changes. Self-organizing maps are similar in concept but the algorithm is refined through the use of a neural network.

Expression analyses with microarrays that have short oligonucleotide probes are similar in principle to those where cDNA probes or long (>50 nucleotide) oligonucleotides are used. However, when using oligonucleotides that are only 25 nucleotides long, the hybridization specificity is not so great. There is a higher tendency for probes to hybridize to other sequences in addition to their expected target sequences, and so additional controls are needed. Accordingly, in the case of Affymetrix GeneChip arrays, each gene is represented by 20 or so different oligonucleotide probes that are selected from different regions along the transcribed sequence. In addition to 20 perfect match (PM) oligonucleotides per gene, a corresponding series of 20 mismatch (MM) oligonucleotides is designed to control for nonspecific hybridization by changing a single base in each of the PM sequences (**Figure 7.11B**). To determine the signal for a particular gene, the signals of all 20 PM oligonucleotides are added together and the signals from all 20 MM oligonucleotides are subtracted from the total.

Modern global gene expression profiling predominantly uses sequencing to quantify transcripts

Microarray hybridization has been a robust and reliable method that has been used for decades, but it has disadvantages. First, a significant amount of input RNA is required and so it is not suited to applications where the starting material is limited, notably in single-cell analyses. Second, it has a poor ability to discriminate between different very weak expression signals and between different very strong expression signals. The amounts of different transcripts in a cell can vary over five orders of magnitude but although microarray hybridization can reliably discriminate

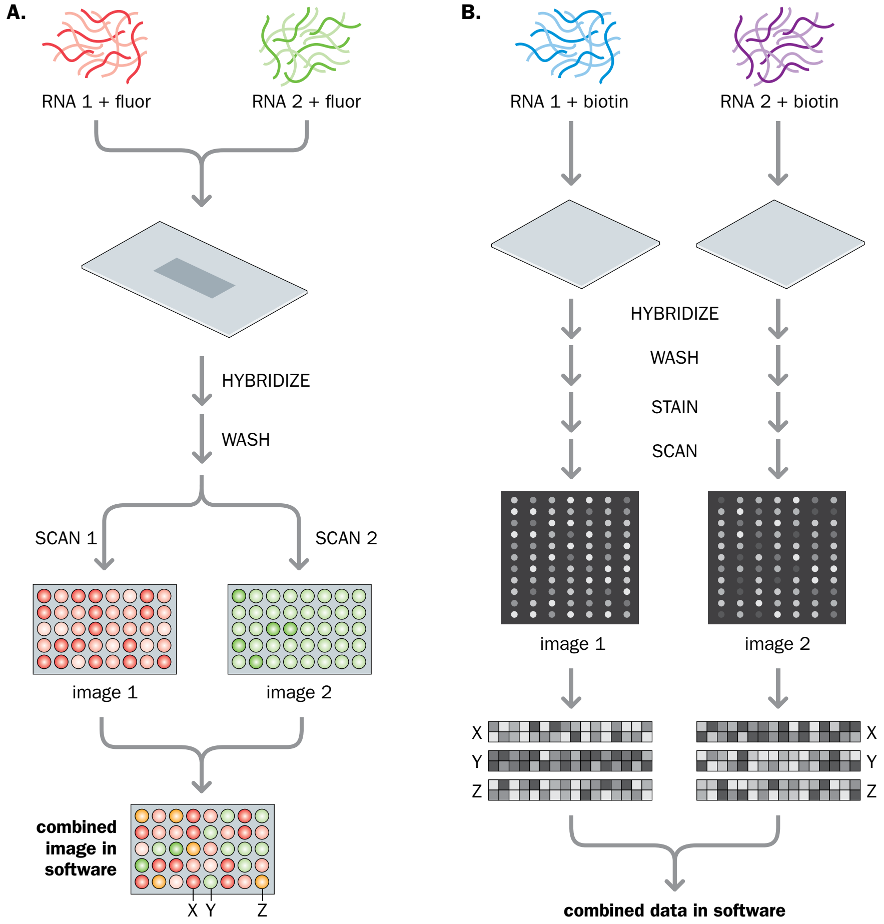


Figure 7.11 Comparative expression analysis with DNA microarrays. (**A**) Using spotted cDNA arrays. Here, comparative expression assays are usually carried out by differentially labeling two RNA or cDNA samples with different fluorophores; the labeled nucleic acids are hybridized to the arrayed cDNAs and then scanned to detect both fluorophores independently. Colored dots labeled X, Y, and Z at the bottom of the image correspond to three hypothetical genes present at increased levels in sample 1 (X, red), increased levels in sample 2 (Y, green), and similar levels in samples 1 and 2 (Z, yellow). (**B**) Using Affymetrix GeneChips. Here RNA is labeled in a two-step process to produce biotinylated cRNA. After hybridization and washing, biotin-cRNA bound to the array is stained by binding a streptavidin-conjugated fluorophore and the bound fluorophore is detected by laser scanning. Each gene is represented by 15–20 different oligonucleotide probe pairs (16 are shown here); one member of each pair is a perfectly matched oligonucleotide probe, the other is a control oligonucleotide with a deliberate mismatch. The example shows expression data for three hypothetical genes, representing genes that are preferentially expressed in sample 1 (X), preferentially expressed in sample 2 (Y), or show equivalent expression in samples 1 and 2 (Z). (Adapted from Harrington CA *et al*. [2000] *Curr Opin Microbiol* **3**:285–291; PMID 10851158. With permission from Elsevier.)

between weak and strong expression, its capacity for quantifying transcription is limited. Microarray hybridization analyses have also been limited by prior knowledge of genes (probes are traditionally designed using the sequences of known genes are so are limited to tracking known genes only).

As an alternative to microarray hybridization, more quantitative, digital profiling is possible by sequencing cDNA copies of transcripts. In the past, methods were devised to retrieve sequence tags from individual cDNAs that could be sequenced quickly, such as the ingenious SAGE (serial analysis of gene expression) method where short sequence tags, up to 25 bp long, from multiple individual cDNAs were concatenated into artificial long constructs that were then sequenced. More recently, high-throughput DNA sequencing has largely supplanted microarray hybridization as the preferred approach for high-throughput transcription profiling.

Transcript profiling by RNA-Seq

For whole-transcriptome profiling, RNA-Seq is the method of choice. An RNA-Seq experiment often involves fragmenting some starting RNA, converting it into cDNA, and then ligating adaptor oligonucleotide sequences to the ends of the fragments; following amplification of the individual sequences using adaptor-specific primers (*adaptor-ligation PCR*), the ends of millions of fragments are sequenced (**Figure 7.12**). The resulting sequence reads can be individually mapped to a reference sequence, the source genome (or the reference transcriptome, when finding novel transcripts is not a high priority).

RNA-Seq is sensitive and offers a way of profiling transcripts of single cells, as described in Section 7.4. It allows quantification of transcripts over five orders of magnitude, and because it does not rely, like microarray hybridization, on prior knowledge of genes, it can be used to identify new transcripts and alternative isoforms, extending genome and gene annotation. Like microarray hybridization, RNA-Seq is often used in differential gene expression analysis, comparing gene expression profiles of different cell sources, and different clustering programs can be used to identify shared groups of

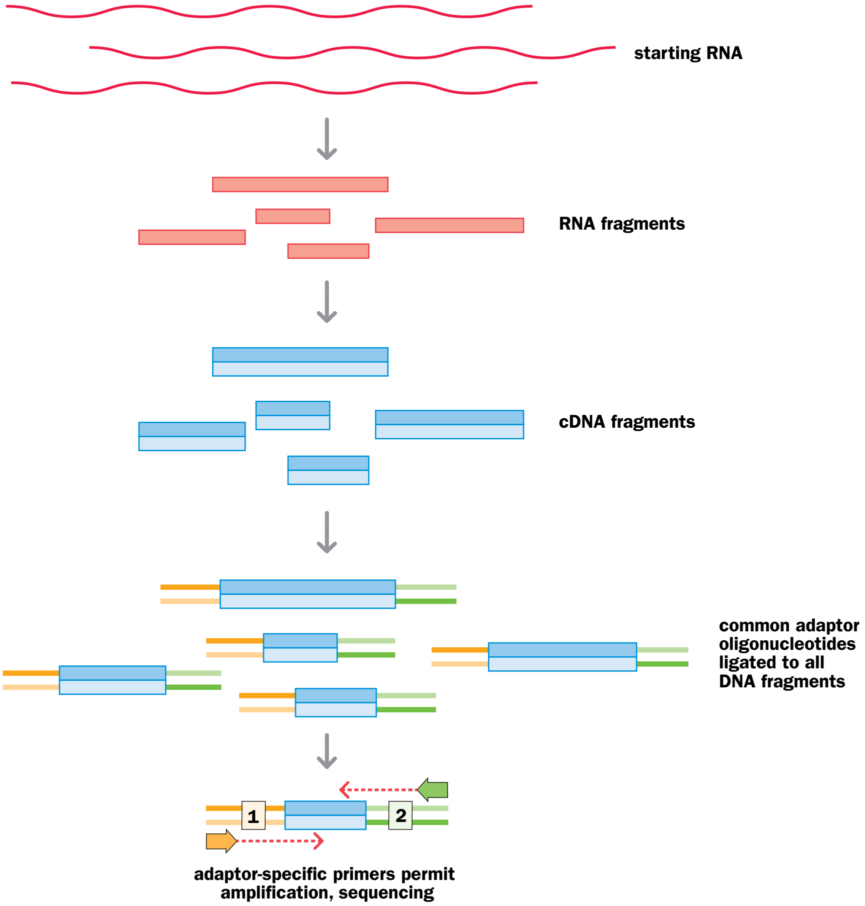

starting RNA
RNA fragments
cDNA fragments
common adaptor oligonucleotides ligated to all DNA fragments
1
2
adaptor-specific primers permit amplification, sequencing

Figure 7.12 RNA-Seq. The method begins by converting input RNA into cDNA fragments, and alternative methods are available. The RNA can first be copied, using a reverse transcriptase and an oligo(dT) or random hexamer primer, and the original RNA strand then destroyed with the enzyme RNaseH, leaving a single DNA strand. Thereafter, the single DNA strand is copied using a DNA polymerase to make a second DNA strand, and the resulting double-stranded DNA is fragmented. Alternatively, it has become common to fragment the input RNA first (typically by RNA hydrolysis or nebulization). The RNA fragments are converted into single-stranded DNA using a reverse transcriptase, and the resulting single-stranded DNA is converted to double-stranded cDNA. Therafter, adaptor oligonucleotides are ligated to the ends of the cDNA fragments. Primers specific for the adaptors can allow amplification of the DNA fragments whose ends can then be sequenced by a high-throughput sequencing method using adaptor-specific primers (adaptor-ligation PCR). In order to retain information on the "strandedness" of the RNA (to distinguish between sense and antisense transcripts), different adaptor oligonucleotides may be attached to the two ends of the fragments (shown here by green and orange coloring); that can be achieved in different ways, such as by using forked (Y-shaped) adaptor oligonucleotides (see **Figure 6.22B**).

genes that show similar expression profiles in the different cell sources, and ones that show dramatically different profiles.

Because of possible **amplification bias**—some sequences may amplify more readily than others—different controls are used. It is common to have external controls: to the input RNA is added a collection of known, external "spike-in" RNAs whose concentrations have been pre-determined. Because long transcripts are fragmented into many smaller pieces, there are often additional internal controls (a single transcript can be represented by several different starting sequences, which are then amplified; consistency in the estimated number of sequence reads ultimately copied from the same transcript provides reassurance that quantification is accurate). The review by Hrdlickova *et al*. (2017) (PMID 27198714) appraises current practice.

After sequence reads from RNA-Seq have been mapped back to the transcription unit/exon of origin, traditional quantification has used the statistic RPKM (reads per kilobase per million mapped reads). That is, RPKM = C/LN, where C = number of mappable reads on a feature (a transcript or exon); L = length of the feature; and N = total number of mappable reads (in millions).

The RPKM statistic is useful for comparing expression profiles from two sources, but more recently, **molecular barcoding** has provided absolute quantification, counting individual RNA molecules. It requires that a primer used for amplification be partially degenerate: the oligonucleotide synthesis is designed so that at a consecutive number of internal nucleotide positions, each of the four nucleotides is made available for synthesis. In that case, the relevant primer is not a single sequence but a heterogeneous collection of very many related sequences. According to which of the many alternative primer sequences is used, cDNA copies of transcripts receive one specific "barcode" sequence that enables direct counting of individual molecules. We show the general principle of molecular barcoding in Figure 7.13, and in Section 7.4 we will describe specific examples of molecular barcoding for whole-transcriptome analysis in single cells.

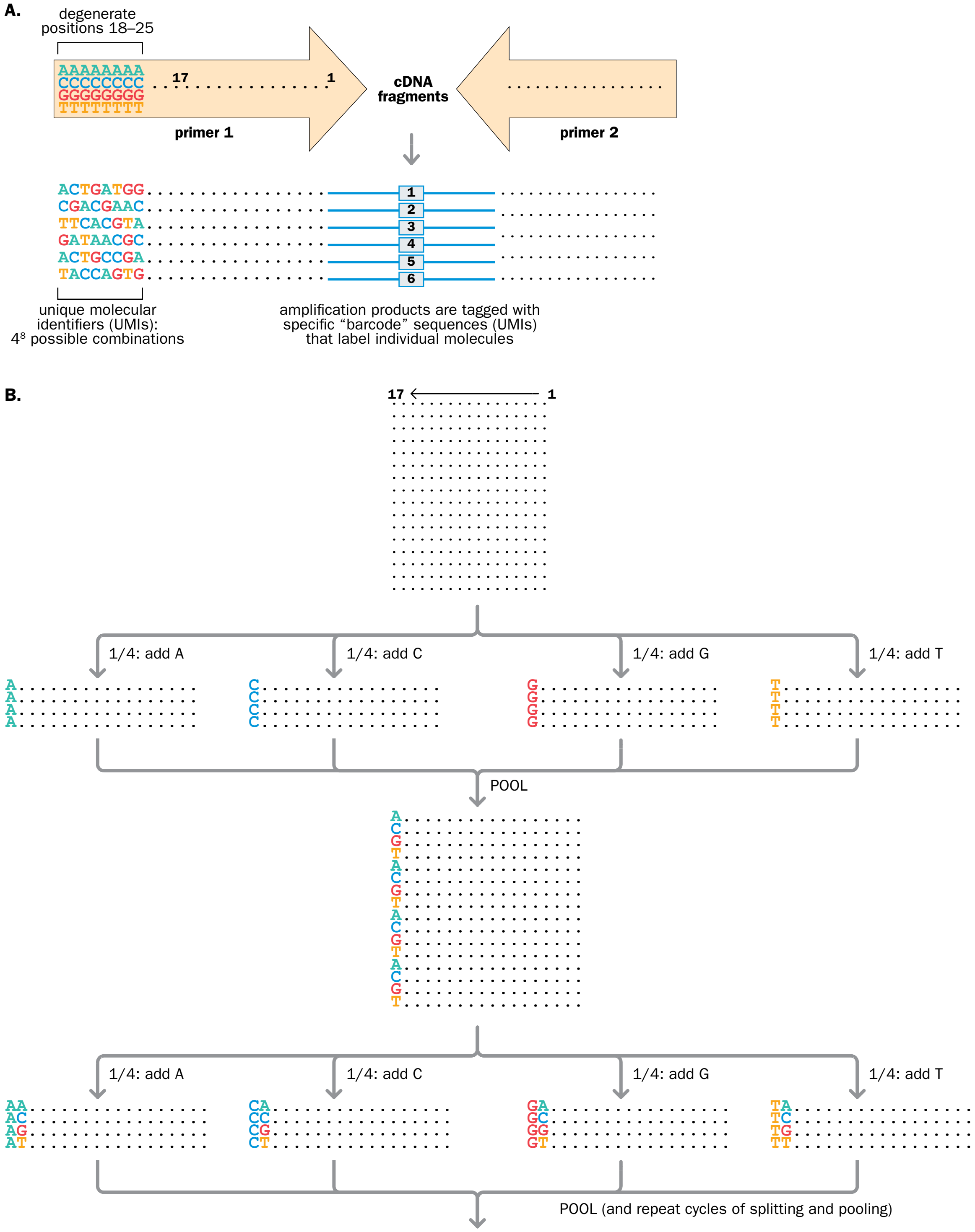


Figure 7.13 Principle of molecular barcoding in RNA-Seq: incorporating random sequence tags from an amplification primer with a partially degenerate sequence. (**A**) Oligonucleotides are synthesized from the 3′ end, and here we imagine standard synthesis for the first 17 nucleotides in the primer on the left. Thereafter, at each position in the eight-nucleotide degenerate region (positions 18–25), all four possible bases are represented in the population of primer molecules (the primer molecules will have positions 1–17 in common but have variable sequences at positions 18–25). Amplification of cDNA sequences using a partially degenerate primer (plus a normal second primer) can allow absolute quantification of the expression of a gene because amplified cDNAs corresponding to individual transcript sequences 1, 2, 3, 4, 5, 6, and so on from the same gene will have incorporated a primer sequence with a randomly selected, variable, eight-nucleotide sequence tag, a "molecular barcode" often described as a unique molecular identifier (UMI). The eight degenerate nucleotide positions in this example would allow a total of 4^8 (65,536) different UMIs. Specific examples of using molecular barcoding are described for whole-transcriptome analysis of single cells in Section 7.4. (**B**) Splitting-and-pooling strategy to create degenerate nucleotide positions in an oligonucleotide. Here, after the seventeenth nucleotide has been incorporated into the growing primer oligonucleotide, the primer population is split into four equal parts that are then exposed to the four different dNTPs: at position 18, one-quarter of the primer population receives an A, one-quarter receives a C, one-quarter receives a G, and one-quarter receives a T. Thereafter, pooling of the four populations creates a mix of primers where the eighteenth nucleotide position is represented by A, C, G, or T. Further splitting-and-pooling contines until the last degenerate position.

Profiling global protein expression using mass spectrometry

Like the transcriptome, the proteome varies widely between different cell types in an organism. Human cells typically contain several thousand proteins differing in abundance over many orders of magnitude. Like nucleic acids, proteins can be detected and identified by specific molecular interactions, in most cases using antibodies or other ligands as probes. However, unlike nucleic acids, there is no procedure for cloning or amplifying rare proteins. Furthermore, the physical and chemical properties of proteins are so diverse that no single, universal methodology analogous to hybridization can be used to study the entire proteome in a single experiment.

As detailed below, proteome profiling essentially involves four steps. First, proteins within the starting protein extract (from cell lysates, tissues, and so on) are fractionated, typically using some form of gel electrophoresis or liquid chromatography. Then, separated proteins are digested with a protease, typically trypsin, to produce a series of peptides. The resulting peptide mixes are analyzed using mass spectrometry, which determines the precise molecular masses (see **Box 7.8**). Finally, the molecular masses

BOX 7.8 MASS SPECTROMETRY (MS) IN PROTEOMICS

THE MASS SPECTROMETER

A mass spectrometer has three components. An ionizer converts the sample to be analyzed (the analyte) into gas-phase ions and accelerates them toward the mass analyzer. The latter separates the ions according to their mass/charge ratio as they travel toward the ion detector, which records the impact of individual ions, presenting these as a mass spectrum.

Traditional mass spectrometry could not be applied to large molecules such as proteins and nucleic acids (which would be broken into random fragments during the ionization process). To overcome this limitation, *soft-ionization* methods were developed, notably the two methods listed below.

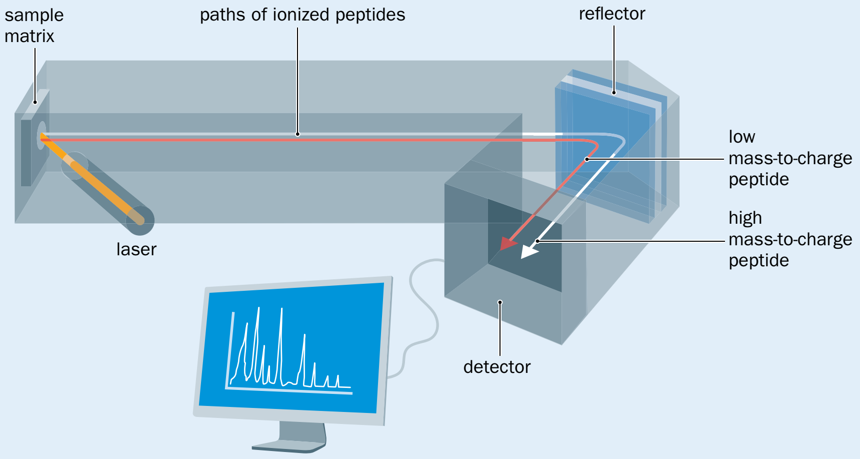
- Matrix-assisted laser desorption/ionization (MALDI) involves mixing the analyte (for example, the tryptic peptides derived from a particular protein sample) with a light-absorbing matrix compound in an organic solvent. Evaporation of the solvent produces analyte/matrix crystals, which are heated by a short pulse of laser energy. The desorption of laser energy as heat causes expansion of the matrix and analyte into the gas phase. The analyte is then ionized and accelerated toward the detector (**Figure 1**).
- In electrospray ionization (ESI), the analyte is dissolved and the solution is pushed through a narrow capillary. A potential difference, applied across the aperture, causes the analyte to emerge as a fine spray of charged particles. The droplets evaporate as the ions enter the mass analyzer.

MASS ANALYZERS

A quadrupole mass analyzer comprises four metal rods, pairs of which are connected electrically and carry opposing voltages that can be controlled by the operator. Mass spectra are obtained by varying the potential difference applied across the ion stream, allowing ions of different mass/charge ratios to be directed toward the detector. A time-of-flight (TOF) analyzer measures the time taken by ions to travel down a flight tube to the detector, a factor that depends on the mass/charge ratio.

TANDEM MASS SPECTROMETRY (MS/MS)

Two or more mass analyzers are in series. Various MS/MS instruments have been described including triple quadrupole and hybrid quadrupole/time-of-flight instruments. The mass analyzers are separated by a collision cell that contains inert gas and causes ions to dissociate into fragments. The first analyzer selects a particular peptide ion and directs it into the collision cell, where it is fragmented. A mass spectrum for the fragments is then obtained by the second analyzer. These two functions may be combined in the case of more sophisticated instruments, such as the ion trap and Fourier transform ion cyclotron analyzers.


sample matrix
paths of ionized peptides
reflector
low mass-to-charge peptide
high mass-to-charge peptide
laser
detector

Box 7.8 Figure 1 Principle of MALDI-TOF mass spectrometry. The analyte (typically a collection of tryptic peptide fragments in proteomic applications) is mixed with a matrix compound and placed near the source of a laser. The laser heats up the analyte/matrix crystals causing the analyte to expand into the gas phase without significant fragmentation. Ions then travel down a flight tube to a reflector, which focuses the ions onto a detector. The time of flight (TOF; the time taken for ions to reach the detector) is dependent on the mass/charge ratio, and allows the mass of each molecule in the analyte to be recorded.

are referenced against known molecular weights of amino acids, and the predicted compositions of all the peptides in a starting sample are then compared against translations of the starting genome sequence to identify coding sequences that could give rise to the predicted peptides.

Proteomes can be analyzed for different purposes, and combining proteome data with data from the corresponding genome can help to elucidate fundamental aspects of our biology, and to permit greater understanding of disease processes that may also be helpful in treating disease (**Figure 7.14**). Of the different approaches used to study proteomes, we are concerned here with *expression proteomics*. That may involve simply collecting quantitative data for different proteins and protein isoforms (which can help in protein and gene annotation). But a powerful additional application is to compare protein expression between related cell samples (differential proteomics). That may be done simply to understand the biology of cells, and will be important in defining subsets of cell types. It can also be done as a way of comparing different cellular states, and here an important application is in dissecting the molecular basis of pathogenesis. We can compare, for example, cells from specific tumor types with the pre-cancerous cell states.

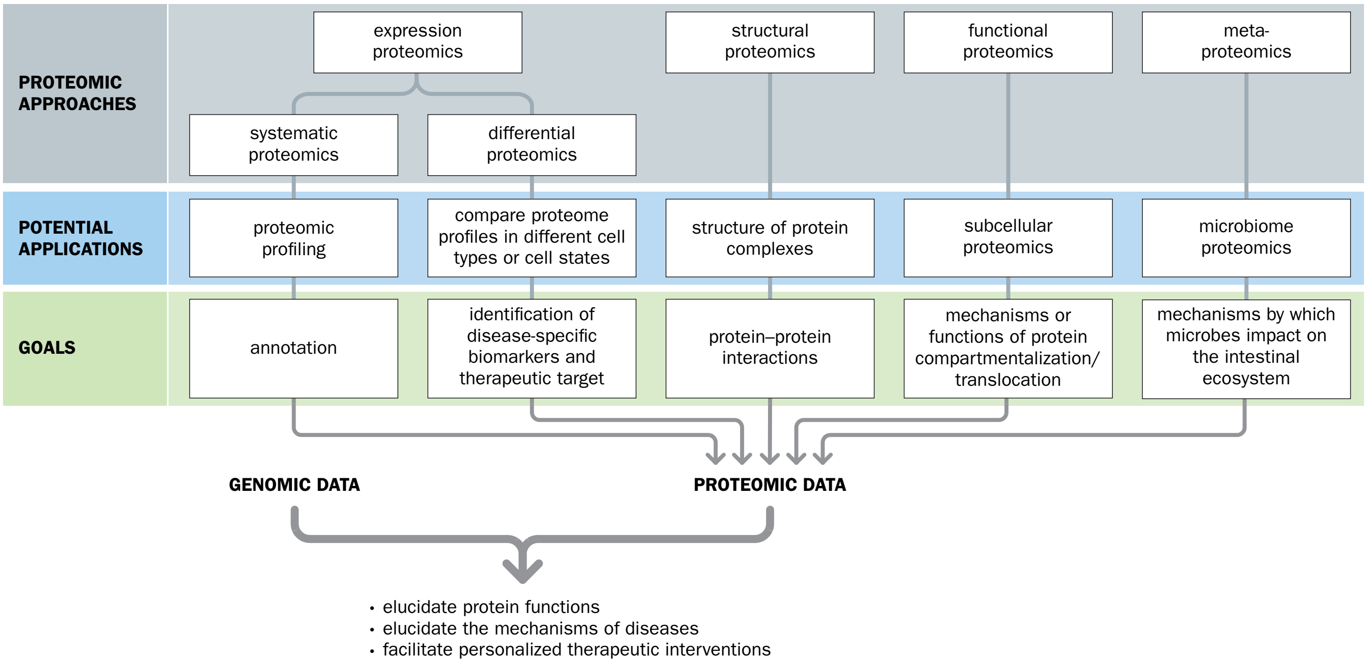

PROTEOMIC APPROACHES
expression proteomics
structural proteomics
functional proteomics
meta-proteomics
systematic proteomics
differential proteomics
POTENTIAL APPLICATIONS
proteomic profiling
compare proteome profiles in different cell types or cell states
structure of protein complexes
subcellular proteomics
microbiome proteomics
GOALS
annotation
identification of disease-specific biomarkers and therapeutic target
protein–protein interactions
mechanisms or functions of protein compartmentalization/ translocation
mechanisms by which microbes impact on the intestinal ecosystem
GENOMIC DATA
PROTEOMIC DATA
• elucidate protein functions
• elucidate the mechanisms of diseases
• facilitate personalized therapeutic interventions

Figure 7.14 Potential applications of proteomics in basic biology and clinical research. In differential proteomics, the proteomes of two different cell types or subtypes are compared, such as two subpopulations of a recognized cell type, or the same cell type under normal and disease states, and so on. (Adapted from Lippolis R & De Angelis [2016] *J Proteomics Bioinform* **9**:63–74. With permission from OMICS International. Published under CC BY license.)

Protein separation

The most widely used methods are gel electrophoresis and liquid chromatography. Two-dimensional (2D) gel electrophoresis uses denaturing polyacrylamide gel electrophoresis (PAGE) for separating proteins. In 2D-PAGE, separation of proteins occurs in the first dimension according to the electrical charge of the protein (isoelectric focusing); thereafter separation occurs according to protein mass in a second dimension, at right angles to the first. 2D-PAGE has the power to resolve up to 10,000 proteins on a single gel, and has been widely used in protein separation. It has its limitations, however. Several classes of protein—strongly basic proteins, membrane proteins, and so on—are underrepresented on standard gels, and the sensitivity is dependent on the detection limit for very scarce proteins (but SYPRO dyes permit detection of protein spots in the nanogram range). Another major limitation of 2D-PAGE is that it is not highly suited for automation, making it difficult to carry out high-throughput analyses of many samples.

The alternative is to use liquid chromatography. High-pressure liquid chromatography can separate a starting protein extract into hundreds of fractions that can then be conveniently processed by easily automated mass spectrometry.

Protein annotation

The masses of peptide fragments can be used to identify the proteins of origin by correlating the experimentally determined masses with those predicted from database sequences, including EST databases and translated nucleotide sequences. Different ways of annotating a protein by mass spectrometry are listed below, and illustrated in **Figure 7.15**.

- Peptide mass fingerprinting (PMF). A simple protein mixture (such as a single spot from a 2D gel) is digested with trypsin. The resulting tryptic peptides are subjected to MALDI-TOF MS (see **Box 7.8**), which returns a set of mass spectra. The spectra are used as a search query against protein sequence databases. The search algorithm carries out virtual trypsin digests of all the proteins in the database and calculates the masses of the predicted tryptic peptides. It then attempts to match these predicted masses against the experimentally determined ones. This method is best suited for simple genomes.
- Fragment ion searching. This method is more suited to the analysis of complex proteomes, and the algorithm can be modified to take into account the masses of known post-translational modifications. The tryptic peptide fragments are analyzed by tandem mass spectrometry (MS/MS; see **Box 7.8**) during which the peptides are broken into random fragments. The mass spectra from these fragments can be used to search against EST databases. Any EST hits can then be used in a BLAST search to identify putative full-length homologs. A dedicated algorithm called MS-BLAST is useful for handling the short sequence signatures obtained from peptide fragment ions.
- De-novo sequencing of peptide ladders. This method is also carried out because it is impossible to account for all variants, either at the sequence level (polymorphisms) or at the protein modification level (for example, complex glycans). Sequencing of peptide ladders may provide sequence signatures that can be used as search queries to identify homologous sequences in the databases. In this technique, the peptide fragments generated by MS/MS are arranged into a nested set differing in length by a single amino acid. By comparing the masses of these fragments to standard tables of amino acids, it is possible to deduce the sequence of the peptide fragment *de novo*, even where a precise sequence match is not available in the database. (In practice, this approach is complicated by the presence of two fragment series, one nested at the N-terminus and one nested at the C-terminus; the two series can, however, be distinguished by attaching diagnostic mass tags to either end of the protein.)

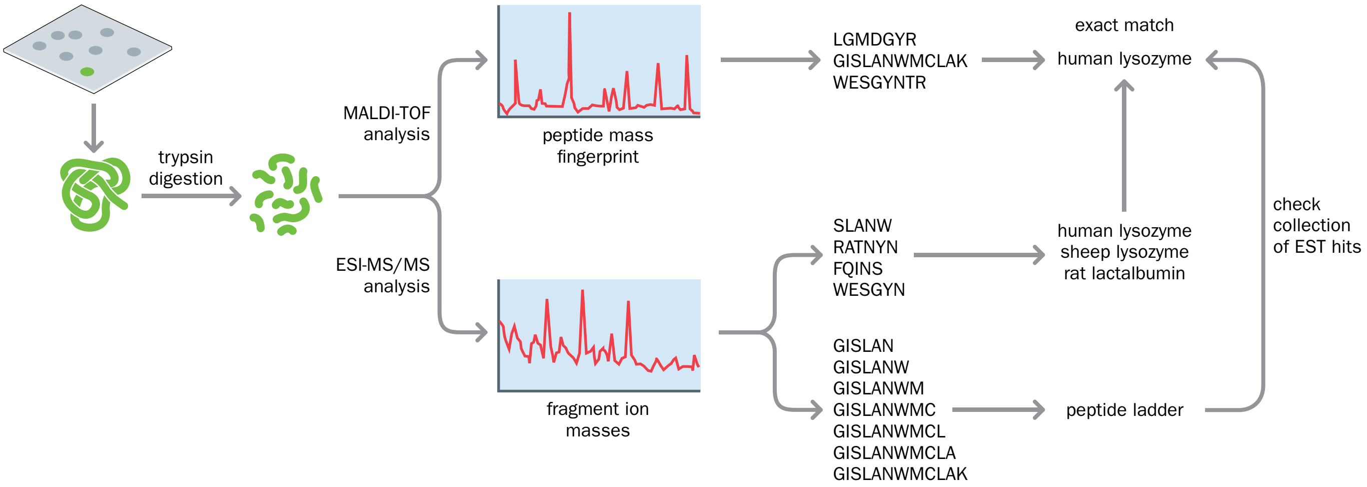
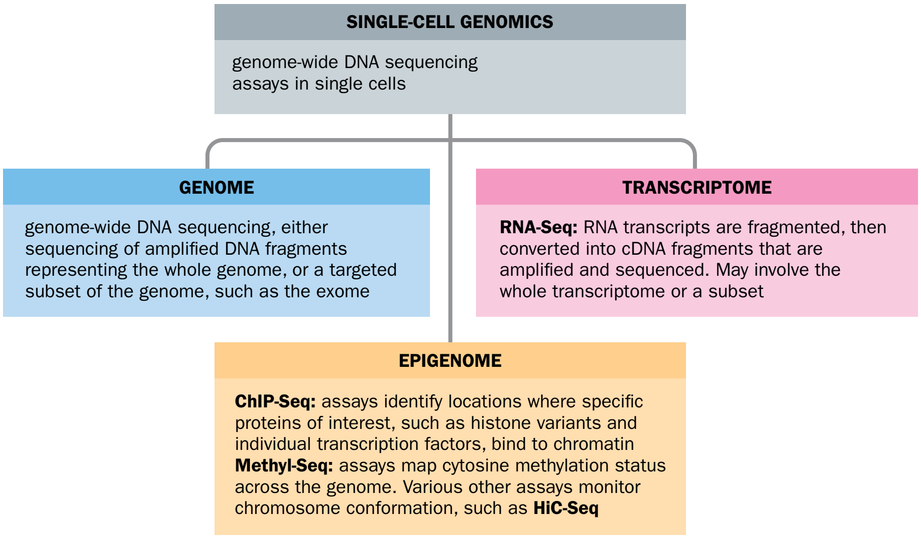


Figure 7.15 Protein annotation by mass spectrometry. Individual protein samples (such as spots from 2D gels) are digested with trypsin, which cleaves on the C-terminal side of lysine (K) or arginine (R) residues (as long as the next residue is not proline). The tryptic peptides can be analyzed as intact molecules by MALDI-TOF (see **Box 7.8**), and the masses used as search queries against protein databases and translated nucleotide databases. Algorithms are used that take protein sequences, cut them with the same cleavage specificity as trypsin, and compare the theoretical masses of these peptides to the experimental masses obtained by MS. Ideally, the masses of several peptides should identify the same parent protein (human lysozyme, in this example). If no hits are recorded (for example, because the protein has been subject to post-translational modification or artifactual modification during the experiment), ESI-tandem mass spectrometry (ESI-MS/MS) can be used to fragment the ions. The fragment ion masses can be used to search sequence databases and obtain partial matches, which may lead eventually to the correct annotation. Alternatively, the masses of peptide ladders can be used to determine protein sequences *de novo*. ESI, electrospray ionization; EST, expressed sequence tag.

7.4 SINGLE-CELL GENOMICS

Improvements in DNA sequencing technology mean that genome-wide DNA sequencing-based analyses can now be carried out using tiny amounts of starting material, and genome-wide DNA sequencing in single human cells and single cells of model organisms is a rapidly advancing field of research. The term **single-cell genomics** is now widely used to cover broad DNA sequencing-based analyses in isolated single cells; it is not limited to analyses of genomic DNA, many of the studies carried out being devoted to analyzing transcriptomes and, to a lesser extent, epigenomes (mostly covering DNA methylation and histone modification states and chromatin conformation). That is, single-cell genomics describes large-scale DNA sequencing-based assays that follow changes in genomic DNA, chromatin, or RNA transcripts, often at a genome-wide level (**Figure 7.16**). Other types of single-cell analysis are also being carried out that use different methods and/or operate on a smaller scale, such as tracking transcripts or proteins expressed by one or often a small number of genes of interest (using fluorescence hybridization with antisense RNA probes or by using specific antibodies).



SINGLE-CELL GENOMICS

genome-wide DNA sequencing assays in single cells

GENOME

genome-wide DNA sequencing, either sequencing of amplified DNA fragments representing the whole genome, or a targeted subset of the genome, such as the exome

TRANSCRIPTOME

RNA-Seq: RNA transcripts are fragmented, then converted into cDNA fragments that are amplified and sequenced. May involve the whole transcriptome or a subset

EPIGENOME

ChIP-Seq: assays identify locations where specific proteins of interest, such as histone variants and individual transcription factors, bind to chromatin **Methyl-Seq:** assays map cytosine methylation status across the genome. Various other assays monitor chromosome conformation, such as **HiC-Seq**

Figure 7.16 Major facets of single-cell genomics.

Given the limiting amounts of starting material (which pushes existing technology to the limits), why should so much effort be expended to apply DNA sequencing methods to study individual human cells? The answer is that traditional cellular analyses have notable limitations because they are typically carried out on cell *populations* (such as bulk tissue samples and cell culture preparations). Because of cell-to-cell variation in phenotype, in gene expression, and even in genomic DNA, the resulting data are necessarily aggregate values: unless single cells are analyzed, all the original intercellular variation is hidden, and important rare cells can be overlooked.

An era has just begun where systems biology will be incrementally applied to understanding the workings of single cells. Cell-to-cell variation has important consequences in health and disease, and we consider first how single-cell analyses will provide exciting new inroads into both basic biology and medical research. We finish by describing some of the technical aspects that have allowed this revolutionary perspective. This is a very fast-moving area, and interested readers are recommended to consult recent reviews.

Understanding cell-to-cell variation: the myriad applications of single-cell genomics in biology

The Human Genome Project paved the way for a biological equivalent of chemistry's periodic table: a periodic table of genes. Admittedly, that table is not a universal one (it varies from organism to organism, but there is a great deal of similarity in the genes of closely related organisms, such as different mammals and vertebrates). There is, however, another fundamental characteristic of living things: cells. What single-cell analyses now offer is the prospect of ultimately establishing a definitive catalog of the cells of multicellular organisms. In addition to the question of defining cell identity, we also briefly outline below how single-cell analyses can offer important insights into other aspects of cell-to-cell variation. Important applications are being found in diverse areas, including neuroscience, development, and also immunology (where, for example, little has previously been known about heterogeneous transcriptional responses in immune cells after activation).

Natural DNA variation in single cells

Though largely stable, the genomic DNA content of our cells does vary as a result of somatic DNA changes. This variation includes programmed changes in the DNA of various normal cell types (and also cancer cells, as described below). In addition, incremental random somatic mutations occur in all cells, so that each cell in our bodies has a unique genome.

Some topical areas of interest include studies of germ-cell DNA (charting aneuploidy, which occurs quite frequently in the gametes of normal humans) and recombination (by comparing the genomes of diploid cells with those of individual gametes from the same individual, crossovers can be mapped across the genome). Certain types of somatic cells are of interest because of a naturally high frequency of large structural changes in their DNA, notably neurons where large deletions are especially frequent, and large-scale duplications are also common. The question of to what extent the changes are related to neuron diversification is being explored. The study of natural DNA variation in human cells also permits, for the first time, detailed tracing of human cell lineages, as described in the next subsection.

Cell lineage tracing

The only complete metazoan cell lineage tree—a cell fate map beginning from the fertilized egg for the nematode *C. elegans*—was reported by John Sulston and colleagues, describing postembryonic lineages in 1976 and embryonic lineages in 1983. This *tour de force* was carried out by time-lapse microscopy, aided by the roundworm's optical transparency, and by the modest size (several hundred cells) and invariant nature of the cell lineage. (Interested readers can find the worm lineages by typing the query "lineage" at http://www.wormatlas.org/ and in the original papers at PMID 838129 and PMID 6684600.) Partial lineage tracing has been carried out in additional model organisms. Often the studies seek to identify progeny of a cell of interest using clonal marking. To do that, the cell of interest is experimentally marked in some way (initially dyes or radiolabeled markers were used; more recently, genetic markers have been introduced into the cell), and descendants of that cell are followed by assaying for the marker.

Experimental clonal marking is not applicable for lineage tracing in humans, but somatic mutation represents a natural way of genetically marking cells: at each cell division, starting from the zygote, new somatic mutations are introduced that are transmitted to progeny (**Figure 7.17**). Sequencing of hypervariable and other highly mutable sites across the genomes of single cells now allows lineage analyses in human cells (see **Table 7.4** for some examples). This new dimension will allow progenitor cells to be defined for multiple cell types where we have limited existing knowledge (see the example of identifying novel candidate stem cells in **Table 7.4**).

Cell identity

Perhaps the most exciting application of single-cell genomics is to clarify our understanding of cell identity. We are likely to have somewhere in the region of 20 to 100 trillion cells, but the division into cell types has been based largely on just anatomical and morphological grounds. About 200 human cell types have been distinguished on this basis, but that number is widely regarded as a gross underestimate (some cell types, notably neurons and T cells, are known to be highly heterogeneous). Not only do we lack a complete catalog of our cells, but there is also limited knowledge of different cell states. (We even lack precise definitions for the intuitive terms "cell type" and "cell state.")

Studying the transcriptomes and epigenomes of single cells holds the promise of fine-scale classification of cells, and identification of novel molecular markers associated with different novel cell subtypes. Rare, functionally important cells can also be identified that are simply not detectable by standard methods designed to study cell populations. While we currently appreciate the more obvious transient cell states, such as when cells transition through different stages of the cell cycle, deep molecular profiling of single cells should also identify more subtle cell states.

Table 7.4 gives some early examples of single-cell analyses that have begun to expand the range of cell types, plus identified previously unappreciated, functionally important rare cells. The greatly enhanced ability to profile single cells at the molecular level has

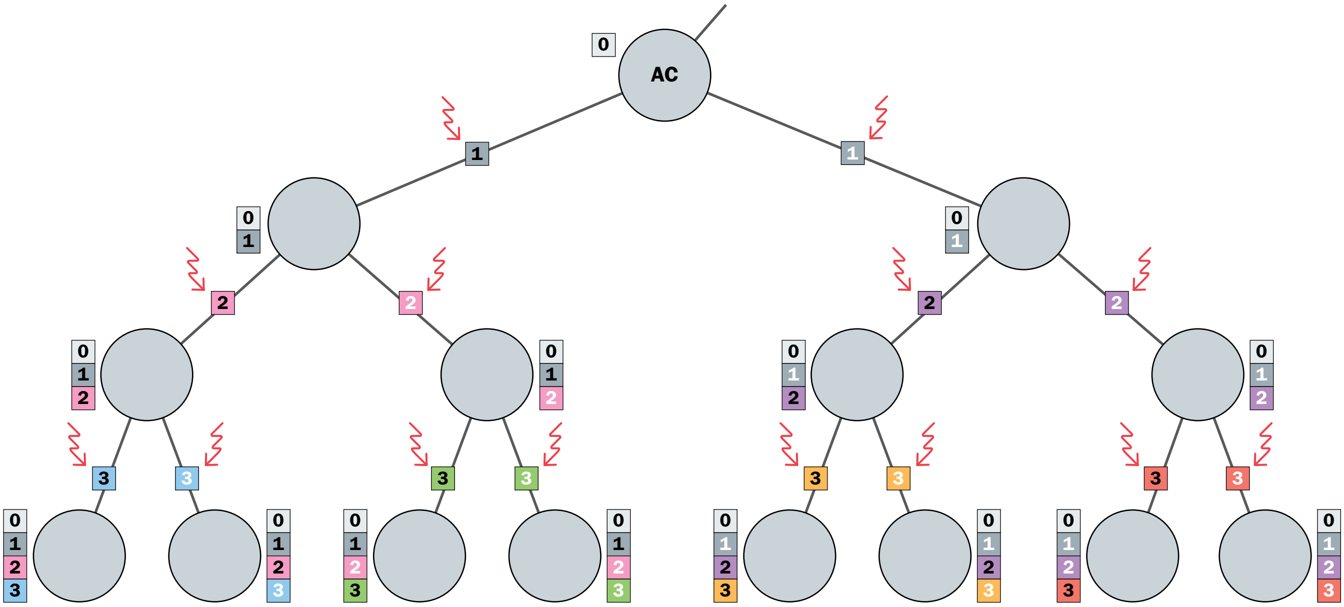


Figure 7.17 Cell lineage tracing using somatic mutations. Each cell in a multicellular organism has a unique genome because, starting from the zygote, many somatic mutations arise *de novo* (shown by red arrows) in a cell's genome prior to each cell division (usually when the DNA replicates), and the two daughter cells inherit different sets of somatic mutations. The figure shows first-, second-, and third-generation descendants of an ancestor cell (AC), and the colored boxes positioned on the lines connecting parent cell to daughter cells indicate new sets of mutations (not present in AC) that have occurred before the first (1), second (2), and third (3) generations. The vertically arranged boxes to the left or right of each descendant indicate acquired somatic mutations not present in AC but accumulated immediately prior to the first, second, and third generations. Thus, for example, the first two descendants of AC have acquired sets of somatic mutations (collectively labeled 1) not present in AC; the black and white numerals indicate that the somatic mutation sets acquired by the two cells are different. The descendants in the second and third generations have acquired additional somatic mutations that were subsequently generated. Although the spectrum of somatic mutations will differ from cell to cell, all eight cells in the third generation will inherit the new somatic mutations acquired by AC (0). By carrying out analyses at hypervariable DNA regions across the genomes of single cells, somatic mutational differences between cells can be identified and analyzed to construct lineage trees.

TABLE 7.4 EXAMPLES OF RESEARCH AREAS WHERE SINGLE-CELL GENOME-WIDE DNA SEQUENCING IS ILLUMINATING UNDERSTANDING OF HUMAN AND MAMMALIAN BIOLOGY

Research area	Example	PMID
Cell identity: subdividing cell types and identifying novel rare cells	Subdividing neurons using single-cell transcriptomics	27565351; 27339989; 27571192
	Identifying novel rare intestinal cells	26287467; 26287456
Cell lineage tracing	Human cortical neuron lineages	26430121
Development	Cell lineage and X-inactivation dynamics in pre-implantation development	27662094
Immunology	Differential responses of individual primary bone-marrow-derived dendritic cells to uniform antigenic stimulus	24919153
Recombination	Comparing the genomes of single germ cells with those of diploid cells from the same individual to map crossover events	22817899
Somatic DNA variation	Mapping structural DNA changes not related to disease in neurons	24179226
Stem cell research	Identifying novel stem cells based on cell lineage tracing	27345837

PMID, PubMed identifier.

recently prompted proposals for an international Human Cell Atlas project (see www.humancellatlas.org and PMID 29206104). The aim is to develop a comprehensive catalog of all human cells based on both their stable properties and transient features, and on cell positions and cell lineages. As well as providing invaluable markers, molecular signatures, and tools for basic research, a Human Cell Atlas should have important clinical applications, as described below.

The prospects of advancing medical research using single-cell genomics

With the exception of pre-implantation diagnosis (where DNA-based diagnosis using single blastomeres from the early *in vitro* fertilization [IVF] embryo has long been available in many countries), single-cell analyses have not been part of medical practice. But the new single-cell genomics technologies offer huge scope for advancing medical research and translational opportunities. One important area is cancer research, where single-cell genomics has been applied since 2011; we outline some applications in the subsection below.

Single-cell genomics offers important new dimensions in various general areas of medical research. One is the cellular basis of disease. Up until recently, cellular analyses of the pathogenic state have been achieved by analyzing mixtures of cells from disease tissues or other heterogeneous cell populations. Inevitably, the picture obtained is clouded by heterogeneity because tissues are made up of different cell types. Cell-to-cell differences in the case of the primary disease cells have not traditionally been explored, and there is no great understanding of the roles of neighboring secondary cells in initiating, promoting, or restraining disease. Analyses of single cells from dissociated tissue or even *in situ* analysis should provide a clearer picture, giving a full description of the expression profiles of individual cells, the dynamic states within each cell type, and the proportion of, and spatial relationships of, the various cell types. Single-cell analyses across many patients can then show how the picture varies during the course of the disease and the responses to treatment.

Other important benefits that might be expected to accrue from single-cell expression profiling include more precisely defined disease markers, expression signatures that identify stages of the disease process and of recovery after treatment, and cell therapy (more precise expression profiling might lead to more accurately defined desired cells to be used in therapy).

Cancer research

Cancers are defined by natural selection acting at the level of the cell to promote abnormal proliferation. During the development of cancerous changes, cancer cells acquire extraordinary genetic and epigenetic changes, and give rise to subpopulations with different cell properties. As a result, tumors are heterogeneous and clonal evolution is important in development of different properties. Rare cancer stem cells may be responsible for regrowth of tumors after cancer treatment. Other initially rare cells in a tumor may give rise to subpopulations of cells that are important in metastasis and in developing resistance to drug treatment. Identifying cell-to-cell variation within (and between) tumors is therefore an especially relevant application of single-cell genomics. We will consider this aspect in Chapter 19.

The technology of DNA-based sequencing assays in single cells

Single-cell genomics involves a succession of methodologies. First, there must be a way of isolating single cells efficiently. Secondly, DNA must be isolated in some way to provide a substrate that can be assayed. It may be done directly—by isolating whole genomic DNA or targeted subsets of genomic DNA from single cells—or indirectly (in transcriptome analysis, RNA is isolated from single cells and converted into cDNA). Because the amount of DNA or RNA isolated from single cells is so tiny, the DNA/cDNA must be amplified to give sufficient material for sequencing. Finally, the end game: high-throughput sequencing and data analysis. We consider some component steps in the subsections below.

Isolating single cells

Different methods have been used to isolate single cells. Manual methods have been used in the past (such as cell capture from tissues by laser microdissection, and manual micromanipulation) but automated methods predominate in single-cell genomics. Some popular methods depend on using labeled antibodies to bind to cell surface proteins characteristic of a cell type of interest; others rely on fluid flow in microchambers—examples are listed below.

- ***Flow cytometry.*** Cells are labeled using fluorescently labeled antibodies, then sorted by the degree of fluorescence they exhibit; because of spectral overlap, however, resolution is limited.
- ***Mass cytometry.*** A cross between flow cytometry and mass spectrometry, the cells are labeled using antibodies conjugated with heavy metal ions; it can discriminate significantly more simultaneous signals than flow cytometry.
- ***Microfluidic cell sorting.*** Cells are suspended in fluid that flows through channels in prefabricated microchambers (where separation can occur according to

inherent physical properties of the cells). Recently developed methods rely on mixing an aqueous solution containing cells with oil to produce an emulsion with very fine droplets that contain single cells (for an example, see the Drop-Seq method for whole-transcriptome analysis, as described below).

Isolating DNA fragments and epigenome analyses

The starting DNA may be fragmented genomic DNA (for genomic analyses) or fragmented RNA that has been converted by reverse transcription to give DNA fragments (for transcriptome analysis). For analysis of the epigenome, different properties—such as the genome-wide locations of methylated cytosines, specific histone variants, and bound transcription factors, and the conformation of chromatin—can each be tracked, ultimately, by a DNA sequencing-based assay. Example methods are described briefly below, and will be detailed in Chapters 9 and 10.

- In the ChIP-Seq method, antibodies specific for DNA-binding proteins of interest, such as specific histone variants and individual types of transcription factor, are used to define DNA binding sites within chromatin of the proteins of interest. That is, they can map all locations across the genome where such a protein of interest is bound. The assay depends on first adding agents that will cause chemical cross-linking of all proteins bound to DNA within cells (proteins bound to chromatin by noncovalent bonds then become *covalently* bound to the DNA). The method is detailed in **Box 9.3**.
- Methyl-Seq involves treating DNA fragments with sodium bisulfite. Nonmethylated cytosines are chemically converted to give uracils, while 5-methylcytosine and hydroxymethylcytosine are unaffected, allowing mapping of methylated cytosines. The method is detailed in **Box 10.3**.
- HiC-Seq and other DNA sequencing-based methods for analyzing chromatin conformation will be described in Section 10.1.

Whole-genome amplification

The total genomic DNA of a human cell is typically less than 10 pg, and current technology requires amplification of the DNA fragments of interest. Three types of method are used for whole-genome amplification, as listed below each with advantages and disadvantages. One problem is *amplification bias*: certain sequences in the starting genome do not amplify very well compared to others. As a result, genome coverage may be comparatively poor (a significant proportion of the genome region is not represented in the amplified DNA fragments), and sometimes specific alleles may be preferentially amplified or not amplified at all (**allele dropout**).

- PCR-based amplification methods. Adaptor-ligation PCR has been used, but more widely used methods use degenerate oligonucleotide primers. That is, instead of having single primers, sets of primers with closely related sequences are used where for each degenerate nucleotide position, some members of the primer set have an A, others have a C, a G, or a T. In degenerate oligonucleotide primer PCR (DOP-PCR) about six or so nucleotide positions, located some distance away from the 3′ end of each primer, are designed to be degenerate. The method is disadvantaged by relatively poor genome coverage, but amplification is otherwise uniform and it is useful for studying copy number variation.
- Multiple displacement amplification (MDA). An isothermal amplification method, it uses random hexanucleotide primers and the φ29 polymerase, which binds very strongly to single-stranded DNA and is highly-processive (it remains bound to DNA over long distances, synthesizing DNA as it goes; other polymerases drop off the DNA more readily). As a result of its highly-processive properties, the φ29 polymerase readily displaces other newly synthesized strands previously formed from primers binding downstream, resulting in branched structures (**Figure 7.18A**) and exponential amplification. With MDA, genome coverage is high, often 80–90% of the genome, but amplification is nonuniform and so it is not well suited to assaying copy number variation.
- Hybrid methods. Multiple annealing and looping-based amplification cycles (MALBAC) begins with a quasilinear MDA-like amplification designed to minimize amplification bias (special primers are used that enable looping of the amplicons to prevent them from being further amplified in subsequent MALBAC cycles). After a number of looping cycles have taken place, PCR amplification is carried out (**Figure 7.18B**). Genome coverage is high and amplification is uniform, but the DNA polymerase used in MALBAC has a relatively high error rate compared to the φ29 polymerase.

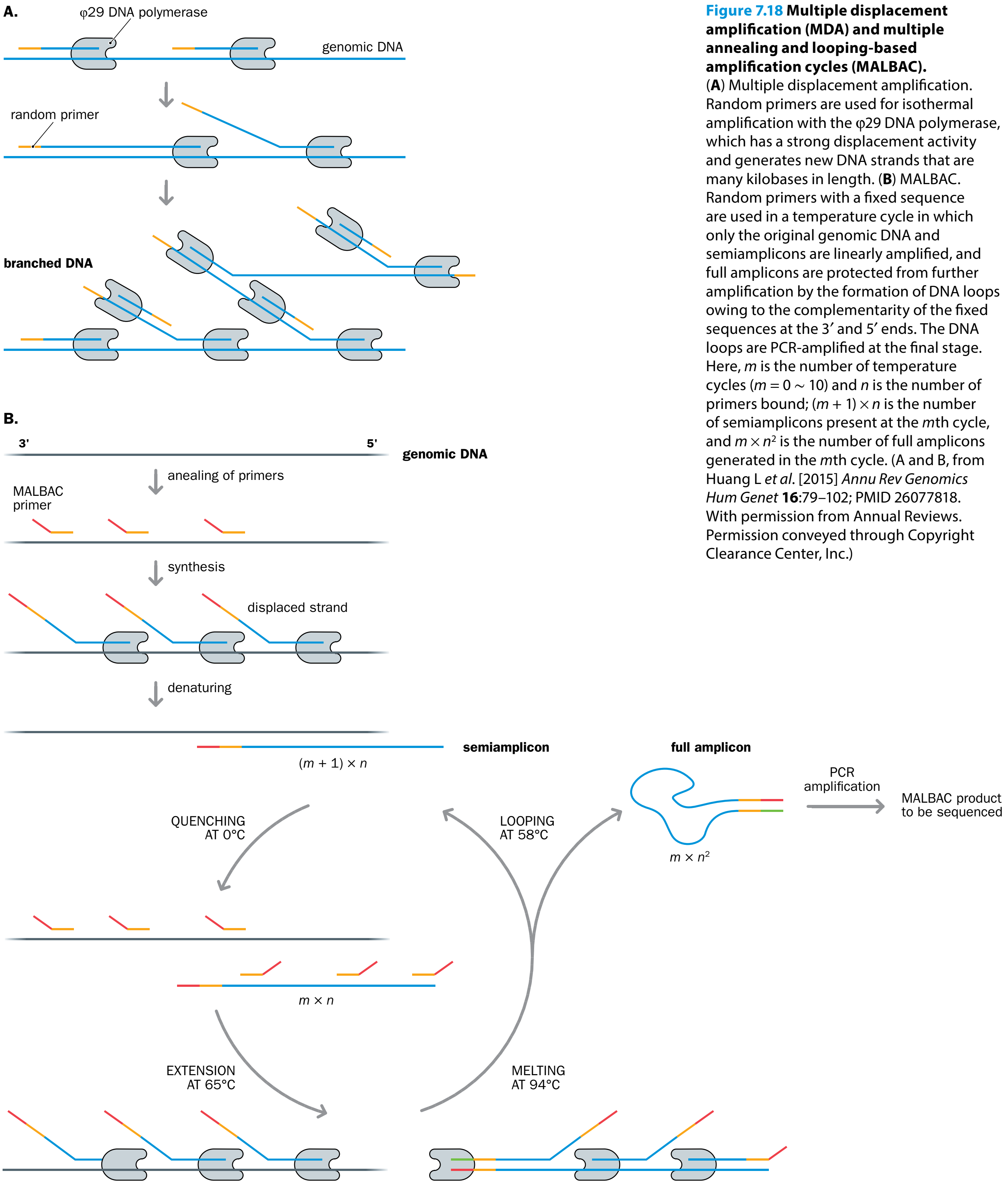


Figure 7.18 Multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC). (**A**) Multiple displacement amplification. Random primers are used for isothermal amplification with the φ29 DNA polymerase, which has a strong displacement activity and generates new DNA strands that are many kilobases in length. (**B**) MALBAC. Random primers with a fixed sequence are used in a temperature cycle in which only the original genomic DNA and semiamplicons are linearly amplified, and full amplicons are protected from further amplification by the formation of DNA loops owing to the complementarity of the fixed sequences at the 3′ and 5′ ends. The DNA loops are PCR-amplified at the final stage. Here, m is the number of temperature cycles ($m = 0 \sim 10$) and n is the number of primers bound; $(m + 1) \times n$ is the number of semiamplicons present at the mth cycle, and $m \times n^2$ is the number of full amplicons generated in the mth cycle. (A and B, from Huang L *et al*. [2015] *Annu Rev Genomics Hum Genet* **16**:79–102; PMID 26077818. With permission from Annual Reviews. Permission conveyed through Copyright Clearance Center, Inc.)

Whole-transcriptome analysis

Until very recently, single-cell transcriptome analysis was hampered because existing methods were not suited to analyzing very large numbers of whole transcriptomes, and required the transcriptomes to be sequenced one after another. In 2015, however, new methods were reported that both sequenced the transcriptomes in parallel and were scalable, allowing highly-parallel whole-transcriptome sequencing.

The new methods involve using primers attached to microbeads and tiny reaction chambers. In one case, a miniature dish with tiny wells is used that contains at most single cells—the volume of the well is just 20 picoliters and the dosage of cells is deliberately kept very low so that most wells do not receive any cells, but those that do can be expected to contain single cells. Other methods rely on microfluidic systems and creating emulsions to trap individual cells in tiny aqueous droplets. An additional feature of the new methods is the use of molecular barcoding systems (see Figure 7.13 for the general principle of molecular barcoding). As an illustration we describe in Figure 7.19 the Drop-Seq

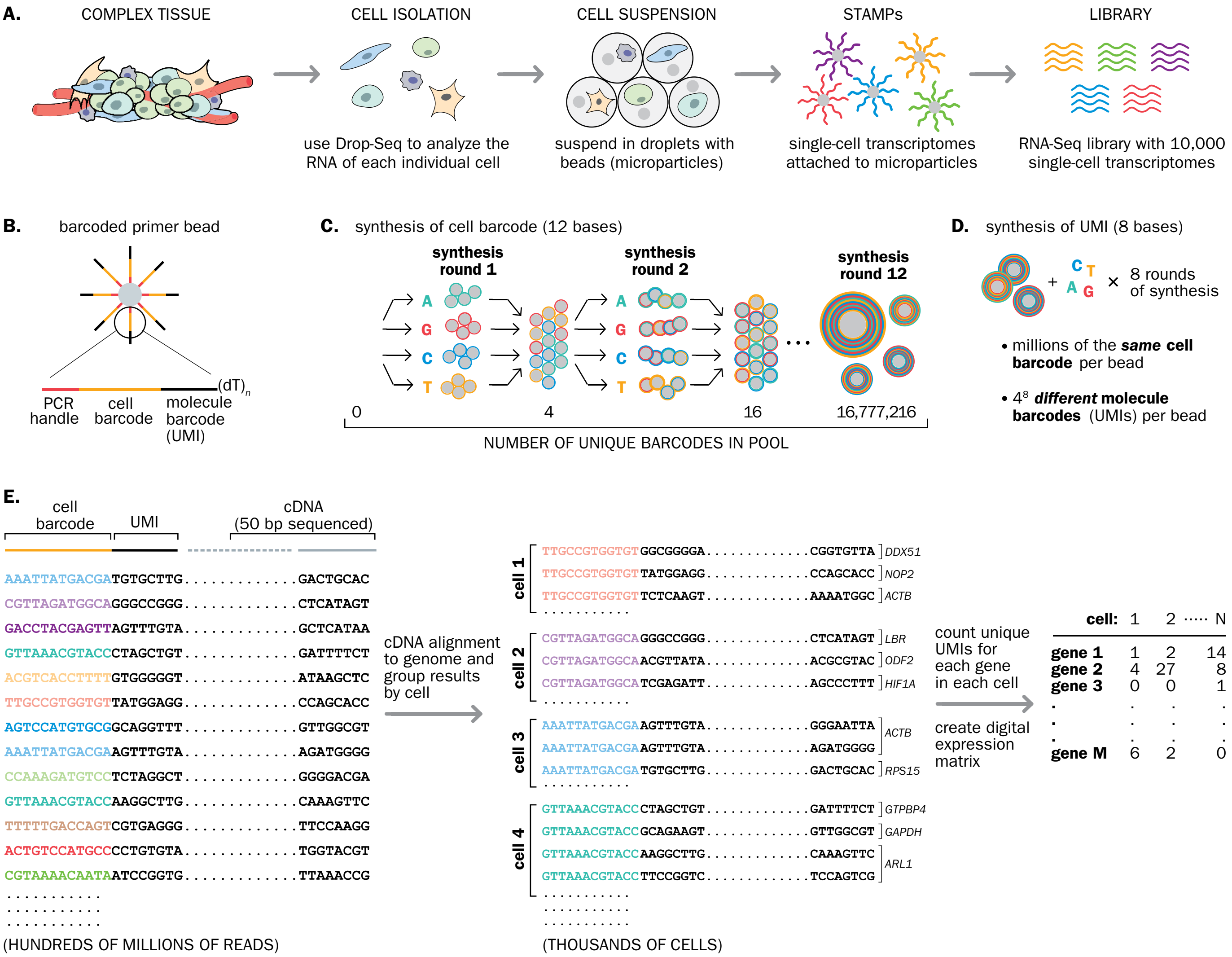

A.
COMPLEX TISSUE
CELL ISOLATION
CELL SUSPENSION
STAMPs
LIBRARY
use Drop-Seq to analyze the RNA of each individual cell
suspend in droplets with beads (microparticles)
single-cell transcriptomes attached to microparticles
RNA-Seq library with 10,000 single-cell transcriptomes
B.
barcoded primer bead
PCR handle
cell barcode
molecule barcode (UMI)
(dT)n
C.
synthesis of cell barcode (12 bases)
synthesis round 1
synthesis round 2
synthesis round 12
A
G
C
T
0
4
16
16,777,216
NUMBER OF UNIQUE BARCODES IN POOL
D.
synthesis of UMI (8 bases)
+ C T A G × 8 rounds of synthesis
• millions of the same cell barcode per bead
• 4^8 different molecule barcodes (UMIs) per bead
E.
cell barcode
UMI
cDNA (50 bp sequenced)
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
GACCTACGAGTTAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
ACGTCACCTTTTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
AGTCCATGTGCGGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CCAAAGATGTCCTCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTTTTGACCAGTCGTGAGGG.............TTCCAAGG
ACTGTCCATGCCCCTGTGTA.............TGGTACGT
CGTAAAACAATAATCCGGTG.............TTAAACCG
(HUNDREDS OF MILLIONS OF READS)
cDNA alignment to genome and group results by cell
cell 1
TTGCCGTGGTGTGGCGGGGA.............CGGTGTTA DDX51
TTGCCGTGGTGTTATGGAGG.............CCAGCACC NOP2
TTGCCGTGGTGTTCTCAAGT.............AAAATGGC ACTB
cell 2
CGTTAGATGGCAGGGCCGGG.............CTCATAGT LBR
CGTTAGATGGCAACGTTATA.............ACGCGTAC ODF2
CGTTAGATGGCATCGAGATT.............AGCCCTTT HIF1A
cell 3
AAATTATGACGAAGTTTGTA.............GGGAATTA ACTB
AAATTATGACGAAGTTTGTA.............AGATGGGG
AAATTATGACGATGTGCTTG.............GACTGCAC RPS15
cell 4
GTTAAACGTACCCTAGCTGT.............GATTTTCT GTPBP4
GTTAAACGTACCGCAGAAGT.............GTTGGCGT GAPDH
GTTAAACGTACCAAGGCTTG.............CAAAGTTC ARL1
GTTAAACGTACCTTCCGGTC.............TCCAGTCG
(THOUSANDS OF CELLS)
count unique UMIs for each gene in each cell
create digital expression matrix
cell: 1 2 ····· N
gene 1 1 2 14
gene 2 4 27 8
gene 3 0 0 1
gene M 6 2 0

Figure 7.19 Drop-Seq: highly-parallel single-cell transcriptome sequencing using droplet reaction chambers and random sequence tags ("barcodes") to indicate cell and molecule of origin for sequence reads. (**A**) Method overview. After dissociation from a tissue, individual cells are encapsulated in droplets along with microparticles (gray circles) that deliver barcoded primers. Each cell is lysed within a droplet; its mRNAs bind to the primers on its companion microparticle. The mRNAs are reverse-transcribed into cDNAs, generating a set of beads called STAMPs (single-cell transcriptomes attached to microparticles). Barcoded STAMPs are amplified in pools for high-throughput mRNA-Seq to analyze any desired number of individual cells. (**B**) Each microparticle contains more than 10^8 individual primers with four types of sequence. The two end sequences are invariant: a "PCR handle" sequence (to allow PCR amplification after STAMP formation) and an oligo(dT) sequence (to capture mRNAs). The two central sequences are variable: a 12-nucleotide cell barcode (common to all primers within one microparticle, but differing between microparticles) and an 8-nucleotide **unique molecular identifier** (UMIs differ from one primer to another within a microparticle, enabling mRNA transcripts to be digitally counted). (**C**) Cell barcode synthesis. The pool of microparticles is repeatedly split into four equally sized oligonucleotide synthesis reactions, to each of which one of the four DNA bases is added, and which are then pooled together after each cycle, in a total of 12 split-pool cycles. The barcode synthesized on any individual bead reflects that bead's unique path through the series of synthesis reactions. The result is a pool of microparticles, each possessing one of 4^{12} (16,777,216) possible 12-nucleotide sequences on its entire complement of primers. (**D**) Molecular barcode synthesis. Following completion of the "split-and-pool" synthesis cycles, all microparticles are together subjected to eight rounds of degenerate synthesis with all four DNA bases available during each cycle, such that each individual primer receives one of 4^8 (65,536) possible 8-nucleotide sequences (UMIs). (**E**) *In silico* reconstruction. Millions of paired-end reads are generated from a Drop-Seq library, representing many thousands of single-cell transcriptomes. The reads are first aligned to a reference genome to identify the gene of origin of the cDNA. Next, reads are organized by their cell barcodes, and individual UMIs are counted for each gene in each cell. The result (shown at bottom extreme right) is a "digital expression matrix" (far right): each column corresponds to a cell, each row corresponds to a gene, and each entry is the integer number of transcripts detected from that gene in that cell. (From Macosko EZ *et al.* [2015] *Cell* **161**:1202–1214; PMID 26000488. With permission from Elsevier.)

method that was used in 2015 to carry out parallel sequencing of the transcriptomes of several tens of thousands of cells. It relies on using molecular barcoding to identify both cell of origin and individual RNA molecules.

SUMMARY

- Framework maps are invaluable assets for sequencing complex genomes for the first time. They are typically constructed using STS (sequence tagged site) markers (DNA sequences, retrieved from genomic and cDNA clones, that must have two key properties: a unique subchromosomal location and the ability to be conveniently assayed, often by PCR).
- For an unexplored genome, most retrieved STS markers are anonymous, but some can be identified to be part of RNA transcripts (having been retrieved from cDNA libraries) and are known as expressed sequence tags (ESTs).
- DNA markers can be mapped to specific chromosomes by labeling larger genomic clones containing their DNA sequence and hybridizing them to preparations of denatured metaphase chromosomes, or by carrying out PCR assays for the marker sequence in panels of somatic cell hybrids containing different chromosomes, or chromosome fragments, for the genome of interest.
- Polymorphic DNA markers are needed to make genetic maps. The different markers are genotyped in individuals from common multigeneration pedigrees to construct maps for each chromosome. The individual genetic maps provide a backbone for building physical maps based on DNA clones.
- Physical maps of chromosomes are based on clone contigs, series of cloned genomic DNA fragments, arranged in the same linear order as the subchromosomal region from which their insert DNAs originated. Overlaps between each DNA fragment allow them to be placed in order, providing important framework maps for genome sequencing.
- Computer-based gene prediction often relies on database searching to identify significant sequence similarity between a test DNA sequence (or derived protein sequence) and a documented gene sequence or protein.
- Transcriptome and proteome describe, respectively, the complete set of RNA transcripts or proteins produced by a cell. Whereas the different nucleated cells of an organism have stable, almost identical genomes, transcriptomes and proteomes are dynamic and each can vary very significantly from one cell type to another, and also from one cell to another of the same type (depending on “cell states” and the interactions between the cells and their environments).
- High-resolution expression analyses are designed to obtain detailed gene expression patterns in cells or tissues, but are limited to analyzing one or a few genes at a time. Nucleic acids are tracked by a hybridization assay; proteins are usually tracked using specific antibodies.
- Highly-parallel expression analyses provide basic gene expression information for many thousands of genes at a time in cell samples. Microarray hybridization used to be the dominant method for highly-parallel analyses of RNA transcripts but has recently been supplanted by RNA sequencing.
- Whole-transcriptome sequencing is aided by molecular barcoding: one of the amplification primers is designed to have a short, highly-variable sequence that will act as a unique molecular identifier. During the amplification step transcripts from each gene acquire unique barcode sequences, allowing direct counting of the number of transcript molecules from individual genes.
- Proteome analyses typically involve fractionating cellular proteins, digesting the separated proteins with trypsin, then using mass spectrometry to determine the precise molecular masses of the resulting peptides, and referencing the data against known molecular weights of amino acids. Predicted peptides are then referenced against known protein sequences and translations of the relevant genome sequence.
- Single-cell genomics means large-scale DNA sequencing-based assays to study properties of the genomic DNA, RNA transcripts, or chromatin of single cells. The motivation is to understand characteristics that depend on cell-to-cell variation but that have not previously been possible to study (when available methods were limited to analyzing cell populations only).
- Single-cell genomics will allow construction of a complete catalog of human cell types, documenting stable cell properties, transient cell features, cell positions, and lineage relationships (leading to identification of novel stem cells).
- Single-cell genomics has been widely employed in cancer research, but more general benefits for medical research will include more thorough understanding of disease processes and greater precision in delivering optimal cell types for cell therapies.

FURTHER READING

Human Genome Project and human genome sequences (see also Box 7.2)

All About The Human Genome Project (HGP). National Human Genome Research Institute. https://www.genome.gov/10001772 (An educational resource maintained by the US National Human Genome Research Institute.)

Human Genome Collection. *Nature* supplement (2006) pp. 1–305. http://www.nature.com/nature/supplements/collections/humangenome/index.html (Contains the 2001 IHGSC draft genome sequence publication and various commentaries, plus papers analyzing the sequences of all 24 human chromosomes.)

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**:931–945; PMID 15496913.

Waterston RH *et al.* (2002) On the sequencing of the human genome. *Proc Natl Acad Sci USA* **99**:3712–3716; PMID 11880605. (Looking back on the draft genome sequences; see also the follow-up publication one year later [PMID 12631699].)

Genome compilations across species

A quick guide to sequenced genomes. Genome News Network. http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_ guide_index.shtml

Mukherjee S *et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and future enhancements. *Nucleic Acid Res* **45** (Database issue):D446–D456; PMID 27794040.

Genome databases and browsers

Hung JH & Weng Z (2016) Visualizing genomic annotations with the UCSC genome browser. *Cold Spring Harb Protoc* 2016:pdb.prot093062; PMID 27574198.

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46** (Database issue):D8–D13; PMID 29140470.

Spudich G *et al.* (2007) Genome browsing with Ensembl: a practical overview. *Brief Funct Genomic Proteomic* **6**:202–219; PMID 17967807.

Wang J *et al.* (2013) A brief introduction to web-based genome browsers. *Brief Bioinform* **14**:131–143; PMID 22764121.

Gene prediction, annotation, and ontology

Brent MR (2007) How does eukaryotic gene prediction work? *Nat Biotechnol* **25**:883–885; PMID 17687368.

Durbin R *et al.* (1998) Biological Sequence Analysis. Cambridge University Press.

Gene Ontology Consortium. http://www.geneontology.org

Huang Y *et al.* (2016) Well-characterized sequence features of eukaryotic genomes and implications for *ab initio* gene prediction. *Comput Struct Biotechnol J* **14**:298–303; PMID 27536341.

Yandell M & Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**:329–342; PMID 22510764.

Genome assembly

Chaisson MJP *et al.* (2015) Genetic variation and the *de novo* assembly of human genomes. *Nat Rev Genet* **16**:627–640; PMID 26442640.

Compeau PE *et al.* (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* **29**, 987–991; PMID 22068540.

Simpson JT & Pop M (2015) The theory and practice of genome sequence assembly. *Annu Rev Genomics Hum Genet* **16**:153–172; PMID 25939056.

Basic gene expression analyses

Real-Time PCR Vs. Traditional PCR. Applied Biosystems. http://www.appliedbiosystems.com/support/ tutorials/pdf/rtpcr_vs_tradpcr.pdf

VanGuilder HD *et al.* (2008) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* **44**:619–626; PMID 18474036.

Ward TH & Lippincott-Schwartz J (2006) The uses of green fluorescent protein in mammalian cells. *Methods Biochem Anal* **47**:305–337; PMID 16335719.

High-throughput gene expression profiling

Allison DB *et al.* (2006) Micorarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**:55–65; PMID 16369572.

Belacel N *et al.* (2006) Clustering methods for microarray gene expression data. *OMICS* **10**:507–531; PMID 17233561.

Hrdlickova R *et al.* (2017) RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA* **8**. doi: 10.1002/wrna.1364; PMID 27198714.

The Chipping Forecast II (2002) *Nat Genet* **32** (Suppl): 465–551.

Wang Z *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**:57–63; PMID 19015660.

High-throughput protein profiling

Altelaar AF *et al.* (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* **14**:35–48; PMID 23207911.

Cox J & Mann M (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem* **80**:273–299; PMID 21548781.

Kolker E *et al.* (2006) Protein identification and expression analysis using mass spectrometry. *Trends Microbiol* **14**:229–235; PMID 16603360.

Lippolis R & De Angelis M (2016) Proteomics and human diseases. *J Proteomics Bioinform* **9**:63–74.

Single-cell genomics: reviews

Gawad C *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**:175–188; PMID 26806412.

Shapiro E *et al.* (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**:618–630; PMID 23897237.

Wagner A *et al.* (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**:1145–1160; PMID 27824854.

Wang Y & Navin NE (2015) Advances and applications of single-cell sequencing technologies. *Mol Cell* **58**:598–609; PMID 26000845.

Single-cell genomics: technologies

Grümlautn D & van Oudenaarden A (2015) Design and analysis of single-cell sequencing experiments. *Cell* **163**:799–810; PMID 26544934.

Huang L *et al.* (2015) Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu Rev Genomics Hum Genet* **16**:79–102; PMID 26077818.

Junker JP & van Oudenaarden A (2015) Single-cell transcriptomics enters the age of mass production. *Mol Cell* **58**:563–564; PMID 26000840.

Kolodziejczyk AA *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**:610–620; PMID 26000846.

Principles of genetic manipulation of mammalian cells

8

Genetic manipulation of animal cells is extremely important for both basic and applied research. From the perspective of human biology, medical research, and clinical applications, genetic manipulation of human cells and other mammalian cells is especially relevant and important, and so the great bulk of this chapter is focused on this aspect. But, on occasions, we describe important applications of genetic manipulation of the cells of other animals.

Genetic manipulation of animal cells often involves transferring genetic material into cultured cells, but sometimes the recipient cells are single cells (such as fertilized oocytes) or cells within a living model organism or person. The genetic material is often a genetically engineered DNA construct, and when transferred into cells, the introduced DNA is known as a **transgene** (even though it may contain multiple genes, or lack any gene). Often, but not always, transgenes are intended to make desired RNA or protein products. Very occasionally, RNA has been transferred into cells (but it is generally preferred to transfer a transgene that is then expressed in the cell to make the desired RNA). In addition, some applications involve transfer of synthetic oligonucleotides into cells.

The aim may be to study the effect on properties of the transfected cells themselves, or to use the transfected cells as a way of introducing genetic material into an experimental organism (transferring genetic material into the germ line to create a transgenic animal), or into a person in an attempt to combat some disease (gene therapy). The motivation may be to understand a basic research question—such as: If I inactivate gene X, what are the phenotypic consequences? Or it may be a question of applied research, a commercial application, or a clinical intervention.

In this chapter we describe the basic principles and technologies for transferring genetic material into intact mammalian cells. We will deal with applications of the methodology in other chapters, notably in Chapter 21, when we consider applications in modeling disease, and in Chapter 22 when we consider gene therapy. Table 8.1 gives a brief, and very selective, list to illustrate some of the applications of genetic manipulation of mammalian cells.

Transporting naked DNA and oligonucleotides into mammalian cells is normally unnatural, and the cell membrane usually poses a formidable barrier to transport of such large, highly charged macromolecules. However, viruses that infect mammalian cells have refined ways of subverting the usual controls regulating transmembrane transport. They are able to efficiently transfer their DNA or RNA genomes into cells, after first surrounding their genomes with a viral protein coat. Viral genomes can be genetically modified—after first converting RNA genomes into complementary DNA, in the case of RNA viruses—and used as vectors to allow the transfer of desired nucleic acid sequences into cells. The alternative is to use nonviral physical or chemical transfer methods. We provide details of the different approaches in Section 8.1.

The introduced genetic material may, or may not, be designed to be incorporated into a cell's genome. If inserted into a chromosomal DNA, an introduced transgene becomes a stable part of the chromosome, and is transmitted to daughter cells after cell division. Alternatively, the genetic material may be intended to have an extrachromosomal location (as is always the case for synthetic oligonucleotides).

According to need, the transfer of genetic material may be intended to produce different effects within the intact cell. Transgenes may have been engineered to produce

TABLE 8.1 SELECTIVE EXAMPLES OF GENETIC MANIPULATION OF MAMMALIAN CELLS

Process	Description/comments
Gene expression for protein purification	A transferred human or animal coding DNA is expressed in large quantities in cells that will have appropriate post-translational modification systems. Mammalian cDNAs can be expressed in well-established mammalian cell lines such as CHO or HEK293 cells. Genetically engineered antibodies are often produced in this way using myeloma cell lines
Directed cell differentiation	Coding DNAs specifying appropriate master transcription factors are expressed in cells of one type causing them to change to a different type. Allows basic studies of cell differentiation and has therapeutic potential (Chapters 21 and 22)
Gene transfer to alleviate genetic deficiency	A coding DNA is expressed to give a product lacking in the recipient cells. Used in some gene therapies where suitable cells of a patient are cultured and then genetically manipulated *in vitro* before being returned to the patient (Chapter 22)
Genome editing to inactivate a specific gene	Often used to investigate gene function or to make animal models of disease (Chapter 21). The genetic deficiency can be extended to the germ line by manipulating the fertilized oocyte of a model organism or cultured pluripotent stem cells that can be implanted in an embryo to give rise to a model organism with the mutation in every nucleated cell
Genome editing to make altered gene product	May be used to investigate gene function or to model disease and has therapeutic potential
Gene silencing to suppress expression of a specific gene	Relies on transferring oligonucleotides (or antisense RNA) to inhibit expression of a gene that makes an RNA containing the same sequence. A fast and simple way to study the cellular function of a gene, it has also been used to model disease (Chapter 21) and in RNA-based therapeutic approaches (Chapter 22)

functional proteins of interest, or noncoding RNAs. We consider the principles of expressing transgenes in mammalian cells in Section 8.2. Other procedures are designed to make specific, fine-scale changes to a pre-determined target sequence within the genome of intact cells or to specifically suppress expression of a pre-determined gene within the cell.

AN OVERVIEW OF GENOME EDITING, GENE SILENCING, AND GERM-LINE TRANSGENESIS

Genome editing involves making desired changes to the sequence of a pre-determined DNA sequence of interest (the *target site*) at some specific location within the genome of an intact cell. (Because the target DNA sequence is usually a gene, genome editing has often been described as *gene targeting*.)

At its most fundamental level, genome editing relies on two components: (1) endonucleases to cut both DNA strands at, or close to, a pre-determined target sequence of interest; and (2) protein complexes that rejoin the broken DNA strands. Following genome editing, screening must be carried out to identify those cells where a desired DNA change has been produced. There must, of course, be some way of ensuring a high degree of sequence specificity so that cleavage is directed to occur at a desired target sequence. And genome editing often (but not always) relies on artificial intervention to direct the desired sequence changes.

Two quite different genome editing approaches have been used (**Figure 8.1**). A traditional method, which has been used in mammalian cells for more than three decades, relies exclusively on homologous recombination using *endogenous* endonucleases that are part of the cell's recombination machinery. The method relies on designing the transgene to have long stretches of DNA that show 100% sequence identity to the target site plus some altered sequence that can be introduced into the target site by recombination. The high degree of sequence homology between the transgene and the target site within the genome promotes recombination between them, in which case endogenous endonucleases cleave the DNA at, or close to, the target site (**Figure 8.1A**). We describe the principle of using standard

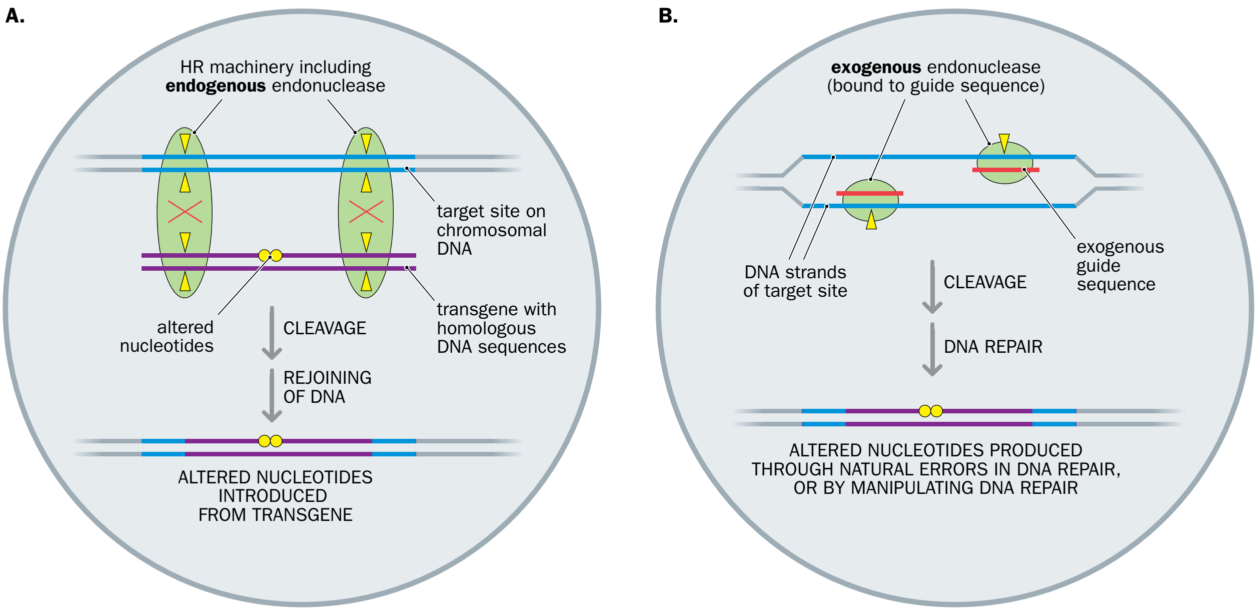


Figure 8.1 Two very different approaches to genome editing. (**A**) Genome editing using homologous recombination (HR) alone. The transgene is designed to have some flanking sequences that are identical to those at a specific target site in the genome but altered sequence (yellow circles) in a central region. Cells are screened to identify rare recombination events between the transgene and target site, notably a double crossover that inserts the altered sequence into the target site (crossover points are indicated by a red X; yellow triangles signify cleavage points). Recombination is carried out by endogenous enzymes, including an endonuclease that cuts double-stranded DNA and a DNA ligase that rejoins the DNA fragments. Details are provided in Section 8.3. (**B**) Genome editing using programmable exogenous nucleases. Introduced transgenes produce endonucleases plus protein or RNA *guide sequences* designed to bind to specific DNA sequences at a pre-determined target site in the genome. The exogenous endonuclease is covalently bound to a protein guide sequence or noncovalently to a hybrid RNA that contains a guide sequence plus binding sites for the endonuclease. A pair of sequence-specific guide sequences is used to target bound endonuclease, to cut both DNA strands at a desired unique site. The resulting double-strand break can be repaired rapidly, in which case errors are introduced that may introduce desired sequence changes (yellow circles), or by manipulating an homologous recombination pathway to introduce the desired changes. Different cells may have different altered sequences and are screened for the presence of a desired sequence alteration. Details are described in Section 8.4.

homologous recombination in genome editing in Section 8.3. Homologous recombination can also be used to direct specific large-scale changes at pre-determined positions in the genome, including very large deletions, large-scale inversions, and translocations (*chromosome engineering*).

A second, quite recently developed method of genome editing uses programmable site-specific nucleases. Transfected transgenes are expressed to make *exogenous* endonucleases plus protein or RNA **guide sequences**. The guide sequences are designed to bind to specific sequences at a pre-determined target site within the genome, and are physically bound in some way to the introduced endonuclease. With the exogenous endonucleases attached, the guide sequences bind to their target sequences whereupon the bound endonucleases cut both DNA strands to make a double break (**Figure 8.1B**).

The resulting double-strand DNA break activates an emergency DNA repair system (unnatural double-strand breaks in DNA can be fatal for a cell—the priority is to effect repairs rapidly). The method then relies on cellular DNA repair systems to repair the broken DNA. Often an error-prone DNA repair mechanism is used, which by itself introduces a desired change in the target DNA sequence. It is also possible to artificially intervene in the DNA repair to direct desired changes in the target DNA. In either case, treated cells are screened to identify those cells where the altered nucleotides are of the desired type. We describe three popular methods of genome editing using programmable endonucleases in Section 8.4.

Genome editing allows all manner of changes to be introduced into target sequences, but it can take some time. Rather than inactivate a gene, it is often more convenient to simply down-regulate the gene. **Gene silencing** is a fast and simple approach in which RNA transcripts of a pre-determined target gene are targeted to suppress gene expression. In Section 8.5 we describe methods of gene silencing that depend on introducing target RNA-specific oligonucleotides or antisense RNAs into cells.

Finally, in Section 8.6, we describe how transgenes can be inserted into the germ line to make transgenic animals. That often involves first transfecting cultured pluripotent cells (embryonic stem cells or induced pluripotent stem cells) or germ-line precursor cells. From the perspective of human genetics, transgenic animals have three particularly valuable research applications, as listed below.

- Investigation of how mammalian (and animal) genes function (to allow understanding of different aspects of cell, tissue, and organ biology).
- Disease modeling. As described in detail in Chapter 21, genetically modified animals are crucially important as models of human diseases, allowing deep insights into the molecular basis of disease.
- As test systems for proposed new disease treatments, especially where genetically modified animals have been shown to be good disease models.

8.1 ARTIFICIAL TRANSFER OF GENETIC MATERIAL INTO MAMMALIAN CELLS

The plasma membranes of our cells are principally organized around hydrophobic lipid bilayers and are semi-permeable: they naturally allow exchange of certain small molecules. In general, the smaller and less polar a molecule is, the higher the chance that it can diffuse across the membrane. Small, nonpolar molecules such as oxygen and carbon dioxide therefore diffuse readily across a lipid bilayer; some small polar molecules, such as water molecules, also diffuse across lipid bilayers, but do so at much reduced rates.

Simple ions, such as Na^+, K^+, and Ca^{2+}, cannot simply diffuse across the plasma membrane. They may be passively transported with the help of concentration gradients or be actively pumped against a concentration gradient. Certain integral membrane proteins known as ion channels help regulate the flow of ions across the cell membrane. Other integral protein-based transporter systems are important in regulating transmembrane transport of some polar molecules such as glucose.

Active transport is needed for macromolecules to cross the plasma membrane and involves a type of **endocytosis**, the general process in which a portion of the plasma membrane invaginates to form a pit and then pinches off to form an *endocytic vesicle*, enclosing some of the extracellular fluid. When endocytosis entraps smaller neighboring cells, such as microbes, the mechanism is known as *phagocytosis*. Another type of endocytosis, known as *receptor-mediated endocytosis* and described later in this section, is responsible for internalizing certain proteins and some polysaccharides that bind to a cell surface receptor, and also some viruses that gain entry to the cell by binding cell surface receptors.

The artificial transfer of genetic material into mammalian cells is known as **transgenesis**. Nonviral or viral methods can be employed to expedite the passage of large, charged nucleic acids and oligonucleotides across plasma membranes. The use of physical or chemical nonviral transfer methods is known as **transfection**. Viral approaches are modeled on viruses that naturally infect animal cells: here, the nucleic acids are transferred after packaging them into a virus protein coat, and the transfer process is referred to as **transduction**.

Note that the terminology differs from that used in bacterial systems, where *transfection* means the uptake of naked virus DNA, while the uptake of naked plasmid or genomic DNA is described as *transformation*. (But for mammalian cells, *transformation* is a term reserved for changes in genotype, with ensuing changes in cell attributes and behavior associated with cancer.)

Depending on the delivery system used to get them into cells, transgenes may be designed to migrate to the nucleus and, in some cases, insert into a chromosomal DNA. Otherwise, a cytoplasmic location may be sufficient for their function. As described below, a cytoplasmic function is often intended also for RNA and oligonucleotides transferred into cells. The cells that are transfected are often cultured cells. As we describe below, different procedures have been developed to make stable permanent cell lines that can be disseminated to laboratories throughout the world. Before examining methods of transfecting genetic material into mammalian cells, we first look at how mammalian cells are cultured.

Culturing mammalian cells: primary cell cultures, cell culture methods, and cell storage

Establishing cultured cells begins by surgically removing cells from an organism and placing them into a suitable culture environment. For that, a suitable vessel is needed, made of glass or treated plastic, and a culture medium containing nutrients to help cells grow. The medium contains a buffer to maintain an appropriate pH (often pH 7.0–7.4) and culture vessels are maintained at a suitable constant temperature (usually 37°C), often within an incubator designed to provide an atmosphere of 5% CO_2. After a short period, cells will attach to the surface or a substrate layer on the surface of the culture vessel, then divide and grow to form a *primary culture*. There are two primary routes toward achieving this objective, as listed below.

- *Explant cultures*. Small pieces of tissue are attached to a glass or plastic vessel and bathed in culture medium. After a few days, individual cells will move from the tissue explant onto the culture-vessel surface where they begin to divide and grow.
- *Enzymatic dissociation*. In this widely used method, proteolytic enzymes (trypsin, collagenase) are first used to dissociate tissue fragments. The ensuing suspension of single cells is placed into culture vessels containing culture medium, and cells are allowed to grow and divide.

After cells in the primary culture vessel have grown to fill up all of the available culture substrate, the cells must be *subcultured* to give them room for continued growth. Cells are gently removed from the substrate using proteolytic enzymes (to break protein bonds attaching the cells to the substrate), and the suspension of recovered cells is then subdivided and placed into new culture vessels.

Two basic cell culturing methods are employed, according to whether or not cells are able to adhere to a glass or treated plastic substrate. If they are able to do so, *monolayer cultures* are established, and different vessels are possible: tissue culture treated dishes, T-flasks, roller bottles, multiple-well plates, and so on. If, instead, the cells float freely, they are often kept actively suspended in the medium (*suspension cultures*—the cells are grown in magnetically rotated spinner flasks or in shaken flasks).

Cells may be maintained by *passaging* (or splitting): a small number of cells from a previously grown culture are seeded into fresh culture medium in a new culture vessel. If cells are passaged regularly, cell senescence (associated with high cell density) can be avoided and the cells can be cultured for longer.

Cell line storage and testing

Cell lines may be stored in a frozen state after adding some kind of cryopreservative agent—often DMSO (dimethylsulfoxide)—to the cell medium. As described in the following section, the most useful cell lines are immortalized cell lines that can be cultured indefinitely (they can be temporarily stored frozen in liquid nitrogen; thawing a frozen sample yields a viable culture). Certain major centers serve as cell line repositories—see **Table 8.2** for some examples. There is a need to periodically check the authenticity of cell lines (by DNA fingerprinting) because proximity of other cell cultures in laboratories often leads to overgrowth of cultured cell lines by other, faster-growing cell lines.

TABLE 8.2 EXAMPLES OF MAJOR CELL LINE REPOSITORIES

Cell line repository	Website
ATCC—American Type Culture Collection	http://www.lgcstandards-atcc.org/
Coriell Institute Biorepositories	https://catalog.coriell.org/
ECACC—European Collection of Authenticated Cell Cultures	https://www.phe-culturecollections.org.uk/collections/ecacc.aspx

The utility and construction of immortalized cell lines

Early cell lines had a limited lifetime because cells develop senescence after a finite number of cell divisions (the Hayflick limit, as described in Section 3.2). Immortalized (permanent) cell lines have the huge advantage of allowing cells to be cultured indefinitely. As well as allowing applications that last over long timescales, and avoiding having to devote significant effort into just keeping the cells alive, there is the convenience of being able to access the desired cells from major cell line repositories.

Cells can become immortal when mechanisms regulating cell division are subverted, allowing cells to keep on dividing. That happens naturally in cancers: normal cell division controls are dysregulated following mutations in certain cancer-susceptibility genes, or after **transformation** of cells by cancer-causing viruses.

Some popular permanent cell lines were originally obtained from naturally occurring tumors (such as the HeLa human cell line that originated from Henrietta Lacks, a patient with cervical cancer). And after being identified as key components in cellular transformation, oncogenes could be artificially expressed in cultured cells to induce transformation, resulting in immortalized cell lines. For example, many immortalized cell lines have been made by transfecting and expressing an oncogene from simian virus 40 (SV40). The gene product, the SV40 large T antigen, binds to and inhibits p53 and the pRb retinoblastoma protein, proteins that normally act as brakes on cell division.

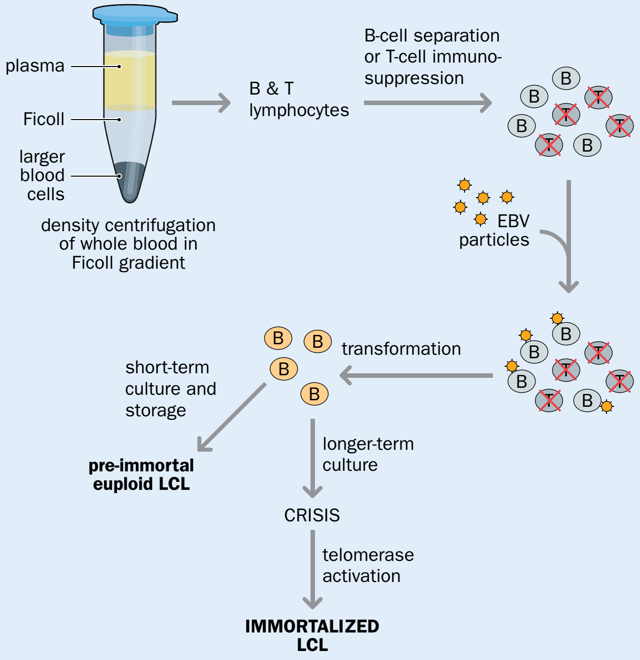
Immortalized cell lines derived from tumors or artificially transformed cells are disadvantaged by genome instability, and aneuploidy is common. Molecular pathways in transformed cell lines, therefore, may not always be representative of those in the original untransformed cells. But for some purposes that is not a major concern. The major utility of immortal human lymphoblastoid cell lines, for example, is to provide an inexhaustible source of DNA from individuals of interest (Box 8.1).

BOX 8.1 MAKING EBV-TRANSFORMED LYMPHOBLASTOID CELL LINES

Most people across the world become infected at some stage in their lives with Epstein–Barr virus (EBV—a gamma herpesvirus) and show no symptoms (after gaining adaptive immunity), but the virus occasionally causes glandular fever. EBV can also transform cells that it infects: it is implicated in different cancers, including certain lymphomas, and it also readily infects resting B lymphocytes *in vitro*, causing growth transformation. A sample of peripheral blood is sufficient, therefore, to prepare transformed B cells. As EBV becomes integrated into B cells, cytotoxic T lymphocytes can be generated that would kill the infected B cells leading to transformation failure. Accordingly, either the T cells are suppressed, or they are removed in some way before the B cells are infected by EBV (Figure 1).

Following transformation, a **lymphoblastoid cell line** (**LCL**) is produced. Initially, the EBV-transformed cells are in a preimmortal stage: the cells are actively proliferating and euploid, with no tumorigenic properties. The telomeres shorten with each round of cell division, but the transformed cells do not undergo senescence at the stage when untransformed B cells would be expected to do so. Nevertheless, after several further rounds of cell division, the cells undergo a stage known as proliferative crisis, during which the vast majority of the cells die. The very rare survivors are cells that have undergone a series of very significant genetic and epigenetic changes, including development of aneuploidy and activation of telomerase. Because these aneuploid cells can proliferate indefinitely, they are described as immortalized LCLs.

LCLs have been used in a variety of functional assays, but as well as providing a continuous *in vitro* source of cells, immortalized LCLs have routinely been used to provide an inexhaustible supply of genomic DNA from individuals of interest. Immortalized LCLs have been developed from large numbers of individuals with genetic disorders and within genetic epidemiology studies, allowing detailed genotype–phenotype correlation studies.



Box 8.1 Figure 1 Constructing lymphoblastoid cell lines (LCLs) by EBV transformation of B lymphocytes. Traditional construction of LCLs has involved separation of lymphocytes by density centrifugation of whole blood. Thereafter, some method is used to remove or suppress T cells. Often an immunosuppressant such as cyclosporin is added to prevent T-cell proliferation, or paramagnetic beads coated with an antibody specific for B cells are used to bind B cells and enrich them using a magnetic sorter. Following mixing with Epstein–Barr virus (EBV) particles (usually maintained and isolated from marmoset cells), B cells may be infected by EBV. Transformed cells are cultured and are initially euploid, but if they go through a sufficient number of population doublings, they go through proliferative crisis after which survivors are aneuploid cells that express telomerase and can proliferate indefinitely (immortalized LCLs). Note that more recently developed methods allow immortalized LCLs to be conveniently isolated from small volumes of cryo-preserved blood (see Amoli MM *et al*. [2008]; PMID 18381392). (Adapted from Sie L *et al*. [2009] *J Neurosci Res* **87**:1953–1959; PMID 19224581. With permission from Wiley-Liss, Inc., © Wiley-Liss, Inc.)

Immortalized euploid cell lines

To make immortalized euploid cell lines that are more representative of the original cells, interest has focused on a final step in cell transformation: activation of TERT, telomerase reverse transcriptase. Recall from Section 2.4 that the DNA of telomeres consists of tandem repeats of the hexanucleotide TTAGGG, and that telomerase extends telomeric DNA by using its RNA component, TERC, to provide an RNA template for the TERT enzyme to make new TTAGGG repeats (see **Figure 2.25**). Only a very few of our cells normally express telomerase because although the TERC RNA is expressed in all cells, the TERT enzyme is normally restricted to unspecialized cells in early development and to immortal stem cells that are needed to replenish body cells. Normal body cells can undergo just a limited number of cell divisions because DNA is lost at the telomeres at each round of DNA synthesis (the chromosome *end-replication problem*; see **Figure 2.24**), causing progressive shortening of telomeres and eventually inducing cell senescence.

Transformed cells do not undergo senescence. But after several rounds of cell division, past the stage where senescence develops in the normal cells, transformed cells undergo a proliferation *crisis*: the very few surviving cells are aneuploid cells that have reactivated telomerase to become immortal. The long route that transformed cells take to achieve immortality can be bypassed by just recreating the last step: by artificially expressing a *TERT* transgene in cultured euploid cells, it is possible to create euploid cell lines with telomerase activity that are effectively immortal.

Nonviral methods of transferring genetic material into mammalian cells

Nonviral methods are used to transfer different types of genetic material into cells. They are quite inefficient by comparison with viral methods, but because they are often quite simple and convenient methods, they have been widely used when the efficiency of transfer is not the highest priority. In the case of gene therapy, concerns about the safety of using virus vectors to transfer transgenes into the cells of a patient have also prompted interest in the alternative of nonviral transfer methods.

Physical methods

Physical methods are used to transfer genetic material into human or animal cells, either by piercing the cell membrane in some way, or by inducing pores in the membrane that allow passage of the nucleic acids or oligonucleotides. Some methods have specialized uses. For example, microinjection of DNA, using a very fine needle to pierce the cell membrane, is limited to transfecting single cells at a time. A common application is to transfer DNA into fertilized oocytes (an important way of delivering genes into the germ line to make transgenic animals). Some of the more generally applicable physical transfer methods are described briefly below.

- *Electroporation*. When exposed to a sufficiently strong electric field, the plasma membrane of a cell undergoes electrical breakdown. Pores form, allowing passage of molecules that normally cannot cross the membrane (**Figure 8.2**). If the exposure is sufficiently short, the membrane can rapidly recover and become semi-permeable again. Electroporation (short for "electric pore formation")

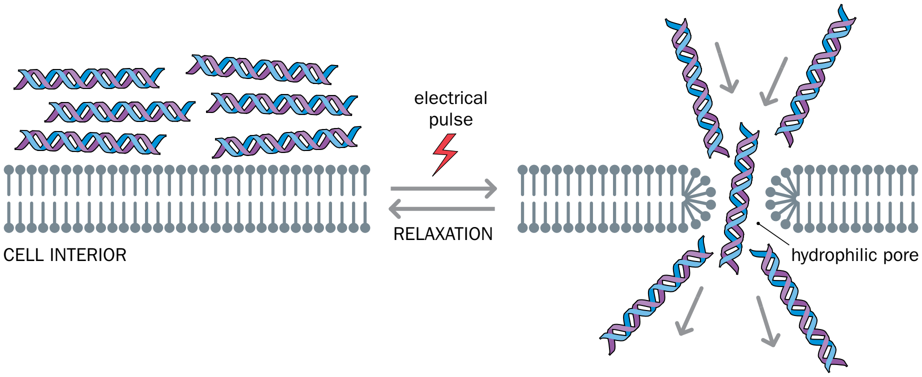


Figure 8.2 Electroporation as a way of permeabilizing the plasma membrane. Although the plasma membrane is a highly fluid structure, it is normally a formidable barrier to nucleic acids and oligonucleotides (which are negatively charged macromolecules). Electroporation involves exposing the cell membrane to very brief pulses of a high voltage electric field, causing pores to form transiently. Hydrophilic pores are bounded by the phosphate groups of membrane phospholipids and facilitate entry of nucleic acids or oligonucleotides into the cell.

involves administering extremely brief pulses of very high voltage to the membranes, allowing entry of desired large molecules and then resealing of the membrane.

- *Sonoporation*. An alternative way of transiently making pores in membranes to increase permeability is to use very brief pulses of high-energy ultrasonic sound.
- *Particle bombardment*. Biolistic methods use a gene gun to fire high-density (gold or tungsten) microparticles coated with nucleic acid; the microparticles are accelerated to very high velocity, usually using compressed gas, allowing efficient transfection of cells, irrespective of cell type. Developed initially to transform plant cells (whose cell walls pose a formidable barrier to passage of macromolecules), particle bombardment has been used to transfect plasmid recombinant DNA into a variety of cultured mammalian and animal cells, as well as tissues *in vivo*. The method is especially useful for transfecting cells that are more resistant to transfection by other methods.

Chemical methods

Certain chemicals can facilitate uptake of genetic material into mammalian cells by endocytosis. Calcium phosphate has long been used as an aid to gene transfer into mammalian cells. It relies on formation of DNA–calcium phosphate co-precipitates that are adsorbed onto the surface of target cells at high concentration, facilitating uptake by the cells through endocytosis. The transfection efficiency is not very high, however.

Various other chemical methods rely on cationic (positively charged) macromolecules that bind the negatively-charged nucleic acid or oligonucleotide molecules and target them to the cell membrane to facilitate their uptake into cells by endocytosis. The vector–nucleic acid complexes have positively-charged surfaces and are attracted to cell membranes whose outer surfaces have numerous negatively-charged phosphate and sulfate groups (within membrane-bound glycoproteins and membrane phospholipids).

Of the wide range of cationic vectors (**Table 8.3**), cationic lipids, as part of artificial lipid bilayers, have been especially widely used because of their comparatively high efficiency. This type of transfer (*lipofection*) uses synthetic spherical vesicles, known as **liposomes**, that have at least one lipid bilayer and form spontaneously when certain lipids are mixed in aqueous solution. After the desired nucleic acids or oligonucleotides are combined with a mixture of a cationic lipid and a helper lipid in water, cationic liposomes spontaneously form with bound nucleic acid/oligonucleotide. After association with the cell membrane, they can be taken up into the cell by endocytosis (**Figure 8.3**). The efficiency may be increased still further when some other chemical vectors, such as polylysine, are also included in the mix.

TABLE 8.3 EXAMPLES OF POSITIVELY-CHARGED CHEMICAL VECTORS FOR TRANSFERRING GENETIC MATERIAL INTO ANIMAL CELLS

Class	Examples	Vector–nucleic acid association
Peptide	Polylysine	Peptide is chemically conjugated to the nucleic acid/oligonucleotide
	Penetratin®	
Polymer	DEAE (diaminoethyl)-dextran	Electrostatic interaction. The complex is known as a *polyplex*
	Polyethyleneimine	
	Poly(amidoamine) dendrimers*	
Lipid	Lipofectin® (1:1 [w/w] liposome formulation of a cationic lipid, N-[1-(2,3-dioleyloxy) propyl]-n,n,n-trimethyl-ammonium chloride, and a helper lipid, dioleoyl phosphatidylethanolamine)	Electrostatic interaction. The complex is known as a *lipoplex*

* So called because they are highly branched structures.

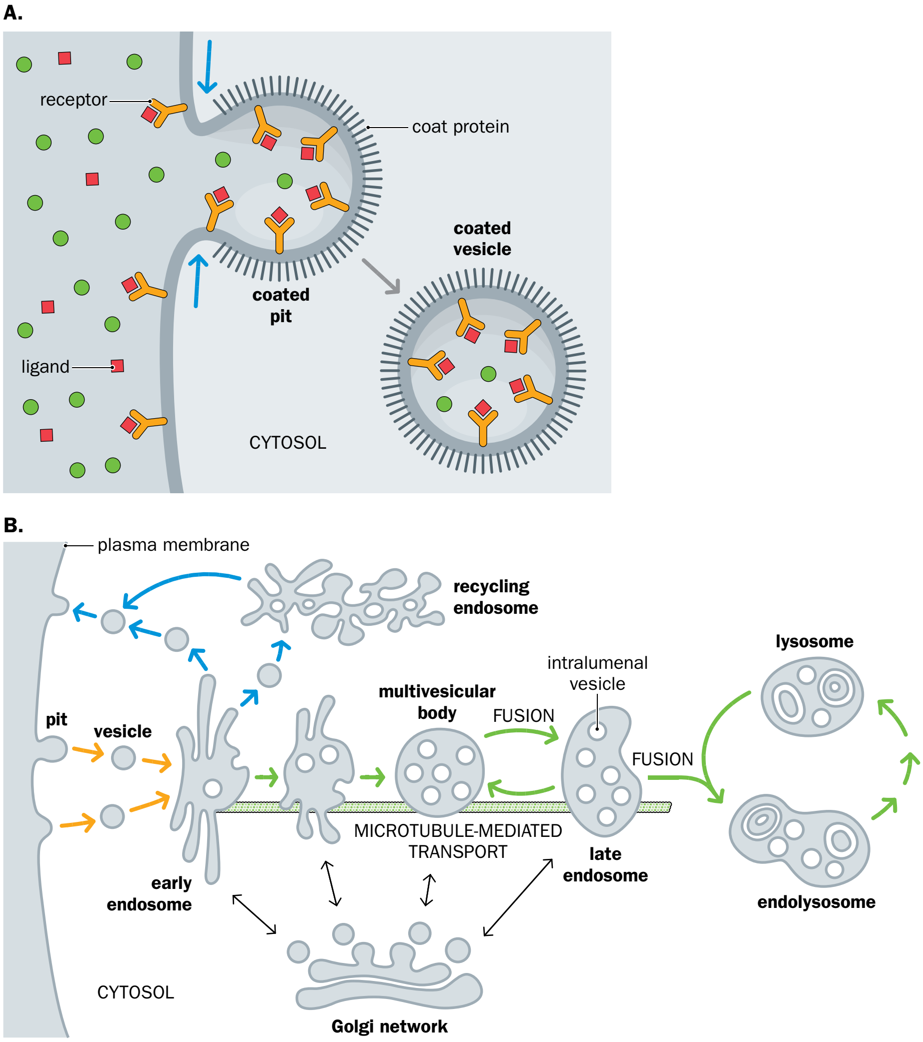


Figure 8.3 Receptor-mediated endocytosis and endosome maturation. (**A**) Receptor-mediated endocytosis. Different classes of protein on the cell surface act as receptors for specific ligands including signaling proteins and viruses. Invagination of the plasma membrane after ligands have bound to receptor proteins leads to formation of pits initially, and then small vesicles that transport the entrapped proteins or viruses within the cell. The pits are coated with a protein, such as clathrin, that plays a major role in vesicle formation, and forms a polyhedral lattice surrounding vesicles. Blue arrows indicate the direction of constriction as the plasma membrane invaginates. (**B**) General scheme for endosome maturation. Endocytic vesicles fuse with an early **endosome**, a sorting station composed of membrane-limited tubules and vesicles. Some membrane proteins, such as receptor proteins, can be returned back to the plasma membrane for reuse, or be stored temporarily in recycling endosomes before being returned. Other membrane proteins and phagocytosed cells are transported via a multivesicular body to a late endosome; further sorting can occur that can lead to fusion with lysosomes and destruction of their contents. (A, adapted from Campbell NA & Reese JB [2008] *Biology* 8th edn. Pearson/Benjamin Cummings; B, from Alberts B *et al.* [2014] *Molecular Biology of the Cell*, 6th edn. Garland Science. With permission from WW Norton.)

Intracytoplasmic passage and nuclear entry

After passage through the cell membrane, transfected nucleic acids/oligonucleotides may need to escape from endosomes (if taken up by endocytosis). When taken up by endosomes, complexes containing bound nucleic acids or oligonucleotides might be expected to be quickly degraded: the endosome containing them would be shunted into the pathway that leads progressively toward formation of a late endosome and fusion with a lysosome (see Figure 8.3B). That fate can be avoided by designing the complex to destabilize the endosome, allowing the transfected genetic material to escape. Figure 8.4 shows how that is achieved in the case of lipofection.

After escaping from the endosome, the genetic material is released from its protective chemical coat and becomes vulnerable to degradation by cytoplasmic nucleases (a defense system against invading viruses). To minimize degradation, oligonucleotides

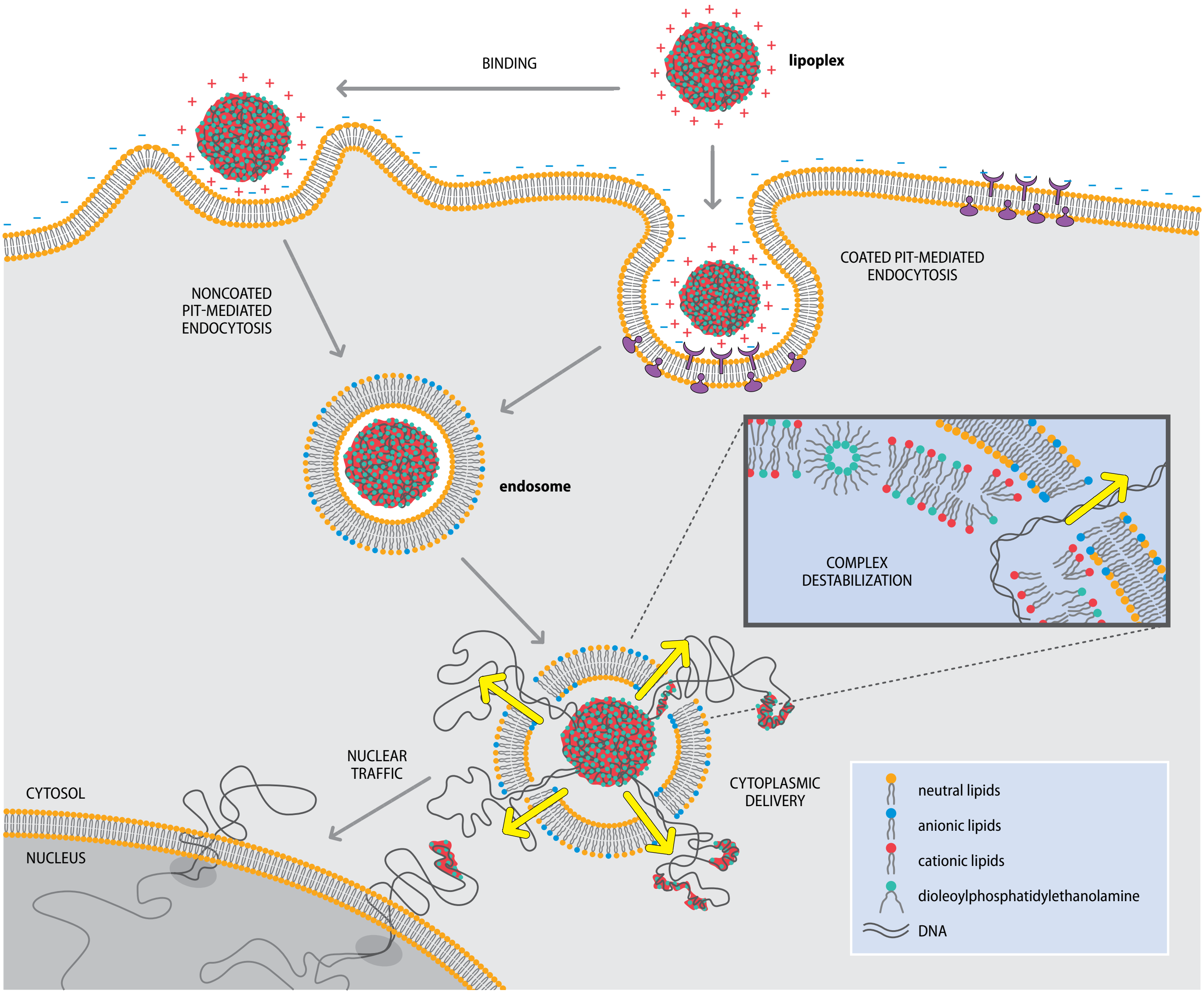


Figure 8.4 Cationic liposomes as vectors for delivery of nucleic acids into mammalian cells. The nucleic acid to be transferred is complexed with liposomes to form *lipoplexes* that have positive charges on the surface, helping interaction with cell membranes (which have multiple negative charges on the surface). The lipoplexes are taken up by cells through different endocytosis pathways in which the cell membrane invaginates to form a pit. Large lipoplexes are taken up by pits coated with clathrin complexes (coated pit-mediated endocytosis at top right); small lipoplexes are taken up by noncoated pits (top left). In either case, the lipoplexes become trapped in *endosomes* (simplified here; see **Figure 8.3B** for a fuller picture) and would be expected to be targeted for destruction by lysosomes. However, the inclusion within the liposomes of certain helper lipids—usually electrically neutral lipids, such as dioleoyl phosphatidylethanolamine—helps to destabilize the endosomal membranes, causing the passenger nucleic acid to escape to the cytoplasm (yellow arrows). For a transferred DNA to be transcribed, it must pass to the nucleus. In dividing cells, the breakdown of the nuclear envelope during mitosis allows the DNA to gain access to the nucleus, but in nondividing cells the precise mechanism of entry into the nucleus is unclear. (From Simões S *et al*. [2005] *Expert Opin Drug Deliv* **2**:237–254; PMID 16296751. With permission from Informa Healthcare.)

that are intended to work in the cytoplasm (to block RNA expression) are designed to be robust, nonstandard, synthetic oligonucleotides that are resistant to nucleases (**Figure 8.5**). For transgenes, the nucleus offers a safer environment, but whereas some virus vectors readily gain access to the nucleus, and even integrate into chromosomal DNA, using nonviral transfer to get a transgene into the nucleus is less straightforward (transport through the nuclear pores is generally inefficient).

To facilitate nuclear targeting, various nuclear localization signal (NLS) peptides were developed to assist active transport of transgenes through nuclear pore complexes. Subsequently, however, *nucleofection*, a proprietary modification of the electroporation method using cell-type-specific reagents, was developed by the Amaxa company and has been used with considerable success to transfect transgenes into both the nucleus and cytoplasm. This method has been particularly useful for transfecting a wide variety of nondividing cell types, including neurons.

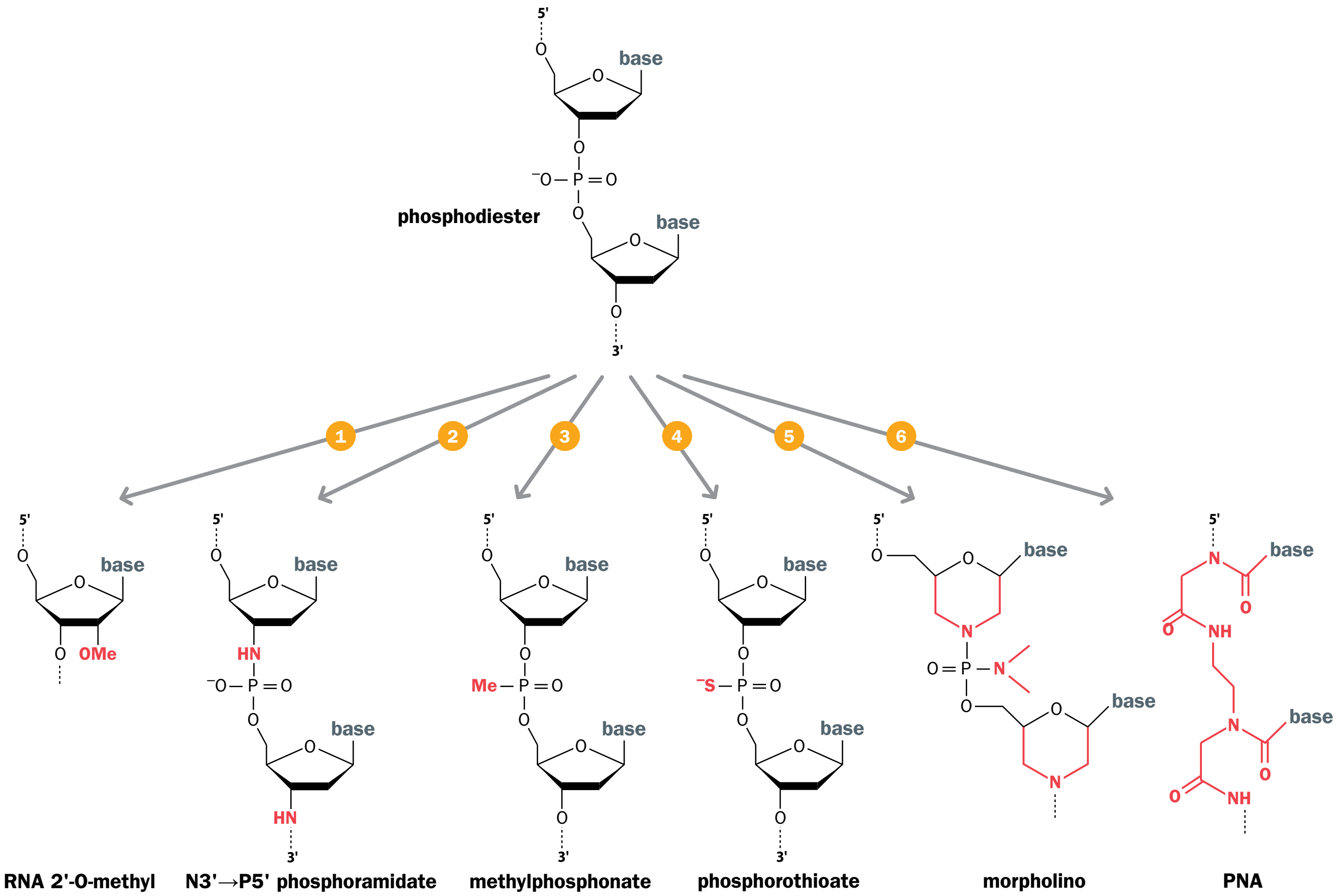


Figure 8.5 Chemical modification can increase oligonucleotide stability. The standard oligodeoxyribonucleotide structure is shown at the top. Four of the six modified oligonucleotides have a minor change, involving either replacement of atoms directly linked to carbons 2′ or 3′ of the deoxyribose (1,2) or replacement of an oxygen ion of the connecting phosphate (3,4). The other two modifications (5,6) are radical alterations to the normal structure, producing morpholino oligonucleotides or peptide nucleic acids (PNA). (Modified from Dias N & Stein CA [2002] *Mol Cancer Ther* **1**:347–355; PMID 12489851.)

Transgene size range

As general methods for delivering genetic material to mammalian cells, nonviral transfer methods have two principal advantages over viral methods. First, they can readily transfer any type of genetic material—including short RNA molecules or chemically-modified oligonucleotides (viral methods are especially used to transfer DNA, but viruses may convert DNA to RNA for propagation purposes). Secondly, nonviral methods can be used to transfer extremely large molecules: it has been possible to introduce transgenes containing megabases of DNA into human cells, where they replicate independently and behave as artificial chromosomes (an example is described in Chapter 21).

Viral methods of transferring DNA into mammalian cells

Over long periods of evolution, viruses have refined ways of packing their genomes into protective protein coats and injecting them into cells. According to the virus, the genome can be DNA or RNA and either single-stranded or double-stranded (**Box 1.2** describes the extraordinary variety of viral genomes). Viruses are most readily manipulated as double-stranded DNA molecules to which a DNA of interest can be covalently attached, forming a transgene that can be packaged into a viral protein coat and transported into cells. In the case of RNA viruses, double-stranded DNA copies of the RNA genome (which naturally exist during the viral life cycle as *replicative form* DNA) are used. If required, the transferred DNA can be a copy of an RNA of interest (artificially made using a reverse transcriptase).

Viral transfer methods offer multiple advantages. First, they allow much higher transfer efficiency than nonviral methods. Secondly, according to the type of virus, transgenes can be ferried into the cytoplasm or nucleus; in the latter case, retroviruses (RNA viruses that replicate through a DNA intermediate) allow integration of a transgene into the genome (retroviral integration into a host-cell chromosome is mandatory for successful completion of the life cycle). As a result, a transgene can shelter in the stable environment of chromosomal DNA and be inherited when cells divide. Additionally, certain strains of viruses are also suited to infecting particular types of cell.

Viral transfer methods do have some downsides. Although their protein coats protect the transferred DNA from nuclease attack, they impose size limits on the DNA that can be transferred—sometimes the maximum limit is just a few kilobases of DNA. And in the case of gene transfer *in vivo* there can be safety concerns. We will explore these in detail when we consider gene therapy in Chapter 22.

Transduction using retroviral vectors

A **retrovirus** has a single-stranded RNA genome but replicates in the host cell through the process of reverse transcription (in which the RNA is converted to DNA). A complete retrovirus particle (*virion*) has two copies of the RNA genome enclosed within a capsid protein coat; in turn, the capsid is surrounded by a lipid bilayer envelope with spike glycoproteins on the outside. In addition to different types of structural protein, the virion contains three key enzymes: a reverse transcriptase, a protease, and an integrase (Figure 8.6).

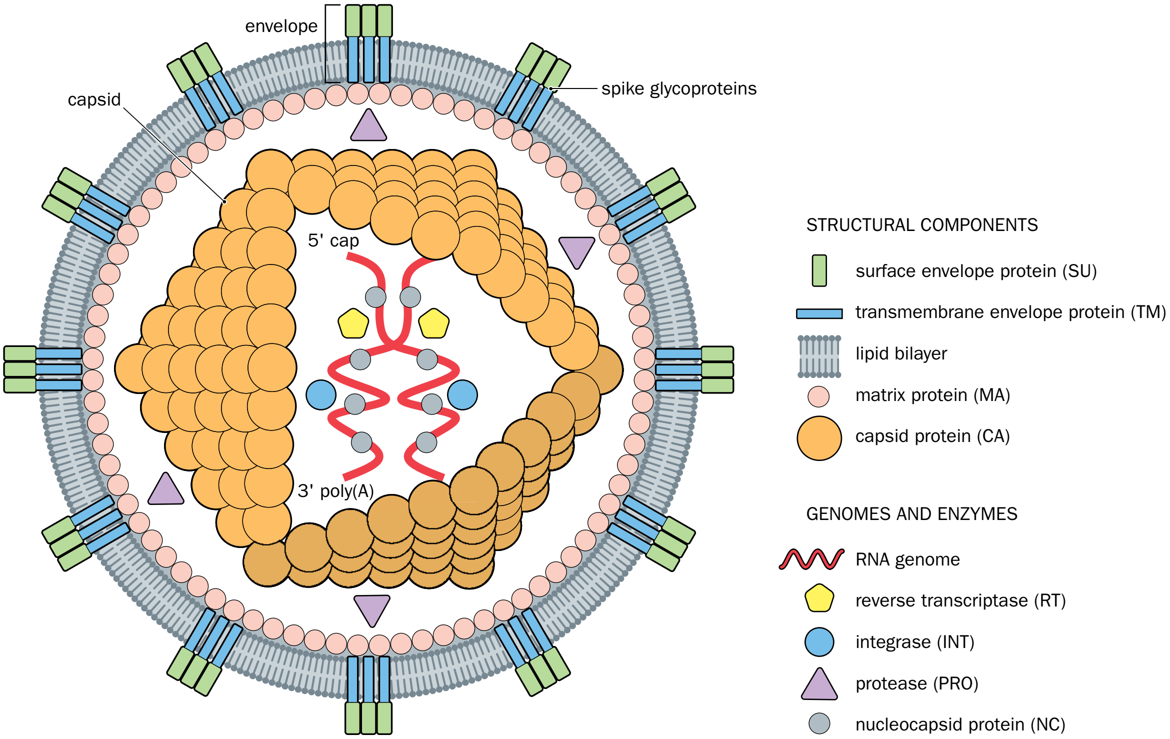


Figure 8.6 Retrovirus structure. An infectious retrovirus particle (virion) has two copies of a single-stranded RNA genome, each with an m7GpppG cap at its 5′ end and a 3′ poly(A) tail (the genome is said to be a *positive* single-stranded RNA because in the cytoplasm the same RNA serves as a sense strand for making proteins). The genomic RNA is bound and structured by a nucleocapsid protein, and is enclosed within an inner capsid along with some viral enzymes (described in **Figures 8.7** and **8.8**). The capsid and its contents constitute the *core particle*, which is surrounded by an envelope consisting of a lipid bilayer with attached proteins. In addition to structural matrix proteins, the envelope periodically has spike glycoproteins consisting of an outer surface glycoprotein (which binds to specific receptors on the surface of cells) and a transmembrane glycoprotein (which aids virus entry into a cell by triggering fusion between the virus lipid bilayer and the cell's plasma membrane).

An infectious retrovirus particle enters a cell after its surface envelope glycoproteins bind to specific receptors on the cell surface (such as the CD4 receptor in the case of human immunodeficiency virus, HIV). Usually the transmembrane envelope glycoprotein then triggers the fusion of the viral and cellular membranes, allowing the virus to enter as a *core particle*, lacking the outer envelope; see Figures 8.6 and 8.7. (But some retroviruses enter by the receptor endocytosis mechanism described in Figure 8.3.) After infecting a cell, the core virus uses its multifunctional DNA polymerase to make a complementary DNA copy of its RNA, degrade the original RNA, and then copy the surviving single-stranded DNA (see Figure 8.7).

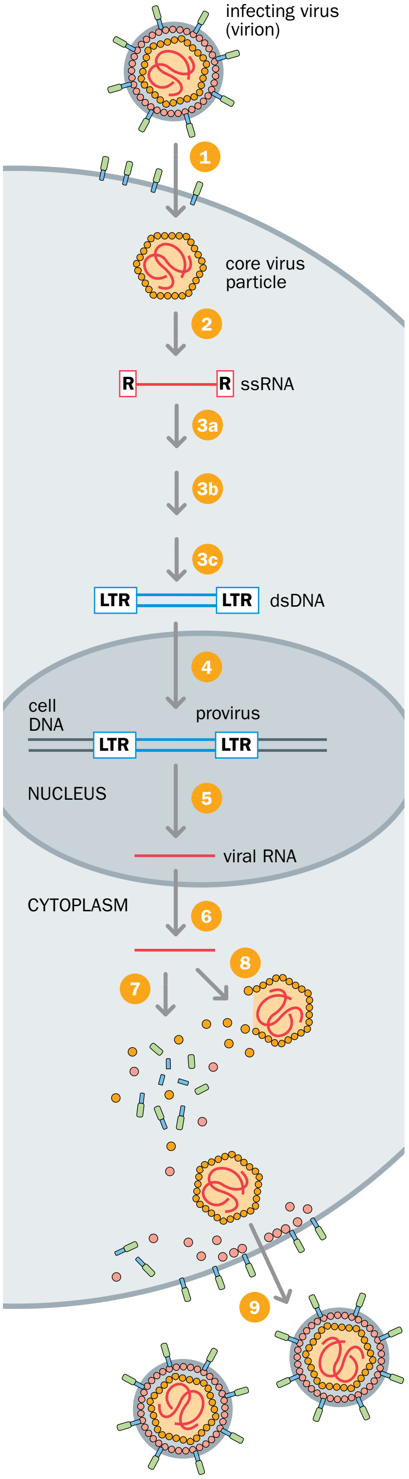


Figure 8.7 Retroviral life cycle. Infection begins when the virus envelope surface glycoprotein recognizes specific receptors located on the cell surface (1) and the virus enters the cell, usually as a core virus particle that lacks the outer envelope. Thereafter, the viral RNA is released from the capsid (2) and the viral DNA polymerase converts the single-stranded (ss) RNA genome into a double-stranded (ds) DNA (3). That is possible because the viral DNA polymerase is multifunctional, having: a reverse transcriptase activity (uses the ssRNA to synthesize a complementary DNA—step 3a); an RNase H activity (degrades the RNA—step 3b); and a standard DNA polymerase activity (converts the ssDNA to a dsDNA—step 3c). After entry into the nucleus, the viral DNA inserts into chromosomal DNA using the viral integrase enzyme (4), enabling the viral genome (now called the *provirus*) to be stably maintained, replicated during DNA synthesis, and passed to progeny cells. Viral RNA is produced (5) and migrates to the cytoplasm (6) where it is translated to produce viral structural proteins and enzymes (7). Viral RNA also associates with some newly-produced viral proteins to form new core particles (8). The core particles then obtain their envelopes and are released from the cell (budding; step 9). Mature progeny virions are then capable of infecting new cells. LTR, long terminal repeat; R, short repeat in LTR (see **Figure 8.8B**).

The resulting double-stranded DNA can be incorporated into the host-cell genome using the viral integrase. The integrated virus, known as a *provirus*, may remain in the host-cell genome and be transmitted to daughter cells following cell division. If germline cells are infected, the virus may be transmitted vertically, but horizontal transmission is the norm: after a virus infects a cell, the transcriptional and translational machinery of the host cell is hijacked to produce new copies of the viral genome plus viral proteins, after which new virus particles are assembled and then exit from the cell to infect other cells (see **Figure 8.7**).

Retroviruses have small genomes, from 7 to 10 kb in length. Simple retroviruses, such as gammaretroviruses, have three genes: *gag*, *pol*, and *env*. They each make protein precursors (polyproteins) that are cleaved to produce two or three individual proteins. In addition, various regulatory sequences are found in the flanking sequences (**Figure 8.8**). More complex retroviruses, including lentiviruses such as HIV, have extra genes involved in replication in addition to the *gag*, *pol*, and *env* genes.

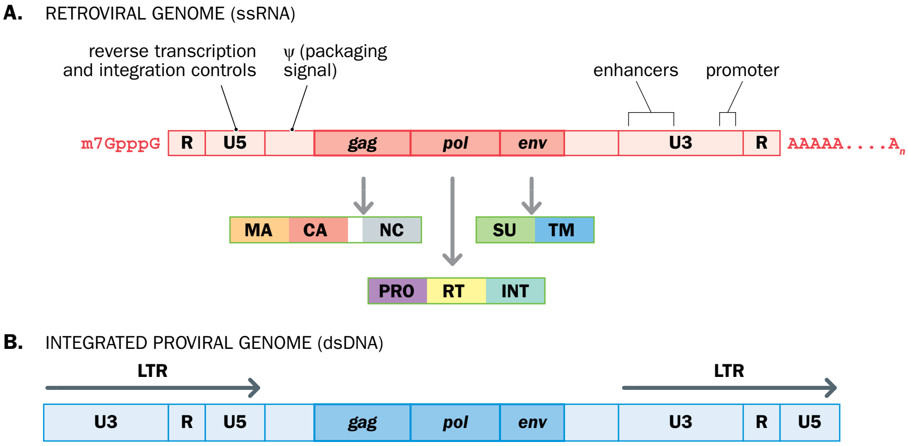


Figure 8.8 Functional components of a simple retroviral genome. (**A**) The positive single-stranded RNA genome in a virion has a 5′ cap and a 3′ poly(A) tail because the same RNA strand is also translated in the cytoplasm to make viral proteins. Each of the three genes makes a polyprotein precursor that is cleaved to give the individual proteins (see **Figure 8.6** for the protein names and locations in the virus). The *gag* (group antigen) gene makes various structural proteins. The *pol* gene gets its name because one of its products is a multifunctional DNA polymerase: it can use both RNA templates—a reverse transcriptase (RT) activity—and also DNA templates. The *pol* gene also encodes a protease (PRO) and an integrase (INT), the enzyme used to insert the viral genome (as double-stranded DNA) into the host cell's nuclear genome. The *env* gene makes the two envelope proteins present in the spike glycoproteins (see **Figure 8.6**). The retroviral protease cleaves the polyproteins encoded by *gag* and *pol*; cellular proteases are responsible for cleaving the *env*-encoded polyprotein. The flanking sequences contain a short repeat (R) plus some regulatory sequences, including U5 and U3 sequences (in the 5′ and 3′ untranslated regions, respectively) and psi (ψ), a crucial packaging signal that directs incorporation of the RNA genome into virions, which is located downstream of U5. (**B**) When the RNA genome is converted to double-stranded DNA, the R repeat is responsible for transfer of DNA synthesis between templates so as to cause duplication of the U5 and U3 sequences. As a result, the proviral genome has long terminal repeats, each containing a U5 and U3 sequence as well as the R sequence.

Different classes of retroviral vector are used to transfer transgenes into cultured mammalian cells, but vectors based on gammaretroviruses, such as murine leukemia viruses, have been widely used. They cannot pass through nuclear pores, and so cannot be used with nondividing cells, but are able to access the nucleus of dividing cells after the nuclear membrane dissolves in preparation for mitosis and then re-forms. Initially, gammaretrovirus vectors were commonly used in gene therapy, but safety concerns associated with their use have prompted the alternative use of lentivirus vectors for this application.

Standard recombinant gammaretrovirus vectors are made by using genetic engineering to modify a double-stranded gammaretroviral DNA, replacing retroviral sequences to create replication-defective vectors. The early steps in the retroviral life cycle—retroviral entry into cells, reverse transcription of the viral genome into DNA, and integration of the viral genome into the host genome—do not require viral protein. As a result, all viral coding regions can be deleted (and replaced by a desired transgene) and the remaining viral sequences can be reduced to the minimum required for high-efficiency transfer. The viral proteins do need to be supplied *in trans* and a **packaging cell line** is required to build a virus coat for a vector containing the desired foreign DNA. Virus particles obtained from such cells can then be used to introduce the desired DNA into a target cell of choice (**Figure 8.9**).

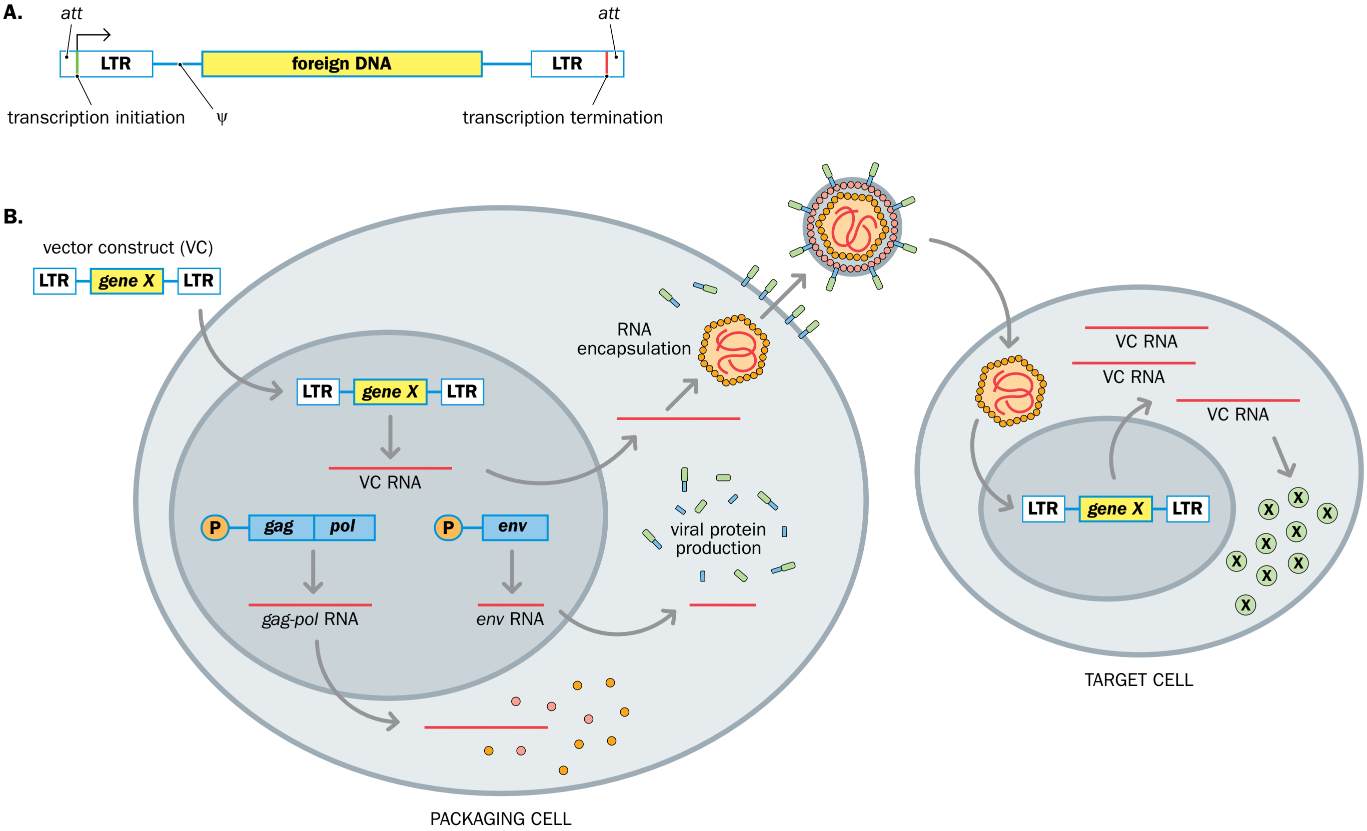


Figure 8.9 Retroviral vectors are produced after first transfecting the recombinant retroviral DNA into a packaging cell line. (**A**) A transgene that is a recombinant retrovirus, often called a *vector construct* (VC). The *gag*, *pol*, and *env* genes have been deleted and replaced by a foreign DNA of interest, but the viral regulatory sequences have been retained. They include promoter/enhancer sequences, transcription termination signals, and the *att* (integration) and ψ (packaging) signals. (**B**) Packaging cell lines are prepared by transfecting a suitable cell with viral genes that can supply suitable viral proteins *in trans* to build a virus particle containing complementary RNA for a foreign DNA. For example, plasmids with the coding sequences for the *gag*, *pol,* and *env* genes plus nonviral upstream promoter/enhancer sequences (P) can be transfected into suitable mammalian cells, where they are stably maintained and where they produce viral structural and enzymatic proteins. When a retroviral vector is introduced into the packaging cell, vector construct RNA can be packaged, resulting in the production of virus particles containing a vector construct RNA genome. This virus can be harvested and then used to infect target cells to introduce foreign *gene X* in the vector construct into the cells, and to make *gene X* product. Because these target cells do not express viral proteins, the vector will not be propagated further. (The viral genes in the packaging cells are not carried along with the vector because they lack the *cis*-acting sequences necessary for propagation.)

Transduction using nonretroviral vectors

Various types of DNA virus have also been used to transduce mammalian cells, notably adenoviruses that normally cause benign infections of the upper respiratory tract in humans. The linear, double-stranded DNA genome remains nonintegrated within the cell nucleus, and unlike retroviruses they are able to infect both dividing cells and nondividing cells. So, in addition to transducing cultured mammalian cells, they can be used to ferry transgenes into nondividing cells *in vivo*.

For use as vectors, adenoviruses have two major advantages over retroviruses. First, the large genome size means that adenoviral vectors can accept quite large insert DNAs—"gutless" adenoviral vectors (which lack any of the adenoviral genes) can accommodate transgenes up to 35 kb in length. (A packaging cell line is required to provide viral proteins *in trans*, just as in the case of retrovirus vectors.) Second, adenoviruses can be produced in very high titers (much higher than retroviruses) and so they offer high levels of transgene expression. Until recently, adenovirus vectors were widely used in gene therapy, but safety concerns have led to the use of alternative vectors, as described in Chapter 22.

Host range and tropism

Viruses gain access to cells after a protein on the outer virus surface recognizes and binds to a specific receptor on the host-cell surface. Different viruses bind different cell surface receptors, and even different strains of the same virus sometimes use different receptors. The virus (and virus vectors) may be limited to infecting cells from one species (the receptor is not so well conserved between species) or may be able to infect cells from a range of species (the receptor may have been very highly conserved during evolution). In the case of Moloney murine leukemia retroviruses, for example, the surface envelope glycoprotein encoded by the *env* gene is the protein responsible for receptor binding, but according to the strain of virus, the virus particles may infect murine cells only (the *env* gene is said to be ecotropic) or a range of mammalian species, including humans.

The **tropism** extends to cell type. For example, HIV is tropic for certain immune system cells that bear the CD4 receptor, notably helper T cells, macrophages, and dendritic cells. Certain strains of the same virus may preferentially infect cells of different types, as in the case of strains of adeno-associated virus.

8.2 PRINCIPLES OF TRANSGENE EXPRESSION IN MAMMALIAN CELLS

Transgenes introduced into mammalian cells often have internal sequences that are designed to be expressed. A minimum requirement is to provide some strong upstream promoter to drive expression to make an RNA product. If the transgene RNA is intended to be an mRNA, a downstream poly(A) addition signal will have been stitched into the transgene.

If a transgene is intended to be incorporated into the germ line and subsequently be present in the cells of a transgenic animal, suitably tissue-specific or stage-specific promoter regulatory sequences may be stitched into the transgene so that its expression may be designed to occur in particular tissues or at particular developmental stages appropriate for its intended function.

The transgene may be designed to exist in an extrachromosomal location in the cell, or it may be intended to be incorporated into a chromosomal DNA molecule. Integrated transgenes have the advantage that they will be replicated and transmitted to descendants of the original cell, but the chromosomal DNA environment may strongly influence the extent to which the transgene is expressed (as described below).

Inducible promoters allow control over transgene expression

Maximum control over transgene expression is provided by inducible promoters that can be switched on and off according to need, usually by controlling the supply of a particular chemical ligand. Typically, the transcription factors that regulate such promoters are structurally modified by this ligand. Naturally inducible promoters have been used, but more robust inducible promoter systems have been developed either by using nonmammalian, often bacterial, components, or by engineering mammalian proteins so that they are incapable of responding to endogenous inducers. Two widely used systems for regulating inducible transgene expression are described below.

Tetracycline-regulated expression

Here, expression is regulated at the transcriptional level. Two transgenes are involved. One has an *E. coli tetR* gene to constitutively express the Tet repressor protein. The other has the sequence of interest (usually a coding DNA) and an upstream promoter with the *tetO* operator in the intervening space. By specifically binding to the *tetO* sequence, the Tet repressor protein blocks expression of the downstream coding sequence. Gene expression can be restored, however, by providing doxycycline, a tetracycline analog that binds to the Tet repressor causing it to change its conformation and be removed from the *tetO* sequence (**Figure 8.10A**). Because expression is induced at the transcriptional level rather than at the protein level, a significant delay may occur before expression is induced in response to providing or removing the induction signal.

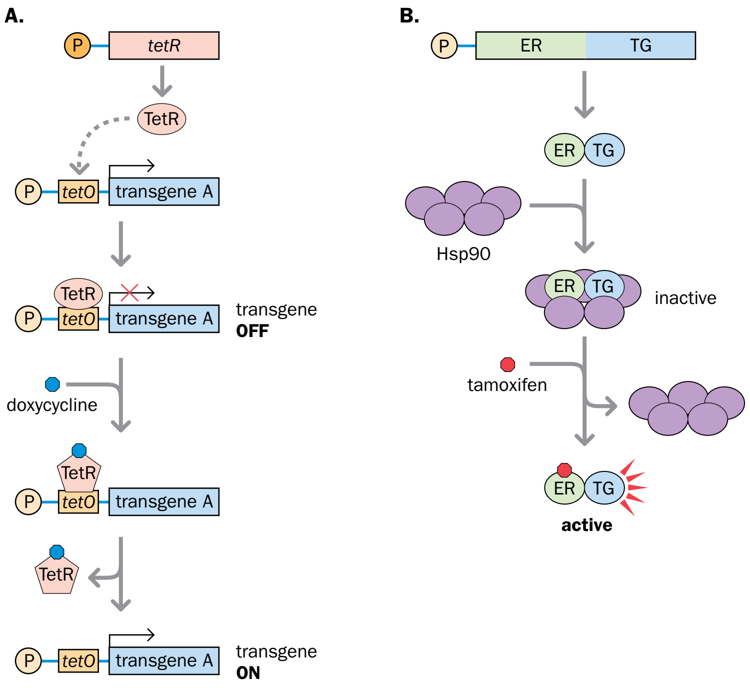


Figure 8.10 Inducible expression of transgenes. (**A**) Tetracycline-inducible expression. A *tetO* operator inserted just upstream of transgene A is a recognition sequence that can be specifically recognized and bound by the *E. coli* Tet repressor protein (TetR). When constitutively expressed from a separately introduced *tetR* transgene, the Tet repressor protein binds to the *tetO* operator, and so prevents the expression of transgene A. However, if doxycycline (an analog of tetracycline) is subsequently provided, it binds to the Tet repressor protein, causing it to change its conformation so that it is no longer capable of binding the *tetO* operator. As a result, expression of the previously silenced transgene A is switched on. Note that toxicity effects reflecting high-level constitutive expression of the repressor limit the use of this system in some cells. (**B**) Tamoxifen-inducible expression. Here, coding sequences for the ligand-binding domain of a mutant mouse estrogen receptor (ER) are fused to a transgene of interest (TG) in an expression vector. The expressed ER–TG fusion protein is bound by the Hsp90 inhibitory protein complex, preventing it from performing its function. The mutant ER ligand-binding domain is not recognized by estrogen, but will bind to tamoxifen, an estrogen analog. Following tamoxifen binding, the fusion protein is released from the Hsp90 complex, and the protein of interest is activated, even though it is part of a fusion protein. P, promoter.

Tamoxifen-regulated expression

Where rapid induction and rapid decay of gene expression are essential, an inducible system that works at the protein level is needed. The estrogen receptor is normally bound by the Hsp90 protein inhibitory complex, keeping it inactive within the cytoplasm until its ligand, estrogen, enters the cell. Estrogen binds to the C-terminal ligand-binding domain of the estrogen receptor, causing a change of conformation. As a result, the Hsp90 inhibitor is released and the activated estrogen receptor translocates to the nucleus to activate certain target genes (**Figures 3.2** and **3.3** give the structure of nuclear hormone receptors and the mechanism of activation by ligands).

Tamoxifen-inducible expression uses a mutant ligand-binding domain of the mouse estrogen receptor, Esr1. The mutant domain does not bind its natural ligand estrogen at physiological concentrations, but will bind the synthetic ligand 4-hydroxytamoxifen. A cDNA for the mutant Esr1 ligand-binding domain is fused to the coding sequence of a gene of interest. Hsp90 binding prevents expression of the fusion protein until tamoxifen is provided (**Figure 8.10B**).

Transient and stable expression in mammalian cells

Recombinant DNA clones containing a mammalian cDNA with appropriate expression signals can be expressed in bacterial cells to give high-level expression of mammalian proteins, but the biological properties of the expressed proteins may not be fully representative of the native molecules. Expressing mammalian proteins in mammalian cells has the obvious advantages that correct protein folding and post-translational modification are generally not an issue, and it is possible to analyze downstream signals and cellular effects. Once transfected into mammalian cells, a transgene with appropriate expression signals can be expressed for short or long periods, according to the expression systems used, as listed below.

- *Transient expression*. Here, the transgene remains as an independent, extrachromosomal genetic element (an *episome*) within the transfected cells. In cultured cells, expression of the introduced gene often reaches a maximal level about two or three days after transfection of the expression vector into the mammalian cell line, but that may be sufficient to carry out certain types of gene analysis. Thereafter, expression can diminish rapidly as a result of cell death or loss of the expression construct. (But in long-lived, nondividing cells, expression of non-integrated transgenes may be maintained for long periods.)
- *Stable expression*. Here, the aim is to have the transgene integrate into the host cell's genome. That often requires a virus vector (retrovirus vectors are highly efficient at integrating into chromosomal DNA) but plasmid expression vectors can also randomly integrate at low frequencies, and stable cell lines can be developed with an integrated

gene of interest. The advantage here is that if the transgene has integrated so that the introduced gene is expressed, all descendant cells will contain this gene and expression can be maintained over many cell generations. Neighboring inhibitory regulatory elements and closed chromatin structure may result in silencing of the transgene ("position effects"). The structure of the transgenic locus may also have an effect when more than one copy of the transgene inserts at the same chromosomal location.

Transient expression: the example of COS cells

Stable expression systems in mammalian cells, typically based on plasmid sequences that integrate within a chromosome, have delivered kilograms of complex proteins in industrial-scale bioreactors, but they have required large investments in time, resources, and equipment. As an alternative, large-scale transient expression systems have been developed for producing "recombinant" proteins in mammalian cells.

In addition to delivering protein expression, some mammalian cell lines have found major applications in screening the effects of *in vitro* manipulations on transcriptional and post-transcriptional control sequences. A good example is provided by COS cells, stable cell lines originally derived from an African green monkey kidney fibroblast cell line, CV-1. When CV-1 cells are infected with the monkey virus SV40, the normal SV40 lytic cycle ensues. However, when CV-1 cells were transformed with a strain of SV40 with a defective replication origin, a segment of the SV40 genome became integrated into the chromosomes of CV-1. The resulting COS cells (**C**V-1 with defective **o**rigin of **S**V40) stably express the SV40-encoded large T antigen, the only viral protein required for activation of the SV40 origin of replication.

By expressing the SV40 large T antigen in a stable way, COS cells permit any introduced circular DNA with a functional SV40 origin of replication to replicate independently of the host cell's chromosomes, with no clear size limitation. When transient expression vectors are transfected into COS cells, permanent cell lines do not result because massive vector replication makes the cells nonviable. Even though only a small proportion of cells are successfully transformed, the amplification of the introduced DNA to high copy numbers in those cells compensates for the low take-up rate.

Stable expression and marker-selection systems

Genes introduced into mammalian cells can stably integrate into host chromosomal DNA, but the process is very inefficient. The rare stably transformed cells must be isolated from the background of nontransformed cells by selection for some marker. Two broad approaches have been used: functional complementation of mutant host cells and dominant selectable markers.

In functional complementation, host cells have a genetic deficiency that prevents a particular function; the original function can be restored if a functional copy of the defective gene is supplied by transformation by a vector. The transgene and the marker can be transferred as separate molecules by a process known as co-transformation. An example is the use of a thymidine kinase gene marker with cells that are genetically deficient in thymidine kinase (TK^-). Thymidine kinase is required to convert thymidine into thymidine monophosphate (TMP or thymidylate), but TMP can also be synthesized by enzymatic conversion from dUMP. The drug aminopterin blocks the dUMP → TMP reaction, so cells cannot grow in the presence of this drug unless they have a source of thymidine and a functional thymidine kinase gene. Selection for TK^+ cells is usually achieved in HAT (**h**ypoxanthine, **a**minopterin, **t**hymidine) medium (**Figure 8.11**).

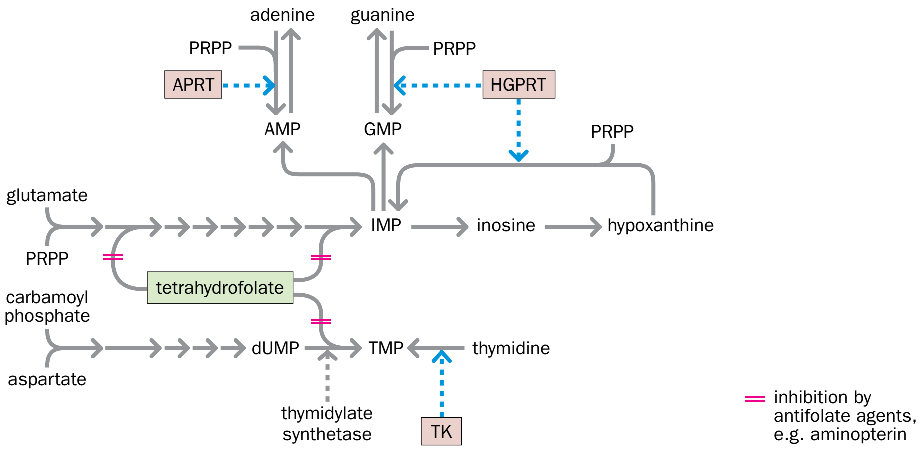


Figure 8.11 Purine and pyrimidine salvage and the basis of HAT selection in thymidine kinase marker assays. Cultured animal cells can normally synthesize purine nucleotides and thymidylate (TMP) by *de novo* pathways that involve certain amino acid substrates. *De novo* synthesis of purine nucleotides begins from glutamate. It involves the eventual synthesis of IMP (inosine monophosphate), which can be converted separately to AMP and GMP. *De novo* synthesis of pyrimidine nucleotides begins from aspartate. It is dependent on synthesizing UMP and dUMP, which is converted to TMP by thymidylate synthetase; UTP is deaminated to give CTP (not shown). An activated form of tetrahydrofolate is crucially required to provide a methyl or formyl group at three key steps that can be blocked by antifolate agents such as aminopterin, shutting down the *de novo* pathways. However, purine nucleotides and thymidylate can also be synthesized from pre-formed bases by salvage pathways that are not blocked by antifolate drugs such as aminopterin. The three salvage-pathway enzymes are shown in pale pink boxes: thymidine kinase (TK), hypoxanthine–guanine phosphoribosyltransferase (HGPRT), and adenine phosphoribosyltransferase (APRT). The use of HAT medium (containing hypoxanthine, aminopterin, and thymidine) therefore selects for cells with active salvage-pathway enzymes. Mutant *tk*⁻ cells are genetically deficient in thymidine kinase and will not survive if grown in HAT medium unless they contain vector molecules containing a functional thymidine kinase gene. PRPP, 5-phosphoribosyl-1-pyrophosphate.

The major disadvantage of complementation markers is that they are specific for a particular mutant host-cell line in which the corresponding gene is nonfunctional. As a result, they have largely been superseded by **dominant selectable markers**, which confer a phenotype that is entirely novel to the cell and hence can be used in any cell type. Markers of this type are usually drug-resistance genes of bacterial origin that can confer resistance to drugs known to affect both eukaryotic and bacterial cells. For example, the aminoglycoside antibiotics (which include neomycin and G418) are inhibitors of protein synthesis in both bacterial and eukaryotic cells. The neomycin phosphotransferase (*neo*) gene confers resistance to aminoglycoside antibiotics, so cells that have been transformed by the *neo* gene can be selected by growth in media containing G418.

8.3 GENOME EDITING USING HOMOLOGOUS RECOMBINATION

During meiosis, homologous chromosomes—maternal and paternal chromosomes that have almost identical DNA sequences—pair up and exchange sequences by recombination. To achieve this, both DNA strands in the participating chromosomes must be cleaved, and then hybrid DNA sequences are formed by joining fragments from the homologous chromosomes. But **homologous recombination** is not just a special characteristic of germ cells: it is also deployed in somatic cells as a way of repairing double-strand DNA breaks, and to rescue stalled or collapsed replication forks.

Unlike in meiosis, there is no mechanism for pairing of homologous chromosomes in somatic cells. However, after the DNA of a single chromosome has replicated, the two double-stranded DNA molecules are held tightly together, forming sister chromatids. *Sister chromatid exchange*—a type of homologous "recombination" —involves formation of double-strand breaks at analogous locations on the paired DNA molecules of sister chromatids, followed by joining of fragments originally located on different DNA molecules. As we explain in Section 11.2, an invading DNA strand from a chromatid can be used as a template to direct repair of the break on the opposing sister chromatid, a form of homologous recombination that allows flawless repair of spontaneous double-strand breaks in cellular DNA (see **Figure 11.6**). Homologous recombination can occasionally also take place between sequences on homologous chromosomes in somatic cells.

The hallmark of homologous recombination is that it takes place only between DNA duplexes having extensive regions of sequence similarity (*sequence homology*). Stable base pairing is a necessary requirement because recombination occurs only after invading single strands from each DNA duplex form heteroduplexes with a sufficiently high degree of base matching.

Homologous recombination is not just confined to naturally occurring DNA molecules in a cell. By introducing a transgene containing sequences that are identical to the sequence of a specific portion of a chromosomal DNA molecule, artificial homologous recombination may take place between the introduced transgene and the homologous target sequence on a chromosomal DNA molecule. That approach enabled the first attempts to alter the genome of intact cells, in a pre-determined way, and at a pre-determined specific target sequence. Often the target sequence is a gene, and this use of genome editing is sometimes known as *gene targeting*.

Genome editing using homologous recombination: general strategies and the need for selection systems

Using artificial homologous recombination, a specific target gene of interest can be modified within an intact cell in any way that we want. Usually, a suitable DNA clone containing a faithful copy of the target sequence is modified in some way so that a central segment has a modified sequence, or an entirely new sequence. The transgene is transfected into suitable cultured cells, and then the cells are screened to identify rare cases where homologous recombination between the transgene and the chromosomal target sequence has resulted in transfer of the altered sequence into the target sequence in chromosomal DNA (**Figure 8.12A**).

The strategy might be to delete an entire gene, a specific exon, or a specific regulatory sequence, such as an enhancer sequence. Very precise changes can also be made, including substituting, deleting, or inserting a single nucleotide at a unique,

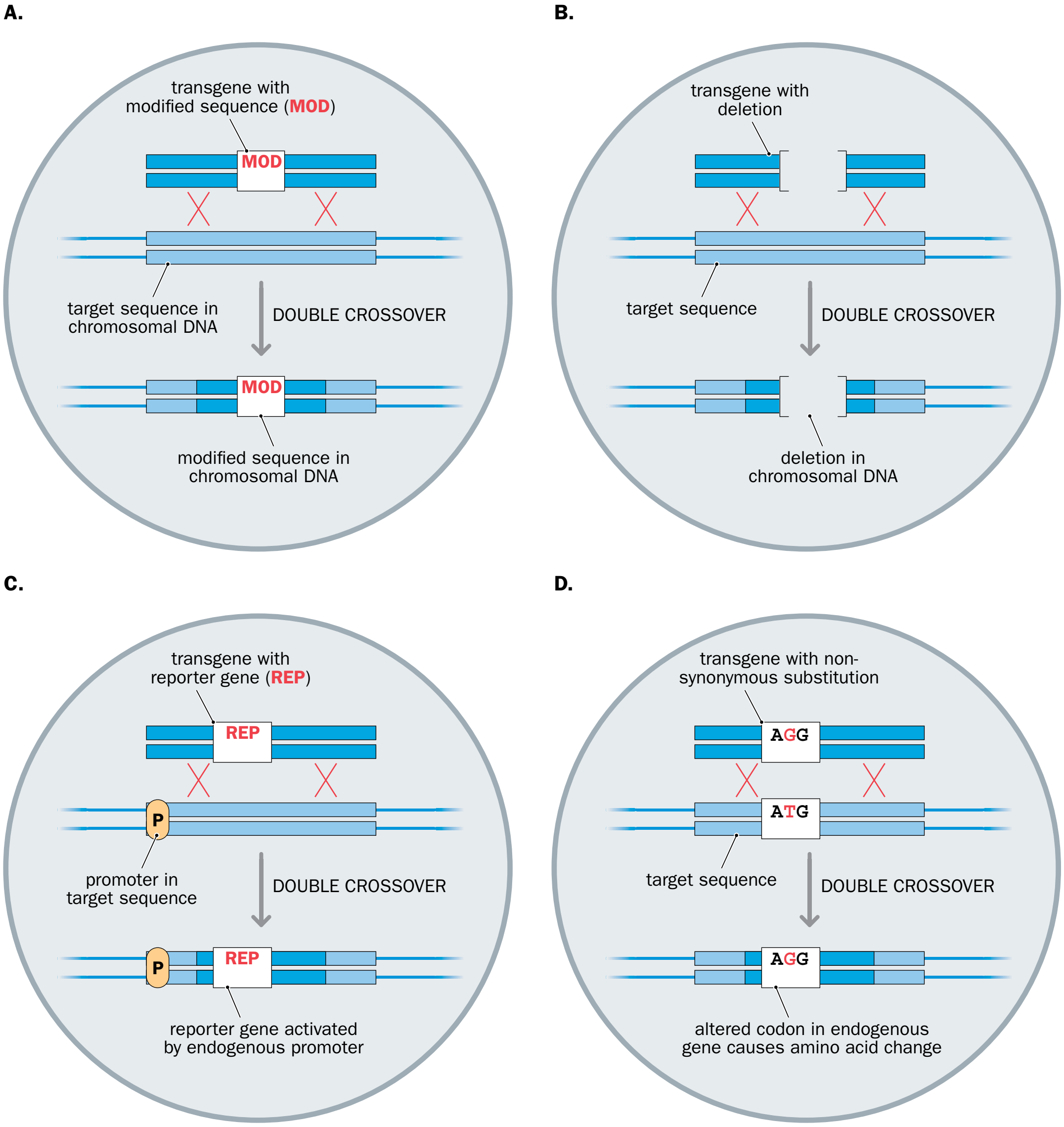

A.
transgene with modified sequence (MOD)
MOD
target sequence in chromosomal DNA
DOUBLE CROSSOVER
MOD
modified sequence in chromosomal DNA
B.
transgene with deletion
target sequence
DOUBLE CROSSOVER
deletion in chromosomal DNA
C.
transgene with reporter gene (REP)
REP
P
promoter in target sequence
DOUBLE CROSSOVER
P
REP
reporter gene activated by endogenous promoter
D.
transgene with non-synonymous substitution
AGG
ATG
target sequence
DOUBLE CROSSOVER
AGG
altered codon in endogenous gene causes amino acid change

Figure 8.12 Homologous recombination can be used to replace a specific chromosomal DNA sequence with a similar but modified transgene sequence. (**A**) The general principle. A transgene, designed to contain sequences identical to those at a desired target site on a chromosomal DNA molecule apart from a central region that has some modified or new sequence (MOD), can recombine with the endogenous target sequence, thereby introducing the desired modified sequence into the target site. (**B**) The modified sequence has a deletion. (**C**) A reporter gene may be introduced into a genomic site and expressed under the control of an endogenous promoter (P). Expression can be tracked in all cells that descend from the transfected cell, which can be important when transfected cells are introduced into the germ line to make transgenic animals (see main text). (**D**) Editing a codon in genomic DNA so that it makes a different amino acid (ATG, specifying methionine, is replaced by AGG, which specifies arginine).

pre-determined position. Or it may be desirable instead to insert some advantageous sequence. Figure 8.12B–D shows some possibilities.

The gene targeting may be designed to produce different outcomes. We may wish to inactivate the gene, for example, the first stage in an exercise where the endpoint is homozygous gene inactivation (**gene knockout**) as a way of trying to establish what the gene normally does in a cell. That may be achieved by deleting the whole gene, but because genes are often large it may be simpler to delete key sequence elements. In the case of a protein-coding gene, a general approach might be to delete one or more of the first few exons so as to produce a frameshift in the translational reading frame. Alternatively, the motivation may be to test whether a candidate pathogenic mutation really is pathogenic. A requirement would be that there is a functional assay for the gene in question.

There is a problem: homologous recombination normally occurs at extremely low frequencies in mammalian cells—about 10^4–10^5 times less frequently than random integration (when the transgene becomes inserted at sites of temporary chromosome breaks). Because the frequency depends on both the length of the homologous regions and the degree of base matching, it is usual for the transgene to have a couple of homologous sequences that are several kilobases long and show 100% identity to sequences at the target site. Even then, genuine homologous recombination events are so infrequent (compared to random integrations) that transgenes also need to carry **marker genes** that allow selection for cells in which homologous recombination has occurred. A widely used approach, *positive-negative selection*, uses a marker gene that is intended to be inserted into the target sequence (positive selection) plus a different marker gene that is located near the end of the transgene and outside of the region of homology; **Figure 8.13** demonstrates the principle.

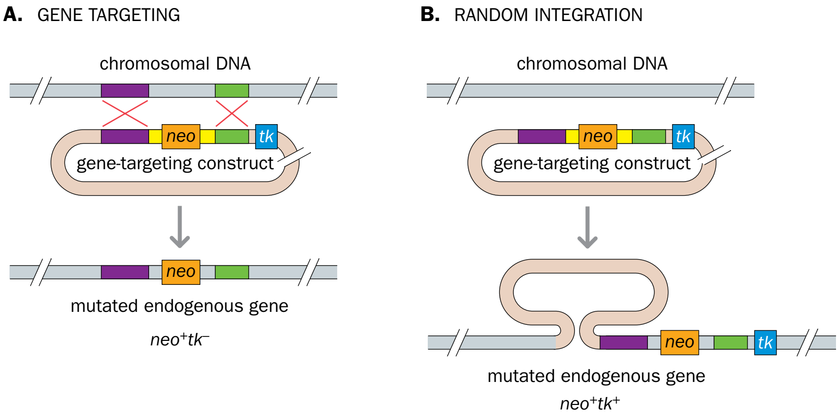


Figure 8.13 Using positive-negative selection to select cells containing a desired gene-targeting event. Here, the linearized plasmid used for gene targeting contains two marker genes: the neomycin phosphotransferase gene (*neo*), which confers resistance to neomycin and its analog G418, and the *Herpes simplex* thymidine kinase gene (*tk*). (**A**) In some gene-targeting events that occur by homologous recombination, a double crossover leads to incorporation of the *neo* gene but not of the *tk* gene. (**B**) Random integration of the gene-targeting construct at a chromosome break leads to integration of both the *neo* and *tk* genes. Suitably modified cells can be identified by selecting for *neo*$^+$*tk*$^-$ cells.

Site-specific recombination allows conditional gene inactivation and chromosome engineering

Site-specific recombination systems are naturally used by several bacteriophages and by bacteria and yeast. In such systems, the recombinase recognizes a specific recognition sequence and induces recombination between two copies of the recognition sequence. Recognition sites for recombinases can be engineered easily into transgenes or targeting vectors. The recombinase enzymes can be provided conditionally by supplying genes that are expressed under the control of regulated or inducible promoters.

Two of these site-specific recombination systems have been widely employed in genetic manipulation of mammalian cells: the Cre–*lox*P system derived from bacteriophage P1, and the FLP-FRT system derived from the 2 µm plasmid of *S. cerevisiae*. Both the Cre (**c**auses **re**combination) recombinase and the FLP (flippase) recombinase recognize specific 34 bp target sequences: respectively, the *lox*P sequence and the FRT sequence (**FLP** **r**ecombinase **t**arget).

Although their sequences are different, *lox*P and FRT have essentially the same structure: inverted 13 bp repeats separated by a central, asymmetric 8 bp spacer (**Figure 8.14A**). If two copies of the recombinase target site are located on the same DNA molecule and in the same orientation, recombination results in excision of the intervening sequence; if the *lox*P sites are in opposite orientations, recombination produces an inversion that can be very large. Site-specific recombination between two target sequences on different DNA molecules is also possible and can produce chromosome translocations (**Figure 8.14B**).

The Cre–*lox*P system has been applied in many different ways. It can allow site-specific integration of transgenes, conditional activation and inactivation of transgenes, and the deletion of unwanted marker genes. Perhaps the most important applications, however, are conditional gene inactivation (described in Section 8.6) and conditional recombination.

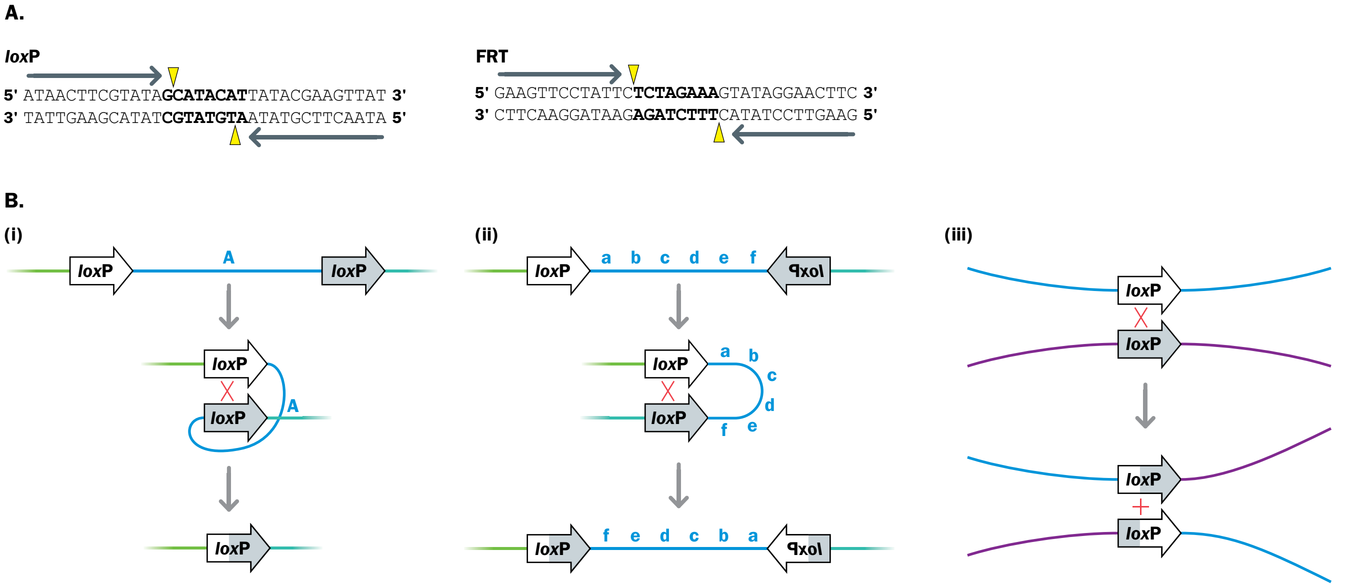

A.
loxP
5' ATAACTTCGTATAGCATACATTATACGAAGTTAT 3'
3' TATTGAAGCATATCGTATGTAATATGCTTCAATA 5'
FRT
5' GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTC 3'
3' CTTCAAGGATAAGAGATCTTTCATATCCTTGAAG 5'
B.
(i)
loxP
A
loxP
loxP
loxP
A
loxP
(ii)
loxP
a b c d e f
loxP
loxP
a b c d e f
loxP
loxP
f e d c b a
loxP
(iii)
loxP
loxP
loxP
loxP

Figure 8.14 Site-specific recombination using Cre–*lox*P and FLP-FRT. (**A**) The 34 bp recognition sequences for *lox*P and FRT. Both sequences contain an asymmetric 8 bp core sequence (shown in bold) flanked by two 13 bp inverted repeats (arrows). The phage P1 Cre recombinase recognizes the *lox*P (**l**ocus **o**f **X**-over, **P**1) sequence and the *S. cerevisiae* FLP (flippase) recombinase recognizes the FRT (**f**lippase **r**ecognition **t**arget) sequence. In each case, a recombinase monomer binds to each inverted repeat, and the two monomers form an active dimer that asymmetrically cleaves the central core sequence; breakpoints following cleavage by Cre recombinase (top) or flippase (bottom) are shown by yellow triangles. The asymmetric central sequence confers orientation: GCATACAT for *lox*P in the orientation shown here, and ATGTATGC for the opposite orientation. (**B**) Standard homologous recombination is first carried out to insert transgenes containing a *lox*P sequence into desired sites in the genome. (i) Two transgenes containing *lox*P sequences can be inserted in the same orientation flanking a sequence of interest (A), whereupon introduction of a *Cre* cDNA transgene to express Cre recombinase results in recombination between the two *lox*P sequences and deletion of sequence A. (ii) Alternatively, *lox*P sequences can be inserted in opposing orientations flanking a sequence of interest (a-b-c-d-e-f). Expressing an introduced *Cre* cDNA transgene then results in an inversion. (iii) Another possibility is to engineer a specific translocation by expressing a *Cre* cDNA transgene in cells after two *lox*P sequences have been inserted into desired target sites on different chromosomes.

8.4 GENOME EDITING USING PROGRAMMABLE SITE-SPECIFIC ENDONUCLEASES

Artificial homologous recombination, which relies on the cellular recombination machinery, is laborious and time-consuming. Alternative methods, using exogenous nucleases that can be programmed to cut genomic DNA at unique sites, originated from genetic engineering studies in the mid-1990s. The aim was to create artificial site-specific endonucleases containing a modular DNA-binding domain joined to a DNA-cleaving domain. The DNA-binding domain would be designed to act as a **protein guide sequence**, one that binds to a desired target DNA sequence.

Initially, the guide sequences were composed of DNA-binding elements from zinc finger transcription factors. They could position a coupled DNA-cleaving domain (from a bacterial endonuclease) at a desired target site in the genome. Thereafter, other types of DNA-binding protein elements were used as guide sequences. Subsequently, methods were developed that used RNA guide sequences instead of protein guide sequences. They have the huge advantage of being very much easier to carry out.

Exploiting repair of double-strand breaks in DNA

In genome editing with programmable endonucleases, the first, crucially important aim is to make a double-strand break in a pre-determined target DNA sequence in a population of desired cells. We describe below how that is carried out. But first we explain the ultimate aim of the genome editing procedures, which is to exploit how cells carry out repair of the unnatural double-strand break, and how to identify cells in which the DNA repair has resulted in a desirable sequence change at the target site. The two major pathways used to repair unnatural double-strand breaks are listed below (we provide details when we describe the mechanics of DNA repair in Section 11.2).

- **Nonhomologous end-joining** (**NHEJ**). Unnatural double-strand breaks can be fatal for cells, and all cells have an emergency repair mechanism in which the priority is to quickly join the two broken ends before the DNA fragments drift apart.

The process is error-prone, however, and joining of the two fragments frequently occurs with loss or gain of a small number of nucleotides.

- **Homologous recombination** (**HR**). This option is most readily available to dividing cells after S phase, when the DNA has duplicated to form closely associated sister chromatids. The cells can make flawless repairs of a double-strand break on a chromatid by using an invading DNA strand from the intact sister chromatid to act as a template DNA for new DNA synthesis (see **Figure 11.6**).

In mammalian cells, NHEJ is the more commonly used method of repairing double-strand breaks, and genome editing strategies often rely on natural errors made during NHEJ-based DNA repair (**Figure 8.15A**). In these cases, the double-strand break can be designed to occur in an early coding exon of a specific target gene in a population of cells. The cells are screened to identify those in which the DNA repair has introduced short deletions or insertions that cause a frameshift in the translational reading frame.

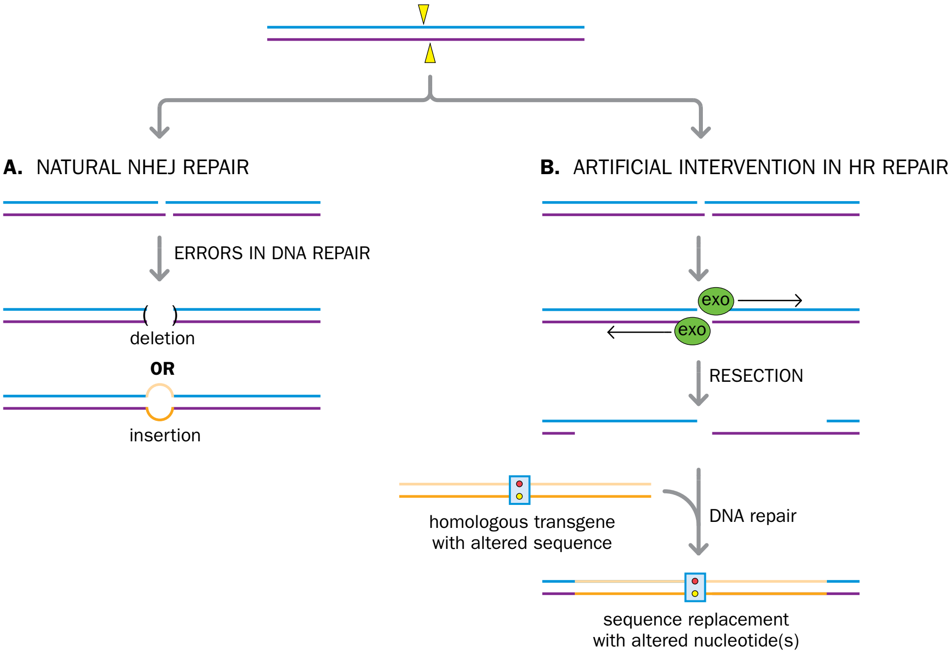


Figure 8.15 Genome editing using programmable nucleases depends on natural errors, or artificial intervention, during repair of an artificial double-strand break in DNA. Programmable nucleases are designed to be able to make a double-strand break at a unique target DNA sequence in the genome (top). (**A**) The break is often repaired by nonhomologous end-joining (NHEJ), which is prone to making errors, notably small deletions or insertions that may be desirable if the intention is to inactivate a gene. (**B**) The break may alternatively be repaired using homologous recombination (HR), and in this case an additional transgene can provide an altered sequence that is incorporated during DNA repair. The repair begins with a 5′ → 3′ exonuclease (exo) that trims back the broken ends (resection). That leaves room for invasion by a donor DNA strand (provided by a homologous DNA sequence) that acts as a template for new DNA synthesis. As a result of the repair, the original sequence is replaced by a copy of the donor DNA (see **Figure 11.6** for the mechanism). By providing a transgene with a homologous sequence but carrying an altered sequence, the repaired sequence at the target site is a faithful replica of the donor sequence. The example here shows introduction of a single nucleotide change, but much more extensive changes can be made.

Alternatively, the genome editing strategies exploit homologous recombination-mediated repair (sometime called *homology-directed repair*). Although artificial homologous recombination normally occurs at very low frequencies, a double-strand break in DNA massively increases the frequency of homologous recombination, by a factor of as much as 10^4 or more. Normally, the invading template strand whose sequence is copied would be from a natural homologous sequence (usually from a sister chromatid). However, the repair can be manipulated by providing an artificial donor sequence—a double-stranded plasmid transgene or a single-stranded oligonucleotide—that is designed to be highly homologous to the target sequence but containing some desired sequence change. By copying this artificial donor sequence during DNA repair, the original target sequence is replaced by the sequence of the donor strand, introducing the desired sequence change (**Figure 8.15B**).

Genome editing using site-specific endonucleases containing a DNA-cleavage domain and a modular DNA-binding domain

A restriction endonuclease works by recognizing a specific short sequence in a DNA molecule, and then cleaving it. In most restriction enzymes, the DNA-binding and DNA-cleavage activities are coincident, mapping to the same part of the protein. But some enzymes, such as *Fok*I, a type IIS restriction endonuclease, have *separable* DNA-binding and DNA-cleavage domains. *Fok*I works as a dimer, cleaving DNA at a short distance from a nonpalindromic and asymmetric recognition sequence.

The two methods described below take advantage of *Fok*I's properties: an isolated *Fok*I DNA-cleavage domain cuts DNA randomly, but joining it to an artificial protein guide sequence that will recognize and bind a long, specific DNA sequence results in a site-specific endonuclease. A significant amount of genetic engineering is required. To make the guide sequence, a series of coding DNA sequences for modular DNA-binding domains with known sequence specificity must be joined together. The assembled guide sequence must then be ligated to a coding sequence for a *Fok*I DNA-cleavage domain. The resulting DNA is incorporated into a transgene and expressed within a cell to make a site-specific endonuclease with a target-specific protein guide sequence covalently joined to a *Fok*I DNA-cleavage domain. A pair of hybrid endonucleases is used to make a double-strand break at the target site (**Figure 8.16A**).

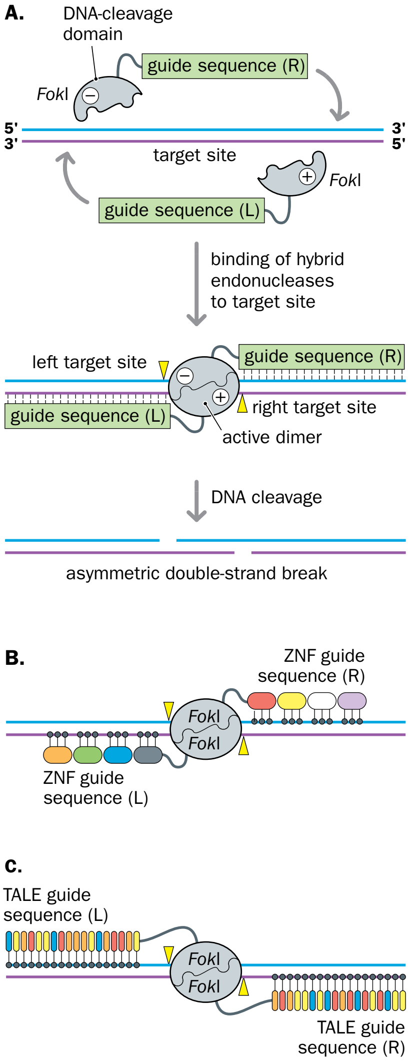


Figure 8.16 Genome editing using programmable nucleases containing zinc finger and TALE DNA-binding modules. (**A**) The aim is to express transgenes to make a pair of site-specific endonucleases, each containing a monomeric *Fok*I DNA-cleavage domain plus a protein guide sequence, an unrelated DNA-binding domain designed to recognize a specific nucleotide sequence. The *Fok*I enzyme naturally works in bacterial cells as a dimer, using interacting DNA-cleavage domain monomers to make an asymmetric double-strand break. Here, the pair of site-specific endonucleases has mutated *Fok*I DNA-cleavage domains (labeled + and –). They are designed to work as heterodimers, and co-operate to make an asymmetric double-strand break at a defined chosen target site, with the DNA-binding domains designed to bind to a sequence on the left (L) part or right (R) part of the target sequence. (**B**) The site-specific endonuclease has guide sequences constructed by combining zinc finger (ZNF) modules, each recognizing a specific trinucleotide sequence. In this example, the target sequence specificity is based on a 24-nucleotide sequence (12 nucleotides on the left and 12 nucleotides on the right, shown as filled black circles). (**C**) The site-specific endonuclease has guide sequences constructed by combining TALE (transcription activator-like effector) modules, each recognizing a single nucleotide.

Zinc finger nucleases

Here, the protein guide sequence is made by using genetic engineering to link together a series of C2H2 (Cys_2/His_2) zinc fingers, the most common DNA-binding motifs in mammalian transcription factors. Each C2H2 zinc finger has about 23 amino acids (the name comes from its finger shape and the role of a central Zn atom that coordinates two cysteine and two histidine residues—see **Box 3.1** for the structure). Individual zinc fingers in C2H2 transcription factors are separated by 7–8 amino acids, and each zinc finger binds to a *specific* trinucleotide sequence. By using genetic engineering, it became possible to prepare site-specific endonucleases with multiple zinc finger modules to produce a combination that can specifically bind to 9- or 12-nucleotide target sequences (**Figure 8.16B**).

There are some major downsides. First, generating individual zinc finger nucleases by combining different modules is laborious. Additionally, not all possible trinucleotides have an available zinc finger module that can bind to them, and there can be sequence-context issues when assembling a modular series of zinc fingers (the sequence specificity may not simply be a combination of the individual trinucleotide specificities). As a result, the choice of target sequences may be limited. Some highly effective zinc finger nucleases have been made, but the general difficulties described above have prompted interest in a more versatile alternative: TALENs.

TALENs (transcription activator-like effector nucleases)

Like zinc finger nucleases, TALENs are site-specific endonucleases that have a *Fok*I DNA-cleavage domain and a protein guide sequence made up of modular DNA-binding sequences. In this case, the DNA-binding sequences come from transcription factors made by certain bacteria that are plant pathogens (notably *Xanthomonas* and related genera). These proteins—conservatively named TALEs (transcription activator-like effectors)—have a DNA-binding domain consisting of a series of tandem 34-amino acid repeats, with each repeat binding to a specific type of nucleotide in DNA. The repeats have highly conserved sequences, but key differences at amino acid residues 12 and 13 dictate the specificity of nucleotide binding.

After TALE proteins were identified that could bind to each of the four nucleotides, TALE guide sequences could be assembled by genetic engineering to specifically bind to any desired sequence. TALENs can therefore be designed to make a double-strand break at any target site of interest (**Figure 8.16C**). Note that although TALENs usually cut to give 5′ overhanging ends (like zinc finger nucleases), the first of the TALE DNA-binding repeats binds to the most 5′ nucleotide of the recognition sequence (whereas the first zinc finger of zinc finger nucleases binds to the most 3′ nucleotide of the recognition sequence).

The much greater versatility of TALENs has meant that they have become the most popular way of carrying out genome editing with a hybrid endonuclease. They also show a high degree of sequence specificity. However, the genetic engineering required to produce TALENs to bind to target sites is still quite laborious.

Genome editing using RNA-guided endonucleases in the CRISPR-Cas system

Genome editing using site-specific endonucleases is greatly disadvantaged by the need to genetically engineer a new endonuclease for each target sequence. However, an alternative method—genome editing using RNA-guided endonucleases—has recently been developed and has rapidly become the method of choice. Its big advantage is that there is no need for complex genetic engineering: a single endonuclease can be used for all applications, and designing RNAs with target-specific guide sequences is comparatively quick and simple. Developed very recently, the method is quickly transforming genome engineering and has huge potential.

The natural function of CRISPR-Cas systems

The CRISPR-Cas system is a type of prokaryotic adaptive immune system used by the great majority of archaea and many bacteria as a defense system against invading viruses and plasmids. Cells have one or more CRISPR loci plus related genes forming a Cas (CRISPR-associated) operon. The central components are short, sequence-specific RNAs that detect foreign nucleic acids and an endonuclease that makes a double-strand break in the genome of the viral/plasmid invader, leading to its destruction. According to the components, there are at least six types of CRISPR-Cas system, but the simple type II system, notably that of *Streptococcus pyogenes*, is the one that has been most exploited in genome editing. The final stage of the defense system—involving recognition and cleavage of invading viruses and plasmids—is known as *interference*, and is preceded by two preparation stages, as listed below.

- *Adaptation* (*spacer acquisition*). The prokaryote captures short DNA segments (20–50 bp) known as "protospacers" from invading viruses and plasmids. It then inserts them into its genome as "spacer" sequences between copies of a similarly sized repeat sequence at certain loci. Containing clustered repeats interspersed with spacer sequences from viruses and plasmids, these loci are called CRISPR loci (CRISPR = **c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats). Individual spacer sequences stored in CRISPR loci act as a type of memory of a previous viral or plasmid invasion, and they allow the host cell to respond effectively to a subsequent invasion by the same type of invader.
- *Expression and maturation*. Transcription of a CRISPR locus produces a precursor RNA with sequences corresponding to each repeat unit and spacer. Thereafter, RNA cleavage and maturation produce short CRISPR RNAs (crRNAs) that have a single 5′ transcribed spacer sequence (guide sequence) plus a 3′ transcribed repeat sequence. Other components needed for interference are transcribed from genes in the Cas operon (**Figure 8.17A**).
- *Interference*. A single crRNA, with its transcribed spacer sequence, acts as a **guide RNA**, recruiting the endonuclease to cleave a viral or plasmid DNA with the same type of spacer sequence and separated by one nucleotide from a short *protospacer-associated motif (PAM)*, such as NGG in the case of the *S. pyogenes* Cas9 system. In the type II CRISPR-Cas system, a crRNA recruits the Cas9 endonuclease with the help of a go-between, a *trans*-activating RNA (tracRNA) that can bind both crRNA and Cas9; the resulting complex is directed to its target by base pairing between the guide sequence of the crRNA and the target site (**Figure 8.17B**).

In most CRISPR-Cas systems, a Cas1–Cas2 protein dimer is important in acquiring spacers and inserting new spacers into a CRISPR locus (at the proximal end, with duplication of one repeat). The spacers appear to be acquired from degraded DNA intermediates in the cell that arise during repair of double-strand breaks. Accidental acquisition of spacers from "self" DNA (the cell's genome) could be detrimental if it results in degradation of self DNA by the CRISPR-Cas system, but the possibility of autoimmunity is minimized by a mechanism that distinguishes self DNA from nonself DNA (viruses and plasmids), as described by Amitai & Sorek (2016) (PMID 26751509) under Further Reading.

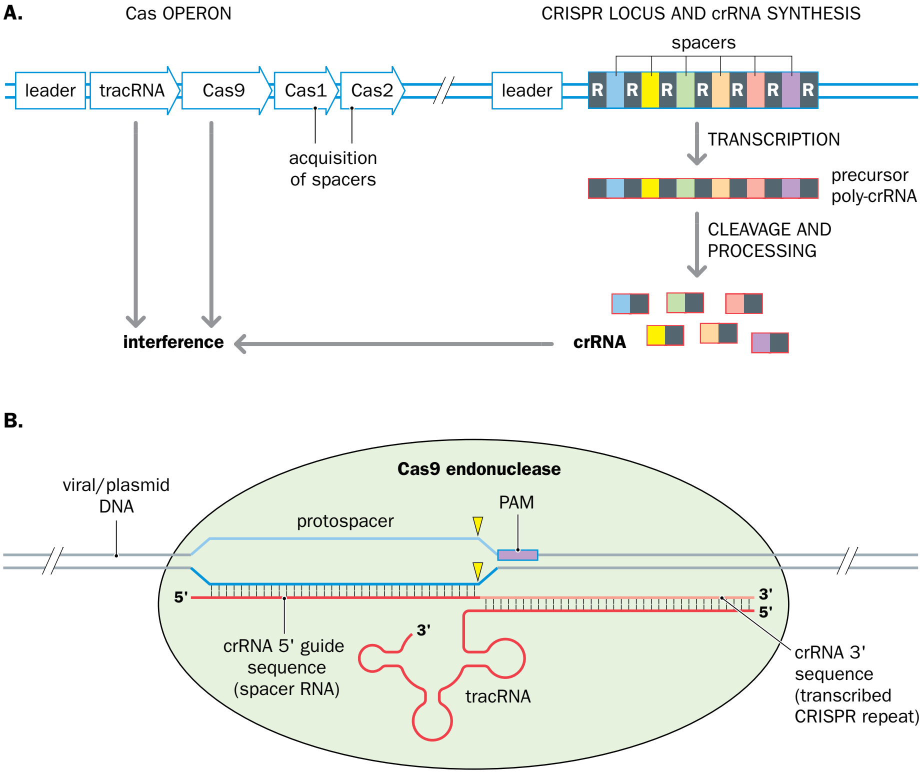


Figure 8.17 Components of the CRISPR-Cas system: a prokaryotic adaptive immune system based on RNA-guided endonucleases. The example here is a generalized type II system using a Cas9 nuclease. (**A**) A CRISPR locus has multiple copies of a specific host-cell repeat sequence (R) with interspersed DNA sequences ("spacers") captured from the genomes of previously invading viruses or plasmids. Transcription and processing of a CRISPR locus produces a series of short CRISPR RNAs (crRNAs), each containing one transcribed spacer sequence at its 5′ end (which will act as a *guide sequence*) plus an adjoining sequence transcribed from a neighboring repeat. (**B**) The interference mechanism. Maturation of each crRNA requires a *trans*-activating RNA (tracRNA) that will act as a common bridge to link a crRNA to the Cas9 endonuclease: the tracRNA 5′ sequence hybridizes to the transcribed CRISPR repeat sequence at the 3′ end of the crRNA, and the tracRNA 3′ sequence has binding sites for the Cas9 endonuclease. The resulting crRNA–tracRNA–Cas9 complex is guided by the 5′ transcribed spacer sequence of the crRNA (guide sequence) to hybridize to a complementary protospacer DNA sequence (of an invading virus/plasmid of the same type as the one from which the spacer was captured). Binding occurs just upstream of a short protospacer-associated motif (PAM) in the virus/plasmid DNA (see text). Thereafter, the Cas9 nuclease cleaves both DNA strands (cleavage sites shown by yellow triangles), leading to degradation of the virus/plasmid DNA.

Genome editing using CRISPR-Cas

The *S. pyogenes* type II CRISPR-Cas system was the first to be modified for genome editing purposes, and has been extensively used because of its simplicity. Transgenes express the Cas9 nuclease and an artificial hybrid guide RNA with features of both a crRNA and a tracRNA. At its 5′ end, the hybrid RNA has a guide sequence ~20 nucleotides long designed to hybridize to a sequence at the target site adjoining a suitable protospacer-associated motif (PAM). At its 3′ end it has a Cas9-binding sequence from the tracRNA sequence, allowing it to recruit the Cas9 nuclease and transport it to the target site. Once deposited, Cas9 cleaves both DNA strands of the target sequence, using different DNA-cleavage domains for the two strands (**Figure 8.18A**).

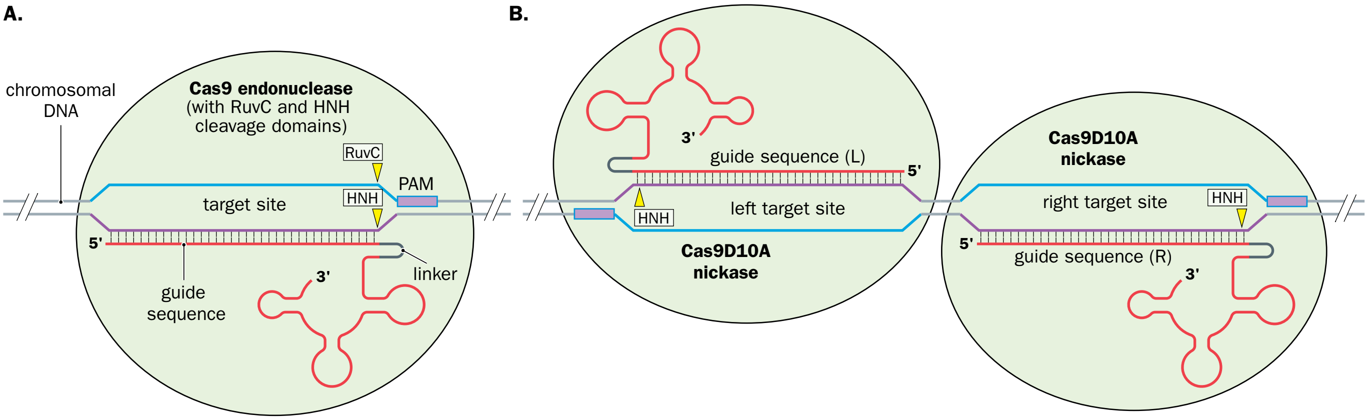


Figure 8.18 CRISPR-Cas genome editing uses an RNA-guided endonuclease to create a site-specific double-strand break in a genome. (**A**) The basic method requires transgenes to express a Cas9 endonuclease (or equivalent) plus a single RNA designed to have a specific ~20-nucleotide guide sequence at its 5′ end and binding sites for the Cas9 endonuclease at its 3′ end (similar to the crRNA–tracRNA combination in natural systems—see **Figure 8.17**). The guide sequence is designed to hybridize to a genome sequence that has a closely flanking protospacer-associated motif (PAM) specific for the endonuclease (NGG in the case of Cas9—see text). The Cas9 endonuclease has two cleavage domains, an N-terminal domain (RuvC) and a centrally located domain (HNH), that are responsible for making cuts on the individual strands, as indicated (yellow triangles indicate cleavage sites). (**B**) Greater specificity is possible using two guide sequences and a pair of mutated Cas9 endonucleases ("nickases") that each cleave just one strand. When Cas9-coding DNA is mutated to produce an aspartate-to-alanine change at amino acid position 10 (D10A), the RuvC cleavage domain is inactivated, but the HNH cleavage domain is able to cut one DNA strand. The pair of guide sequences are selected to bind to opposing DNA strands at neighboring positions (left and right target sites, just like when using zinc finger nucleases and TALENs).

The CRISPR-Cas9 genome editing system is both simple and quick but the target specificity is not so high as that of TALENs. Correct base pairing at nucleotides near the 5′ end of the guide sequence does not seem to be important but the *seed sequence*, the sequence closest to the PAM, is very important and correct base pairing here is critically important. As a result, the effective region for base pairing may be substantially less than 20 nucleotides (simply extending the length of a single guide sequence does not increase the sequence specificity).

Various modifications have been made to reduce the chances of *off-target effects* (where double-strand breaks occur elsewhere in the genome in addition to the desired target site). One way is to mutate one of the two cleavage domains of the Cas9 nuclease so that the modifed enzyme becomes a *nickase*; that is, it cuts a single DNA strand instead of cutting both strands. For example, the RuvC cleavage domain located at the N-terminal region of the Cas9 nuclease can be made catalytically inactive by a D10A substitution, and an H840A substitution inactivates the central HNH cleavage domain. A pair of guide RNAs and a pair of nickases can be used to cleave the two DNA strands at neighboring regions of a target DNA sequence to increase the specificity (**Figure 8.18B**).

8.5 GENE SILENCING

Inactivating a specific, pre-determined gene within cells is the most popular approach to working out how the gene functions, but gene knockout methods used to be rather time-consuming. An alternative approach is to carry out **gene silencing** (sometimes called a *gene knockdown*) by suppressing the activity of a pre-determined gene at the RNA level. That is, the RNA transcripts of the gene are targeted instead of the gene itself. And, depending on the approach, the aim is to block transcripts in some way to prevent them being expressed, or to cleave the transcripts, causing them to degrade.

Antisense technology as a way of suppressing the expression of a pre-determined target gene

Developed in the 1980s, antisense technology was the first general approach to use the specificity of base pairing to selectively inhibit the expression of a pre-determined gene. Initially that meant providing specific antisense RNA molecules that could hybridize to the transcripts of a pre-determined gene of interest to block its function. That is most easily done by transfecting a transgene that has been engineered so that the usual sense strand is used instead as a template strand for RNA synthesis. The resulting **antisense RNA** can hybridize to RNA transcripts from the corresponding chromosomal gene and prevent them carrying out their functions. Subsequently, oligodeoxynucleotides came to be used instead of antisense RNA because of their greater stability.

Antisense technology initially was very much a hit-or-miss affair depending on which gene was targeted. However, the design of chemically-modified, highly stable antisense oligonucleotides has resulted in consistent inhibition of gene expression. Morpholino oligonucleotides have a particularly robust structure (see **Figure 8.5**) and are widely used to make antisense oligonucleotides to knock down the expression of specific genes in various vertebrate models, notably zebrafish embryos. That is possible because the morpholino oligonucleotide hybridizes to mRNAs from the gene of interest and blocks them from being translated. For cultured cells, an alternative technology has rapidly transformed the field, as described in the next subsection.

Gene silencing using RNA interference provides a rapid way of evaluating gene function in cultured cells

The most widely used method to assess gene function in cultured mammalian cells exploits **RNA interference** (**RNAi**), a natural cellular pathway where the formation of double-stranded RNA induces specific gene inactivation. RNA interference is thought to have evolved to protect cells against the accumulation of potentially dangerous nucleic acid sequences, notably viruses (**Box 8.2**).

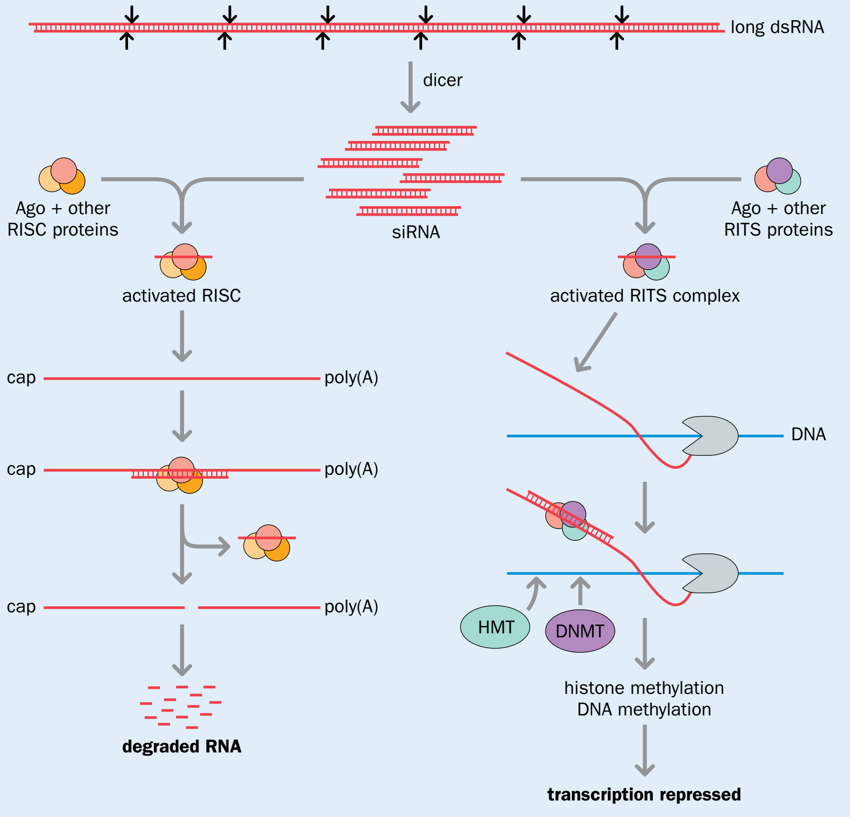
To knock down a pre-determined gene transcript, a suitably specific double-stranded RNA needs to be provided. In some animal systems, such as the nematode *Caenorhabditis elegans*, long double-stranded RNA (dsRNA) is experimentally used to inactivate genes. After it has been transferred into the worm's cells, the dsRNA is chopped by an endogenous cytoplasmic ribonuclease called dicer into uniformly small pieces of dsRNA (close to 20 bp in length and with 3′ dinucleotide overhangs) called

BOX 8.2 RNA INTERFERENCE AS A NATURAL CELL DEFENSE MECHANISM

RNA interference (*RNAi*) is an evolutionarily ancient mechanism that is used in animals, plants, and even single-celled fungi to protect cells against viruses and transposable elements. Both viruses and active transposable elements can produce long double-stranded RNA at least transiently during their life cycles. Long double-stranded RNA is not normally found in cells and, for many organisms, it triggers an RNA interference pathway. A cytoplasmic endoribonuclease called dicer cuts the long RNA into a series of short double-stranded RNA pieces known as short interfering RNA (siRNA). The siRNA produced is on average 21 bp long, but asymmetric cutting produces two-nucleotide overhangs at their 3′ ends (Figure 1).

The siRNA duplexes are bound by different complexes that contain an argonaute-type endoribonuclease (Ago) and some other proteins. Thereafter, the two RNA strands are unwound and one of the RNA strands is degraded by argonaute, leaving a single-stranded RNA bound to the argonaute complex. The argonaute complex is now activated: the single-stranded siRNA acts as a **guide RNA**, guiding the argonaute complex to its target RNA by base-pairing with complementary RNA sequences in the cells. Different argonaute complexes work in different ways, as listed below.

- ***RNA-induced silencing complex (RISC)***—see Figure 1 bottom-left panel. In this case, after the single-stranded guide RNA binds to a complementary long single-stranded RNA, the argonaute enzyme will cleave the RNA, causing it to be degraded. Viral and transposon RNA can be inactivated in this way.
- ***RNA-induced transcriptional silencing (RITS)*** complex—see Figure 1 bottom-right panel. Here, the single-stranded guide RNA binds to complementary RNA transcripts as they emerge during transcription by RNA polymerase II. This allows the RITS complex to position itself on a specific part of the genome and then attract proteins such as histone methyltransferases (HMT) and sometimes DNA methyltransferases (DNMT), which covalently modify histones in the immediate region. This process eventually causes heterochromatin formation and spreading; in some cases, the RITS complex can induce DNA methylation. As a result, gene expression can be silenced over long periods to limit, for example, the activities of transposons.



Box 8.2 Figure 1 RNA interference. Long double-stranded (ds) RNA is cleaved by cytoplasmic dicer to give siRNA. siRNA duplexes are bound by argonaute (Ago) complexes that unwind the duplex and degrade one strand to give an activated complex with a single siRNA that then works as a guide RNA. By base-pairing with complementary RNA sequences, the siRNA guides argonaute complexes to recognize target sequences. Activated RNA-induced silencing complexes (RISCs) cleave any RNA strand that is complementary to their bound siRNA. The cleaved RNA is rapidly degraded. Activated RNA-induced transcriptional silencing (RITS) complexes use their siRNA to bind to any newly synthesized complementary RNA being synthesized on the chromosomal DNA. They then attract proteins, such as histone methyltransferases (HMT) and sometimes DNA methyltransferases (DNMT), that can modify the chromatin to repress transcription.

short interfering RNA (**siRNA**) (see pathway 1 in Figure 8.19). The siRNA associates with a multisubunit protein complex, the RNA-induced silencing complex (RISC), and the duplex siRNA is unwound and one strand is degraded by a RISC ribonuclease called argonaute. The RISC with one siRNA strand attached to it is now activated whereupon the siRNA strand will guide it to bind RNA transcripts containing complementary sequences. An RNA transcript that binds to an antisense siRNA is then targeted for destruction by the argonaute ribonuclease.

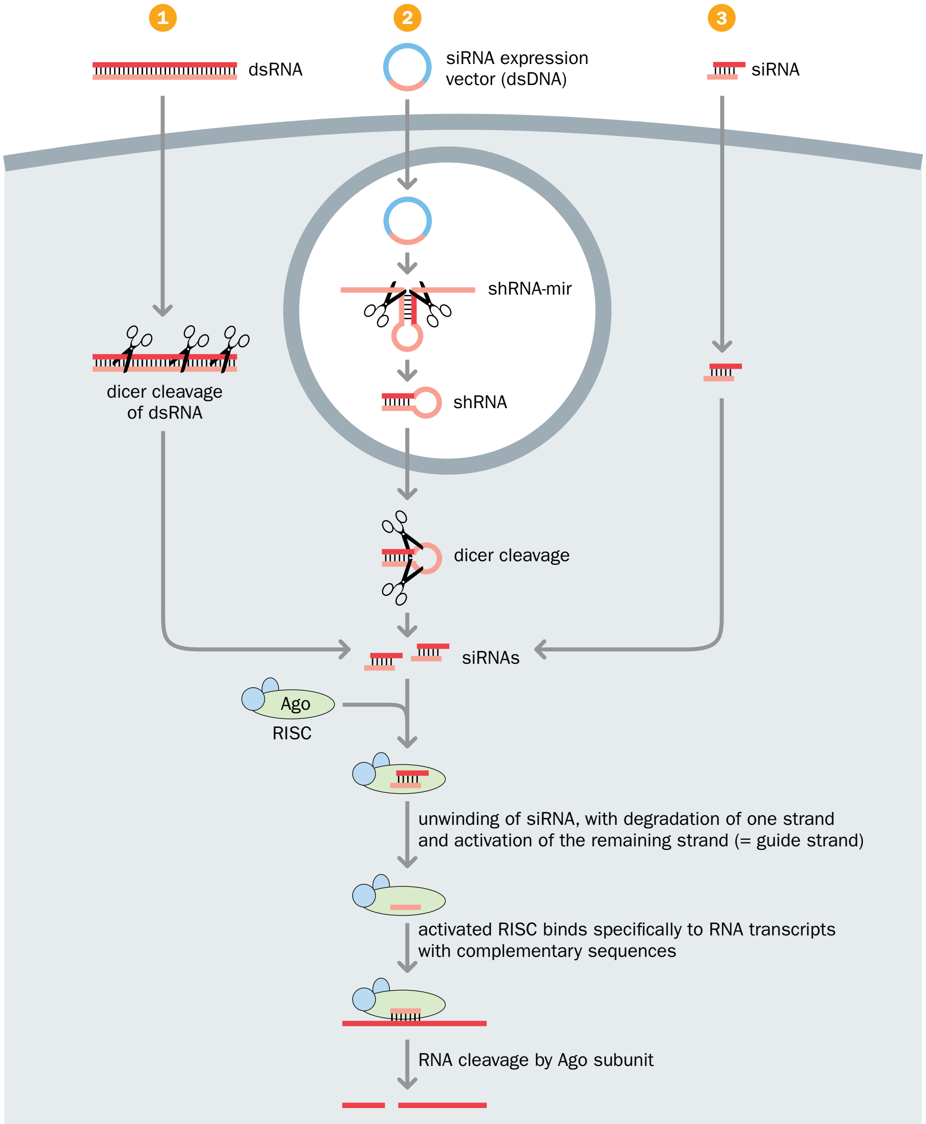

1
dsRNA
2
siRNA expression
vector (dsDNA)
3
siRNA
shRNA-mir
dicer cleavage
of dsRNA
shRNA
dicer cleavage
siRNAs
Ago
RISC
unwinding of siRNA, with degradation of one strand
and activation of the remaining strand (= guide strand)
activated RISC binds specifically to RNA transcripts
with complementary sequences
RNA cleavage by Ago subunit

Figure 8.19 Three routes for experimentally inducing RNA interference (RNAi) in animal cells. To induce RNAi, cells are provided with double-stranded RNA (dsRNA) whose nucleotide sequence is complementary to a sequence in a pre-determined target gene. The RNA is either transfected into cells (pathways 1 and 3) or produced from a transfected expression vector (pathway 2). Pathway 1 shows the use of long dsRNA (typically longer than 200 bp) that can be used in invertebrate systems such as that in *C. elegans*. In mammalian systems, however, long dsRNA triggers *nonspecific* RNA cleavage (and hence alternative pathways 2 and 3 need to be used in mammalian cells). Short dsRNA can be produced from an expression vector (pathway 2) transcribed in the nucleus to give a single RNA molecule with inverted repeats. The repeats base-pair to each other to produce a short *hairpin RNA* that resembles initial miRNA transcripts (shRNA-mir). The 5′ and 3′ tails of this primary transcript are then cleaved in the nucleus to give a more compact, short hairpin RNA (shRNA) that is exported to the cytoplasm where it is processed by the dicer endonuclease to give double-stranded short interfering RNA (siRNA). Alternatively, two short oligoribonucleotides are chemically synthesized to have complementary sequences and are allowed to anneal to form double-stranded siRNA (pathway 3). Whichever pathway is used, the end result is to provide an siRNA duplex that is bound by the argonaute endoribonuclease subunit (Ago) of the RNA-induced silencing complex (RISC). Unwinding of the siRNA and degradation of one of its strands activates the RISC. The activated RISC binds to any mRNAs having a complementary sequence, and the associated argonaute subunit cleaves the bound mRNAs.

Unlike in *C. elegans*, addition of long double-stranded RNA (dsRNA) to mammalian cells does not inhibit the expression of specific genes. Instead, long dsRNA induces a general antiviral pathway in mammalian cells that produces global, nonspecific gene silencing. It does this by activating the PKR protein kinase (which is also induced by interferon). Two important PKR targets are the eukaryotic translation initiation factor 2 (eIF2α) and 2′,5′-oligoadenylate synthetase (2′,5′-AS). Phosphorylated eIF2α causes a generalized inhibition of translation, and activated 2′,5′-AS induces a ribonuclease that causes nonspecific mRNA degradation.

As an alternative to adding long dsRNA to mammalian cells, two approaches are used (see **Figure 8.19**). In one method, transgenes are added that mimic the way miRNA works. Another widely used method involves chemically synthesizing two short complementary oligoribonucleotides and allowing them to anneal to form a synthetic equivalent of siRNA. Often, the synthesized oligonucleotides are chosen to be about 20–30 nucleotides long and the sequences are chosen to form a duplex that has two deoxythymidine bases added as 3′ overhangs (duplexes with 3′ overhangs are more potent at inducing RNAi than blunt ones).

As the RNAi efficiency can vary depending on the sequence, it is usual to design three or more complementary pairs of siRNA oligonucleotides corresponding to different sequences for any transcript. The efficiency of gene knockdown is assessed by real-time PCR or, where antibodies are available for the relevant protein product, by western blotting. These kinds of gene silencing approaches, which depend on RNAi, are sometimes also known as **RNA silencing**.

8.6 GERM-LINE TRANSGENESIS AND TRANSGENIC ANIMALS

A **transgenic animal**, one in which genetic material has been artificially inserted into its cells, is usually most useful when fully transgenic (all cells contain the same transgene in the same context). Fully transgenic animals must develop from a transgenic zygote and that requires that transgenes be transferred into the germ line.

Transgenes can be inserted into the germ line by direct transfer into the zygote, gametes, and embryonic or somatic cells that eventually contribute to the germ line (**Figure 8.20**). Of the possible routes, those beginning with transfer of transgenes into fertilized oocytes and pluripotent stem cells have been particularly popular, as described in the two main subsections below, but gene transfer into gametes and germ-cell precursors has also been carried out.

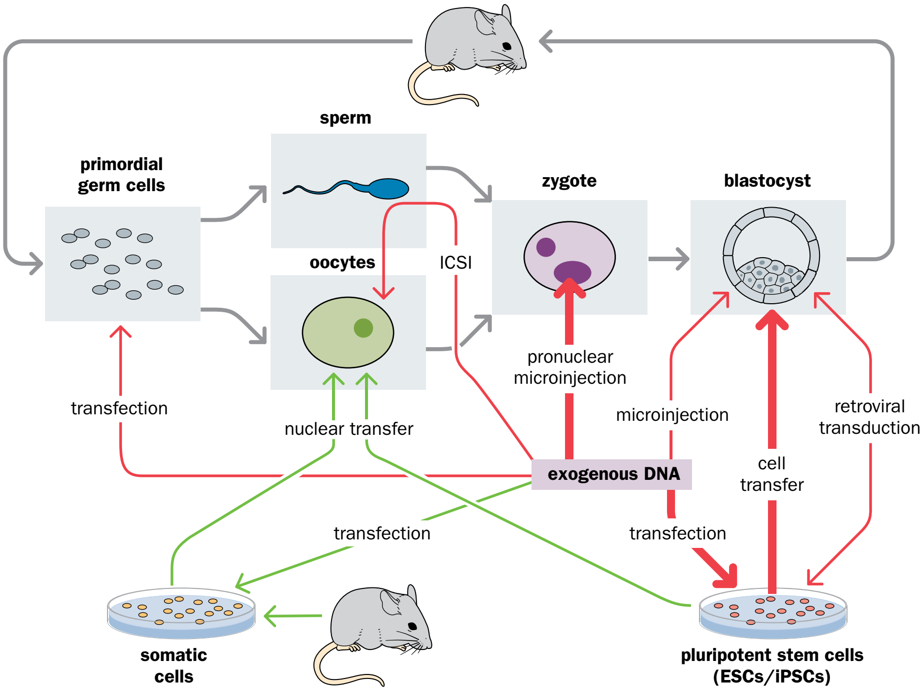


Figure 8.20 Genetically modified mice can be produced by a variety of routes. The boxes linked by gray arrows at the top show components of the mouse life cycle and represent the multiple stages at which, potentially, genetic modification can be performed. Shown at bottom are sources of donor cells for nuclear transfer procedures (process shown in green arrows). Red arrows show the input and transport of exogenous DNA for all other routes. The most widely used gene transfer methods—microinjection into the male pronucleus of the fertilized oocyte, and transfection of embryonic stem cells (ESCs) followed by injection of genetically modified ESCs into a blastocyst—are shown by thick red arrows. ICSI, intracytoplasmic sperm injection; iPSCs, induced pluripotent stem cells. For the sake of clarity, some methods described in the text are not shown.

Gene transfer into gametes and germ-cell precursors

Germ-cell precursors are a favorite target for making transgenic fruit flies. However, although mammalian primordial germ cells are relatively easy to isolate, culture, and transfect, it is difficult to persuade the modified cells to contribute to the germ line when re-introduced into a host animal.

A less conventional way of genetically modifying zygotes is to use sperm-mediated transgene delivery into unfertilized eggs. Sperm heads bind spontaneously to DNA *in vitro*, and so sperm can be used simply as vehicles for delivering DNA—the sperm genome itself is not modified. **Intracytoplasmic sperm injection** (**ICSI**), a method used in infertility treatment where sperm heads are injected into the cytoplasm of the egg, has been adapted to produce transgenic mice. To do this, sperm heads are coated with plasmid DNA before being introduced into mouse eggs. However, the transgene often fails to integrate into the genome in this method.

More recently, the injection of recombinant retroviruses into the peri-vitelline space of isolated oocytes has allowed the production of transgenic cattle and the first ever transgenic primate, a rhesus monkey named ANDi.

Pronuclear microinjection: an established method for making transgenic animals

Transgenic animal technologies came of age in the early 1980s when the first transgenic mice and transgenic fruit flies were reported. Subsequently, a wide variety of other transgenic animals—including worms, birds, frogs, fish, and many types of mammal—have been produced. For multiple reasons (explained below), transgenic mice have been the principal animal model.

Transgenic mice are often produced by injecting DNA into a newly fertilized zygote, usually by the most efficient route: microinjection of a transgene into the large male pronucleus. The transgene then randomly integrates into chromosomal DNA, usually at a single site, and often as multiple tandemly repeated copies (Figure 8.21).

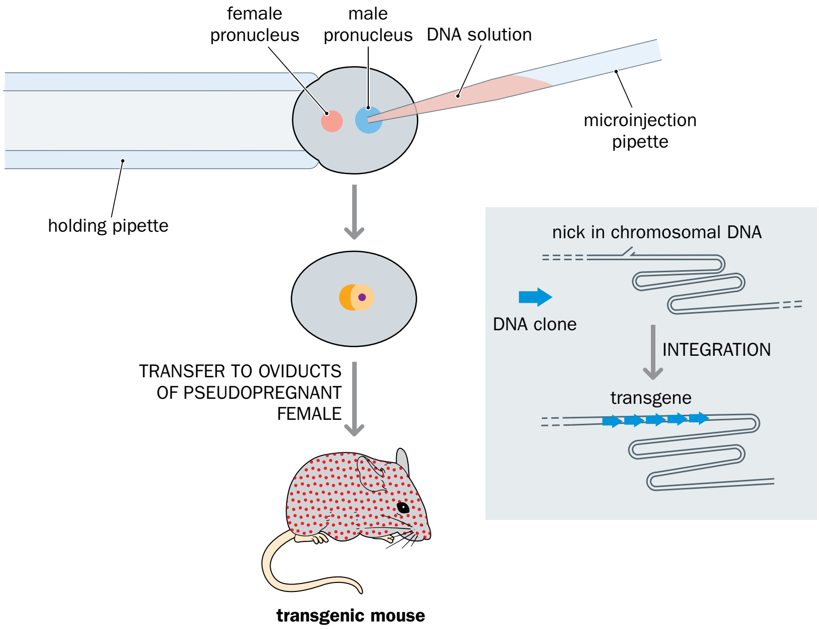


Figure 8.21 Construction of transgenic mice by pronuclear microinjection. A fertilized oocyte is held in place with a holding pipette. A microinjection pipette with a very fine point is then used to pierce first the oocyte and then the male pronucleus (which is bigger than the female pronucleus), delivering an aqueous solution of a desired DNA clone. The introduced DNA integrates at a nick (single-stranded DNA break) that has occurred randomly in the chromosomal DNA. The integrated transgene usually consists of multiple head-to-tail copies of the DNA clone. Surviving oocytes are re-implanted into the oviducts of *pseudopregnant* foster females (they will have been mated with a vasectomized male to initiate physiological changes in the female that stimulate the development of the implanted embryos). DNA analysis of tail biopsies from resulting newborn mice checks for the presence of the desired DNA sequence.

If the transgene integrates into the chromosomal DNA of the zygote, the resulting mouse will be fully transgenic. However, it is more common for the DNA to integrate after one or two cell divisions, in which case the resulting mouse is a genetic mosaic that contains both transfected and nontransfected cells. Where the transfected cells contribute to the germ line, the transgene is passed to the next generation (as verified by PCR or Southern blotting or some test for transgene expression), resulting in fully transgenic mice.

It is possible to achieve germ-line transmission in up to 40% of microinjected mouse eggs. In other mammals, the transmission rate is generally much lower—often <1%—partly because of the difficulty in handling eggs, and partly due to lower survival rates.

Genetic modification of pluripotent stem cells as a route to germ-line transmission of genetically altered target genes

Genes cloned into retroviral vectors have been transferred into unselected pluripotent cells of very early embryos, notably in mice and birds. In mice, recombinant retroviruses can be used to infect pre-implantation embryos or can be injected into early post-implantation embryos. Following the stable transformation of some cells, mosaics are produced that may give rise to transgenic offspring.

A convenient and widely used alternative is to use cultured pluripotent stem cells. The first successful approach was made possible when mouse **embryonic stem cells (ESCs)** were derived from cells taken from the inner cell mass of a blastocyst (Figure 8.22A). The resulting availability of ESC lines then offered the possibility of homologous recombination-based genome editing, that is, inserting transgenes into the germ line using homologous recombination between a transgene and a target site in the genome of cultured ESCs. Thereafter, cells could be selected that contained the inserted transgene (see above and Figure 8.13 for the principle). Cells with integrated transgenes could then be injected into isolated blastocysts that would be re-implanted into suitably treated female mice. Finally, any resulting chimeric offspring could be bred to produce mutants (Figure 8.22B and C).

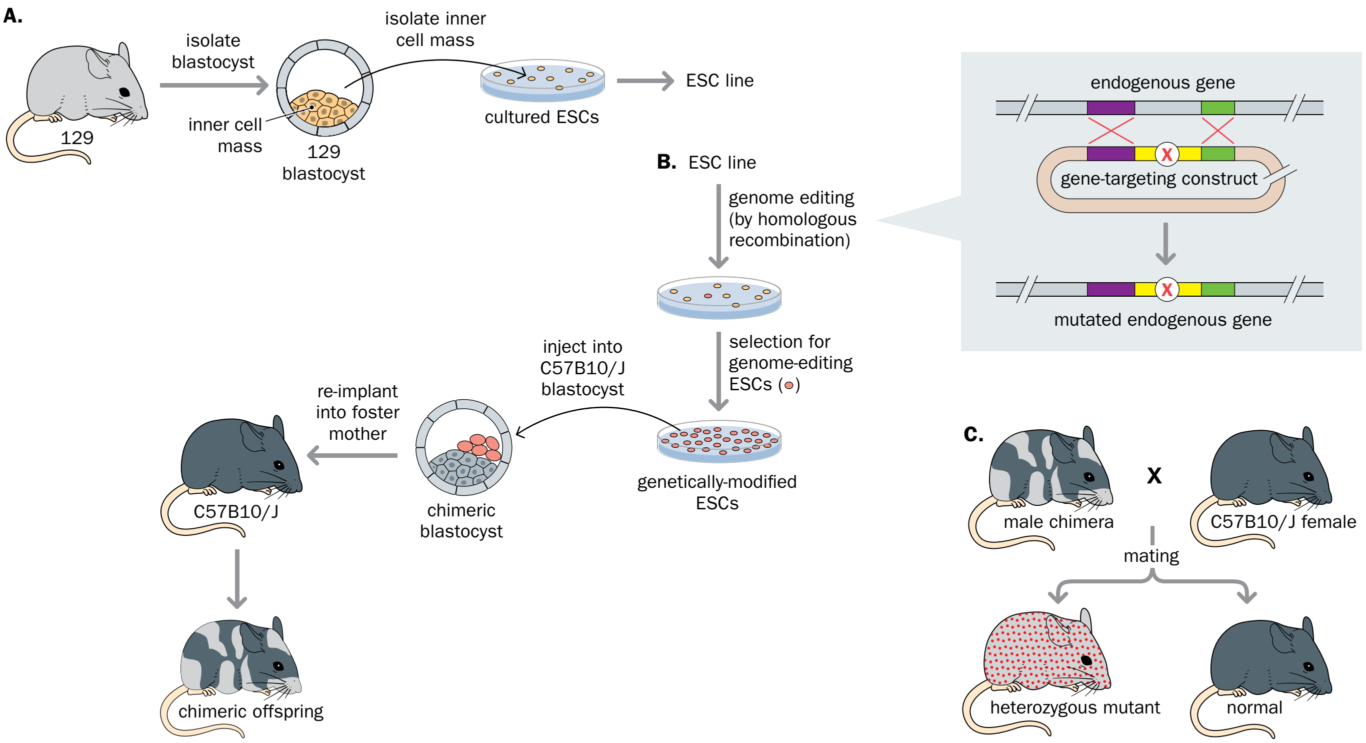


Figure 8.22 Genome editing of embryonic stem cells to introduce mutations into the mouse germ line. (**A**) Embryonic stem cell (ESC) lines have been made by excising blastocysts from the oviducts of a suitable mouse strain (such as 129). Cells from the blastocyst's inner cell mass are cultured to eventually give a pluripotent ESC line. (**B**) An ESC line can be genetically modified by transfection of a linearized recombinant plasmid containing DNA sequences that are identical to certain sequences in a predetermined endogenous gene (shown by purple and green bars), plus a distinct central region with a desired genetic modification (shown by the white circle with the red X). Homologous recombination involving a double crossover, allows sequence swapping so that the desired sequence change is incorporated into the target site in the ESC's genome. Only a very few cells will undergo homologous recombination to produce the expected modification of the targeted gene, but they can be selected for and amplified. The modified ESCs are then injected into isolated blastocysts of another mouse strain such as C57B10/J, which has black coat coloration that is recessive to the agouti color of the 129 strain; the cells are then implanted into a pseudopregnant foster mother of the same strain as the blastocyst. Subsequent development of the introduced chimeric blastocyst results in a chimeric offspring containing two populations of cells that derive from different zygotes (here, 129 and C57B10/J, normally evident by the presence of differently colored coat patches). (**C**) Backcrossing of chimeras can produce mice that are heterozygous for the genetic modification (bottom left). Subsequent interbreeding of heterozygous mutants generates homozygotes (not shown).

Certain strains of mouse ESCs have been particularly suitable for allowing gene transfer into the germ line, and have been available since the early 1980s. The *in vitro* step—growing ESCs in culture—allows sophisticated genetic manipulations to be carried out. The ESC route has, therefore, long been the primary route for inactivating or otherwise modifying specific genes in the mouse germ line, and the resulting modified mice have been invaluable for understanding gene function and making disease models.

Because isolating equivalent stem cells from other mammals proved to be extremely difficult, the mouse became, by a long distance, the premier model organism. As described in Section 4.2, recent advances have led to the identification of equivalent ESCs in some other mammalian model organisms, including the rat. Additionally, it has recently been possible to create **induced pluripotent stem cells (iPSCs)** from multiple species. They have many of the same properties as ESCs and can be used as an alternative route to transfer transgenes into the germ line.

Gene-targeting approaches to create transgenic animals with null alleles or subtle point mutations

We have illustrated above the traditional gene-targeting approach to modify the germ line of mice, involving genome editing using standard homologous recombination in ESCs. However, recent advances are transforming the picture. Now, induced pluripotent stem cells are available from many species and can substitute for ESCs. And CRISPR-Cas-based genome editing of the germ line of model organisms is becoming a popular alternative method of genome editing. Here, we now consider how different strategies to alter the germ line of model organisms are carried out in an effort to understand human gene function, and to model disease.

Making gene knockouts

A common strategy in gene targeting is to completely inactivate a pre-determined target gene, creating a *null allele*. The objective may be to investigate the function of a gene; having null alleles in all cells of the model organism will offer much greater insight into how a gene functions than simply knocking it out in cultured cells. An alternative objective may be to model a human disease caused by loss of gene function—analyzing animals that are heterozygous and homozygous for the null allele can be expected to offer insights into the disease process.

Designing a gene knockout may involve deleting a gene (if it is small) or, more commonly, deleting a small component that is crucially important to its function. For a protein-coding gene, it is common to delete one or a few exons at the 5′ end of the target gene in order to induce a shift in the translational reading frame so that an early premature termination codon will be introduced. For some genes, a standard gene knockout may not be very informative (homozygous null alleles may result in embryonic lethality and the heterozygous knockout may appear to have a normal phenotype). Often, therefore, a conditional gene knockout is needed, as described in **Box 8.3**.

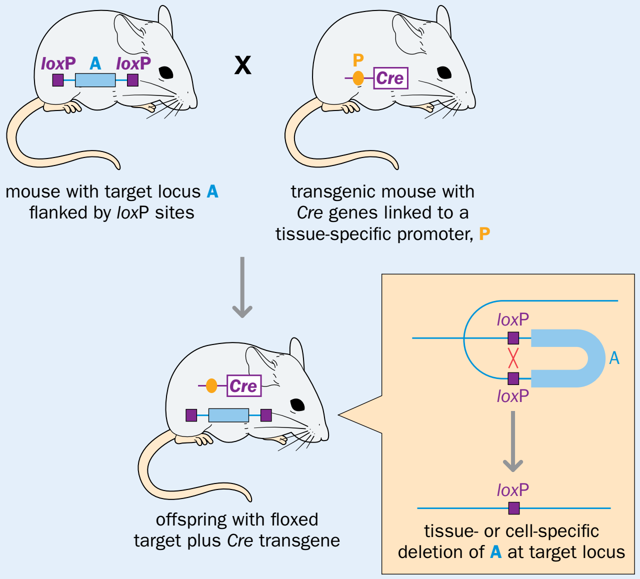
BOX 8.3 CONDITIONAL GENE INACTIVATION IN TRANSGENIC ANIMALS

Producing a full gene knockout often results in embryonic lethality. In order to study genes that are essential for the viability of an organism *conditional knockouts* are made. Here, the gene is designed to be inactivated in a selected tissue or group of cells, or at a desired developmental stage.

Conditional gene inactivation typically involves using a bacterial site-specific recombination system. Often, for example, an early exon (or exons) of the endogenous mouse target gene is replaced with a *floxed* homologous gene segment (*floxed* means that the sequence is **f**lanked by ***lox***P sequences—see **Figure 8.14**). Mice carrying the floxed target sequence are then mated with a strain of mouse that carries a *Cre* transgene under the control of a tissue-specific or developmental stage-specific promoter (**Figure 1**).

Offspring of the cross that contain both a floxed target sequence and a *Cre* transgene are identified and analyzed. According to the promoter regulating the *Cre* transgene, the Cre recombinase should be expressed only in certain cells or at certain stages of development so that offspring can be analyzed. Note, however, that tissue-specific promoters may often not be expressed exclusively in the tissue or cells of interest. For example, pancreatic-specific expression in mice often uses a promoter from the *Pdx1* gene that is, however, also expressed in the developing stomach and duodenum.

Recently, inducible *Cre* transgenic animals have become widely used. Here, Cre expression is induced using tetracycline/doxycycline, or where Cre is constitutively expressed as a fusion protein with the estrogen receptor and is activated at the protein level by tamoxifen (see **Figure 8.10** for the principle).


loxP A loxP
X
P
Cre
mouse with target locus A flanked by loxP sites
transgenic mouse with Cre genes linked to a tissue-specific promoter, P
Cre
offspring with floxed target plus Cre transgene
loxP
A
loxP
loxP
tissue- or cell-specific deletion of A at target locus

Box 8.3 Figure 1 Conditional gene inactivation using Cre–*lox*P site-specific recombination. The mouse shown at the top left has been subjected to gene targeting so as to introduce two *lox*P sequences at positions flanking a specific genomic DNA sequence, A, that may be a small gene or an exon that, if deleted, could be expected to cause gene inactivation. The floxed A sequence will be specifically deleted in the presence of Cre recombinase. To introduce Cre recombinase, a mouse is required that has previously been constructed to have a *Cre* transgene ligated to a tissue-specific promoter, P, of interest (top right). By breeding the two types of mouse it is possible to obtain offspring containing both the *lox*P-flanked target locus plus the *Cre* transgene (bottom left). Cre recombinase will be produced in the desired type of tissue and will cause recombination between the two *lox*P sequences in cells of that tissue, leading to deletion of the A sequence and tissue-specific gene inactivation.

Modern strategies for knocking out genes are often designed to also introduce a reporter gene, such as *lacZ*, or a gene that encodes an autofluorescent protein such as GFP (green fluorescent protein), RFP (red), or YFP (yellow). The targeting construct, containing a reporter gene, is usually designed to integrate at the 5′ end of the endogenous gene, and the procedure is designed to both inactivate the endogenous gene while at the same time using the endogenous gene's promoter and other expression controls to activate the reporter gene (*knocking in*). Because the expression of the knocked-in reporter gene is regulated by the endogenous target gene's regulatory sequences, it should faithfully mimic the expression of the target gene. The principle underlying a **gene knock-in** is illustrated in **Figure 8.23**; a practical example of knocking in a reporter gene while simultaneously knocking out the endogenous gene is shown in **Figure 8.24**.

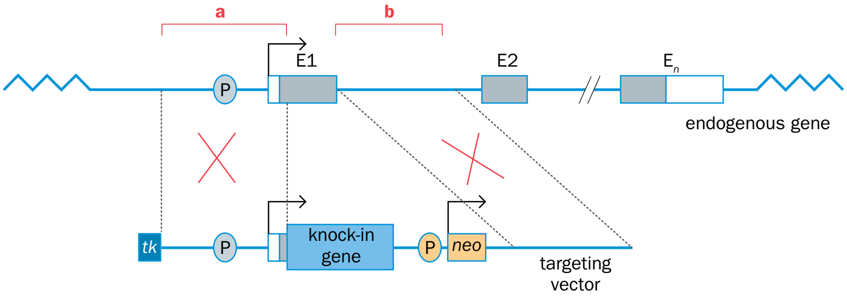


Figure 8.23 A gene knock-in replaces the activity of one chromosomal gene by that of an introduced transgene. In this example, the endogenous gene is envisaged to have multiple exons (E1, E2, … E_n); coding sequences are in filled boxes and untranslated sequences are in open boxes. The gene-targeting vector contains two sequences from the endogenous target gene: **a**, a 5′ flanking sequence including the promoter, and a very small 5′ component of exon 1 that is mostly made up of the 5′ untranslated sequence; and **b**, an internal segment that spans much of intron 1. Between these two sequences is the coding sequence of the gene to be knocked in plus a marker gene whose expression is regulated by its own promoter. Arrows indicate expression driven by upstream promoters (P). The targeting procedure results in a double crossover (X) and in this case an inactivating deletion of a large part of exon 1 that is usually designed to result in a shift in the translational reading frame. The knocked-in gene of interest comes under the control of the endogenous promoter. The *neo* marker helps identify cells with the correct gene-targeting event, and an upstream thymidine kinase (*tk*) gene marker (far left in the targeting vector) helps select against random integration as illustrated in **Figure 8.13**.

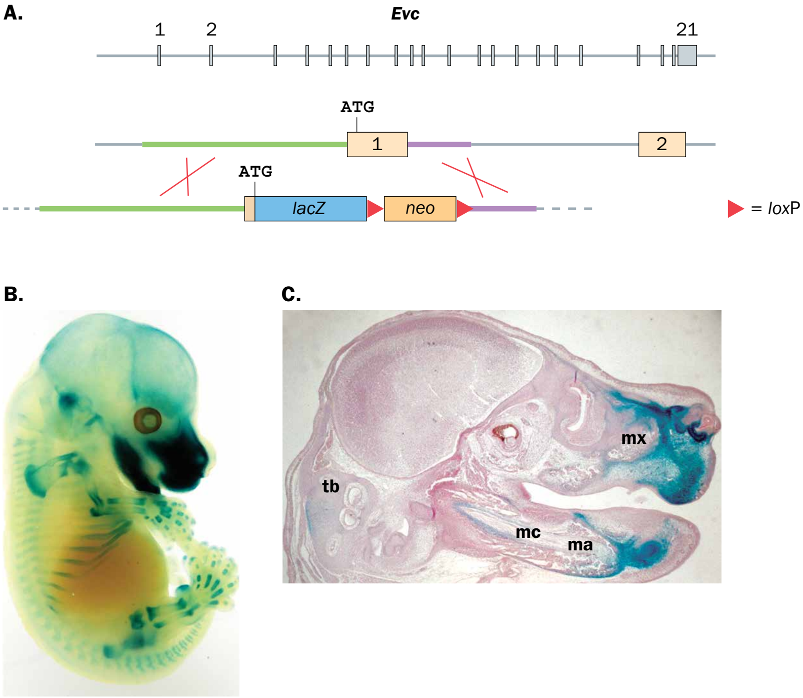


Figure 8.24 Knocking in a *lacZ* reporter transgene to simultaneously inactivate a gene and monitor its normal expression. The example illustrates inactivation of the mouse *Evc* gene, the ortholog of the Ellis–van Creveld syndrome gene, by replacement of exon 1 (which contains the initiator methionine codon) by a *lacZ* reporter gene and a *neo* marker gene. The latter is flanked by *loxP* sequences for easy removal once the selection system shows that the transgene has integrated. Flanking homology sequences include a 6 kb upstream region preceding exon 1 (green line) plus a 5′ region of exon 1 and a 1 kb region immediately following exon 1 (lilac line). Integration of the transgene resulted in inactivation of the *Evc* gene (by deleting the rest of the exon 1 coding sequence), and also brought the *lacZ* gene under the regulation of the endogenous *Evc* promotor. (**B**, **C**) Expression images show β-galactosidase activity in $Evc^{+/-}$ embryos and represent (**B**) expression in the whole E15.5 embryo (15.5 days *post coitum*) and (**C**) in a sagittal paraffin section of the head of an E15.5 embryo. Note the strong expression in the mouth and tooth-forming areas in addition to the developing skeleton. mx, maxilla; ma, mandible; mc, Meckel's cartilage; tb, temporal bone. (Courtesy of Judith Goodship, Newcastle University; from Ruiz-Perez VL *et al.* [2007] *Development* **134**:2903–2912; PMID 17660199. Reproduced with permission from The Company of Biologists.)

Creating point mutations

Sometimes the gene targeting is designed to alter just one or a small number of nucleotides at a pre-determined position within a target gene. The objective may be to investigate the contribution of a small sequence to gene function, such as a specific codon, or to replicate a known or suspected pathogenic human mutation. In either case, it is important that the end product is a mutant allele that does not have any marker gene (which could possibly affect its function). To achieve this, two rounds of recombination are typically required. In the tag-and-exchange strategy, for example, the desired gene is first tagged by inserting a selectable marker during one gene-targeting round. The ESCs containing the suitably tagged target gene are then identified so that a second round of targeting can be carried out to introduce the desired point mutation with a high degree of specificity while at the same time removing the marker gene.

Nuclear transfer has been used to produce genetically modified domestic mammals

Until the advent of induced pluripotent stem cells, the principal methods used to generate transgenic mice were not readily applicable in some animals. However, **somatic cell nuclear transfer** (**SCNT**), a technically very demanding method that allows animals to be cloned, has occasionally been used to make transgenic mammals other than mice.

Somatic cell nuclear transfer involves the replacement of an oocyte nucleus with the nucleus of a somatic cell, which is then re-programmed by the oocyte. Despite the differentiated state of the donor cell, the re-programming can make the nucleus *totipotent* so that it is able to recapitulate the whole of development.

The technology itself is not new. It has been used for over 50 years to clone amphibians and it has been possible to produce cloned mammals from embryonic cells since the 1980s. In 1995, it was shown for the first time that mammals could be cloned from the nuclei of cultured cells, and in the following year the first mammal cloned from an adult cell was first reported, a sheep called Dolly.

Dolly was produced by transferring a nucleus from a mammary gland cell from a Finn Dorset sheep into an enucleated oocyte taken from a Scottish blackface sheep. The resulting artificially fertilized oocyte was allowed to develop to the blastocyst stage before implantation in the uterus of a foster mother (**Figure 8.25**). Out of a total of 434 oocytes, only 29 developed to the transferable blastocyst stage, and of these only one developed to term, giving rise to Dolly. Dolly had the same nuclear genome as the Finn Dorset sheep donor of the mammary gland cell and so they were considered to be clones. However, their mitochondrial genomes were different because Dolly inherited her mtDNA from the Scottish blackface sheep from which the enucleated oocyte was derived. Subsequently, a variety of different mammals have been cloned by the nuclear transfer procedure, including mice, rats, cats, dogs, horses, mules, and cows.

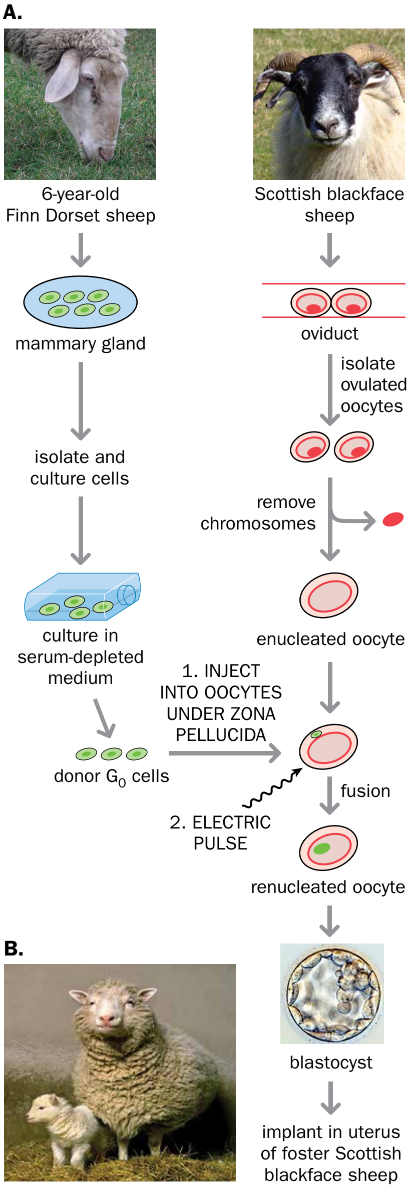


Figure 8.25 The first successful attempt at mammalian cloning from adult cells resulted in a sheep called Dolly. (**A**) Experimental strategy. The donor nuclei were derived from a cell line established from adult mammary gland cells of a Finn Dorset sheep. The donor cells were deprived of serum before use, forcing them to exit from the cell cycle and enter a quiescent state known as G_0, in which only minimal transcription occurs. Nuclear transfer was accomplished by fusing individual somatic cells to enucleated, metaphase II-arrested oocytes from a Scottish blackface sheep. Eggs are normally fertilized by transcriptionally inactive sperm whose nuclei are presumably programmed by transcription factors and other chromatin proteins available in the egg, and so the G_0 nucleus may represent the ideal basal state for reprogramming. Note that in the original cloning strategy leading to the birth of Dolly (Wilmut I *et al.* [1997]; PMID 9039911) the nuclear transfer occurred by cell fusion, but it is now customary to isolate the nucleus from donor cells, after which it is fused to the enucleated oocyte (after applying an electric current) or microinjected into the oocyte. (**B**) Dolly with her firstborn, Bonnie. (B, courtesy of the Roslin Institute, Edinburgh.)

We will consider here the nuclear transfer method as a way to generate transgenic animals. The essential point is that if the donor nucleus is taken from a somatic cell that has been genetically manipulated to contain a transgene of interest, the animal that develops from the manipulated oocyte will be transgenic. Transgenic livestock have been produced in this way with transgenes that produce therapeutic proteins, as described in Chapter 22. It is also possible to modify a specific pre-determined endogenous gene in a somatic cell, such as a fibroblast, and then use SCNT to make an animal with the mutation in all its cells, as in the pig and ferret models of cystic fibrosis that will be described in Chapter 21.

SUMMARY

- Genetic manipulation of mammalian cells involves transporting genetic material—often engineered DNA constructs (transgenes) or synthetic oligonucleotides—into cells.

- Genetic manipulation of mammalian cells is carried out for different purposes, including: to understand how genes and other functional DNA sequences work; to model diseases and investigate the molecular basis of disease; to make therapeutic proteins and other valuable gene products; and to treat certain types of disease (gene therapy).

- At a molecular level, the motivation may be to express a human or mammalian gene to give some desired product, to alter the genome of the cell in some defined way (genome editing), or to block the expression of a pre-determined gene of interest.

- The genetic manipulation can be carried out *in vitro* using cultured cells, fertilized oocytes, and so on. Subsequently, the genetically modified cells may be transferred into an embryo of a model organism, or into a person or model organism.

- The most convenient type of cultured cells are cell lines that are immortal (because they derive from cells with naturally high telomerase expression, such as pluripotent stem cells) or that have been made immortal in some way (leading to high telomerase expression).

- Genetic manipulation can also be carried out *in vivo* (when the intention is to transport genetic material into the cells of a tissue within a person or model organism).

- Transfer of transgenes into mammalian cells is most efficiently accomplished using viruses (transduction): the transgene is inserted into a virus vector and becomes packaged within a virus protein coat (which, however, places size limits on the amount of foreign DNA that can be accommodated).

- Physical and chemical methods can also be used to transport transgenes into mammalian cells (transfection). They are comparatively less efficient than viral transfer methods, but safer and can often transport very large transgenes.

- For some purposes it is sufficient for transgenes to reside in the cytoplasm. In dividing cells, entry into the nucleus becomes possible when the nuclear membrane begins to break down in preparation for cell division.

- Many viruses naturally enter the nucleus and certain viral vectors can efficiently transport foreign DNA into the nucleus. Certain physical methods are also quite successful in transporting transgenes to the nucleus.

- Retroviruses are RNA viruses that can insert a DNA copy of their genome into the genome of the host cell. Using retroviral vectors, a transgene can be integrated into the cell's genome; the integrated transgene is duplicated during S phase then transmitted to daughter cells.

- Genome editing often involves making small, precise changes to a pre-determined gene within intact cells (gene targeting). It can also involve inserting reporter genes to be expressed by an endogenous promoter (to track expression of the endogenous gene). Occasionally, it involves making larger changes: large-scale deletions and inversion, plus translocations (chromosome engineering).

- Genome editing using homologous recombination involves sequence exchange between a genome sequence of interest (the target site) and an introduced transgene. The transgene is designed to have two segments identical in sequence to segments of the target site but separated by a central segment with an altered or novel sequence that is stitched into the target site by recombination.

- Heterologous site-specific recombination systems can be used in genome editing by using a transgene to express the heterologous recombinase (such as the bacterial Cre protein) and by using homologous recombination to stitch in two copies of its recognition sequence (such as *lox*P sequences). The expressed recombinase will promote recombination between its two target sequences to produce a deletion or inversion of the intervening DNA, or a translocation.

- Genome editing using programmable endonucleases begins by designing a protein or RNA with a target-specific guide sequence (to bind to the pre-determined target site) and the capacity to transport an exogenous endonuclease to make a double-strand break at that site. Thereafter, desirable sequence changes can be introduced during DNA repair, either as a result of random DNA repair errors, or by copying a desired sequence from an introduced donor DNA.

- Some programmable endonuclease systems use a pair of hybrid proteins to cleave each DNA strand. Each hybrid protein has a target-specific protein guide sequence covalently joined to a nonspecific DNA-cleavage domain (from a restriction endonuclease with separable DNA-binding and DNA-recognition domains).

- Protein guide sequences for programmable endonucleases are constructed using genetic engineering to assemble a series of modular DNA-binding units of known sequence specificity, such as C2H2 zinc fingers (which recognize specific trinucleotides) and TALEs (which bind to just one type of nucleotide).

- CRISPR-Cas genome editing uses RNA-guided endonucleases. A hybrid RNA is constructed that consists of a target-specific RNA guide sequence (designed to hybridize to the target site) joined to an RNA sequence that can bind a Cas endonuclease. The Cas endonuclease has two DNA-cleavage domains, one for cleaving each DNA strand.

- Genome editing can be used to target a gene and inactivate it (a gene knockout). An alternative is to specifically target the RNA from a gene of interest in order to down-regulate it (a gene knockdown = gene silencing).

- Gene silencing can be carried out by designing antisense oligonucleotides or RNA to hybridize specifically to RNA produced from a gene of interest (target RNA), inhibiting its expression. The alternative is to produce target RNA-specific short interfering RNAs (complementary short oligonucleotides with 2 bp overhanging 3′ ends) that trigger a cellular

RNA interference pathway to cleave target RNAs having the same sequence (RNA silencing).

- Inserting transgenes into the germ line allows transgenic animals to be made that have a desired genetic modification in all their cells. They are mostly used for understanding gene function or for modeling a disease.
- A wide range of transgenic animals have been made by a method that begins with microinjecting a transgene into a pronucleus of a recently fertilized oocyte.
- Genetically modified mice have also frequently been obtained after first genetically modifying embryonic stem cells in culture, and then selecting suitably modified stem cells to be inserted into the inner cell mass of a blastocyst that is then implanted into the uterus of a foster mother.
- To avoid a problem with embryonic lethality, it is often desirable to make conditional knockouts in which the genetic modification is induced to occur in defined tissues or at a defined stage of development.

FURTHER READING

Establishing immortalized cell lines

Hui-Yuen J *et al.* (2011) Establishment of Epstein-Barr virus growth-transformed lymphoblastoid cell lines. *J Vis Exp* **57**:e3321; PMID 22090023. (A video showing technical details is available at http://www.jove.com/video/3321/.)

Ouellette MM *et al.* (2000) The establishment of telomerase-immortalized cell lines representing human chromosome instability syndromes. *Hum Mol Genet* **9**:403–411; PMID 10655550.

Sie L *et al.* (2009) Utility of lymphoblastoid cell lines. *J Neurosci Res* **87**:1953–1959; PMID 19224581.

Methods for transferring genetic material into mammalian cells

Heiser WC (ed) (2004) *Gene Delivery to Mammalian Cells, Vol. 1: Nonviral Gene Transfer Techniques*. Humana Press.

Heiser WC (ed) (2004) *Gene Delivery to Mammalian Cells, Vol. 2: Viral Gene Transfer Techniques*. Humana Press.

Miller AD (1997) Development and applications of retroviral vectors. In: *Retroviruses* (Coffin JM, Hughes SH & Varmus HE eds). Cold Spring Harbor Laboratory Press. (Freely available at http://www.ncbi.nlm.nih.gov/books/NBK19428/.)

Yarmush ML *et al.* (2014) Electroporation-based technologies for medicine: principles, applications, and challenges. *Annu Rev Biomed Eng* **16**:295–320; PMID 24905876.

Homologous recombination mechanism

Greene EC (2016) DNA sequence alignment during homologous recombination. *J Biol Chem* **291**:11572–11580; PMID 27129270. (Focuses on what is known about how homologous recombination is initiated.)

San Filippo J *et al.* (2008) Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem* **77**:229–257; PMID 18275380.

Genome editing using homologous recombination and Cre/*lox*P recombination

Capecchi M (1989) Altering the genome by homologous recombination. *Science* **244**:1288–1292; PMID 2660260.

Kühn R & Torres RM (2002) Cre/loxP recombination system and gene targeting. *Methods Mol Biol* **180**:175–204; PMID 11873650.

Le Y & Sauer B (2000) Conditional gene knockout using cre recombinase. *Methods Mol Biol* **136**:477–485; PMID 10840735.

Wallace HA *et al.* (2007) Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**:197–209; PMID 17218265.

Yu Y & Bradley A (2001) Engineering chromosomal rearrangements in mice. *Nat Rev Genet* **2**:780–790; PMID 11584294.

Genome editing using programmable (targetable) endonucleases

Amitai G & Sorek R (2016) CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* **14**:67–76; PMID 26751509.

Carroll D (2014) Genome engineering with targetable nucleases. *Annu Rev Biochem* **83**:409–439; PMID 24606144.

Sander JD & Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* **32**:347–355; PMID 24584096.

Shalem O *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. Science **343**:84–87; PMID 24336571.

Antisense technology and gene silencing

Gao X & Zhang P (2007) Transgenic RNA interference in mice. *Physiology* **22**:161–166; PMID 17557936.

Heasman J (2002) Morpholino oligos: making sense of antisense? *Dev Biol* **243**:209–214; PMID 11884031.

Nature Insight Guide on RNA silencing (2009) *Nature* **457**:396–425. (Various reviews on RNA silencing.)

Transgenic animals and mouse genetics

Houdebine LM (2007) Transgenic animal models in biomedical research. *Methods Mol Biol* **360**:163–202; PMID 17172731.

Silver L (1995) *Mouse Genetics. Concepts and Applications*. Oxford University Press. (Now freely available electronically at http://www.informatics.jax.org/silver/.)

Uncovering the architecture and workings of the human genome

9

In this chapter we cover two areas. First, we describe what is currently known about the architecture of the human genome—the different classes of DNA sequences, their characteristics, and how they are organized. Secondly, we describe recent and ongoing global efforts to build comprehensive catalogs of human genes, gene products, and regulatory sequences, and to carry out genome-wide investigations to dissect how our genome functions.

We present an overview of the human genome in Section 9.1, covering the basic characteristics of the nuclear and mitochondrial genomes and providing some genome statistics, before concluding with a section on accessing relevant electronic resources. In Section 9.2 we examine how genes are organized in our genome, and the extent and organization of repetitive gene sequences and gene families. Thereafter, we turn our attention in Section 9.3 to two classes of very highly repetitive, mostly noncoding DNA: heterochromatic DNA (the DNA underlying constitutive heterochromatin) and transposon repeats.

As detailed in Section 7.1, most of our knowledge of the human genome was made possible by the Human Genome Project, an international effort that lasted about 14 years, concluding in 2003 with an almost complete sequence of the euchromatic portion of the human genome. The resulting human genome reference sequence was a monumental achievement at the time. It went a long way to answering the question: *What is the sequence of the human genome?* But in the quest to understand our genome, it had just got us to first base (excuse the pun!).

In 2003 the challenge moved to addressing the much more difficult question: *What does the sequence mean?* As a first step toward answering that question, the international ENCODE Project sought to establish a catalog of functional human DNA elements. In Section 9.4 we examine how the ENCODE Project and ongoing studies, including global analyses of gene function and of human transcriptomes and proteomes, seek to define the functional elements in our genome, and how our genome works.

Other aspects of the human genome are covered in later chapters. Gene regulation is examined in Chapter 10, and the evolutionary aspects (of the genome, genes, and humans) are covered in Chapters 13 and 14. The principles of genome sequence variation and population genetics are introduced in Chapters 11 and 12; multiple later chapters are devoted to disease-associated sequence variation.

9.1 AN OVERVIEW OF THE HUMAN GENOME

The human genome consists of 25 different DNA molecules and comprises two physically separate genomes: a complex nuclear genome that contains the vast majority of our genes, and a very simple mitochondrial genome with just 37 genes.

The protein-coding genes of the nuclear and mitochondrial genomes are expressed using independent protein-synthesis capacities. The mRNA transcripts from nuclear genes are translated on 80S cytoplasmic ribosomes with the assistance of cytosolic tRNAs (made by nuclear genes). Some of the 80S cytoplasmic ribosomes exist freely in the cytosol; others are attached to the outer membrane of the nuclear envelope and to physically continuous endoplasmic reticulum (rough endoplasmic reticulum). According to their function, the resulting proteins may be exported to different cell compartments (including mitochondria), or secreted.

The mRNAs produced from mitochondrial DNA (mtDNA) are translated on 55S mitochondrial ribosomes (also called mitoribosomes) located on the inner mitochondrial membrane, using mitochondrial tRNAs (all transcribed from mitochondrial DNA). The mitoribosomes have 80 proteins and three RNAs: a 16S rRNA in the large subunit, a 12S rRNA in the small subunit, and in addition a mitochondrial $tRNA^{Val}$ molecule that serves as the equivalent of the 5S rRNA of cytoplasmic ribosomes. The mitochondrial mRNAs make certain of the membrane-bound proteins that work in oxidative phosphorylation, as described below.

As we detail below, the human nuclear and mitochondrial genomes are very different in many respects—see **Table 9.1** for a summary.

TABLE 9.1 THE HUMAN NUCLEAR AND MITOCHONDRIAL GENOMES

Characteristic	Nuclear genome	Mitochondrial genome
Size	3.1 Gb	16.6 kb
Number of different DNA molecules	23 (in XX cells); or 24 (in XY cells); all linear	One circular DNA molecule
Total number of DNA molecules per cell	Varies according to ploidy, but 46 in diploid cells and 23 in gametes	Often several thousand copies (copy number varies in different cells)
Associated protein	Several classes of histone and non-histone protein	Largely free of protein
Number of protein-coding genes	Close to 20,000	13
Number of RNA genes	Uncertain; possibly ~20,000 (see text)	24
Gene density	~1/80 kb	1/0.45 kb
Number of pseudogenes	~15,000	None
Repetitive DNA	At least 50% of genome, and perhaps 67%	Very little
Transcription	Many genes are independently transcribed	Multigenic transcripts are produced from both DNA strands
Introns	Found in most protein-coding genes and many RNA genes	Absent
Amount of protein-coding DNA	~1.1%	~66%
Codon usage	61 amino acid codons plus three stop codons[a]	60 amino acid codons plus four stop codons[a]
Recombination	At least once for each pair of homologs at meiosis	Not evident
Inheritance	Mendelian for DNA sequences on X chromosome and autosomes; paternal for most DNA sequences on the Y chromosome	Exclusively maternal

[a] For details see **Figure 1.29**.

The mitochondrial genome resembles a stripped-down bacterial genome and its copy number varies very significantly between cells

Like most bacterial genomes, the human mitochondrial genome consists of a single type of circular, double-stranded DNA. Its sequence of 16,569 nucleotides was first reported by Fred Sanger and colleagues at Cambridge, UK, in 1981, almost a decade before the Human Genome Project began. A revision of the sequence, published in 1999, provided the revised Cambridge reference sequence (GenBank: NC_012920). This tiny genome—less than 0.4% of the size of the *E. coli* genome (and even less than 10% of the size of the human Epstein–Barr virus genome)—resembles bacterial genomes in being packed with genes. There are no introns, and significantly more than 90% of the genome directly specifies a protein or functional RNA.

Whereas the nuclear DNA is tightly complexed with proteins that dictate its conformation, mitochondrial DNA is, like bacterial DNA, comparatively protein-free.

(It is, however, complexed with certain proteins to form *nucleoids*, nucleoprotein structures that have a diameter of around 100 nm; see **Figure 2.4**). The two strands of mitochondrial DNA are transcribed to give long RNA transcripts that, like bacterial polycistronic RNAs, are cleaved to generate individual functional mRNAs and noncoding RNAs. And, as described above, the mitochondrial mRNAs are translated on mitochondrial ribosomes using mitochondrial tRNAs.

These and other observations support the *endosymbiont hypothesis* to explain the origin of eukaryote cell lineages by a cell-fusion event. In its modern form, it envisages that an aerobic α-proteobacterium was engulfed by a larger eukaryotic precursor cell, almost certainly a type of archaeon (see the cell evolution description in Section 2.1 for further details). The resulting single cell had, therefore, two genomes and two sets of protein-synthesis machinery. The genome of the engulfed α-proteobacterium progressively shed DNA sequences, becoming much smaller in size, eventually giving rise to the mitochondrial genomes of eukaryotic cells. Many of the shed DNA sequences integrated into the genome of the engulfing cell. The latter genome also increased in size during evolution by undergoing various types of DNA duplication, and gave rise to the large nuclear genomes of eukaryotic cells.

Replication and transmission of mtDNA

The replication of both the heavy (H) and light (L) strands of mtDNA is unidirectional and starts at specific origins (**Figure 9.1**). Although mtDNA is principally double-stranded, repeat synthesis of a small segment of the H-strand DNA produces a short third DNA strand called 7S DNA. The 7S DNA strand can base-pair with the L-strand, displacing the H-strand, which forms a loop, the displacement loop (D-loop). This region contains many of the mtDNA control sequences (including the major promoter regions) and so it is referred to as the CR/D-loop region (CR = control region). The origin of replication for the H-strand lies in the CR/D-loop region, and that of the L-strand is sandwiched between two tRNA genes some distance from the control region.

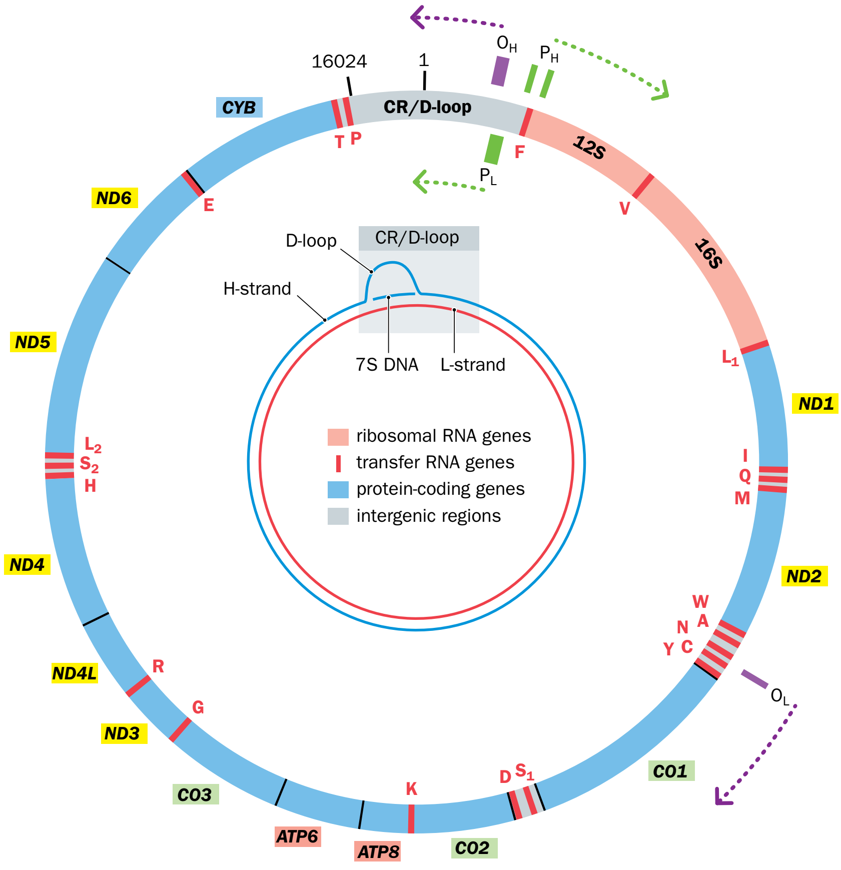


Figure 9.1 Organization of the human mitochondrial genome. The circular 16,569 bp genome has a base composition of 44% GC, but has a heavy (H) strand rich in guanines, and a light (L) strand rich in cytosines. Twenty-four RNA genes make 12S and 16S rRNAs and 22 tRNAs (tRNA genes are shown as thin red bars with a letter indicating the amino acid; note there are two tRNA^{Leu} genes, L_1 and L_2, and two tRNA^{Ser} genes, S_1 and S_2). Thirteen protein-coding genes make components of the oxidative phosphorylation system: seven NADH dehydrogenase subunits (*ND1–ND6* and *ND4L*), two ATP synthase subunits (*ATP6* and *ATP8*), three cytochrome *c* oxidase subunits (*CO1–CO3*), and cytochrome *b* (*CYB*). Two promoters, green boxes labeled P_H (in two segments) and P_L, transcribe respectively the H- and L-strands in opposite directions, generating large multigenic transcripts from each strand that are subsequently cleaved. O_H and O_L (purple boxes) signify replication origins (dashed arrows mark the direction of DNA synthesis). A roughly 500 bp control region (CR), commencing at position 16024, has multiple regulatory sequences. It has an internal displacement loop (D-loop), a triple-stranded structure formed by repeat synthesis of a short piece of heavy-strand DNA, 7S DNA, that base-pairs with the L-strand, causing local looping of the H-strand. Gene symbols for protein-coding genes are abbreviated by omitting the prefix *MT-*. For further information, see the MITOMAP database at http:// www.mitomap.org.

Unlike chromosomal DNA molecules that are normally subject to strict control of copy number, mtDNA replication is more flexible and varies between cells. According to the cell type, between 1000 and 10,000 mtDNA copies are found within the inner mitochondrial compartment (matrix), but oocytes are exceptional in having about 100,000 mtDNA copies. During mitotic cell division, the multiple mtDNA molecules in a dividing cell segregate in a purely random way to the two daughter cells.

During zygote formation, a sperm cell contributes its nuclear genome, but not its mitochondrial genome, to the egg cell. Originating exclusively from the unfertilized egg, the mitochondrial DNA of the zygote is maternally inherited: males and females both inherit their mitochondria from their mother, but males do not transmit mitochondrial DNA to subsequent generations.

The limited autonomy of the mitochondrial genome and the use of a variant genetic code

As detailed in **Figure 9.1**, the human mitochondrial genome contains just 37 genes. Whereas nuclear genes often have their own dedicated promoters, transcription of mtDNA results in large multigenic transcripts, just as in transcription of bacterial DNA. The large multigenic transcripts produced by transcribing the H- and L-strands are subsequently cleaved to generate a mix of coding and noncoding RNAs, as described below.

- Thirteen mRNAs. They are translated on mitoribosomes to make 13 protein components of mitochondrial respiratory complexes, the membrane-bound enzymes of oxidative phosphorylation used to make ATP.
- Two rRNAs and 22 tRNAs. (Note that the mitochondrial tRNA^{Val} has a dual role: as well as carrying valine to be incorporated into proteins, some of the tRNA^{Val} molecules act as structural and functional components of the mitoribosome, serving a role analogous to that of 5S rRNA in cytoplasmic ribosomes.)

Like bacterial genomes, the mitochondrial genome is extremely compact. In addition to lacking introns, the mitochondrial genes are very tightly packed. In most cases the sequences of neighboring genes are contiguous, or separated by just one or two noncoding bases, and two of the protein-coding genes, *MT-ATP6* and *MT-ATP8*, have overlapping coding sequences with different reading frames. Space is conserved to such an extent that some of the protein-coding genes even lack termination codons; to overcome this deficiency, UAA codons have to be introduced at the post-transcriptional level.

The limited autonomy of the mitochondrial genome

The mitochondrial genome has retained a measure of independence when it comes to protein synthesis because all the ribosomal RNAs and tRNAs needed for protein synthesis on mitoribosomes are made by transcribing mtDNA. However, it is a different matter for mitochondrial proteins. Only 13 out of the 80 proteins required for oxidative phosphorylation are specified by the mitochondrial genome (**Table 9.2**). None of the core nucleoid proteins—including critically important proteins needed to replicate, transcribe, and repair mtDNA—are specified by the mitochondrial genome. And although all the RNAs needed for translating mitochondrial mRNAs are made by mtDNA, all the proteins of the protein-synthesis machinery (mitochondrial ribosomal proteins and aminoacyl tRNA synthetases) are made by nuclear genes. In total, nearly 1700 nuclear genes are listed as making mitochondrial proteins in the February 2016 release of the MitoProteome database at www.mitoproteome.org, and all but 13 of the mitochondrial proteins—that is, more than 99%—are specified by nuclear genes, being synthesized on cytoplasmic ribosomes and then imported into the mitochondria (see **Table 9.2**).

The variant mitochondrial genetic code

Prokaryotic genomes and the nuclear genomes of eukaryotes encode many hundreds to usually many thousands of different proteins. They are subject to a “universal” genetic code that is kept invariant: mutations that could potentially change the genetic code are likely to produce at least some critically malfunctional proteins and so are strongly selected against. However, the much smaller mitochondrial genomes make very few polypeptides. As a result, the mitochondrial genetic code has been able to drift by mutation to be slightly different from the universal genetic code.

In the human mitochondrial genetic code, 60 codons specify amino acids, one fewer than in the nuclear genetic code. There are four stop codons: UAA and UAG (also stop codons in the nuclear genetic code) and AGA and AGG (which specify arginine in the

TABLE 9.2 THE LIMITED AUTONOMY OF THE MITOCHONDRIAL GENOME

Mitochondrial proteins and RNAs	Specified by mitochondrial genome	Specified by nuclear genome
OXIDATIVE PHOSPHORYLATION SYSTEM PROTEINS	13	84
I NADH: ubiquinone oxidoreductase complex	7	37
II Succinate dehydrogenase	0	4
III Ubiquinol–cytochrome C reductase complex	1	9
IV Cytochrome *c* oxidase complex	3	16
V ATP synthase complex	2	18
PROTEIN SYNTHESIS COMPONENTS	24	>100
rRNA	2	0
tRNA	22	0
Ribosomal proteins	0	80
Aminoacyl tRNA synthetases*	0	19
Translation factors and others	0	ALL
CORE NUCLEOID PROTEINS**	0	ALL
OTHER MITOCHONDRIAL PROTEINS	0	ALL

*There are only 19 mitochondrial aminoacyl tRNA synthetases because $tRNA^{Gln}$ is a dual-purpose tRNA: it can be misacylated and sometimes carries glutamate and sometimes glutamine.

** Mitochondrial nucleoid proteins are involved in crucially important basic functions such as replication, transcription, and repair of mtDNA. They include DNA polymerase λ subunits 1 and 2, mitochondrial RNA polymerase, mitochondrial transcription factors A and B, Twinkle (mtDNA helicase), mitochondrial single-stranded DNA-binding protein, Lon peptidase, and ERCC6 (excision repair 6, chromatin remodeling factor).

nuclear genetic code; see **Figure 1.29**). UGA, a nuclear stop codon, encodes tryptophan in mitochondria, and AUA specifies methionine in mitochondria, not isoleucine.

There are only 22 different types of human mitochondrial tRNA, but the individual tRNA molecules can each interpret several different codons because of *third-base wobble*. Eight of the 22 tRNA molecules have anticodons that each recognize families of four codons differing at the third base only. The other 14 tRNAs recognize pairs of codons that are identical at the first two base positions and share either a purine or a pyrimidine at the third base. Between them, therefore, the 22 mitochondrial tRNA molecules can recognize a total of (8 × 4) + (14 × 2) = 60 codons.

Evolutionarily recent and ongoing transfer of mtDNA sequences into the nuclear genome

As described above, the endosymbiont hypothesis proposes that mitochondrial genomes arose from an α-proteobacterium in an ancient cell-fusion event that occurred around 1.5 billion years ago. To explain the very small size of current mitochondrial DNAs, DNA sequences were envisaged to have been shed by the original genome, but with many of them subsequently integrating into what would become the nuclear genome. That probably happened comparatively quickly after the ancient cell-fusion event (at least when referenced against evolutionary timescales) and had the advantage of removing most of the genes from close exposure to harmful free radicals generated by the oxidative phosphorylation system.

In addition to this ancient sequence transfer between the two precursor genomes, there has also been evolutionarily recent and ongoing transfer of mitochondrial DNA sequences to the nuclear genome. Analysis of the human genome reference sequence using standard BLAST programs reveals over 750 nuclear sequences that are imperfect copies of mtDNA sequences, with sizes ranging from tens of nucleotides up to 14,654 nucleotides in length. The transferred mtDNA sequences have acquired inactivating mutations over time and so are sometimes described as nuclear mitochondrial pseudogenes, but they are more generally known as nuclear mitochondrial DNA sequences (NUMTs). In total they account for over 627 kb of the nuclear genome; see **Table 9.3** for five prominent examples. Many human NUMT sequences are present in some haplotypes but not in others and so constitute insertion/deletion polymorphisms.

TABLE 9.3 EXAMPLES OF NUCLEAR MITOCHONDRIAL SEQUENCES (NUMTs) THAT ARE CLOSELY RELATED IN SEQUENCE TO mtDNA

NUMT length (bp)	Chromosome location (in GRCh38 reference sequence)			Corresponding coordinates on the human mtDNA reference sequence		% NUMT–mtDNA sequence identity
11,115	Chr 17:	22,521,401	22,532,515	1	11,112	85.1
9108	Chr 5:	100,055,045	100,045,983	6117	15,183	89.2
5841	Chr 1:	629,084	634,924	3914	9755	98.6
5219	Chr 5:	134,928,527	134,923,309	10,269	15,487	94.1
12,611	Chr 7:	57,185,765	57,198,375	3819	16,475	79.3

NUMTs are found in the nuclear genomes of a wide range of eukaryotes, and their existence means that even in complex eukaryotic cells, mitochondrial DNA sequences can somehow move to the nucleus, and then integrate into the nuclear genome. Integration appears to occur using the nonhomologous end-joining DNA repair mechanism that specializes in repairing double-strand DNA breaks. There is less certainty about how mitochondrial DNA sequences get to the nucleus, but it most likely begins with natural degradation of damaged or abnormal mitochondria, and release of mitochondrial DNA fragments into the cytoplasm.

The great majority of human NUMTs are not of recent origin (the oldest may have entered the nuclear genome about 60 million years ago). However, the insertion process is ongoing: a small proportion of NUMT loci are polymorphic, being present in some individuals but absent in others (note that the human genome reference sequence lacks some NUMTs present in the human population). And occasional *de novo* insertion of mtDNA sequences into the nuclear genome is known to disrupt gene expression, causing disease.

The human nuclear genome is comparatively gene-poor and has significant amounts of heterochromatic DNA

The principal aim of the Human Genome Project was to obtain the (almost) complete sequence of the nuclear genome. Unlike mitochondrial DNA, chromosomal DNA molecules have regions with long clusters of consecutive repeated DNA sequences (tandem repeats) that are difficult to sequence (assembly of clone contigs across long regions of repetitive DNA is often a challenge). The constitutive heterochromatin of somatic cells is essentially devoid of genes, and it has very long clusters of repetitive DNA whose sequences are difficult to obtain.

Because of the above considerations, heterochromatin was accorded a low priority for DNA sequencing; the primary goal of the Human Genome Project was to sequence the euchromatin fraction of the genome. That meant, however, that essentially all genes would be sequenced. (A notable exception was the large arrays of tandemly repeated ribosomal RNA genes; that was not considered a problem because individual sequences had already been obtained for each of the different rRNA genes.) By 2003, a virtually complete sequence had been obtained for the euchromatic region of human DNA.

By summing the lengths of the DNA molecules in our 24 different chromosomes (22 autosomes, X and Y), the size of our nuclear genome is now estimated to be very close to 3.1 Gb (3100 Mb) in size, and currently approximately 40,000 human genes are recognized, as described below. That gives a comparatively low gene density of roughly one gene per 80 kb, and there is a very low percentage, about 1.1%, of coding DNA. By comparison, the 16.6 kb mitochondrial genome, like bacterial genomes, is packed with genes (with one gene per 450 bp, on average), and has a high proportion (66%) of coding DNA. The nuclear genome is also radically different from the mitochondrial genome in having a very large amount of repetitive DNA, which is especially prevalent in heterochromatin, but also very prominent in euchromatin, as described in Section 9.3. Note that unlike the mitochondrial genome, the chromosomal DNA molecules have large amounts of bound protein, both histones and non-histone proteins, and certain associated RNAs can serve structural roles too within the heterochromatic regions, as described below.

All egg cells, and sperm cells bearing an X chromosome, have a haploid genome of 3.04 Gb of DNA (3.1 Gb minus Y-chromosome DNA); sperm cells bearing a Y chromosome have a haploid genome of 2.94 Gb. Most somatic cells are diploid, having two copies of the nuclear genome, but have a much larger and more variable number of copies of the mitochondrial genome. Because the size of the nuclear genome is about 190,000 times the size of an mtDNA molecule, however, the nucleus of a human cell typically contains more than 99% of the DNA in the cell. (The oocyte is a notable exception: its 100,000 mtDNA molecules comprise a total of about 1.6 Gb of DNA, about one-third of the DNA in the cell.)

The DNA of human chromosomes varies considerably in length, from 249.7 Mb for chromosome 1 to 46.7 Mb for chromosome 21 (**Table 9.4**). There is also significant variability between chromosomes in the proportions of DNA underlying euchromatin and constitutive heterochromatin. Each chromosome has several megabases of heterochromatic DNA at the centromere, but many chromosomes also have significant amounts of non-centromeric heterochromatin. Thus, chromosomes 1, 9, 16, and 19 have substantial amounts of heterochromatin in regions close to the centromeres (*pericentromeric heterochromatin*)—see the chromosome banding figure in the inside back cover. The acrocentric chromosomes (having a centromere very close to one end of the chromosome; that is, chromosomes 13, 14, 15, 21, and 22) each have two sizeable heterochromatic regions in their short arms. But the Y chromosome has by far the largest proportion of heterochromatic DNA, with a particularly large segment of non-centromeric heterochromatin on the long arm (see the figure on the inside back cover).

TABLE 9.4 REFERENCE SEQUENCES AND SOME CHARACTERISTICS OF THE 24 HUMAN CHROMOSOMAL DNA MOLECULES

Name	Size (Mb)	RefSeq ID[a]	GenBank ID	Heterochromatin component (Mb)	%GC
1	249.70[b]	NC_000001.11	CM000663.2	19.5	42.3
2	242.51[b]	NC_000002.12	CM000664.2	2.9	40.3
3	198.45[b]	NC_000003.12	CM000665.2	1.5	39.7
4	190.42[b]	NC_000004.12	CM000666.2	3.0	38.3
5	181.63[b]	NC_000005.10	CM000667.2	0.3	39.5
6	170.81	NC_000006.12	CM000668.2	2.3	39.6
7	159.35	NC_000007.14	CM000669.2	4.6	40.7
8	145.14	NC_000008.11	CM000670.2	2.2	40.2
9	138.69[b]	NC_000009.12	CM000671.2	18.0	42.3
10	133.80	NC_000010.11	CM000672.2	2.5	41.6
11	135.19[b]	NC_000011.10	CM000673.2	4.8	41.6
12	133.28	NC_000012.12	CM000674.2	4.3	40.8
13	114.36	NC_000013.11	CM000675.2	17.2	40.2
14	108.14[b]	NC_000014.9	CM000676.2	17.2	42.2
15	102.44[b]	NC_000015.10	CM000677.2	18.3	43.4
16	92.21[b]	NC_000016.10	CM000678.2	10.0	45.1
17	83.84[b]	NC_000017.11	CM000679.2	7.5	45.3
18	80.37	NC_000018.10	CM000680.2	1.4	39.8
19	58.62	NC_000019.10	CM000681.2	0.3	47.9
20	64.44	NC_000020.11	CM000682.2	1.8	43.9

(*Continued*)

TABLE 9.4 (*Continued*) REFERENCE SEQUENCES AND SOME CHARACTERISTICS OF THE 24 HUMAN CHROMOSOMAL DNA MOLECULES

Name	Size (Mb)	RefSeq ID[a]	GenBank ID	Heterochromatin component (Mb)	%GC
21	46.71	NC_000021.9	CM000683.2	11.6	42.2
22	51.86[b]	NC_000022.11	CM000684.2	14.3	47.7
X	156.04	NC_000023.11	CM000685.2	3.0	39.6
Y	57.26[b]	NC_000024.10	CM000686.2	31.6	45.4
Unplaced[c]	4.46				
ALL	**3099.71**				

The GRCh38.p12 reference sequence and the sequences of individual chromosomes can be downloaded by following instructions at https://www.ncbi.nlm.nih.gov/genome/guide/human/. The sequences of chromosome regions can be exported from genome browsers, such as the ENSMBL browser (select chromosome co-ordinates and use the *Export data* function at left of screen). Heterochromatin DNA estimates are approximate values abstracted from the International Human Genome Sequence Consortium (2004) *Nature* **431**:931–945, PMID 15496913; see inside back cover for locations of heterochromatic DNA. [a] RefSeq = NCBI reference sequence database. [b] In each of these chromosomes the sequence assembly includes some *unlocalized* scaffolds (the position and orientation within the chromosome had not been established). The corresponding chromosomal DNA sequences (listed in columns three and four are therefore slightly shorter in size. [c] *Unplaced* scaffolds are sequences found in the primary assembly but not associated with a specific chromosome

Variable base composition in the nuclear genome and the significance of CpG islands

The base composition of the nuclear genome averages out at 41.5% GC, significantly less than the 44.4% GC of human mitochondrial DNA. There is considerable variation between chromosomes, from 38.3% GC for chromosome 4 to 47.9% GC for chromosome 19 (see **Table 9.4**) and also across the lengths of chromosomes. Chromosomes and chromosomal regions with high and low %GC have, respectively, comparatively high and low gene densities, as described below.

The proportion of some combinations of nucleotides can vary considerably. Like other vertebrate nuclear genomes, the human nuclear genome has a conspicuous shortage of the CG dinucleotide, often denoted as CpG (to distinguish it from a CG base pair; the p in CpG signifies the intervening phosphate). However, certain small regions of transcriptionally active DNA have the expected CpG density and, significantly, are unmethylated or hypomethylated (**CpG islands**; see **Box 9.1**).

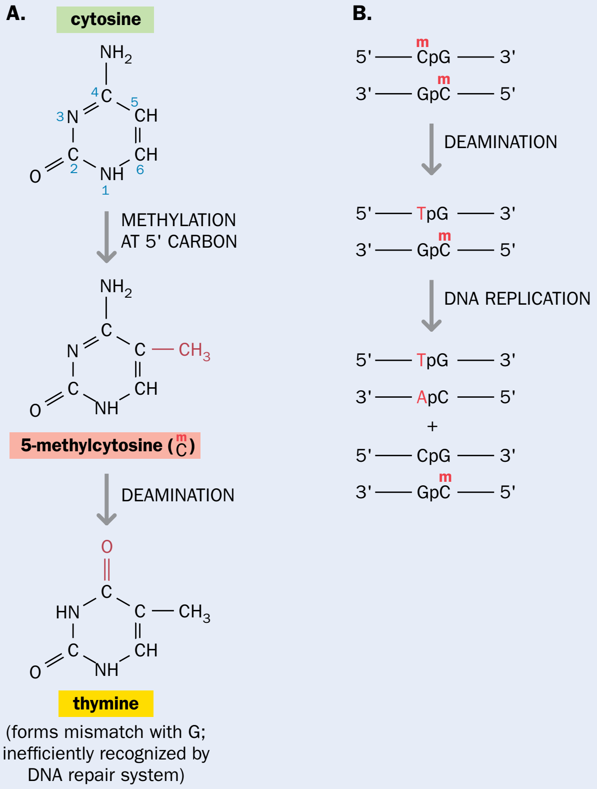
BOX 9.1 VERTEBRATE CpG ISLANDS

As detailed in Section 1.4, many different types of chemically-modified nucleotides are found in vertebrate RNAs, but chemical modification in vertebrate DNA seems to be limited to methylation of carbon atom 5 of a small percentage of cytosines. The resulting 5-methylcytosines often have a guanine as their 3′ neighbor (because the dinucleotide CpG is a common target for cytosine methylation by specific cytosine methyltransferases; but some methylated cytosines have other nucleotides as their 3′ neighbors—see Section 1.2).

5-Methylcytosine is chemically unstable and is prone to deamination (**Figure 1A**). Other deaminated bases produce derivatives that are identified as abnormal nucleotides in DNA—unmethylated cytosine is deaminated to give uracil, for example—and are removed by the DNA repair machinery. However, 5-methylcytosine is deaminated to give thymine, a natural base in DNA that is not so readily recognized as being abnormal by cellular DNA repair systems. Over evolutionarily long periods, therefore, the number of CpG dinucleotides in vertebrate DNA has gradually fallen because of the slow but steady conversion of CpG to TpG (and to CpA on the complementary strand; **Figure 1B**).

Although the overall frequency of CpG in the vertebrate genome is low, small segments of unmethylated or hypomethylated DNA extending over hundreds of nucleotides have the normal, expected CpG frequency and are comparatively GC-rich (typically more than 50% GC). Within the sea of CpG-poor DNA, these islands of normal CpG density are known as CpG islands.

CpG islands are associated with transcriptionally active regions, and so are gene markers. That is the case because highly methylated DNA regions are prone to adopting a condensed chromatin conformation; but for actively transcribing DNA, the chromatin needs to be comparatively free of methylated cytosines (so that it can adopt an extended, open conformation to allow various regulatory proteins to bind readily to control regions such as promoters and enhancers).



Box 9.1 Figure 1 Instability of vertebrate CpG dinucleotides. (**A**) The cytosine in CpG dinucleotides is a target for methylation at the 5′ carbon atom. The resulting 5-methylcytosine is deaminated to give thymine (T), which is inefficiently recognized by the DNA repair system and so tends to persist (however, deamination of unmethylated cytosine gives uracil, which is readily recognized by the DNA repair system). (**B**) Vertebrate CpG dinucleotides are gradually replaced by TpG and by CpA (after DNA replication).

The human genome contains about 20,000 protein-coding genes, but the total number can never be an exact one

A major motivation of the Human Genome Project (HGP) was to define a human gene catalog. But many years after the completion of the project in 2003, we still do not know how many genes there are in the human genome! As described in Section 9.2, much of the uncertainty concerns identifying RNA genes, but even in the case of protein-coding genes, getting an accurate number has not been straightforward.

At the time of the HGP, attempts to build a human gene catalog focused on identifying protein-coding genes, which were widely assumed to make the overwhelming genetic contribution to dictating how our cells work. RNA genes, by contrast, were viewed primarily as making ubiquitous accessory molecules that assisted in a general way to guide the expression of protein-coding genes. And although some noncoding RNAs had long been known to be more specific gene regulators, they were widely regarded at the time as quirky exceptions.

When the first draft human genome sequences began to be studied, therefore, the initial analyses primarily focused on looking for protein-coding genes. At least, protein-coding genes are comparatively easy to identify in a genome sequence: there is an open reading frame, the predicted protein sequence has frequently been quite highly conserved during evolution, and the reading frame is normally as long, or almost as long, in closely related species.

Preliminary analyses of draft human genome sequences reported in 2001 suggested about 30,000 (protein-coding) genes. Most were gene predictions made without any supportive experimental evidence and heavily reliant on identification of sizeable open reading frames. The original gene number was recognized to be inflated when the genome sequences of other mammals were obtained: evolutionary comparisons failed to identify counterparts of many predicted human genes in the mouse and dog genomes.

And in several cases, two or more predicted neighboring genes turned out to be components of a much larger gene.

Recent protein-coding gene estimates

Analyses of the human genome GRChp38.12 reference sequence (released in December 2017) suggest that there are around 19,900–20,500 human protein-coding genes (**Table 9.5**). The variation in the number exists because there remain some difficulties in how to define genes. For example, read-through transcription has been found to occur between well-characterized neighboring genes on the same DNA strand, leading to the concept of *conjoined genes* (**Figure 9.2**).

TABLE 9.5 HEADLINE HUMAN GENE STATISTICS FROM THE GRCh38.p12 REFERENCE SEQUENCE (NOVEMBER 2017 FREEZE)

Feature	NCBI Eukaryotic Genome Annotation[a]		GENCODE Genome Annotation[b]	
Total genes and pseudogenes	54,274		58,381*	
Protein-coding genes	20,070		19,901	
Transcripts		113,224		82,335
Noncoding (RNA genes)[c]	17,710		23,348	
Transcripts (all noncoding)		44,965		nd
Transcripts (long noncoding only)		nd		28,468
Pseudogenes	16,085		14,723	

[a] Data from the NCBI *Homo sapiens* Annotation release 109 at https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/109/. [b] Data from GENCODE's Stats page at https://www.gencodegenes.org/stats/current.html. * Note that the GENCODE estimate for total genes and pseudogenes also includes 411 immunoglobulin/T-cell receptor gene segments not separately listed here. [c] Version 28 of the human GENCODE release recognized 15,779 long noncoding RNA genes and 7569 small noncoding RNA genes, but defining an RNA gene is not easy (see text). nd, not disclosed.

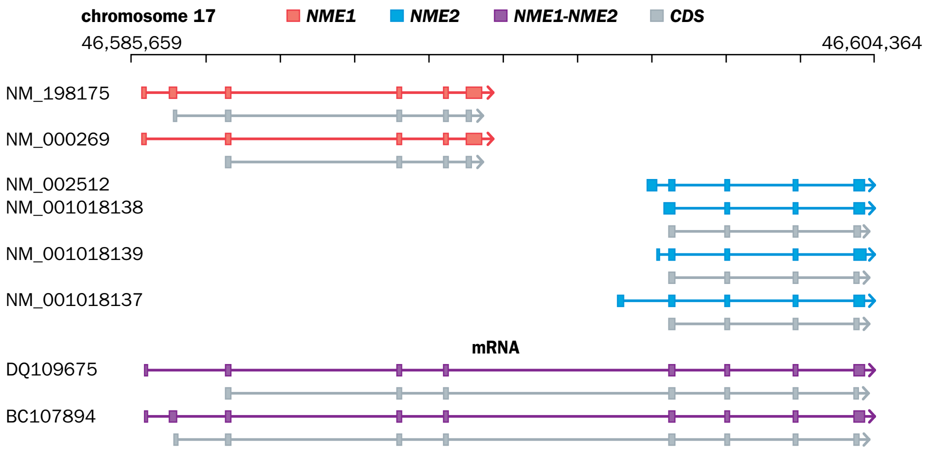


Figure 9.2 An example of read-through transcription leading to the concept of a conjoined gene. The neighboring *NME1* and *NME2* genes on chromosome 17q21 are members of a family of genes making nucleoside diphosphate kinases. Red and blue boxes indicate exons incorporated in, respectively, *NME1* and *NME2* transcripts (connecting lines are introns and gray boxes represent coding sequences (CDS); the identification numbers for reference mRNAs in the RefSeq database are shown at left). In addition to transcripts from the parent genes, other transcripts are found that link exons of the two genes, including two transcripts with GenBank reference sequences DQ109675 and BC107894 (exons shown here as magenta boxes) and numerous EST (expressed sequence tag) sequences (not shown here). (Adapted from Prakash T *et al*. (2010) *PLoS One* **5**:e13284; PMID 20967262.)

In the example in **Figure 9.2**, one might conceivably count two parental genes plus a conjoined gene. There are close to 500 examples of read-through transcription, and depending on whether conjoined genes are counted as additional genes or not, estimates of the number of human protein-coding genes can vary (see **Table 9.5**).

Further complications stem from how we classify genes. A few genes, for example, seem to make both coding transcripts (mRNA) and also transcripts that make a functional noncoding RNA. Additionally, genes originally classified as making long noncoding RNAs have subsequently been found to have very short, central coding DNA sequences that make micropeptides. We return to consider these points in Section 9.2.

The total number of human protein-coding genes is likely to stabilize somewhere in the region of 20,000, but the number of nuclear protein-coding genes can never be exact, unlike the invariant 13 protein-coding genes in our mitochondrial genome. That is so because of structural variation in the nuclear genome: for many families of protein-coding genes, the copy number varies between haploid genomes. Differences in gene number can be expected from one individual to the next, and between the maternal and paternal genomes inherited by a person.

There are numerous different human noncoding RNAs but identifying, and even defining, functional RNA genes is not straightforward

In the post-genome era, there has been a major re-evaluation of the importance of RNA genes. Currently, despite decades of not seriously investigating RNA genes, the number of recognized RNA genes is already on a par with the number of protein-coding genes (see **Table 9.5** for analyses of the GRChp38.12 reference genome sequence). And it is now clear that RNA genes perform a wide variety of functional roles, with many RNAs working as regulators of specific target genes—see **Figure 9.3** for a brief illustration of some RNA functions.

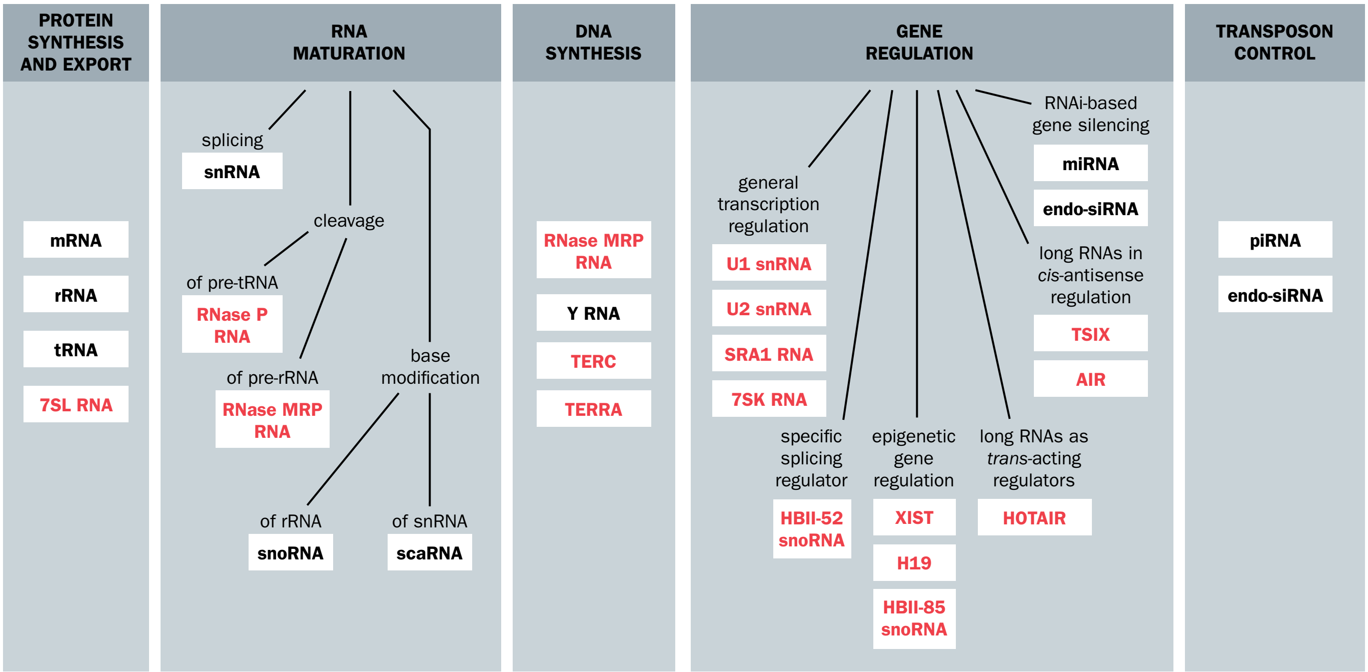


Figure 9.3 An illustration of the functional diversity of RNA. Different RNAs within the white boxes include individual RNAs (red font) or RNA families (black font). 7SL RNA is the RNA component of the signal recognition particle that regulates protein export from cells. RNAs involved in RNA maturation include spliceosomal small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), small Cajal body RNAs (scaRNAs), and RNA components of two RNA ribonucleases: RNase P, which cleaves tRNA precursors, and RNase MRP, which cleaves rRNA precursors and also has a crucial role in initiating mtDNA replication. The Y RNA family is involved in replicating chromosomal DNA. Telomere DNA synthesis is performed by a ribonucleoprotein complex of TERC (telomerase RNA component) and a reverse transcriptase (see **Figure 2.25**), and TERRA regulates certain telomere characteristics, including length. Some RNAs have general accessory roles in transcription, but many regulate specific target genes. Three classes of tiny RNA use RNA interference pathways to act as regulators: individual microRNAs (miRNAs) regulate the expression of defined target genes; piRNAs (PiWi protein-interacting RNAs) regulate the activity of transposons in germ-line cells; and endogenous short interfering RNAs (endo-siRNAs) act as gene regulators and also regulate some types of transposon. A large number of long noncoding RNAs regulate genes, often at the transcriptional level; some are involved in epigenetic gene regulation, in imprinting, X-inactivation, and so on.

Short and long noncoding RNA

It has become a convention to divide RNA genes into two classes, according to whether their mature RNA products are up to 200 nucleotides in length (short noncoding RNA) or over 200 nucleotides in length, when the RNA is described as a **long noncoding RNA** (**lncRNA**). This extremely arbitrary division was designed simply as a way of separating most of the long-familiar ubiquitous RNAs (which were thought to assist in general mechanisms used to express genes) from long RNAs that acted as regulators of specific target genes. In many ways this is a very unsatisfactory separation. First, many important regulatory RNAs are now known to be small noncoding RNAs. Additionally, this convention would split the ribosomal RNAs into two classes, identifying the longer ribosomal RNAs (cytoplasmic 28S and 18S rRNA; mitochondrial 16S and 12S rRNA) as long noncoding RNAs, while classifying the cytoplasmic 5.8S and 5S rRNAs as short noncoding RNAs. If we disregard rRNAs for the moment, the compositions of the two classes are listed below.

- ***Short noncoding RNA.*** There are seven major subfamilies (**Table 9.6**). Four are well-understood classes of ubiquitous RNAs; three are recently discovered classes of tiny regulatory RNAs. MicroRNAs are often tissue-specific and/or developmental stage-specific regulators; they work by binding to short recognition sequences in RNA transcripts produced by the target genes. piRNAs (PiWi protein-interacting RNAs) and endogenous siRNAs (short interfering RNAs) are important in germline cells as part of genome defense mechanisms (described in Section 9.3).
- ***Long noncoding RNA (lncRNA).*** The interest in lncRNAs is predominantly because this noncoding RNA class contains many important regulatory RNAs. Some lncRNAs work in the nucleus to regulate target genes, either through chromatin modification (epigenetic regulation), or through transcriptional regulation (often as natural antisense RNA transcripts). Many lncRNAs resemble mRNAs in having 5′ caps and 3′ poly(A) tails, but others do not, including many thousands of circular RNAs (which can be produced in different ways, including “back-splicing” where the 5′ splice acceptor preceding an upstream exon splices to the 3′ splice donor of a downstream exon) and stable linear RNAs processed from intronic RNA sequences. Both of the latter lncRNA categories include some examples that have been claimed to regulate target genes.

TABLE 9.6 THE ARBITRARY DIVISION OF NONCODING NONRIBOSOMAL RNAs INTO SHORT AND LONG CATEGORIES

RNA class	Major RNA families	Functions
Short noncoding RNA (up to 200 nucleotides)	snRNA (small nuclear RNA)	RNA splicing
	snoRNA (small nucleolar RNA)	Maturation of rRNAs (by chemically modifying specific bases)
	scaRNA (small Cajal body RNA)	Maturation of snRNAs (by chemically modifying specific bases)
	tRNA (transfer RNA)	Required for protein synthesis
	miRNA (microRNA)	Regulation of specific target genes by base pairing with transcripts
	piRNA (PiWi protein-interacting RNA)	Genome defense: they are important in limiting the mobilization of transposons in germ cells
	siRNA (short interfering RNA)	
Long noncoding RNA (lncRNA) (>200 nucleotides)	Nuclear lncRNA[a]	Regulation of specific target genes by modifying chromatin or transcriptional regulation
	Cytoplasmic lncRNA	Some are known to regulate translation and protein production; some others modulate protein localization or protein activity

[a] Includes antisense RNA transcripts.

Problems in identifying functional RNA genes

When it became clear that noncoding RNAs had diverse functions, searches for novel human RNA genes were stepped up. But identifying and verifying novel RNA genes has generally been much more difficult than identifying protein-coding genes for various reasons. Two general explanations for the difficulty in identifying novel RNA genes apply to all RNA genes: there are no open reading frames to search for, and by comparison with proteins, RNA sequences have often been poorly conserved during evolution.

Functional noncoding RNAs are comparatively poorly conserved for two reasons. First, unlike any protein, the sequence of some functional RNAs may be largely unimportant for their function. Because both enhancers and promoters are now known to initiate transcription bidirectionally, many long noncoding RNAs might not perform sequence-specific functions. Most antisense RNAa, for example, might have the sequence-independent function of simply being transcribed to disrupt transcription of the sense RNA. Even where it is clear that the sequence of an RNA is important for its function, natural selection may be predominantly focused on a limited number of nucleotide positions that are important in maintaining important functional features (such as binding sites for other molecules) and structural features (such as the stems of stem-loop structures—the structures of RNAs are often highly conserved, if not their sequences).

An additional reason that hindered identification of novel RNAs is that some RNAs, such as microRNAs, are tiny. Using computer programs alone to scan the genome and transcriptomes to identify tiny RNAs was too challenging. Instead, human microRNAs were discovered only after experimental studies revealed that microRNAs were functionally important gene regulators in model organisms such as *C. elegans* and *Drosophila melanogaster*, prompting strenuous efforts to identify human sequences that resembled them.)

Problems in defining RNA genes

An additional problem concerns how one defines RNA genes. It is now accepted that the great majority of the genome—at least 80%—is transcribed to produce a rather complex profile of RNA transcripts dominated by noncoding transcripts. Transcription can also occur in DNA sequences that we would not regard as genes, such as highly repetitive centromere and telomere DNA sequences, as described in Section 9.3. It is also often difficult to classify transcription units that make natural antisense transcripts as RNA genes: some antisense transcripts can extend over very large regions corresponding to several genes transcribed on the opposing DNA strand.

Many other functional RNAs are produced from transcription units that are not easily viewed as genes. For example, hundreds of thousands of different tiny piRNAs are made as part of general genome defense systems, acting through RNA interference pathways to limit the mobilization of endogenous transposons in germ cells. We further consider their role in Section 9.3. Additionally, the traditional view that intronic RNA sequences are excised and then degraded has been challenged by the finding of many thousands of circular RNAs containing intronic sequence and some stable linear intronic RNA sequences. For both of these types of noncoding RNAs, some examples are known that appear to play a role in gene regulation. As described below, intronic RNAs produced during the expression of some protein-coding genes are processed to produce small, functional noncoding RNAs. And, finally, some protein-coding genes are known to produce both mRNAs and functional noncoding RNAs that significantly overlap in sequence (**Table 9.7**).

TABLE 9.7 FEATURES THAT CAUSE UNCERTAINTY ABOUT WHETHER A GENE SHOULD BE LABELED AS PROTEIN-CODING OR NONCODING, OR BOTH

Feature	Example(s)	Reference
Hosting of small sequences that specify noncoding RNAs	Some introns in certain protein-coding genes, such as *HTR2C*, are transcribed and processed to yield snoRNAs and/or miRNAs	**Figure 9.5B**
Ability to produce a micropeptide	A gene that was long thought to make a long noncoding RNA is now known to make a 46-amino-acid micropeptide, myoregulin (which regulates muscle performance)	PMID 25640239
Functional "pseudogenes"	*PTEN* makes a tumor suppressor protein. A closely related copy, *PTENP1*, was labeled a pseudogene (it cannot make a protein). But *PTENP1* makes a regulatory noncoding RNA closely related in sequence to *PTEN* mRNA	PMID 20577206
Dual coding–noncoding genes (with overlapping functional sequences)	*SRA1* makes a regulatory noncoding RNA, steroid receptor RNA activator (SRA). Through alternative splicing, however, it is now known to make mRNA that shares much of the SRA sequence and is translated to make a protein, SRAP	PMID 21807064, 27095489
An mRNA also serves as a regulatory RNA	*TP53* makes the p53 tumor suppressor protein. But *TP53* mRNA also serves as a regulatory RNA by binding to the inhibitor protein MDM2 (which under normal conditions targets the p53 protein for degradation)	PMID 26823446

Partly because of the above complexities, it has been proposed that mature noncoding transcripts should be considered the functional units, not the DNA sequences from which they are transcribed. But that inevitably raises difficult challenges. DNA sequences are at least stable entities, being present in all nucleated cells, and, difficult though it has been, one can imagine eventually producing a definitive catalog of genes. But the sheer complexity of noncoding transcripts, their variable expression in different cell types and developmental stages, and the huge range in the expression levels makes the challenge of producing a definitive list of functional noncoding RNAs a formidable one.

The long trek to finish the sequence of the nuclear genome

At the time of writing, 15 years since the completion of the Human Genome Project in 2003, the sequence of the nuclear genome remains incomplete: the "finished sequence" is not finished! Very long stretches of tandem repeats offer a particularly difficult challenge to both DNA cloning (because of instability of DNA inserts) and to genome assembly (because of limitations in the lengths of insert DNA that can be sequenced).

Long arrays of tandem repeats are predominantly found in heterochromatin DNA, and so because of the difficulty in sequencing this DNA (and the general lack of interest in gene-lacking heterochromatin), the Human Genome Project was limited to sequencing just the euchromatin DNA (which accounts for just over 93% of the human genome). For technical reasons, too, no attempt was made to fully sequence the multi-megabase ribosomal DNA clusters (the arrays of tandem 45S rRNA repeats that specify the 28S, 5.8S, and 18S rRNAs). Because of various difficulties there are also significant gaps (missing sequence) within the euchromatin DNA sequence plus unplaced contigs.

Recall from Section 7.1 that the nuclear genome component of the human genome reference sequence – abbreviated as the nuclear reference sequence below – is not a single haploid genome from a single individual. That would have allowed relatively easy assembly of the genome sequence (the ideal assembly involves a **golden path**, a unique clone tiling path that can be reduced to one nonredundant haploid representation of the genome). Instead, the DNA clones contributing to the initial reference sequence came from blood-cell samples provided by a selected set of anonymous human donors, each with two different haploid genomes. As a result, the reference nuclear sequence is a mosaic with pieces of sequence from different individuals, each with two parental haplotypes (but is nevertheless dominated by the contribution from a single individual—see **Box 7.3**). And, of course, 24 chromosomal DNA molecules are represented in the nuclear reference sequence, instead of the 23 DNA molecules in haploid human cells.

The ideal nuclear reference sequence should be made up exclusively of *clone contigs*, ordered series of clones having overlapping DNA sequences without gaps. The reality, however, is that the reference sequence consists of multiple *scaffolds*, each consisting of an ordered and oriented set of contigs, but with some gaps between constituent contigs. Although a scaffold has gaps, however, there must be some evidence to support the order and orientation of its constituent contigs, with approximate estimates made for the gap size(s) within the scaffold.

The "finished" human genome sequence reported in 2003 had a few hundred gaps in the euchromatin DNA sequence, often occurring at regions prone to structural variation—large-scale changes involving deletions, duplications, insertions, or inversions. At subchromosomal regions, where the reference sequence is derived from DNA coming from different haplotypes, structural differences between the haplotypes may cause problems: one haplotype may have a sizeable deletion, for example, so that there is no sequence to form an overlap with a neighboring clone from a different haplotype. We now know that structural variation is much more common than first anticipated: the original human genome reference sequence was subsequently found to lack many sequences present in some humans because the source DNAs for genome sequencing had some large deletion alleles at several loci. About 1.4% of the genome sequence of James D Watson, for example, could not be matched to the current reference sequence at that time, hg18.

To address these difficulties, the Genome Reference Consortium (http://genomereference.org) was formed in 2007, and various improvements to the reference sequence have been made. Some gaps have been closed, and revisions are made periodically by adding scaffold sequence *patches* that are of two types:

(a) FIX patches, which are intended to correct errors in an assembly;

(b) NOVEL patches, which add alternative loci. Certain regions of the genome show highly significant and frequent structural variation, and so additional alternative sequences are included at some loci in order to have a more representative reference sequence. In certain subchromosomal regions, therefore, the reference sequence now has multiple alternative haplotypes, coming from additional human donors.

At the time of writing, the most recent human genome reference sequence is GRCh38.p12 (**G**enome **R**eference **C**onsortium **h**uman assembly **38**, **p**atch release **12**) which was released in December 2017 and can be downloaded at https://www.ncbi.nlm.nih.gov/genome/guide/human/. As well as gaps between contigs within scaffolds, there are 349 gaps between scaffolds. Additionally, there remain two types of unassigned DNA sequence: *unlocalized sequences* (DNA sequences associated with a specific chromosome but whose order or orientation on that chromosome is unknown) and *unplaced sequences* (sequences found in the assembly but not associated with a specific chromosome). In the GRCh38.p12 reference sequence, the ***primary assembly***—the collection of assembled chromosomal DNA sequences plus unlocalized and unplaced sequences—has a total length of 3,099,706,404 nucleotides including the length of all gaps (the cumulative length of gaps is 151,122,679 nucleotides).

The total GRCh38.p12 reference sequence has 3,257,319,537 nucleotides because in addition to the nearly 3.1Gb in the primary assembly sequence and the 16,569 nucleotides in mtDNA, it has just over 48 Mb of sequence representing 70 FIX patches and 109.5 Mb of multiple alternate reference loci. In the latter case, for example, variation in the major histocompatibility complex (MHC) is represented by eight different haplotypes, not just one. For some information on accessing alternate human sequence, see the Ensembl blog at www.ensembl.info/2011/05/20/accessing-non-reference-sequences-in-human/

The number of gaps is gradually being reduced, notably with the help of ultra long-read nanopore sequencing (PMID 29431738). Limitations in DNA technology has meant that progress has been slow, but next-generation optical mapping techniques (such as Bionano Genomics's Saphyr system) are designed to offer high-throughput, high-resolution mapping of individual megabase-sized DNA molecules. Progress has also recently been made in assembling DNA sequences in heterochromatic regions of our genome (which were off limits in the Human Genome Project). Lack of human centromere sequence assemblies used to be simply acknowledged in the reference sequence as standard gaps of 3 Mb, but in 2014, the first centromere reference models were produced when two haploid human satellite arrays were modeled on the X and Y chromosomes (PMID 24501022), and reference centromere models have now been produced for each human chromosome. A significant recent advance has been the linear assembly of the centromere DNA on the Y chromosome (PMID 29553574).

Progress is also being made in developing sequence assemblies for the short arms of the acrocentric chromosomes (13, 14, 15, 21 and 22), including the *nucleolar organizer regions* (*NORs*) that contain tandem transcription units specifying 28S, 18S and 5.8S rRNA (ribosomal DNA repeats). Because the very similar NORs on the five acrocentric chromosomes associate to form a chromosomal structure around which a nucleolus can form, sequence exchanges are common between the ribosomal DNA repeats on different acrocentric chromosomes. To avoid the complication of interchromosomal sequence exchange, human–mouse hybrid cells with a single acrocentric human chromosome are used to prepare DNA for sequence assemblies. Thereafter, very long human DNA templates can be prepared using methods such as transformation-associated recombination in yeast, followed by long-read DNA sequencing of selected template DNAs (PMID 29788454).

A quick tour of some electronic resources used to interrogate the human genome sequence and gene products

Numerous databases and genome browsers allow users to gain information about the human genome, and this section will necessarily be selective. There are different electronic gateways to access information on the human genome, human genes, and the gene products (**Table 9.8**). If the starting point of interest is a gene or gene product, the HGNC portal organized by the HUGO Gene Nomenclature Committee has a simple, user-friendly architecture; we describe some features and components of that gateway below. More comprehensive coverage of electronic resources is available through NCBI's Human Genome Resources (see **Table 9.8**).

Gene nomenclature and the HGNC gateway

Gene symbols for human genes are allocated by the HGNC. They typically have between three and seven characters, and are displayed in italicized uppercase, such as *HBB* (hemoglobin beta subunit), *CFTR* (cystic fibrosis transmembrane regulator), or *RN7SL* (7SL RNA). Mitochondrial genes are prefixed by *MT-* (for example, *MT-RNR1* is the mitochondrial 12S rRNA gene). Pseudogenes normally have a symbol that is the same as a related functional gene, but followed by a P, or by a P followed by a number (for example, *CFTRP3* is one of three pseudogenes related to the *CFTR* gene). Note that gene symbols

TABLE 9.8 SOME OF THE PRINCIPAL ELECTRONIC RESOURCES FOR INTERROGATING THE HUMAN GENOME, HUMAN GENES, AND GENE PRODUCTS

Resource use	Popular resources	Website address
Gateways to multiple electronic resources	Human Genome Resources HGNC portal Entrez – NCBI	https://www.ncbi.nlm.nih.gov/genome/guide/human/ https://www.genenames.org/ https://www.ncbi.nlm.nih.gov/gquery/
Reference nucleotide and protein sequences	RefSeq RefSeqGene	https://www.ncbi.nlm.nih.gov/refseq/ https://www.ncbi.nlm.nih.gov/refseq/rsg/
Identifying related sequences and homologs	BLAST programs BLAT HomoloGene HCOP (orthology predictions)	https://blast.ncbi.nlm.nih.gov/Blast.cgi https://genome.ucsc.edu/cgi-bin/hgBlat?command=start https://www.ncbi.nlm.nih.gov/homologene https://www.genenames.org/cgi-bin/hcop
Protein sequence analysis	UniProt InterPro	https://www.uniprot.org https://www.ebi.ac.uk/interpro/
Genome browsers	Ensembl NCBI Genome Data Viewer UCSC Genome Browser	https://www.ensembl.org/ https://www.ncbi.nlm.nih.gov/genome/gdv/ https://genome.ucsc.edu/
Genome annotation	GENCODE NCBI Annotation (release 108)	https://www.gencodegenes.org/ https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/108/

For further descriptions of individual resources, see main text. HGNC, the HUGO (Human Genome Organization) Gene Nomenclature Committee; NCBI, US National Center for Biotechnology Information; UCSC, University of California at Santa Cruz.

for mouse and rat genes have just the first letter in uppercase (for example, the mouse and rat orthologs of the human *CFTR* gene are each given the symbol *Cftr*).

The HGNC portal at www.genenames.org has links to many assorted databases and browsers. It can be interrogated by using as a query a gene symbol, if known, or descriptive text for an associated gene product or disease. For example, entering hemoglobin as the search term returns a list of genes encoding the different subunits of all forms of human hemoglobin, and entering cystic fibrosis yields different results with the first entry being the cystic fibrosis gene *CFTR*. Selecting a gene symbol such as *CFTR* opens the way to an extraordinary amount of related information through linked databases and genome browsers (**Figure 9.4**). The HGNC portal can also be used to identify members of gene families: see the Gene Families Index at https://www.genenames.org/cgi-bin/genefamilies/

General databases storing nucleotide and protein sequences

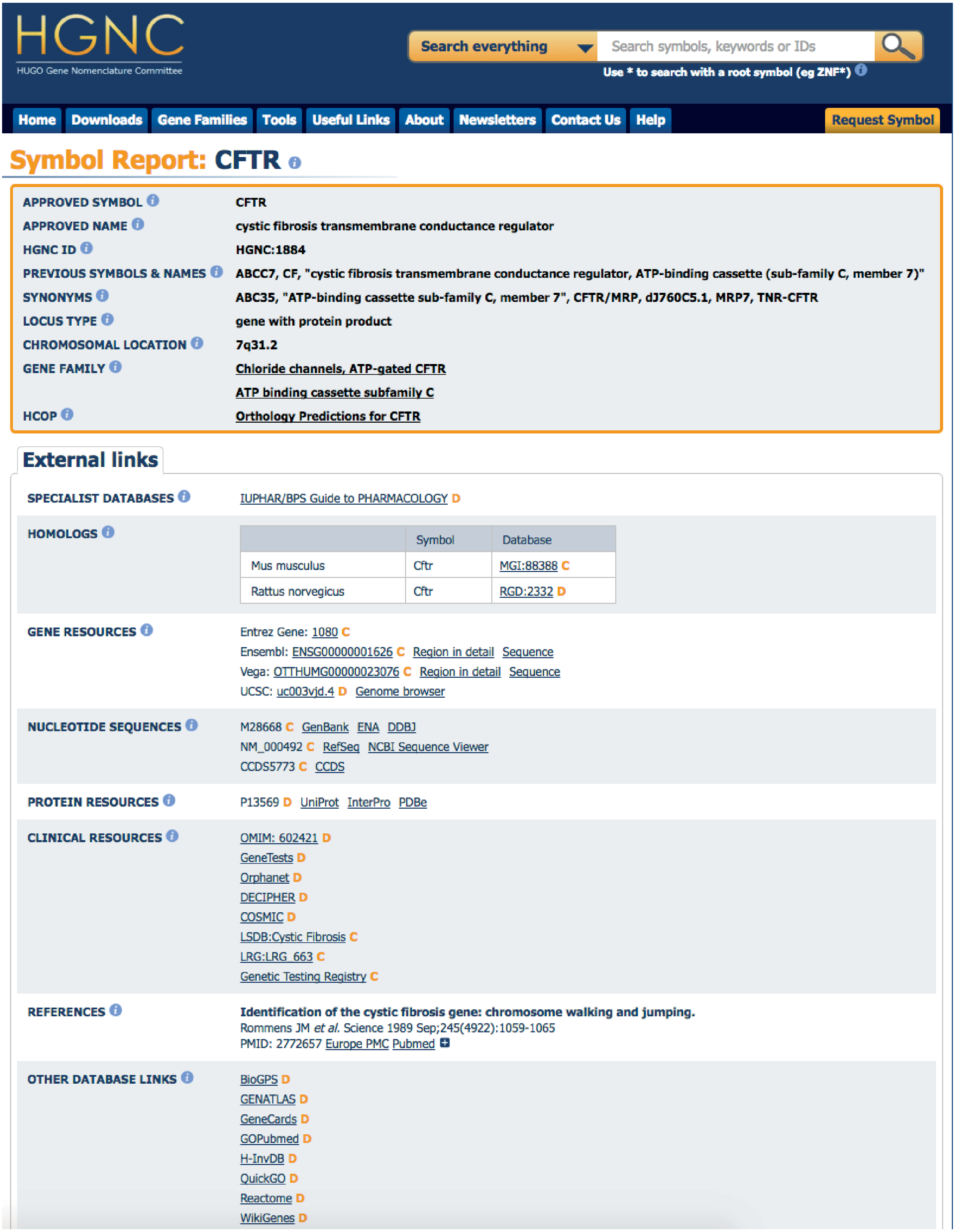
In the example of **Figure 9.4**, the external link for "NUCLEOTIDE SEQUENCES" begins with general nucleotide sequence databases that are part of the International Nucleotide Sequence Database Collaboration (INSDC), comprising GenBank, the European Nucleotide Archive (ENA), and the DNA Database of Japan (DDBJ). The ID number for the *CFTR* mRNA sequence in these databases is M28668, and the 6129-nucleotide sequence is presented with coordinates for the coding sequence and a translation of the open reading frame presented in the single-letter amino acid code.

The general nucleotide sequence databases contain redundant sequences (with sometimes many entries for the same sequence from independent DNA clones, some having partial sequences, some full length). To make it easier to find a complete sequence of interest, the NCBI RefSeq database was established to provide a comprehensive, nonredundant, and well-annotated set of reference sequences for different species. RefSeq reference transcripts have ID numbers beginning with NM_ for mRNA, and NR_ for noncoding RNA; proteins associated with NM_ transcripts have ID numbers beginning with NP_ . The RefSeqGene database stores gene reference sequences.

Finding related nucleotide and protein sequences

Sequences evolutionarily related to a query sequence are identifiable by sequence comparison. The suite of BLAST programs at the NCBI can scan all nucleotide sequences in the general nucleotide databases and translations of nucleotide sequences to identify other nucleotide and protein sequences that are significantly related to a nucleotide

Figure 9.4 Getting a wealth of information on a selected human gene starting from the HGNC portal at www.genenames.org. The figure shows the output displayed after selecting *CFTR*, the human cystic fibrosis gene. The box outlined in orange at top shows basic items of information maintained by HGNC, including related sequences (chloride channels of the same type as the CFTR protein, other members of the ATP binding cassette subfamily C) plus equivalent genes identified in other species by HCOP, the HGNC's Comparison of Orthology Predictions tool. The "External links" section allows access to numerous databases and genome browsers. The former includes databases of nucleotide and protein sequences, and databases with associated clinical information (including the general OMIM resource and a locus-specific database, LSDB, that is devoted to cystic fibrosis). The genome browsers, which allow detailed examination of the structure of genes and their products and the ability to download selected sequence components, are listed under the "GENE RESOURCES" catalog. They include the NCBI's Map Viewer available via the "Entrez Gene" subheading, the Ensembl browser, and the UCSC (University of California at Santa Cruz) Genome Browser—see text for more detail.



Symbol Report: CFTR

APPROVED SYMBOL	CFTR
APPROVED NAME	cystic fibrosis transmembrane conductance regulator
HGNC ID	HGNC:1884
PREVIOUS SYMBOLS & NAMES	ABCC7, CF, "cystic fibrosis transmembrane conductance regulator, ATP-binding cassette (sub-family C, member 7)"
SYNONYMS	ABC35, "ATP-binding cassette sub-family C, member 7", CFTR/MRP, dJ760C5.1, MRP7, TNR-CFTR
LOCUS TYPE	gene with protein product
CHROMOSOMAL LOCATION	7q31.2
GENE FAMILY	Chloride channels, ATP-gated CFTR ATP binding cassette subfamily C
HCOP	Orthology Predictions for CFTR

External links

SPECIALIST DATABASES	IUPHAR/BPS Guide to PHARMACOLOGY D
HOMOLOGS	

	Symbol	Database
Mus musculus	Cftr	MGI:88388 C
Rattus norvegicus	Cftr	RGD:2332 D

GENE RESOURCES	Entrez Gene: 1080 C Ensembl: ENSG00000001626 C Region in detail Sequence Vega: OTTHUMG00000023076 C Region in detail Sequence UCSC: uc003vjd.4 D Genome browser
NUCLEOTIDE SEQUENCES	M28668 C GenBank ENA DDBJ NM_000492 C RefSeq NCBI Sequence Viewer CCDS5773 C CCDS
PROTEIN RESOURCES	P13569 D UniProt InterPro PDBe
CLINICAL RESOURCES	OMIM: 602421 D GeneTests D Orphanet D DECIPHER D COSMIC D LSDB:Cystic Fibrosis C LRG:LRG_663 C Genetic Testing Registry C
REFERENCES	**Identification of the cystic fibrosis gene: chromosome walking and jumping.** Rommens JM *et al.* Science 1989 Sep;245(4922):1059-1065 PMID: 2772657 Europe PMC Pubmed ⊞
OTHER DATABASE LINKS	BioGPS D GENATLAS D GeneCards D GOPubmed D H-InvDB D QuickGO D Reactome D WikiGenes D

or protein query. Significant homology may be apparent across the length of the query sequence or be limited to a region, such as a conserved protein domain or some other shared sequence.

The BLAT program, available at sites such as the UCSC (University of California at Santa Cruz) Genome Browser, allows rapid sequence searching across whole genomes. Query nucleotide sequences (up to a total of 25,000 nucleotides) and query protein sequences (up to 10,000 amino acids) can be entered to search for homologous sequences across the human genome, or the genome of any of multiple model organisms. The output lists significant hits, with given chromosome coordinates and sequence alignments.

The sequence comparisons include searching to find orthologs in other species using programs such as HCOP (the HGNC's Comparison of Orthology Predictions tool) and HomoloGene. The latter provides pairwise and multisequence alignments of protein sequences from identified orthologs.

Protein characteristics and analyses

The "PROTEIN RESOURCES" component in **Figure 9.4** provides access to programs that record or analyze different functional characteristics of the predicted protein of a gene of interest, including:

- UniProt—records associated functional information including presence of conserved domains, disease-associated variants, expression characteristics, protein–protein interactions, and so on;
- InterPro—scans for the presence of related sequences including conserved domains and repeats that might be shared with other proteins, and seeks to assign proteins to membership of protein families.

Genome browsers

Navigating a sequenced genome is effectively assisted by programs with graphical user interfaces to portray genome information for selected chromosomes and subchromosomal regions. The characteristics (genes, exons, transcripts, and so on) of a selected human chromosome or chromosome region can be tracked, moving from large scale to nucleotide scale, with click-over facilities to identify the characteristics and download the sequences of genes and associated exons, RNAs, and proteins. The principal browsers are listed in **Table 9.8** and are described in greater detail in Section 7.1.

9.2 GENE ORGANIZATION AND DISTRIBUTION IN THE HUMAN GENOME

In this section we focus on genes, and the repetition of gene segments and duplication of whole genes, forming **gene families**. DNA sequence repetition is generally rare in the small genomes of prokaryotes and mitochondria (which are typically tightly packed with genetic information that is presented in extremely economical forms). But in the large nuclear genomes of eukaryotes, and especially in complex multicellular organisms, there is less constraint in packing the genetic information, and repetitive DNA sequences are a striking feature of these genomes, both in terms of abundance and importance. Many human genes, both protein-coding genes and RNA genes, are members of multigene families that can vary enormously in terms of copy number and distribution. And, as described in Section 9.3, many non-genic sequences are repeated. As detailed below, repeated DNA sequences arise by one or more of a variety of different mechanisms that result in gene duplication.

Human genes show enormous variation in size and internal organization

Genes in simple organisms such as bacteria are comparatively similar in size and are usually very short (typically about 1 kb long). The genes of complex eukaryotes, however, can show huge size variation. As measured by the distance from the beginning of the first exon to the end of the last exon, the longest human protein-coding gene is *CNTNAP2,* which at 2.305 Mb just pips the dystrophin gene, the previous record holder at 2.242 Mb. Using the same definition of gene length, the mean length of a human protein-coding gene is 64.9 kb (but the median length is 26.0 kb), but there is enormous diversity as shown in **Table 9.9**. Although there is generally a direct correlation between gene and product sizes, there are some striking anomalies. For example, the 2.24 Mb dystrophin

TABLE 9.9 HUMAN PROTEIN-CODING GENES SHOW MARKED VARIATION IN SIZE AND ORGANIZATION

Human protein	Number of amino acids[a]	Gene size (kb)[b]	Number of exons	Coding DNA (%)	Average size of exon (bp)	Average size of intron (bp)
SRY	204	0.9	1	94	850	n/a
β-globin	146	1.6	3	38	150	490
p16 tumor suppressor	156	7.4	3	17	406	3064
Collagen type VII	2928	31	118	29	77	190
p53 tumor suppressor	393	39	10	6	236	3076
Apolipoprotein B	4563	45	29	31	487	1103
Phenylalanine hydroxylase	452	90	26	3	96	3500
Factor VIII	2351	186	26	3	375	7100
Huntingtin	3144	189	67	8	201	2361
RB1 tumor suppressor	928	198	27	2.4	179	6668
CFTR	1480	250	27	2.4	227	9100
Titin	34,350	283	363	40	315	466
Utrophin	3433	567	74	2.2	168	7464
Dystrophin	3685	2242	79	0.6	180	28,560

[a] For the largest isoform. [b] As measured from the start of the first exon to the end of the last exon for multiexon genes. Note the extraordinarily high exon content and comparatively small intron sizes in the genes encoding type VII collagen and titin. In addition to *SRY*, the major male-determining gene, other intronless protein-coding genes in the nuclear genome include genes encoding other SOX proteins, interferons, histones, many G-protein-coupled receptors, heat shock proteins, many ribonucleases, various neurotransmitter receptors and hormone receptors, and also retrogenes (see **Box 9.2**). CFTR, cystic fibrosis transmembrane receptor; n/a, not applicable.

gene, 50 times the length of the apolipoprotein B gene, makes a protein that is almost 900 amino acids smaller than apolipoprotein B.

For human protein-coding genes, the mean exon size is 268 nucleotides (but median size is 129 nucleotides), with an average of nine exons per gene. As genes get larger, exon size remains fairly constant. Note, however, that the terminal exon can occasionally be very long; if we discount the first and last exons (which quite often are mostly untranslated sequence), the mean length of an internal exon in a human protein-coding gene is 147 nucleotides (median length 121 nucleotides). Unlike exons, the introns of human protein-coding genes vary enormously in size (from a few nucleotides to over 100,000 nucleotides long), and intron size generally correlates with gene length (but there are some notable exceptions—see **Table 9.9**).

Human protein-coding genes generally contain introns, but a small minority lack introns and are generally small genes (see the footnote to **Table 9.9** for some examples). For those that do possess introns, there is an inverse correlation between gene size and fraction of coding DNA because large genes tend to have large introns while exons tend to be more uniform in size. Because transcribing long introns is costly in both energy and time (16 hours are needed to transcribe the 2.24 Mb dystrophin gene), very highly expressed genes often have short introns, or no introns at all.

Sizes and intron content of RNA genes

Many RNA genes have been identified very recently, and statistics on RNA gene size and intron content are being accumulated. Genes encoding lncRNAs can be tens of kilobases long and often contain introns (the exons of these genes show a mean length of 683 nucleotides, but a median length of 129 nucleotides). At the transcript level, lncRNA genes show less diversity (a mean of 1.8 transcripts per gene) than protein-coding genes (close to 7 transcripts per gene). At the other end of the scale are tiny RNA genes, including those specifying miRNAs that often lack introns. However, as described above, there is some uncertainty as to how to define an RNA gene.

Variable gene density and overlapping transcription units in the nuclear genome

Simple genomes have high gene densities (roughly one per 0.5, 1, and 2 kb for the genomes of human mitochondria, *Escherichia coli*, and *Saccharomyces cerevisiae*, respectively) and often show examples of partly overlapping genes. Different reading frames may be used, sometimes from the same sense strand. In complex organisms, such as humans, genes are much bigger and there is less clustering of protein-coding sequences.

Gene density varies enormously from chromosome to chromosome. The constitutive heterochromatin regions of somatic cells are almost devoid of genes (but a few genes in these regions may be expressed in germ-line cells). Within the euchromatic portion of the genome, gene density can also vary substantially between chromosomal regions, and also between whole chromosomes.

The first general insight into how genes are distributed across the human genome was obtained when purified CpG island fractions were hybridized to metaphase chromosomes (CpG islands have long been known to be strongly associated with genes—see **Box 9.1** and **Figure 7.7**). The results indicated that gene density is high in subtelomeric regions, and that chromosomes with high %GC (for example, chromosomes 19 and 22) are gene-rich, whereas others with low % GC (for example, X and 18) are gene-poor. The predictions of differential CpG island density and differential gene density were subsequently confirmed by analyzing the human genome sequence.

This difference in gene density can also be seen with Giemsa staining (G-banding) of chromosomes. Regions with a low %GC content correlate with the darkest G-bands (G-positive bands), and those with a high %GC content correlate with pale bands. Pale G-negative bands are comparatively rich in genes; dark G-positive bands are gene-poor. In chromosomal regions with high gene density, overlapping genes may be found; they are typically transcribed from opposing DNA strands. For example, the class III region of the HLA complex at 6p21.3, located within a G-negative band, has an average gene density of about one gene per 15 kb and is known to contain several examples of partly overlapping genes. In striking contrast, the mammoth dystrophin gene extends over 2.24 Mb of DNA in a dark G-band at Xp21.2 without evidence for any other protein-coding gene in this region.

Genes-within-genes

Whole-genome analyses show that about 9% of human protein-coding genes overlap another such gene, and >90% of the overlaps involve genes transcribed from opposing strands. Sometimes the overlaps are partial, but in other cases small genes are located within larger genes. The *NF1* (neurofibromatosis type I) gene, for example, has three small internal genes transcribed from promoters on the opposite strand (**Figure 9.5A**).

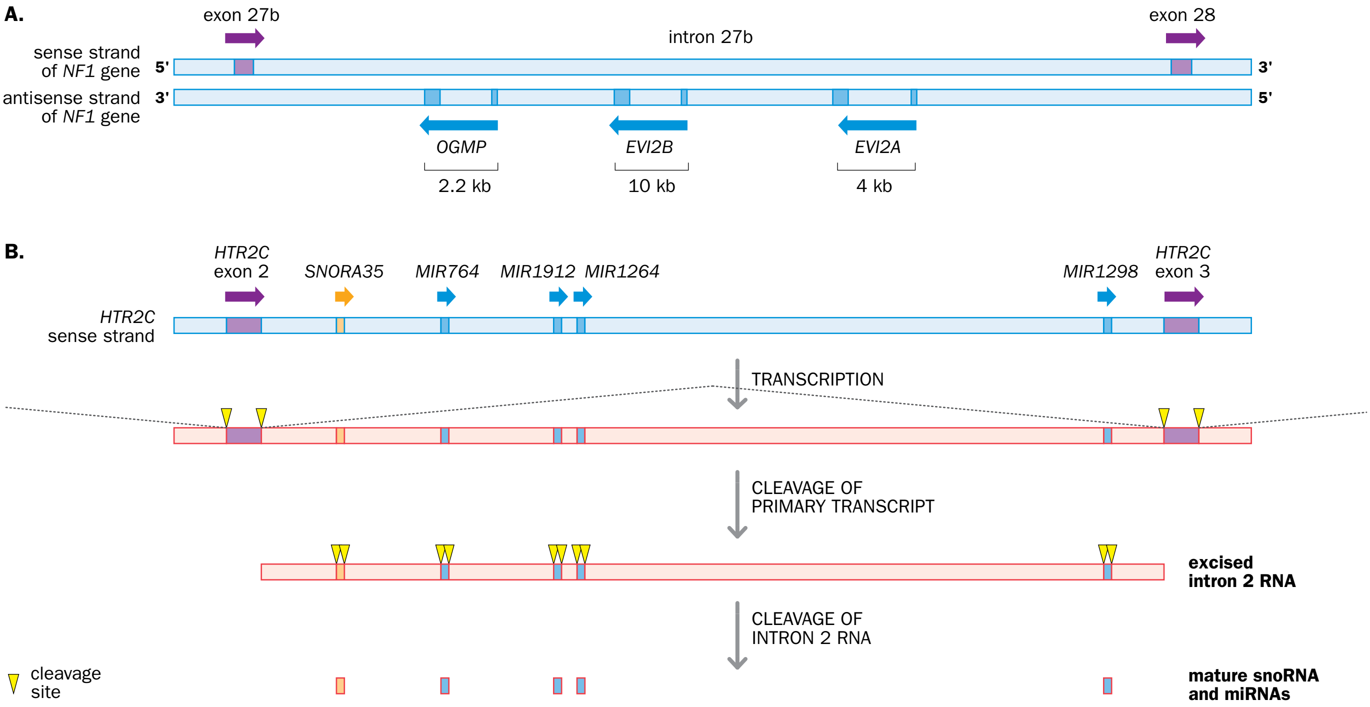


Figure 9.5 Different types of organization for genes-within-genes. (**A**) Small protein-coding genes can be located on the antisense strand of a larger gene. The 60.5 kb intron 27b of the *NF1* (neurofibromatosis type I) gene contains three small internal genes, all transcribed from the opposite DNA strand. The internal genes, each with two exons and not drawn to scale, are *OGMP* (oligodendrocyte myelin glycoprotein) and *EVI2A* and *EVI2B* (human homologs of murine genes thought to be involved in leukemogenesis and located at ecotropic viral integration sites). (**B**) Small RNA genes can be located on the sense strand of a larger gene. Intron 2 of the *HTR2C* gene (which encodes the 5-hydroxytryptamine receptor 2C protein) contains one snoRNA gene and four miRNA genes, each of which is co-transcribed as part of the primary *HTR2C* transcript, and then processed from the excised intron 2 RNA sequence.

Small RNA genes are also frequently found within both protein-coding genes and long noncoding RNA genes, usually within introns. These genes are often located on the *sense strand* of the larger gene (usually described as a *host gene*), and their expression is dictated by the host gene's promoter. For example, almost all snoRNAs, and many miRNAs and scaRNAs, are made by processing of the intron transcripts of larger host genes (see **Figure 9.5B** for an example).

Different origins for duplicated genes in the human genome

Gene duplication has been an important driver in the evolution of functional complexity and the origin of increasingly complex organisms. As a result, duplicated genes (and duplicated segments of coding sequence) are a common feature of animal genomes, especially large vertebrate genomes. In the human genome, the resulting multigene families have from two to many hundreds of gene copies that may be clustered together in one subchromosomal location, or dispersed over several chromosomal locations—we consider different types of gene-family organization below.

One type of DNA duplication arose at the birth of eukaryotes when the fusion of two cells gave rise to two genomes that would become the nuclear and mitochondrial genomes. A few of the original gene duplicates were retained. Subsequently, a variety of different DNA duplication mechanisms have led to more recent gene duplication and the formation of gene families, as listed below.

- *Whole-genome duplication (WGD)*. The most recent WGDs in mammalian lineages appear to have occurred during the early evolution of chordates. A couple of WGD events occurring within the approximate interval of 525–875 million years ago can explain why there are four vertebrate HOX gene clusters, but many of the duplicated genes were not retained after whole-genome duplication.
- *Tandem duplication*. Crossover can occasionally occur between unequally aligned chromatids, either on homologous chromosomes (**unequal crossover**) or on the same chromosome (unequal sister chromatid exchange). The repeated segment may be just a few hundred nucleotides long or may be quite large and contain from one to several genes. Initial duplication events produce tandem repeats (that is, the head of one repeat joins the tail of its neighbor: →→). **Figure 9.6** shows the general mechanism of tandem gene duplication. Over a long (evolutionary) timescale the duplicated sequences can be separated on the same chromosome (by a DNA insertion or inversion).
- *Duplicative transposition by recombination*. Certain euchromatic regions, notably those close to the centromeres and telomeres (pericentromeric and subtelomeric regions, respectively), are comparatively unstable, and are prone to recombination with other chromosomes. As a result, segments of DNA containing multiple genes can be transposed to another location (but with copies of the transposed genes retained on nonrecombining homologs). We give the examples of the *NFI* and *PKD1* multigene families below.
- *RNA-directed duplicative transposition*. Reverse transcriptases within the cell are used to make a cDNA copy of an RNA transcript, whereupon the cDNA copy integrates into a new chromosomal location. The same type of mechanism can often lead to defective gene copies, and will be described below.

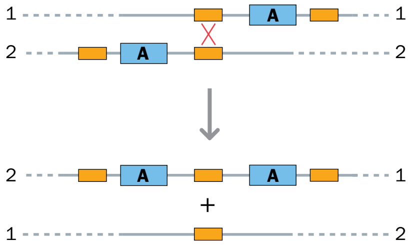


Figure 9.6 Tandem gene duplication arising by unequal crossover. Homologous chromatids 1 and 2 have paired out of register, so that gene A on one chromatid is not directly opposite gene A on its partner chromatid. As shown here, the mispairing of the chromatids might sometimes be stabilized by closely related members of an interspersed repeated DNA family such as Alu repeats (small orange boxes). A crossover event (red X) in the mispaired region can result in tandem duplication of gene A, by joining of the left part of chromatid 2 to the right part of chromatid 1. (The other product of the crossover, formed by joining of the left part of chromatid 1 to the right part of chromatid 2, would lack gene A). A crossover event like this may result from unequal crossover (chromatids 1 and 2 are misaligned non-sister chromatids on homologous chromosomes) or from unequal sister chromatid exchange (chromatids 1 and 2 are misaligned sister chromatids).

Segmental duplication

More than 5% of the euchromatic portion of the human genome is accounted for by 400 large blocks of duplicated DNA where the sequence identity between the DNA copies is very high (often more than 95% identity). The very high degree of sequence identity across such long sequences is the result of evolutionarily recent DNA duplication, occurring within the past 40 million years of primate evolution, and this type of recent duplication is often described as **segmental duplication**. It includes both intrachromosomal duplications (see **Figure 9.7** for an example) and also interchromosomal duplications. Segmental duplications are important contributors to copy number variation and to chromosomal re-arrangements leading to disease and rapid gene innovation.

Functionally inactive copies of gene sequences

Gene duplication is driven by evolutionary pressure to develop specialized functions. We owe our developed sense of smell to having a family of close to 1000 olfactory receptor genes that collectively provide a large range of different olfactory receptor proteins. However, duplication is not always successful in producing additional functional genes. Some duplicated genes degenerate and become nonfunctional gene copies called **pseudogenes** (in the case of the olfactory receptor gene family, for example, the majority of the sequences are inactive—see **Box 9.2** for an overview of human pseudogenes).

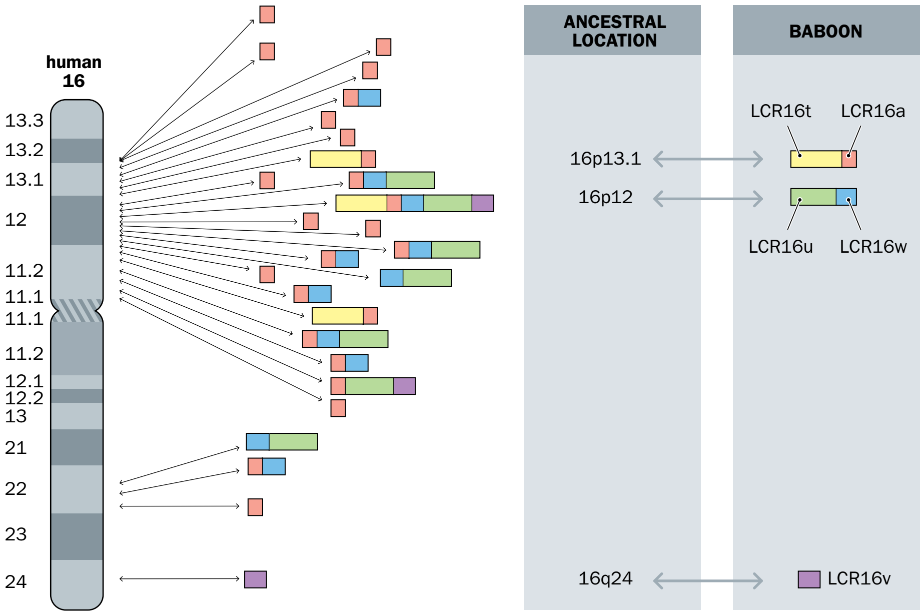


Figure 9.7 An example of intrachromosomal segmental duplication in the human genome. Shown at the left are locations on human chromosome 16 of five types of chromosome 16 low-copy-number repeats (LCR16). The repeats vary in size from 19 kb (LCR16a) to 75 kb (LCR16t), and in copy number from 3 repeats (LCR16t, LCR16v) to over 20 repeats (LCR16a). Each type of repeat unit has a single counterpart in the baboon genome (where the LCR16t and LCR16a are neighboring sequences, as are the LCR16u and LCR16w repeats). Since divergence of the baboon lineage from lineages leading to humans and the great apes, approximately 30 million years ago, there has been a rapid series of duplications of the five sequences in the human and great ape genomes. In the human genome, the five LCR16 repeat families are believed to have arisen by duplication of single repeats originally located at 16p13.11, 16p12.3, and 16q24.2. Note that there has also been an expansion of these sequences in the chimpanzee and gorilla, but there are species differences regarding both the copy number of the repeats and also their chromosomal locations. (Adapted from Marques-Bonet T & Eichler EE [2009] *Cold Spring Harb Symp Quant Biol* **74**:355–362; PMID 19717539. With permission from Cold Spring Harbor Laboratory Press.)

BOX 9.2 PSEUDOGENES, RETROGENES, AND RNA-DIRECTED COPYING OF PROTEIN-CODING GENES

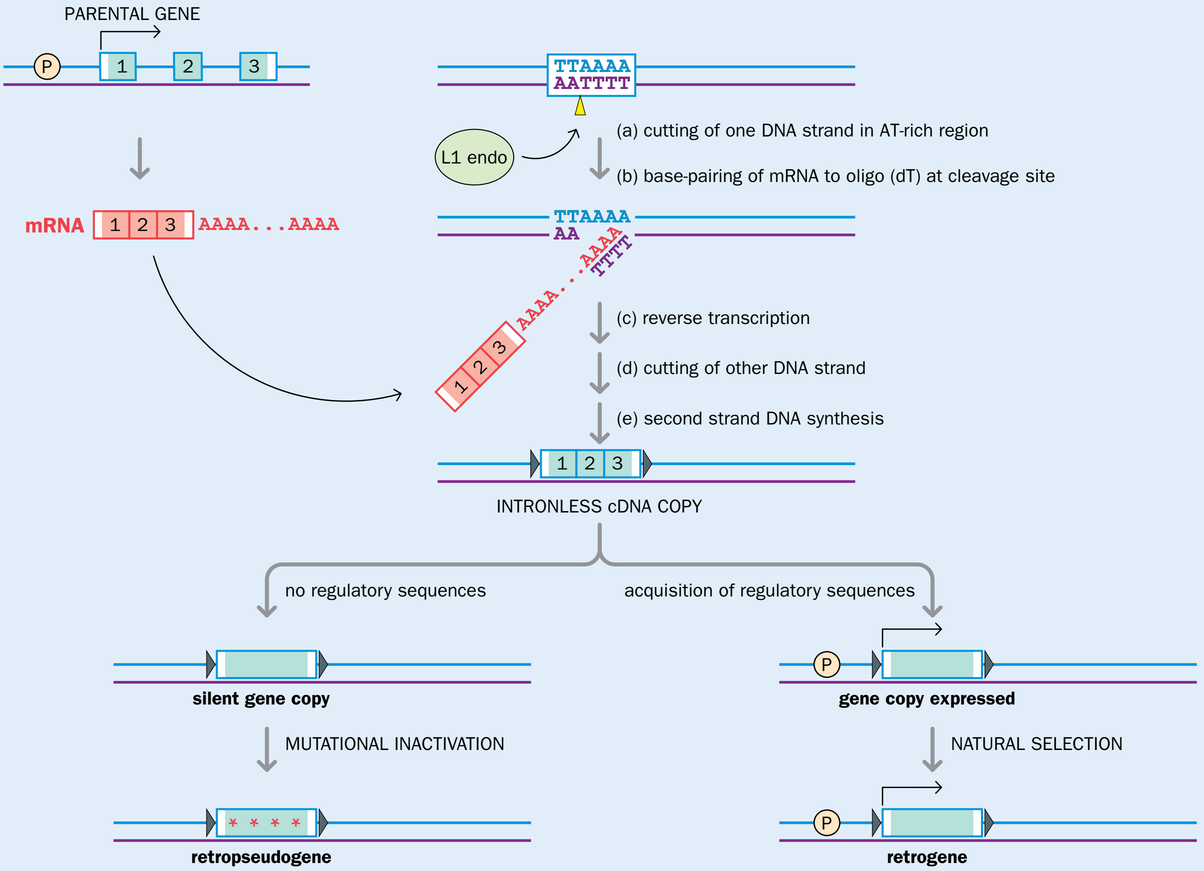
At the time of writing (May 2018), approximately 15,000 pseudogenes were recognized in the human genome (the Pseudogene.org database has a component devoted to human pseudogenes at www.pseudogene.org/Human/). Nearly all of them are copies made of sequences derived from a parent gene still present in the genome, but acquisition of deleterious mutations has meant that the pseudogene is unable to carry out the same function as the parent gene. A pseudogene that acquires deleterious mutations, and becomes functionless, will not be maintained by natural selection. Through periodic deletions, it can lose segments of its sequence to give a truncated sequence and ultimately a gene fragment (examples can be found in Figure 9.9). Not all pseudogenes are functionless, however; some that originated by copying a protein-coding gene are transcribed and, although they cannot make a protein, they may nevertheless be functional. The example of *PTENP1*, which produces a regulatory RNA, is described in Table 9.7; for a list of transcribed pseudogenes with published functions, see http://www.genenames.org/cgi-bin/genefamilies/set/859.

The catalog of currently recognized pseudogenes is focused on sequences copied from protein-coding genes (inactivating mutations, such as frameshifting and nonsense mutations, are more readily identified than those in RNA pseudogenes; however, some RNA pseudogenes can reach very high copy numbers—see Table 9.12). Three classes of pseudogene are recognized, as listed below.

- ***Non-processed pseudogenes.*** They account for nearly one-quarter of human pseudogenes, and are direct copies made at the DNA level (such as by tandem duplication; see Figure 9.6). A full-length pseudogene of this type has sequences corresponding to any introns and immediately neighboring regulatory sequences such as the promoter. Note that the nuclear mitochondrial DNA sequences (NUMTs) might be considered members of this class, even though they originated from intronless mtDNA copies.
- ***Processed pseudogenes (retropseudogenes).*** They account for nearly three-quarters of human pseudogenes, and are copies of gene transcripts. For protein-coding genes, copies are made of a processed sequence whereby the RNA transcript is first processed to make mRNA and then converted into a cDNA that integrates elsewhere in the genome (Figure 1). As a result, the full-length pseudogene will lack a promoter and other non-exonic regulatory sequences, as well as lacking intron sequences.
- ***Unitary pseudogenes.*** The pseudogene status is simply due to inactivation in recent evolutionary times of a formerly functional gene. The equivalent gene may be functional in multiple related species, such as great apes or other mammals, but the gene has become inactivated in the human lineage. (As described in Chapter 13, gene loss, beginning with natural gene inactivation, is an important driver in evolution in addition to the emergence of new genes).

A variety of intronless **retrogenes** are known to have testis-specific expression patterns and are typically autosomal homologs of an intron-containing X-linked gene (Table 1). In those cases, expression of the retrogene can compensate for lack of expression of the X-linked parental sequences in the testis during male meiosis when the paired X and Y chromosomes form the highly condensed, transcriptionally inactive XY body. Although gene expression on both the X and Y chromosomes is shut down in the XY body, autosomal retrogenes continue to supply important enzymes and proteins needed by the cell that would normally be synthesized by their parent genes on the X chromosome.

In addition, a small percentage of human intronless genes are thought to be "orphan retrogenes," retrogenes that have supplanted the parent gene (see Ciomborowska J *et al*. [2013]; PMID 23066043).



Box 9.2 Figure 1 Processed pseudogenes (retropseudogenes) and retrogenes originate by reverse transcription from RNA transcripts. In this example, a protein-coding gene with three exons is transcribed from an upstream promoter (P), and introns are excised from the transcript to yield an mRNA. The resulting mRNA can undergo retrotransposition to produce an intronless cDNA copy at a position elsewhere in the genome, often on a different chromosome from the parental gene. LINE-1 repeats provide the machinery for retrotransposition. The precise details of the mechanism are unclear, but the following sequence is likely. First, a LINE-1 endonuclease (L1 endo) cuts one DNA strand within a preferred target sequence (often TTAAAA/TTTTAA). The poly(A) sequence of the mRNA allows pairing with a T-rich sequence at the target sequence and the LINE-1 reverse transcriptase makes a cDNA copy of the mRNA. Thereafter the LINE-1 endonuclease cuts the other strand about 7–20 nucleotides downstream, second strand DNA synthesis occurs and the cDNA integrates and becomes flanked by short direct repeats (black arrowheads; for a detailed model of the integration process interested readers should consult Figure 1 of PMID 19030023). For the integrated cDNA to be transmitted to future generations, the integration event must occur in germ-line DNA. Because the integrated cDNA lacks the promoter and regulatory sequences of the parental gene, it usually remains transcriptionally silent, acquires deleterious mutations (red asterisks) in the coding sequence and degenerates into a processed pseudogene (retropseudogene). On some occasions, however, the cDNA copy acquires or develops regulatory sequences (for example, by inserting close to or within another gene) that allow it to be transcribed (arrow). If expression of the cDNA copy is advantageous, it can then be preserved by natural selection as a functional retrogene (see text).

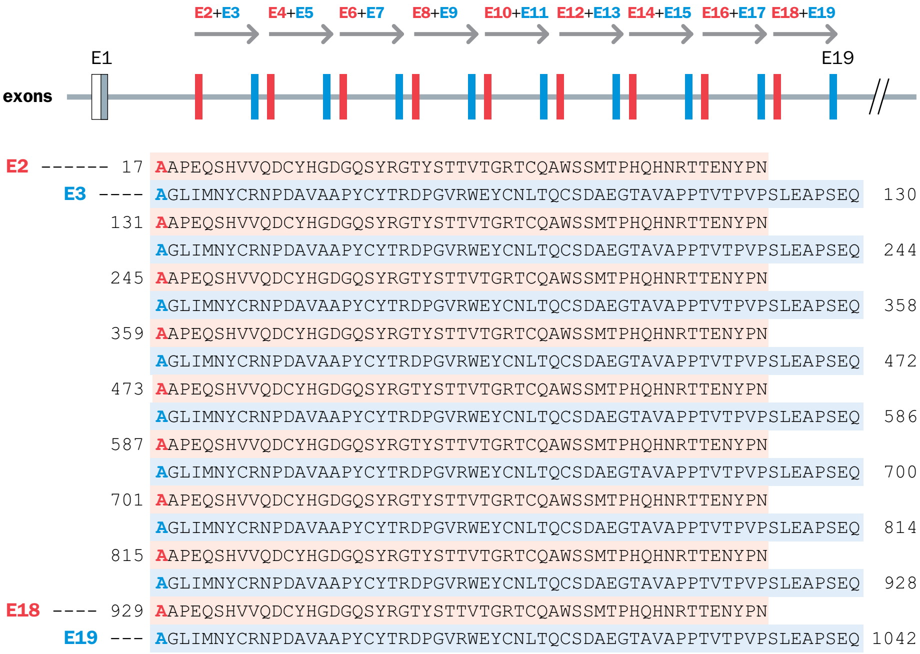
BOX 9.2 TABLE 1 EXAMPLES OF HUMAN INTRONLESS RETROGENES AND X-LINKED INTRON-CONTAINING PARENT GENES

Retrogene	Parent gene	Product
GK2 at 4q13	*GK* at Xp21	Glycerol kinase
GLUD1 at 10q23	*GLUD2* at Xq25	Glutamate dehydrogenase
PDHA2 at 4q22	*PDHA1* at Xp22	Pyruvate dehydrogenase E1 component, alpha subunit
PGK2 at 6p12	*PGK1* at Xq13	Phosphoglycerate kinase
TAF1L at 9p12	*TAF1* at Xq13	TATA box binding protein associated factor, 250 kD

Repetitive functional sequences within genes

Having genes split into exons and introns facilitates tandem exon duplication using mechanisms such as unequal crossover. Tandem repetition of sequences encoding known or assumed protein domains is quite common, and it may be functionally advantageous by providing a more available biological target. Sometimes that leads to a series of repeated exons (or group of exons) that specify a protein domain or other functional unit.

Although the sequence identities between the repeated protein domains are often quite low, they can sometimes be high. A classic example is provided by tandemly repeated kringle domains present in the lipoprotein Lp (a) protein encoded by the *LPA* gene on chromosome 6q26. These domains, which are especially common in blood-clotting and fibrinolytic proteins, are formed as multiple loop structures constrained by three disulfide bridges, and get their name from a Scandinavian pastry of the same shape. Each kringle domain in the Lp (a) protein is about 114 amino acids long and is encoded by a pair of exons; tandem repetition of the exon pair has produced a large number of almost identical kringle repeats (**Figure 9.8**). The high level of sequence identity between the two-exon repeats promotes unequal crossover and striking length polymorphism.



```
E2 ------ 17 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
     E3 ---- AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  130
         131 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  244
         245 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  358
         359 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  472
         473 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  586
         587 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  700
         701 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  814
         815 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
             AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ  928
E18 ---- 929 AAPEQSHVVQDCYHGDGQSYRGTYSTTVTGRTCQAWSSMTPHQHNRTTENYPN
    E19 ---  AGLIMNYCRNPDAVAAPYCYTRDPGVRWEYCNLTQCSDAEGTAVAPPTVTPVPSLEAPSEQ 1042
```

Figure 9.8 Tandem repetition of kringle domains in lipoprotein Lp (a) as a result of exon duplication. The *LPA* gene that encodes lipoprotein Lp (a) is implicated in coronary heart disease and stroke. The Lp (a) protein shows extensive size variation, having variable numbers of tandem 114-amino-acid repeats representing kringle IV domains. For simplicity we show just the first nine kringle domains, which are identical in sequence and span amino acids 17 to 1042, being encoded by exon pairs 2+3, 4+5, 6+7, and so on, up to 18+19. The first amino acid encoded by each exon is always an alanine (shown in bold red or bold blue), and the codon for each such alanine is interrupted by an intron. There are additional, subsequent kringle domains (not shown, but up to as many as 32 copies) and they can show some limited sequence variation.

Different organizations for protein-coding gene families

Protein-coding genes that duplicated recently in evolution have a strong tendency to be clustered at a subchromosomal location and produce proteins with clear structural and functional similarity. By contrast, genes that make proteins with little sequence similarity are typically found at distinct chromosomal locations, and usually on different chromosomes, even if they work very closely together in the same functional pathway, or as part of a common protein. The insulin gene is located on chromosome 11 but the insulin receptor gene is on chromosome 19, for example, and genes encoding the five subunits of RNA polymerase I are located at 2p, 2q, 6p, 9p, and 13q.

Different classes and subclasses of gene family can be distinguished according to the degree of sequence similarity of the protein products, the gene copy number, the locations of gene members in the genome, and whether duplication events have occurred by making DNA copies of genomic DNA or mRNA. Many protein-coding gene families have pseudogenes. If the genes are clustered at a subchromosomal region, pseudogene copies are of the nonprocessed type. A large-copy-number gene family dispersed on multiple chromosomes is often the result of retrotransposition, and may contain many processed pseudogenes. Some dispersed gene families arise by alternative methods, such as through recombination; in that case, nonprocessed pseudogenes are evident. Different classes of gene family are outlined in **Table 9.10** and in **Figures 9.9** and **9.10**.

TABLE 9.10 DIFFERENT CLASSES OF HUMAN GENE FAMILY ACCORDING TO CHROMOSOMAL DISTRIBUTION AND WHETHER SEQUENCES WERE COPIED AT THE DNA LEVEL OR AT THE RNA LEVEL

Type of gene family	Gene-copying mechanism	Example	Characteristics
Single subchromosomal location	DNA-directed	Class I HLA gene family at 6p21.3	Six genes plus many non-processed genes and gene fragments within 2 Mb (see **Figure 9.9**)
Multiple clusters on different chromosomes	DNA-directed	Globin gene family	14 members on five chromosomes with two major clusters containing at least three non-processed pseudogenes (see **Figure 13.12**)
		HOX gene family	A total of 39 genes organized in four clusters of 9–11 genes at 2q, 7p, 12q, and 17q
		Olfactory gene family	Nearly 1000 genes dispersed across genome, with many of them located in at least 25 large clusters; less than half of the genes are functional
Interspersed on different chromosomes	RNA-directed	Ferritin heavy-chain family	One gene at 11q and at least 27 processed pseudogenes dispersed across the genome
	DNA-directed	PAX gene family	Nine functional genes at nine distinct locations on eight chromosomes; no recognizable pseudogenes
		NF1 (neurofibromatosis type I)	One functional gene at 17q11.2; more than 10 related nonprocessed pseudogenes or gene fragments (see **Figure 9.10**)

Note that whereas the protein products of some gene families show sequence homology across their lengths, some other families are defined simply by highly conserved domains (such as ~60-amino-acid homeodomains in HOX proteins, and ~124-amino-acid paired domains in PAX proteins), or even collections of conserved short sequence motifs.

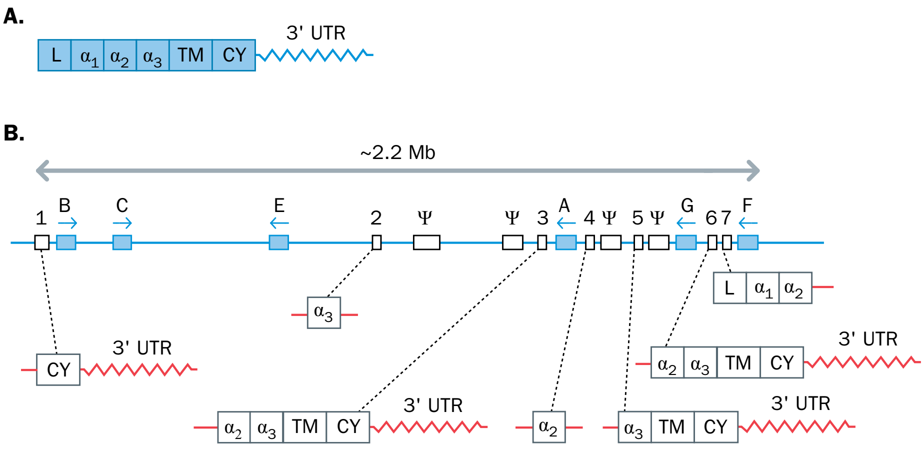


Figure 9.9 The class I HLA gene family at 6p21.3 has multiple nonprocessed pseudogenes and gene fragments. (**A**) Structure of a class I HLA heavy-chain mRNA. The full-length mRNA contains a polypeptide-encoding sequence specifying a leader sequence (L), three extracellular domains (α_1, α_2, α_3,), a transmembrane sequence (TM), a cytoplasmic tail (CY), and a 3′ untranslated region (3′ UTR). The three extracellular domains are each encoded essentially by a single exon. The very small 5′ UTR is not shown. (**B**) Labeled items are: six functional genes (blue boxes; A–C, E–G); four full-length nonprocessed pseudogenes (Ψ); and a variable number of partial gene copies (unfilled boxes), seven in this example, some truncated at the 5′ end (1,3,5,6), some truncated at the 3′ end (7), and some containing single exons (2 and 4).

Organization and distribution of RNA gene families

Long noncoding RNAs are highly heterogeneous, but even if we consider long regulatory RNAs, the examples that we know are divergent in sequence; families of evolutionarily related genes are not so obvious. Genes encoding small noncoding RNAs are a different matter. There are several well-recognized gene families with moderately large numbers of genes, often several hundred genes, that specify RNAs of a particular class with conserved structures.

Some genes present at high copy number are used to make large amounts of essential noncoding RNAs. Take the example of genes making the four cytoplasmic ribosomal RNAs. The 28S, 5.8S, and 18S rRNAs are encoded by a single 13 kb multigenic (28S+5.8S+18S) transcription unit that, along with a 27 kb spacer region, is tandemly repeated to form large ribosomal DNA (rDNA) arrays on the short arms of each of the five human acrocentric chromosomes: 13, 14, 15, 21, and 22. These rDNA arrays (known as **nucleolar organizer regions** because they associate to form a chromosomal structure around which a nucleolus forms), vary significantly in repeat copy number. A haploid

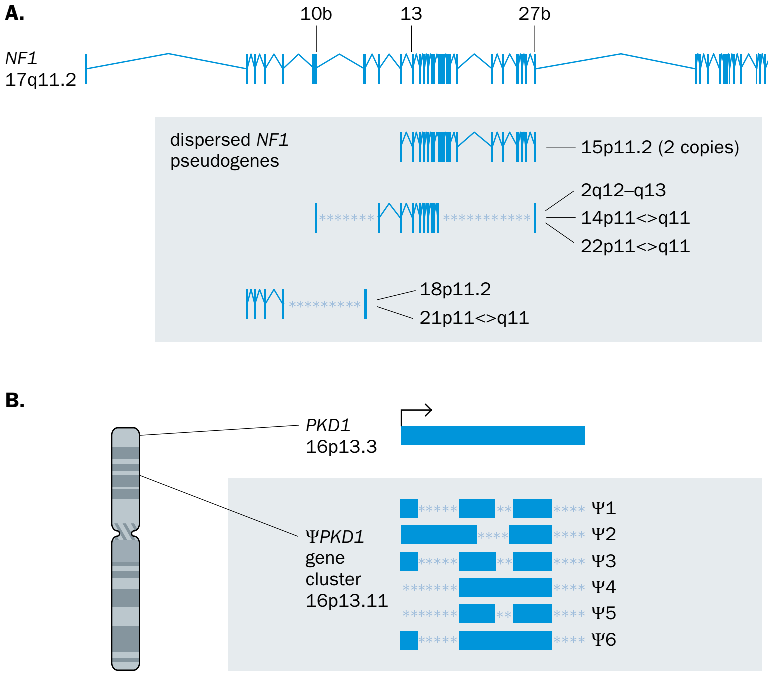


Figure 9.10 Dispersal of nonprocessed *NF1* and *PKD1* pseudogenes as a result of pericentromeric or subtelomeric instability. (**A**) The 58-exon *NF1* neurofibromatosis type I gene is located close to the centromere of human chromosome 17—exons and introns are shown as, respectively, thin vertical boxes and chevrons (^). Highly homologous defective copies of the *NF1* gene are found at nine or more other genome locations, mostly in pericentromeric regions. The pseudogene copies have segments of the full-length gene, with both exons and introns. Seven examples are shown here, such as two copies on 15p that have intact genomic sequences spanning exons 13 and 27b. Sometimes, rearrangements have caused deletion of exons and introns (shown by asterisks). (**B**) The 48-exon *PKD1* polycystic kidney disease gene is located close to the telomere of 16p. As a result of segmental duplication events during primate evolution (see **Figure 9.7**), large components of this gene have been duplicated and six *PKD1* pseudogenes (Ψ) are located at 16p13.11, with large blocks of sequence (shown as boxes) copied from the *PKD1* gene (asterisks represent the absence of counterparts to *PKD1* sequences).

human genome may have around 400 rDNA repeats (of which only 20–50% may be transcriptionally active in most cells), but individual nucleolar organizing regions can have from as low as 1-3 copies up to 140 rDNA repeats (spanning 5.6 Mb). Because the sequence arrays are in close proximity, there are frequent sequence exchanges between arrays on different acrocentric chromosomes as well as between arrays on homologous chromosomes or sister chromatids. The sequence exchanges (unequal crossover or unequal sister chromatid exchange) cause the individual sequences to be almost identical. Tandemly duplicated 5S rRNA genes are clustered at 1p42 and can also exchange sequences by unequal crossover, but in this family there are also many dispersed pseudogenes.

As well as the mtDNA sequences specifying mitochondrial tRNAs, there are close to 600 members of the human tRNA gene family dispersed over several chromosomes (including multiple pseudogenes), but there are some large clusters, notably one at 6p22.1.

Gene families making short guide RNAs that work in RNA maturation or regulation

In addition to genes that make tRNAs and the short ribosomal RNAs, some other moderately large gene families make short noncoding RNAs that serve to modify RNAs during RNA maturation or to regulate gene expression at the RNA level. In each case, the RNA works as a **guide RNA** that binds to RNA target sequences (rather than ones in the genome). Like other guide RNAs, they have binding sites for executive proteins that they transport to their target sequences to carry out an action, in this case a desired RNA modification. They include the RNAs listed below (with the number of human genes specifying them shown in brackets–gene details can be found by querying individual gene families at https://www.genenames.org/cgi-bin/genefamilies/).

- Small nuclear RNAs (snRNAs) are needed for RNA splicing (37 genes).
- Small nucleolar RNAs chemically modify rRNA at specific nucleotide sites (537 genes). There are two subfamilies: C/D box snoRNAs guide site-specific 2′-O-ribose methylations, with 105–107 varieties of this methylation in rRNA (355 genes; the SNORD family); H/ACA snoRNAs guide site-specific pseudouridylations, where uridine is isomerized to give pseudouridine at 95 different positions in the pre-rRNA (182 genes; the SNORA family).
- Small Cajal body RNAs (scaRNAs) chemically modify snRNAs (27 genes).
- MicroRNAs (miRNAs) regulate the expression of certain target genes by binding to the mRNAs they produce (1776 genes).

The genes specifying these RNAs are dispersed across the genome, but are occasionally grouped in small clusters, notably in the case of snoRNA genes. Note that most snoRNA genes are located within the introns of larger genes transcribed by RNA polymerase II; in these cases, snoRNAs are produced by processing of the intronic RNA, and so regulation of snoRNA synthesis is coupled to that of the host gene (see **Figure 9.5B** and **Figure 10.34** for examples). The recent burgeoning interest in RNA has led to establishment of a wide range of RNA databases (see **Table 9.11** for examples).

TABLE 9.11 EXAMPLES OF RNA DATABASES

Database	Description	Website address
RNAcentral	A portal that provides access to a wide range of electronic resources on RNA	http://rnacentral.org/
Rfam	A database of RNA families	http://rfam.xfam.org/
NONCODE	Integrated database of all noncoding RNAs except rRNA and tRNA	http://www.noncode.org
DASHR	Database of small human noncoding RNAs. In addition to sequences it includes annotation of precursor and mature small noncoding RNAs, and expression and RNA processing information across multiple normal human tissues and cell types	http://lisanwanglab.org/DASHR/smdb.php
miRBase	The microRNA database. Can be searched by species. The address in the column on the right is for human sequences	http://www.mirbase.org/cgi-bin/mirna_summary.pl?org=hsa
snoRNABase	A database of human small nucleolar RNAs	https://www-snorna.biotoul.fr/
GtRNAdb	A database of predicted tRNA genes. Can be searched by species. The address in the column on the right is for human sequences	http://gtrnadb.ucsc.edu/genomes/eukaryota/Hsapi19/
lncRNAdb	A reference database for lncRNAs known to be functional	http://www.lncrnadb.org/

RNA pseudogenes

Some RNA gene families have large numbers of pseudogenes—see **Table 9.12** for some examples. Whereas all protein-coding genes are transcribed by RNA polymerase II using an upstream promoter, some genes specifying some types of small noncoding RNAs are transcribed by RNA polymerase III, in which case transcription often occurs from an *internal promoter*. That is, the promoter is located within the transcribed DNA. (That is possible because the job of the promoter is to be a landing ground for a protein complex that binds to the promoter and then binds other proteins whose job is to position the RNA polymerase at the required position upstream. By the time the RNA polymerase is correctly positioned upstream, the protein complex that was bound to the promoter has been removed, leaving the way clear for transcription.)

TABLE 9.12 EXAMPLES OF RNA GENE FAMILIES HAVING LARGE NUMBERS OF PSEUDOGENES

RNA family	Number of human genes	Number of related pseudogenes
U6 snRNA	49	~800
Y RNA	4	~1000
[Alu repeats]	1 (*RN7SL*)	~1.5 million

Alu repeats may not be an obvious RNA family, but they originated in primates by retrotransposition of transcripts of the *RN7SL* gene that makes 7SL RNA (a component of the signal recognition particle that controls protein export from the cell).

Genes that have an internal promoter—including tRNAs, some snRNA genes, and some other types of RNA gene—are prone to making large numbers of copies by retrotransposition. That happens because the transcripts carry a copy of the promoter sequence and when they are converted into cDNA and then integrate into the genome elsewhere, the retrocopies can be transcribed from their promoter to make more transcripts, and so on. As detailed in Section 9.3, the Alu repeat, the most common repeat in the human genome, and the mouse B1 repeat originated, independently, by retrotransposition of 7SL RNA transcripts prior to the primate–rodent split about 90 million years ago. Other highly repetitive human and mouse DNA sequences arose by retrotransposition of tRNAs.

9.3 HETEROCHROMATIN DNA AND TRANSPOSON REPEATS

As listed below, two main DNA sequence classes are present in very high copy numbers in the human genome:

- Tandemly repeated short DNA sequences located at centromeres and other regions of constitutive heterochromatin (which remains condensed throughout the cell cycle);
- Interspersed transposon repeats that are distributed across the nuclear genome and vary in size from hundreds of base pairs to several kilobases in length.

Constitutive heterochromatin is very largely defined by long arrays of tandem DNA repeats

Constitutive heterochromatin is the highly condensed chromatin that is usually located at the periphery of the nucleus, attached to the nuclear membrane (the euchromatic DNA tends to be concentrated in the center of the nucleus where it can be actively transcribed). The underlying DNA accounts for around 200 Mb (~6.5%) of the human genome (see **Table 9.4**) and encompasses megabase-sized regions at the centromeres, multiple kilobases of DNA at the telomeres of all chromosomes, and large additional regions on some chromosomes, including: the majority of the Y chromosome; most of the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22); and very substantial regions of pericentromeric heterochromatin, notably on chromosomes 1, 9, 16, and 19 (see the chromosome banding image on the inside back cover).

The DNA of the telomeres is very highly conserved in sequence: in vertebrates (and in many other eukaryotes) it consists of long arrays of tandem repeats of a hexanucleotide sequence TTAGGG (or CCCTAA on the opposing strand); according to the length of the arrays, the telomere repeats belong to the minisatellite class of noncoding tandem repeats (see **Table 9.13**). Each array extends over several kilobases but the length varies with age, being reduced each time the DNA of a cell replicates in preparation for cell division (because of the *chromosome end-replication problem* illustrated in **Figure 2.24**). The telomere hexanucleotides are not represented in the GRCh38.p12 reference sequence); instead, telomere DNA is simply acknowledged by an arbitrary length of 10 kb, with the nucleotides simply represented by the letter N. (In human chromosome 1, for example, the location of telomere sequences in the reference sequence is marked by the following coordinates: short arm—chr1:1–10,000; and long arm—chr1: 248,946,423–248,956,422.) Immediately proximal to the telomere repeats are simple-sequence sub-telomeric DNA repeats (illustrated in **Figure 2.23**).

The DNA underlying constitutive heterochromatin at the centromeres and other regions mostly consists of very long arrays of high-copy-number tandemly repeated DNA sequences, known as **satellite DNA**. Large tracts of heterochromatin are typically composed of a mosaic of different satellite DNA sequences that are occasionally interrupted by transposon repeats. There are different satellite DNA organizations, and the repeated unit may be a very simple sequence (less than 10 nucleotides long) or a moderately complex sequence extending to over 100 nucleotides long; see **Table 9.13** for a classification of human satellite DNA and other high-copy-number tandem repeats.

TABLE 9.13 MAJOR CLASSES OF HIGH-COPY-NUMBER TANDEMLY REPEATED HUMAN DNA

Class[a] and subclasses	Size or sequence of repeat unit	Major chromosomal location(s)
SATELLITE DNA[b] (arrays often >100 kb)		
α (alphoid DNA)	171 bp	Centromeric heterochromatin of all chromosomes
β (*Sau*3A family)	68 bp	Notably the centromeric heterochromatin of 1, 9, Y, and the five acrocentric chromosomes[c]
Satellite 1	25–48 bp (AT-rich)	Centromeric heterochromatin of most chromosomes and other heterochromatic regions
Satellite 2	Diverged forms of ATTCC/GGAAT	Most, possibly all, chromosomes

(*Continued*)

TABLE 9.13 (*Continued*) MAJOR CLASSES OF HIGH-COPY-NUMBER TANDEMLY REPEATED HUMAN DNA

Class[a] and subclasses	Size or sequence of repeat unit	Major chromosomal location(s)
Satellite 3	ATTCC/GGAAT	Short arms of the five acrocentric chromosomes[c] and heterochromatin on 1q, 9q, and Yq12
DYZ19	125 bp	Yq11; comprising around 400 kb
DYZ2	AT-rich	Yq12; higher periodicity of ~2470 bp
MINISATELLITE DNA (arrays 0.1–20 kb)		
Telomeric minisatellite	TTAGGG	All telomeres
Hypervariable minisatellites	9–64 bp	All chromosomes, associated with euchromatin, notably in sub-telomeric regions
MICROSATELLITE DNA (arrays <100 bp)	Often 1–4 bp	Widely dispersed throughout all chromosomes

[a] The distinction between satellite, minisatellite, and microsatellite is made on the basis of the total *array length*, not the size of the repeat unit. [b] Satellite DNA arrays that consist of simple repeat units often have base compositions that are radically different from the average 41% GC (and so they could be isolated by buoyant density gradient centrifugation, which separates them from the bulk of the DNA, causing them to appear as *satellite bands*—hence the name). [c] The five human chromosomes that are acrocentric (with the centromere very close to one end) are chromosomes 13, 14, 15, 21, and 22.

Various satellite DNA families are associated with human centromeres (**Figure 9.11**), but only the α-satellite is known to be present at all human centromeres, and its repeat units often contain a binding site for a specific centromere protein, CENPB. Cloned α-satellite arrays have been shown to seed *de novo* centromeres in human cells, indicating that the α-satellite must have an important role in centromere function.

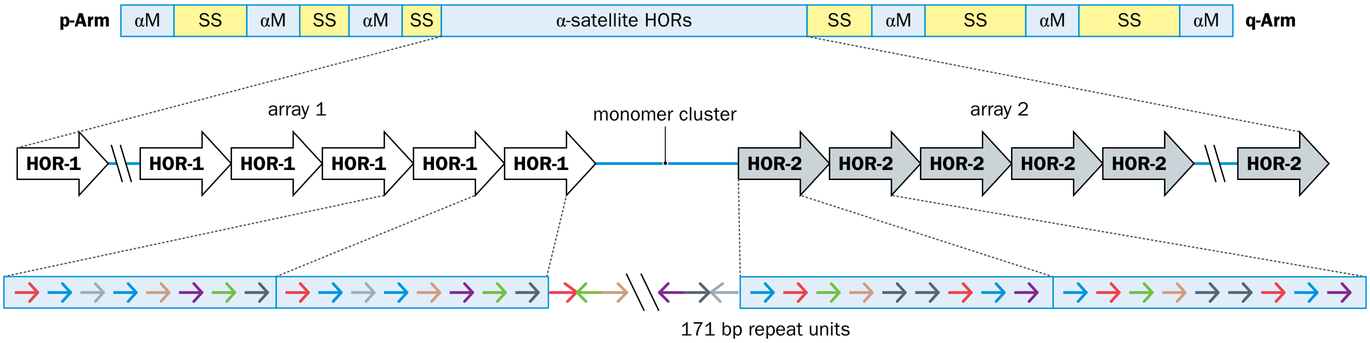


Figure 9.11 Human centromere and α-satellite DNA organization. α-satellite DNA (alphoid DNA) is a prominent component of the centromere of all human chromosomes. It is composed of repeats of a 171 bp sequence but significant sequence differences are found between different 171 bp repeats (represented here by different colored arrows at bottom; the monomers can vary at up to 30–40% of nucleotide positions). Higher-order repeats (HOR) mark regions where a series of similarly orientated monomer repeats has been tandemly repeated to form long multimer arrays with very high levels of sequence identity (97–100%) between the multiple HORs in an array. In this hypothetical example, we imagine two arrays composed respectively of HOR-1 repeats (each consisting of a sequence of eight monomers) and HOR-2 repeats (with a nine-monomer sequence). Clusters of simple monomers that can be in different orientations are typically found in the interval between the multimer arrays. Outside the α-satellite HORs, the centromere sequences often include α-satellite monomer clusters (αM) and simple sequence (SS) satellite DNAs. For a specific example—the centromere of human chromosome 10—see Figure 4 in the paper of Aldrup-Macdonald & Sullivan (2014) (PMID 24683489) listed under Further Reading.

Transcription of heterochromatin DNA

We typically think of heterochromatin DNA sequences as being transcriptionally inactive, but RNA transcripts can be produced from the tandemly repeated DNA sequences underlying constitutive heterochromatin. Telomeric repeat-containing RNA (TERRA) transcripts of variable length (100 bp–9 kb) are notably produced in the G_1 phase of the cell cycle and contain both subtelomeric sequences and C-rich telomere hexanucleotide repeats. TERRA transcripts have multiple roles, including regulation of telomere length and telomere capping and replication. Satellite DNA sequences in pericentromeric and centromeric regions can also be transcribed. The output varies according to developmental stage and cell type, but is amplified in response to cellular stresses, such as heat shock, exposure to hazardous chemicals and heavy metals, and so on.

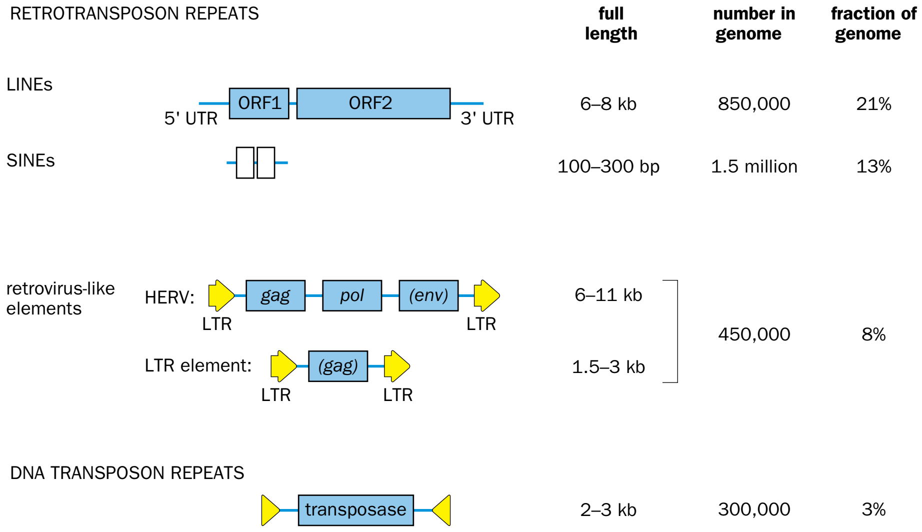
Transposon-derived repeats make up the majority of the human genome and arose very largely through retrotransposition

The majority of the human genome is made up of interspersed repetitive noncoding DNA sequences derived from **transposons** (also called *transposable elements*), mobile DNA sequences that can migrate to different regions of the genome. About 45% of the genome can readily be seen to be made up of transposon repeats, but much of the remaining "unique" DNA is likely to have been derived from ancient transposon copies that have diverged extensively over long evolutionary timescales. (The most sensitive computer programs indicate that at least two-thirds of the human genome arose in this way.)

In humans, and other mammals, only a tiny minority of transposon repeats are actively transposing. Both the frequency of transposition and the percentage of full-length repeats in a repeat family depend on the family's evolutionary age: recently evolved repeats have a comparatively high percentage of full-length repeats and of actively transposing members, and conversely, more ancient transposon repeats are frequently truncated copies or have inactivating mutations. Transposons that can transpose independently are described as *autonomous*. Others are *nonautonomous*: they can transpose only with the help of an autonomous transposon.

DNA transposons versus retrotransposons

A small minority of human transposon repeats originated from the DNA transposon class. Transposons of this type have terminal inverted repeats and migrate directly without any copying of the sequence using a "cut-and-paste" mechanism (they make a transposase to excise the DNA sequence, which then re-inserts elsewhere in the genome). There are two major human DNA transposon superfamilies (**Figure 9.12**) plus a variety of less frequent families. Although these repeats were actively transposing in the past, there is much less evidence of recent transposition; they are often described, therefore, as transposon fossils.



RETROTRANSPOSON REPEATS	full length	number in genome	fraction of genome
LINEs	6–8 kb	850,000	21%
SINEs	100–300 bp	1.5 million	13%
retrovirus-like elements HERV:	6–11 kb	450,000	8%
LTR element:	1.5–3 kb		
DNA TRANSPOSON REPEATS	2–3 kb	300,000	3%

Figure 9.12 Human transposon repeat classes. Note that full-length repeats are rare; most are truncated copies. Some full-length LINEs (of the LINE-1 subfamily) are able to transpose independently because they can make a functional reverse transcriptase. SINEs are nonautonomous: they need a reverse transcriptase to be supplied. LTR transposons resemble retroviruses and have long terminal repeats (LTR) characteristic of retroviruses. They include endogenous retroviruses (with *gag*, *pol,* and *env* genes), plus truncated LTR elements that have lost key retroviral sequences. DNA transposons use a cut-and-paste transposition, but human DNA transposon repeats seem to be unable to transpose, being truncated or having a mutated transposase gene. Various human DNA transposon superfamilies exist, of which the most numerous are the hAT and the Tc1/*mariner* superfamilies (see PMID 17339369). LINE, long interspersed nuclear element; SINE, short interspersed nuclear element; HERV, human endogenous retrovirus. (Adapted from The International Human Genome Sequencing Consortium [2001] *Nature* **409**:860–921; PMID 11237011. With permission from Springer Nature. Copyright © 2001.)

The great majority of human transposon repeats belong to the **retrotransposon** class (also called retroposons). They can transpose using a reverse transcriptase to convert an RNA transcript into a cDNA copy that then integrates into the genomic DNA at a different location (a "copy-and-paste" mechanism). There are three major types of mammalian retrotransposon repeat, as listed below and illustrated in **Figure 9.12**.

- LINEs (long interspersed nuclear elements; over 6 kb when full length) have a comparatively long evolutionary history: equivalent sequences are present in other mammals, such as mice. Human LINEs, consisting of three distantly related families (LINE-1, LINE-2, and LINE-3), are located primarily in euchromatic regions, preferentially in the dark AT-rich G-bands of metaphase chromosomes. The LINE-1 (or L1) family is the predominant LINE family, accounting for 17% of the genome, and is detailed below. It continues to have actively transposing members that are the only autonomous transposon repeats in the human genome.

- SINEs (short interspersed nuclear elements; full-length members are less than 400 nucleotides long). SINEs cannot transpose independently. However, SINEs and LINEs share sequences at their 3′ end, and SINEs have been shown to be mobilized by neighboring LINE repeats. By parasitizing on the LINE transposition machinery, SINEs can attain high copy numbers. The primate-specific Alu repeat is the most abundant SINE in the human genome, and is detailed below; the next most common human SINE family are mammalian-wide interspersed repeats, known as MIR elements.
- LTR transposons (repeats that resemble retroviruses, minimally having the long terminal repeats [LTR] characteristic of retroviruses). There are two subclasses: human endogenous retroviruses (HERVs) and LTR elements. HERVs have the *gag*, *pol,* and *env* genes of retroviruses, and arose when infectious retroviruses repeatedly entered the germ line of hosts over many tens of millions of years. LTR elements are effectively truncated HERVs: they retain LTR sequences but have lost key retroviral sequences.

In addition to the major classes above, some repeats are composites of different classes, notably the SVA family, as described below.

The LINE-1 (L1) family

Full-length functional LINE-1 elements make two proteins: an RNA-binding protein, p40, and a protein with both endonuclease and reverse transcriptase activities (**Figure 9.13**). Full-length copies bring with them their own promoter (located in the 5′ untranslated region) that can be used after integration in a permissive region of the genome. After translation, the LINE-1 RNA assembles with its own encoded proteins and moves to the nucleus.

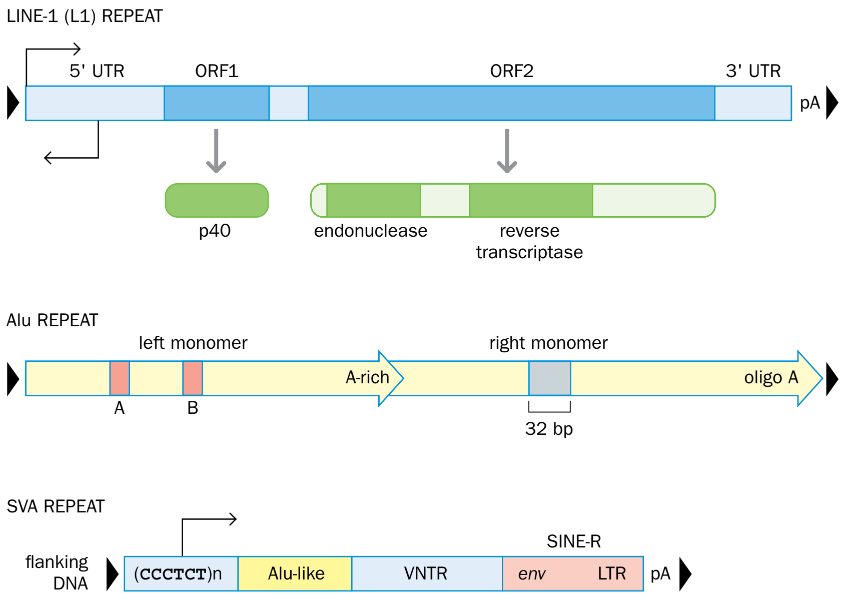


Figure 9.13 Structure of human LINE-1, Alu, and SVA repeats. The LINE-1 p40 protein is an RNA-binding protein with a nucleic acid chaperone activity. Converging arrows mark potential transcription from a bidirectional internal protein within the 5′ UTR (untranslated region) of LINE-1 elements. At the other end is an A_n/T_n sequence, often described as the 3′ poly(A) tail (pA). The LINE-1 endonuclease cuts one strand of a DNA duplex, preferably within the sequence TTTT↓A, and the reverse transcriptase uses the released 3′-OH end to prime cDNA synthesis. New insertion sites are flanked by a small target-site duplication (flanking black arrowheads). Alu repeats often consist of two monomer repeats that have similar sequences terminating in an A-rich or A_n/T_n sequence (oligo A) but differ in size because of the insertion of a 32 bp element within the larger repeat. The smaller repeat has internal components of an RNA polymerase III promoter (boxes A and B). Nonautonomous SVA repeats are usually more than 2 kb in length and have both an Alu-like sequence and a 3′ HERV fragment (called SINE-R), separated by a VNTR sequence. They may often be transcribed from promoters in flanking DNA sequence. VNTR, variable number of tandem repeats.

To integrate into genomic DNA, the LINE-1 endonuclease cuts a DNA duplex on one strand, leaving a free 3′ OH group that serves as a primer for reverse transcription from the 3′ end of the LINE RNA. The endonuclease's preferred cleavage site is TTTT↓A; hence the preference for integrating into AT-rich regions. During integration, the reverse transcription often fails to proceed to the 5′ end, resulting in truncated, nonfunctional insertions. Accordingly, only about 1 in 100 copies are full length, and most LINE-derived repeats are short (the average size for all LINE-1 copies is 900 bp).

The LINE-1 machinery is responsible for reverse transcription of all retroelements in the genome, including nonautonomous SINEs and SVA repeats, and also copies of mRNA transcripts that integrate in the genome to create processed pseudogenes and retrogenes, as illustrated in **Box 9.2**. Of the 6000 or so full-length LINE-1 sequences, about 60–100 are still capable of transposing, and they occasionally cause disease as a result of aberrant gene expression after insertion.

Alu repeats

The human Alu repeat is the most abundant sequence in the human genome. The full-length repeat is about 280 bp long and consists of two tandem repeats, each about 120 bp in length followed by an A-rich or A_n/T_n sequence (see **Figure 9.13**). Monomers, containing only one of the two tandem repeats, and various truncated versions of dimers and monomers are also common, giving a genome-wide average of 230 bp. Alu repeats are primate specific, but subfamilies of different evolutionary ages can be identified, of which the Y and S subfamilies contain the most mobile Alu sequences.

Alu repeats have a relatively high GC content and, although dispersed mainly throughout the euchromatic regions of the genome, are preferentially located in the GC-rich and gene-rich R chromosome bands, in striking contrast to the preferential location of LINEs in AT-rich DNA. When located within genes they are, like LINE-1 elements, almost always confined to introns and the untranslated regions.

Like other mammalian SINEs, Alu repeats originated from cDNA copies of small RNAs transcribed by RNA polymerase III that re-integrated into germ-line DNA. Genes transcribed by RNA polymerase III often have internal promoters, and so the cDNA copies of transcripts carry with them their own promoter sequences that can activate transcription of the cDNA copy in a permissive chromatin location. Both the Alu repeat and, independently, the mouse B1 repeat originated from cDNA copies of 7SL RNA, the short RNA that is a component of the signal recognition particle, using a retrotransposition mechanism. Some other SINEs, such as MIR elements and the mouse B2 repeat, are known to be retrotransposed copies of tRNA sequences.

SVA repeats

SVA repeats are composite repeats, having both an Alu-like sequence and a truncated HERV component (the 3′ end of the *env* gene plus an LTR) that was originally named SINE-R. These two elements are separated by a VNTR region with a variable number of tandem copies of a 35–50 bp sequence (the name SVA is short for SINE-R–VNTR–Alu). The Alu sequence is preceded by tandem CCCTCT repeats, and the HERV component is followed by an $(A/T)_n$ tail, called a poly(A) sequence (see **Figure 9.13**). This recently evolved, hominid-specific family contains mostly full-length sequences (~2 kb in length), and although there are only 2700 human SVA repeats, they are the third most actively transposing human transposon repeat sequence (after Alu repeats and LINE-1 repeats).

The double-edged nature of transposons: both friends and foes

As detailed in Chapter 13, transposons have been crucially important in genome evolution, and copies of transposons have been valuable sources of novel functional sequences, including not just new regulatory sequences, but also new exons, and, very occasionally, even new genes. Transposon repeats may also assist gene and exon duplication (which are also important in genome evolution) by stabilizing local mispairing of chromatids (see **Figure 9.6**), and they can alter expression of host genes in different ways, such as by offering new regulatory sequences or new splice sites. Retrotransposons also actively transpose during neurogenesis, creating an additional level of genetic diversity that may be valuable in promoting neuron diversity.

While transposons offer many advantages, active transposons pose a threat to the genome. Effectively, they are mutagens and potentially harmful, and excessive mobilization of transposons can result in chaos. To contain this threat, the first line of defense is epigenetic regulation: the transcription of active retrotransposons is down-regulated by setting epigenetic marks to alter the chromatin state (typical epigenetic marks are DNA methylation and histone modifications). However, epigenetic marks across the genome are erased in early development (in preparation for re-setting of epigenetic marks, including imprinting marks); that allows an opportunity for transient transcription of the retrotransposons, leading to their rapid proliferation.

Because of the importance of the germ line, and because the chromatin state of germ-line cells may be more inherently permissive to retrotransposon activity than that of somatic cells, a variety of different genome defense strategies have evolved to limit the potentially dangerous spread of retrotransposons in the germ line. Many of these mechanisms are often used also to combat virus infections (viruses originated from ancient transposons). Interested readers wishing to explore this topic further might like to consult the reviews by Molaro & Malik (2016) (PMID 26821364) and Goodier (2016) (PMID 27525044) listed under Further Reading. We briefly describe below one important method of containing the threat posed by retrotransposons, which is based on RNA silencing.

Controlling retrotransposons using piRNA-mediated RNA silencing

Both piRNAs and, to a lesser extent, endogenous short interfering RNAs (siRNAs) are important in transposon control in the germ line using RNA silencing. We outlined siRNA-based RNA silencing in **Box 8.2**, and piRNA-based silencing is similar in some respects, but different in others. piRNAs, resembling a type of repeat-associated siRNA, are typically 24–31 nucleotides long, slightly longer than siRNAs, and they have a distinct preference for a U at the +1 position. The name is a contraction of PiWi protein-interacting RNAs (they were first discovered in *Drosophila* where they bind to the PiWi protein and two related proteins; the equivalent mammalian proteins are the cytoplasmic proteins MILI and MIWI and a nuclear protein MIWI2).

Whereas both siRNAs and miRNAs are produced by cleavage of double-stranded RNA precursors into functional small RNAs, piRNAs are produced from single-stranded RNAs, and predominantly in germ cells. The majority of piRNAs originate from long (50–100 kb) RNA polymerase II transcripts produced from numerous regions of the genome with a high density of truncated transposon repeats. The capped and polyadenylated transcripts are exported to the cytoplasm where they are each subject to processing and ultimately the production of thousands of piRNAs from transcripts.

Like siRNAs, piRNAs act as guide RNAs, recognizing and binding to complementary DNA and RNA sequences, and like siRNAs, their job is to bind executive silencing proteins and deliver them to the target sequences. For example, in the embryonic mouse male germ line, piRNAs originating from transposons can bind the MIWI2 protein and transport it to the sites of complementary sequences within transposons across the genome; the deposited MIWI2 proteins recruit transcriptional silencing complexes to the target sequences, often repressing their transcription by CpG methylation. Similarly, the cytoplasmic MILI protein can be bound by piRNAs that then bind to transposon transcripts with the complementary sequence, whereupon the slicer endonuclease cleaves the transcript, leading to its degradation, a form of post-transcriptional silencing.

9.4 A START ON WORKING OUT HOW OUR GENOME FUNCTIONS

In retrospect, getting the genome sequence was the easy part. In the "post-genome era," now comes the hard part: working out what the sequence means, and how our genome functions. The recent availability of the human genome sequence plus the genome sequences of other organisms has promoted different types of genomic strategies (**Figure 9.14**), and in this section we start with **functional genomics**. An important development here has been the attempt to define all the functional DNA elements in the human genome, beginning with the ENCODE (Encyclopedia of DNA Elements) Project that commenced in 2003 at the conclusion of the Human Genome Project.

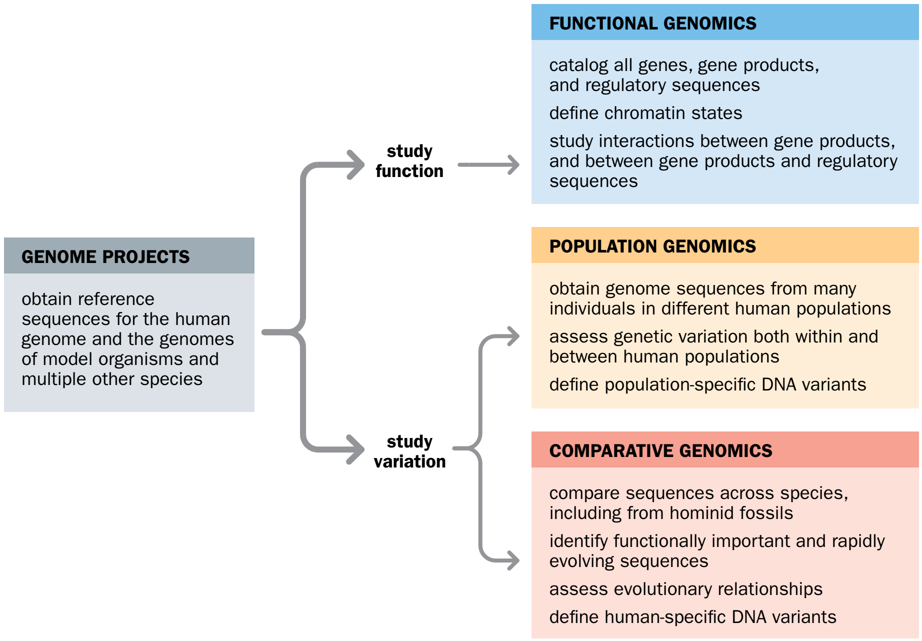

FUNCTIONAL GENOMICS
catalog all genes, gene products, and regulatory sequences
define chromatin states
study interactions between gene products, and between gene products and regulatory sequences
study function
GENOME PROJECTS
obtain reference sequences for the human genome and the genomes of model organisms and multiple other species
POPULATION GENOMICS
obtain genome sequences from many individuals in different human populations
assess genetic variation both within and between human populations
define population-specific DNA variants
study variation
COMPARATIVE GENOMICS
compare sequences across species, including from hominid fossils
identify functionally important and rapidly evolving sequences
assess evolutionary relationships
define human-specific DNA variants

Figure 9.14 Three directions for genomics projects in the "post-genome era". Genome projects have produced reference sequences for the human genome and the genomes of many vertebrate and invertebrate animals. Functional genomics projects are attempting to define all functional DNA elements in genomes and how they work; projects investigating human and mouse genome function are described in Section 9.4. Population genomics projects to provide a comprehensive understanding of human genetic variation are described in Section 11.3. Comparative genomics projects are described in Section 13.1.

In addition to defining genes and regulatory sequences across the genome, systematic genome-wide efforts have been made to track human gene products at both the RNA and protein levels. The transcriptome represents the combined output of transcription, RNA processing, and RNA turnover; similarly, the proteome is the combined output of translation, post-translational processing (including protein cleavage and protein modification), and protein turnover. Because processes such as transcription, RNA processing, protein synthesis, and post-translational processing are all regulated, the transcriptome and proteome differ significantly between different cell types, and even between cells of the same type in response to different environmental conditions.

Efforts in research laboratories across the world also seek to understand gene function. That can mean manipulating genes in cultured cells, or producing genetically modified animals with desired mutations at pre-determined positions, allowing insights into how human genes work. In addition, international collaborative projects are carrying out parallel gene manipulations on a genome-wide scale.

The effort needed to understand how our genome works should not be underestimated. The genome is a comparatively stable entity, and because it is essentially the same in all nucleated cells, we typically refer to *the* human genome as if it were a constant (there is, of course, genetic variation). However, transcriptomes and proteomes vary enormously depending on cell type. That begs the question of how many different human cell types should be explored, and under what conditions, in order that we can get a handle on how the genome functions. The task of defining interactions between the huge number of functional DNA elements and the vast numbers of transcripts and proteins (which vary from one cell type to another) is an immense one.

The ENCODE Project: the first systematic attempt to catalog functional DNA elements in the human genome

The international ENCODE Project (https://www.encodeproject.org/) was established to define and catalog the functional DNA elements in our genome. Operationally, a functional DNA element was defined as a discrete genome segment that makes a defined product (protein or noncoding RNA) or displays a reproducible biochemical signature (such as a protein-binding capacity or a specific chromatin structure). Begun in September 2003 as a pilot project, the ENCODE Project provided a full report in 2012 in a series of papers in *Nature* and *Genome Research*.

The ENCODE Project involved global (genome-wide) assays of functional sequences as interpreted by analyzing multiple characteristics that affect gene expression and are dependent on cell type, including: transcript profiles; histone modification states; transcription factor-binding states; chromatin conformation states; and DNA methylation states (**Figure 9.15**)—we describe the different types of assay below. The assays were carried out on various cell types from a dedicated resource of over 100 different human cell types, including immortal stem cells, immortalized cell lines, and primary (nontransformed) cells.

For many assays, only a small subset of the cell types was used, the different cell types being classified into different tiers according to priority. Tier 1 (highest priority) included the H1 human embryonic stem cell line and two well-studied immortalized cell lines: GM12878, a B-lymphoblastoid cell line, and K562, an erythroleukemic cell line. Tier 2 cell types comprised the HepG2 hepatoblastoma cell line, HeLa cells, and primary (nontransformed) human umbilical vein endothelial cells. Tier 3 was represented by more than 100 cell types, including different cell lines and cultured cells, plus cells from some primary tissues.

The ENCODE Project comprised a variety of experimental assays plus annotation efforts. The headings below give a brief description of the types of approach used, and also the conclusions that were drawn.

Annotation

An important component of the ENCODE Project was devoted to *annotation*—descriptive commentary of the characteristics of genes and their products, and of other functional DNA sequences. It encompassed protein-coding genes, pseudogenes, and noncoding transcribed loci across the genome, and was also required in cataloging the products of transcription including alternative splice forms. Some automated annotation, such as by using the Ensembl gene build pipeline, was carried out, but much of the annotation was carried out manually by the HAVANA (human and vertebrate annotation) team at the Wellcome Trust Sanger Institute. GENCODE, the result of ENCODE gene annotation, provides a comprehensive catalog of transcripts and gene models (which continues to be regularly updated), and is available at http://www.gencodegenes.org/.

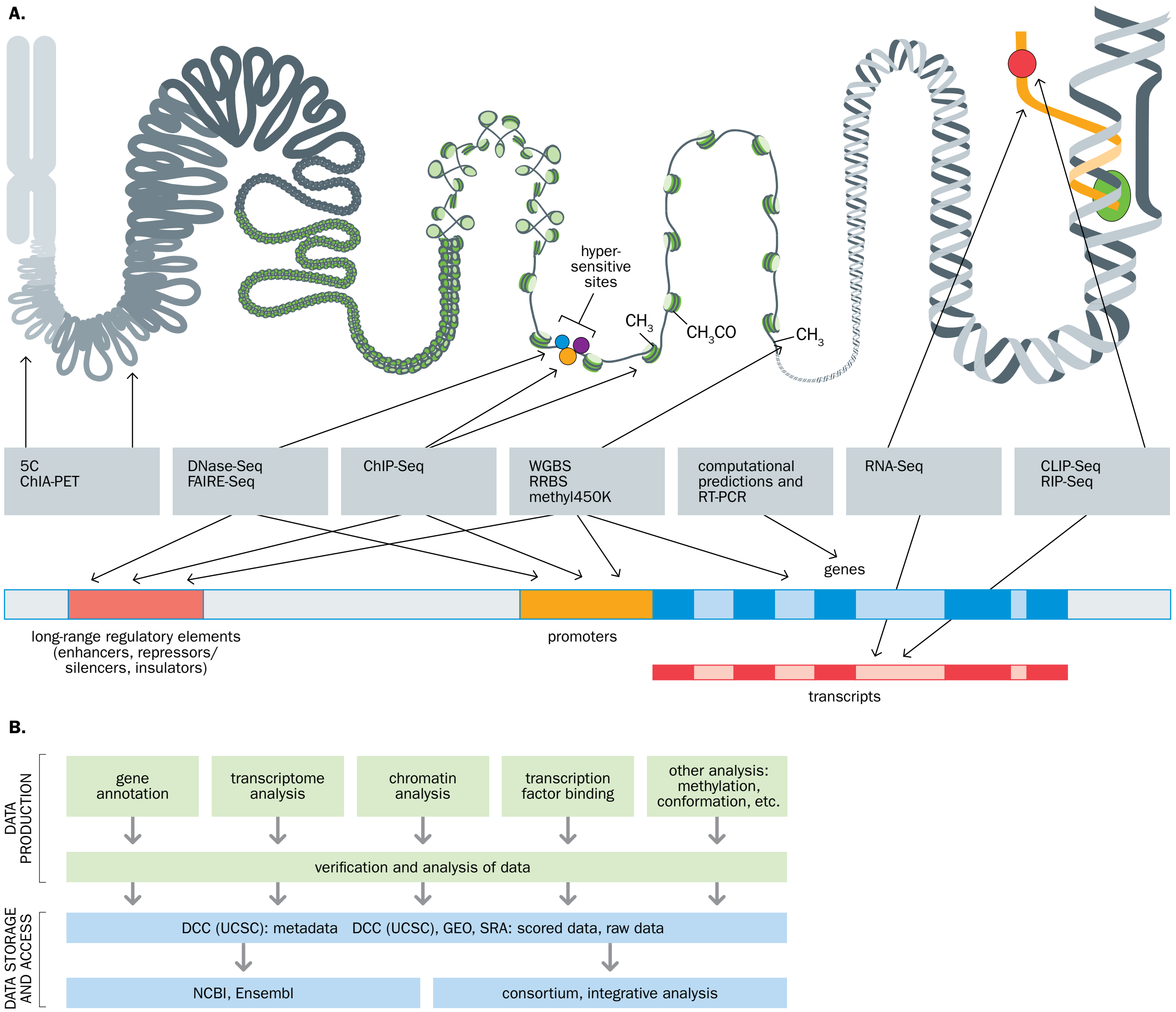


Figure 9.15 Principal methods and data flow in the ENCODE Project. (**A**) Major methods used to detect functional DNA elements (gray boxes), represented on an idealized model of mammalian chromatin and a mammalian gene. 5C, ChiA-PET and Hi-C are methods used to study chromatin conformation and are described in **Box 10.1**. The DNAse-Seq, FAIRE-Seq and ChIP-Seq methods seek to identify regions of chromatin where the DNA is readily accessible (and sensitive to DNAse I), and are described in Section 10.1. ChIP-Seq helps define binding sites for DNA-associated proteins (**Box 9.3**). Whole genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS) and methyl arrays are used to survey DNA methylation profiles across the genome. Methods focused on RNA include RNA-Seq, RNA immunoprecipitation (RIP; using a specific antibody to immunoprecipitate a protein that binds to RNA) and cross-linking immunoprecipitation (CLIP; similar to RIP but with an extra cross-linking step). (**B**) The overall data flow from the production groups after reproducibility assessment to the Data Coordinating Center (DCC) for public access and to other public databases. (Adapted from The ENCODE Project Consortium [2011] *PLoS Biol* **9**:e1001046; PMID 21526222; and associated website at https://www.encodeproject.org/.) GEO, Gene Expression Omnibus; SRA, Sequence Read Archive; UCSC, University of California at Santa Cruz.

Transcript analysis

Different methods were used to analyze RNA transcripts as described in Table 9.14. After surveying 15 different human cell lines, the ENCODE consortium concluded that about 75% of the human genome is transcribed in at least one of the cell types studied (but the percentage is significantly lower in any one cell type). RNA transcripts from both strands were found to be common (both within genes and in intergenic regions close to genes), and a protein-coding gene is now known to produce, on average, about six to seven different RNA transcripts, including some noncoding transcripts—see Figure 7.8 for a graphical display of transcripts from the *CFTR* (cystic fibrosis transmembrane regulator) gene. Some genes are known to produce functional noncoding transcripts as well as functional mRNAs, but in general the functional status of noncoding transcripts is underexplored.

(Note: high-resolution spatial expression of transcripts, which is most efficiently carried out in the miniaturized tissues of the embryo, was not carried out by the ENCODE Project. Various other projects have focused on large-scale *in situ* hybridization studies against primarily mouse embryonic tissue sections, and gene expression data are deposited in resources such as the e-Mouse Atlas at http://www.emouseatlas.org/ and the Allen brain atlas at http://mouse.brain-map.org/.)

TABLE 9.14 PRINCIPAL TRANSCRIPTION ANALYSES CARRIED OUT IN THE ENCODE PROJECT

Method	Description
RNA-seq	The principal method used to analyze transcriptomes. RNA transcripts are fragmented and converted to cDNA, and then sequencing adaptors are ligated to enable high-throughput DNA sequencing (see **Figure 7.12**)
CAGE	The Cap Analysis Gene Expression method was used to identify transcriptional start sites. It begins by isolating RNA, converting it with reverse transcriptase to give an RNA–cDNA hybrid, and then adding a biotin group to the methylated cap structure (found at the natural 5′ ends of RNA transcripts). Biotinylated RNA–cDNA hybrids are then captured using streptavidin-coated magnetic beads. RNase I is added to destroy any part of the RNA that is not bound to the cDNA, causing removal of the 5′ biotinylated cap from any RNAs bound to noncomplete cDNAs
RNA-PET	The RNA-Paired End Tag method allows capture of full-length RNA transcripts by selecting for RNAs with both a 5′ methyl cap and a poly(A) tail. Follow-up high-throughput sequencing allows short sequences (sequence tags) to be determined at the two ends of each of the transcripts
RT-PCR	Carried out as a check to validate alternative splice forms suggested from other transcript analyses

DNA methylation and chromatin chemical modification analysis

Both the DNA of our cells and the histone proteins of chromatin are subject to patterns of chemical modification that vary between different cell types; in each case, as detailed in Chapter 10, the chemical modification is a form of epigenetic control of gene expression. Different types of chemical modification occur in the side chains of N-terminal residues of histone proteins—including methylation of certain lysines (K) and arginines (R), and acetylation of certain lysines—but chemical modification of DNA is limited to methylation of certain cytosines.

The patterns of DNA methylation were tracked by comparing DNA samples treated with sodium bisulfite with untreated DNA samples (the bisulfite treatment converts unmethylated cytosines to uracil, but methylated cytosines are not affected—see **Box 10.3** for the general principle). In addition to expensive whole-genome bisulfite sequencing (WGBS), more efficient and high-throughput reduced representation bisulfite sequencing (RRBS) allowed analysis of genome-wide methylation profiles at single-nucleotide resolution in gene-rich sequences. (RRBS combines restriction enzymes and bisulfite sequencing to enrich for CpG-rich areas of the genome, and quantitative DNA methylation profiles were obtained for 1.2 million CpGs, on average, in each of 82 cell lines and tissue cells.) Methyl450K bead chips allowed assay of 450,000 methylation sites selected from genes, CpG islands, and enhancers across the genome. Two principal findings were: (1) 96% of CpGs exhibited differential methylation in at least one of the cell types assayed; and (2) the most variably methylated CpGs are found more often in gene bodies (comprising all exons and introns) and intergenic regions rather than in promoters and upstream regulatory regions.

To assay chemical modification of histone side chains, the **ChIP-Seq** method, combining chromatin immunoprecipitation with DNA sequencing, was used (see **Box 9.3** for the method). Various antibodies specific for the histone variants were used. Up to 11 types of histone modification were sampled in 46 cell types, and eight selected variants were assayed across all cell types in tier 1 and tier 2 categories. As expected, the global patterns of histone modification were found to be highly variable across different cell types.

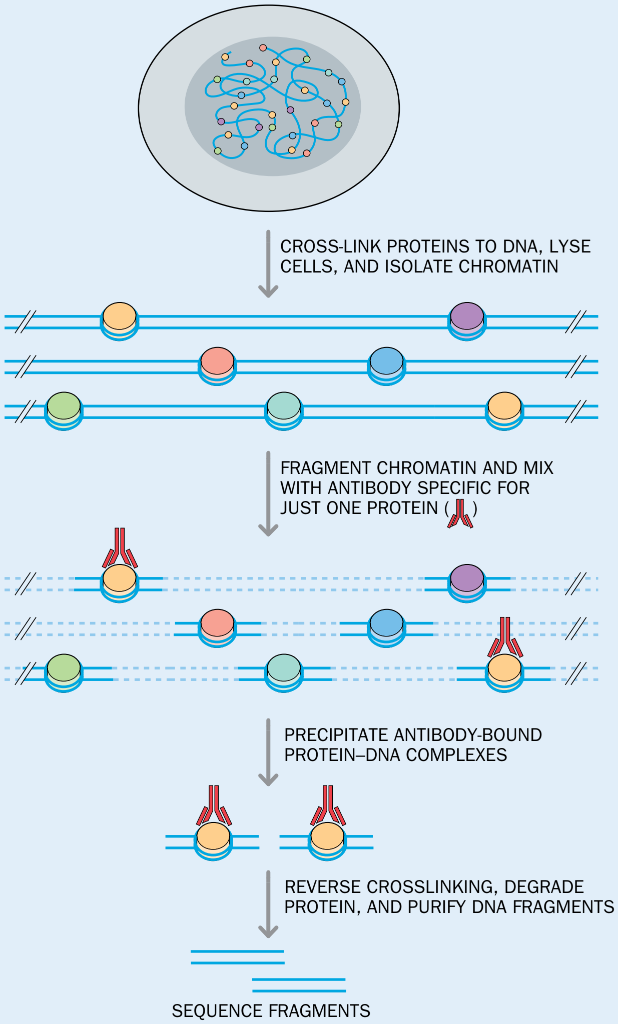
Transcription factor binding, regulatory sequences, and chromatin accessibility

ChIP-Seq was also used to map the 6–25-nucleotide-long genomic positions bound by many different transcription factors across the genomes of 72 different types of living cell. The data, together with that of binding sites for some other DNA-binding proteins

BOX 9.3 ChIP-Seq, A METHOD FOR DEFINING THE BINDING SITES OF DNA-ASSOCIATED PROTEINS

Certain DNA-binding proteins, such as transcription factors, bind transiently by noncovalent bonding to specific sequences along the DNA; at any moment in the intact cell, the DNA will be associated with a number of such proteins that have bound to their target sequences. Treatment of the cells with a cross-linking agent effectively “freezes” these transient DNA–protein interactions in time. As a result of cross-linking, the DNA-binding proteins that are attached to their target sequences in chromatin become covalently bonded and irreversibly fixed to the target DNA.

The cells are lysed, the chromatin is fragmented, and antibodies against a protein of interest can be used to immunoprecipitate the chromatin fragment that includes the bound protein (chromatin immunoprecipitation or ChIP). Incubation at 65° reverses the protein–DNA cross-linking, and proteinase treatment removes the protein, allowing the DNA fragments to be purified and sequenced (**Figure 1**).


CROSS-LINK PROTEINS TO DNA, LYSE CELLS, AND ISOLATE CHROMATIN
FRAGMENT CHROMATIN AND MIX WITH ANTIBODY SPECIFIC FOR JUST ONE PROTEIN ( )
PRECIPITATE ANTIBODY-BOUND PROTEIN–DNA COMPLEXES
REVERSE CROSSLINKING, DEGRADE PROTEIN, AND PURIFY DNA FRAGMENTS
SEQUENCE FRAGMENTS

Box 9.3 Figure 1 The ChIP-Seq method identifies DNA sequences that are bound by specific proteins. In the ENCODE Project, ChIP-Seq was used across the genomes of different cell types to identify the DNA sequences in nucleosomes that carry specific histone modifications (using antibodies that are specific for different modified histones), and to identify DNA-binding sites for transcription factors.

(including components of RNA polymerases), are integrated with the ENCODE data for histone modification and nucleosome occupancy profiles, and are available at the Factorbook website at http://www.factorbook.org/human/chipseq/tf/.

Regulatory DNA sequences, including the sequences recognized by transcription factors, must be in regions of chromatin that are accessible to protein factors. Accessible chromatin regions are more sensitive to cleavage by deoxyribonuclease I (DNase I) and the ENCODE Project also mapped nearly 3 million unique, non-overlapping DNase I hypersensitive sites by a technique known as DNase-Seq in 125 cell types. In addition, regulatory elements were isolated by the FAIRE-Seq method: FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates nucleosome-depleted genomic regions by exploiting the difference in cross-linking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low). The method involves cross-linking, phenol extraction, and then sequencing the DNA fragments in the aqueous phase.

Chromatin conformation

Cis-acting regulatory elements, such as enhancers, can control genes from long distances (tens to hundreds of kb) on the same DNA molecule through DNA looping: the enhancer and its bound regulatory proteins are brought into close proximity to the gene’s promoter and nearby regulatory elements. To map these long-range interactions, the 5C method, an enhanced version of chromosome conformation capture (3C), was used in the ENCODE Project. We detail these methods in Chapter 10 when we look at aspects of gene regulation.

Overall conclusions and the functional fraction of the genome

Among the big surprises of the ENCODE Project was the finding of *pervasive* transcription of the human genome. Not only is a large fraction of the genome transcribed (with about 75% of the sequence in the genome being transcribed in one or more of the sampled cell lines), both DNA strands are often transcribed, and the great majority of the transcripts are noncoding RNAs. The analyses also showed evidence for an initial set of nearly 400,000 regions with features resembling enhancers and over 70,000 promoter-like regions. A big surprise from various transcript analyses was that bidirectional transcription commonly proceeds from both promoters and enhancers.

A headline conclusion was that at least 80% of the human genome participates in at least one biochemical RNA and/or chromatin-associated event in at least one cell type. That was interpreted in commentary articles in *Nature* and *Science* to mean that most of our genome is functionally important after all, and that it was time to abandon the concept of "junk DNA." (Previous studies based largely on species comparisons had estimated that a small fraction, perhaps just 3–8%, of the euchromatic genome was functionally significant; much of the rest was sometimes labeled as junk DNA.) However, the idea that the great majority of the genome is functionally important, and that junk DNA is now an outmoded concept, has been strongly resisted by evolutionary geneticists. We revisit the arguments when we consider comparative genomics in Section 13.1.

Approaching the functions of human genes through genetic manipulation of cultured cells and extrapolation from model organisms

The function of a gene is actually the function of its product(s), and can be conveniently classified at three levels:

- *The biochemical level*. For example, a gene product may be described as a kinase or a calcium-binding protein. This reveals little about its wider role in the organism;
- *The cellular level*. This builds in information about intracellular localization and biological pathways. For example, it may be possible to establish that a protein is located in the nucleus and is required for DNA repair, even if its precise biochemical function is unknown;
- *The organismal level*. This may include information about where and when a gene is expressed in different tissues, and its role in the processes of development and/or physiology.

To obtain a complete picture of gene function, information is required at all three levels. Defined vocabularies have been developed to describe gene function across all genomes. The most commonly used information resources here are the Gene Ontology (GO) system (http://www.geneontology.org/) and the Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg/).

Useful information about a gene's function can be obtained by using specific labeled probes or antibodies to track expression at the RNA and/or protein levels in cells and tissues, and to identify interactions with other proteins and nucleic acid sequences. However, the most comprehensive insights into gene function come from different types of genetic screens. In principle, genetic screens can be classified into two major headings:

- *Forward genetic screen*—the starting point is an abnormal phenotype (often induced by exposure to a mutagen), and the task is to identify the locus of the genetic change that has produced the phenotype, and so connect genotype with phenotype;
- *Reverse genetic screen*—the starting point is a defined gene or genes, and the task is to induce genetic changes in a defined gene(s) to produce an abnormal phenotype that provides information on how the gene(s) normally function.

In the post-genome era, large gene collections are available, and because the functions of many of them are not known or poorly understood, reverse genetic screens are now commonly deployed. They can be carried out in cultured cells, as described below. Inevitably, assaying gene function in cultured cells has its limitations. In particular, no detailed information can be obtained about how a gene functions in regulating processes that involve an interaction between different types of cell in the body, whether it be in the context of physiology (such as in the nervous system) or embryonic development. Defining gene function in this wider context requires the genetic manipulation of model organisms.

Understanding gene function in model organisms

The most powerful way to understand gene function is to manipulate genes in the germ line, and then analyze the phenotype of the whole organism so that information can be obtained at each of the three levels listed above. That is not an option in humans (although we may gain valuable information from studying people with genetic mutations). Instead, we rely heavily on extrapolation from genetic manipulation of various model organisms to infer the functions of human genes. That may be done in different ways: by introducing precise changes at the DNA level to alter a gene; by blocking gene expression at the RNA or protein level; or by introducing and expressing transgenes so that the gene product is either produced in abnormally large quantities (overexpression) or produced in cells or tissues where it is not normally expressed—see **Figure 9.16** for an overview, and see Section 8.6 for the technology.

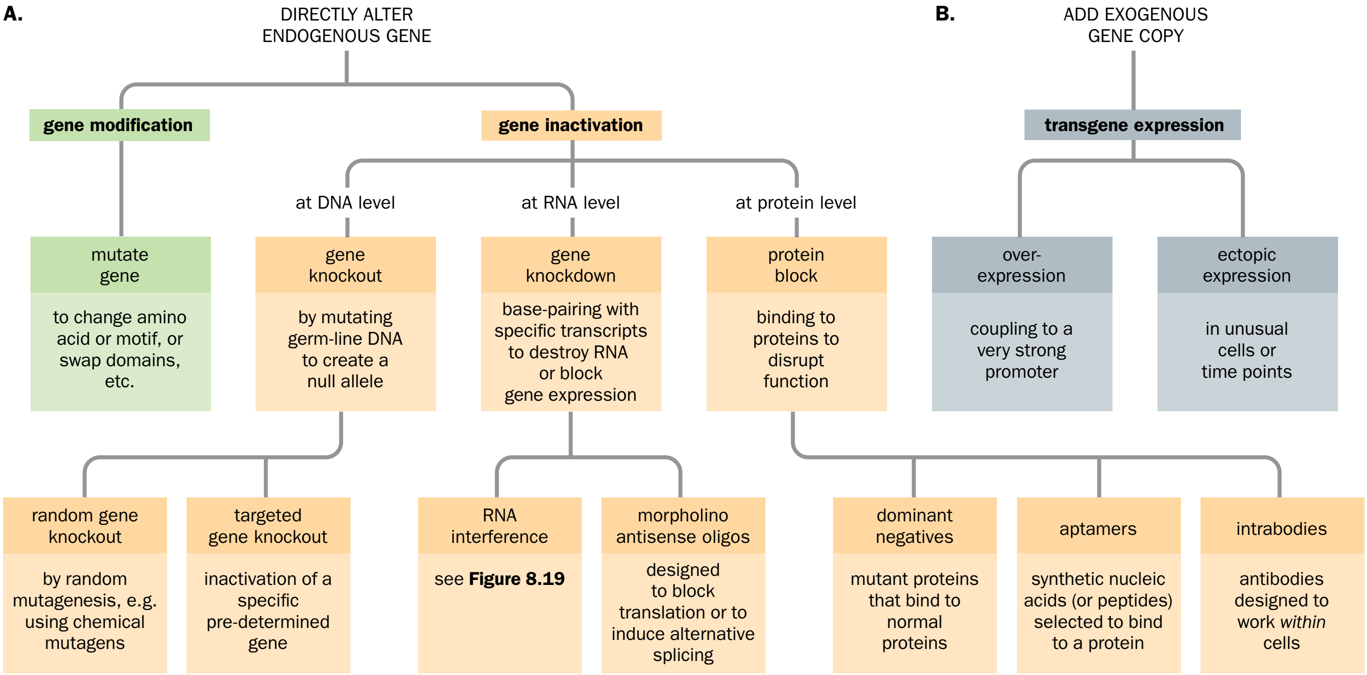

A.
DIRECTLY ALTER ENDOGENOUS GENE
gene modification
gene inactivation
mutate gene
to change amino acid or motif, or swap domains, etc.
at DNA level
at RNA level
at protein level
gene knockout
by mutating germ-line DNA to create a null allele
gene knockdown
base-pairing with specific transcripts to destroy RNA or block gene expression
protein block
binding to proteins to disrupt function
random gene knockout
by random mutagenesis, e.g. using chemical mutagens
targeted gene knockout
inactivation of a specific pre-determined gene
RNA interference
see Figure 8.19
morpholino antisense oligos
designed to block translation or to induce alternative splicing
dominant negatives
mutant proteins that bind to normal proteins
aptamers
synthetic nucleic acids (or peptides) selected to bind to a protein
intrabodies
antibodies designed to work *within* cells
B.
ADD EXOGENOUS GENE COPY
transgene expression
over-expression
coupling to a very strong promoter
ectopic expression
in unusual cells or time points

Figure 9.16 Principal ways of studying gene function *in vivo* by genetically modifying model eukaryotes. (**A**) Most methods seek to inactivate a gene to produce a mutant phenotype that can be correlated with a specific gene (peach boxes). Usually this is done at the level of DNA, where large-scale random mutagenesis screens have been performed in many models, including both invertebrates (such as *D. melanogaster*) and vertebrates (notably zebrafish and mice). Specific knockouts of pre-determined genes have also been conducted, notably by using homologous recombination in mice. The expression of a specific gene can also be selectively inhibited by targeting its transcripts—large-scale RNA interference-based genetic screens have been conducted in *C. elegans* and *D. melanogaster,* and gene knockdown with antisense morpholino oligonucleotides is commonly performed in zebrafish embryos. Blocking proteins with dominant-negative mutant proteins, aptamers, or intrabodies also has potential roles in treating disease (by specifically down-regulating harmful genes) and is discussed in more detail in Chapter 22. (**B**) An alternative approach is to use a transgene to overexpress a specific gene or to express it in tissues or developmental stages in which it is not normally expressed (*ectopic expression*) in an attempt to produce a phenotype that will provide clues to the function of the gene.

Germ-line gene inactivation is the most common approach to defining the function of a pre-determined gene in model organisms. Genetically amenable invertebrate models such as *Drosophila melanogaster* and *Caenorhabditis elegans* have been invaluable for understanding gene function, and even unicellular yeasts have provided valuable information that has illuminated how some well-conserved genes work. Inevitably, however, because these models are evolutionarily distantly related to us, they cannot provide information on many functionally important human genes with vertebrate-specific functions.

Of the vertebrate model organisms, two models stand out because of their amenability to genetic manipulation. The zebrafish is advantaged by a short generation time, production of a large number of eggs at mating, external fertilization (so that all aspects of development are readily accessible), and a transparent embryo (facilitating identification of developmental mutants). However, the most useful model organism to infer human gene function has been the mouse: being a mammal, its physiology is very similar to ours, it has a short generation time and litter sizes are large, and it has long been amenable to genetic experimentation (mutants have been produced and studied over decades). In mice, inactivation of a pre-determined gene has typically been done by mutating the gene in embryonic stem cells (ESCs) to produce a null allele, a **gene knockout**, after which the mutant gene is introduced into the germ line by transferring the mutant ESCs into the inner cell mass of a blastocyst that is then implanted in the uterus of a foster mother (see **Figure 8.22**).

Many research labs have produced mutant mice by knocking out a pre-determined gene of interest. Recently, however, the field has been transformed by a systematic and comprehensive approach toward understanding gene function that seeks to make the mouse the first mammal for which there will be a truly comprehensive functional catalog. Two key developments toward achieving this aim are listed below.

- The International Knockout Mouse Consortium (IKMC). Various USA and European research centers are collaborating to make a public resource of mouse embryonic stem cells that collectively contains a null mutation for every gene in the mouse genome, and the targeting strategies allow for producing both standard null mutations and *conditional gene knockouts*, where a null mutation can be activated later in development once the embryo has progressed through critical stages of early embryogenesis (see **Box 8.3**). For a general background to the IKMC, see http://www.mousephenotype.org/about-ikmc, and for the targeting strategies used to make the gene knockouts, see http://www.mousephenotype.org/about-ikmc/targeting-strategies.
- The International Mouse Phenotyping Consortium (IMPC; at http://www.mousephenotype.org). The IMPC is generating a knockout mouse strain for every protein-coding gene by using the ES cell resource generated by the IKMC. Systematic broad-based phenotyping is performed by each IMPC center using standardized IMPReSS (International Mouse Phenotyping Resource of Standardised Screens) procedures (http://www.mousephenotype.org/impress).

Genetic screens in cultured cells

Although germ-line manipulation is the gold standard in functional screens, useful information can also be obtained by carrying out global screens of genetic function in cultured human and mammalian cells, and a variety of different reverse genetic screens can be carried out (for a background, see the review by Tochitani & Hayashizaki [2007], PMID 17308666, under Further Reading).

Until quite recently, the most commonly used genetic screens in human and mammalian cells depended on gene silencing using RNA interference to suppress gene expression. Sometimes that involves making siRNA libraries: pairs of complementary siRNAs are synthesized for each gene in the genome and deposited as individual pairs in wells of huge microtiter plate arrays. Alternatively, one makes libraries of genes encoding short hairpin RNA; the idea is to mimic the way in which miRNA is naturally made (it forms a short transcript that spontaneously forms a hairpin RNA that then undergoes cleavage—for the details see **Figure 8.19**).

Because RNAi-based silencing does not completely disrupt gene expression, the modern preference now is to use an alternative: genome-wide gene knockout screens using genome editing with the CRISPR-Cas9 system. As an example, see Shalem *et al*, (2014) (PMID 24336571).

International efforts to produce human proteome maps and human protein interactomes

Human proteomes are nowhere near so complex as human transcriptomes. There are, however, large differences in both protein abundance and also expression (both cell type-dependent and time-dependent). Isoforms are also quite common as a result of alternative processing (notably alternative splicing and alternative promoter usage). Although proteomes are dependent on the type of cell, it is possible to sum the proteomes of different cell types to produce a composite draft human proteome, a catalog of human proteins and peptides.

Draft human proteomes were published in 2014 by two research teams using high-resolution mass spectrometry to analyze multiple human tissues and cell lines. In both cases, because of limitations of current technologies, the number of proteins identified (>17,000) was significantly less than the expected number. This type of approach has been extended to include the first assays of the abundance and expression patterns of proteins in human tissues at the proteome level. A tissue-based map of the human proteome, published in 2015, was obtained by combining quantitative transcriptomics (using RNA-Seq on 32 human tissues) with protein profiling using microarray-based immunohistochemistry in 44 human tissues and organs. The latter method was able to achieve spatial localization of human proteins in tissues down to the single-cell level and employed 24,028 antibodies to track the proteins produced by 16,975 human genes. And a landmark study published in 2017 has produced a subcellular map of the human proteome, defining the proteomes of 13 major organelles.

Interaction proteomics

The ENCODE Project mapped out some protein–DNA interactions in cells, notably binding of transcription factors, but protein–protein interactions are also key to how cells function. Systematic large-scale approaches have begun to define the network of all protein–protein interactions within cells, the protein interactome, and often use yeast two-hybrid screening (see **Box 9.4** for the basic yeast two-hybrid method). Individual screens can suffer from a high rate of false positives, but data are typically validated by different confirmatory methods and may be further backed up by supporting evidence from genetic interactions and gene expression data (interacting proteins need to be co-expressed in the same tissue or cell).

Extensive protein interactomes have been built up for various simple model eukaryotes, and concerted efforts are being made to define major protein interaction networks in human cells. A proteome-scale map of the human interactome network, published in 2014, documented ~14,000 high-quality, human binary protein–protein interactions. Various databases have been established to assimilate and present these and other human proteome data (**Table 9.15**). However, although the goal remains to link all the proteins in cells into functional pathways, protein interaction networks are not simple to represent. One estimate is that there may be up to 600,000 human protein–protein interactions; if so, clear data presentation will become a significant challenge (**Figure 9.17**). Another significant challenge is that there is no single human interactome: protein interaction networks are dynamic and context dependent.

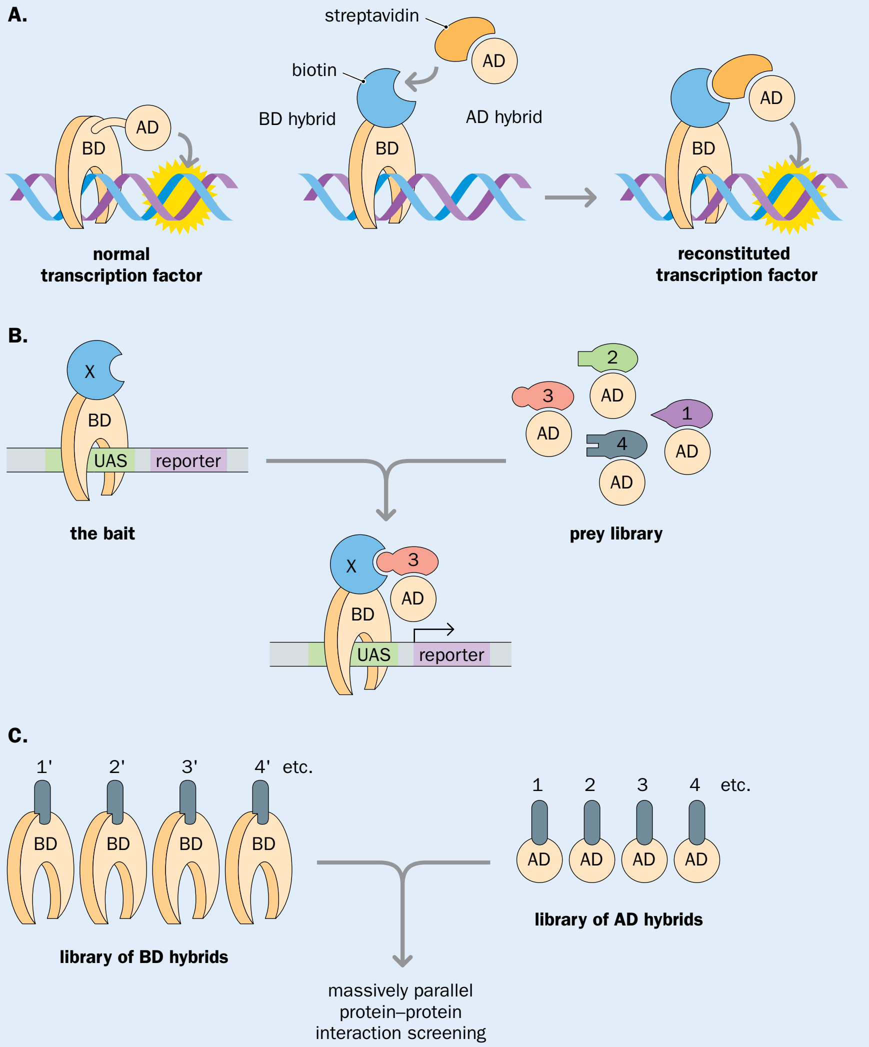
BOX 9.4 IDENTIFYING PROTEIN–PROTEIN INTERACTIONS USING YEAST TWO-HYBRID SCREENING

In yeast two-hybrid assays, specific protein–protein interactions are detected by generating a functional transcription factor that is not normally made in the host yeast cell. The newly generated transcription factor then activates a reporter gene and/or selectable marker.

The method relies on the observation that transcription factors have two key domains that can maintain their function when separated: a DNA-binding domain (BD) and a transcription activation domain (AD). Coding sequences for individual BD and AD domains of a specific transcription factor can be joined to other cDNAs to produce hybrid (fusion) proteins carrying either a BD or AD domain. If a protein that is fused to the BD domain of the transcription factor interacts with another that is coupled to the AD domain of the same kind of transcription factor, the close association of BD and AD domains can produce a functional transcription factor even though the two domains are located in different proteins (**Figure 1**).

The object of a standard two-hybrid screen is to use a specific protein of interest as a bait for specific recognition by an interacting protein selected from a large library. A haploid yeast strain is designed to express the bait protein of interest coupled to a DNA-binding domain of a transcription factor. It is mated with yeast strains from a library of haploid yeast cells that each express a single type of protein fused to the appropriate transcription factor activation domain but collectively express thousands of different hybrid proteins (the prey library; see **Figure 1B**).

The target cells are also engineered to carry a reporter gene and/or a selectable marker gene that cannot be activated until the intended transcription factor has been assembled. Interactions are tested in the resulting diploid yeast cells. If the bait and prey do not interact, the two transcription factor domains remain separate; if they do interact, a new transcription factor is assembled and the marker gene is then activated, facilitating the visual identification and/or selective propagation of yeast cells containing interacting proteins. From these cells, the cDNA sequence of the prey construct can be identified.



Box 9.4 Figure 1 Yeast two-hybrid (Y2H) screening. (**A**) An active transcription factor comprises complementary DNA-binding (BD) and transcription activation (AD) domains. A cDNA for either domain can be spliced to cDNA for other proteins, producing AD and BD hybrid (fusion) proteins. Binding of the attached proteins, as for biotin and streptavidin as shown here, will associate the AD and BD hybrid proteins, reconstituting the transcription factor. (**B**) In a standard Y2H screen, one yeast strain expresses the *bait*, a BD hybrid protein with a specific test protein (X) fused to a DNA-binding domain for a specific transcription factor. The BD hybrid protein binds to a specific upstream activating sequence (UAS) positioned before a reporter gene but is inactive in the absence of the relevant AD domain. Possible protein partners for protein X are sought from within a *prey library* of numerous yeast strains containing different proteins linked to the relevant AD domain. Crossing of the bait strain with a prey strain that makes a protein that binds to protein X (protein 3 in this example) can produce a diploid cell in which AD and BD domains associate. As a result, the transcription factor is reconstituted and drives expression of the reporter gene. (**C**) Crossing of yeast cells from a library of cells expressing BD hybrid proteins with cells from a library expressing AD hybrid proteins permits massively parallel protein–protein interaction screening.

TABLE 9.15 SOME IMPORTANT PUBLICLY ACCESSIBLE ELECTRONIC RESOURCES FOR HUMAN PROTEOME AND PROTEIN INTERACTION DATA

Database	Description	Web site
Human Proteome Map	Integrates the massive peptide sequencing results from the draft human proteome maps	www.humanproteomemap.org
ProteomicsDB	Draft human proteome data from the Technical University of Munich's human proteome project	https://www.proteomicsdb.org/
Human Protein Atlas	High resolution images showing the spatial distribution of proteins in normal human tissues and cancer types and human cell lines	www.proteinatlas.org
Human Proteinpedia	Community portal for sharing and integration of human protein data; as well as protein–protein interaction (PPI) data, it also holds data on post-translational modification, tissue and cell line expression, subcellular localization, etc.	http://www.humanproteinpedia.org/
HPRD (Human Protein Reference Database)	Integrates data deposited in the Human Proteinpedia along with literature information curated in the context of individual proteins	http://www.hprd.org/

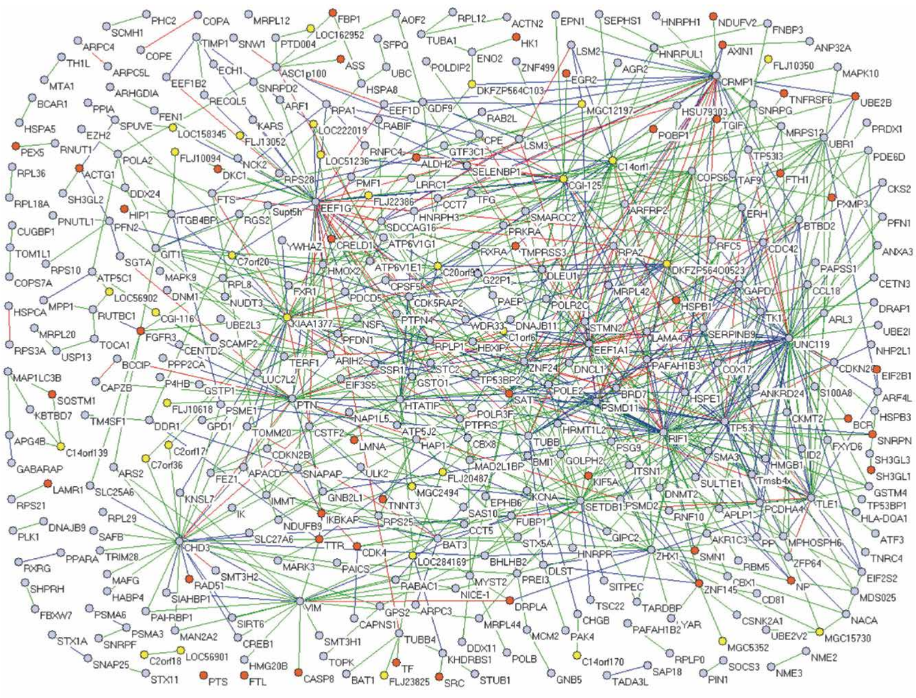


Figure 9.17 A small subset of the human protein interactome. Shown here is an early example, displaying a human protein interaction network with just 401 human proteins linked via 911 interactions. Orange, disease proteins (according to OMIM Morbid Map; NCBI); light blue, proteins with GO (gene ontology) annotation; yellow, proteins without GO and disease annotation. Interactions connecting the nodes are represented by color-coded lines according to their confidence scores: green, 3 quality points; blue, 4 quality points; red, 5 quality points; purple, 6 quality points. A human protein interactome published in 2014 charts a described ~14,000 binary protein–protein interactions and the full interactome can be expected to involve hundreds of thousands of interactions. That will not be easy to display! (From Stelzl U *et al*. [2005] *Cell* **122**:957–968; PMID 16169070. With permission from Elsevier.)

SUMMARY

- The human genome comprises the set of 24 different chromosomal DNA molecules (the nuclear genome, where the vast majority of our genes reside), plus the tiny circular mitochondrial genome (which, although gene-dense, has just 37 genes).

- Mitochondria have their own ribosomes, and mitochondrial ribosomal RNAs and transfer RNAs are all made by genes in mtDNA; >99% of mitochondrial proteins, however, are encoded by nuclear genes and translated on cytoplasmic ribosomes.

- Most of the genome is transcribed, and the great majority of the transcripts are noncoding; coding DNA sequences, which ultimately specify a protein or peptide, account for just over 1% of the genome.

- The human genome reference sequence is a composite: different genome regions are represented by DNA from different individuals. Although the reference sequence is mostly haploid, some genome regions are represented by multiple different haplotypes to take into account structural variation. The sequence has gaps because of difficulty in clone assembly and sequencing for some regions.

- Human genes are not always discrete entities: the transcripts of some genes overlap those from neighboring genes, sometimes on both strands.

- Genes are traditionally divided into those that have coding DNA to make proteins, and those that make functional noncoding RNA (RNA genes). However, an average human protein-coding gene makes about six to seven transcripts including some noncoding transcripts, and some genes are known to make both a protein and a functional noncoding RNA.

- The number of genes in the human genome can never be an exact number: RNA genes can be difficult to define, and even protein-coding genes can vary in number between haplotypes because of structural variation.

- Periodic duplication of single genes and subchromosomal regions, plus evolutionarily ancient whole-genome duplication, has given rise to families of related genes, increasing functional diversity.

- Some copies of a functional gene have acquired inactivating mutations and cannot make the expected gene product. These pseudogenes originate either by copying genomic DNA (nonprocessed pseudogenes) or by retrotransposition: a processed RNA transcript is reverse-transcribed and re-integrates into the genome (processed pseudogene).

- Retrogenes are intronless genes that originated by retrotransposition, acquired the ability to be expressed, and retained function because of selection pressure.

- Long arrays of high-copy-number tandem repeats, known as satellite DNA, are associated with highly condensed, constitutive heterochromatin, but these sequences can be transcribed. The DNA underlying telomeric heterochromatin is also transcribed to give RNA transcripts that regulate various facets of telomere biology.

- The majority of the human genome is made up of sequences originating from transposons, DNA elements that can jump from one genomic location to another. A small percentage of the originating transposons were DNA transposons that moved by a "cut-and-paste" mechanism, but the great majority were retrotransposons that jumped by a "copy-and-paste" mechanism.

- The vast majority of human transposon repeats are now incapable of transposing: during evolution they picked up inactivating mutations or have lost some functional sequence. The most evolutionarily recent transposon families are the ones with the highest proportion of full-length repeats and actively transposing repeats.
- The ENCODE Project was a 10-year international project that was the first systematic attempt to define the functional DNA sequences in our genome, including both genes and regulatory elements. Although the project reported its conclusions in 2012, this type of work is unending, and continues in many research centers across the globe.
- Genome annotation involves using descriptive text to illustrate the characteristics of genes, gene products, and regulatory elements in genome databases and browsers.
- Genome-wide screens of human gene function are limited to genetic screens (by RNA silencing or gene knockout) in cultured cells. More comprehensive insights into human gene function are possible by extrapolating from studies of model organisms, notably mice with artificially designed gene knockouts.
- Unlike the human genome, the transcriptome and proteome vary enormously in different cell types. Proteomes are much less complex than transcriptomes. The Human Proteome Project seeks to obtain a human proteome map by defining all the peptides in multiple cell types using mass spectrometry analysis and aggregating them.
- Large-scale protein interaction studies seek to define interactomes, global functional protein networks in cells.

FURTHER READING

Human mitochondrial genome

Anderson S *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**:457–465; PMID 7219534.

Andrews RM *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**:147; PMID 10508508.

Falkenberg M *et al.* (2007) DNA replication and transcription in mammalian mitochondria. *Annu Rev Biochem* **76**:679–699; PMID 17408359.

MITOMAP. Human mitochondrial genome database. http://www.mitomap.org

Human nuclear genome

Church DM *et al.* (2015) Extending reference assembly models. *Genome Biol* **16**:13; PMID 25651527.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**:860–921; PMID 11237011.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**:931–945; PMID 15496913.

Miga KH (2015) Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Res* **23**:421–426; PMID 26363799.

Nature Collections: Human Genome Supplement, 1 June 2006 issue. (Includes papers analyzing the sequence of each chromosome plus reprints of commentaries and the IHGSC paper reporting the 2001 draft sequence; available at http://www.nature.com/nature/supplements/collections/humangenome/)

Schneider VA *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**:849–854; PMID 28396521.

Gene duplication, pseudogenes, and retrogenes

Bailey JA *et al.* (2002) Recent segmental duplications in the human genome. *Science* **297**:1003–1007; PMID 12169732.

Conrad B & Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* **8**:17–35; PMID 17386002.

Kaessmann H *et al.* (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**:19–31; PMID 19030023.

Sasidharan R & Gerstein M (2008) Protein fossils live on as RNA. *Nature* **453**:729–731; PMID 18528383.

Zheng D & Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet* **23**:219–224; PMID 17382428.

Noncoding RNAs

Chen L-L (2016) The biogenesis and emerging role of circular RNAs. *Nat Rev Mol Cell Biol* **17**:205–211; PMID 26908011.

He Y *et al.* (2008) The antisense transcriptomes of human cells. *Science* **322**:1855–1858; PMID 19056939.

Iyer MK *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**:199–208; PMID 25599403.

Kopp F, Mendell JT (2018) Functional classification and experimental dissection of long noncoding RNAs. *Cell* **172**:393–407; PMID 29373828.

Matera AG *et al.* (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**:209–220; PMID 17318225.

Ponting CP *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell* **136**:629–641; PMID 19239885.

Sampath K & Ephrussi A (2016) CncRNAs: RNAs with both coding and non-coding roles in development. *Development* **143**:1234–1241; PMID 27095489.

Heterochromatin DNA and transposon repeats

Aldrup-Macdonald ME & Sullivan BA (2014) The past, present, and future of human centromere genomics. *Genes* **5**:33–50; PMID 24683489.

Biscotti MA *et al.* (2015) Transcription of tandemly repetitive DNA: functional roles. *Chromosome Res* **23**:463–477; PMID 26403245.

de Koning APJ *et al.* (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**:e1002384; PMID 22144907.

Deininger P (2011) Alu elements: know the SINEs. *Genome Biol* **12**:236; PMID 22204421.

Dewannieux M & Heidmann T (2013) Endogenous retroviruses: acquisition, amplification and taming of genome invaders. *Curr Opin Virol* **3**:646–656; PMID 24004725.

Goodier JL (2016) Restricting retrotransposons. *Mob DNA* **7**:16; PMID 27525044.

Hancks DC & Kazazian HH Jr (2012) Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* **22**:191–203; PMID 22406018.

Hayden KE *et al.* (2013) Sequences associated with centromere competency in the human genome. *Mol Cell Biol* **33**:763–772; PMID 23230266.

Molaro A & Malik HS (2016) Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. *Curr Opin Genet Dev* **37**:51–58; PMID 26821364.

ENCODE Project and functional genomics

ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**:e1001046; PMID 21526222.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74; PMID 22955616. (The *Nature* ENCODE Explorer at http://www.nature.com/encode/#/threads collates the many papers from the ENCODE project.)

Kellis M *et al.* (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA* **111**:6131–6138; PMID 24753594.

Shalem O *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**:84–87; PMID 24336571.

Skarnes WC *et al.* (2011) A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**:337–342; PMID 21677750.

Tochitani S & Hayashizaki Y (2007) Functional screening revisited in the postgenomic era. *Mol Biosyst* **3**:195–207; PMID 17308666. (The focus is on functional screens in cultured mammalian cells.)

Human Proteome Project and Interactome mapping

Baker MS *et al.* (2017) Accelerating the search for the missing proteins in the human proteome. *Nat Commun* **8**:14271; PMID 28117396.

Lawrence RT & Villén J (2014) Drafts of the human proteome. *Nat Biotechnol* **32**:752–753; PMID 25101745.

Legrain P *et al.* (2011) The Human Proteome Project: current state and future direction. *Mol Cell Proteomics* **10**:M111.009993; PMID 21742803.

Rolland T *et al.* (2014) A proteome-scale map of the human interactome network. *Cell* **159**:1212–1226; PMID 25416956.

Thul PJ, Lindskog C (2018) The human protein atlas: a spatial map of the human proteome. *Protein Sci.* **27**:233–244; PMID 28940711.

Thul PJ *et al.* (2017) A subcellular map of the human proteome. *Science* **356**:eaa13321; PMID 28495876.

Uhlén M *et al.* (2015) Tissue-based map of the human proteome. *Science* **347**:1260419; PMID 25613900.

10 Gene regulation and the epigenome

The basics of transcription and translation were covered in Chapter 1; the subject of this chapter is how those processes are regulated. Two numbers underscore the central importance of gene regulation in making us who we are. First, 20,338—the number of protein-coding genes in the human genome according to Ensembl (November 2017; see **Table 9.5** for some alternative but not very different numbers). That number is amazingly small. Before the Human Genome Project most scientists assumed the number would be 100,000 or maybe much higher, commensurate with our far greater complexity compared to organisms like yeast, flies, or worms. Second, 20,362—the number of protein-coding genes in the 1 mm long nematode worm *Caenorhabditis elegans*, a creature widely studied as one of the very simplest multicellular organisms. Since we have no more protein-coding genes than that simple worm, our far greater complexity must be because we use the same genes in a smarter way. In other words, gene regulation is the essence of what makes us human.

Regulation can be short- or long-term. Short-term regulation allows cells to respond to a fluctuating environment. Long-term patterns of selective gene expression govern the permanent tissue-specific identity of a cell, and are stable through mitosis so that tissue identity can be transmitted from cell to daughter cell. Mechanisms that alter gene expression without altering the underlying gene sequence are collectively termed **epigenetic** (literally, above genetics), although that word is usually reserved for changes that are heritable through mitosis. Sometimes epigenetic changes can be transmitted across generations, from parent to child, although it is controversial how great a role transgenerational epigenetic changes play in human life. At the DNA level, the mechanisms underlying short-term and long-term regulation of gene expression overlap.

All the cells of our body, apart from gametes, are derived by repeated mitosis from the original fertilized egg cell. All those cells, with a few minor exceptions, therefore contain exactly the same set of genes. The differences in anatomy, physiology, and behavior between cells, described in Section 2.1 (see also **Figure 3.17**), are the result of differing readouts of that fixed set of genes. During normal development, cells follow branching trajectories in which successively more specialized patterns of gene expression define the transitions from totipotency through pluripotency and onward to terminal differentiation. These developmental decisions are the result of epigenetic mechanisms starting in the earliest stages of development, as described in Section 4.1. Most of them are irreversible in the context of normal human physiology, although there is great interest in reversing them by artificial manipulation, to produce induced pluripotent stem cells (see Section 4.2).

Sophisticated control of gene expression is central to the way our genome functions. Thus it should come as no surprise that human gene regulation involves many interacting players and mechanisms, operating on both transcription and translation. The ENCODE project, described in Section 9.4, was the first systematic attempt to catalog all functional elements of the human genome, most of which function in gene regulation. The reports from that project (accessible through the Nature ENCODE Explorer, www.nature.com/encode/) show how far we have come from a picture of our genome as comprising thinly scattered protein-coding sequences in a vast sea of largely nonfunctional DNA. Controversially, the ENCODE reports ascribed some biochemical function to up to 80% of all nucleotides in the genome. That figure is widely felt to have used an unrealistically broad definition of function, but certainly the majority of all sequences are transcribed, at least in some cells and at some times. We consider in Chapter 13 the extent to which the ENCODE data have changed perceptions on the amount of functionally important DNA in our genome.

The regulatory landscape varies between tissues and cell types. We have only two genomes (remember we are diploid), but potentially hundreds of epigenomes. Fully documenting epigenomes is a much larger task than documenting the basic DNA sequence of the human genome. The Roadmap Epigenomics Consortium (www.roadmapepigenomics.org/) and the International Human Epigenome Consortium (www.ihec-epigenomes.org/) are among endeavors to move forward from ENCODE. The 2015 publication from the Roadmap Epigenomics Consortium (see Further Reading) gives a foretaste of the complexity to come. Perhaps the central insight from all these studies is that gene expression is not regulated by a series of one-dimensional linear processes, but by a network of interacting systems. In this chapter we necessarily describe individual components of the network separately—chromatin accessibility, histone modifications, DNA-binding proteins, DNA methylation, promoters, and enhancers—but it is important to bear in mind that no one of them is independent of the others. A key overall consideration, and our starting point here, is accessibility. A regulatory element can only function if its DNA is accessible, to allow regulatory proteins or RNAs to bind it.

10.1 CHROMATIN ACCESSIBILITY AND CONFORMATION

Active regulatory sequences cannot be occluded in tightly packaged chromatin; they must be free to interact with DNA-binding proteins and/or RNA species. Three main methods can be used to identify regions of chromatin where the DNA is easily accessible:

- Identifying DNase hypersensitive sites: as mentioned in Section 9.4, when intact cell nuclei are lightly digested with the nonspecific endonuclease DNase I, multiple cuts are made preferentially in the regions of the DNA that are most accessible to the enzyme. Sequencing the fragments this generates allows genome-wide identification of regulatory regions;
- FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) uses the fact that formaldehyde preferentially cross-links nucleosome-bound DNA. Intact cells are treated with formaldehyde, the chromatin sheared by sonication, and DNA extracted. Free extracted DNA comes preferentially from nucleosome-free regulatory regions, and can be sequenced to identify these (Giresi & Lieb [2009], PMID 19303047). FAIRE is an alternative and rather simpler method to obtain similar information to DNase hypersensitive site mapping;
- ATAC (Assay for Transposase-Accessible Chromatin) uses a hyperactive mutant Tn5 bacterial transposase to insert primers for next-generation sequencing into regions of accessible chromatin (Buenrostro *et al.* [2013], PMID 24097267). Amplification and sequencing of the tagged fragments generates data similar to DNase hypersensitive site mapping or FAIRE, but with simpler protocols and using much smaller numbers of cells—potentially even single cells.

These techniques have been used to generate genome-wide maps of regions of highly accessible DNA, which include putative promoters and enhancers. Accessibility is determined by a set of interacting and mutually reinforcing factors, including nucleosome positioning, covalent modifications of histones in nucleosomes, and DNA methylation. These are described in turn below.

The two meters of DNA in a diploid cell nucleus is not simply crammed in; chromatin is highly structured. At the resolution obtainable under the light microscope, chromosomes occupy distinct and largely non-overlapping territories within the nucleus. These can be revealed by fluorescence *in situ* hybridization using chromosome paints (**Figure 10.1**; see also **Figure 2.20**).

Zooming in, the spatial relationship of individual sequences can be investigated using either fluorescence *in situ* hybridization of pairs of loci or a set of related techniques collectively called chromosome conformation capture (**Box 10.1**). These have shown that interphase chromosomes are organized into **topologically-associated domains** (TADs) typically 500 kb in size. TADs are highly conserved across cell types and species, and appear to be fixed structural elements of chromosomes.

Functionally, the genome can be partitioned into active and repressed segments; these are typically of the order of 5 Mb in size and contain different TADs in different cell types. Segments of each type tend to cluster in the nucleus. Central locations and positions on the outside of a chromosome territory are associated with transcriptionally

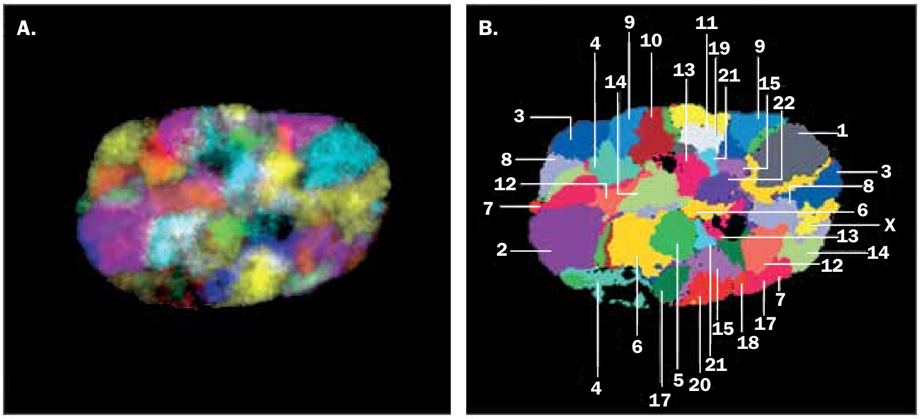
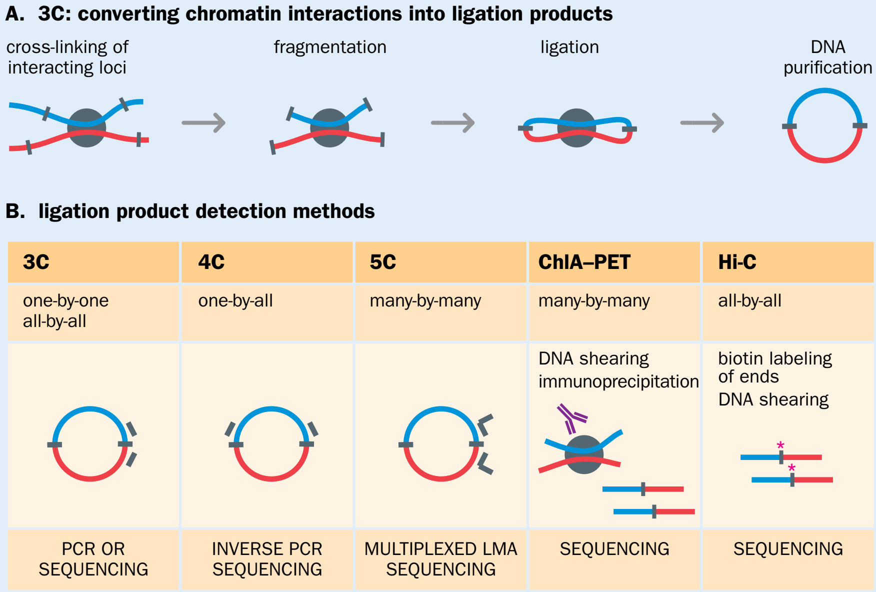


Figure 10.1 Chromosome territories within the cell nucleus. When chromosomes are labeled with fluorescent chromosome paints, they are seen to occupy discrete regions of the nucleus, although with contacts at the boundaries. (**A**) False-color representation of chromosomes differentially labeled with combinations of fluorophores. (**B**) Identification of individual chromosomes in a nucleus. (Reprinted from Speicher MR & Carter NP [2005] *Nat Rev Genet* **6**:782–792; PMID 16145555. With permission from Springer Nature. Copyright © 2005.)

BOX 10.1 CHROMOSOME CONFORMATION CAPTURE

Chromosome conformation capture (3C) is a method for identifying DNA sequences that may be widely separated in the genome sequence but lie physically close together within the interphase cell nucleus. Cells are treated with formaldehyde to cross-link regions of chromatin that lie physically close to one another. The cross-linked chromatin is solubilized and digested with a restriction enzyme. Interacting sequences will be represented by cross-linked chromatin particles containing a DNA fragment from each of the interacting sequences. The fragments are ligated by treatment with DNA ligase in very dilute solution, to favor ligation of fragments contained in the same chromatin particle. The ligated DNA fragments can then be identified by various methods (Figure 1).

- 3C explores the interaction of a predefined “anchor” sequence, such as a promoter, with sequences located a few tens or hundreds of kilobases away, using one PCR primer specific for the anchor and others chosen to amplify candidate targets nearby. Alternatively, an entire 3C library can be sequenced *en masse*, giving all-by-all interaction data.
- 4C (chromosome conformation capture-on-chip) uses inverse PCR to generate genome-wide interaction profiles for single loci.
- 5C (chromosome conformation capture carbon copy) combines 3C with hybrid capture approaches to identify up to millions of interactions in parallel between two large sets of loci: for example, between a set of promoters and a set of distal regulatory elements.
- HiC includes a biotin-labeling step that allows selective purification and mass sequencing of ligation junctions, to give an unbiased genome-wide view of interactions.
- ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) uses chromatin immunoprecipitation to give a genome-wide analysis of long-range interactions between sites bound by a protein of interest.


A. 3C: converting chromatin interactions into ligation products
cross-linking of interacting loci
fragmentation
ligation
DNA purification
B. ligation product detection methods

3C	4C	5C	ChIA–PET	Hi-C
one-by-one all-by-all	one-by-all	many-by-many	many-by-many	all-by-all
			DNA shearing immunoprecipitation	biotin labeling of ends DNA shearing
PCR OR SEQUENCING	INVERSE PCR SEQUENCING	MULTIPLEXED LMA SEQUENCING	SEQUENCING	SEQUENCING

Box 10.1 Figure 1 Chromosome conformation capture. (**A**) Interacting regions of chromatin are cross-linked using formaldehyde, fragmented, and the associated two DNA fragments ligated together. (**B**) The ligated DNA can be analyzed in different ways to detect specific or general interacting partners. See text for details. LMA, ligation-mediated amplification; PCR, polymerase chain reaction. (Adapted from Dekker J *et al*. [2013] *Nat Rev Genet* **14**:390–403; PMID 23657480. With permission from Springer Nature. Copyright © 2013.)

active genes; positions near the nuclear lamina or within the interior of a chromosome territory are associated with inactive, repressed states. Transcription is concentrated in localized transcription factories.

To understand why some sequences are active and others inactive, it is necessary to look in detail at the local chromatin structure. A first distinction is between euchromatin and heterochromatin. As described in Section 2.4, heterochromatin is closed and repressed, while euchromatin is open and potentially transcriptionally active (see **Figure 2.19**). Heterochromatin has tightly packed nucleosomes that carry distinctive histone modifications (see below). These attract proteins like HP1 (CBX1) that help mediate gene silencing. Recent studies have developed a more nuanced description of chromatin conformation, documenting a dozen or more chromatin flavors that correlate with different functions and activities of the DNA, as described below.

Nucleosome positioning and chromatin remodeling complexes

Active transcription start sites typically have about 150 bp of nucleosome-free DNA, with flanking nucleosomes highly ordered. Binding sites for other proteins, such as the insulator protein CTCF (see below), are similarly flanked by regularly positioned nucleosomes. Nucleosome positioning can be altered by **chromatin remodeling complexes**. These are large, ATP-powered multiprotein machines that can physically shuffle nucleosomes along the DNA. As well as changing the position of nucleosomes, some remodeling complexes can swap variant histone molecules into and out of nucleosomes, altering their stability or function. Overall, they allow chromatin to transition between different states (for example, active versus repressed). Complexes are grouped into families such as BAF, INO80, ISWI, CHD, and so on, primarily by their ATPase subunit (humans have 29 different ATPase subunits), but they contain overlapping sets of other subunits. **Figure 10.2** illustrates how changing certain subunits of a complex can affect cell identity.

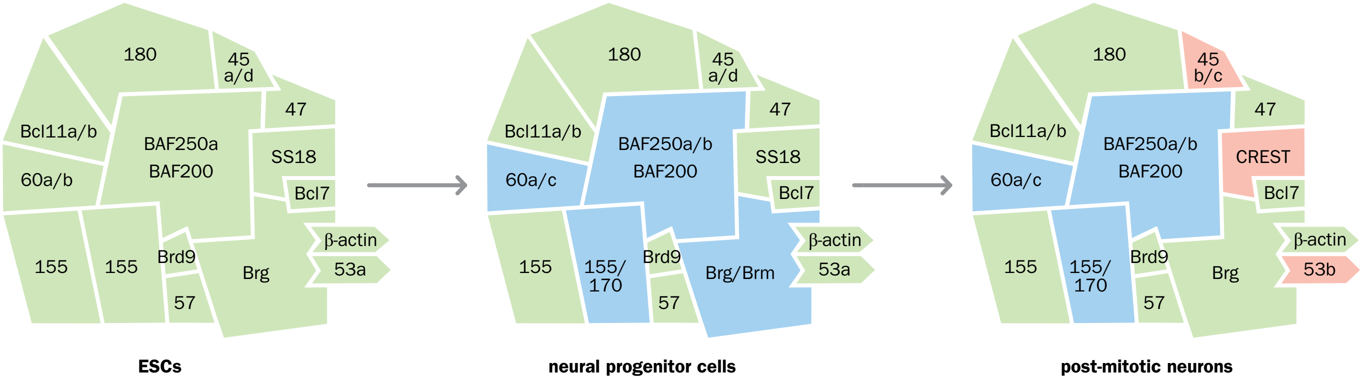


Figure 10.2 Changing subunits of a chromatin remodeling complex can change patterns of gene expression. The BAF complex always contains the same number of subunits from the same sets of families, but different family members are found in the complex in different cell types. The figure shows the complex in ESCs (embryonic stem cells), neural progenitors, and post-mitotic neurons. These differences help orchestrate different patterns of gene expression, leading to functional differences as shown. (Adapted from Son EY & Crabtree GR [2014] *Am J Med Genet C Semin Med Genet* **166C**:333–349; PMID 25195934. With permission from John Wiley & Sons, Inc., © 2014 Wiley Periodicals, Inc.)

As well as control of transcription, chromatin remodeling complexes are involved in other processes where changes in chromatin state may need managing, such as DNA replication, DNA damage repair, and chromosome segregation. Mutations in components of remodeling complexes are important in human disease, especially cancer, but also congenital conditions such as the Coffin-Siris group of syndromes (OMIM #135900). Relevant human genes are often members of the SMARC (SWI/SNF-related, matrix-associated, actin-dependent regulators of chromatin) family. Nomenclature of subunits is confusing, with subunits often having several alternative names and homologous subunits in different organisms having different names; of the subunits shown in **Figure 10.2**, Brg, Brm, and BAFs 47, 57, 60a,b, and c, 155, and 170 all have alternative SMARC names—see the OMIM entries.

10.2 HISTONES AND OTHER DNA-BINDING PROTEINS

We saw in Chapter 2 (see **Figure 2.18**) that nucleosomes consist of 146 bp of DNA wrapped around an octamer of eight molecules of histones. These small basic proteins carry a positive charge that gives them an affinity for the negatively-charged sugar–phosphate backbone of DNA. There are normally two molecules each of histones H2A, H2B, H3, and H4, although some nucleosomes contain variant histones such as H2A.X, H2A.Z, H3.3, or CENP-A. As shown in **Figure 2.18B**, the N-terminal tails of the histone molecules protrude from the core nucleosome and are accessible to external proteins. An extensive suite of enzymes attaches different small groups to specific amino acid residues in the tails (**Figure 10.3**). Common modifications include acetylation and mono-, di-, or trimethylation of lysines. Other important modifications include phosphorylation of serines H3S10 and H3S28, mono-, di-, or trimethylation of arginines H3R2, H3R8, H3R17, and H4R3, phosphorylation of threonines H3T3, H3T6, and H3T11, and ubiquitylation of lysines H2AK119 and H2BK120. See **Box 10.2** for an explanation of this nomenclature.

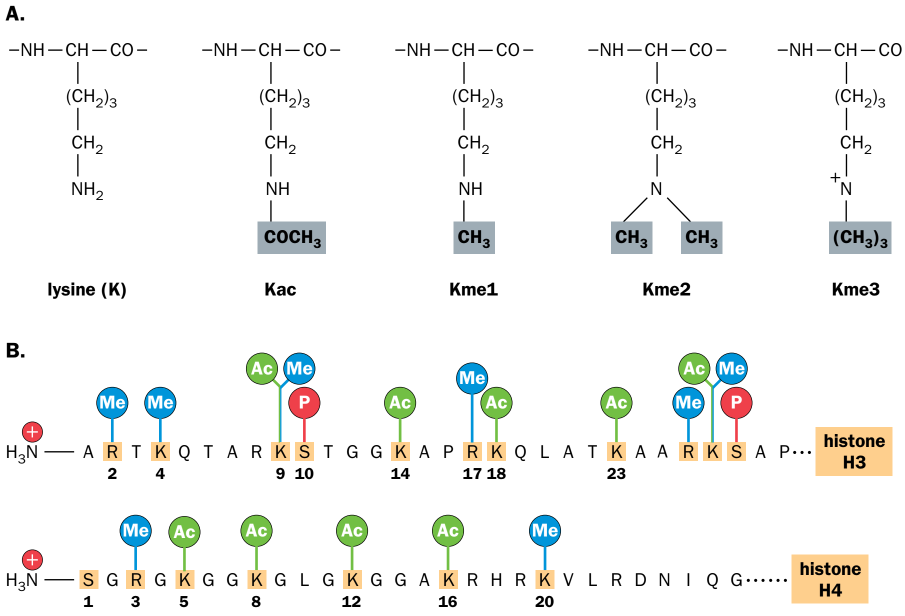


Figure 10.3 Modification of histone tails. (**A**) The ε-amino group of lysine residues can be modified by acetylation or the addition of one to three methyl groups. The standard nomenclature of the modified lysines is shown. (**B**) The N-terminal tails of histones H3 and H4 are the sites of many of the modifications that control chromatin structure. The amino acid sequence is shown in single-letter code, and a selection of potential modifications are indicated. Ac, acetylation; Me, methylation; P, phosphorylation.

BOX 10.2 NOMENCLATURE OF HISTONE MODIFICATIONS

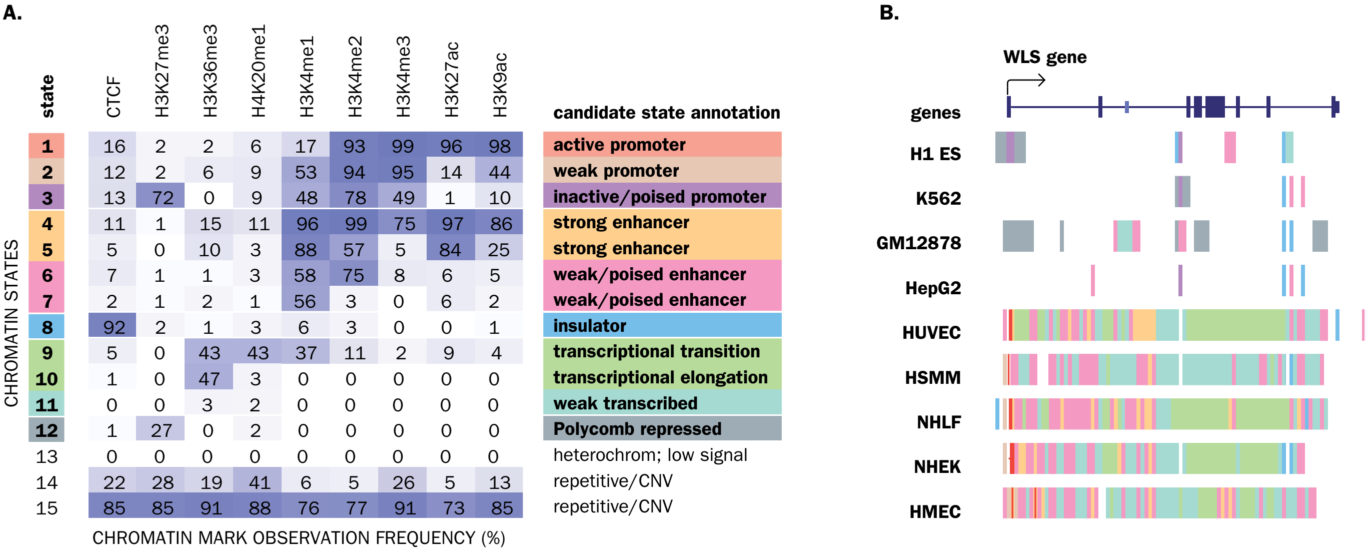
A common shorthand is used for histone modifications. Specific amino acid residues are identified by the type of histone, the one-letter amino acid code, and the position of the residue, counting from the N-terminus. Thus, H3K9 is lysine-9 in histone H3. Modifications are then described using ac for acetylation, me for methylation, p for phosphorylation, ub for ubiquitylation, and su for sumoylation. For example, H4K12ac is acetylated lysine-12 of histone H4 and H3K4me3 is trimethylated lysine-4 of histone H3.

Histone modifications can be mapped across the genome by chromatin immunoprecipitation followed by sequencing (ChIP-seq; see Section 9.4 for details). This procedure identifies the genome-wide locations of specific DNA-bound proteins—in this case, of nucleosomes carrying specific histone modifications. These data can then be compared to RNA sequencing data that show which sequences are being transcribed in a given cell type, and to data obtained by DNase-seq, FAIRE, or ATAC (see above) that identify regulatory sequences. Combining these sources of data allows a series of chromatin flavors to be defined that relate to specific states or activities in the genome. **Figure 10.4** shows one among several possible statistical clusterings.

Large numbers of enzymes are concerned in these modifications. Many act only on one particular residue in one particular class of histone. They can be grouped into three classes:

- Writers add groups. They include histone methyltransferases, histone acetyltransferases, and histone phosphokinases;
- Erasers remove groups. They include histone demethylases, histone deacetylases, and histone phosphatases;
- Readers bind to specific modified residues and initiate some action. For example, chromodomain proteins bind methylated histones, bromodomain proteins bind acetylated lysines.

Defects in one or another of these enzymes underlie many human clinical syndromes. **Table 10.1** shows selected examples.



state	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	candidate state annotation
1	16	2	2	6	17	93	99	96	98	active promoter
2	12	2	6	9	53	94	95	14	44	weak promoter
3	13	72	0	9	48	78	49	1	10	inactive/poised promoter
4	11	1	15	11	96	99	75	97	86	strong enhancer
5	5	0	10	3	88	57	5	84	25	strong enhancer
6	7	1	1	3	58	75	8	6	5	weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	insulator
9	5	0	43	43	37	11	2	9	4	transcriptional transition
10	1	0	47	3	0	0	0	0	0	transcriptional elongation
11	0	0	3	2	0	0	0	0	0	weak transcribed
12	1	27	0	2	0	0	0	0	0	Polycomb repressed
13	0	0	0	0	0	0	0	0	0	heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	repetitive/CNV
15	85	85	91	88	76	77	91	73	85	repetitive/CNV

Figure 10.4 Chromatin flavors and their relationship to function. (**A**) Machine-learning clustering of data from 90 chromatin immunoprecipitation experiments across 9 cell types identified 15 recurrent combinations (states 1–15, left). These could be linked to known functional annotations (right). Note that H3K9me3, characteristic of heterochromatin, is not shown here. CNV copy number variant (see **Figure 11.8**). (**B**) Example analysis of chromatin state maps across the *WLS* gene in 9 cell types (H1 ES, human embryonic stem cells; K562, erythrocytic leukemia cells; GM12878, B-lymphoblastoid cells; HepG2, hepatocellular carcinoma cells; HUVEC, umbilical vein endothelial cells; HSMM, skeletal muscle myoblasts; NHLF, normal lung fibroblasts; NHEK, normal epidermal keratinocytes; HMEC, mammary epithelial cells). The analysis shows how chromatin states (color-coded as in part A) vary between cell types depending on the pattern of gene expression. *WLS* encodes a G-protein-coupled cell surface receptor required for Wnt signaling that is "poised" in ESCs (see main text), repressed in GM12878 cells, and expressed in the five primary cell types. (Adapted from Ernst J *et al*. [2011] *Nature* **473**:43–48; PMID 21441907. With permission from Springer Nature. Copyright © 2011.)

TABLE 10.1 SELECTED EXAMPLES OF CLINICAL SYNDROMES CAUSED BY MUTATION IN WRITERS OR ERASERS OF HISTONE MODIFICATIONS

	Gene	Target	Syndrome	OMIM #
WRITERS				
Lysine methyltransferases	*KMT2D*	H3K4	Kabuki 1	147920
	EHMT1	H3K9	Kleefstra	610253
	EZH2	H3K27	Weaver	277590
	NSD1	H3K36	Sotos 1	117550
Histone acetyltransferases	*CREBBP*	*	Rubinstein-Taybi 1	180849
	EP300	*	Rubinstein-Taybi 2	613684
	KAT6B	*	Genitopatellar	606170
Histone phosphokinase	*RPS6KA3*	H3	Coffin-Lowry	303600
ERASERS				
Lysine demethylases	*KDM5C*	H3K4	Claes-Jensen	300534
	PHF8	H3K9, H4K20	Siderius	300263
	KDM6A	H3K27	Kabuki 2	300867
Histone deacetylases	*HDAC4*	*	Brachydactyly-ID	600430
	HDAC8	SMC3[1]	Cornelia de Lange 5	300882

* Histone acetyltransferases and deacetylases have broader substrate specificities than the lysine methyltransferases and demethylases.

[1] SMC3 is part of the cohesin complex that holds sister chromatids together, but is also involved in promoter–enhancer interactions (see main text). Brachydactyly-ID, brachydactyly with intellectual disability.

Transcription factors and other DNA-binding proteins

Transcription factors are DNA-binding proteins that control gene expression. The basal or general factors TFIIA, B, D, E, F, and H bind promoters and are required for transcription from most, if not all, RNA polymerase II promoters (see Section 1.3). However, the majority of the 2000 or so human transcription factors bind enhancers (see below) and are responsible for cell-type-specific gene expression. They act in hierarchies. A small number of master factors define cell identity by binding to so-called super-enhancers (unusually long and complex enhancers containing multiple binding sites; see Hnisz *et al*. [2013], PMID 24119843, in Further Reading). Subordinate transcription factors then refine the pattern of gene expression within each cell type.

Many years ago, long before any of this was understood, CH Waddington put forward the idea of an epigenetic landscape (**Figure 10.5**). He conceived a model of a ball rolling down a tilted three-dimensional surface with hills and bifurcating valleys. As the ball rolls down, its options are limited to the valleys that open up from the particular valley it is currently occupying. The further down the surface it rolls, the fewer its options are. This is a very nice metaphor for the progressive epigenetic restriction of differentiation potency as embryonic development proceeds. In 2017 we can put flesh on Waddington's concept. Each valley is defined by the battery of genes a cell expresses, and this depends on the transcription factors present. Among those genes are genes for further transcription factors, which in turn define the secondary valleys. Choices between valleys can depend on signals from the surrounding cells or medium, or they can be generated within a cell by asymmetric cell division, or simple chance (see Section 4.1). Transcription factors active in higher valleys may be actively turned off as differentiation proceeds so that they are diluted out as the cells multiply. Replacing them may reverse differentiation—simply adding the four transcription factors OCT4, SOX2, c-MYC, and KLF4 to differentiated somatic cells can cause them to revert to a pluripotent state (induced pluripotent stem cells; see Section 4.2). Short-term responses to external signals often work by activating pre-existing transcription factors, for example by relocating them from the cytoplasm to the nucleus (see **Figure 3.3**); long-term tissue identity is more likely to depend on the overall repertoire of transcription factors in a cell.

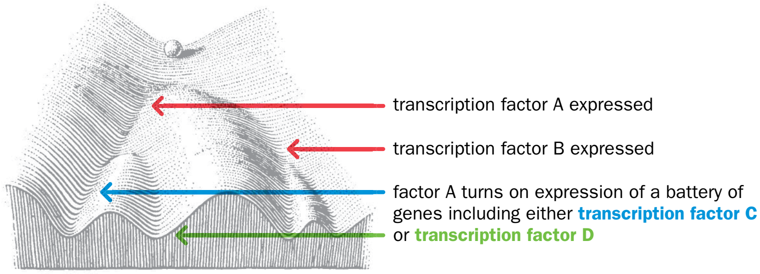

transcription factor A expressed
transcription factor B expressed
factor A turns on expression of a battery of genes including either transcription factor C or transcription factor D

Figure 10.5 Waddington's concept of an epigenetic landscape, interpreted in terms of transcription factors.

The main protein motifs responsible for sequence-specific DNA binding were described in Chapter 3 (see **Box 3.1**): helix-loop-helix, helix-turn-helix, leucine zipper, and zinc finger motifs. Binding sites for most transcription factors are quite short, and some variation is tolerated (**Figure 10.6**). Optimum binding sequences can be identified by systematically modifying oligonucleotides to find the one that is most strongly bound *in vitro*, but chromatin immunoprecipitation studies show that the binding sites actually occupied often differ from the optimum. Binding that is too strong could be just as deleterious to the cell as binding that is too weak—it could lead to cells overreacting to possibly spurious weak signals. Weaker binding allows more flexible responses. Although thousands of possible sites for any one factor are present across the genome, only a tiny minority are actually occupied, therefore DNA sequence cannot be the sole determinant of transcription factor binding. Binding is often combinatorial, so that groups of factors bind to nearby sequences, and the binding is stabilized by protein–protein interactions.

In addition to a DNA-binding domain, transcription factors contain activation domains that mediate their functional effect by recruiting effector proteins. These include co-activators and co-repressors—proteins or protein complexes that do not themselves bind DNA but are brought to regulatory sequences by protein–protein interactions with the DNA-bound proteins at the site.

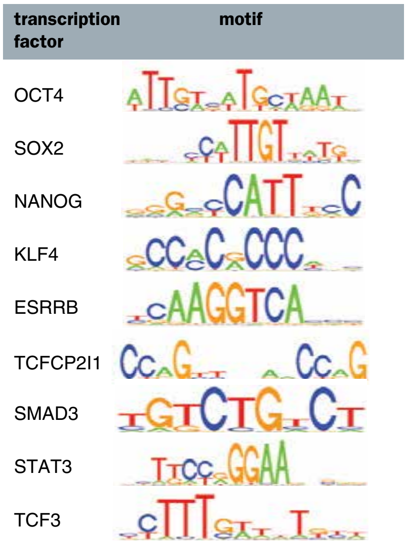


Figure 10.6 Transcription factor binding sites. Binding site preferences for selected transcription factors are shown using motif notation, where the height of a letter shows the frequency with which a particular nucleotide is found at each position. (Adapted from Hnisz D *et al*. [2013] *Cell* **155**:934–947; PMID 24119843. With permission from Elsevier.)

Insulators and the CTCF protein

Insulators are DNA sequences that block the interaction of promoters and enhancers that lie either side of the insulator. They mark the boundaries of the topologically associated domains (TADs) described above. Insulators act by binding CTCF (CCCTC binding factor) protein, and maybe other proteins. CTCF can dimerize when it is bound to different DNA sequences, thus mediating long-range chromatin looping. CTCF is also involved in positioning of nucleosomes, and in delimiting the boundaries of heterochromatin, which has a natural tendency to spread.

10.3 REGULATION BY DNA METHYLATION AND NONCODING RNAS

Methylation of DNA, along with modification of histones and positioning of nucleosomes, is one of the main epigenetic mechanisms operating across the genome. As mentioned in Section 1.4, the nucleotide bases in RNA are subject to a wide range of chemical modifications. DNA, however, normally suffers only one major modification, methylation of cytosine to produce 5-methylcytosine (5-meC; see **Figure 1.9**), although 6–7 adenines per million, averaged over the genome, are also methylated as N^6-methyladenine. 5-methylcytosine base-pairs with guanine in exactly the same way as unmethylated cytosine, so any genetic information encoded in the nucleotide sequence is unchanged. However, the methyl group, located in the major groove of the double helix, can attract methyl-binding proteins that affect gene expression, either directly or through binding partners. Hence DNA methyltransferases and methyl-DNA-binding proteins are important components of the regulatory landscape—for more detail, see the review by Schübeler (2015), PMID 25592537. Humans have three DNA methyltransferases, DNMT1, DNMT3A, and DNMT3B, plus two related proteins DNMT2 and DNMT3L (**Table 10.2**).

TABLE 10.2 HUMAN DNA METHYLTRANSFERASES

Enzyme	OMIM #	Major functions	Associated proteins
DNMT1	126375	Maintenance methyltransferase	PCNA (replication forks); histone methyltransferases; histone deacetylases; HP1 (heterochromatin); methyl-DNA-binding proteins
[DNMT2][a]	602478	Methylation of cytosine-38 in $tRNA^{Asp}$; no DNA-methylation activity	
DNMT3A	602769	*de novo* methyltransferase	Histone methyltransferases; histone deacetylases
DNMT3B	602900	*de novo* methyltransferase	Histone methyltransferases; histone deacetylases
DNMT3L[b]	606588	Binds to chromatin with unmethylated H3K4 and stimulates activity of DNMT3A/B	DNMT3A; DNMT3B; histone deacetylases

[a] DNMT2 turned out to have RNA rather than DNA as its substrate, but its structure is that of a DNA methyltransferase.

[b] DNMT3L is a co-factor rather than an active methyltransferase. See the main text for discussion of maintenance methylation. PCNA, proliferating cell nuclear antigen; HP1, heterochromatin protein 1.

There is no corresponding DNA demethylase to act as an eraser. Instead two mechanisms are available to demethylate DNA:

- Passive dilution—an original methylation signal can be diluted out if cells proliferate without methylating newly synthesized DNA;
- Oxidative demethylation—enzymes of the TET (ten-eleven translocation) family convert 5-meC into 5-hydroxymethylcytosine (5-hmC). The process can continue through 5-formylcytosine to 5-carboxycytosine, which is then excised from

the DNA by thymine-DNA glycosylase or methyl-CpG-binding domain protein 4 (MBD4; **Figure 10.7**). Humans have three TET-family enzymes; they use molecular oxygen together with α-ketoglutarate and ferrous iron to oxidize 5-meC. There is ongoing debate about how far 5-hmC is simply an intermediate in this process and how far it constitutes a separate epigenetic signal in DNA.

Techniques for studying DNA methylation are outlined in **Box 10.3**.

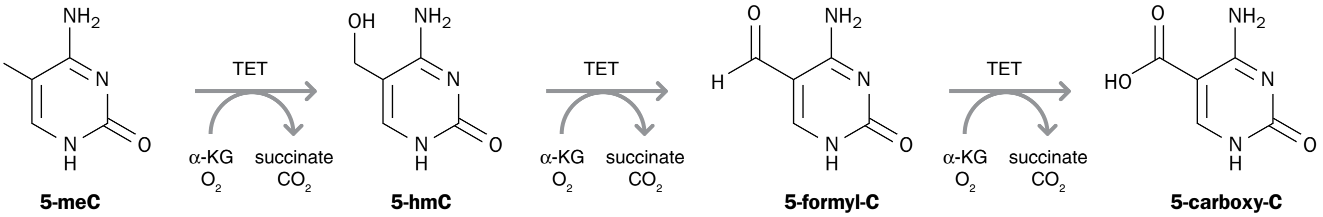


Figure 10.7 DNA demethylation. TET enzymes oxidize 5-methylcytosine (5-meC) in stages to 5-carboxycytosine (5-carboxy-C); this is excised from the DNA by thymine-DNA glycosylase to complete the removal of methylated cytosines. 5-hmC, 5-hydroxymethylcytosine; 5-formyl-C, 5-formylcytosine; α-KG, α-ketoglutarate.

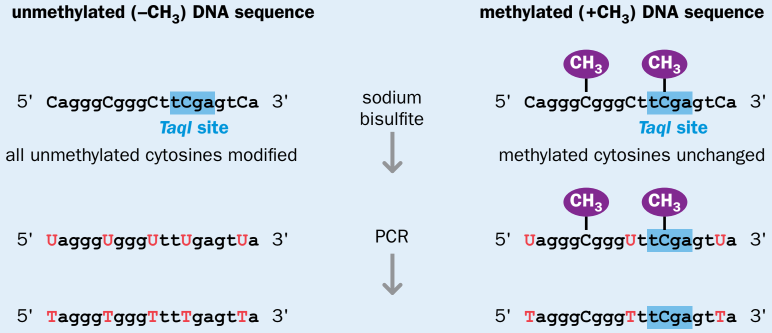
BOX 10.3 STUDYING DNA METHYLATION

Whatever method is used, it must be used on raw genomic DNA. Sequences to be studied cannot be first amplified by PCR as in most genetic tests because the PCR product would contain entirely normal, unmethylated, cytosine. The two main general methods are bisulfite sequencing and methyl-DNA immunoprecipitation (MeDIP). In addition, some next-generation sequencing systems may be able to detect 5-meC directly (see Section 6.5).

When single-stranded DNA is treated with sodium bisulfite ($NaHSO_3$) or metabisulfite ($Na_2S_2O_5$) under carefully controlled conditions, cytosine is deaminated to uracil but 5-methylcytosine (and 5-hydroxymethylcytosine) remain unchanged (**Figure 1**). Adaptor oligonucleotides with PCR primer sites are ligated to each end of the resulting fragments and the whole product is PCR-amplified. Uracil in the treated DNA is represented by thymine in the PCR product. Massively parallel sequencing and comparison with the normal (untreated) sequence shows which cytosines were methylated, at single-nucleotide resolution and across the whole genome. Specific sequences can be interrogated by PCR, using primers matching either the original or the bisulfite-converted sequence (methylation-sensitive PCR). Whole-genome bisulfite sequencing is expensive; a cheaper alternative is reduced representation bisulfite sequencing. Genomic DNA is first digested to completion with a restriction enzyme whose recognition sequence includes CG. Only size-selected small (for example 500–600 nt) fragments are used for the bisulfite sequencing. These will come predominantly from CpG islands. Only a small fraction of the whole genome is covered, but it should be a reproducible fraction, so the technique can be used in comparative studies.

MeDIP lacks the base-level resolution of bisulfite sequencing, but is cheaper and can be used to distinguish 5-meC and 5-hmC. Methylated DNA fragments are immunoprecipitated from fragmented genomic DNA using antibodies against 5-meC or 5-hmC, and then sequenced.

As an alternative to genome-wide analyses, specific sequences can be checked using the restriction enzymes *Msp*I and *Hpa*II. Both cut the same CCGG sequence but are sensitive to different methylation states. *Msp*I recognizes and cleaves sites containing unmodified, methylated, or hydroxymethylated cytosine equally well, but *Hpa*II cleaves only a completely unmodified site. If raw genomic DNA is digested with *Hpa*II, only sequences containing modified CCGG sites will remain intact and so be PCR-amplifiable. A variant of the technique can be used to distinguish methylated and hydroxymethylated sites. Treating the raw DNA with UDP-glucose and T4-β-glucosyltransferase glucosylates all hydroxymethylcytosines. Glucosylation prevents *Msp*I cutting, so comparison of glucosylated and non-glucosylated reactions identifies hydroxymethylated sites specifically.


unmethylated (–CH_3) DNA sequence
methylated (+CH_3) DNA sequence
CH_3
CH_3
5' CagggCgggCttCgagtCa 3'
sodium bisulfite
5' CagggCgggCttCgagtCa 3'
TaqI site
TaqI site
all unmethylated cytosines modified
methylated cytosines unchanged
CH_3
CH_3
5' UagggUgggUttUgagtUa 3'
PCR
5' UagggCgggUttCgagtUa 3'
5' TagggTgggTttTgagtTa 3'
5' TagggCgggTttCgagtTa 3'

Box 10.3 Figure 1 Modification of DNA by sodium bisulfite. Note the Taq1 restriction site in this sequence that is differentially affected by methylation, providing an additional method for checking the methylation status of that cytosine.

The significance of CpG sequences

Methylation is largely restricted to cytosines that lie immediately upstream of guanines in so-called CpG sequences (the p represents the phosphate linking the two nucleosides). CpG dinucleotides occur symmetrically in the double helix. Opposite 5′-CpG-3′ in one strand there will be 3′-GpC-5′ in the complementary strand, so that reading in the usual 5′ → 3′ direction, both strands carry CpG (**Figure 10.8**).

```
5' ...AGTCACGATCC... 3'
3' ...TCAGTGCTAGG... 5'
```

Figure 10.8 CpG sequences occur symmetrically in double-stranded DNA.

The significance of this is that it provides a mechanism for epigenetic memory. The DNMT1 methyltransferase specifically methylates cytosines in CpG sequences where the CpG in the complementary strand is already methylated (**Figure 10.9**). Thus when a cell divides, it reproduces the pattern of methylation from the mother cell in both daughter cells.

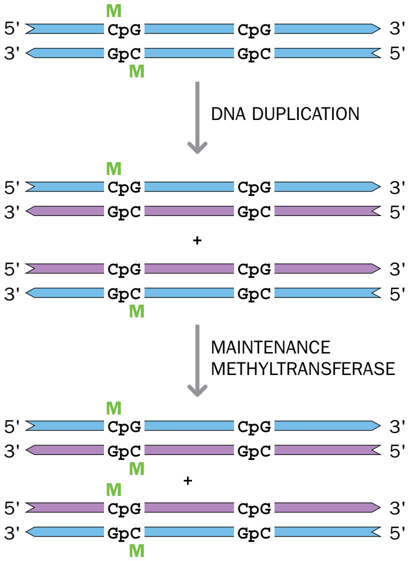


Figure 10.9 Maintenance methylation. When DNA containing a methylated CpG sequence is replicated, the CpG in the newly synthesized strand is initially unmethylated. The DNMT1 DNA methyltransferase specifically methylates CpG sequences that are paired with methylated CpG, thus preserving the pattern of methylation that was present in the DNA before replication.

This cannot be the only mechanism that can allow epigenetic marks to be preserved from mother to daughter cell. Epigenetic memory functions in flies, worms, and yeast, which do not methylate their DNA. There is some evidence for transmission of specific histone marks through DNA replication and mitosis. The most likely candidates are patterns of lysine methylation, which are known to be stable over hours or days in cells; acetylation marks, by contrast, often have half-lives of only minutes. It is known that Polycomb proteins (responsible for repressive histone marks) can remain bound to chromatin and DNA during DNA replication, at least *in vitro*. Histones from nucleosomes seem to be randomly distributed between parental and new DNA strands when the replication fork passes. Some proteins that recognize and bind to specifically modified histones also possess the ability to impose those same modifications on adjacent nucleosomes; this explains the tendency of some chromatin marks to spread, but could also suggest a mechanism for epigenetic memory independent of DNA methylation.

In some cell types, cytosines in other contexts may also be methylated (so-called CH, CHG, and CHH sequences, where H represents any base except G). This is particularly seen in embryonic stem cells and oocytes, but also in neurons, where CpA methylation builds up over time to equal the level of CpG methylation; it occurs at a lower level in other tissues. Note that non-CpG methylation is not symmetrical across the two strands of the double helix. Its significance is unclear, but non-CpG methylated sequences may bind methyl-DNA-binding proteins differently, and so carry a specific functional signal.

Methylation of DNA is normally a repressive signal. In plants and many invertebrates, CpG methylation is concentrated on repetitive sequences, including dispersed transposons and the satellite repeats characteristic of pericentric heterochromatin. DNA methylation at repeated sequences probably serves to repress transcription. This functions as a defense mechanism since unbridled transcription of transposons is a threat to the integrity of the genome. In mammals, CpG methylation is more pervasive, for example in the body of genes, especially in exons, and in intergenic sequences. Of the roughly 28 million CpGs in the human genome 60–90% are generally methylated. Less than 10% are located in **CpG islands**.

CpG islands are stretches of a few hundred base pairs of DNA where cytosines are unmethylated and, as a result, there has not been the depletion of CpG sequences over evolutionary time that is seen in the bulk of the genome (see **Box 9.1**). CpG islands are associated with genes. They are found at around 70% of promoters (high-CpG-density promoters), where they are normally unmethylated whether or not the associated gene is expressed. On occasion CpG islands can become methylated. This shuts down transcription, as observed on the inactive X chromosome or at imprinted genes (see below). Aberrant methylation of CpG islands at promoters, particularly those associated with tumor suppressor genes, is a general characteristic of cancer cells (this is discussed further in Chapter 19). In promoters without CpG islands, such CpGs as are present are normally methylated, without any apparent effect on transcription. Around 15% of CpG sequences show variable methylation that often correlates with tissue-specific functions. **Figure 10.10** shows examples of the distributions of CpG methylation in different tissues across one 340 kb region.

DNA methylation sometimes affects gene expression directly, in that some transcription factors such as YY1 fail to bind to methylated DNA. More generally the effect is mediated by proteins that contain a methyl-CpG-binding domain (MBD). One such protein, MeCP2, has been studied closely because loss of function causes Rett syndrome (OMIM #312750). This is a strange X-linked condition in which heterozygous girls develop normally for their first year but then regress in a very characteristic way. *MECP2* mutations are normally lethal in males. DNA methylation proceeds normally in patients with Rett syndrome but the signals are not read correctly. MeCP2 protein is normally very abundant in neurons where it seems to be particularly involved in repressive regulation of unusually long protein-coding genes. Interestingly, MeCP2 recognizes methylated

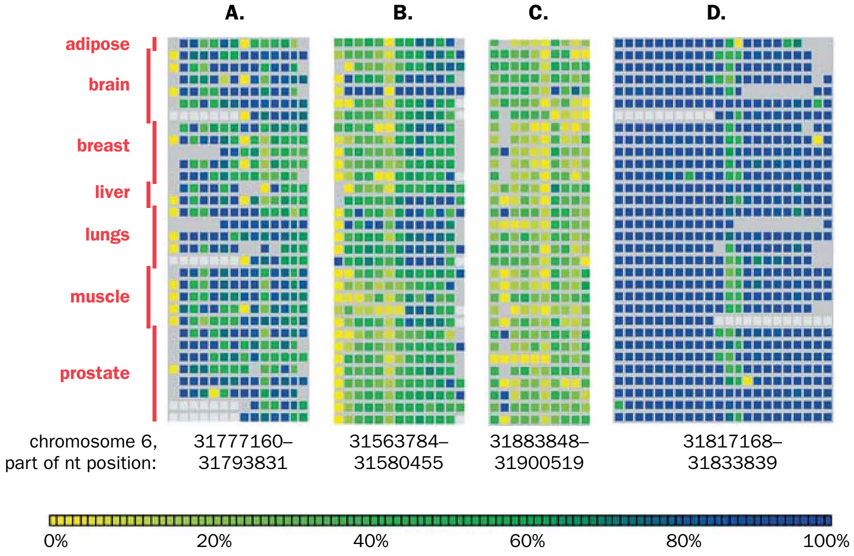


Figure 10.10 Patterns of CpG methylation may be person-specific, tissue-specific, or locus-specific. Each square represents a CpG sequence from a subregion (labeled A–D) of the major histocompatibility complex on chromosome 6, with the intervening nucleotides omitted. Squares are color-coded to indicate the extent of methylation (see scale at bottom). Rows show the methylation pattern seen in different tissue samples. Region A shows diverse patterns of methylation both between different tissues and between samples of the same tissue from different individuals, whereas region B shows more tissue-specific methylation. Region C is largely unmethylated, and region D is almost completely methylated. (From Hermann A *et al*. [2004] *Cell Mol Life Sci* **61**:2571–2587; PMID 15526163. With permission from Springer International Publishing AG.)

CpA but apparently not hydroxymethylcytosine, both of which accumulate in neurons over the timescale in which Rett symptoms appear. In a mouse model of Rett syndrome, restoring *Mecp2* function caused even established Rett-like symptoms to regress, holding out hope that the human syndrome might be treatable.

DNA methylation shows striking changes during embryonic development

A recently fertilized oocyte has methylation differences at paternal and maternal alleles of many genes, reflecting the very different methylation profiles of the sperm and oocyte. The methylation patterns then show striking changes during embryonic development, summarized in Figure 10.11. The erasure of parental epigenetic settings that occurs shortly after fertilization is substantial, resetting the epigenetic information originating from the previous generation to achieve the relatively blank and versatile pluripotent state necessary at the onset of development. However, it cannot be complete because at imprinted loci, genes retain an epigenetic memory of their parental origin (see below).

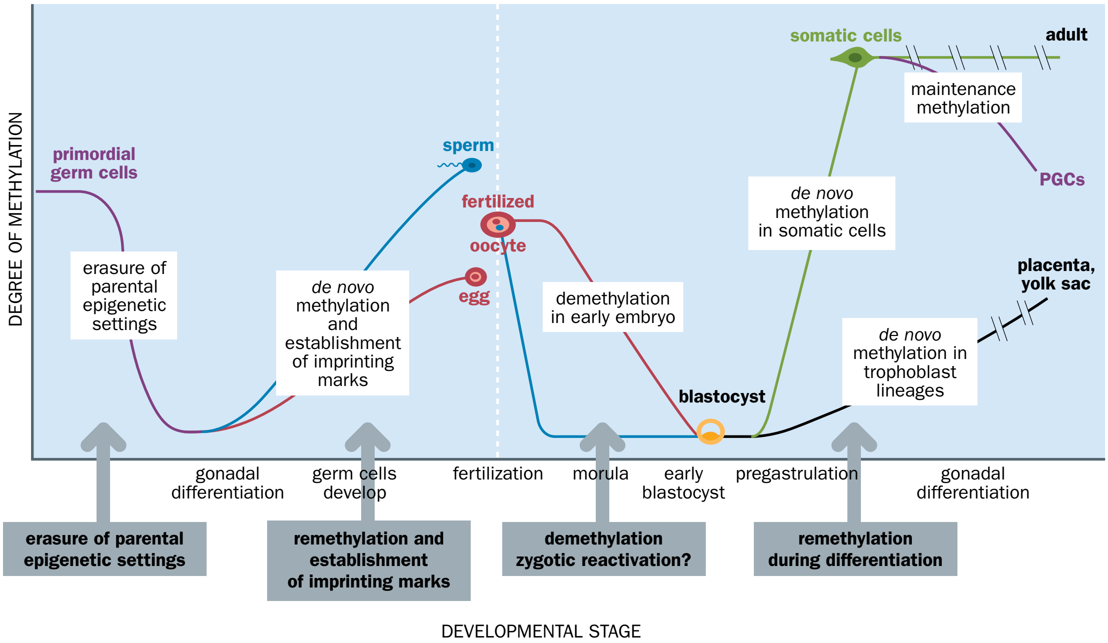

DEGREE OF METHYLATION
primordial germ cells
erasure of parental epigenetic settings
de novo methylation and establishment of imprinting marks
sperm
fertilized oocyte
egg
demethylation in early embryo
blastocyst
somatic cells
adult
maintenance methylation
PGCs
de novo methylation in somatic cells
placenta, yolk sac
de novo methylation in trophoblast lineages
gonadal differentiation
germ cells develop
fertilization
morula
early blastocyst
pregastrulation
gonadal differentiation
erasure of parental epigenetic settings
remethylation and establishment of imprinting marks
demethylation zygotic reactivation?
remethylation during differentiation
DEVELOPMENTAL STAGE

Figure 10.11 Changes in DNA methylation during mammalian development. Drastic and often tissue-specific changes in overall methylation accompany gametogenesis and early embryonic development. Note the breaks (slashes) in the developmental lines. PGCs, primordial germ cells.

The roles of RNA in regulation of gene expression

Table 9.6 detailed the many classes of noncoding RNA (ncRNA). Many of these have roles in gene regulation. Among the short ncRNAs, microRNAs are involved at the level of translation, as described below, while piRNAs and siRNAs have specialized roles in keeping transposons quiescent (see **Box 8.2** and Section 9.3). Human cells contain many thousands of different long (>200 nt) ncRNAs. ENSEMBL (November 2017) listed 14,720 lncRNA genes. lncRNAs have many diverse roles in gene regulation; **Table 10.3** shows some examples, and the paper by Ponting and colleagues (Ponting *et al*. [2009], PMID 19239885; see Further Reading) gives others. Some, like HOTTIP, HOTAIR, and FOXCUT, are *cis*-acting regulators of a nearby gene. They may activate or repress expression. lncRNA-COX2, and probably many others, forms *trans*-acting complexes that can affect expression of large numbers of genes. Some guide epigenetic regulators such as histone-modifying enzymes or DNA methyltransferases to specific sites in the genome. Many are antisense transcripts that overlap a protein-coding gene and inhibit its expression. They may act by recruiting repressive protein complexes (ANRIL; see **Table 10.3**), but often the simple act of transcription prevents transcription of the target gene (KCNQ1OT1, SNHG14). XIST is an example of an RNA that acts in a spatially defined part of the nucleus, in this case on the future inactive X chromosome (see below). A frequent role of noncoding RNAs may be to act as scaffolds for assembling multiprotein complexes, and to define substructure within the cell nucleus.

TABLE 10.3 EXAMPLES OF LONG NONCODING RNAS

Species	Comments
HOTTIP	3764 nt spliced and polyadenylated RNA. *cis*-acting transcriptional activator of *HOXA* gene. Works in *cis* by looping and requires close proximity to target genes. Brings H3K4 methyltransferase to its target gene, yielding a broad domain of H3K4me3 and transcription activation
FOXCUT	*cis*-acting RNA that stimulates expression of nearby *FOXC1* gene
HOTAIR	2158 nt *trans*-acting spliced and polyadenylated transcriptional repressor of *HOXD* genes. Scaffold for PRC2 and coREST repressive complexes
ANRIL (CDKN2B-AS1)	3834 nt spliced and polyadenylated antisense transcript that represses transcription of *CDKN2B* gene on 9p21. Recruits PRC1/2 to silence co-located genes
KCNQ1OT1 (LIT1)	Unspliced antisense transcript that represses transcription of *CDKN1C* gene on the opposite strand of an imprinted region at 11p15
SNHG14 (SNRPN)	Huge, spliced (460 kb, 148 exons) imprinted ncRNA, part of which is antisense to the *UBE3A* gene and represses it
lncRNA-COX2	*trans*-acting spliced RNA up-regulated by Toll-like receptors. It binds to heterogeneous ribonucleoprotein A/B and modulates expression of many immune response genes across the genome
H19*	2700 nt spliced RNA, which includes the miR675 microRNA sequence; imprinted
XIST	19 kb RNA responsible for X-inactivation. XIST selectively coats the inactive X in XX cells
XACT	250 kb single or multiple transcript that selectively coats the active X in human embryonic stem cells

* Note that although H19 is classed as a noncoding RNA, it includes code for a microRNA, and many of its physiological actions may be due to the miRNA. PRC1/2, Polycomb repressive complexes 1 and 2.

Interactions among epigenetic mechanisms

As mentioned in the introduction to this chapter, the different processes involved in gene regulation are not independent linear systems, but form interacting networks. Methyl-DNA-binding proteins can recruit other proteins associated with repressive structures, such as histone methyltransferases or histone deacetylases, while methylated histones can recruit DNA methyltransferases (Figure 10.12). Transcription factor binding inhibits DNA methylation. Some long noncoding RNAs guide epigenetic regulators such as histone-modifying enzymes or DNA methyltransferases to specific sites in the genome.

This raises the question of which epigenetic changes are the primary causes, and which are downstream effectors, or just consequences, of changes in gene expression. For example, the methylation changes shown in Figure 10.11 are very dramatic, but it is debatable how far they are causative of the events they accompany, rather than being downstream results of the mechanisms that cause differentiation. DNA methylation is certainly required for vertebrate development: animals deficient in DNA methyltransferase activity die at various stages of development, and mice that are specifically unable to methylate sperm DNA are infertile. Similarly, although pluripotent embryonic stem cells derived from blastocyst-stage embryos can grow normally even in the complete absence of DNA methylation, these cells immediately undergo apoptosis when stimulated to differentiate. However, although DNA methylation is a feature of vertebrates, flowering plants, and some fungi, as mentioned above it is virtually absent from many of the best-studied model organisms, including yeast, *C. elegans*, and *Drosophila*—yet the mechanisms of differentiation seem fundamentally similar in all these organisms.

This introduces something of a chicken-and-egg question—what is the primary determinant? The review by Smith & Meissner (2013) (PMID 23400093; see Further Reading) underscores the complexity of the interactions involved. Some evidence suggests that binding of transcription factors may be a primary event, particularly of so-called pioneer factors that seem able to bind to unmodified chromatin. This then triggers a self-reinforcing process. It has also been reported that housekeeping genes, which are relatively independent of external signals, rely more on histone modification, while signal-responsive genes depend more on transcription factor binding, which can act rapidly by relocating pre-existing transcription factors from the cytoplasm to the nucleus (see Figures 3.3 and 3.4).

The new technologies of genome editing (Chapter 8) will allow answers to this question. The CRISPR/Cas system can be modified by fusing a defective Cas9 nuclease to one of the epigenetic writers or erasers. The guide RNA will then direct the modification to one specific location, allowing just the effect of that one change to be investigated. It would not be surprising if it turned out that a small minority of epigenetic modifications were causal (instructive), while the majority were downstream consequences of the feedback mechanisms that reinforce the initial events to produce stable states.

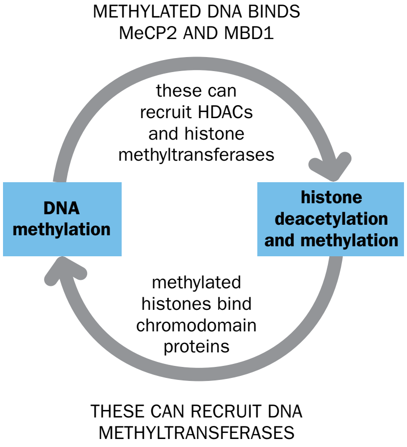

METHYLATED DNA BINDS MeCP2 AND MBD1
these can recruit HDACs and histone methyltransferases
DNA methylation
histone deacetylation and methylation
methylated histones bind chromodomain proteins
THESE CAN RECRUIT DNA METHYLTRANSFERASES

Figure 10.12 Mutually reinforcing histone and DNA methylation in inactive chromatin. Methylation of DNA attracts proteins that modify associated histone proteins. Histone deacetylation and methylation attract proteins that methylate associated DNA. In addition to the effects shown here, some histone methyltransferases and some long noncoding RNAs can interact directly with DNA methyltransferases. HDACs, histone deacetylases; MeCP2/MBD1, methylcytosine-binding proteins.

10.4 X-INACTIVATION, IMPRINTING, AND EPIGENETIC MEMORY

X-inactivation: an epigenetic change that is heritable from cell to daughter cell, but not from parent to child

In humans, females normally have two X chromosomes (46,XX) while males have one X and one Y (46,XY); see Figure 2.10. The fact that normal healthy people can have different numbers of X and Y chromosomes requires explanation. Animals, including humans, do not readily tolerate having wrong numbers of chromosomes. Chromosomal aneuploidies (extra or missing chromosomes, as in Down syndrome—see Chapter 15) have severe, usually lethal, consequences. Nevertheless, in organisms with an XX/XY sex determination system, males and females must be able to develop normally despite having different sex chromosomes. For the human Y chromosome, the solution is to carry very few genes. Some of these have counterparts on the X chromosome so that both XX and XY individuals have two copies, while many are related to male sexual function and so are dispensable in females (see Section 13.3).

The human X chromosome, on the other hand, carries many essential genes. Conceptuses that lack an X chromosome cannot survive, while the many X-linked diseases show the importance of individual X-linked genes. Different organisms solve the problem of coping with either XX or XY chromosome constitutions in different ways. In male *Drosophila* flies, genes on the single X chromosome are transcribed at double the rate of those on the X chromosomes of females. Mammals, including humans, take a different approach: they use **X-inactivation** (sometimes called **lyonization** after its discoverer, Dr Mary Lyon).

Early in embryogenesis each cell somehow counts its number of X chromosomes, and then permanently inactivates all X chromosomes except one in each somatic cell. At very early stages in development, both X chromosomes are active, but X-inactivation is initiated at the late blastula stage as cells begin to differentiate. Inactivated X chromosomes are still physically present, and on a standard mitotic karyotype they look entirely normal, but the inactive X fails to decondense after mitosis. It remains condensed throughout the cell cycle and most genes on the chromosome are permanently silenced in somatic cells. In interphase cells the inactive X may sometimes be seen under the microscope as a **Barr body** or **sex chromatin** (**Figure 10.13**). Regardless of the karyotype, each somatic cell retains a single active X:

- XY males keep their single X active (no Barr body);
- XX females inactivate one X in each cell (one Barr body);
- Females with Turner syndrome (45,X) do not inactivate their X (no Barr body);
- Males with Klinefelter syndrome (47,XXY) inactivate one X (one Barr body);
- 47,XXX females inactivate two X chromosomes (two Barr bodies).

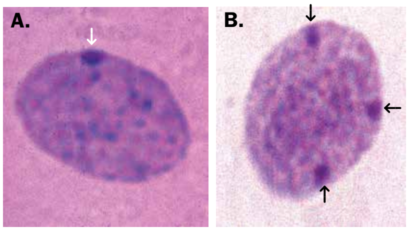


Figure 10.13 Barr bodies. (**A**) A cell from a 46,XX female has one inactivated X chromosome and shows a single Barr body (arrow). (**B**) A cell from a rare 49,XXXXY male has three inactivated X chromosomes and shows three Barr bodies. (Courtesy of Malcolm Ferguson-Smith, University of Cambridge.)

A 46,XX cell may inactivate the maternal or the paternal X chromosome—it is a random choice, made independently by each cell—but whichever is chosen for inactivation, that same one is inactivated in all daughter cells (**Figure 10.14A**). The body of an adult female is thus a mosaic of cell clones, each clone retaining the pattern of X-inactivation that was established in its progenitor cell early in embryonic life. This can have implications for women who are heterozygous for an X-linked pathogenic loss-of-function mutation. A woman carrier of hemophilia A, for example, has one X chromosome with an intact Factor VIII gene and one with a nonfunctional copy. On average around half her cells will have the intact copy active. Factor VIII is a circulating protein and there is an averaging effect: she will have around half the normal level of the clotting factor, but that is normally sufficient for her to avoid clinical consequences. On the other hand, a woman heterozygous for the cell-autonomous condition X-linked ectodermal dysplasia (OMIM #305100) has some patches of skin with normal sweat glands, clonal progeny of an embryonic cell that inactivated the mutation-bearing X, and patches lacking sweat glands, derived from a cell that inactivated the normal X. **Figure 10.14B** shows a somewhat analogous example from the calico cat.

X-linked severe combined immunodeficiency (X-SCID; OMIM #300400) shows another effect. X-SCID is caused by loss of function of the *IL2RG* gene. This X-linked gene is required for development of both B and T lymphocytes. Affected males have a lethal immunodeficiency. In heterozygous women, the descendants of any precursor cell that inactivated the normal X are unable to give rise to lymphocytes, and so all the lymphocytes she does have are progeny of cells that inactivated the mutation-bearing X. Thus X-inactivation in lymphocytes is 100% skewed, although other tissues show the normal random X-inactivation. Highly skewed X-inactivation in any tissue can be a pointer to heterozygosity for an X-linked condition. In Section 15.2 we discuss the way X-inactivation can affect a woman carrying a chromosomal translocation between one X chromosome and an autosome.

X-inactivation is an epigenetic process: the DNA sequence is unaltered by inactivation. Inactivation is stable through mitosis but not across the generations. During oogenesis the inactive X is reactivated and all memory of X-inactivation status is erased. The previously inactive X has the same chance as the previously active X of being passed on to any child, and in a daughter it has the same chance as the paternal X of being inactivated in any particular cell.

Initiating X-inactivation: the role of XIST

Inactivation is initiated at the 1 Mb X-inactivation center (*XIC*) at Xq13. The earliest observed event in XX cells is a transient pairing of the two *XIC* sequences. This is probably the mechanism by which the X chromosomes are counted, because counting is disrupted if the *XIC* sequences are deleted, duplicated, or translocated. *XIC* encodes a large noncoding RNA, XIST (X-inactivation-specific transcript; see **Table 10.3**), which is expressed only from the inactive X chromosome. The primary transcript undergoes

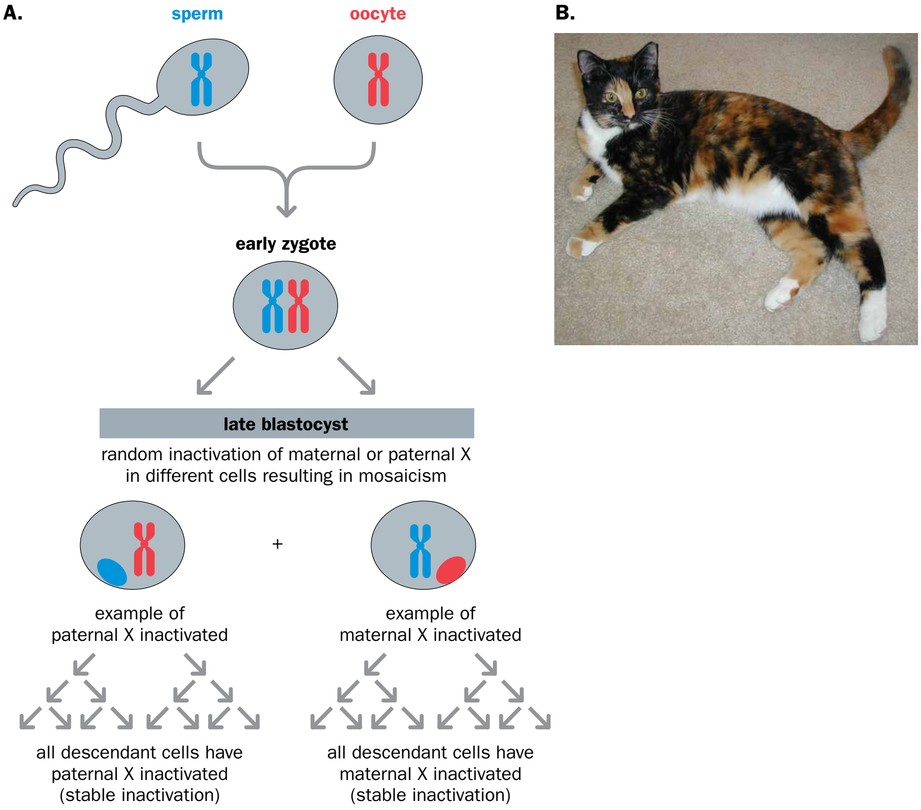

A.
sperm
oocyte
early zygote
late blastocyst
random inactivation of maternal or paternal X in different cells resulting in mosaicism
+
example of paternal X inactivated
example of maternal X inactivated
all descendant cells have paternal X inactivated (stable inactivation)
all descendant cells have maternal X inactivated (stable inactivation)
B.

Figure 10.14 X-inactivation. (**A**) In human embryos a randomly chosen X chromosome is inactivated in each cell of a 46,XX embryo, but once made, the choice is transmitted through all subsequent rounds of mitosis. (**B**) The calico (tortoiseshell and white) cat is heterozygous at an X-linked coat color locus. One allele specifies black coat color, the other orange. The different color patches reflect clones in which different X chromosomes are inactivated. The white patches are the result of an unrelated coat color gene. Calico cats are always female, apart from occasional XXY males. (Adapted from Migeon BR [1994] *Trends Genet* **10**:230–235; PMID 8091502. With permission from Elsevier.)

splicing and polyadenylation to generate a 19 kb mature noncoding RNA. XIST is required to establish X-inactivation, but not to maintain it: in differentiated cells that have already undergone X-inactivation, loss of XIST does not cause reactivation.

Most detailed studies of the mechanism have been performed in the mouse, looking at the onset of random X-inactivation as pluripotent embryonic stem cells (ESCs) start to differentiate. This has revealed a complex set of noncoding RNAs that regulate Xist expression. However, the timing and initial mechanism of X-inactivation are very different in mice and humans (and, confusingly, papers do not always make clear whether they are talking about mice or humans). Little of the organization of the mouse *Xic* is conserved in humans. In the mouse, expression of Xist from the active X chromosome is silenced by an antisense transcript (*Tsix*) that alters the chromatin configuration at the Xist locus. There is a human *TSIX* gene, but compared to mouse *Tsix* it is truncated at the 5′ end so that it does not cover the *XIST* promoter and does not have the CpG island that is essential for X-inactivation function in the mouse. The papers by Chang & Brown (2010) and by Migeon (2016) (PMID 20211024 and 26805440, respectively; see Further Reading) give more detail.

One way or another, in both species the XIST/Xist RNA comes to coat the whole inactive X. Spreading depends on physical continuity of the chromosome. On an X chromosome that is split in two by an X-autosome translocation (see Chapter 15), inactivation is limited to the segment that includes *XIC*. It cannot jump to the detached portion. Exactly how this causes silencing is not clear. Numerous interacting partners of the XIST RNA have been identified but their roles, if any, are largely undefined (reviewed by Moindrot & Brockdorff [2016], PMID 26816113; see Further Reading). Directly or indirectly, XIST recruits repressive proteins including the PRC1 and PRC2 Polycomb complexes that organize the chromatin into a closed, transcriptionally inactive conformation. The chromatin of the inactive X comes to carry modifications typical of heterochromatin (H3K9me3, H3K27me3, unmethylated H3K4). In addition, many nucleosomes carry macro-H2A, a variant of histone H2A. A major stabilizer of inactivation is methylation of the normally unmethylated CpG islands at promoters of genes. However, this cannot be the sole force maintaining inactivation because around 40% of the genes have low-CpG promoters that lack the islands.

The action of the maintenance DNA methyltransferase DNMT1, perhaps together with some of the histone modifications, ensures that whichever X (maternal or paternal) is initially inactivated in a cell of the inner cell mass, that same X is inactivated in all daughter cells derived from it. As described above, females are thus mosaic for clones of cells expressing either the maternal or the paternal X chromosome.

Escaping X-inactivation

X-inactivation is not a blanket inactivation of the entire chromosome. The two pseudoautosomal regions (detailed in Section 13.3) escape inactivation, but even outside these regions X-inactivation is patchy. Around 15% of human X-linked genes escape inactivation, at least in some tissues and some individuals. X-inactivation is tighter in the mouse, where only 3–6% of genes consistently escape inactivation in cell lines. In one study, different hybrid cells containing independent copies of a human inactive X were used to investigate the transcription of 612 X-linked genes: 458 of the genes were inactivated in all or most of the cell lines, but 94 were expressed. The remaining 60 genes showed a variable pattern of expression in different cell lines (**Figure 10.15**). Despite the general requirement for physical continuity, the spreading X-inactivation is evidently able to jump over these escape genes and continue its progress. Some of the genes that escape inactivation are genes that have a functional counterpart on the Y chromosome, for which there would be no need for dosage compensation. Many, however, have no such counterpart.

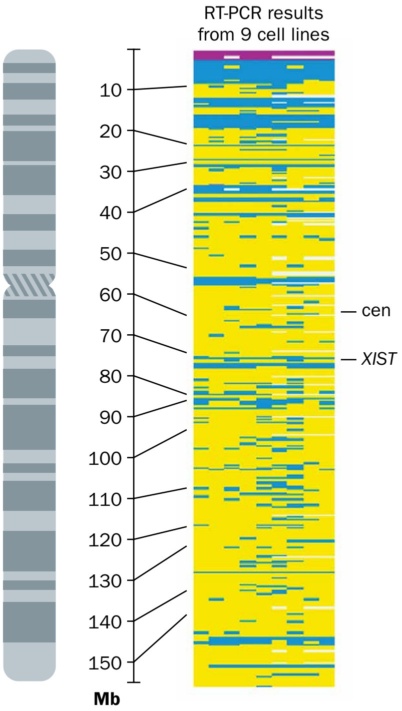


Figure 10.15 Genes that escape X-inactivation. Columns in the rectangle show results of systematic reverse transcription-PCR tests for expression of X-linked genes in nine independent somatic cell hybrids. Each hybrid contained a single inactive human X chromosome. Blue bars identify genes that were expressed from the inactive X chromosome in a particular hybrid, and yellow bars mark genes that were not expressed. Many genes, scattered all along the X chromosome, escape inactivation in one or more of the hybrids. Some genes, such as *XIST,* escape inactivation in all nine hybrids, whereas others show a more patchy pattern of inactivation. cen, centromere. (Reprinted from Carrel L & Willard HF [2005] *Nature* **434**:400–404; PMID 15772666. With permission from Springer Nature. Copyright © 2005.)

At imprinted loci, expression depends on the parental origin

The whole of Mendelian pedigree interpretation is based on the premise that the parental origin of a gene is irrelevant: a heterozygous person is the same, regardless whether the mutated allele came from their father or their mother. However, for about 100 human genes this is not true. These genes retain a memory—an imprint—of their parental origin. The imprint may persist through all the mitoses that generate the whole adult body (although with some genes only certain tissues retain the imprint). In all cases the imprint must be erased in the germ line and replaced by one appropriate to the sex of the person. A man may receive a gene with a maternal imprint from his mother, but if he passes it on to his child it must then carry a paternal imprint (**Figure 10.16**). Thus imprinting is a classic reversible epigenetic process.

Early evidence that parental genomes are not wholly equivalent came from the observation that zygotes with two maternal or two paternal genomes develop abnormally, as ovarian teratomas or hydatidiform moles, respectively. Hydatidiform moles are abnormal conceptuses that lack an embryo and consist of just hypertrophic extra-embryonic membranes; for teratomas, see Section 4.2. In mice, chromosomal manipulations make it possible to produce animals that have the correct number of chromosomes, but where both members of one chosen pair come from the same parent (**uniparental disomy**, UPD). Systematic exploration showed that for some chromosomes, although not others, UPD is pathogenic. The abnormal phenotypes are sometimes complementary for different parental origins; for example, overgrowth in paternal UPD and growth retardation in maternal UPD. For some chromosomes, UPD is lethal. These effects are caused by UPD for small numbers of specific genes on the relevant chromosome, rather than by the whole chromosome. Further work has identified about 100 imprinted loci on 11 of the mouse chromosomes.

In humans, a similar systematic investigation is not possible, but genotyping occasionally reveals UPD in patients or in healthy individuals, thus highlighting chromosomal regions where UPD matters, and others where it does not. The mechanism that

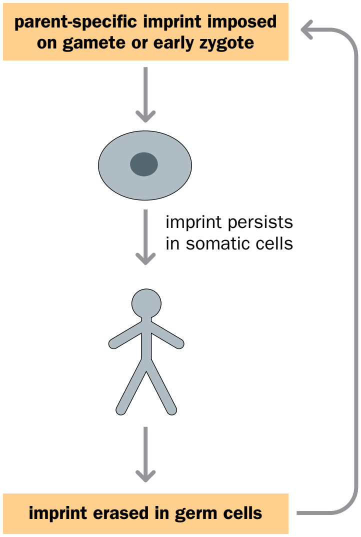

parent-specific imprint imposed
on gamete or early zygote
imprint persists
in somatic cells
imprint erased in germ cells

Figure 10.16 Imprinting is epigenetic and reversible. A man might receive a gene carrying a maternal imprint from his mother. The imprint persists in his somatic cells, but if he passes that gene on to a child, it will then carry a paternal imprint.

produces most cases of human UPD, trisomy rescue, is described in Section 15.2. Some microdeletion or microduplication syndromes show parent-of-origin effects, for example Prader–Willi and Angelman syndromes with a microdeletion of 15q11q13, and Beckwith–Wiedemann syndrome (BWS) with various abnormalities of 11p15 (see below). Many of the mouse imprinted genes are now known also to be imprinted in humans, although the correspondence is not perfect. The Otago University catalog of parent-of-origin effects (http://igc.otago.ac.nz/home.html) gives a comprehensive list of definitely or possibly imprinted genes in humans and other species.

To confirm imprinting of a gene, it is necessary to identify an individual who is heterozygous for a sequence variant present in the mature mRNA. Messenger RNA from different tissues can then be checked for monoallelic or biallelic expression, and the origin of each allele can be determined by typing the parents. Random monoallelic expression is surprisingly common, so for a claim of imprinting to be convincing it should be based on consistent data from many independent samples. In many cases imprinting is partial, with one parental allele expressed at a higher level than the other; often it is confined to certain tissues or to certain stages of development. In humans, imprinted genes have been identified on chromosomes 6q24–q26, 7q21–q22, 7q32 (and maybe 7p12), 11p13, 11p15, 14q32, 15q11–q13, 19q13, and 20q13 (see **Box 15.1** for explanation of cytogenetic nomenclature).

Mechanisms underlying imprinting

Examining imprinted regions, one finds short sequences where the DNA is differentially methylated on the paternal and maternal chromosomes, and often noncoding antisense transcripts (**Table 10.4**). The imprint is not a special mark present in gametes; rather it is a failure to remove methylation from specific sequences in the wave of genome-wide demethylation in the early zygote (**Figure 10.11**). In most cases the signal is methylation of a gene-associated CpG on the maternal chromosome; the few examples of paternal-specific methylation affect CpGs in intergenic DNA.

TABLE 10.4 EXAMPLES OF HUMAN IMPRINTED CLUSTERS SHOWING GENES AND NONCODING RNAS

Chromosome	Protein-coding genes	ncRNAs	Comments
6q25	*IGF2R* *SLC22A2* *SLC22A3*	AIRN	Imprinting strong in mice but variable among humans
11p15	*IGF2* *ASCL2* *TRPM5* *KCNQ1* *CDKN1C* *SLC22A18* *PHLDA2*	H19 KCNQ1OT1	Complex gene-rich region with two adjacent, separately imprinted domains. The *IGF2*/H19 domain is growth-promoting, the *KCNQ1* domain includes the growth-suppressing *CDKN1C* gene. The large *KCNQ1* gene is imprinted in several tissues but biallelically expressed in heart (see **Figures 10.17** and **10.18A**)
14q32	*DLK1*	MEG3/8 (GTL2)	MEG3/8 is a long ncRNA with many splice variants. Like the H19 RNA, it contains miRNA genes
15q11	*UBE3A* *NDN* *MAGEL2* *MKRN3(ZNF127)*	SNHG14	*UBE3A* imprinted only in brain. Complex alternative promoters and splicing of the 460 kb SNHG14 antisense transcript; snoRNAs in introns (see **Figure 10.18B**)
20q13	*GNAS* *XLAS* *NESP55*	GNASAS1 (NESPAS)	*GNAS* biallelic except in a few tissues. Coding transcripts have alternative promoters and first exons, but share downstream exons

Maternally expressed genes are shown in red, paternally expressed in blue, and biallelically expressed in black. Genes are often imprinted only in certain tissues, and the ncRNAs often show many alternative isoforms.

One mechanism by which differential methylation can affect gene expression is illustrated by the *IGF2*/*H19* cluster. As previously mentioned (see **Table 10.3**), H19 is a noncoding RNA that also includes the gene for microRNA miR675. IGF2 (insulin-like growth factor 2) is an important fetal growth factor. The basic mechanism influencing gene expression is competition for an enhancer, controlled by a differentially methylated imprinting control region (ICR; **Figure 10.17**). The *DLK1*/*MEG3* cluster at 14q32 functions in a similar way.

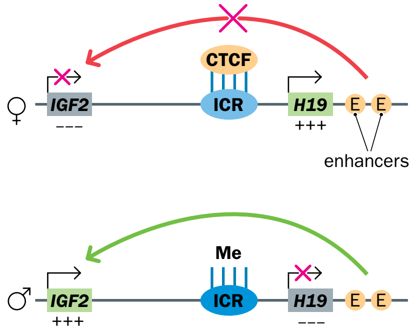


Figure 10.17 Paternal methylation of an intergenic insulator governs competition for an enhancer. Binding of the CTCF insulator protein to a differentially methylated region determines the outcome of the competition. On the maternal chromosome, the imprinting control region (ICR) is unmethylated, allowing it to bind CTCF (beige oval). This prevents the *IGF2* gene from accessing the enhancers, and so the enhancers drive *H19* expression. On the paternal chromosome, methylation of the imprinting control region prevents binding of CTCF, thereby allowing *IGF2* to outcompete *H19* for access to the enhancers. (Adapted from Wallace JA & Felsenfeld G [2007] *Curr Opin Genet Dev* **17**:400–407; PMID 17913488. With permission from Elsevier.)

An alternative mechanism is exemplified by the second imprinted domain on 11p15 and by the Prader–Willi/Angelman imprinted region on 15q11 (**Figure 10.18**). Again there is differential methylation, but here maternal-specific methylation of a CpG island prevents expression of a long noncoding antisense RNA. The paternal-specific RNA shuts off expression of nearby genes. At least for some of those genes, the effect seems to be due simply to the fact of transcription of the noncoding RNA, not to any property of the transcript. Probably there are other interactions at work as well, because in both cases there is imprinted expression of neighboring genes that do not overlap the antisense RNA and there is a great deal of alternative splicing.

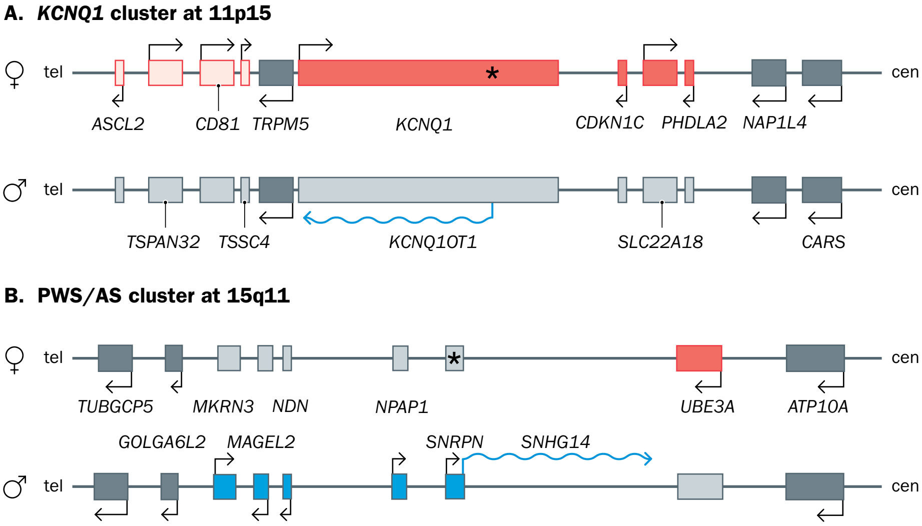


Figure 10.18 Maternally methylated promoters prevent transcription of long noncoding antisense RNAs. Maternally expressed genes are shown in red, paternally expressed in blue, and biallelically expressed in dark gray (there are some reports that *ATP10A* is only maternally expressed). Nonexpressed genes are shown in pale gray. Colored boxes with pink infill are imprinted only in extra-embryonic tissue. Blue wavy arrows show long noncoding RNAs. Asterisks show the differentially methylated imprinting control centers. tel, telomere; cen, centromere. (A, adapted from Pauler FM *et al.* [2012] *Curr Opin Genet Dev* **22**:283–289; PMID 22386265. With permission from Elsevier.)

Clinically, abnormalities in these regions manifest as developmental syndromes with parent-of-origin effects. On 15q11, lack of a maternal *UBE3A* product causes Angelman syndrome (OMIM #105830). The lack can be due to a microdeletion, paternal UPD, or a point mutation in *UBE3A*. Lack of the paternal SNHG14 RNA because of a microdeletion or maternal UPD causes Prader–Willi syndrome (OMIM #176270). The immediate cause is probably lack of the SNORD116 snoRNA cluster encoded in an intron of the SNHG14 RNA. With both syndromes, cases due to microdeletions and cases due to UPD are indistinguishable, suggesting that overexpression of the relevant gene has no adverse effects. This is different from the situation on 11p15. Overexpression of *IGF2* or underexpression of *CDKN1C* cause BWS (OMIM #130650), an overgrowth condition, while the reciprocal underexpression of *IGF2* or overexpression of *CDKN1C* cause Silver–Russell syndrome (SRS, OMIM #180860), in which there is growth retardation. A confusing variety of molecular events including UPD, deletions, duplications, or defective methylation can cause these effects. With all these syndromes we have given a rather simplified description of causes here; interested readers should consult OMIM for more details and references.

The regions described here show features common to several imprinted regions:

- Imprinting is controlled by one or more small regions where the DNA is differentially methylated on the maternal and paternal chromosomes;
- Imprinted regions are often complex, with clusters of genes, some paternally imprinted and others maternally imprinted. There may also be nonimprinted (biallelically expressed) genes in the cluster. Imprinting of some genes can be tissue-specific;
- There are often overlapping and oppositely imprinted sense and antisense transcripts, only one of which can be expressed at any one time. This points to some sort of flip-flop mechanism, in which transcription from one DNA strand prevents transcription from the opposite strand. The actual antisense transcript often seems to have no function of its own; what matters is the fact of transcription.

When a pathogenic variant in an imprinted gene is compatible with survival and fertility, the resulting pedigrees show unusual features (Figure 10.19). The variant may be inherited from a parent of either sex, and may affect persons of either sex, but the resulting phenotype will be apparent only when it is inherited from one sex of parent. So, if an imprinted gene is expressed only from the paternal chromosome, persons inheriting a pathogenic variant of the gene would show the resulting phenotype only if they inherited it from their father. A man who inherited it from his mother would be phenotypically normal, but would be at risk of having affected children.

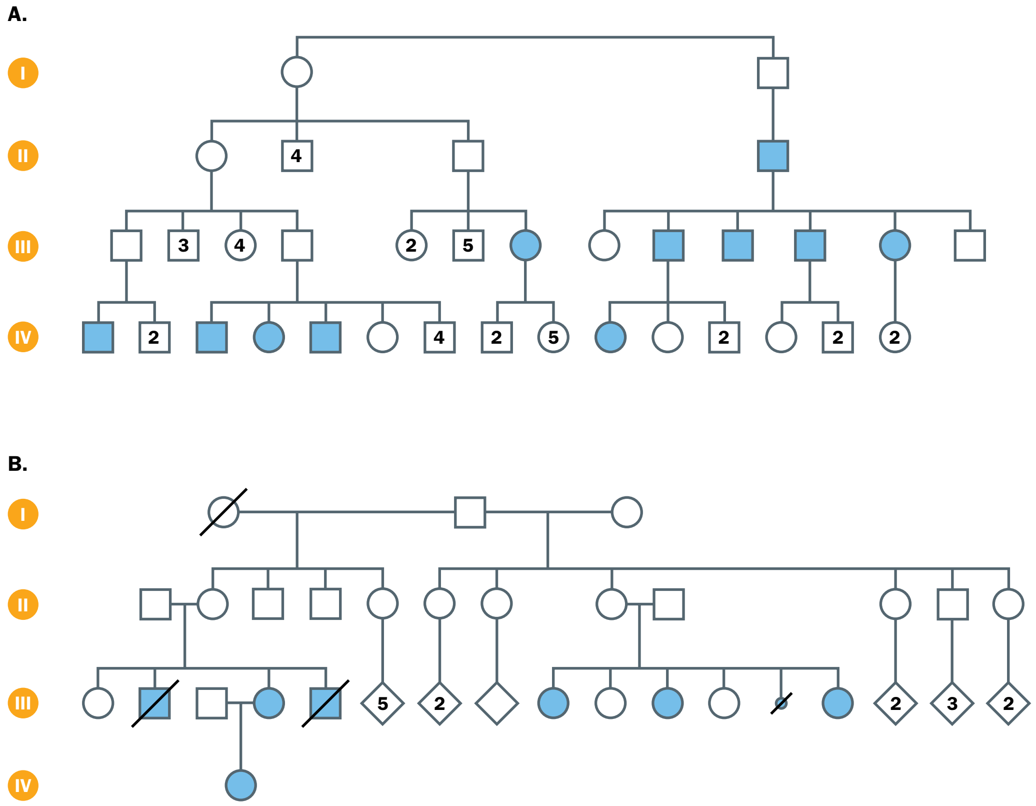


Figure 10.19 Pedigrees of conditions with imprinted gene expression. (**A**) In this family, autosomal dominant glomus tumors (OMIM #168000) manifest only when the gene is inherited from the father. (**B**) In this family, autosomal dominant Beckwith-Wiedemann syndrome (OMIM #130650) manifests only when the gene is inherited from the mother. (A, family reported in Heutink P *et al.* [1992] *Hum Mol Genet* **1**:7–10; PMID 1301144. With permission from Oxford University Press; B, family reported in Viljoen D & Ramesar R [1992] *J Med Genet* **29**:221–225; PMID 1583639. With permission from the BMJ Publishing Group Ltd.)

One might reasonably ask what is the function of such complicated mechanisms? One popular theory is based on a conflict of evolutionary interest between fathers and mothers. Selfish gene theory suggests that paternal genes might be best propagated by ensuring that offspring are born as robust as possible, even at the expense of the mother—a man can father children by many different women. Maternal genes, in contrast, are best propagated if the mother remains capable of further pregnancies. So, the theory goes, paternal genes program the fetus to extract nutrients at the greatest possible rate from the mother through the placenta, whereas maternal genes act to limit the depredations of a parasitic fetus. This does fit many imprinted loci, for example those at 11p15 described above, or those responsible for the abnormal development of conceptuses with two paternal or two maternal genomes. Hydatidiform moles, with two paternal genomes, are a mass of extra-embryonic membranes with no embryo, while ovarian teratomas, with two maternal genomes, are a disordered mass of fetal tissues with no membranes. It is less clear why imprinting should often be variable between individuals or tissues, or only involve mild allelic imbalances. Perhaps such variability suggests that one should not always try too hard to find a clear function in every case. Gene expression in single cells is often rather chaotic, with random monoallelic expression; maybe some of the weaker or more variable imprinting effects reflect a similar lack of tight control. It is amazing that cells work at all; natural selection ensures that they work well enough, but that does not require perfection and absolute precision in every detail.

Transgenerational epigenetic memory is a controversial and poorly understood subject

Epigenetic modifications are remembered through mitosis, by definition, but not normally through meiosis—consider X-inactivation for example. Reported exceptions need careful analysis: there are many possible reasons for parent-offspring resemblances. No fancy mechanism is required to explain why ten generations of a family all speak Chinese. Intrauterine environment can be important. Rats nurtured by a stressed mother are more likely themselves to be stressed because of intrauterine hormonal effects. Epigenetic marks made by a fetus in response to its environment affect metabolism and health in later life so, for example, underweight babies born to starved mothers after the 1944–1945 Dutch "hunger winter" developed into adults with increased adiposity. Given that a fetus already has the primordial germ cells that will produce the second generation, even maternal effects on grandchildren's health have many possible interpretations.

Paternal effects are more interesting. For example, in the Överkalix region of northern Sweden, the risk of cardiovascular and diabetes-related death of individuals could be related to increased food supply during the prepubertal growth period of their grandfathers. These and similar studies are reviewed by Pembrey *et al*. (2014) (PMID 25062846; see Further Reading). Mice provide a tractable experimental system for investigating such effects, particularly since the involvement of the fathers can be limited to providing sperm for *in vitro* fertilization. For example, Chen and colleagues (2016) (PMID 26721680; see Further Reading) fed male mice on either a normal diet (ND) or a high-fat diet (HFD). As expected, the HFD mice became obese, glucose intolerant, and insulin resistant. Sperm heads of ND and HFD mice were injected into normal mouse oocytes and the embryos transferred into surrogate mothers. All male offspring from both groups were fed a normal diet. Although there were no obvious differences in body weight over 16 weeks, offspring produced by the HFD-group sperm had developed severely impaired glucose tolerance and insulin resistance by 15 weeks. Other groups have reported similar findings and several studies have demonstrated second-generation paternal effects (summarized by Patti [2013], PMID 23435955; see Further Reading).

The causative factor in the experiment of Chen and colleagues must have been in the sperm heads, so they set out to identify it. They found they could replicate the glucose intolerance, though not the insulin resistance, by injecting normal zygotes with purified RNA from sperm heads of HFD, but not ND, mice. Further fractionation showed the effect was due to small (30–34 nt) RNA fragments derived from transfer RNAs. Work by other groups has confirmed the role of RNA fragments in similar effects.

A second area where RNA is involved concerns the strange phenomenon of **paramutation**. This describes the situation when an organism shows the phenotype of a mutation that was present in an ancestor, despite not having inherited any of the mutant DNA. A French group studied mice with a loss-of-function allele at the *Kit* locus. Heterozygotes have white spotting. When these heterozygotes were intercrossed or backcrossed with wild-type homozygotes, a proportion of the homozygous wild-type offspring showed the white spotting characteristic of heterozygotes, despite having the wild-type genotype (**Figure 10.20**). The effect persisted for a few generations,

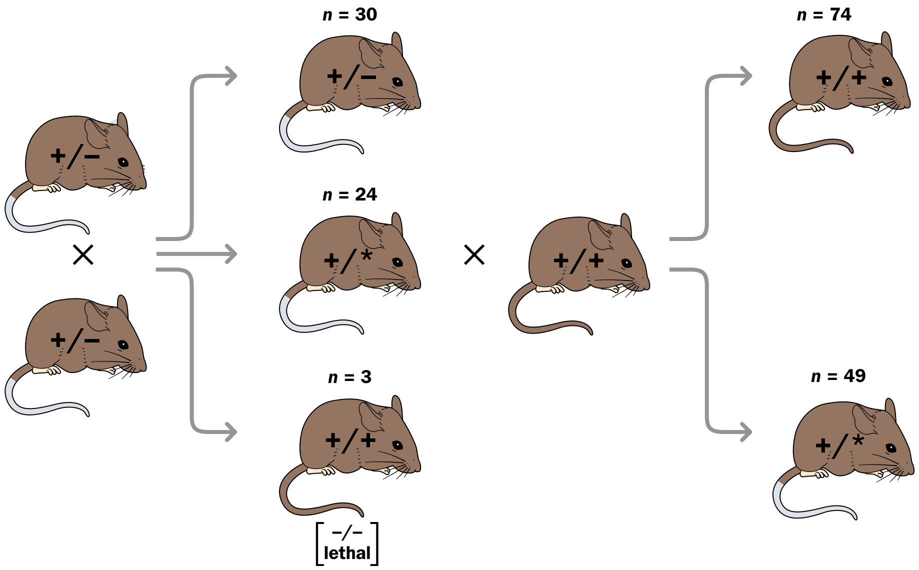


Figure 10.20 Paramutation at the *Kit* locus in mice. Mice heterozygous for a *Kit* mutation (+/–) show white spotting (seen here on the tail). When heterozygous mice were crossed, most (24 out of 27) of the genotypically wild-type (+/+) offspring nevertheless showed the white spotting characteristic of heterozygous animals. The wild-type allele inherited from the heterozygous parent has been somehow changed (paramutated, asterisk). The change is unstable: when +/* mice were crossed with wild-type animals, fewer than the predicted 50% of offspring showed the expected +/* phenotype. (Data from Rassoulzadegan M *et al*. [2006] *Nature* **441**:469–474; PMID 16724059.)

but with diminishing intensity. It could be replicated by microinjecting sperm RNA from a heterozygous mouse into the pronuclei of normal fertilized eggs. Surprisingly, expression of the paramutation phenotype required the DNMT2 RNA methyltransferase (see **Table 10.2**). This and some related transgenerational effects depend in some way on transposons, as only alleles with transposon insertions show the effects. It is intriguing that transposons are silenced by an RNA-dependent mechanism (piRNAs, see Section 9.3).

In summary, transgenerational epigenetic effects are a confusing topic. Such effects undoubtedly occur, but whether they are isolated oddities or part of a whole major, underappreciated substratum of genetics is hard to say. There are many individual reports, but they do not fit easily into a single narrative. The wide-ranging review by Miska & Ferguson-Smith (2016) (PMID 27846492; see Further Reading) attempts to provide a conceptual framework.

10.5 MAKING THE TRANSCRIPT: PROMOTERS AND ENHANCERS

Transcription requires the pre-initiation complex to be assembled at the promoter

As we saw in Chapter 1 (see **Table 1.3**), humans have four RNA polymerases that make RNA copies of a DNA template. Pol-mt transcribes only mitochondrial DNA; Pol I is specialized for the sequences encoding the main ribosomal RNAs; and Pol III transcribes the genes for some of the numerous small noncoding RNAs (**Figure 1.17** illustrates some Pol III promoters). In this chapter we are concerned with RNA polymerase II, which transcribes all protein-coding genes and most noncoding RNA genes. As described in Chapter 1 (Section 1.3), transcription starts with assembly of a pre-initiation complex. First TFIID and TFIIA bind to the core promoter, followed by TFIIB, TFIIF, and Pol II, and finally TFIIE and TFIIH. **Figure 1.16** showed the sequence motifs of a canonical core promoter—TATA box, Inr, DPE, and BRE—to which one could add the CCAAT box at –50 to –80 nt relative to the transcription start site. There is a nucleosome-free region of around 150 bp, and flanking nucleosomes carry H3K4me3 near active promoters, or H3K27me3 at repressed promoters. In undifferentiated cells a proportion of promoters are enriched for both the active H3K4me3 and the repressive H3K27me3 mark (see **Figure 10.4**). It is thought that these bivalent or poised promoters are inactive but poised to become active as the cell differentiates. ENCODE identified 70,292 sequences in the human genome with promoter-like chromatin signatures.

However, few real promoters completely fit this standard description. As mentioned above, about 70% of promoters are associated with CpG islands, and these tend to lack most of the core elements. Promoters differ in the precision with which they specify the transcription start site. Focused or sharp promoters have one or a few fixed transcription start sites. They tend to have TATA, Inr, BRE, and DPE motifs, and they tend to regulate transcription of tissue-specific or signal-responsive genes. Dispersed or broad promoters allow transcription to initiate at many positions over a stretch of maybe 100 nucleotides. They are less likely to show any of the canonical core elements apart from multiple Inr-related motifs. The genes they regulate tend to be housekeeping genes that are expressed at a relatively steady level in many cell types (**Figure 10.21**).

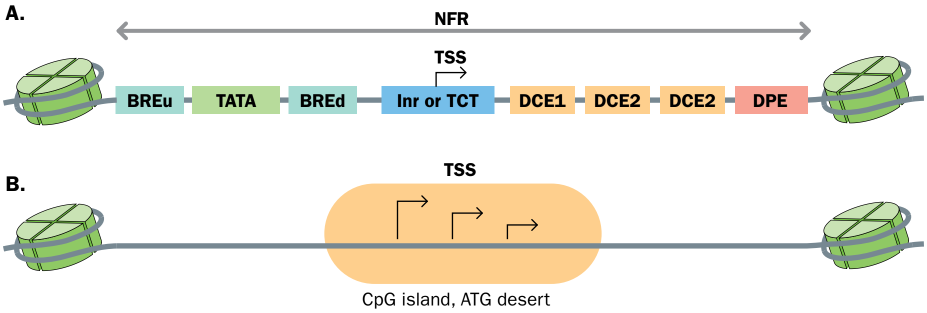


Figure 10.21 Canonical and noncanonical core promoters. (**A**) A canonical promoter showing the TATA box, upstream (u) and downstream (d) BRE elements, and Inr, TCT, DCE, and DPE elements. TCT is used by the TATA-binding protein-related factor TRF2 in ribosomal protein-coding genes. See **Figure 1.16** and Roy & Singer (2015) for descriptions of these elements. (**B**) A noncanonical promoter marked by CpG islands and a region depleted in ATG sequences. TSS, transcription start site; NFR, nucleosome-free region. (Adapted from Roy AL & Singer DS [2015] *Trends Biochem Sci* **40**:165–171; PMID 25680757. With permission from Elsevier.)

Contrary to the simple picture, it would appear that many promoters, especially those lacking TATA boxes, are bidirectional, initiating divergent transcription from both DNA strands in opposite directions (see Wu & Sharpe [2013], PMID 24267885, in Further Reading). Only a small proportion of the transcripts are stable polyadenylated mRNAs; many are part of the pervasive transcription identified by ENCODE. Also, many promoters are occupied by RNA polymerase II in a pre-initiation complex even when the downstream gene is not being expressed. It appears that the decisive control on gene expression is not so much assembly of the pre-initiation complex, as whether the polymerase is able to move into elongation mode, as described below. Roy & Singer (2015) (PMID 25680757; see Further Reading) review the whole topic.

Many genes have more than one promoter

At least half of all mammalian genes have two or more alternative promoters. These will drive transcription from alternative versions of a first exon, which may or may not be coding. As we saw in Chapter 1 (**Figure 1.19**), introns are defined by consensus sequences at either end: a 5′ donor sequence including the invariant GT (GU in the RNA), and a 3′ acceptor sequence with the invariant AG. Alternatively viewed, exons are flanked by 5′ acceptor and 3′ donor sequences. However, the 5′ end of the first exon does not have an acceptor splice site. Thus splicing of the transcript from an upstream exon 1 can pass over the sequence of any number of downstream alternative first exons to splice on to exon 2 (**Figure 10.22**). Some genes have whole batteries of alternative first exons. The UDP glycosyltransferase gene *UGT1A1* on chromosome 2 (position 2q37) has 13 alternative first exons. On chromosome 5, the protocadherin α and γ genes each consist of large tandem arrays of 2400 bp alternative first exons that encode the bulk of the protein. These are spliced onto three small invariant exons that encode the C-terminal part of the protein. There are some analogies to the immunoglobulins (see Section 11.5), in that both have variable N-terminal but constant C-terminal regions, although this is achieved by entirely different means.

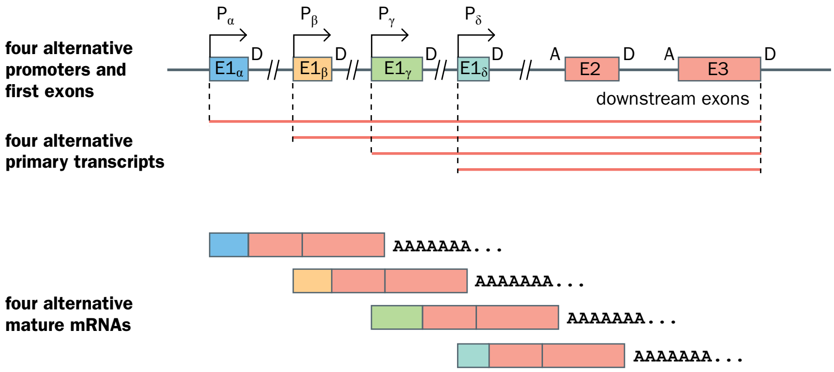


Figure 10.22 A gene with several alternative promoters. This gene has four alternative promoters (P_α–P_δ) and first exons (1_α–1_δ, shown as different-colored boxes), plus two downstream exons (E2, E3; pink boxes). The positions of splice donor (D) and splice acceptor (A) sites are shown. Each splice junction has to be made by joining a donor and an acceptor site, so exon 1_α is spliced on to exon E2 and not exon 1_β. The mature mRNA has a poly(A) tail.

Some alternative promoters are internal to a gene, within an intron, rather than upstream. In such cases, exons upstream of the internal promoter are not included in that transcript. The 79-exon dystrophin gene has several examples (**Figure 10.23**).

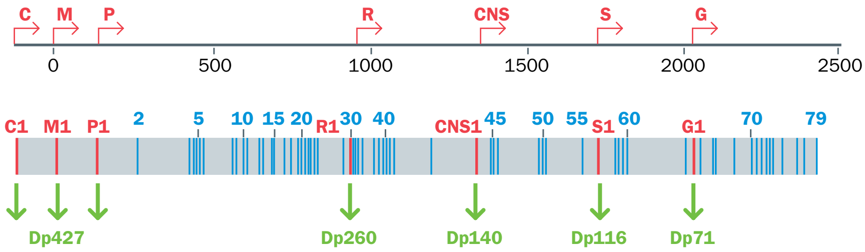


Figure 10.23 Alternative promoters for the dystrophin gene. The positions of seven alternative promoters are shown at the top. C, cortical; M, muscle; P, Purkinje cells; R, retinal, brain, and cardiac muscle cells; CNS, central nervous system; S, Schwann cells; G, general (almost ubiquitously expressed, but not detectable in fully differentiated skeletal muscle). Each promoter uses its own first exon, shown in red on the diagram of the gene structure, together with the shared downstream exons shown and numbered in blue. The protein product from each is named for its molecular mass; for example, Dp427 is a 427 kDa dystrophin protein.

The alternative promoters can serve two purposes. First, they can involve different regulatory elements. Tissue-specific alternative promoters can allow different regulation of gene expression in different tissues. The various promoters in the dystrophin gene are an example. Differential regulation may be specific to particular developmental stages. Even imprinting can be differentially regulated; for example, the *PLAG1* gene on chromosome 6 has two promoters located 55 kb apart. The downstream promoter is imprinted and is responsible for imprinted expression of the gene in many tissues. However, the upstream promoter is not imprinted, and allows biallelic expression in peripheral blood leukocytes. It is perhaps significant that the first intron of a gene is often by far the longest (genome-wide mean size 14,186 bp, compared to 4,847 bp for internal introns). Thus, promoters often lie far upstream of the internal exons of a gene, and alternative promoters are often well separated from each other. This would allow promoters to sit in a different chromatin environment from each other and from the bulk of the gene.

A second value of alternative promoters is that they allow functionally significant sequence differences between alternative first exons. Alternative 5′ untranslated regions may contain different regulatory elements. If the alternative exons contain coding sequence, they can result in protein isoforms with different properties. The examples of UDP glycosyltransferase and protocadherin, described above, show how this mechanism can allow a single gene to generate a whole family of proteins. Isoforms may have different subcellular locations (for example, soluble or membrane-bound) or different functions. For example, the progesterone receptor gene *PGR* on chromosome 11 uses alternative promoters to produce two isoforms, PRA and PRB, that differ by 165 N-terminal amino acids. Both isoforms are transcription factors, members of the nuclear receptor family described in Chapter 3, but they target different response genes and have different physiological effects. An extreme example is the *CDKN2A* gene (see **Figure 19.13**). Here, two alternative first exons each contain a translational start site. Depending on which one is used, the sequence of the shared downstream exons is translated in different reading frames. Thus, the same downstream exons encode totally different proteins, depending on which promoter is used.

A quick check with any genome browser shows that most human genes encode more than one transcript. In December 2017, ENSEMBL listed 200,310 human transcripts but only 20,338 protein-coding genes. The ENCODE pilot project identified 2608 transcripts from the 487 loci examined, an average of 5.4 per locus. In addition to the use of multiple promoters, most transcripts are subject to alternative splicing, as described below. An additional mechanism, RNA editing, which operates only on certain genes, is also mentioned below. A further example, the remarkable series of complex rearrangements that generate the vast diversity of immunoglobulin and T-cell receptor molecules from a small number of genes, is described in Chapter 11 (Section 11.5).

Enhancers

Enhancers are regulatory elements that are located some distance away from the gene whose expression they control. They may be upstream or downstream of their target gene and may be as far as a megabase away. In some cases there may be other genes located in the DNA between an enhancer and its target. Enhancers have much in common with promoters—both are regions of accessible chromatin containing binding sites for transcription factors. Enhancers tend to carry the H3K4me1 mark on associated nucleosomes, compared to H3K4me3 on promoters. Sequences that carry the chromatin signature of enhancers in one cell type may lack this in other cell types, and may not work in them in functional studies. Overall, ENCODE identified 399,124 enhancer-like sequences in the cell types studied. Using a more stringent definition, the study by Andersson *et al*. in 2014 (PMID 24670763; see Further Reading) identified 43,011 putative enhancers in a detailed analysis of 800 human cell types, and characterized their activity across cell and tissue types.

DNA looping brings enhancers into close proximity to the promoters they control, probably with the assistance of the Mediator and cohesin complexes (**Figure 10.24**). Mediator, a 26-subunit protein complex, is involved in all aspects of gene expression; with its many subunits it can bind RNA polymerase and bridge across to transcription factors. Cohesin forms a ring that can enclose two DNA double helices. It is best known for its role in holding sister chromatids together (see Section 2.3) but it is also involved in stabilizing promoter–enhancer interactions. Chromosome conformation capture (see **Box 10.1**) confirms the reality of this looping, though it also reveals promoter–promoter and enhancer–enhancer interactions. As mentioned above, promoters and enhancers have a great deal in common. Unexpectedly, RNA sequencing reveals that enhancer sequences are transcribed, in both directions, producing short-lived noncoding RNAs. The function, if any, of these eRNAs is unknown, but the study by Andersson and colleagues (2014) showed that they mark active enhancers. Possibly they help engage Mediator and assist enhancer–promoter looping. An important paper by Lupiáñez and colleagues in 2015 (PMID 25959774; see Further Reading) showed that enhancer–promoter interactions are controlled by the topologically associated domain (TAD) structure of chromosomes; changes in TAD boundaries cause mis-expression of genes (see **Figure 16.9**).

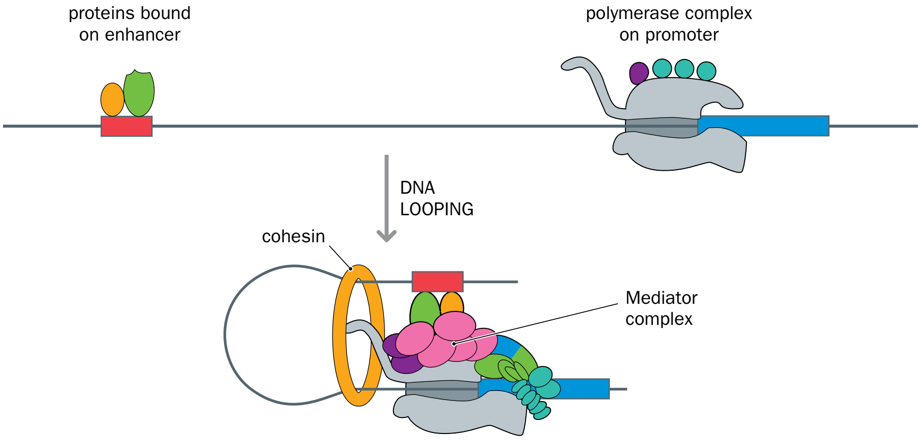


Figure 10.24 Enhancers and promoters are brought together by DNA looping. Regulatory elements such as enhancers (red box) may be located hundreds of kilobases upstream or downstream of the gene they control (blue box). DNA looping allows direct physical interactions between proteins bound to these distal elements and some of the many proteins bound to the promoter. The Mediator and cohesin multiprotein complexes stabilize the interaction. For clarity, only the RNA polymerase is shown at the promoter.

Enhancers are responsible for most of the tissue- and cell-specificity of gene expression. Important developmental genes are often located in gene deserts—long stretches of genomic DNA containing no protein-coding genes but many enhancers. Probably the presence of other expressed genes would excessively complicate the chromatin interactions necessary for enhancer function. A given gene may have a whole battery of enhancers (**Figure 10.25**). Individual enhancers control expression in specific tissues, as revealed by experiments in which transgenic mice are made that carry a *lacZ* reporter gene driven by the enhancer in question (see **Figure 13.7B**, for an example). Some human diseases illustrate the role of tissue-specific enhancers. Total loss of function of an important developmental gene can cause multisystem abnormalities—but if just one enhancer is inactivated, while the other gene functions remain intact, the result may be just one particular facet of the usual syndrome (**Table 10.5**). The most important developmental genes, which specify the identity of cell lineages, are often associated with so-called super-enhancers or stretch enhancers. These are exceptionally long (>3 kb) and complex enhancer sequences that bind Mediator and a large number of transcription factors.

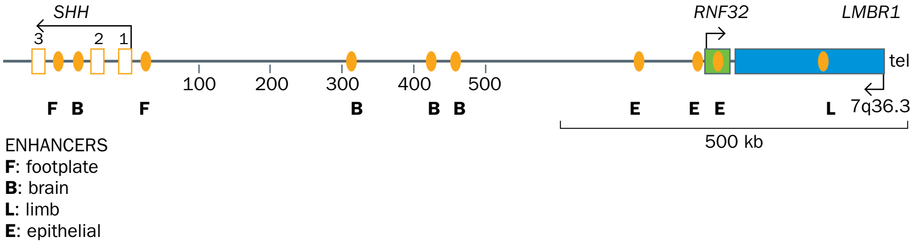


Figure 10.25 Developmental genes may have several distant enhancers, each controlling one aspect of tissue-specific expression. Arrows indicate the direction of transcription of the *SHH* (Sonic Hedgehog) developmental gene (first three exons shown) and of neighboring genes. Tissue-specific expression of the *SHH* gene is controlled by a series of enhancers (orange ovals) located up to 1 Mb upstream. The limb enhancer lies in an intron of the unrelated *LMBR1* gene; point mutations in it produce polydactyly. Although these changes occur within the *LMBR1* gene, that gene is actually irrelevant to the phenotype. tel, telomere. (Adapted from Lettice LA *et al*. [2011] *Hum Mutat* **32**:1492–1499; PMID 21948517. With permission from John Wiley & Sons, Inc., © 2011 Wiley Periodicals, Inc.)

TABLE 10.5 EFFECT OF INACTIVATING AN INDIVIDUAL ENHANCER OF A HUMAN DEVELOPMENTAL GENE WHOSE EXPRESSION IS CONTROLLED BY MULTIPLE ENHANCERS

Gene	Variant	Phenotype
SOX9	Coding mutation	Campomelic dysplasia
	Deletion or duplication of noncoding element 500 kb from promoter	Disorders of sex development
	Duplication of noncoding element 1200 kb from promoter	Brachydactyly-anonychia
	Point mutation or deletion of mandibular enhancer 1450 kb from promoter	Pierre–Robin sequence
SHH	Coding mutation	Holoprosencephaly
	Point mutation of brain enhancer 460 kb from promoter	Holoprosencephaly
	Point mutation or duplication of limb enhancer 1000 kb from promoter	Polydactyly
TBX5	Coding mutation	Holt–Oram syndrome
	Point mutation in heart enhancer 90 kb from promoter	Congenital heart defects
PTF1A	Coding mutation	Agenesis of pancreas and cerebellum
	Point mutation or deletion of pancreatic enhancer 25 kb from promoter	Agenesis of pancreas

Mutations of the coding sequence produce a multisystem syndrome. Mutations of specific enhancers affect just a single system. Data assembled by Gordon CT & Lyonnet S (2014) *Nat Genet* **46**:3–4; PMID 24370740.

Elongation of the transcript

Unlike DNA polymerases, RNA polymerases do not need a primer to get started, but further actions of the TFII transcription factors are needed before transcription can start. One subunit of TFIIH is a DNA helicase. This uses energy from the hydrolysis of ATP to open up the DNA double helix, giving RNA polymerase II (pol II) access to the template strand. TFIIH-dependent phosphorylation of serine residues in the C-terminal domain of pol II allows the polymerase to escape from the transcription start site and start RNA synthesis. However, two factors associated with the pol II protein, DSIF (DRB sensitivity-inducing factor) and NELF (negative elongation factor), cause it to pause 20–60 nucleotides downstream of the transcription start site. Short oligoribonucleotides are produced in a process of abortive initiation until further phosphorylation by P-TEFb (positive transcription elongation factor b) allows the polymerase to move into elongation mode. The average speed is then about 20 nucleotides per second, although there are pauses and spurts. At this speed it would take more than 24 hours to transcribe the 2.4 Mb dystrophin gene! Some data suggest that the transition from paused into elongation mode is the most critical part of transcription control. Most promoters, whether active or inactive, are said to be occupied by a pre-initiation complex, but only polymerases on active genes are able to move into elongation mode. As mentioned above, certain promoters carry both the activating H3K4me3 and repressive H3K27me3 marks. RNA polymerase at these promoters is said to be poised: it is thought that poised promoters may be particularly responsive to external signals.

Termination of transcription

In Section 1.4 we briefly outlined the process. Specific signals govern termination of transcription by polymerases I and III. For polymerase II, the only signal seems to be the poly(A) addition site—AAUAAA or a closely similar sequence. Two models have been proposed for how this causes termination.

- The allosteric model proposes that the polyadenylation signal induces a change in the RNA polymerase that commits it to termination.
- The torpedo model proposes that the polyadenylation signal triggers endonucleolytic cleavage of the nascent transcript some small distance downstream. The Xrn2 exonuclease attaches to the 5′ end of the RNA at the cleavage site, degrading it in a 5′ → 3′ direction. A sort of race ensues between the polymerase, continuing along the DNA and elongating the tail of the transcript, and the exonuclease coming up behind the polymerase and eating up the transcript (**Figure 10.26**). The exonuclease is the faster of the two, and when it catches up with the polymerase, transcription terminates.

Porrua & Libri (2015) provide an extensive review, while Libri (2015) briefly summarizes recent evidence favoring each model (PMID 25650800 and 26474063, respectively; see Further Reading). The truth may lie in some combination of the two.

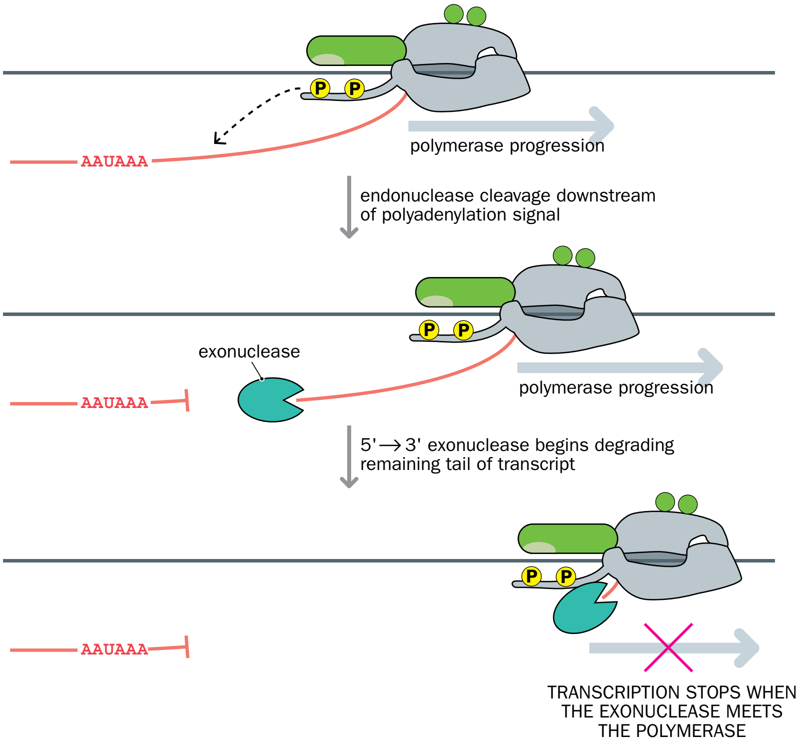


Figure 10.26 The torpedo model for termination of transcription. The transcript (red line) is cleaved (dotted line) downstream of the AAUAAA polyadenylation signal. An exonuclease attaches to the free 5′ end and works its way along the tail of the nascent transcript. Transcription ceases when the exonuclease reaches the body of the polymerase.

10.6 POST-TRANSCRIPTIONAL REGULATION

Alternative splicing allows one primary transcript to encode multiple protein isoforms

The basic mechanism of splicing was described in Chapter 1 (see **Figures 1.18–1.21**). However, for the great majority of human genes, there is more than one way of splicing. Alternatively spliced transcripts (splice isoforms) can be identified for almost every human gene. These may skip one or more exons, include additional internal exons, or vary the length of an exon by repositioning the exon–intron junction (**Figure 10.27**). It may be that some of these isoforms just reflect imprecision in the complex splicing machinery and are not functionally significant—but numerous examples are known in which alternative splicing clearly is functional.

Functional alternative splicing can have a variety of results. Alternatively spliced exons may encode signals governing different intracellular localizations. Some proteins have soluble and membrane-bound isoforms, produced by the inclusion or omission of an exon that encodes a transmembrane domain. The two forms may compete. Competition for a ligand between soluble and membrane-bound isoforms can regulate

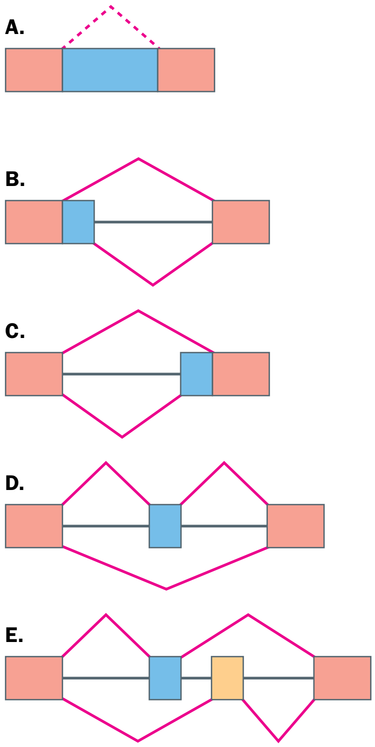


Figure 10.27 Types of alternative splicing event. Red-colored boxes represent exons that are always included in the mature mRNA. (**A**) An intron (blue) is either retained or excluded. (**B**) Use of alternative splice donor sites results in the inclusion or exclusion of the sequence in blue. (**C**) Use of alternative splice acceptor sites results in the inclusion or exclusion of the sequence in blue. (**D**) The exon in blue may be either included or skipped (a cassette exon). (**E**) Alternative exons: the mature mRNA includes either the exon in yellow or the exon in blue, but not both or neither.

cell surface receptors. The splicing may be tissue-specific, so that different tissues contain different variants. Alternatively spliced exons may introduce sites for important post-translational modifications, such as serine phosphorylation. Some genes are regulated by inclusion or exclusion of a poison exon that includes a premature stop codon.

As the most complex part of the human body, the central nervous system may need the most complex proteome, and alternative splicing is particularly marked in neurons. Many widely expressed genes have neuron-specific splice isoforms. As long ago as 1994 a compilation listed almost 100 examples of neuron-specific splicing, including every type shown in **Figure 10.27**. Many of the variant isoforms of ion channels and receptors are known to be functionally important. Some genes encode an extraordinary number of different transcripts. The example of protocadherins α and γ, with their batteries of alternative promoters, was mentioned earlier. **Figure 10.28** shows the neurexin 3 (*NRXN3*) gene. This large gene on chromosome 14q encodes a cell adhesion and receptor molecule that is present at synapses in the nervous system. It has two promoters and 24 downstream exons. Seven of the exons can be each individually included or excluded in transcripts; one exon has alternative splice donor sites, and three have alternative splice acceptors. One of the alternatively spliced exons includes a stop codon, the use of which would produce a protein lacking transmembrane and cytoplasmic domains. Potentially this one gene could encode 1000 different proteins. Genes expressed in the central nervous system also often contain micro-exons, 6–30 nt exons that tend to be inefficiently spliced, thus adding to the variety of mature mRNAs.

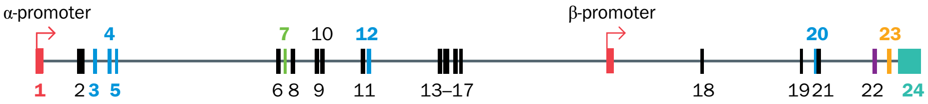


Figure 10.28 Alternative splicing of the neurexin 3 transcript in the nervous system. There are two alternative promoters, α and β (red bars). Exons 3, 4, 5, 12, 20, and 24 (blue) can each be either included or skipped. Exon 7 (light green) can be included, using either of two alternative 5′ splice acceptor sites, or it can be completely skipped. Exon 22 (purple) has two alternative 3′ splice donor sites. Exon 23 (yellow) has two alternative 5′ splice acceptor sites that use different reading frames, one of which leads to an in-frame stop codon within this exon. The protein produced from this variant lacks the transmembrane and cytoplasmic domains encoded by exon 24. Exon 24 (dark green) has three alternative 5′ splice acceptor sites. By using different combinations of variants, this single gene could potentially encode about 1000 different proteins.

What controls alternative splicing?

The basic splicing machinery was described in Chapter 1. However, not all splice sites are equal. The choice of where the spliceosome is assembled on the primary transcript depends on a balance of positive and negative factors. The sequence surrounding the invariant GU...AG sequences may be a better or worse fit to the optimum sequence. Nearby splicing enhancers bind SR (serine-arginine) proteins that help anchor the spliceosome in place, whereas splicing suppressors bind hnRNP (heterogeneous ribonucleoprotein) proteins that have the opposite effect. Thus, splice sites can be strong or weak. Weak sites may be skipped in favor of an alternative. Splicing patterns are often tissue-specific, presumably because enhancers or suppressors of splicing bind tissue-specific proteins. Attempts have been made to identify a splicing code that would predict this. A large study of 10,689 alternatively spliced human exons by Xiong and colleagues in 2015 (PMID 25525159; see Further Reading) used machine learning to identify combinations of 1393 sequence features (sizes of exons and introns, binding sites, structural features, and so on) that predicted the effect of sequence variants on splicing. Applying the predictor to 650,000 variants resulted in many successful predictions of pathogenicity. Evidently alternative splicing is predictable, but not by examination of just a few features.

Epigenetic marks can also affect splicing. Exons show epigenetic differences relative to introns, including increased CpG methylation and different histone modifications. Although these marks are present on the DNA and not on the primary transcript, they can affect the progress of the RNA polymerase along a gene, either directly or by binding proteins that block the polymerase. Splicing is co-transcriptional, and polymerase pausing can affect the splicing machinery, for example allowing more time to assemble a spliceosome on a weak splice site.

Alternative sites for 3′ cleavage and polyadenylation produce additional variation

About half of human genes use alternative cleavage and polyadenylation to generate messenger RNA transcripts that differ in the length of their 3′ untranslated regions (3′ UTRs), while producing the same protein. The use of specific alternative sites often depends on the particular cell type and can change upon proliferation or differentiation. As described below, 3′ UTRs include recognition sites for RNA-binding regulatory proteins and small RNAs, particularly microRNAs. The alternative 3′ UTRs can affect the cellular location or stability of a transcript and the abundance of the encoded protein.

RNA editing can change the sequence of the mRNA after transcription

In contravention of the central dogma, there are examples in which the DNA sequence of a gene does not fully determine the sequence of its transcript. RNA editing involves the insertion, deletion, or modification of specific nucleotides in the primary transcript. In humans the main types of event are deamination of cytosine or adenine resulting in C>U or A>I conversions (I is inosine; see below).

A>I editing is performed by members of the ADAR (adenosine deaminase acting on RNA) family of deaminases (**Figure 10.29**). Inosine base-pairs with cytosine rather than thymine. Over 99% of A>I edits occur in Alu sequences. Editing in human protein-coding genes is particularly seen in the central nervous system. Often the editing converts CAG codons, encoding glutamine (Q), into CIG codons that, like CGG, encode arginine (R). So-called Q/R editing alters the function of several genes encoding neurotransmitter receptors or ion channels (*GABRA3*, *GRIA2*, and *GRIK2*). A>I editing sometimes makes other coding-sequence changes (I/V, Y/C, N/S, and so on). In the *HTR2C* serotonin receptor gene, A>I editing at splice sites modulates alternative splicing.

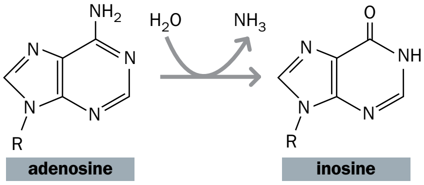


Figure 10.29 Deamination of adenosine. Enzymes of the ADAR family deaminate the amino group at carbon 6 of adenosine to produce inosine. R, ribose.

C>U editing is performed by enzymes of the APOBEC family. The human apolipoprotein gene *APOB* encodes the large ApoB100 protein in the liver. In the intestine, however, C>U editing at nucleotide position 6666 of the mRNA causes replacement of the CAA glutamine codon by a UAA stop codon (**Figure 10.30**). The mRNA now encodes a shorter polypeptide, ApoB48. Uncontrolled APOBEC editing is a major source of mutations in tumors (see Chapter 19)—for example, converting arginine codon 3916 to a stop codon in the *NF1* (neurofibromin) gene.

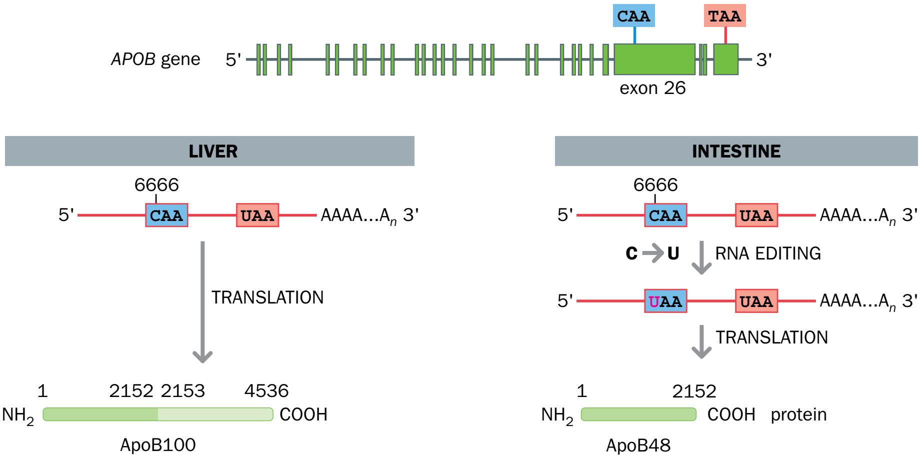


Figure 10.30 APOBEC RNA editing: the two products of the *APOB* gene. In the liver, the *APOB* mRNA encodes a 4536-residue protein, ApoB100. In the intestine, the APOBEC1 cytosine deaminase specifically converts cytosine 6666 in the mRNA to uridine, changing the CAA glutamine codon 2153 into a UAA stop codon. The mRNA now encodes ApoB48, consisting of just the first 2152 amino acids of ApoB100.

Regulation of translation

For protein-coding genes, the layers of regulation extend from transcription and maturation of the transcript through to translation. The 5′ and 3′ UTRs of mRNAs have important regulatory functions. As described below, the 3′ UTR is the main site of miRNA binding. Two features in the 5′ UTR can affect initiation of translation.

- Between 10 and 30% of human mRNAs have additional open reading frames (ORFs) upstream of the main coding sequence. These can inhibit translation of the main reading frame. For example, upstream ORFs of the *TPO* gene limit production of the encoded protein, thrombopoietin. Mutations that prevent recognition of the upstream ORF lead to excessive production of thrombopoietin, causing hereditary thrombocythemia. The *HR* (hairless) gene has four upstream ORFs in its 691 bp 5′ UTR (**Figure 10.31**). Loss-of-function mutations in the inhibitory *U2HR* cause overexpression of the *HR* protein. *HR* is a regulator of Wnt signaling; the disturbed Wnt signaling affects the cycling of hair follicles, resulting in an autosomal dominant hair-loss syndrome, Marie Unna hypotrichosis (OMIM #146550).

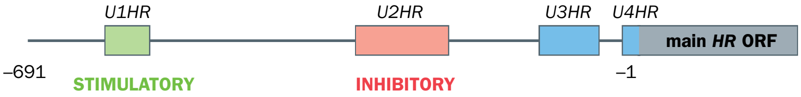


Figure 10.31 Upstream open reading frames in the 5′ untranslated region of the *HR* (hairless) gene control production of the HR protein. ORF open reading frame. (Reprinted from Wen Y *et al.* [2009] *Nat Genet* **41**:228–233; PMID 19122663. With permission from Springer Nature. Copyright © 2009.)

- Stem-loop structures formed by the single-stranded mRNA can directly affect the level of the gene product by impeding the progress of ribosomes, and they act as recognition sites for RNA-binding regulatory proteins. **Figure 10.32** shows how stem-loop iron-responsive elements in the ferritin and transferrin receptor genes orchestrate the response to dietary iron.

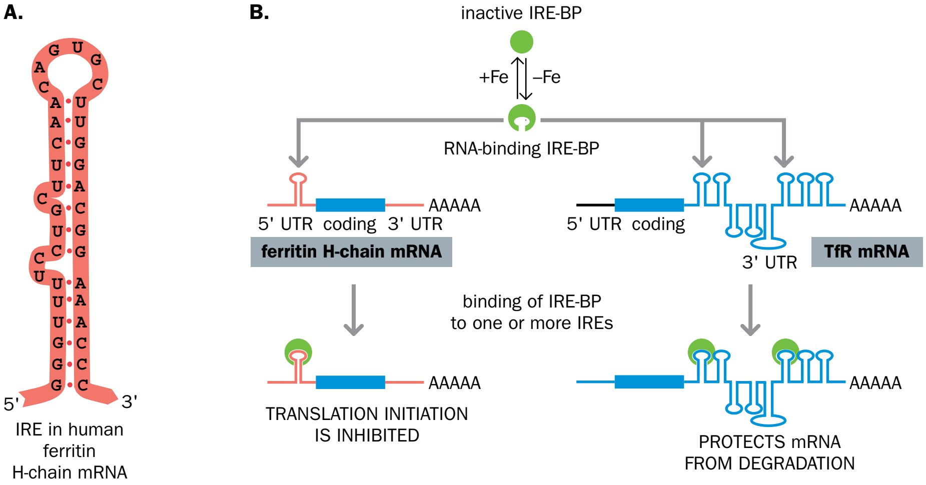


Figure 10.32 Iron-response elements in the ferritin and transferrin receptor mRNAs. (**A**) Stem-loop structure of an iron-response element (IRE) in the 5′ untranslated region (UTR) of the ferritin heavy (H)-chain mRNA. (**B**) Low iron concentrations activate a specific IRE-binding protein (IRE-BP), enabling it to bind the IRE in the ferritin heavy-chain gene and also IREs in the 3′ untranslated region of the transferrin receptor (TfR) mRNA. Binding inhibits the translation of ferritin but protects the transferrin receptor mRNA from degradation. When the iron concentration is high, IRE-BP is degraded, releasing translational repression of stored ferritin mRNA but inhibiting production of the transferrin receptor.

The discovery of many small RNAs that regulate gene expression caused a paradigm shift in cell biology

Two seemingly disparate lines of research in the 1990s, both involving the *C. elegans* worm, alerted biologists to the unsuspected importance of very small RNAs. Andrew Fire and Craig Mello received the 2006 Nobel Prize in Physiology or Medicine for their roles in understanding RNAi—the specific inhibition of gene expression by short double-stranded RNA molecules (see Section 8.5 and **Box 8.2**). Their Nobel lectures, describing the process of discovery, can be read at http://nobelprize.org/nobel _prizes/medicine/laureates/2006/. Meanwhile Victor Ambros and Gary Ruvkun, among others, opened up the world of microRNAs (miRNAs) in development. Their accounts of how they came to make these discoveries can be read in two Commentaries published in a *Cell* supplement (see Further Reading).

Not all small RNAs are rare. An adult *C. elegans* cell contains more than 50,000 molecules each of miRNAs 2, 52, and 58. Why were they not discovered earlier? When researchers ran RNA gels, they assumed—usually with good reason—that the smear of very-low-molecular-weight RNA at the bottom of the gel consisted of degradation products and was of no interest. In addition, such short molecules are difficult to study: standard bioinformatics approaches overlook them, and such tiny targets were seldom hit in mutagenesis experiments. Thus, although they must have been observed in many laboratory experiments, they were not recognized.

MicroRNAs as regulators of translation

The various categories of small RNAs were listed in **Table 9.6**. MicroRNAs, siRNAs, and piRNAs are all short RNAs that act in related ways to repress expression of their targets (see **Figure 8.19**). MicroRNAs are processed from capped and polyadenylated precursor pri-miRNAs through the Drosha and Dicer ribonucleases (**Figure 10.33**). One strand, the guide strand, of the resulting 21–22 nt RNA complexes with Argonaute and GW182 proteins to form the RISC (RNA-induced silencing complex, see **Figure 8.19**). miRNAs hybridize to target sequences in mRNAs, primarily in the 3′ UTRs. Hybridization

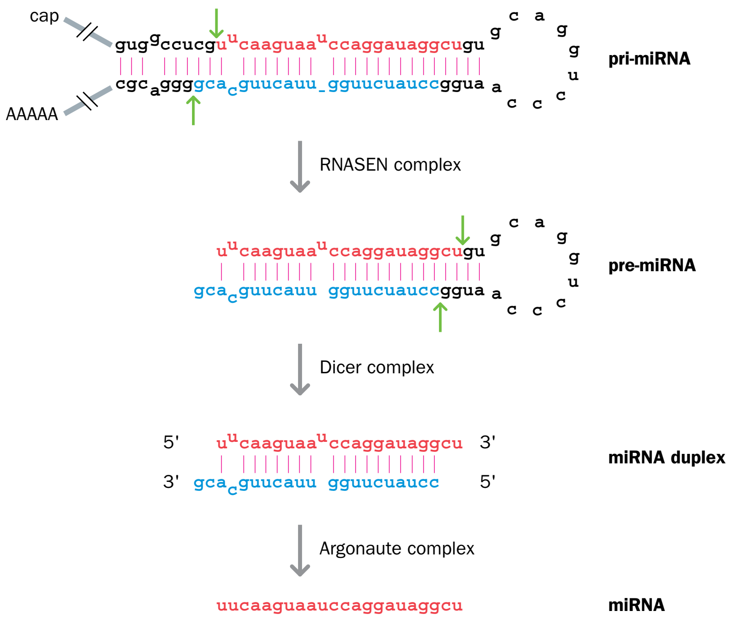

cap
gugᵍccucguᵘcaaguaaᵘccaggauaggcugu g c a g g u
cgcₐggggcaᶜguucauu_gguucuauccggua a c c c
AAAAA
pri-miRNA
RNASEN complex
uᵘcaaguaaᵘccaggauaggcugu g c a g g u
gcaᶜguucauu gguucuauccggua a c c c
pre-miRNA
Dicer complex
5' uᵘcaaguaaᵘccaggauaggcu 3'
3' gcaᶜguucauu gguucuaucc 5'
miRNA duplex
Argonaute complex
uucaaguaauccaggauaggcu
miRNA

Figure 10.33 Maturation of a miRNA. miRNA26a1 is initially synthesized as a capped and polyadenylated pri-miRNA. Inverted repeats in the pri-miRNA cause it to adopt a hairpin structure. Asymmetric cleavage by the Drosha (RNASEN) and Dicer enzymes (green arrows) produces a short double-stranded RNA with 3′ overhangs. Argonaute proteins complex with the guide strand (red) to form the RNA-induced silencing complex (compare with **Figure 8.19**).

depends on imperfect base pairing, with nucleotides 2–8 (the seed region) providing the main specificity. Normally miRNAs initially repress translation of an intact mRNA, although subsequently the target mRNA is often deadenylated and degraded. In some cases miRNAs have been reported to activate, rather than repress, gene expression.

The miRBase database (www.miRBase.org; consulted January 2017) lists 1881 precursor and 2588 mature miRNAs in the human genome. The genes occur singly or in clusters in a variety of genomic contexts (**Figure 10.34**).

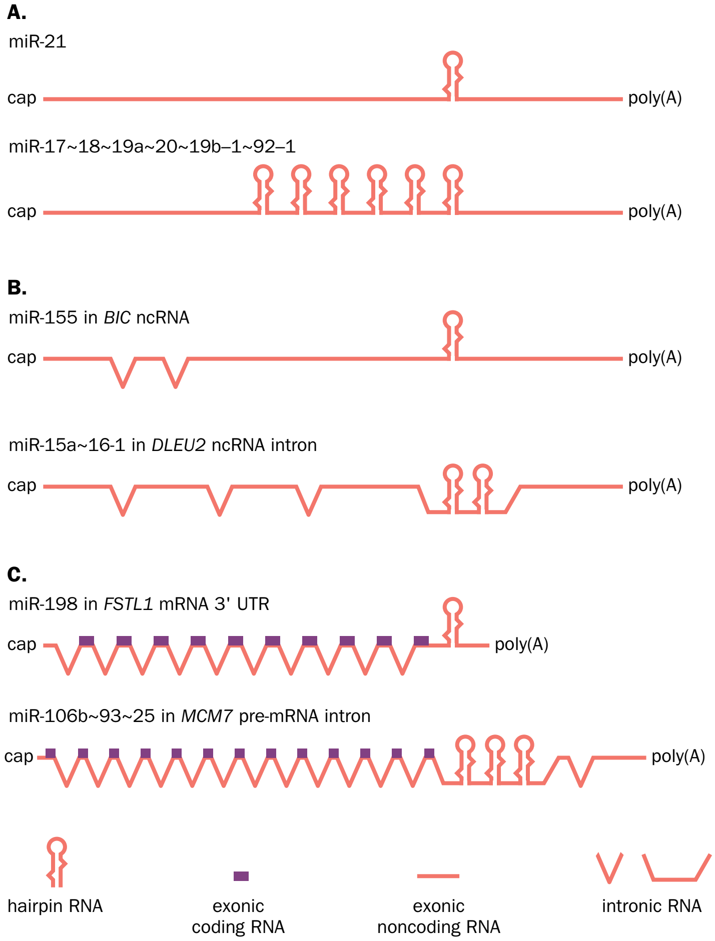


Figure 10.34 MicroRNA genes. (**A**) Stand-alone primary transcripts. (**B**) miRNA genes within exons or introns of a long noncoding RNA. Expression will depend on expression of the host RNA. (**C**) miRNA genes within the 3′ UTR or introns of a host protein-coding mRNA.

There is no one-to-one relation between miRNAs and their targets. Each miRNA can affect translation of many targets, and many messenger RNAs are targeted by multiple microRNAs. Thus miRNAs form a very broad-based regulatory system. However, individual miRNA effects are usually modest. Few proteins show an increase as great as twofold when an inhibitory miRNA is knocked down, or repression greater than 50% when the cell is transfected with an miRNA. The general picture that emerges is of very wide-ranging but relatively small effects compared to controls on transcription.

Because an individual miRNA can bind to many different target mRNAs, the different targets compete for the miRNA. Unless the miRNA is very abundant in a cell, higher transcription of target A would mop up more of the miRNA, leading to less binding to target B. Since miRNA effects are almost always repressive, this would increase translation of target B. The result is a dense network of regulatory interactions. Regulation by competing endogenous RNAs may be an important general phenomenon; there can also be competition for RNA-binding proteins as well as for miRNAs. Two examples illustrate these effects:

- Cells are very sensitive to the level of the PTEN tumor suppressor protein (see Chapter 19). *PTEN* has a pseudogene, *PTENP1*, which is transcribed although not translated. The PTENP1 transcript competes with PTEN mRNA for binding of miRNAs 17, 21, 214, 19, and 26. Expression of the pseudogene suppresses cell growth by reducing miRNA-based repression of PTEN and hence up-regulating the level of the growth-inhibitory PTEN protein;
- Circular RNAs had generally escaped detection, but were recently shown by Salzman and colleagues to be the predominant transcript isoform of hundreds of human genes (Salzman *et al.* [2012], PMID 22319583; see Further Reading). They are produced by abnormal splicing of multiexon transcripts, where a downstream splice donor site is spliced on to an upstream acceptor. One well-studied circular RNA, CDR1as, contains 77 binding sites for miR-7. It is a stable and abundant miR-7 sponge, and hence is a powerful up-regulator of other miR-7 target transcripts.

SUMMARY

- Humans have only about the same number of protein-coding genes as the 1 mm long nematode worm *Caenorhabditis elegans*. The difference in complexity must be almost entirely due to much more elaborate and sophisticated control mechanisms in humans.
- Cells continuously regulate expression of their genes, both in short-term responses to external signals and long term as part of their specific cell-type and tissue identity. The mechanisms that govern stable tissue-specific identity are generally called epigenetic, but they overlap with the mechanisms responsible for short-term regulation.
- The characteristic forms and behaviors of different cell types result from different epigenetically-determined readouts of the same genome.
- The DNA in an interphase cell nucleus is highly organized. At the largest scale chromosomes occupy discrete territories that contain domains of active and repressed genes. On a smaller scale the chromatin is organized in a hierarchy of loops, including topologically-associated domains (TADs) that constrain regulatory interactions.
- For a gene to be transcribed, the DNA must be accessible, not buried in densely packed chromatin. The local chromatin conformation is governed by a complex interplay between enzymes that methylate cytosine in DNA, enzymes that modify histones in nucleosomes, proteins that bind methylated DNA or modified histones, and ATP-driven chromatin remodeling complexes that can change the positions of nucleosomes.
- RNA polymerase II transcribes all protein-coding genes, and many that specify noncoding RNAs. Transcription of a gene requires a multiprotein transcription initiation complex to be assembled on the promoter.
- Promoter activity is affected by sequence-specific DNA-binding proteins that bind either directly to the promoter or at more distant positions (enhancers and repressors). DNA looping brings distant sites close to the promoter. Co-activators and co-repressors do not bind DNA but are recruited to promoters or enhancers by protein–protein interactions.
- Large-scale studies of human transcripts have shown that the majority of all human genomic DNA is transcribed, at least in some cells and at some times. The function (if any) of most of the transcripts is unknown. This suggests a substantial revision of the traditional view of discrete genes sparsely scattered along genomic DNA.
- Humans have very many more long noncoding RNAs than the *C. elegans* worm, and many of them have roles in regulating gene expression.
- X-inactivation is a major example of an epigenetic process. In a 46,XX cell, one X, selected at random, is inactivated at the blastocyst stage of embryogenesis. Whichever X is chosen in a given cell, all progeny of that cell inactivate the same X.

The effect depends on the XIST long noncoding RNA, which is expressed only from the inactive X and physically coats it, repressing gene expression. X-inactivation is erased at meiosis.

- For a few dozen human genes, expression depends on parental origin because of imprinting. This is an epigenetic process whereby alleles at a locus carry an imprint of their parental origin that governs the pattern of expression.
- Epigenetic marks are normally erased at meiosis but various instances are known where they are transmitted from parent to offspring. How general such transgenerational effects are remains unclear.
- Most protein-coding genes encode multiple polypeptides because of the use of alternative promoters and/or alternative splicing. In many cases, the resulting protein isoforms have different tissue specificities, biological properties, and functions.
- Gene expression is also regulated by controlling whether or not an mRNA will be translated. This often depends on microRNAs that bind to sequences in the 3′ untranslated region of the mRNA.
- MicroRNAs form a complex regulatory network, with each miRNA regulating many genes and each gene regulated by many miRNAs. Competition for miRNAs may be an important regulatory mechanism.

FURTHER READING

Overviews of epigenomics

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**:799–816; PMID 17571346.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74; PMID 22955616. (The Nature ENCODE Explorer [www.nature.com/encode/] is a portal giving access to the large set of 2012 reports from the ENCODE project.)

Roadmap Epigenomics Consortium, Kundaje A *et al*. (2015) Integrative analysis of 111 reference human epigenomes. *Nature* **518**:317–330; PMID 25693563.

Identifying regulatory DNA

Buenrostro JD *et al*. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**:1213–1218; PMID 24097267. (The ATAC technique.)

Giresi PG & Lieb JD (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods* **48**:233–239; PMID 19303047.

Chromatin analysis

Dekker J *et al*. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**:390–403; PMID 23657480.

Ernst J *et al*. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**:43–49; PMID 21441907.

Kadoch C *et al*. (2013) Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* **45**:592–601; PMID 23644491.

Lupiáñez DG *et al*. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**:1012–1025; PMID 25959774.

Lupiáñez DG *et al*. (2016) Breaking TADs: how alterations of chromatin domains result in disease. *Trends Genet* **32**: 225–237; PMID 26862051.

Olivares-Chauvet P *et al*. (2016) Capturing pairwise and multiway chromosomal conformations using chromosomal walks. *Nature* **540**:296–300; PMID 27919068.

Schoenfelder S *et al*. (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**:53–61; PMID 20010836.

Son EY & Crabtree GR (2014) The role of BAF (mSWI/SNF) complexes in mammalian neural development. *Am J Med Genet C Semin Med Genet* **166C**:333–349; PMID 25195934.

Tsurusaki Y *et al*. (2012) Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nat Genet* **44**:376–378; PMID 22426308.

Methylation of DNA and RNA

Kinde B *et al*. (2015) Reading the unique DNA methylation landscape of the brain: non-CpG methylation, hydroxymethylation, and MeCP2. *Proc Natl Acad Sci USA* **112**:6800–6806; PMID 25739960.

Liu N *et al*. (2015) N^6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518**:560–564; PMID 25719671.

Schübeler D (2015) Function and information content of DNA methylation. *Nature* **517**:321–326; PMID 25592537.

Schultz MD *et al*. (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**: 212–216; PMID 26030523.

Shukla S *et al*. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**: 74–79; PMID 21964334.

Smith ZD & Meissner A (2013) DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**:204–220; PMID 23400093.

X-inactivation

Chang SC & Brown CJ (2010) Identification of regulatory elements flanking human XIST reveals species differences. *BMC Mol Biol* **11**:20; PMID 20211024.

Deng X *et al*. (2014) X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet* **15**: 367–378; PMID 24733023.

Migeon BR (2016) An overview of X inactivation based on species differences. *Semin Cell Dev Biol* **56**:111–116; PMID 26805440.

Moindrot B & Brockdorff N (2016) RNA binding proteins implicated in Xist-mediated chromosome silencing. *Semin Cell Dev Biol* **56**:58–70; PMID 26816113.

Imprinting

Soellner L *et al*. (2017) Recent advances in imprinting disorders. *Clin Genet* **91**:3–13; PMID 27363536.

Transgenerational epigenetic effects

Chen Q *et al*. (2016) Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science* **351**:397–400; PMID 26721680.

Miska EA & Ferguson-Smith AC (2016) Transgenerational inheritance: models and mechanisms of non-DNA sequence-based inheritance. *Science* **354**:59–63; PMID 27846492.

Morgan HD *et al*. (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23**:314–318; PMID 10545949.

Patti ME (2013) Intergenerational programming of metabolic disease: evidence from human populations and experimental animal models. *Cell Mol Life Sci* **70**:1597–1608; PMID 23435955.

Pembrey M *et al*. (2014) Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research. *J Med Genet* **51**:563–572; PMID 25062846.

Rassoulzadegan M *et al*. (2006) RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* **441**:469–474; PMID 16724059.

Promoters

Roy AL & Singer DS (2015) Core promoters in transcription: old problem, new insights. *Trends Biochem Sci* **40**:165–171; PMID 25680757.

Wu X & Sharp PA (2013) Divergent transcription: a driving force for new gene origination? *Cell* **155**:990–996; PMID 24267885.

Enhancers

Andersson R *et al*. (2014) An atlas of active enhancers across human cell types and tissues. *Nature* **507**:455–461; PMID 24670763.

Farley EK *et al*. (2015) Suboptimization of developmental enhancers. *Science* **350**:325–328; PMID 26472909.

Gordon CT & Lyonnet S (2014) Enhancer mutations and phenotype modularity. *Nat Genet* **46**:3–4; PMID 24370740.

Hnisz D *et al*. (2013) Super-enhancers in the control of cell identity and disease. *Cell* **155**:934–947; PMID 24119843.

Kearns NA *et al*. (2015) Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat Methods* **12**:401–403; PMID 25775043.

Transcription termination

Libri D (2015) Endless quarrels at the end of genes. *Mol Cell* **60**:192–194; PMID 26474063.

Porrua O & Libri D (2015) Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* **16**:190–202; PMID 25650800.

Splicing of the primary transcript

Xiong HY *et al*. (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**:1254806; PMID 25525159.

Regulation of translation

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* **136**:215–233; PMID 19167326.

Hughes TA (2006) Regulation of gene expression by alternative untranslated regions. *Trends Genet* **22**:119–122; PMID 16430990.

Lee R *et al*. (2004) A short history of a short RNA. *Cell* **116**(2 Suppl):S89–S92; PMID 15055592.

Poliseno L *et al*. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**:1033–1038; PMID 20577206.

Ruvkun G *et al*. (2004) The 20 years it took to recognize the importance of tiny RNAs. *Cell* **116**(2 Suppl):S93–S96; PMID 15055593.

Salzman J *et al*. (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* **7**:e30733; PMID 22319583.

Tay Y *et al*. (2014) The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**:344–352; PMID 24429633.

Noncoding functions of RNA

Holoch D & Moazed D (2015) RNA-mediated epigenetic regulation of gene expression. *Nat Rev Genet* **16**:71–84; PMID 25554358.

Ponting CP *et al*. (2009) Evolution and functions of long noncoding RNAs. *Cell* **136**:629–641; PMID 19239885.

Rinn J & Guttman M (2014) RNA and dynamic nuclear organization. *Science* **345**:1240–1241; PMID 25214588.

GENETIC VARIATION BETWEEN INDIVIDUALS AND SPECIES

PART THREE

11 An overview of human genetic variation

The description of the human genome given in Chapter 9 was largely based on outputs from the Human Genome Project. The human genome reference sequence, however, is just a single representative snapshot, and an artificial one (different parts of the sequence originated from different people). We may speak about *the* human genome, but, in reality, there are many billions of different human genomes that owe their differences to genetic variation.

Genetic variation is mostly inherited, transmitted between generations in gametes. During life, every fertile man makes billions of sperm cells, but each sperm cell—and each egg cell—is genetically unique (pre-existing genetic variation is shuffled at meiosis by recombination and independent chromosome assortment). Every one of us, therefore, arose from a single fertilized egg cell that contained a unique diploid genome. Occasionally, however, splitting of the very early embryo generates genetically identical embryos that give rise to siblings.

Because each of us inherits two different haploid nuclear genomes, one from mum and one from dad, inherited genetic variation occurs within, as well as between, individuals. At any genetic **locus** (DNA region having a unique chromosomal location), the maternal and paternal DNA sequences (**alleles**) are normally identical or slightly different (we are said to be **homozygotes** if the maternal and paternal alleles are identical, or **heterozygotes** if they differ by even a single nucleotide).

Two regions of the human genome are always inherited from a single parent. The nonrecombining portion of the Y chromosome has no sequence counterparts on the X chromosome, and is transmitted exclusively by fathers to sons. Men are said to be **hemizygous** for sequences in this region, having inherited just a single allele in each case. All of us inherit our mitochondrial DNA (mtDNA) exclusively from our mothers. The transmitted maternal egg, however, contains about 100,000 mtDNA molecules that may show some differences in sequence. This type of mitochondrial DNA sequence variation is described as **heteroplasmy**.

In addition to the genetic variation that we inherit from our parents, DNA changes occur in the DNA of our cells throughout life. This post-zygotic genetic variation includes changes to the DNA of both somatic cells and germ-line cells: each of us, therefore, carries cells with different genomes. Although most post-zygotic DNA changes occur in a rather random fashion, some do not. In some cells, notably maturing B and T lymphocytes, certain DNA rearrangements occur in nonrandom, and quite defined, ways that are functionally important.

Individuals differ from each other mostly because our DNA sequences differ, but genetic variation is not the only explanation for differences in **phenotype** (our observable characteristics). During development, additional effects on the phenotype are made by a combination of stochastic (random) factors, differential gene–environment interactions, and additional epigenetic variation that is not attributable to changes in base sequence.

In this chapter we are not particularly concerned with the very small fraction of human genetic variation that causes disease (we cover that in Chapters 15–19). Instead, we focus here on general principles of human genetic variation. In Section 11.1 we consider the origins of DNA sequence variation. DNA repair mechanisms seek to minimize the effects of DNA sequence variation, and in Section 11.2 we outline the different DNA repair mechanisms that work in our cells. Much of our knowledge of the frequencies of different types of DNA variants is being derived from population-based genome sequencing. As described in Section 11.3, the resulting sequence data enable a comprehensive assessment of the extent of human genetic variation, and the different forms in which this variation manifests.

The great majority of genetic variation has no effect on the phenotype (*neutral mutations*), but in Section 11.4 we turn our attention to the small fraction of genetic variation that affects how genes function ("functional genetic variation"). Proteins are the primary endpoint of gene function, and in this section we examine, in a general way, how variation in the sequences of protein products is determined. We consider how different forms of natural selection affect genetic variation, but aspects of population genetics that relate to the spread of advantageous and harmful disease-associated DNA variants will be described more fully in later chapters.

Functional genetic variation is most highly developed in genes encoding immune system proteins that must recognize antigens of invading microorganisms and viruses. In Section 11.5 we explain how programmed DNA rearrangements at immunoglobulin and T-cell receptor loci, in maturing B and T cells, respectively, allow each person to produce extraordinarily diverse populations of antibodies and T-cell receptors, and we describe the basis for the quite exceptional sequence variation of classic HLA proteins (and their medical significance).

11.1 ORIGINS OF DNA SEQUENCE VARIATION

Underlying genetic variation are changes in DNA sequences. Traditionally, the term **mutation** has been used in two ways. It can describe an event or process that produces either a change in the base sequence or in the copy number of a DNA sequence, or it can describe the outcome of that process, the altered DNA sequence.

As events, mutations can occur at a wide variety of levels, and can have different consequences. The altered DNA may contribute to a normal phenotype (such as height) or to a disease phenotype. Very rarely, a mutation has some beneficial effect that produces an altered phenotype conferring a competitive advantage over other phenotypes in the population. The great majority of mutations, however, are neutral: they have no discernible effect on the phenotype. As a result, there has been an increasing trend toward using the more neutral term **DNA variant** instead of *mutation* to describe, in a general way, a DNA change produced by mutation. But DNA variants associated with an altered phenotype continue to be widely described as mutations.

DNA variants originate as a result of changes in our DNA that have not been corrected by cellular DNA repair systems. The DNA changes are occasionally induced by radiation and chemicals in our environment, but the great majority arise from endogenous sources: spontaneous errors in normal cellular mechanisms regulating chromosome segregation, recombination, DNA replication, and DNA repair, plus spontaneous chemical damage to DNA.

Mutations are unavoidable. They may have adverse effects on individual organisms, causing aging and underlying many human diseases. But they also provide the raw fuel for natural selection of beneficial adaptations that allow evolutionary innovation and, ultimately, the origin of new species.

Genetic variation arising from endogenous errors in chromosome and DNA function

Natural errors in various processes that affect chromosome and DNA function—chromosome segregation, recombination, and DNA replication—are important contributors to genetic variation. That happens because no cellular function can occur with 100% efficiency—mistakes happen. Endogenous errors in the above processes may often not have harmful consequences, but some of them make important contributions to disease. We will examine in detail how they can cause disease in later chapters; in this section we take a broad look into how they affect genetic variation in general.

DNA replication errors

Each time the DNA of a human diploid cell replicates, six billion nucleotides need to be inserted in the correct order in the newly synthesized DNA molecules. No enzyme is 100% efficient, and DNA polymerases will occasionally make mistakes, inserting the wrong nucleotide to produce mispaired bases (base mismatches). The likelihood of such an error simply reflects the relative binding energies of correctly paired bases and mispaired bases.

In the great majority of cases, the errors are quickly corrected by the DNA polymerase itself. The major DNA polymerases engaged in replicating our DNA have an intrinsic $3' \rightarrow 5'$

exonuclease activity with a proofreading function. If, by error, the wrong base is inserted, the 3′ → 5′ exonuclease is activated and degrades the newly synthesized DNA strand from its 3′ end, removing the wrongly inserted nucleotide and a short stretch before it. Then the DNA polymerase resumes synthesis again. If mispaired bases are not eliminated by the DNA polymerase, a DNA mismatch repair system is activated (as described below).

Another type of DNA replication error commonly occurs within regions of DNA where there are sequential repeats of a specific nucleotide or short oligonucleotide, such as a run of cytosines or multiple tandem repeats of the CA dinucleotide. When the advancing DNA polymerase encounters a series of short tandem repeats, there will be an increased chance that during DNA replication a mistake is made in aligning the growing DNA strand with its template strand. The two strands can pair-up out of register, causing the growing DNA strand to have fewer or more repeat units (**replication slippage**—see Figure 11.1). Errors like this are also often repaired successfully by the DNA mismatch repair system.

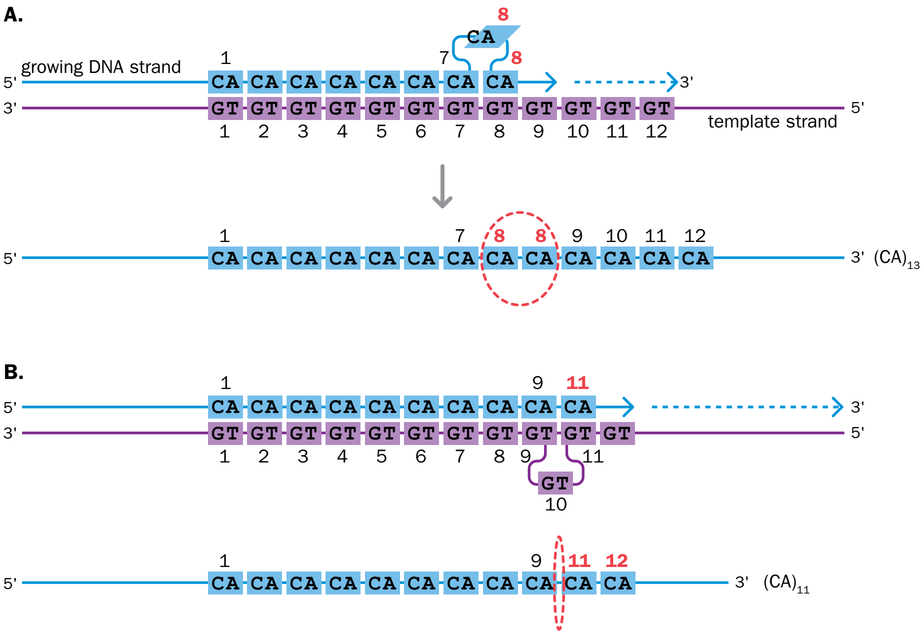


Figure 11.1 Insertion and deletion of one or a few nucleotides are often the result of replication slippage at an array of tandem mononucleotide or oligonucleotide repeats. In this example, the parental DNA strand is envisaged to have had 12 tandem CA repeats. After strand separation, a new DNA strand (the *nascent* or growing strand) is being synthesized, using as a template the DNA strand with the complementary TG repeats. During normal DNA replication, the growing strand often partly dissociates from the template, and then reassociates with the template strand. When there are tandem repeats, mispairing of the growing strand and template strand is facilitated by base pairing between misaligned repeats. (**A**) The mispairing can cause looping-out of the growing DNA strand, inducing an insertion of, in this case, a single CA dinucleotide, to give a $(CA)_{13}$ strand. (**B**) The alternative is that the template strand loops out, inducing a deletion of, in this case, a single CA dinucleotide, to give a $(CA)_{11}$ strand.

Although the vast majority of DNA changes caused by DNA replication errors are identified and corrected, some persist. That happens because although we have many very effective DNA repair pathways, inevitably DNA repair is also not 100% effective, and unrepaired changes in DNA sequence are an important source of mutations. However, the great majority of the mutations introduced in this way do not cause disease: almost 99% of the genome is noncoding DNA, many mutations in coding DNA do not change an amino acid, and functional noncoding DNA sequences (such as those specifying noncoding RNAs) can often tolerate many kinds of sequence change without compromising the function of the sequence.

Chromosome segregation and recombination errors

Errors in chromosome segregation result in abnormal gametes, embryos, and somatic cells that have fewer or more chromosomes than normal, and so have altered numbers of whole DNA molecules. Changes in chromosomal DNA copy number are not uncommon. If they occur in the germ line they often cause embryonic lethality or a congenital disorder (such as Down syndrome, which is commonly caused by an extra copy of chromosome 21); changes in copy number of sex chromosomes are more readily tolerated. In somatic cells, changes in chromosomal DNA copy number are a common feature of many cancers.

Various natural errors can also give rise to altered copy number of a specific sequence within a DNA strand that may range up to megabases in length. That can occur by different recombination and recombination-like mechanisms in which non-allelic (but often related) sequences align so that chromatids are paired with their DNA sequences locally out of register, and subsequent crossover or sister chromatid exchange produces chromatids with fewer or more copies of the sequences, for example. The ensuing duplication or deletion of sequences may or may not have functional consequences—we cover the mechanisms and how they result in disease in Chapter 15.

Various endogenous and exogenous sources can cause damage to DNA by altering its chemical structure

DNA is a comparatively stable molecule, but nevertheless there are constant threats to its integrity, causing breakage of covalent bonds within DNA or inappropriate bonding of chemicals to DNA. Most of the damage originates spontaneously within cells (normal cellular metabolism generates some chemicals that are harmful to cells). A minority of the damage is induced by external sources.

Chemical damage to DNA can involve cleavage of covalent bonds in the sugar-phosphate backbone of DNA causing single-strand or double-strand breaks (**Figure 11.2A**). Alternatively, bases are deleted (by cleavage of the *N*-glycosidic bond connecting a base to a sugar; **Figure 11.2B**), or they are chemically modified in some way.

Many base modifications involve replacing certain groups on bases or adding methyl or larger alkyl groups or other large chemicals (**Figure 11.2C**). Sometimes covalent bonds form between two bases (**base cross-linking**) that may be on the same strand (**Figure 11.2D**, part [i]) or on complementary DNA strands (**Figure 11.2D**, part [ii]). Chemically modified bases may block DNA or RNA polymerases, causing base mispairing and, if not repaired, mutations.

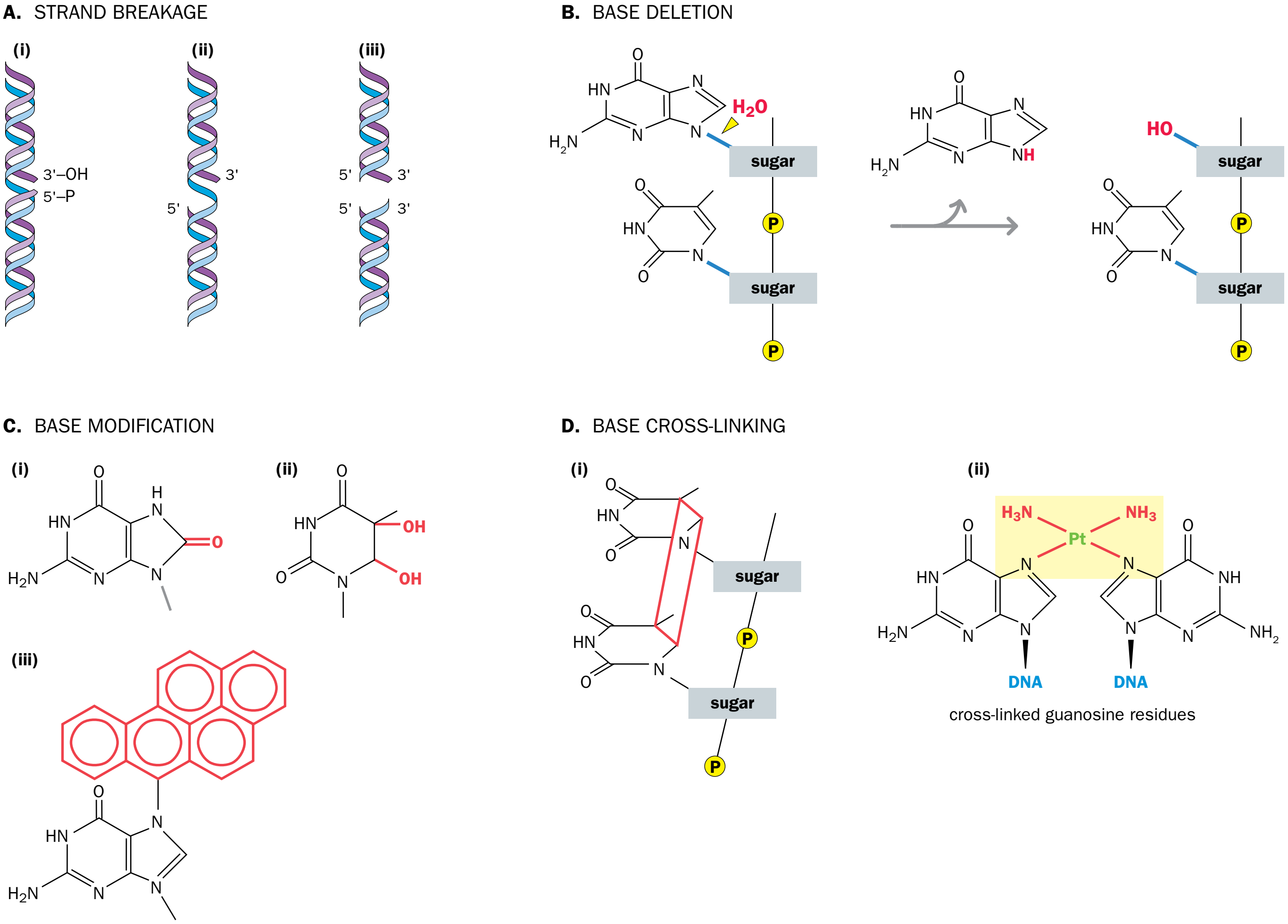


Figure 11.2 Four classes of chemical damage to DNA. (**A**) *DNA strand breakage.* A single strand may be broken by simple cleavage of a phosphodiester bond (i) or by a more complex single-strand break (ii) where the ends are damaged and sometimes one or more nucleotides are deleted. Double-strand DNA breaks (iii) occur when both strands are broken at sites that are in very close proximity. (**B**) *Base deletion.* Hydrolysis cleaves the covalent *N*-glycosidic bond (shown in deep blue) connecting a base to its sugar. (**C**) *Base modification.* Altered bonding or added chemical groups are shown in red. Examples are: 8-oxoguanine (i), which base-pairs to adenine and so induces mutations; thymidine glycol (ii), which is not a mutagen but blocks DNA polymerase; and a DNA adduct (iii) formed by covalent bonding, in this case bonding of an aromatic hydrocarbon, benzo(*a*)pyrene, to N7 of a guanine residue. (**D**) *Base cross-linking.* This involves formation of new covalent bonds linking two bases that may be on the same DNA strand (an intrastrand cross-link) or on complementary DNA strands (an interstrand cross-link). The former includes cyclobutane pyrimidine dimers: linked carbon atoms 4 and 5 on adjacent pyrimidines on a DNA strand (i). This is the most prevalent form of damage incurred by exposure to ultraviolet light from the sun. The anticancer agent cisplatin, $(NH_3)_2PtCl_2$, causes interstrand cross-links by covalently bonding the N7 nitrogen atoms of guanines on opposite strands (ii).

Endogenous chemical damage to DNA

Most of the chemical damage to our DNA arises spontaneously and is unavoidable. Every day, under normal conditions, around 20,000–100,000 lesions are generated in the DNA of each of our nucleated cells. Three major types of chemical change occur, as listed below. Hydrolytic and oxidative damage are particularly significant, breaking various covalent bonds in the nucleotides of DNA (Figure 11.3).

- ***Hydrolytic damage***. Hydrolysis is inevitable in the aqueous environment of cells. It can disrupt bonds that hold bases to sugars, cleaving the base from the sugar to produce an abasic site (see Figure 11.2B); loss of purine bases (depurination) is particularly common. Hydrolysis also strips amino groups from some bases (deamination), leaving a carbonyl (C=O) group. Cytosines are often deaminated to give uracil, which base-pairs with adenine (see left part of Figure 11.7); adenine is occasionally deaminated to produce hypoxanthine, which effectively behaves like guanine by base-pairing with cytosine.
- ***Oxidative damage***. Normal cellular metabolism generates some strongly electrophilic (and, therefore, highly reactive) molecules or ions. The most significant are **reactive oxygen species** (**ROS**) formed by the incomplete one-electron reduction of oxygen, including superoxide anions (O_2^-), hydrogen peroxide (H_2O_2), and hydroxyl radicals ($OH\cdot$). ROS are generated in different cellular locations and play important roles in certain intercellular and intracellular signaling pathways, but they mostly originate in mitochondria (where electrons can prematurely reduce oxygen). Endogenous ROS attack covalent bonds in sugars, causing damage to the sugar–phosphate backbone of DNA. They also attack DNA bases, especially purines (see Figure 11.3), and many derivatives are produced from each base. Some of the base derivatives are highly mutagenic, such as 7,8-dihydro-8-oxoguanine (also called 8-oxoguanine or 8-hydroxyguanine), which base-pairs with adenine; others are not mutagenic but nevertheless block DNA and RNA polymerases (see Figure 11.2C, parts [i] and [ii]).
- ***Aberrant DNA methylation***. Many cytosines in our DNA are methylated by methyltransferases. Cells also use *S*-adenosylmethionine (SAM) as a methyl donor in a nonenzymatic reaction to methylate different types of molecule, but sometimes SAM can inappropriately methylate DNA to produce harmful bases. Each day about 300–600 adenosines in each nucleated cell are converted to 3-methyladenosine, a cytotoxic base that distorts the double helix, disrupting crucial DNA–protein interactions.

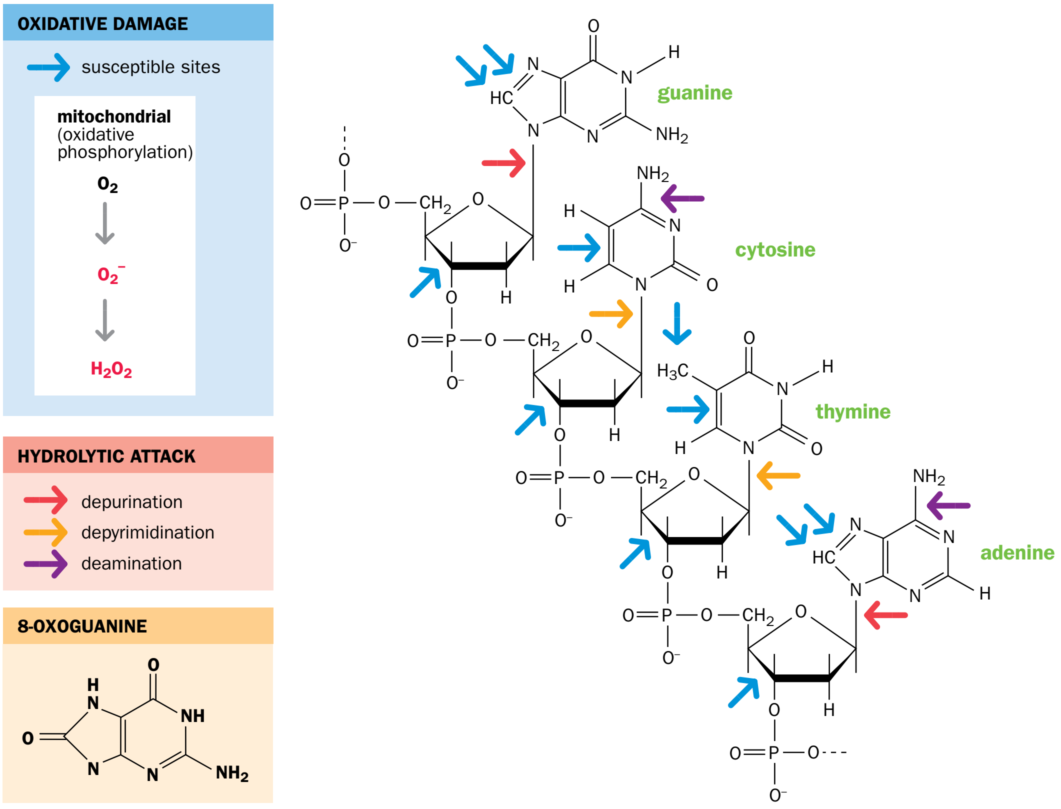


Figure 11.3 DNA sites that are susceptible to spontaneous hydrolytic attack and oxidative damage. Each day, every nucleated human cell loses 5000 or more purines (A and G) and about 300 pyrimidines (C and T) as a result of hydrolytic attack. The *N*-glycosidic bond connecting a base to its deoxyribose sugar is cleaved by H_2O (see **Figure 11.2B** for the reaction). In deamination, an amino group is replaced by an oxygen atom to give a carbonyl group. About 100–500 cytosines are replaced by uracil in each cell every day (see left part of **Figure 11.7**); a smaller number of adenines are replaced by hypoxanthines. In addition, normal metabolism generates reactive oxygen species (see text) that cleave certain chemical bonds not just in bases but also in sugar residues, causing breakage of DNA strands.

Chemical damage to DNA caused by external mutagens

A minority of the chemical damage to our DNA is caused by external agents that can induce mutation (**mutagens**), including radiation and harmful chemicals in the environment. Ionizing radiation (X-rays, gamma rays, and so on) interacts with cellular molecules to generate ROS that break chemical bonds in the sugar–phosphate backbone, breaking DNA strands (see below). Non-ionizing ultraviolet radiation causes covalent bonding between adjacent pyrimidines on a DNA strand (see **Figure 11.2D**, part [i]).

Our bodies are exposed to many harmful environmental chemicals—in our food and drink, and in the air that we breathe. Some chemicals interact with cellular molecules to generate ROS. Other chemicals covalently bond to DNA forming a DNA adduct that may be bulky, causing distortion of the double helix. Cigarette smoke and automobile fumes, for example, have large aromatic hydrocarbons that can bond to DNA (see **Figure 11.2C**, part [iii]). Electrophilic alkylating agents can result in base cross-linking.

11.2 DNA REPAIR

According to the type of DNA damage, cells have different systems for detecting and repairing DNA lesions. Some types of DNA damage may be minor, resulting simply in an altered base, for example. Others, such as DNA cross-linking, may be more problematic and block DNA replication (the replication fork stalls) or block transcription (the RNA polymerase stalls).

Different molecular sensors identify different types of DNA damage, triggering an appropriate DNA repair pathway. If the DNA lesion is substantial and initial repair is not effective, cell cycle arrest may be triggered. The arrest may be temporary (we consider this later in the context of cancer) or be more permanent. In other cases, as often happens in lymphocytes, apoptosis may be triggered.

The DNA repair process occasionally involves simple reversal of the molecular steps that cause DNA damage. This, however, is rare in human cells. Examples include the use of O-6-methylguanine DNA methyltransferase to reverse methylation of guanine at the O6 position, and the use of DNA ligase to repair broken phosphodiester bonds (DNA nicks).

Normally, DNA repair pathways do not directly reverse the damage process. Instead, according to the type of DNA lesion, one of several alternative DNA repair pathways is used. Most of the time, the repair needs to be made to one DNA strand only, but sometimes there is a need to repair both strands, as in the case of interstrand cross-linking (see **Figure 11.2D**, part [ii]) and double-strand DNA breaks (see **Figure 11.2A**, part [iii]).

Errors in DNA replication and chemical damage to DNA are a constant throughout life. Inevitably, however, some mistakes are made in repairing DNA and there are also inherent weaknesses in detecting some base changes (as described below). Inefficiency in detecting and repairing DNA damage is an important contributor to generating mutation. We consider the major DNA repair mechanisms in the next two sections—see **Table 11.1** for a road map of how individual types of DNA changes are repaired.

TABLE 11.1 FREQUENT TYPES OF DNA DAMAGE/ALTERATION AND HOW THEY ARE REPAIRED IN HUMAN CELLS

DNA damage/alteration	DNA repair mechanism	Comments
Base mismatches caused by replication errors	Mismatch repair (**Figure 11.5**)	May be less efficient in condensed, late-replicating DNA, leading to higher frequencies of replication errors
Small insertions/deletions due to replication slippage (**Figure 11.1**)		
Small-scale, single base modification (oxidation, deamination, methylation)	Base-excision repair (BER) (**Figure 11.4A**)	For modified bases, a DNA glycosylase* cuts out a base to produce an abasic site. At all abasic sites, the remaining sugar–phosphate is eliminated and the gap sealed by inserting the appropriate nucleotide
Single base deletion—an *abasic site*—resulting from hydrolysis		
Bulky, helix-distorting DNA lesions (large DNA adducts; DNA intrastrand cross-links)	Nucleotide-excision repair (NER) (**Figure 11.4B**)	Involves removal and re-synthesis of a sequence of several nucleotides spanning the altered site
Single-strand DNA breaks (other than DNA nicks)	Variant of base-excision repair	Initiated by poly(ADP-ribose) polymerase binding to cleavage site
Double-strand DNA breaks	Homologous recombination (HR)-mediated DNA repair (**Figure 11.6**)	Accurate DNA repair but needs an intact homologous DNA strand to use as a template (limited to post-replication in S phase or occasionally G_2)
	Nonhomologous end-joining (NHEJ)	Does not rely on a template strand and so is available throughout cell cycle. Less accurate than HR-mediated DNA repair
DNA interstrand cross-links	HR and Fanconi anemia DNA repair pathway	Some uncertainty about mechanism

* Different DNA glycosylases are specific for different types of modified base, such as uracil DNA glycosylase and N-methylpurine DNA glycosylase.

Repair of DNA damage or altered sequence on a single DNA strand

DNA damage or an error in DNA replication mostly results in one strand having a DNA lesion or a wrongly inserted base but leaves the complementary DNA strand unaffected at that location. In that case, the undamaged complementary strand may be used as a template to direct accurate repair. Different repair mechanisms are available as listed below.

Base-excision repair (BER)

This pathway is specifically aimed at lesions where a single base has either been modified or excised by hydrolysis to leave an abasic site (~20,000 such events occur in each nucleated cell every day). To replace a modified base by the correct one, a specific DNA glycosylase cleaves the sugar–base bond to delete the base, producing an abasic site. For all abasic sites, the residual sugar–phosphate residue is removed using a dedicated endonuclease and phosphodiesterase. The gap is filled using a DNA polymerase (to insert the correct nucleotide) and DNA ligase (**Figure 11.4A**). Note that the same DNA repair machinery occasionally makes a more substantial long-patch repair that involves replacing more than a single base.

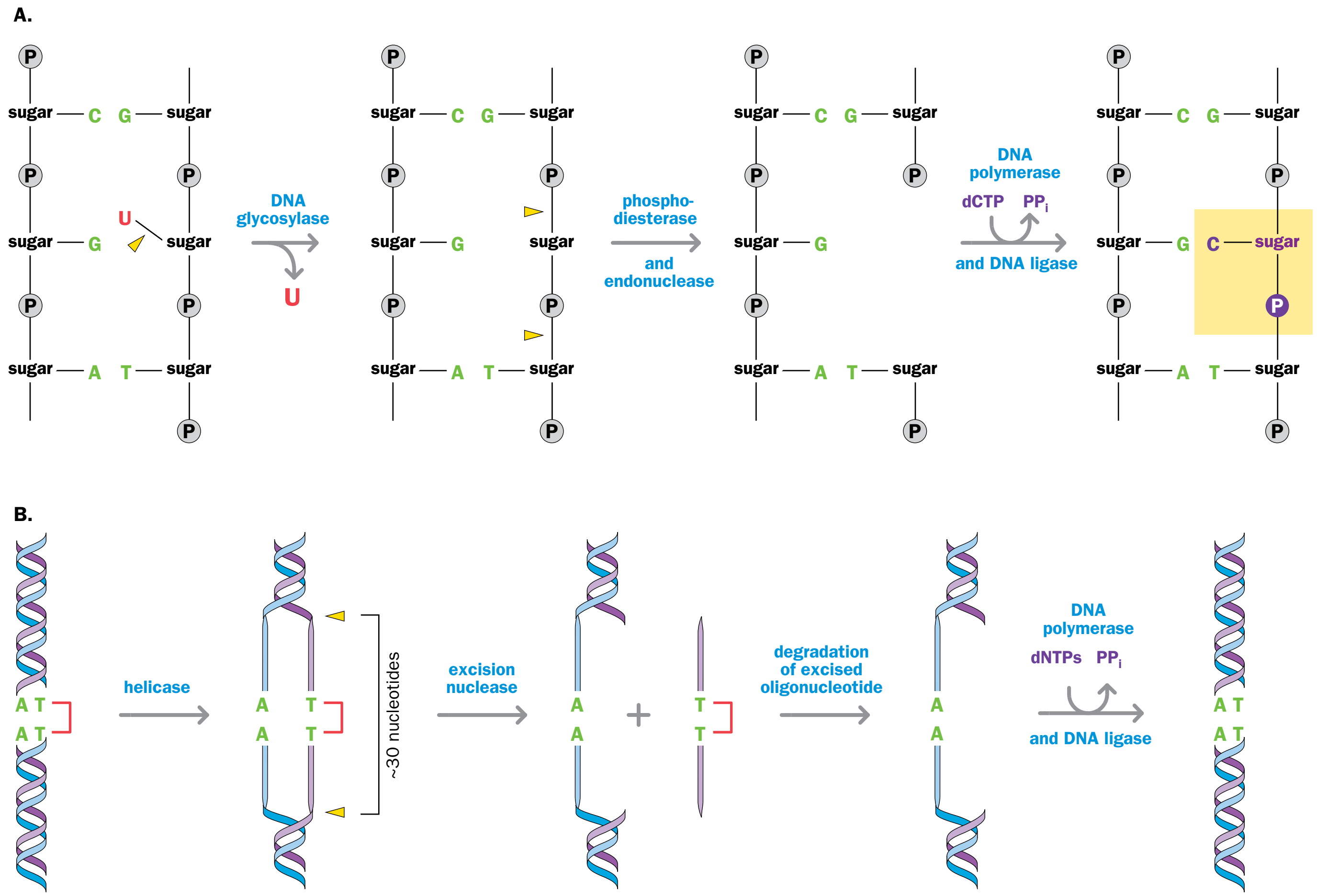


Figure 11.4 Base-excision and nucleotide-excision repair. (**A**) Base-excision repair. This pathway repairs modified bases and abasic sites produced by depurination or depyrimidination. Modified bases are first removed by a DNA glycosylase that cleaves the *N*-glycosidic bond that connects the base to the sugar, producing an abasic site. The cell has a range of different DNA glycosylases that are specific for common modified bases, such as 8-oxoguanine DNA glycosylase and uracil DNA glycosylase (as shown here). Abasic sites are usually repaired by excising the remaining sugar–phosphate residue (using a phosphodiesterase and specialized endonuclease) then inserting the correct nucleotide to match the undamaged complementary DNA strand. (**B**) Nucleotide-excision repair. After identification of a bulky DNA lesion, this type of repair involves opening-out the double helix containing the lesion over a considerable distance (using a helicase to unwind the DNA). Subsequently, an excision nuclease makes cuts on either side of the lesion on the damaged DNA strand, generating an oligonucleotide of about 30 nucleotides containing the damaged site, which is then degraded. The resulting gap is repaired by DNA synthesis, using the undamaged strand as a template, and sealed using a DNA ligase.

Single-strand break repair

Single-strand breaks (SSBs)—also called DNA nicks—are very common. They are easily reversed by DNA ligase but oxidative attack can cause deoxyribose residues to disintegrate, producing more complex strand breakage. A type of base-excision repair is employed: strand breaks are rapidly detected and briefly bound by a sensor molecule, poly(ADP-ribose), that initiates repair by attracting suitable repair proteins to the site. The 3′ or 5′ termini of most SSBs are damaged and need to be restored. The gap is then filled using a DNA polymerase and DNA ligase.

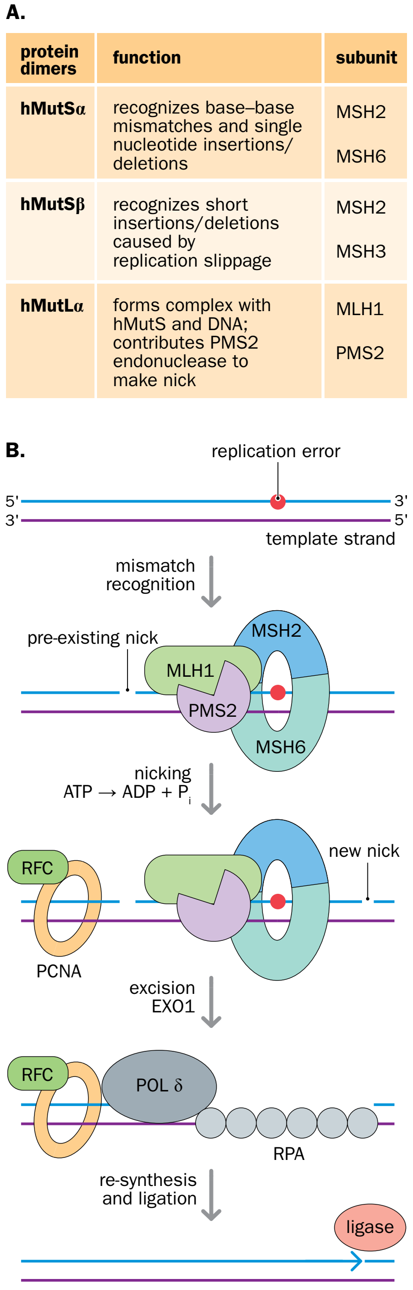
Nucleotide-excision repair (NER)

This mechanism allows repair of bulky, helix-distorting DNA lesions. After the lesion is detected, the damaged site is opened out and the DNA is cleaved some distance away on either side of the lesion, generating an oligonucleotide of about 30 nucleotides containing the damaged site that is discarded. Resynthesis of DNA is carried out using the opposite strand as a template (Figure 11.4B). The priority is to rapidly repair bulky lesions that block actively transcribed regions of DNA. A specialized subpathway, transcription-coupled repair, initiates this type of repair after detection of RNA polymerases that have stalled at the damaged site. Otherwise, an alternative global, genome nucleotide-excision repair pathway is used.

Base mismatch repair

This mechanism corrects errors in DNA replication. The mismatch repair (MMR) components work closely with the DNA replication machinery. In human cells, three types of protein dimer carry out most of the repairs (Figure 11.5A). Two of them—hMutSα and hMutSβ—are needed to identify base mismatches. The former identifies base–base mismatches but can also handle mismatching due to single nucleotide insertions or deletions; hMutS can spot base mismatching for different sizes of very short insertions or deletions (which frequently occur at short tandem repeats as a result of *replication slippage*, the tendency for DNA polymerase to stutter or skip forward at tandem repeats).

The MMR machinery cannot simply repair one of the two strands at random: there has to be a way of distinguishing the original (correct) strand from the newly replicated strand with the incorrect sequence that needs to be repaired. Before being repaired by DNA ligase, nicks (single-strand breaks) are common on a freshly replicated DNA strand, and in human (and other eukaryotic) cells the strand distinction is achieved by identifying a nearby nick on the newly replicated DNA strand. Then hMutLα cleaves the newly replicated strand close to the mismatch and recruits an exonuclease to excise a short stretch of DNA containing the replication error so that the DNA can be re-synthesized and repaired (Figure 11.5B). Errors in base mismatch repair are important in some cancers, as described in Chapter 19.



A.

protein dimers	function	subunit
hMutSα	recognizes base–base mismatches and single nucleotide insertions/ deletions	MSH2 MSH6
hMutSβ	recognizes short insertions/deletions caused by replication slippage	MSH2 MSH3
hMutLα	forms complex with hMutS and DNA; contributes PMS2 endonuclease to make nick	MLH1 PMS2

Figure 11.5 Mismatch repair for correcting replication errors. (**A**) Major classes of MutS or MutL dimers in human mismatch repair. (**B**) Mechanism of 5′-directed mismatch repair in eukaryotic cells. Replication errors on a newly synthesized strand result in base mismatches that can be recognized by a MutS–MutL complex. The MutS component works as a clamp that can slide along the DNA, allowing it to scan for a base–base mismatch (using MutSα, shown here as subunits MSH2 and MSH6) or an unpaired insertion/deletion loop (often using MutSβ). MutLα, which has an endonuclease function, can form a ternary complex with MutS and DNA. After the newly replicated DNA has been identified (by having a pre-existing nick in the DNA), PCNA (proliferating cell nuclear antigen) and RFC (replication factor C) are loaded onto the newly replicated DNA, where they help trigger the endonuclease function of PMS2 to make a new nick close to the replication error. EXO1 exonuclease is recruited to excise the sequence containing the replication error, making a gapped DNA. The resulting stretch of single-stranded DNA (stabilized by binding the RPA protein) is used as a template for the re-synthesis of the correct sequence using high-fidelity DNA polymerase δ (POL δ), followed by sealing with DNA ligase I. (Adapted from Geng H & Hsieh P [2013] In: *DNA Alterations in Lynch Syndrome: Advances in Molecular Diagnosis and Genetic Counseling* [M Vogelsang, ed.]. With permission from Springer Science and Business Media. Copyright © 2013.)

Repair of DNA lesions that affect both DNA strands

Double-strand DNA breaks (DSBs) are normally rare in cells. They do occur naturally, however, and are necessary for specialized DNA rearrangements in B and T cells that maximize immunoglobulin and T-cell receptor diversity (as described below).

DSBs also occur by accident, as a result of chemical attack on DNA by endogenous or externally-induced reactive oxygen species (but at much lower frequencies than single-strand breaks). In these cases, DNA repair is required, but can sometimes be difficult to perform. For example, when the two complementary DNA strands are broken simultaneously at sites sufficiently close to each other, neither base pairing nor chromatin structure may be sufficient to hold the two broken ends opposite each other. The DNA termini will often have sustained base damage and the two broken ends are liable to become physically dissociated from each other, making alignment difficult.

Unrepaired DSBs are highly dangerous to cells. The break can lead to inactivation of a critically important gene, and the broken ends are liable to recombine with other DNA molecules, causing chromosome rearrangements that may be harmful or lethal to the cell. Cells respond to DSBs in different ways. Two major DNA repair mechanisms can be deployed to repair a DSB, as listed below, but if repair is incomplete, apoptosis is likely to be triggered.

- ***Homologous recombination (HR)-mediated DNA repair***. This highly accurate repair mechanism requires a homologous intact DNA strand to be available to act as a template strand. Normally, therefore, it operates after DNA replication (and before mitosis), using a DNA strand from the undamaged sister chromatid as a template to guide repair (**Figure 11.6**). It is important in early embryogenesis, when many cells are rapidly proliferating, and in the repair of proliferating cells; it occurs after the DNA has replicated, notably in S phase.
- ***Nonhomologous end-joining (NHEJ)***. No template strand is needed here: the broken ends are simply fused together quickly. Specific proteins bind to the exposed DNA ends and recruit a special DNA ligase, DNA ligase IV, to rejoin the broken ends. Unlike HR-mediated DNA repair, NHEJ is, in principle, always available to cells, but it is most important for the repair of differentiated cells and of proliferating cells in G_1 phase before the DNA has replicated. We describe the mechanism in Chapter 15.

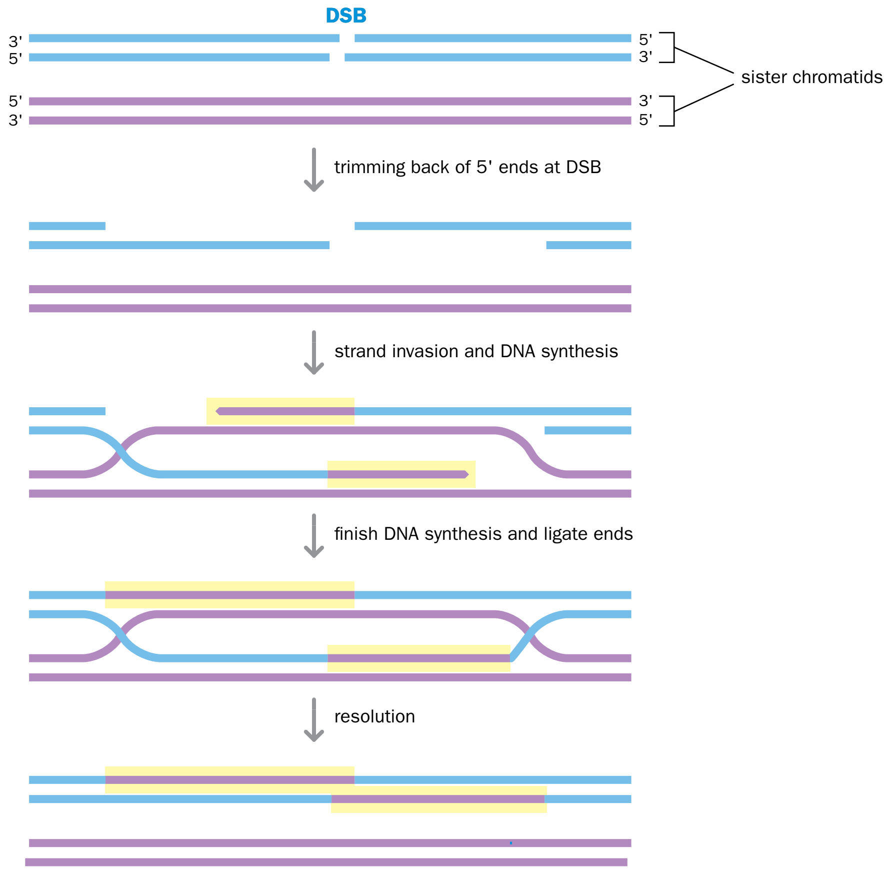

DSB
3'
5'
5'
3'
5'
3'
3'
5'
sister chromatids
trimming back of 5' ends at DSB
strand invasion and DNA synthesis
finish DNA synthesis and ligate ends
resolution

Figure 11.6 Homologous recombination-mediated repair of double-strand DNA breaks. The double-strand break (DSB) in the chromatid at top is repaired using as a template the undamaged DNA strands in the sister chromatid (*note*, to make the mechanism easier to represent, the upper chromatid is shown in an unusual format; the 3′ → 5′ strand is placed above the 5′ → 3′ strand). The first step is to cut back the 5′ ends at the double-strand break to leave protruding single-strand regions with 3′ ends. Following strand invasion, each of the single-strand regions forms a duplex with an undamaged complementary DNA strand from the sister chromatid, which acts as a template for new DNA synthesis (newly synthesized DNA copied from the sister chromatid DNA is highlighted in yellow). Following DNA synthesis, the ends are sealed using DNA ligase. The repair is highly accurate because for both broken DNA strands, undamaged sister chromatid DNA strands act as a template to direct incorporation of the correct nucleotides during DNA synthesis.

Repair of DNA interstrand cross-links

Cross-linking can occur between bases on complementary strands of a double helix, either as a result of endogenous metabolites or through exogenously supplied chemicals, notably many anticancer drugs (see the example of cisplatin in **Figure 11.2D**, part [ii]). Interstrand cross-links seem to be repaired using a combination of nucleotide-excision repair, translesion synthesis, homologous recombination, and a complex of multiple different protein subunits that are encoded by genes mutated in Fanconi anemia (Fanconi anemia DNA repair pathway).

Undetected DNA damage, DNA damage tolerance, and translesion synthesis

DNA damage may sometimes go undetected. For example, cytosines that occur within the dinucleotide CpG are highly mutable as a result of inefficient DNA repair. In vertebrates, the dinucleotide CpG is a frequent target for DNA methylation, converting the cytosine to 5-methylcytosine (5-meC). Deamination of cytosine residues normally produces uracil, which is efficiently recognized as a foreign base in DNA, but deamination of 5-methylcytosine produces thymine that may go undetected as an altered base (**Figure 11.7**). As a result, C→T substitutions are the most frequent type of single nucleotide change in our DNA.

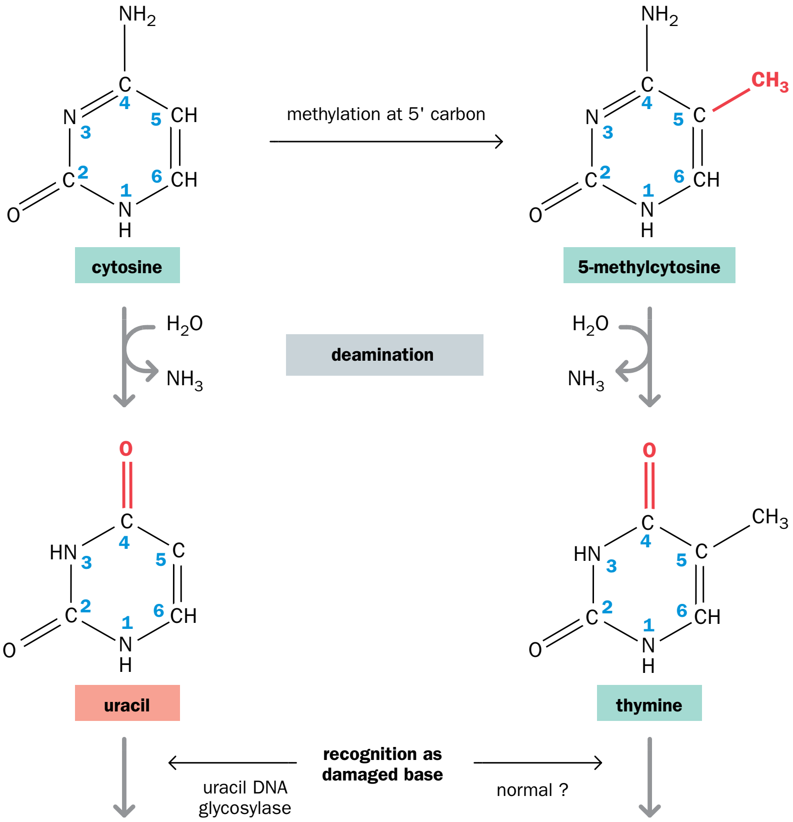


Figure 11.7 Why C→T transitions are so common in human DNA. Deamination of cytosine is a common event in our cells and normally produces uracil, a base that is usually found in RNA, not DNA. A dedicated enzyme, uracil DNA glycosylase, recognizes uracil residues in our DNA and removes them as part of the base-excision DNA repair pathway (see **Figure 11.4A**). However, as in the DNA of other vertebrates, many of our cytosines are methylated at the 5′ carbon. Deamination of 5′-methylcytosine produces thymine, a base normally found in DNA. Although a stable C-G base pair has been replaced by a T-G base mismatch, the base mismatch may often escape detection by the base mismatch repair system (which focuses on DNA replication events). At the subsequent round of DNA replication, the thymine will form a T-A base pair, effectively producing a C→T mutation.

Sometimes, DNA lesions may be identified but not repaired before DNA replication (damage tolerance). For example, DNA lesions that block replication may be bypassed rather than repaired, and nonclassic DNA polymerases are required to resume DNA synthesis past the damaged site (**translesion synthesis**). Subsequently, the gap in the daughter strand opposite the lesion is filled-in, and later on the lesion can be repaired using the daughter strand as a template in nucleotide-excision repair. The nonclassic DNA polymerases used in translesion synthesis exhibit a low fidelity in DNA replication (**Table 11.2**). They have a higher success in incorporating bases opposite a damaged site, but they are prone to error by occasionally inserting the wrong base. As a result, replication forks are preserved, but quite often at the cost of mutagenesis.

TABLE 11.2 NONCLASSIC DNA-DEPENDENT MAMMALIAN DNA POLYMERASES

DNA polymerases	Roles
ζ (zeta); η (eta); Rev1	Translesion synthesis[a] (to bypass obstructive DNA lesion)
ι (iota); κ (kappa)	Translesion synthesis[a] plus other roles in DNA repair
ν (nu)	Interstrand cross-link repair?
θ (theta)	Translesion synthesis[a]; base-excision repair; somatic hypermutation[b]
λ (lambda); μ (mu)	V(D)J recombination[c]; double-strand break repair; base-excision repair
Terminal deoxynucleotide transferase	V(D)J recombination[c]

Classic DNA-dependent DNA polymerases, notably the α and δ DNA polymerases, are mostly used in general DNA synthesis and DNA repair, and have very low error rates. By contrast, the nonclassic polymerases listed here exhibit comparatively low fidelity of DNA replication and are reserved for specialized mechanisms. They include DNA repair mechanisms used to bypass damaged DNA, and also B-cell-specific and T-cell-specific mechanisms that are employed to make diverse immunoglobulins and T-cell receptors.

[a] Used to bypass an obstructive DNA lesion.

[b] A specialized mechanism used in B cells to maximize the variability of immunoglobulins.

[c] A specialized mechanism used in both B and T cells to maximize the variability of, respectively, immunoglobulins and T-cell receptors.

11.3 POPULATION GENOMICS AND THE SCALE OF HUMAN GENETIC VARIATION

Genetic variation is caused by changes to the base sequence. As described below, they can be placed into different variant classes, but all sequence changes can be classified within two broad categories.

- ***Changes that do not affect the DNA content.*** Here the number of nucleotides is unchanged. Quite often a single nucleotide is replaced by a different nucleotide. More rarely, multiple nucleotides at a time may be sent to a new location (by chromosome breakage and rejoining) without net loss or gain of DNA content. The great majority of the latter events are balanced translocations and inversions; they often have no effect on the phenotype, but can be harmful if the change significantly alters the expression of a gene at, or close to, a chromosome breakpoint.
- ***Changes in copy number of a nucleotide or DNA sequence.*** Simple deletion/insertion of a unique sequence can change the haploid copy number from, respectively, 1 to 0 or 0 to 1. Or there may be a change in copy number of a repeated DNA sequence, resulting in fewer or more copies of that sequence. The result is a net loss or gain of a DNA sequence that may be large or small. At one extreme are copy number changes resulting from abnormal chromosome segregation, producing fewer or more chromosomes than normal, and therefore a change in the copy number of whole nuclear DNA molecules. They are almost always harmful: many result in spontaneous abortion, and some give rise to developmental syndromes. At the other end of the scale is deletion or insertion of a single nucleotide. In between are copy number changes for sequences that range in size from specific short oligonucleotide sequences to DNA segments hundreds of kilobases long.

DNA variants, polymorphisms, and population genomics

Alternative forms of DNA produced by mutation are generally known as DNA variants. In the past, it was traditional to describe a common DNA variant, with a population frequency >0.01, as a **polymorphism**; DNA variants with frequencies <0.01 were traditionally described as *rare variants*. The 0.01 cut-off was an arbitrary one, however (it was initially proposed in order to exclude recurrent mutation). In the HapMap project—which produced detailed genetic maps based on polymorphisms in which a single nucleotide is changed (**a single nucleotide polymorphism** or **SNP** [pronounced “snip”])—a *common* SNP was defined as a SNP where there was a minority DNA variant with a frequency >0.05.

The general use of the term *polymorphism* is now declining, partly because of the arbitrary nature of the 0.01 or 0.05 cut-offs, and partly because of ambiguity in how the term is used. In medical disciplines, for example, *polymorphism* is often used to denote any sequence variation that does not cause disease, whereas *mutation* is used to describe a disease-causing sequence variant. In modern times, the term *polymorphism* is largely avoided in population genomics projects; instead, DNA variants are often classified as: common (frequency >5%); low frequency (from 0.5% to 5%); and rare (<0.5%). The preference now, for example, is to describe a change to a single nucleotide as a single nucleotide variant.

Detailed knowledge of human genetic variation comes from comprehensive analysis of DNA sequences from multiple individuals. The Human Genome Project produced a human genome reference sequence that was a patchwork composite: sequences present in different genome regions came from different human donors (detailed in **Box 7.3**). As next-generation sequencing technologies became available, whole genome sequences could readily be obtained from individuals. The era of population-based genome sequencing began in 2008 with the 1000 Genomes Project that began to sample variation in 26 different human populations across the world. Follow-up projects have included studies of national populations and of individuals with genetic disorders and their families (see **Table 11.3**). We describe the outcome of the completed 1000 Genomes Project later in this section.

TABLE 11.3 EXAMPLES FROM THE FIRST WAVE OF HUMAN POPULATION GENOMICS PROJECTS

Project	Genomes sequenced	Publications
1000 Genomes (www.internationalgenome.org)	Genomes of, eventually, 2504 individuals from diverse ethnic groups. Two-parent-and-child trios allowed assessment of the frequency of *de novo* mutations	PMID 26432245 and 26432246
Icelander Genomes	Genomes of 2636 Icelanders by the deCode company	PMID 25807286
Genome of the Netherlands (www.nlgenome.nl/)	Genomes of 769 individuals from 250 Dutch families	PMID 24974849
ToMMo Japanese Reference Panel Project	Genomes of 1070 healthy Japanese individuals	PMID 26292667
Human Longevity Inc Genomes Project	Genomes of 10,545 people from different ethnic populations at 30× to 40× genome coverage	PMID 27702888
Wellcome Trust UK10K (www.uk10k.org/)	Genomes of 4000 people from TwinsUK study/ALSPAC study (individuals followed since birth in 1991–92), plus 6000 people with a severe, genetically predisposed condition	Interim report: PMID 26367797
100,000 Genomes Project (https://www.genomicsengland.co.uk)	Genomes of 70,000 people (UK National Health Service [NHS] patients with a rare disease and their families), plus 25,000 genomes of tumors from cancer patients	Progress report: PMID 29720373

PMID, PubMed identifier. Exome projects are covered in Section 11.4.

Small-scale genetic variation and structural variation

The vast majority of DNA changes are errors of DNA replication and repair, and typically affect one or a very small number of nucleotides (>99.9% of the DNA changes are accounted for by base substitutions and small-scale insertions and deletions). Small-scale changes like this often have no obvious effect on the phenotype (neutral mutations). That happens because significantly more than 90% of our DNA is poorly conserved and of questionable functional value to the organism, appearing to tolerate DNA sequence changes without effect.

Because of the predominance of small-scale mutations, the study of human genetic variation was largely focused on small-scale (<50 bp) variation until quite recently. However, additional moderate- and large-scale DNA changes (that include specific types of DNA breakage and rejoining mechanisms) are also highly significant. Such **structural variation** can involve very large changes, and although structural variants are infrequent, significantly more nucleotides across the genome are altered as a result of structural variation than as a result of small-scale mutations.

The borderline between small-scale genetic variation and structural variation is an arbitrary one. In the past, structural variation used to be applied to sequences that were one kilobase or longer, but the modern tendency has been to include smaller variants. In the 1000 Genomes Project, for example, changes affecting 1–50 nucleotides were classified as small-scale, and anything larger was classified as structural variation (because commonly used massively parallel DNA sequencing methods used to produce quite short sequences, and were often not suited to detecting deletions/insertions of greater than 50 nucleotides).

Nomenclature and databases

The recommendations of the Human Genome Variation Society (HGVS) for how to describe sequence variants were most recently updated in 2016. **Table 11.4** provides illustrative examples of the variation nomenclature. **Table 11.5** lists some of the general human DNA variation databases. There are also many databases that relate variants to clinical information, as described in later chapters.

TABLE 11.4 HUMAN GENETIC VARIATION NOMENCLATURE: HGVS DEFINITIONS OF VARIATION CLASSES WITH EXAMPLES OF VARIANTS

DNA change	Definition	Example
Substitution (>)	A change where one nucleotide is replaced by one other nucleotide	g.1318G>T
Deletion (del)	A change where one or more nucleotides are not present (deleted)	g.3661_3706del
Inversion (inv)	A change where more than one nucleotide replaces the original sequence and is the reverse-complement of the original sequence (e.g., CTCGA to TCGAG)	g.495_499inv
Duplication (dup)	A change where a copy of one or more nucleotides is inserted directly 3′ of the original copy of that sequence	g.3661_3706dup
Insertion (ins)	A change where one or more nucleotides are inserted in a sequence and where the insertion is not a copy of a sequence immediately 5′	g.7339_7340insTAGG
Conversion (con)	A specific type of deletion/insertion where a range of nucleotides replacing the original sequence is a copy of a sequence from another site in the genome	g.333_590con1844_2101
Deletion/insertion (delins/indel)*	A change where one or more nucleotides are replaced by one or more other nucleotides and which is not a substitution, inversion, or conversion	g.112_117delinsTG

Under the Definition heading, the phrase "A change where" should be read as "A change where in a specific sequence compared to the reference sequence." In the examples, the prefix g. denotes a recognized subgenomic reference sequence that may be the sequence of a whole chromosome such as the X chromosome (NC_000023.11), the mitochondrial DNA sequence, or a recognized reference sequence for an individual locus (such as a gene locus within the RefSeq database at https://www.ncbi.nlm.nih.gov/refseq).

* Note that *delins* or *indel* gives a precise but narrow meaning, but *indel* has been consistently used in a much broader way in genome sequencing projects (see Text and **Figure 11.8**). HGVS, Human Genome Variation Society. For more information on nomenclature, see the dedicated HGVS nomenclature Website at http://www.hgvs.org/mutnomen/. (Reproduced from den Dunnen JT *et al*. [2016] *Hum Mutat* **37**:564–569; PMID 26931183.)

TABLE 11.5 GENERAL HUMAN GENETIC VARIATION DATABASES

Database	Description	Website
dbSNP	Covers single nucleotide variations, microsatellites, and small-scale insertions and deletions. It contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for neutral variants and clinical mutations	http://www.ncbi.nlm.nih.gov/SNP/index.html
dbVAR	A database of genomic structural variation. It stores information associated with large-scale genomic variation (including large insertions, deletions, translocations, and inversions). It also stores associations of defined variants with phenotype information	http://www.ncbi.nlm.nih.gov/dbvar/
DGV	Database of Genomic Variants. A curated collection of structural variation in the human genome	http://dgv.tcag.ca
ALFRED	The Allele Frequency Database. Has information on allele frequencies in over 700 human populations	https://alfred.med.yale.edu/alfred/index.asp

Databases focusing on mutations that relate to phenotypes and disease are described in later chapters.

Small-scale variation: single nucleotide variants and small insertions and deletions

Base substitution is the most common type of point mutation and results in **single nucleotide variants (SNVs)**. For example, most sampled sequences might have a G at a defined position, but a minority might have a T at that position. (Previously, where a minor DNA variant has a population frequency of 0.01 or more, the DNA variation has traditionally been described as a single nucleotide polymorphism or SNP.) As outlined in **Box 11.1**, the pattern of single nucleotide variation in the human genome is nonrandom for various reasons.

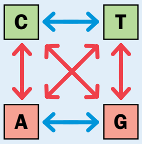
BOX 11.1 THE PATTERN OF SINGLE NUCLEOTIDE VARIATION IN THE HUMAN GENOME IS NONRANDOM FOR MULTIPLE REASONS

Various factors cause different regions and DNA sequences in the human genome to undergo different mutation rates. First, the human genome is divided between two cellular compartments with different environments. Mitochondrial DNA has a mutation rate that is at least one order of magnitude greater than the mutation rate of nuclear DNA (partly because of proximity to the high levels of damaging reactive oxygen species generated in mitochondria).

In the case of base substitutions, transitions are much more common than might have been expected (**Figure 1**). There is also a general mutational bias toward A-T base pairs (as also observed in a wide range of eukaryotic and prokaryotic species). Chromatin structure has an effect. Open chromatin is replicated earlier, and with higher accuracy, than regions of condensed chromatin, which tend to replicate later and under tighter time constraints (the condensed structure may also possibly afford reduced access to the mismatch repair machinery, making replication errors more frequent).

Single nucleotide variation is also nonrandom across the genome because genes and other functional DNA sequences are subject to natural selection. As detailed in Section 11.4, natural selection can have different effects on single nucleotide variation in functionally important DNA sequences, but the most common type of natural selection results in suppression of genetic variation.

A final reason for nonrandom variation in germ-line DNA comes from our evolutionary ancestry. One might reasonably wonder why just certain nucleotides should be polymorphic and be surrounded by stretches of nucleotides that only rarely show variants. In general, the nucleotides found at SNP sites are not particularly susceptible to mutation (the germ-line single nucleotide mutation rate is about 1.2×10^{-8} per generation, roughly 1 nucleotide per 100 Mb, and SNPs are stable over evolutionary time). Instead, the alternative nucleotides at SNP sites simply mark alternative ancestral chromosome segments that are common in the present-day population.



Box 11.1 Figure 1 Transversions should theoretically be twice as frequent as transitions. For clarity, pyrimidines and purines are depicted within pale green and pale pink boxes, respectively. Blue arrows represent transitions (replacement of one pyrimidine by another, or of one purine by another); red arrows represent transversions (replacement of a pyrimidine by a purine, or of a purine by a pyrimidine). There are four possible transitions (C → T, T → C, A → G, and G → A) and eight possible transversions. In the nuclear genome, however, transitions are twice as common as transversions (see text), and in our mitochondrial genome the ratio is biased even more toward transitions (see Kivisild T [2015]; PMID 25798216).

Small insertions and deletions frequently arise as a result of replication errors. They are especially frequent in regions containing repetition of a single nucleotide (a *homopolymer*) or tandem repetition of short oligonucleotide sequences, where they arise as a result of replication slippage. They can often readily be identified when comparing a human genome sequence against the reference human genome sequence. However, the sequences of some variants indicate that they are related by a combination of insertion and deletion events at the same position, such as that shown in **Figure 11.8**.

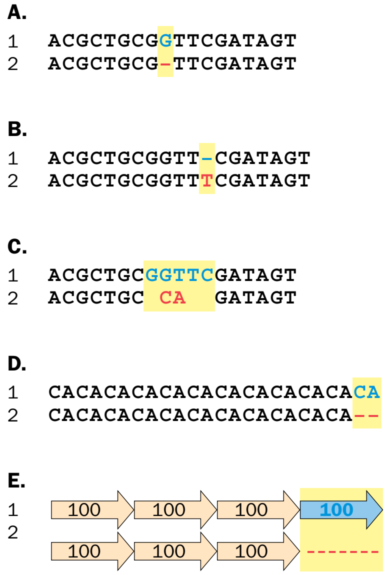


Figure 11.8 Different meanings for the terms *indel* and *copy number variant*. In each case, imagine that sequence 1 derives from the human genome reference sequence and sequence 2 is a test sequence under comparison. Boxes in color shading frame the differences: blue, reference sequence; red, test sequence. In (**A**) we might infer a deletion of a G, and in (**B**) an insertion of a T. But in (**C**) the situation is more complex, suggesting that variant 2 is related to variant 1 by a compound event involving deletion (of GGTTC) plus insertion (of CA). This type of variant, classified by the Human Genome Variation Society as an *indel*, might be expected to be quite rare. However, the term *indel* is also used widely in evolutionary genetics as a broad term to include all sites that show insertions, duplications, or deletions. In genome sequencing projects, *indel* is also used in the broad sense but is further qualified by being restricted to small size changes only (up to 50 nucleotides in the 1000 Genomes Project), sometimes described as short indels. The term *copy number variant* used to be applied to all variants that had a variable number of tandem repeats, including short tandem repeats, such as the microsatellite in (**D**) where there are 12 or 11 copies of the CA dinucleotide. In genome sequencing projects, the term is reserved for large size changes only, such as variable numbers of repeats exceeding 50 nucleotides in the case of the 1000 Genomes Project. In that case, the test sequence variants shown in (**A**) through (**D**) would be classified as (short) indels, but the deletion of 100 nucleotides in (**E**) would be classified as a copy number variant (CNV) and represents a type of structural variation.

The type of variation shown in **Figure 11.8C** has been described as an *indel* (or *delins*) but the term **indel** has also been widely used in a much broader sense to include all types of DNA change that cause a size change at a specific position: insertions, duplications, deletions, and compound insertion/deletion. (In evolutionary genetics, if species 1 has TTTT at a specific position, and the counterpart in closely related species 2 is TTTTT, it may be difficult to decide whether there has been an insertion or deletion of a T without having DNA from a common ancestor.) In population genome sequencing projects, such as the 1000 Genomes Project, the term indel is reserved for small insertion/deletion events involving 1–50 nucleotides (the size restriction is sometimes emphasized by using the term *short* indel).

Short tandem repeat variants/polymorphisms (also known as **microsatellites**) vary in size by small numbers of nucleotides, and so although they vary in the copy number of repeats, they are classified as indels. In genome sequencing projects, the term *copy number variant* (*CNV*) is often now reserved for variants that differ in copy number of larger (>50 bp) repeat sequences (see **Figure 11.8**), and is included as a component of structural variation, as described in the next section.

Structural variation: inversions, translocations, large-scale insertions and deletions, and copy number variants

In the past, structural variation was concerned with large sequence changes typically 1 kb or more, but in the era of rapid whole-genome sequencing it has come to include smaller changes, typically ones that involve changes to more than 50 nucleotides at a time.

In balanced structural variation, the DNA variants have the same DNA content but differ in that some DNA sequences are located in different positions within the genome. Here the variants are formed when DNA molecules sustain double-strand breaks and the resulting fragments are incorrectly rejoined, but without loss or gain of DNA. There are two significant mechanisms. In **inversions**, two double-strand breaks on a DNA molecule release a central fragment that rotates through 180° before being rejoined to the end fragments. Alternatively, **translocations** occur in which two different chromosomal DNA molecules each receive a double-strand break and exchange fragments without any change in DNA content (**Figure 11.9A**).

In unbalanced structural variation, the DNA variants differ in DNA content. Rare cases occur where a person has gained or lost certain chromosomal regions (as when a parent with a balanced reciprocal translocation passes one of the two translocation chromosomes to a child); the gain or loss of substantial chromosomal segments often results in disease. However, by far the most common cause of unbalanced structural variation occurs when variants differ in the number of copies of a moderately long (at least >50 nucleotides) to very long DNA sequence (**Figure 11.9B**).

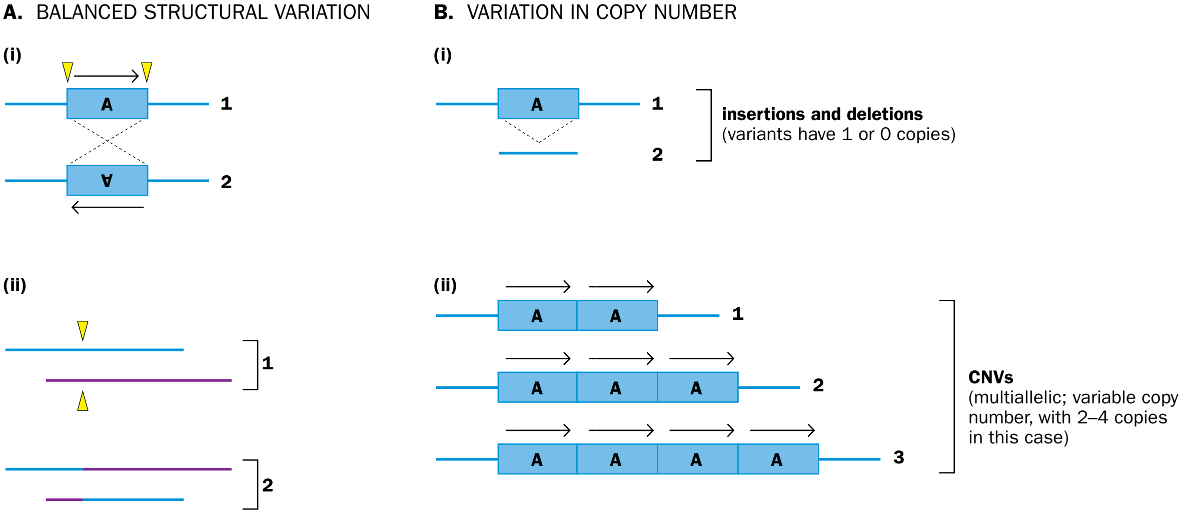

A. BALANCED STRUCTURAL VARIATION
(i)
A
1
A
2
(ii)
1
2
B. VARIATION IN COPY NUMBER
(i)
A
1
2
insertions and deletions
(variants have 1 or 0 copies)
(ii)
A A
1
A A A
2
A A A A
3
CNVs
(multiallelic; variable copy number, with 2–4 copies in this case)

Figure 11.9 The most common forms of structural variation. The numbers 1, 2, and 3 refer to allelic variants. (**A**) Balanced structural variation involves large-scale changes that produce variants with the same number of nucleotides, including many inversions (i) and balanced translocations (ii). (**B**) Unbalanced structural variation includes rare unbalanced translocations (not shown) plus different types of low-copy-number variation. The latter includes variants that differ by having one copy or zero copies as a result of large-scale insertion (for example of a mobile element) or large-scale deletion (i). Another subclass differs in possessing variable numbers of a moderately long sequence (represented by the box marked A). Such variants, called copy number variants (CNVs), arise by tandem duplications. Note that tandem duplication of a sequence is quite often followed by additional insertion and inversion events, resulting in interspersed duplication with normal or inverted orientation of the copies (not shown here).

The variation in copy number can take different forms. One class involves large-scale insertions and deletions so that DNA variants either lack a specific sequence (0 copies) or possess that sequence (1 copy—see **Figure 11.9B**, part [i]). Other forms result from tandem duplication of sequences greater than 50 nucleotides in length (**Figure 11.9B**, part [ii]) that may be complicated by subsequent insertion and/or inversion events. Variants arising in this way are described as **copy number variants** (**CNVs**) and quite often have multiple alleles. In some CNVs the DNA sequence that varies in copy number can include part of a gene sequence or regulatory sequences and sometimes multiple genes. As a result, some CNVs are important contributors to disease, as described in later chapters, but many common CNVs do not affect gene function and are not implicated in disease.

Common types of insertion

The most common type of large-scale insertion in germ-line DNA arises through retrotransposition: transcripts of members of highly repetitive interspersed repeat families are converted into cDNA copies (using reverse transcriptases) and then integrate into a new site in the genome (see **Box 9.2** for the principle of retrotransposition). Most of the inserts arise from SINE families (notably Alu repeats) and LINE families (notably the LINE-1 = L1 repeat), but some also arise from SVA (SINE–VNTR–Alu) repeats. (The structures of these repeats are given in **Figure 9.13**.)

Some insertions also arise after copies of mitochondrial DNA sequences are inserted into the nuclear genome (nuclear mitochondrial insertions or NUMTs; see **Table 9.3** for some examples). The reference human genome has over 700 sequence copies of segments of human mtDNA, but variation arises because of ongoing insertion events.

Whole-genome sequencing allows direct measurement of germ-line mutation rates and comprehensive analysis of germ-line DNA variation

Advances in DNA sequencing technology have meant that personal genomes can now be sequenced quite rapidly. By sequencing large numbers of human genomes, including those of parents and their children, it has been possible to acquire detailed information on human germ-line mutation rates, the extent of human genetic diversity, and the degree to which haploid human genomes differ from each other.

Measuring human germ-line mutation rates

The human germ-line mutation rate has long been a subject of intense interest, but until recently all estimates were achieved using indirect methods. Long before the structure of DNA was discovered, human geneticists were able to infer the spontaneous mutation rate by studying how some highly penetrant single gene disorders are inherited. Reporting a study of hemophilia in 1935, JBS Haldane gave the first careful, but approximate, estimate of the human mutation rate: between 1 and 5×10^{-5} per locus per generation (but nothing was known about the structure of the locus). Such a study was made possible by a balance between mutation, which introduces new pathogenic DNA variants, and natural selection, which removes them. (We introduce aspects of natural selection in Section 11.4, but the mutation–selection balance for inherited disorders will be covered in Chapter 12; interested readers can find an electronic version of the Haldane paper, reprinted in 2004, at PMID 15689625.)

After the emergence of molecular genetics, new indirect methods of inferring human mutation rates became possible. In particular, phylogenetic methods were extensively used to look at the rate of base substitution in DNA sequences that might not be subject to natural selection, such as transcriptionally inactivate pseudogenes. That meant comparing human DNA and protein sequences with their counterparts in closely related species, notably chimpanzees or other apes, and then dividing the number of differences by the estimated time since the last common ancestor. This indirect approach assumes a constant mutation rate in the separate lineages descending from the common ancestor, and is disadvantaged by difficulties in correctly dating the last common ancestors (the fossil record is incomplete). Until very recently, the consensus estimate was that, on average, each nucleotide mutates once every billion years. If we were to assume a historical 30-year generation time, that gives a mutation rate of 3×10^{-8} mutations per nucleotide per generation. Recent comparisons between the DNA of modern and ancient humans suggest a rate less than one-half of the one above, supporting the idea that human mutation rates may have changed in evolution.

The indirect approaches above have recently been superseded by comprehensive methods for directly measuring germ-line mutation rates. As whole-genome sequencing took off, various studies have included whole-genome analyses of family groups, permitting identification of *de novo* germ-line mutations. (Family trios, composed of a child plus the two biological parents, have been widely investigated, but some studies have investigated multisibling families.) The mean genome-wide human germ-line base-substitution rate is now considered to be roughly 1.0–1.2 × 10^{-8} per nucleotide per generation.

Variation in germ-line mutation rates

The estimate for the base-substitution rate is the headline figure in human germ-line mutation, simply because base substitutions are so common. As expected, there are significant differences in the germ-line mutation rate for different components of the genome (**Table 11.6**). The germ-line mutation rate can also vary between populations and families, and there is a clear sex difference: the male germ-line mutation rate is significantly higher than the female rate, and increases with paternal age (see **Box 11.2**).

TABLE 11.6 GERM-LINE MUTATION RATES FOR DNA VARIATION CLASSES

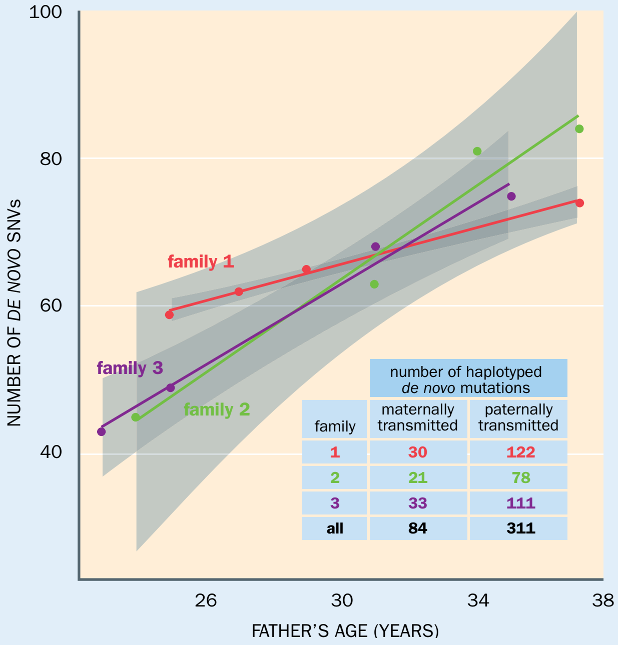
Type of DNA change		Human germ-line mutation rate
Base substitution	Across all autosomes	~1.2 × 10^{-8} per nucleotide per generation
	In mtDNA	>10× increase over genome-wide rate
	CpG transitions	~10–18× increase over genome-wide rate
Indels	1 bp deletion	~3.2 × 10^{-10} per site per generation
	1 bp insertion	~1.1 × 10^{-10} per site per generation
	Short (2–4 bp) tandem repeat	~2.7–10 × 10^{-4} per site per generation
Copy number variation (>50 ntds repeat)		~2.5 × 10^{-6}–1.0 × 10^{-4} per site per generation

BOX 11.2 SEX DIFFERENCES AND PARENTAL AGE EFFECTS IN GERM-LINE MUTATION RATES

Most mutations arise as a result of unrepaired replication errors. However, the number of human male germ-cell divisions greatly outnumbers female germ-cell divisions. And because each mitotic cell division is preceded by DNA replication, the male germ-line mutation rate might be expected to greatly exceed the female germ-line rate.

Using the estimates reported by Drost and Lee (1995) (PMID 7789362), the journey from human zygote to primary oocyte requires approximately 31 cell divisions. In males, by contrast, 34 cell divisions are required for germ-cell development before spermatogenesis; then, after puberty, sperm cells are continuously produced by asymmetric division of spermatogonial stem cells every 16 days—that is, 23 cell divisions per year. A further four cell divisions are required for the stem cells to undergo differentiation. If we take an average age of onset of male puberty as 13 years and an average paternal age of 30 years, a total of 34 + (23 × [30 – 13]) + 4 = 429 male germ-cell divisions are needed. Even more germ-cell divisions would be required to produce sperm in older fathers.

In many studies, the observed rates of *de novo* single nucleotide variants in families show that the male germ-line mutation rate exceeds the female mutation rate by a factor of about four, but is dependent on paternal age (see **Figure 1** for the results from one study). The SNV studies include a



family	number of haplotyped *de novo* mutations	
	maternally transmitted	paternally transmitted
1	30	122
2	21	78
3	33	111
all	84	311

Box 11.2 Figure 1 ***De novo* mutations increase in number with paternal age and are mostly of paternal origin.** The graph shows the total number of *de novo* mutations recorded by genome-wide sequencing for each of four children in three families plotted against the age of the father. Gray areas denote the region covered by the 95% confidence interval of the intercept and slope of the linear regression line for each family. As shown in the table, informative haplotyping was possible for a proportion of the *de novo* mutations, showing a roughly 4:1 bias in favor of paternal origin. (Reprinted from Rahbari R *et al*. [2016] *Nat Genet* **48**:126–133; PMID 26656846. With permission from Springer Nature. Copyright © 2016.)

sizeable proportion of **C**pG → **T**pG transitions (which occur independently of DNA replication). *De novo* length variants at microsatellite (short tandem repeat) loci always originate by replication slippage, and they reveal a slightly higher proportion of paternal variants.

The increase in paternal germ-line mutation rate with age is quite steep, doubling roughly with each additional 17 years of age. However, it appears to differ between different families (see **Figure 1**), and it is not as high as might be predicted by the number of additional cell divisions per year in males after puberty. Recent data suggest a model where the mutation rate is not constant during germ-cell development, being comparatively high before puberty (especially after the primordial germ-cell stage) but significantly reduced after male puberty (see Rahbari R *et al.* [2016] PMID 26656846, under Further Reading).

A maternal age effect has also recently become evident in a large study (36,441 high quality *de novo* mutations in 816 family trios) reported by Goldmann *et al.* in 2016 in *Nature Genetics* (PMID 27322544; see Further Reading). It now seems that one new mutation is contributed for every additional four years of maternal age, about ¼ of the paternal age effect (one additional mutation for each additional year of father's age). Thus, a 40-year-old mother would be expected to contribute 5 more *de novo* mutations to her child than she would have done at age 20, and a 40-year-old father should contribute 20 additional mutations than when aged 20 years.

De novo mutations

Each of us displays some DNA variants that are not apparent in either of our biological parents (***de novo* mutations** or variants). Taking an average human base-substitution rate of 1.0–1.2 × 10^{-8} per generation, and 6000 Mb of DNA in a zygote, the average number of *de novo* SNVs (single nucleotide variants) in a child might be expected to be roughly 60–70 (1.0–1.2 × 10^{-8} × 6 × 10^9). Studies that have directly assessed the frequency of *de novo* mutation (by examining the DNA of parents and their children) reveal that children often show from 30–80 *de novo* mutations, the number being very dependent upon paternal age, and to a lesser extent maternal age (see **Box 11.2**).

How similar are human genomes?

The 1000 Genomes Project data show that genetic variation in 2504 individuals from 26 different populations ranges from approximately 4.1 million to 5.0 million variant sites, with significant variation between populations. In particular, African populations show considerably greater genetic diversity than all other populations, consistent with the out-of-Africa model of human origins. Because of recent population admixture, individual populations from Central and South America also show elevated within-population genetic diversity that roughly correlates with the degree of recent African ancestry (**Figure 11.10**).

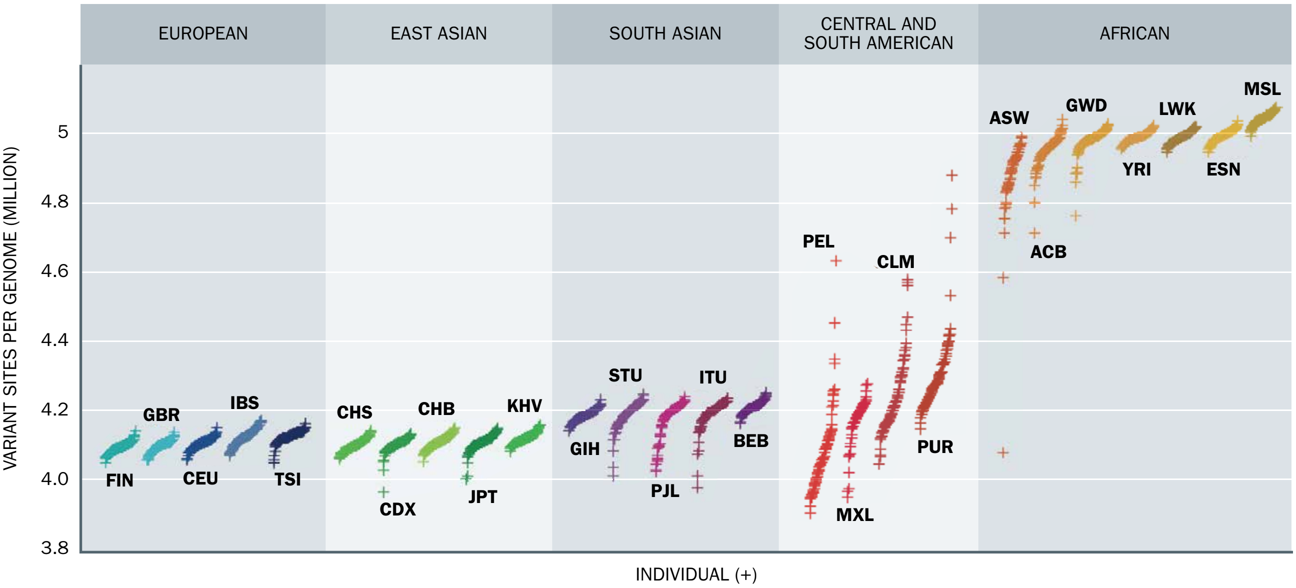


Figure 11.10 Genetic diversity in 26 different human populations sampled in the 1000 Genomes Project. Despite its name, the 1000 Genomes Project sampled DNA from 2504 genomes obtained from 26 different human populations. Each + sign indicates a single individual. Individuals showed from 3.90 to 5.05 million variant sites per genome. Note the consistently low values for genetic variation in *European populations* (FIN, Finnish; GBR, British; CEU, French; IBS, Spanish; TSI, Italian), *East Asian populations* (CHS, Han Chinese South; CDX, Dai Chinese; CHB, Han Chinese Beijing; JPT, Japanese; KHV, Vietnamese), and *South Asian populations* (GIH, Gujerati Indians; STU, Sri Lanka Tamils; PJL, Pakistani Punjabis; ITU, Indian Telugi; BEB, Indian Bengalis). There is a wider spread of variation in the *Central and South American populations* (PEL, Peruvians; MXL, Mexicans; CLM, Colombians; PUR, Puerto Ricans), but by far the highest variation is seen in *African populations* (ASW, African Americans; ACB, Afro-Caribbeans; GWD, Gambians; YRI, Nigerian Yoruba; LWK, Kenyan Luhya; ESN, Nigerian Esan; MSL, Sierra Leone Mende). (Adapted from the 1000 Genomes Project Consortium [2015] *Nature* **526**:68–74; PMID 26432245. With permission from Springer Nature. Copyright © 2015.)

Single nucleotide variants account for 87% of all human DNA variants but short indels are significant, accounting for close to 13% of variants (see **Figure 11.11** for some examples). Although SNVs and short indels together account for more than 99.9% of the number of genetic variants, the great majority of that variation falls outside coding DNA and other functionally important sequences, and is thought to represent neutral mutations.

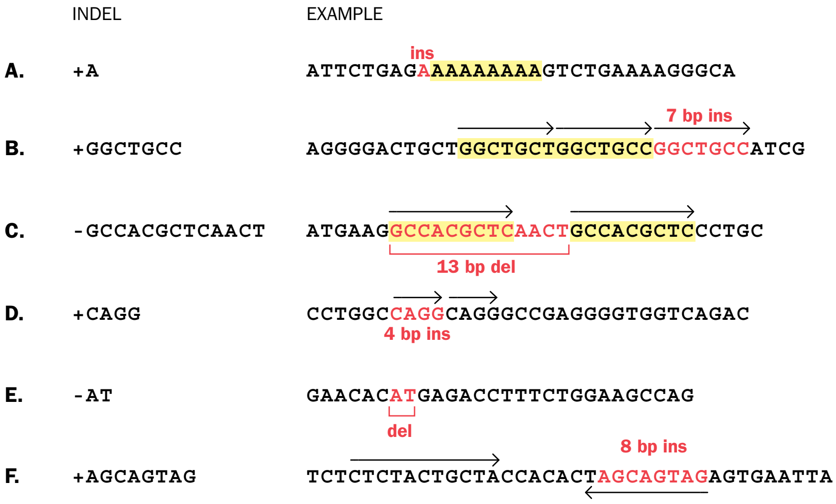

INDEL
EXAMPLE
A.
+A
ins
ATTCTGAGAAAAAAAAAGTCTGAAAAGGGCA
B.
+GGCTGCC
7 bp ins
AGGGGACTGCTGGCTGCTGGCTGCCGGCTGCCATCG
C.
-GCCACGCTCAACT
ATGAAGGCCACGCTCAACTGCCACGCTCCCTGC
13 bp del
D.
+CAGG
CCTGGCCAGGCAGGGCCGAGGGGTGGTCAGAC
4 bp ins
E.
-AT
GAACACATGAGACCTTTCTGGAAGCCAG
del
F.
+AGCAGTAG
8 bp ins
TCTCTCTACTGCTACCACACTAGCAGTAGAGTGAATTA

Figure 11.11 Different classes of short indels according to the presence or absence of repeats at the indel site. Yellow shading indicates direct repeats in the human genome reference sequence. Large letters in red font indicate altered nucleotides in each indel, either because of an insertion (ins) or deletion (del). (**A**) Homopolymer. The insertion of an adenine occurs in a run of As. (**B**) Tandem oligonucleotide repeat. The reference sequence has two almost identical tandem heptanucleotide repeats GGCTGCT/GGCTGCC and the insertion results in a third repeat that is a perfect copy of the second repeat. (**C**) Almost contiguous tandem repeats. The deletion removes one of the GCCACGCTC repeats plus the intervening sequence. (**D**) No repeats, but the insertion results in tandem duplication. The insertion results in duplication of a CAGG tetranucleotide. (**E**) Simple deletion in repeat-free site. (**F**) No repeats, but insertion creates an inverted repeat of nearby sequence. All these examples came from within genes—for further information, see Montgomery SB *et al*. (2013) *Genome Res* **23**:749–761; PMID 23478400.

Structural variants account for just 0.05% of the variants, but some of them, notably many individual CNVs and large deletions, are very large-scale changes. As a result they involve considerably more nucleotides in the genome, at least nearly twice as many, than do small-scale changes (**Table 11.7**). Structural variation is also more highly correlated with functional DNA sequences than is small-scale mutation, and structural variants contribute very significantly to altered gene expression and disease.

The overall data suggest that if we compare one haploid human genome against another, differences might be expected, on average, at 4–5/1000 nucleotide sites (at least for the sampled DNA sequences, which are predominantly from euchromatin: a total of ~4.9 Mb of small-scale changes plus 8.7 Mb of structural variation superimposed on the 3000 Mb euchromatic portion of the reference genome sequence = 13.6 Mb/3000 Mb).

TABLE 11.7 EXTENT OF HUMAN GENETIC VARIATION IN THE 1000 GENOMES PROJECT ACCORDING TO VARIANT CLASS

Type of DNA variation	Number of variants per individual[a]	% of all DNA changes	Amount of DNA involved per haploid genome
SMALL-SCALE DNA CHANGES (<50 bp)	4,076,000–4,935,000	99.95	~4.9 Mb
Single nucleotide variants (SNV)	3,530,000–4,310,000	86.96	3.92 Mb
Indels	546,000–625,000	12.99	1 Mb
STRUCTURAL VARIATION (>50 bp)	2114–2507	0.05	~18.4 Mb[c]
Mobile element insertion[b]	1012–1225		
Large deletions	939–1100		5.6 Mb
Copy number variants (CNV)	153–170		11.3 Mb
Inversions	9–12		

[a] Numbers were the mean of average estimates for the five geographical groupings of human populations shown in boxes in **Figure 11.10**.
[b] 90% were Alu insertions (83.3%) or L1 insertions (7.1%).
[c] Note that the raw number of 18.4 Mb is misleading because it does not take into account sequence repetition. When homozygous structural variants and CNVs carrying multiple sequence copies are collapsed onto the haploid reference assembly, a median of 8.9 Mb is affected by structural variation.
(Data extracted from 1000 Genomes Project Consortium [2015] [PMID 26432245] and Sudmant PH *et al*. [2015] [PMID 26432246]; see under Further Reading.)

Common and rare variants

Only about 10% of the variants sampled across all genomes in the 1000 Genomes Project are common (frequency >0.05); 90% are rare (~76% with a frequency <0.005 and ~14% with a frequency of between 0.005 and 0.05). However, the majority of variants within a single individual are common (only ~1–4% have a frequency of <0.05). That happens because most common variants are shared by all human populations.

Rarer variants are typically restricted to a single continental group, and extremely rare variants are much more likely to be restricted to the same population or to related populations within a continental group. Because sequencing of protein-coding exomes has been carried out on very large numbers of humans there is a wealth of data about extremely rare coding DNA variants and protein variants. We consider this in Section 11.4.

Post-zygotic DNA changes: *de novo* mutations, high somatic mutation rates, and extraordinary structural variation

The DNA in each of our cells carries a common pattern of inherited genetic variation. Superimposed upon that are some post-zygotic DNA changes that vary between cells. Because most mutations arise from base misincorporation during replication, and since DNA replication precedes each mitotic cell division, there are many opportunities for post-zygotic mutation. In total, there are an estimated 10^{16} mitotic divisions in an average person's lifetime: those needed for the zygote to develop into an adult human with about 10^{14} cells, plus extra divisions in some cells that continue to undergo turnover throughout life. Some mutations are independent of DNA replication, however, such as common **C**pG → **T**pG transitions. They simply increase in number in proportion to time, even in *post-mitotic* cells (cells no longer capable of undergoing mitosis, such as the differentiated neurons and muscle cells in brain, heart, and skeletal muscle).

Cells that sustain post-zygotic mutations early in development can give rise to cell lineages containing that mutation, but as development proceeds, more and more additional mutations are progressively introduced into the DNA. Mutations acquired very early in development may be present in many cells; those acquired much later on will be restricted to small numbers of cells. Our cells therefore have different genomes, and each of us is a genetic **mosaic**. That may not be apparent, however, when a blood or tissue sample containing thousands or millions of cells is submitted for DNA analysis (genetic variation is dominated by inherited DNA variants that are common to all nucleated cells). But deep sequencing of somatic cell biopsies and single-cell genomics reveal the complexity.

Origin of *de novo* mutations

Despite the name, an apparently *de novo* mutation may in reality have been inherited from one parent. Different possibilities exist. A new mutation may arise during meiotic recombination in gamete formation (the process of recombination may induce point mutations at crossover points). Alternatively, a post-zygotic DNA change may have occurred in one parent. If such a mutation occurs quite early in development, both germ-line and somatic cells may be affected, and the individual is described as a **gonosomal mosaic**. But a late-occurring mutation may be confined to just germ-line cells, in which case a person may be described as a **gonadal mosaic** (or *germ-line mosaic*). In the case of a parent who is a gonadal mosaic or a gonosomal mosaic, a new variant can be transmitted in gametes but be absent in the parental cells used for DNA testing, such as blood cells.

Occasionally, a *de novo* mutation arises as a post-zygotic DNA change in the child. If the DNA change occurs at one of the early cleavage divisions of the embryo, the new DNA variant may even be present in all cells of the child (recall that a small minority of cells in the early embryo give rise to body cells). Usually, the DNA change occurs later in development; the DNA variant may exist in a proportion of cells, including the cells used for DNA testing, and may be confined to somatic tissues (*somatic mosaicism*). We consider the pathogenic contribution of *de novo* mutations in later chapters. Note that while all of us are mosaics (we each have multiple genetically different cell lines originating from a single zygote), some rare individuals are *chimeras*, having cells derived from two zygotes (see **Figure 5.16**).

High somatic mutation rates

De novo germ-line mutations account for a tiny proportion of the variation transmitted to children (~60 *de novo* SNVs compared to ~4,000,000 inherited but pre-existing parental SNVs). However, in humans (and in all other species where investigated), the mutation rates in normal somatic cells are generally substantially higher than the germ-line mutation rate. Indirect estimates based on marker loci for phenotypes in four human tissues suggested mutation rates that were 4× to 17× greater than the germ-line mutation rate; more recent genome-wide sequencing studies suggest even higher rates in highly proliferative cells and cells in tissues exposed to environmental mutagens, such as skin—see Lynch (2016) (PMID 26953265) and Martincorena *et al.* (2015) (PMID 25999502) under Further Reading.

The high mutation rate in somatic cells might be expected to pose a health risk. Even when a somatic mutation is identical to an inherited pathogenic mutation, however, that is usually not a concern: the consequences are limited to just that cell and any descendants, and the gene in question might not even be expressed in that cell. A mutation that inactivates the LDL receptor gene in a lymphocyte, for example, will have no consequences.

Somatic variation is, nevertheless, important in medicine for different reasons. Somatic mutations in genes that regulate cell proliferation or apoptosis are important in initiating uncontrolled growth of cells leading to cancer, and the great majority of mutations contributing to cancers are somatic mutations. Early somatic mutations may predispose to some neurodevelopmental disorders, and may also be important in neurodegenerative disorders. And post-zygotic mutations can also give rise to germ-line mosaicism: an unaffected person may carry harmful mutations in gonadal tissue that can be transmitted to successive children. We will consider this in detail in later chapters when we consider how genetic variation contributes to disease. Finally, the sheer load of accumulated somatic mutations that have built up over many decades leads to impaired gene function with progressive decline in cell and tissue function, contributing to the aging process.

Extraordinary structural variation in somatic genomes

Recently, it has become clear that copy number variation for large DNA sequences is very frequent in somatic tissues, and somatic genomes appear to be much more variable than formerly thought. In one study where single-cell genomics was used to sequence the genomes of 110 individual human frontal cortex neurons, over 40% of neurons were found to have very large deletions and insertions, ranging from about 3 Mb of DNA to an entire chromosome; no two changes were the same.

The extent of structural variation in neurons has attracted much attention for different reasons. First, a popular idea imagines that there must be special genetic mechanisms to explain extensive neuronal functional diversity (just as there are to explain the functional diversity of lymphocytes). Second, there have been reports of very frequent mobilization of L1 retrotransposons in neurons. Retrotransposition of L1 repeats is known to be suppressed by various cellular defense systems in germ-line cells, and active mobilization of the L1 repeats in neurons has been imagined to expand the genetic diversity of neurons. However, the extent of retrotransposition has been controversial (see Evrony *et al*. [2016], PMID 26901440).

Neurons are also unusual in expressing the recombination activating gene *RAG1* (which is deployed in lymphocytes to allow the cell-specific DNA rearrangements required to diversify antibodies and T-cell receptors). Neural stem and progenitor cells have also been found to undergo very frequent DNA breaks in a very restricted set of genes involved in neural cell adhesion and synapse function. That has raised the possibility that targeting double-strand DNA breaks to certain genes important in neuron function drives somatic brain mosaicism and endows neurons with individual identity—see the commentary by Weissman and Gage (2016) (PMID 26871622) under Further Reading.

11.4 FUNCTIONAL GENETIC VARIATION AND PROTEIN VARIATION

In this section we give a broad overview of functional genetic variation, the fraction that affects how genes or other functional DNA sequences work. Some genetic variants in functionally important DNA sequences have a significant impact on biological **fitness** (the capacity for reproductive success and transmission of the genotype to descendants), and as we will see, natural selection can have negative and positive impacts on the frequency of different DNA variants.

Protein sequence variation

The bulk of protein sequence variation is due to genetic variation. The major contribution is made by **nonsynonymous** base substitutions, causing amino acid replacement. Short non-frameshifting indels in coding DNA may produce fewer or more amino acids. But protein sequence variation is more limited than that of coding DNA. Redundancy in the genetic code means that many base substitutions in coding DNA are **synonymous (silent)** and do not affect the protein sequence (see **Figure 11.12**).

Sequence variation in a protein does not simply correspond to the sequence variation in its coding DNA for other reasons as well as silent substitutions. First, protein sequence

```
       F   V   N   Q   H   L   C   G   S   H   L   V   E   A   L   Y   L   V   C   G   E   R   G   F
HUMAN TTT GTG AAC CAA CAC CTG TGC GGC TCA CAC CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TTC
      --- --- --* --* --- --* --* --* --* --- --- --- --* --- --- --- --* --- --* --- --* --* --- ---
MOUSE TTT GTC AAG CAG CAC CTT TGT GGT TCC CAC CTG GTG GAG GCT CTC TAC CTG GTG TGT GGG GAG CGT GGC TTC
       F   V   K   Q   H   L   C   G   S   H   L   V   E   A   L   Y   L   V   C   G   E   R   G   F
```

Figure 11.12 Redundancy in the genetic code means that protein sequence variation is less than variation of the respective coding DNA. The example shown here is the first 24 amino acids of the insulin beta chain from human (top row) and mouse (bottom row) with the respective coding sequences. The protein sequences are identical except for one conservative substitution at amino acid number 3 (23/24 = 95.83% sequence identity). The coding sequences, however, show 11 differences (highlighted by asterisks) out of 72 positions (61/72 = 84.72% sequence identity). All 11 nucleotide differences occur at the third base of codons, and all are synonymous (silent) changes except the nonsynonymous one in codon 3 that gives chemically related amino acids: asparagine (in human) and lysine (in mouse).

variation can arise post-transcriptionally: alternative RNA splicing and RNA editing can produce alternative **isoforms** of a protein (and of functional noncoding RNAs). A second type of isoform can occur when duplicated genes make extremely similar proteins, causing some confusion in how to label protein variants.

(Note: protein sequence variation has traditionally been described as *polymorphism* when there are at least two genetic variants in the population, and although the term *allele* strictly refers to DNA variants at a single locus, it has often been used loosely to describe protein variants originating from genetic variation, irrespective of frequency. The recently duplicated *HLA-DRB1* and *HLA-DRB5* genes each make polymorphic HLA-DRβ chains, and the different HLA-DRβ variants have also been described as alleles, even though they originate from two gene loci. To avoid confusion, there is an increasing tendency to use the term *protein variant*.)

Although individual human proteins generally show limited sequence variation, there is a sliding scale. Some proteins, such as histones and developmental gene regulators, play fundamentally important cellular or developmental roles and are extremely highly conserved: sequence variation is strongly suppressed because even small changes may often result in loss of function and harm to cells and tissues. Some other proteins, such as fibrinopeptides, show significant sequence variation simply because they are under few structural or functional constraints.

The proteins showing the highest sequence variation are ones that work in systems where there is some kind of genetic conflict driving competition between genes. Here, the sequence variation is actively promoted. One type of competition occurs between host defense genes and genes making a foreign protein that has been introduced into the body and is associated with some threat. In the case of microbial pathogens, for example, there is an arms race between pathogen genes that propel sequence variation in the surface proteins of bacteria and viruses (to avoid being detected and destroyed by the immune system), and immune system genes that promote sequence variation in proteins dedicated to recognizing foreign proteins (in an attempt to deal with the pathogenic threat). We describe the extraordinary variation in the adaptive immune system in Section 11.5.

Rare protein variants

The protein-coding exome, accounting for just over 1% of the genome, is easier to sequence than a whole genome, and provides valuable information on coding DNA variants and protein variants. A study of variation in the exomes of 60,706 humans has provided unprecedented resolution of human genetic variation: an average of one variant every eight nucleotides was recorded, and mutational recurrence was found to be widespread. The rare variants have arisen very recently in evolution following a dramatic increase in the rate of human population growth in the last 10,000 years, and especially in the last millennium. We consider pathogenic variants in later chapters.

Most genetic variation has no effect on the phenotype, but a small fraction is harmful and is subject to purifying selection

Functional genetic variation accounts for a rather small fraction of human genetic variation (the great majority of the variation does not affect cellular function in any way—see **Table 11.8**). Because it affects how genes work, functional genetic variation can result in differences between individuals in health status and survival rates. That can have consequences for the reproductive fitness of individuals and their ability to transmit genotypes to future generations.

TABLE 11.8 THREE REASONS WHY MOST HUMAN MUTATIONS ARE NEUTRAL, HAVING NO EFFECT ON THE PHENOTYPE

Reason	Examples
Functionally important DNA sequences account for a quite small percentage of our genome	The total amount of coding DNA accounts for just over 1% of our genome. Noncoding regulatory DNA sequences and DNA sequences transcribed to make functional noncoding RNA account for just a few more percent of the genome
Genetic redundancy (at the level of the genetic code, gene duplication, and internal tandem repetition)	Most nucleotide substitutions at the third base position of a codon, and a few at the first base position, do not cause a change in the amino acid
	Some crucially important functions are performed by repeated sequences, including rRNA and classic histone proteins. Thus, for example, there are no functional consequences for point mutations that inactivate a ribosomal RNA gene because each of our different ribosomal RNAs is made by hundreds of almost identical genes
Functionally unimportant amino acid or nucleotide positions within proteins or within functionally important noncoding sequences	In flexible linker regions connecting functional domains of proteins, amino acid substitutions and small insertions and deletions are readily tolerated. Some intragenic repeats coding for repeated structures within proteins are disposable
	In noncoding RNAs the structure of the RNA is often critical, as are binding sites for interaction with other molecules. Nucleotides involved in intramolecular base pairing and in binding to other molecules are important, but many other nucleotides often undergo substitutions without apparent functional consequences

A small fraction of genetic variation is harmful and affects how genes work in two major ways. An abnormal gene product may be made that works in a way that is positively harmful. Alternatively, the *amount* of the normal gene product is altered, causing cells to function abnormally. Harmful DNA variants can reduce the biological fitness of at least some individuals who carry them (depending on the DNA variant, some individuals may be normally healthy carriers of that variant).

Transmission of a harmful DNA variant to future generations is reduced because the reproductive success of at least some individuals with that variant is reduced (as a result of associated impaired health or premature death, for example). This is an example of natural selection operating on the phenotype to affect genetic variation. In this case, **purifying selection** (also called *negative selection*) works toward elimination of the harmful DNA variant from the population (over multiple generations).

Purifying selection maintains the function of all functionally important sequences: coding DNA, regulatory sequences, and sequences transcribed to make functional noncoding RNA. Not all the nucleotide positions in these sequences are crucial for function, but, overall, these sequences are much more likely to be conserved than sequences not subject to purifying selection (where mutation is much more readily tolerated). Acting over evolutionary time periods, purifying selection suppresses genetic variation in functionally important DNA sequences in an effort to maintain function (the sequences are said to be *functionally constrained*). Sequences like this can most readily be detected as sequences that have been comparatively highly conserved during evolution (by, for example, comparing human sequences with corresponding sequences from other species; we consider this in detail within the context of comparative genomics in Chapter 13).

Positive selection underpins adaptive changes by promoting the frequency of advantageous DNA variants

Occasionally, a DNA variant has a beneficial effect on the phenotype that can be transmitted to offspring. DNA variants like this become prevalent through a form of natural selection called **positive selection** (or Darwinian selection). Individuals who possess the advantageous DNA variant have increased survival and reproductive success rates compared with other individuals in a population. Because they will be more able to transmit their DNA to future generations, the advantageous DNA variant will increase in frequency over many generations and spread through the population. Unlike purifying selection, therefore, which acts to eliminate a harmful DNA variant and suppresses genetic variation, positive selection acts to promote the spread of an advantageous DNA variant.

Positive selection has occurred in all animal lineages including the human lineage, where it has been responsible for fostering different features that distinguish us from the great apes, notably human innovations in brain development (detailed in Chapter 14). But it is much less widespread than purifying (negative) selection. It is much easier for harmful variants to arise as a result of mutation than for beneficial ones to do so. If we take the classic interpretation of mutation at coding sequences, nearly 30% of all substitutions causing an amino acid change have been thought to be severely deleterious (having a major negative effect on protein function), and another 33% or so are considered to be mildly deleterious. As a result, purifying selection is operating on a significant fraction of coding DNA and is continually required to maintain the functions of important DNA sequences. Advantageous mutations are rare: there are fewer opportunities for mutation to transform how genes function so as to confer some significant advantage, permitting increased reproductive success.

Although less widespread than purifying selection, positive selection is extremely important in helping organisms to adapt to a change of environment, or to a new environment. In principle, positive selection can act on *standing variation* (that is, pre-existing variation) or on novel mutations. Adaptation can be expected to occur more rapidly from standing variation than from new mutation (beneficial alleles are immediately available from standing variation, and can start at higher frequencies).

Selective sweeps

Although expected to be much more frequent than positive selection working on novel variants, cases of positive selection working on standing variation are very difficult to identify. However, recent strong positive selection acting on a novel variant can leave telltale DNA signatures. One likely example appears to have occurred when human populations first moved out of equatorial Africa and migrated to northern latitudes where sunlight is reduced. Selection for reduced skin pigmentation is thought to have compensated for the reduced sunshine by enhancing transmission of ultraviolet (UV) light through the skin. The most significant contributor is believed to have been a nonsynonymous change in the *SLC24A5* gene, resulting in replacement of alanine at position 111 by a threonine (A111T).

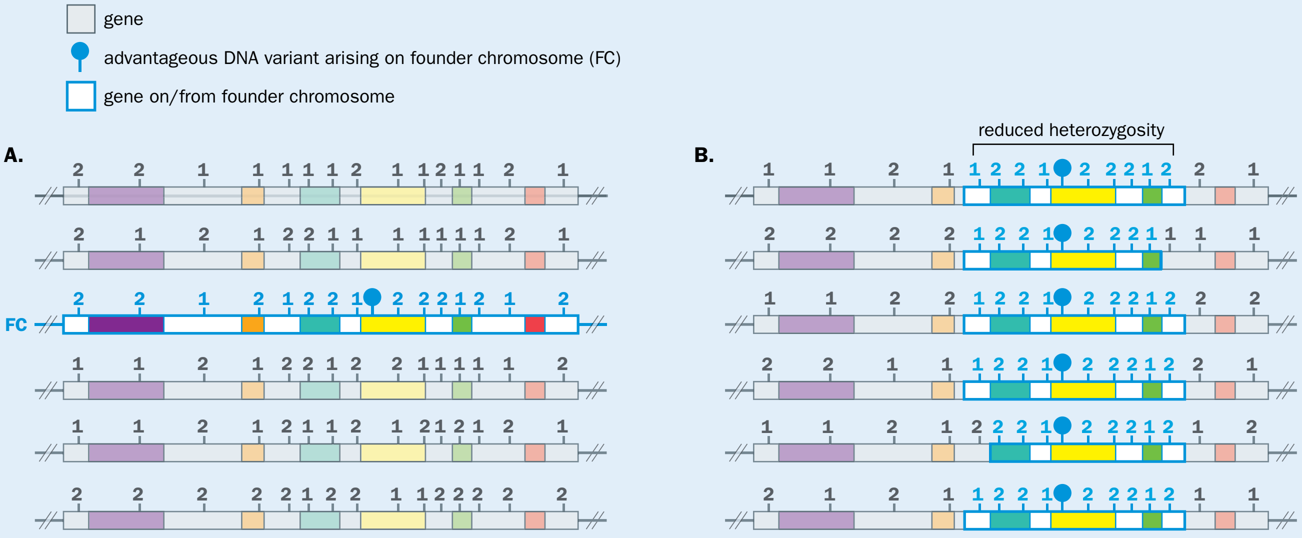
The SLC24A5 protein is a type of calcium transporter that regulates melanin production, and the A111T change results in defective melanogenesis, enhancing UV light transmission through the skin. Although positive selection can foster genetic diversity (by promoting the spread of a new advantageous variant), as the advantageous new variant spreads through the population, the subchromosomal region containing the new variant replaces the pre-existing equivalent regions. As a result, genetic variation can also be suppressed over a sizeable chromosomal region containing the new variant. In **Box 11.3** we illustrate this phenomenon, called a **selective sweep**, with reference to the A111T variant at the *SLC24A5* locus.

BOX 11.3 RECENT STRONG POSITIVE SELECTION CAN LEAD TO A SELECTIVE SWEEP WITH LOCAL SUPPRESSION OF GENETIC VARIATION

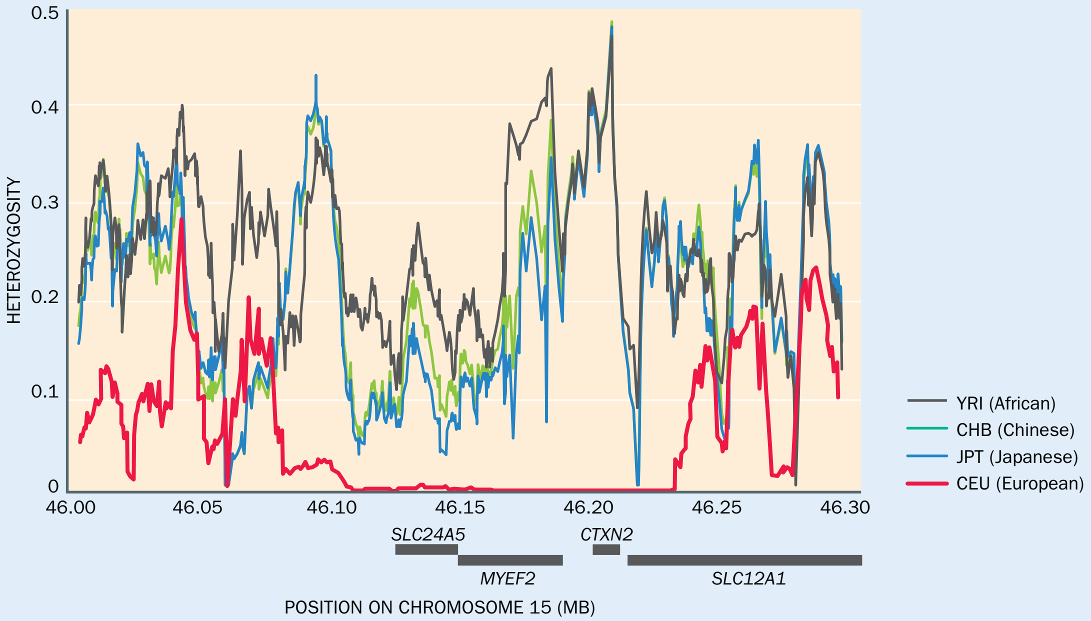
Positive selection for an advantageous DNA variant can leave distinctive signatures of genetic variation in the DNA sequence. Imagine a large population of individuals before positive selection for some advantageous DNA variant on a region of, say, chromosome 22. If we were able to scan each chromosome 22 in the population before selection we might expect to find hundreds of thousands of different combinations of genetic variants.

Now imagine that an advantageous DNA variant arises by mutation on one chromosome 22 copy (**Figure 1A**), and then gets transmitted through successive generations. If the advantageous variant is subject to strong positive selection, people who carry it will have significantly higher survival and reproductive success rates. As descendants of the original chromosome 22 copy carrying the variant become more and more common, the selected DNA variant will increase in frequency to become a common allele.

The entire chromosome 22 copy is not passed down as a unit: recombination will result in replacement of some original segments by equivalent regions from other chromosome 22s. However, a short segment from the original chromosome 22 copy containing the favorable DNA variant and neighboring alleles (***hitchhiking alleles***) will increase in prevalence in what is termed a selective sweep. The segment will be marked by very low levels of heterozygosity (**Figure 1B**), but it will be slowly reduced in size by recombination. The region on chromosome 15 containing the *SLC24A5* locus in Europeans (described in the main text) provides a good example—see **Figure 2**.



Box 11.3 Figure 1 General effect of a selective sweep for an advantageous DNA variant. (**A**) Heterozygosity profile before selection. Imagine that an advantageous DNA variant has just occurred in the gene shown in yellow on a founder chromosome 22 (with DNA outlined in blue, and genes shown as colored boxes and intergenic space represented by white boxes). Now imagine assaying genetic variation over each copy of chromosome 22 in the population using intronic and extragenic single nucleotide variant (SNV) markers known to map this region of chromosome 22. We might expect significant heterozygosity, as shown by the six representative chromosome 22s. (**B**) Heterozygosity profile after positive selection over many generations. Vertical transmission of the founder chromosome, recombination, and continued positive selection for the advantageous variant will result in increased frequency of the advantageous DNA variant plus very closely linked variants, causing reduced heterozygosity for that chromosome segment. Some tightly linked neighboring genes will also increase in frequency in the population because of selection for the variant. They are often described as *hitchhiking alleles* (shown here in blue and green).



Box 11.3 Figure 2 A strong selective sweep acting on an advantageous DNA variant in the *SLC24A5* gene in European populations. Heterozygosity levels in the region containing the *SLC24A5* gene on chromosome 15 were determined for a high density of common single nucleotide polymorphisms and averaged over 10 kb windows. The observed heterozygosity profiles for this chromosome region are unremarkable in African, Chinese, and Japanese populations. However, in the European population a strong selective sweep for a specific *SLC24A5* variant associated with reduced skin pigmentation (see main text) has meant that almost all European chromosome 15s share a segment containing the favorable *SLC24A5* variant and hitchhiker alleles at the neighboring *MYEF2* and *CTXN2* loci. The result is a sharp reduction in heterozygosity for this chromosome region. (Adapted from Lamason RL *et al.* [2005] *Science* **310**:1782–1786; PMID 16357253. Reprinted with permission from the AAAS.)

Protein diversity generated by gene duplication and alternative processing of a single gene

Genetic variation at a single locus is the major source of sequence variation in a protein, but additional protein diversity can be generated as a result of different mechanisms, notably gene duplication and post-transcriptional processing.

Diversity through gene duplication

Gene duplication offers the possibility of generating many slightly different forms of a protein. After each gene duplication event, random mutations cause the repeated genes to diverge in sequence. Initially highly similar proteins are produced with slightly different properties, but eventually, over long evolutionary time periods, more divergent proteins are formed with rather different functions.

Among the different selective advantages for gene duplication is the advantage of conferring additional diversity for proteins working in host defense against external threats. Examples include duplication of major histocompatibility complex genes (described in Section 11.5) and of genes encoding some cytochrome P450 enzymes involved in dietary metabolism.

The most extraordinary example of diversity through gene duplication is the olfactory receptor gene family, our largest protein-coding gene family. Extensive gene duplication events have formed 396 olfactory receptor (OR) genes plus about 425 OR pseudogenes. Most odorants bind to several OR variants: precise identification of any one odor seems to depend on a combinatorial code of binding of different receptors, allowing us to potentially sense many thousands of different odors. There is, however, pronounced variation between individuals in the ability to detect specific odors, and the OR gene family demonstrates the greatest variation in gene content of any human gene family. In addition to the pseudogenes, alleles for deleterious mutations at functional OR loci are both common and highly variable between individuals (**Figure 11.13**). An average person is heterozygous at about one-third of the OR loci and makes in the region of 500 variant olfactory receptors.

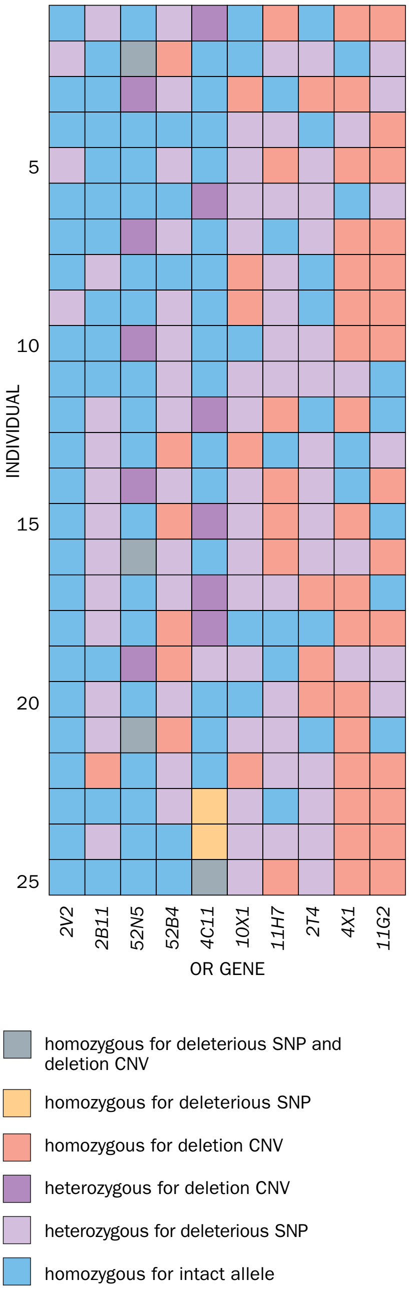


Figure 11.13 Genetic variation in some olfactory receptor (OR) genes. Genotype calls of ten OR genes, for which both an intact and inactive allele are present in the population. Each row represents an individual; each column represents a gene. (Data courtesy of Doron Lancet and Tsviya Olender, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot.)

In addition to gene duplication, the special case of intragenic somatic recombination within clusters of serially repeated segments of coding sequence is a major contributor to the sequence variation seen in immunoglobulins and T-cell receptors—we provide details in Section 11.5.

Post-transcriptionally induced variation

Not all of the sequence variation in our proteins and functional noncoding RNAs is due to variation at the DNA level. Some variants are attributable to alternative post-transcriptional processing so that a single allele at the DNA level can produce different variants (**isoforms**) of a protein or noncoding RNA. Alternative RNA splicing is very common and can lead to skipping of exons from some transcripts and other forms of sequence difference. Specific nucleotides in transcripts can also occasionally be altered by processes known as RNA editing (described in Section 10.6).

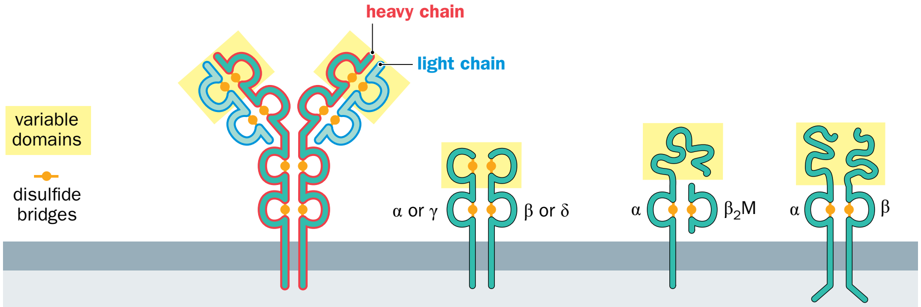
11.5 EXTRAORDINARY GENETIC VARIATION IN THE ADAPTIVE IMMUNE SYSTEM

Our immune systems have a tough task. They are engaged in a relentless battle to protect us from potentially harmful microbes and other parasites. Not only must we be protected against a bewildering array of pathogens, but, in addition, new forms of a pathogen can rapidly develop by mutation to provide new challenges to which human populations must continuously adapt. Different types of proteins work in the immune system, some as sentinels dedicated to recognizing antigens on the surface of invading microbes, some as signaling molecules that allow different immune system cells to communicate and co-operate, and some as executor proteins involved in mechanisms to deal with the invaders.

In Section 3.4 we covered the detail about the biology of the vertebrate immune system: the cells involved and some of the principal mechanisms used in immune system host defense. Here, our priority is to explain the extraordinary sequence diversity of some immune system proteins that act as sentinels by seeking out and binding foreign antigen. Although we will revisit some of the concepts involved to set the scene, interested readers might wish to reacquaint themselves with some of the wider background given in Section 3.4.

Recall that the vertebrate immune system has two arms. The innate immune system provides a first line of defense: it uses different types of barrier to reduce access by harmful pathogens to body tissues, and develops a rapid response based on general pattern recognition of invading microbes. The more sophisticated **adaptive immune system** is mobilized by components of the innate immune system when the latter fails to provide adequate protection against an invading pathogen. It has exquisite specificity, being able to discriminate between tiny differences in molecular structure of invading cells and viruses that pose a threat to the body. At the forefront of the adaptive immune system's ability to recognize foreign antigen are four types of protein that show extraordinary sequence variation, as described below and in **Figure 11.14**.

- ***Immunoglobulins.*** Expressed on the surface of B cells in the bone marrow or secreted as soluble immunoglobulins (antibodies) by activated B cells, their main task is to recognize and bind specific foreign antigens. They can bind and neutralize toxins released by microbes, inhibit viruses from infecting host cells, and activate complement-mediated lysis of bacteria and phagocytosis.
- ***T-cell receptors.*** Displayed on the surface of T cells, they work in cell-mediated immunity, along with proteins encoded by the major histocompatibility complex (MHC; known in humans as the HLA complex).
- ***Class I MHC (HLA) proteins.*** They are expressed on the surface of almost all nucleated cells and enable cytotoxic T lymphocytes to recognize and kill host cells that are infected by a virus or other intracellular pathogen.
- ***Class II MHC (HLA) proteins.*** They are displayed on the surface of very few types of cell, notably immune system cells that present foreign antigen to be recognized by helper T lymphocytes.



	Ig = B-cell receptor (or soluble antibody)	T-cell receptor	classic class I MHC (HLA)	classic class II MHC (HLA)
haploid number of human genes	3	4	3*	6*
number of different proteins	huge numbers in one person**	huge numbers in one person**	up to 6 in one person; ~4000 in population	up to ~12 in one person; ~1200 in population
source of genetic variation	somatic recombination and hypermutation	somatic recombination	extensive polymorphism	extensive polymorphism

Figure 11.14 Extreme variation exhibited by four types of protein that are used to recognize foreign antigens. Immunoglobulins (Igs), T-cell receptors, and MHC (major histocompatibility complex) proteins are heterodimers with similar structures—globular domains (maintained by intrachain disulfide bridges) and N-terminal *variable domain* that bind foreign antigens (but they otherwise have conserved sequence, mostly *constant domains* in Igs). They are cell surface receptors (except that Igs in activated B cells become secreted antibodies). Only a few human genes encode each protein chain, but nevertheless many different proteins are made because of special genetic mechanisms in B and T lymphocytes and because of selection for heterozygosity of HLA proteins (see main text). β_2-microglobulin (β_2M), the non-polymorphic light chain of class I HLA proteins, is encoded by a gene outside the major histocompatibility complex. * See **Box 11.4 Figure 1** for genes encoding classic HLA proteins. ** The recombination mechanisms can generate a theoretical diversity of at least 5 × 10^{13} naive B-cell receptors and 10^{18} α:β T-cell receptors, but only a subset of this diversity is expected to be physiologically present in the actual repertoire of an individual.

The mechanisms that produce the extraordinary variety of immunoglobulins and T-cell receptors are different from those responsible for producing the different types of MHC (HLA) proteins. The extensive variety of MHC proteins is apparent at a *population* level. MHC proteins are highly polymorphic but a single person has a limited number of them, having at most two alleles at each of a small number of polymorphic MHC loci. By contrast, the huge sequence variation of immunoglobulins and T-cell receptors is generated at the level of the *individual*, as a result of post-zygotic DNA changes in B cells and T cells.

Somatic mechanisms allow cell-specific production of immunoglobulins and T-cell receptors

Human sperm cells and egg cells each have just three immunoglobulin genes—*IGH*, *IGK*, and *IGL*—and four T-cell receptor genes—*TRA*, *TRB*, *TRD*, and *TRG*. Nevertheless, each person can make huge numbers of different immunoglobulin proteins, and huge numbers of different T-cell receptors. That happens because immunoglobulin genes in maturing B cells, and T-cell receptor genes in maturing T cells, are programmed to undergo extraordinary DNA changes. There is a high degree of flexibility in how the changes occur, and they occur *in a cell-specific way*: in each person the precise DNA changes vary from one maturing B cell to the next, and from one maturing T cell to the next. Different mechanisms are responsible as described below.

Combinatorial diversity via somatic recombination

In germ-line DNA, each immunoglobulin and T-cell receptor gene is made up of a series of repeated gene segments that specify discrete segments of the protein. Take the example of an immunoglobulin heavy chain. The constant domain defines the functional class of immunoglobulin (IgA, IgD, IgE, IgG, or IgM) and is encoded by repeated C gene segments (each with coding sequences split by introns). The variable domain is encoded by three types of repeated gene segment containing coding DNA only: a V (variable region) gene segment (encoding the majority of the variable domain, starting from the N-terminus) plus a D (diversity region) gene segment and a J (joining region) gene segment (**Figure 11.15A**). A similar arrangement applies to light-chain domains, except that the variable domain is encoded by two repeated gene segments only: V gene segments and J gene segments.

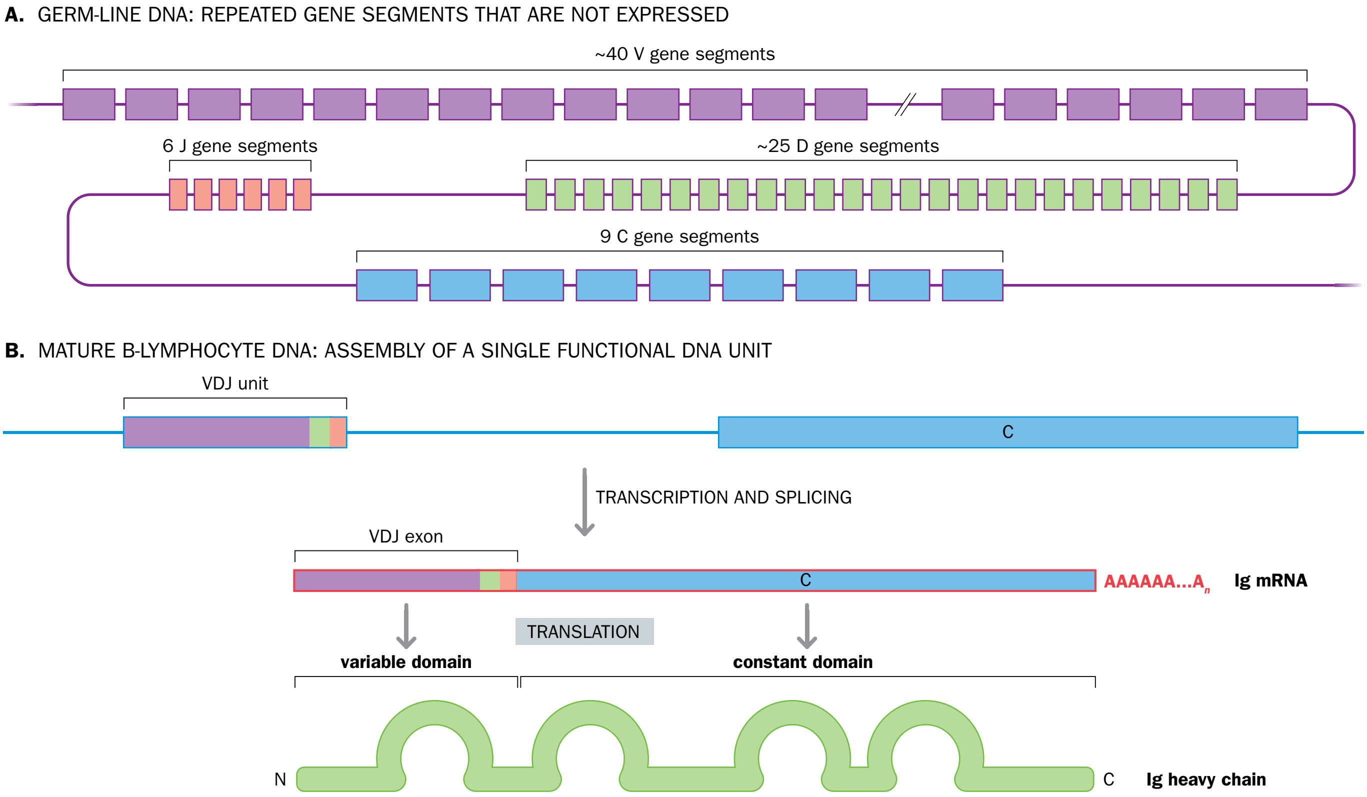


Figure 11.15 Variation in the structure of the immunoglobulin heavy-chain gene in germ-line DNA and maturing B lymphocytes. Somatic DNA rearrangements occur in immunoglobulin genes and T-cell receptor genes in B and T cells, respectively, resulting in a change in gene organization, as in the human *IGH* gene shown here. (**A**) In germ-line DNA, and in all non-B cells, the *IGH* gene has multiple but slightly different repeats for each of four types of gene segment and is not expressed. V (variable region), D (diversity region), and J (joining region) gene segments have sequences that together can encode the variable domain of the immunoglobulin heavy chain; C (constant domain) gene segments have sequences that can encode the constant domain of the protein (note: individual C domain gene segments have multiple exons that are not shown here for clarity). (**B**) Somatic recombination in maturing B cells produces a change in gene organization that allows gene expression and the production of a single type of immunoglobulin heavy chain. As a result of recombination events, a single V gene segment is fused to individual D and J gene segments to produce a functional VDJ unit that activates transcription and splicing. The resulting VDJC mRNA is translated to give the heavy chain, with the VDJ sequence specifying the variable domain of the protein, and the C sequence specifying the constant domain. For the sake of clarity, the other gene segments are not shown. See **Figure 11.16** for the detail of the process.

Cells other than B lymphocytes have the same immunoglobulin gene organization as in germ-line DNA, and the immunoglobulin genes are not expressed. In maturing B cells, however, a single type of immunoglobulin protein is made. In the case of the immunoglobulin (Ig) heavy chain, somatic recombination brings about joining of one V gene segment to individual D and J gene segments to form a functional VDJ unit, essentially a novel exon, that activates transcription and RNA splicing. As a result, a VDJ exon is joined to a C sequence at the RNA level; the VDJ component of the mRNA specifies the variable domain of the protein and the C component specifies the constant domain (**Figure 11.15B**).

In much the same way, an immunoglobulin light chain is made after somatic recombination forms a novel functional VJ exon from the light-chain gene segments. T-cell receptor chains are formed in maturing T cells after individual gene segments are brought together by somatic recombination to form a functional VDJ exon (in the case of a T-cell receptor β chain) or a VJ exon (for a T-cell receptor α chain).

The key point about the somatic recombination events used to produce the functional VDJ or VJ exons is that they occur *in a cell-specific way*. From the two or three clusters of repeated segments encoding sequences for the variable domain, one segment only from each cluster is selected to be included to make a functional VDJ or VJ exon, and that choice is made *randomly* in each B or T cell in a person. We use the example of making an immunoglobulin heavy chain to illustrate this point in detail in **Figure 11.16**.

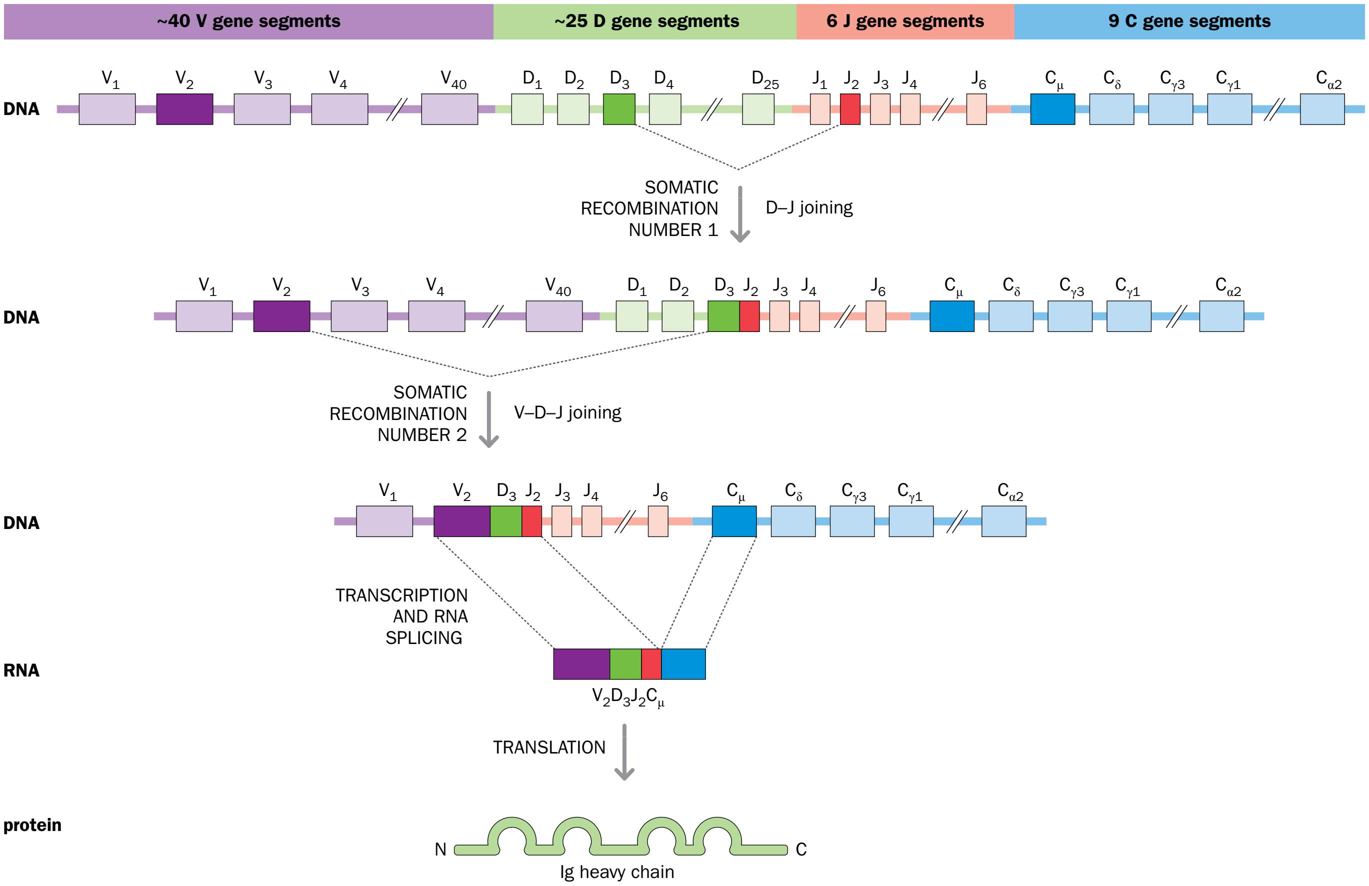

~40 V gene segments
~25 D gene segments
6 J gene segments
9 C gene segments
DNA
V_1 V_2 V_3 V_4 V_{40} D_1 D_2 D_3 D_4 D_{25} J_1 J_2 J_3 J_4 J_6 C_μ C_δ $C_{\gamma3}$ $C_{\gamma1}$ $C_{\alpha2}$
SOMATIC RECOMBINATION NUMBER 1
D–J joining
DNA
V_1 V_2 V_3 V_4 V_{40} D_1 D_2 D_3 J_2 J_3 J_4 J_6 C_μ C_δ $C_{\gamma3}$ $C_{\gamma1}$ $C_{\alpha2}$
SOMATIC RECOMBINATION NUMBER 2
V–D–J joining
DNA
V_1 V_2 D_3 J_2 J_3 J_4 J_6 C_μ C_δ $C_{\gamma3}$ $C_{\gamma1}$ $C_{\alpha2}$
TRANSCRIPTION AND RNA SPLICING
RNA
$V_2D_3J_2C_\mu$
TRANSLATION
protein
N
C
Ig heavy chain

Figure 11.16 Cell-specific immunoglobulin chains are produced by somatic recombination in B cells. Here, as in **Figure 11.15**, we use the example of the human *IHG* gene. The fusion of single V, D, and J gene segments to make an active VDJ unit is cell-specific, and the chosen single gene segments are shown in dark coloring for emphasis. In this B cell we imagine that the second of the V gene segments fuses with the third D gene segment and the second J gene segment. This comes about by two sequential somatic recombinations: first, D–J joining; then addition of the chosen V gene segment. The mature, functional VDJ coding sequence unit is effectively a large novel exon that will encode the variable domain of the protein. In this example, the successful combination is $V_2D_3J_2$, but the choice of combinations is *cell-specific*: in a neighboring maturing B cell it could be $V_{24}D_{19}J_5$, for example. Once a functional VDJ exon has been assembled, transcription is initiated starting with this exon and RNA splicing joins the VDJ coding sequence to coding sequences initially in the constant domain (C) gene segment, in this case C_μ and then by alternative splicing to C_δ. Another type of somatic recombination (known as *class-switching*, but not shown here) can change the position of C gene segments so that other C gene segments can be used to give alternative classes of immunoglobulin (see text).

RNA splicing fuses the VDJ sequence initially to a C sequence from the nearest C gene segment, initially C_μ and then, through alternative splicing, either C_μ or C_δ. The first immunoglobulins to be made by a B cell, therefore, are membrane-bound IgM and then IgD. Subsequently, as B cells are stimulated by foreign antigen and helper T lymphocytes, they secrete IgM antibodies.

Later in the immune response, B cells undergo another type of somatic recombination called *class-switching* (or isotype-switching) to produce different antibody classes. The recombinations (not shown here) can position an alternative C gene segment to be nearest to the J gene segments: either a C_γ, C_ε, or C_α gene segment to produce, respectively, an IgG, IgE, or IgA antibody.

Additional diversity generation

The use of somatic recombination alone would allow each person to produce diverse sets of both immunoglobulins (potentially over 2 million types; see **Table 11.9**) and T-cell receptors. As listed below, two or three additional mechanisms are responsible for generating sequence diversity; together with VDJ and VJ recombination, they endow each of us with the potential for making huge numbers of different antigen-binding sites, both for immunoglobulins and T-cell receptors. As required, individual B cells and T-cell receptors that successfully recognize foreign antigen are induced to proliferate to make identical clones with the same antigen specificity as the original cell.

- ***Protein chain combinatorial diversity***. Immunoglobulins and T-cell receptors are heterodimers and diversity is compounded by unique combinations of two unique protein chains. Take the example of immunoglobulins. Although a B cell, being diploid, has six immunoglobulin genes, each B cell is said to be monospecific: it makes only one type of antibody. That happens because in each B cell only one of the two *IGH* alleles is (randomly) selected to make a heavy chain (**allelic exclusion**), and because only one of the four light-chain genes is ever used to make a light chain (a combination of *light-chain exclusion*—to select either a kappa or lambda light chain—plus allelic exclusion). That raises the number of different immunoglobulins by a factor of eight (see **Table 11.9**).
- ***Junctional diversity***. The somatic recombination mechanisms that bring together different gene segments in immunoglobulin or T-cell receptor genes variably add or subtract nucleotides at the junctions of the selected gene segments.
- ***Somatic hypermutation***. This mechanism applies to immunoglobulins and is used to further increase variability in the variable domain after somatic recombinations have produced functional VDJ or VJ exons. When B cells are stimulated by foreign antigen, an activation-induced cytidine deaminase is produced by the activated B cell that deaminates cytidine to uridine. The uridines are variably repaired so that multiple nucleotides in the variable domain are mutated.

TABLE 11.9 MILLIONS OF DIFFERENT IMMUNOGLOBULINS CAN BE MADE USING JUST CELL-SPECIFIC SOMATIC RECOMBINATION AND PROTEIN CHAIN COMBINATORIAL DIVERSITY

Gene	Gene segment	Number of functional copies	Number of VDJ or VJ combinations	Number of different immunoglobulin chains	Number of genes that can make a chain	Total number of different immunoglobulins
IGH (14q32)	Variable region (V)	~40	40 × 25 × 6 = 6000 VDJ combinations	6000 heavy chains	Two alleles can make a heavy chain	6000 × 2 × 335 × 4 = ~16,080,000 immunoglobulins
	Diversity region (D)	~25				
	Joining region (J)	6				
IGK (2p11)	Variable region (V)	~40	40 × 5 = 200 VJ combinations	335 light chains	Two alleles of *IGK* + two alleles of *IGL* can make a light chain	
	Joining region (J)	5				
IGL (22q11)	Variable region (V)	~30	30 × 4.5 = 135 VJ combinations			
	Joining region (J)	4 to 5				

Somatic recombination alone generates up to 6000 × 335 = 2,010,000 different immunoglobulins. But in each B cell just one type of immunoglobulin is made. Only one of the two *IGH* alleles is (randomly) selected to make a heavy chain, and only one of the four possible light-chain genes (from the two alleles each of *IGK* and *IGL*) is randomly selected to make a light chain. If we factor in protein chain combinatorial diversity, the number of different possible immunoglobulins made by a B cell therefore increases to over 16 million. In addition, but not shown here, is the added contribution of junctional diversity, and also somatic hypermutation (in the case of immunoglobulins)—see text.

Functions of major histocompatibility complex (MHC) proteins and the concept of MHC restriction

The HLA complex is the human major histocompatibility complex (MHC). The latter name came from the observation that certain MHC genes are the primary determinants in transplant rejection. That is an artificial situation: the normal function of MHC genes largely consists of helping T cells to identify host cells harboring an intracellular pathogen such as a virus.

Some MHC genes—called classic MHC genes—are extremely polymorphic. They are deployed on the cell surface as heterodimers (see Figure 11.14), where they serve to bind peptide fragments derived by intracellular degradation of pathogen proteins. After being displayed on the surface of host cells, individual MHC–peptide fragment complexes can be recognized by T cells with a suitable T-cell receptor (**antigen presentation**). Appropriate immune reactions are then initiated to destroy infected host cells. There are two major classes of classic MHC protein, as detailed below.

Class I MHC proteins

Class I MHC proteins are expressed on almost all nucleated host cells. Their job is to help cytotoxic T cells (CTLs) to recognize and kill host cells that have been infected by a virus or other intracellular pathogen. When intracellular pathogens synthesize protein within host cells, a proportion of the protein molecules get degraded by proteasomes in the cytosol. The resulting peptide fragments are transported into the endoplasmic reticulum (Figure 11.17A). Here, a newly formed class I MHC protein binds a peptide and is exported to the cell surface where it is recognized by a CTL with a suitable receptor.

Because of cell-specific somatic recombinations, individual CTLs make unique T-cell receptors that recognize *specific* class I MHC–peptide combinations (Figure 11.17B). If the bound peptide is derived from a pathogen, the CTL induces killing of the host cell. Note that a proportion of normal host-cell proteins also undergo degradation in the cytosol and the resulting self-peptides are bound by class I MHC proteins and displayed on the cell surface. But there is normally no immune response (starting in early fetal life, CTLs that recognize MHC–self-peptide are programmed to be deleted to minimize autoimmune responses).

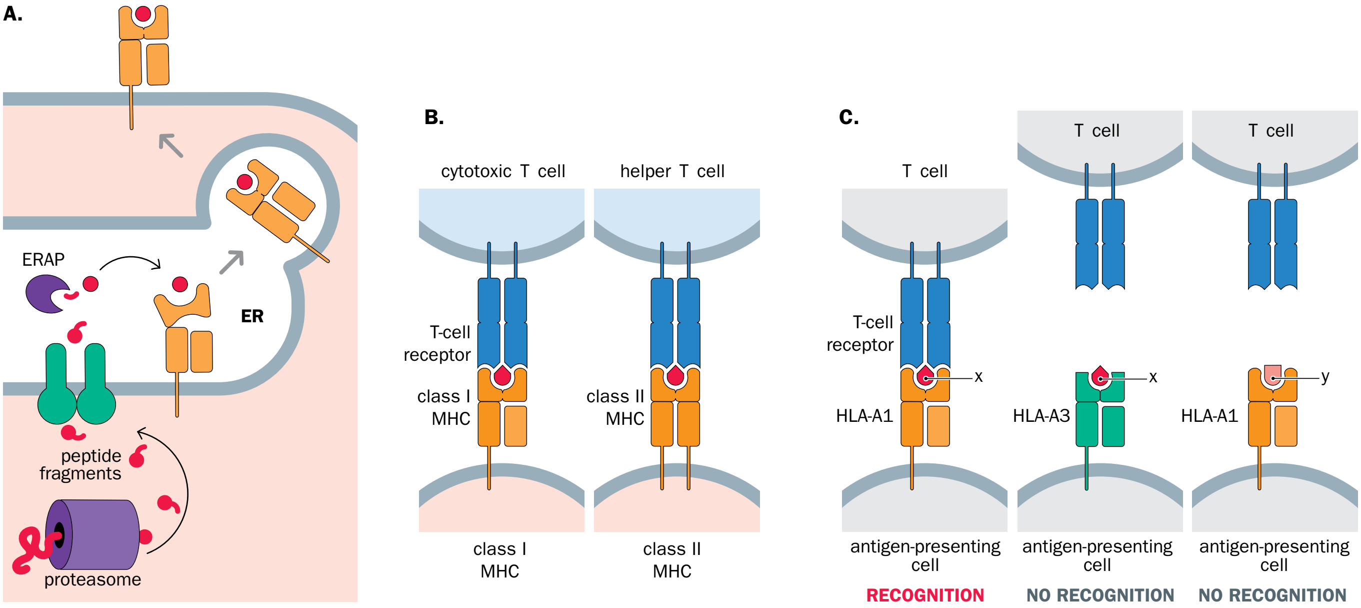


Figure 11.17 **MHC peptide binding, antigen presentation, and MHC restriction.** (**A**) Class I MHC proteins (known as class I HLA proteins in human cells) serve to bind peptides and display them on the cell surface. The peptides are produced by the degradation of any protein synthesized within the cell (either a host-cell protein or one made by a virus or other intracellular pathogen). Peptide fragments are produced within the proteasome and transported into the endoplasmic reticulum (ER). Here they are snipped by an endoplasmic reticulum aminopeptidase (ERAP) to the proper size needed for loading on to a partly unfolded class I HLA protein. Once the peptide has been bound, the HLA protein completes its folding and is transported to the plasma membrane with the bound peptide displayed on the outside. (**B**) Receptors on cytotoxic T cells bind class I MHC–peptide complexes; those on helper T cells bind class II MHC–peptide complexes. (**C**) MHC restriction. T cells have cell-specific receptors that recognize a combination of a specific peptide and a specific MHC protein. (Adapted from Murphy K [2011] *Janeway's Immunobiology*, 8th edn. Garland Science. With permission from WW Norton.)

Class II MHC proteins

Class II MHC proteins are expressed in professional antigen-presenting cells: dendritic cells, macrophages, and B cells. These cells also express class I MHC proteins but, unlike most cells, they make **co-stimulatory molecules** needed to initiate lymphocyte immune responses.

Whereas class I MHC proteins bind peptides from *endogenous* proteins (that were made within the cytosol, such as a viral protein made after infection of that cell), class II MHC proteins bind peptides derived from *exogenous* proteins that were transported into the cell (by endocytosis of a bacterium or other microbial cell or its products) and delivered to an endosome where limited proteolysis occurs. The resulting peptide fragments are bound by previously assembled class II MHC proteins and transported to the cell surface so that a helper T lymphocyte with an appropriate receptor recognizes a specific class II MHC–peptide combination (see **Figure 11.17B**). (Helper T cells play critical roles in co-ordinating immune responses by sending chemical signals to other immune system cells.)

MHC restriction

T cells recognize foreign antigens only after they have been degraded and become associated with MHC molecules (**MHC restriction**; **Figure 11.17C**). A proportion of all normal proteins in a cell are also degraded and the resulting peptides are displayed on the cell surface, complexed to MHC molecules. MHC proteins cannot distinguish self from nonself, and even on the surface of a virus-infected cell the vast majority of the many thousands of MHC proteins are associated with bound peptides derived from host cell proteins rather than from virus proteins.

The rationale for MHC restriction is that it provides a simple and elegant solution to the problem of how to detect intracellular pathogens—it allows T cells to survey a peptide library derived from the entire set of proteins in a cell *but only after the peptides are displayed on the cell surface*.

What is the basis of the extraordinary polymorphism of MHC proteins?

The polymorphism of the classic MHC genes is unusual in several respects. There is a very high number of alleles at each locus (**Table 11.10**), and because of the generally low allele frequencies, individuals are often heterozygotes at the classic MHC loci. There is also a very high frequency of nonsynonymous base substitution, consistent with some type of positive selection for new variants. Importantly, the substituted amino acids are not just restricted to the variable domains (which form the peptide-binding groove), but are clustered at specific sites, occupying positions that line the peptide-binding groove.

TABLE 11.10 SEQUENCE VARIATION AT THE SIX MOST POLYMORPHIC HLA LOCI

HLA LOCUS	-A	-B	-C	-DPB1	-DQB1	-DRB1
Alleles (DNA variants)	4340	5212	3930	1014	1257	2593
Protein variants	2980	3700	2661	692	838	1878

Data were derived from the European Bioinformatics Institute's IMGT/HLA database (release 3.33.0, July 2018). The statistics for these and additional loci are available at http://www.ebi.ac.uk/ipd/imgt/hla/stats.html.

In the past, selection may have driven gene duplication in the MHC: by producing multiple classic class I and class II loci, more scope was provided for generating protein diversity. But it is clear that natural selection also works to promote the very high levels of polymorphism at the classic MHC loci, generating protein variants that exhibit different peptide-binding specificities.

The prevailing theory is that selection is pathogen-driven. **Balancing selection** (also called overdominant selection) has been invoked to favor heterozygotes (*heterozygote advantage*). Heterozygosity at classic MHC loci might be expected to confer greater protection from pathogens than homozygosity, and so a higher chance of transmitting genotypes to future generations. Diversifying *frequency-dependent selection* has alternatively been proposed, whereby selection is skewed toward initially rare MHC alleles that confer a fitness advantage in response to new pathogenic variants.

Any selection operating on MHC genes must be longstanding, and is likely to have originated before the speciation event leading to evolutionary divergence from the great apes. MHC polymorphism is exceptional in showing **trans-species polymorphism**: some HLA protein variants are more closely related to the equivalent chimpanzee protein variants than they are to other protein variants at the same HLA locus. For example, human HLA-DRB1*0701 and HLA-DRB1*0302 show 31 amino acid differences out of 270 amino acid positions, but human HLA-DRB1*0701 and chimpanzee Patr-DRB1*0702 show only two differences out of the 270 positions.

An additional component seems to be MHC-driven disassortative mating. In many species, the choice of mate seems to be influenced by the MHC: individuals tend to choose MHC-different mates to have progeny with. If the differences in the MHC of the two parents are great, there is a greater chance that progeny are heterozygous at MHC loci. The basis of this type of *sexual selection* is not well understood. If it were to be based on olfaction, it may not be so significant in some contemporary human cultures where perfumes, aftershave lotions, and so on may be expected to obscure the natural olfactory cues.

The medical importance of the HLA system

The HLA system is medically important for two principal reasons. First, the high degree of HLA polymorphism poses problems in organ and cell transplantation. Second, certain HLA alleles are risk factors for individual diseases, notably many autoimmune diseases and certain infectious diseases; other HLA alleles are protective factors, being negatively correlated with individual diseases.

Transplantation and histocompatibility testing

Following organ and cell transplantation, the recipient's immune system will often mount an immune response against the transplanted donor cells (the *graft*), which carry different HLA proteins to those of the host cells. The immune reaction may be sufficient to cause rejection of the transplant (but corneal transplants produce minimal immune responses—the cornea is one of a few **immune privileged sites** that actively protect against immune responses in several ways, including having much-reduced expression of class I HLA proteins).

Bone marrow transplants and certain stem cell transplants can also result in graft-versus-host disease (GVHD) when the graft contains competent T cells that attack the recipient's cells. GVHD can even occur when donor and recipient are HLA-identical because of differences in minor (non-HLA) histocompatibility antigens.

Immunosuppressive drugs are used to suppress immune responses following transplantation, but transplant success largely depends on the degree of HLA matching between the cells of the donor and the recipient. Histocompatibility testing (also called tissue typing) involves assaying HLA alleles in donor tissues so that the best match can be found for prospective recipients. The key HLA loci are the most polymorphic ones: *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1,* and *-DPB1* (see **Table 11.10** and **Box 11.4**).

HLA disease associations

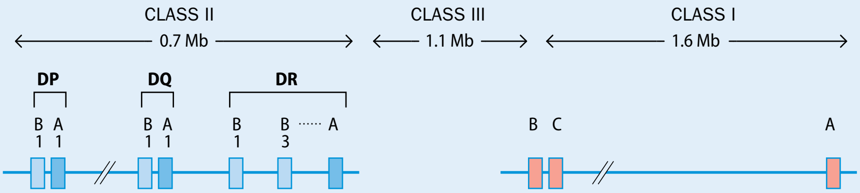
By displaying peptide fragments on host-cell surfaces, HLA proteins direct T cells to recognize foreign antigens and initiate an immune response against cells containing viruses or other intracellular pathogens. Because HLA proteins differ in their ability to recognize specific foreign antigens, people with different HLA profiles might be expected to show different susceptibilities to some infectious diseases.

In autoimmune diseases, the normal ability to discriminate self-antigens from foreign antigens breaks down and autoreactive T cells launch attacks against certain types of host cell. Certain HLA proteins are very strongly associated with individual diseases, such as type 1 diabetes and rheumatoid arthritis; in general, genetic variants in the HLA complex are the most significant genetic risk factors that determine susceptibility to autoimmune diseases. Determining to what extent HLA variants are directly involved in the pathogenesis and how much is contributed by other variants that lie in the immediate vicinity of the HLA genes (and outside of the HLA complex) is a major area of research—we consider HLA associations with individual diseases in some detail in Chapter 18.

BOX 11.4 HLA GENES, ALLELES, AND HAPLOTYPES

HLA GENES

The HLA complex spans 3.6 Mb on the short arm of chromosome 6. The 253 genes in the complex include 10 highly polymorphic classic HLA genes (**Figure 1**). They comprise seven class II HLA genes and three class I HLA genes. The intervening class III region does not contain any HLA genes; but it does contain multiple genes with an immune system function, including several complement genes.



Box 11.4 Figure 1 Classic (polymorphic) HLA genes within the HLA complex at 6p21.3. Genes in the class II HLA region encode α-chains (dark shading) and β-chains (pale shading) that pair up to form heterodimers of the HLA proteins HLA-DP, HLA-DQ, and HLA-DR. Classic class I HLA genes encode a polymorphic class I α-chain that forms a heterodimeric protein with the non-polymorphic β_2-microglobulin chain encoded by a gene on chromosome 15. Within the class I and class II HLA regions are several other nonpolymorphic HLA genes and many HLA-related pseudogenes not shown here. The class III region includes certain complement genes. Some additional genes with an immune system function are found within the HLA complex plus some functionally unrelated genes such as the steroid 21-hydroxylase gene.

HLA ALLELES (PROTEIN VARIANTS)

Because of their extraordinary polymorphism, alleles of the classic, highly polymorphic HLA genes have been typed for many decades at the protein level (using serological techniques with panels of suitably discriminating antisera). For example, the following can be distinguished this way: 28 HLA-A alleles; 50 HLA-B alleles, and 10 HLA-C alleles (called Cw for historical reasons; *w* signifies "workshop" because nomenclature was updated at regular HLA workshops).

Serological HLA typing is still used when rapid typing is required (for solid organ transplants, the time between chilling the organ and warming it after the blood supply is restored needs to be minimized). Much of modern HLA typing, however, is carried out at the DNA level where very large numbers of alleles can be identified (see **Table 11.10**). The complexity means that the nomenclature for HLA alleles identified at the DNA level is quite cumbersome—see **Table 1** for examples.

BOX 11.4 TABLE 1 HLA NOMENCLATURE

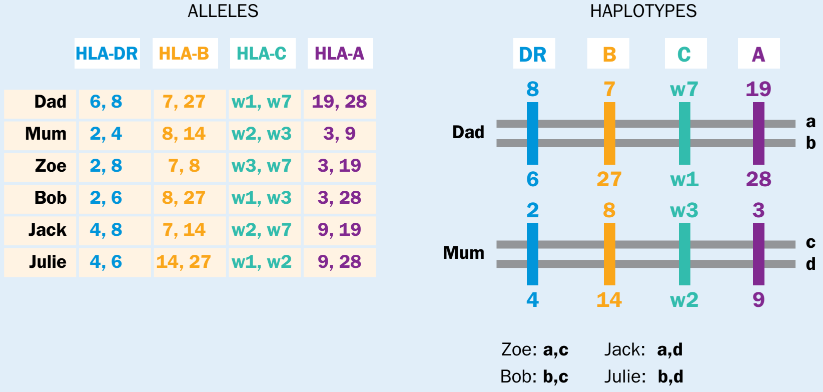
Nomenclature	Meaning
HLA-DRB1	An HLA gene (encoding the beta chain of the HLA-DR antigen)
*HLA-DRB1*13*	Alleles that encode the serologically-defined HLA-DR13 antigen
*HLA-DRB1*13:01*	One specific HLA allele that encodes the HLA-DR13 antigen
*HLA-DRB1*13:01:02*	An allele that differs from *DRB1*13:01:01* by a synonymous mutation
*HLA-DRB1*13:01:01:02*	An allele differing from *DRB1*13:01:01* by having a mutation outside the coding region
*HLA-A*24:09N*	A null allele that is related by sequence to alleles that encode the HLA-A24 antigen

For more details, see hla.alleles.org.

HLA HAPLOTYPES

The genes in the HLA complex are highly clustered, being confined to an area that represents only about 2% of chromosome 6. Genes that are close to each other on a chromosome are usually inherited together because there is only a small chance that they will be separated by a recombination event occurring in the short interval separating the genes. Such genes are said to be *tightly linked*.

A **haplotype** is a series of alleles at linked loci on an *individual* chromosome and the term was first used widely in human genetics with reference to the HLA complex. **Figure 2** shows how haplotypes are established by tracking inheritance of alleles in family studies. Note that because the HLA genes are very closely linked, recombination occurring within the HLA complex is rare.


ALLELES

	HLA-DR	HLA-B	HLA-C	HLA-A
Dad	6, 8	7, 27	w1, w7	19, 28
Mum	2, 4	8, 14	w2, w3	3, 9
Zoe	2, 8	7, 8	w3, w7	3, 19
Bob	2, 6	8, 27	w1, w3	3, 28
Jack	4, 8	7, 14	w2, w7	9, 19
Julie	4, 6	14, 27	w1, w2	9, 28

Zoe: **a,c** Jack: **a,d**
Bob: **b,c** Julie: **b,d**

Box 11.4 Figure 2 Deriving HLA haplotypes from family studies. Dad, mum, and their two daughters, Zoe and Julie, and two sons, Bob and Jack, have been tissue-typed using serological reagents for four HLA antigens as shown at left. By tracking which parental alleles have been passed on to individual children it is possible to deduce the parental HLA haplotypes. Dad has one chromosome 6 with the HLA haplotype DR8, B7, Cw7, A19 (haplotype **a**) and another chromosome 6 with the HLA haplotype DR6, B27, Cw1, A28 (haplotype **b**). Similarly, mum has haplotypes **c** (DR2, B8, Cw3, A3) and **d** (DR4, B14, Cw2, A9). Dad has transmitted haplotype **a** to Zoe and Jack, and haplotype **b** to Bob and Julie. Mum has transmitted haplotype **c** to Zoe and Bob, and haplotype **d** to Jack and Julie.

SUMMARY

- DNA damage and spontaneous mutations mean that the DNA in our cells accumulates changes (DNA variants).
- DNA damage can arise by chemical attacks on DNA from external agents, but mostly originates after highly reactive chemicals naturally produced inside our cells attack certain chemically unstable bonds and chemical groups in DNA. One or both DNA strands may be broken, bases or nucleotides may be deleted, or inappropriate chemical groups may be covalently bonded to the DNA.
- Spontaneous mutations largely result from unrepaired errors in DNA replication involving changes of one or a very few nucleotides.
- According to the type of DNA change, different cellular pathways are used to repair a DNA lesion. Direct reversal of chemical steps that cause a DNA lesion is rare, and some DNA changes are not successfully repaired.
- In population genomics, whole diploid genomes from multiple individuals from defined populations are sequenced, providing comprehensive data on human genetic variation.
- DNA changes can vary enormously in length, from whole chromosomal DNAs (by abnormalities in chromosome segregation and recombination) to single nucleotides.
- Most DNA variants in a population sample have low frequencies (rare variants); more common variants have traditionally been described as DNA polymorphisms (with a frequency >0.01) or minor alleles (frequency >0.05).
- A single nucleotide variant (SNV) or polymorphism (SNP) primarily denotes substitution of one nucleotide by another. It is sometimes also taken to include the variable presence of a single nucleotide, but the modern trend is to call that a single nucleotide indel (arising by insertion or deletion).
- The average human germ-line base-substitution rate is about 1.2×10^{-8} per nucleotide per generation, but it varies across the genome (C → T substitutions are particularly common in vertebrate DNA, for example), and it varies between families and the sexes.
- Because of the higher number of male germ-cell divisions and continuing production of sperm after puberty, the male germ-line mutation rate is much higher than the female germ-line rate, and doubles approximately every 17 years of age past puberty. On average, about 75% or so of *de novo* mutations are paternal in origin.
- A (short) indel is a site where variants differ by lacking or possessing one or a few (typically up to 50) nucleotides. It includes length variation produced by replication slippage at short tandem repeats (microsatellites).
- Structural variation results from larger-scale changes in DNA, typically changing >50 nucleotides at a time. In balanced structural variation, the variants do not differ in DNA content. In unbalanced structural variation, there is substantial length variation between variants. That sometimes arises by insertion of transposon repeats. More frequently, large deletions are involved, or the number of large repeated DNA sequences at a site changes, resulting in copy number variation (CNV).
- Of DNA changes, 99.9% are small scale (mostly base substitutions) but affect fewer nucleotides in the genome than structural variation. Comparing any two human haploid genomes, about 1–2 nucleotides out of every thousand will show differences on average because of small-scale changes, and a further 3 nucleotides per 1000 are changed by structural variation.
- Some mutations adversely affect functionally important DNA sequences. By affecting how genes work or are expressed, the mutations can contribute to disease. Because at least some people who carry deleterious DNA variants have reduced reproductive capacity (biological fitness), the variants tend to be removed from populations (purifying selection).
- Very occasionally, a DNA variant may result in some benefit and may accumulate in frequency if it endows individuals with increased reproductive capacity (positive selection).
- Protein variants arise not just as a result of allelic variation, but can also arise as a result of gene duplication and post-transcriptional changes.
- Extensive post-zygotic DNA variation means that we are all mosaics, with cells that have different genomes. Small-scale post-zygotic DNA changes occur mostly at random, but in each person there are extensive differences in structural variation in somatic cells, notably neurons.
- We inherit only three immunoglobulin and four T-cell receptor genes from each parent, but cell-specific DNA rearrangements in maturing B and T cells allow each of us to make huge numbers of different immunoglobulins and T-cell receptors.
- Our most polymorphic proteins are produced by genes in the HLA complex (the human major histocompatibility complex or MHC). HLA proteins recognize and bind peptides from processed foreign antigens and present them on cell surfaces so that they can be recognized by specific T-cell receptors.
- The extreme polymorphism of HLA proteins means that recipients of tissue or organ transplants often mount strong immune responses to the foreign tissue. Tissue typing seeks to find reasonable matches between HLA proteins expressed by donor tissue and prospective recipients.
- The polymorphism of MHC proteins is believed to be pathogen-driven and driven by balancing selection to favor heterozygotes (heterozygote advantage).

FURTHER READING

DNA damage and DNA repair

Barnes DE & Lindahl T (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet* **38**:445–476; PMID 15568983.

Ciccia A & Elledge SJ (2010) The DNA damage response: making it safe to play with knives. *Mol Cell* **40**:179–204; PMID 20965415. (Has detailed tabulation of frequencies of different types of DNA damage and of inherited disorders of DNA damage responses/DNA repair.)

Kowalczykowski SC (2015) An overview of the molecular mechanisms of recombinational DNA repair. *Cold Spring Harb Perspect Biol* **7**:a016410; PMID 26525148.

Kunkel TA & Erie DA (2015) Eukaryotic mismatch repair in relation to DNA replication. *Annu Rev Genet* **49**:291–313; PMID 26436461.

DNA variation classes and human genetic variation maps

Alkan C *et al*. (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**:363–376; PMID 21358748.

Conrad DF *et al*. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**:704–712; PMID 19812545.

Montgomery SB *et al*. (2013) The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res* **23**:749–761; PMID 23478400.

Spielmann M *et al*. (2018) Structural variation in the 3D genome. *Nat Rev Genet* **19**:453–467; PMID 29692413.

Sudmant PH *et al*. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* **526**:75–81; PMID 26432246.

Zarrei M *et al*. (2015) A copy number variation map of the human genome. *Nat Rev Genet* **16**:172–183; PMID 25645873.

1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* **526**:68–74; PMID 26432245.

Human mutation and germ-line mutation rates

Acuna-Hidalgo R *et al*. (2016) New insights into the generation and role of *de novo* mutations in health and disease. *Genome Biol* **17**:241; PMID 27894357.

Aggarwala V & Voight BF (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* **48**:349–355; PMID 26878723. (A heptanucleotide context is proposed to explain >81% of variability in substitution probabilities; new mutation-promoting dinucleotide and tetranucleotide motifs are described.)

Campbell CD & Eichler EE (2013) Properties and rates of germline mutations in humans. *Trends Genet* **29**:575–584; PMID 23684843.

Goldmann JM *et al*. (2016) Parent-of-origin-specific signatures of *de novo* mutations. *Nat Genet* **48**:935–939; PMID 27322544. (See also the “News and Views” article in the same issue by Anne Goriely [pp. 823–824].)

Kivisild T (2015) Maternal ancestry and population history from whole mitochondrial genomes. *Investig Genet* **6**:3; PMID 25798216.

Lynch M (2016) Mutation and human exceptionalism: our future genetic load. *Genetics* **202**:869–875; PMID 26953265.

Makova KD & Hardison RC (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet* **16**:213–223; PMID 25732611.

Rahbari R *et al*. (2016) Timing, rates and spectra of human germline mutation. *Nat Genet* **48**:126–133; PMID 26656846.

Ségurel L *et al*. (2014) Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* 15:47–70; PMID 25000986.

Shendure J & Akey JM (2015) The origins, determinants, and consequences of human mutations. *Science* **349**:1478–1483; PMID 26404824.

Postzygotic (somatic) DNA variation

Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* **107**:961–968; PMID 20080596.

Martincorena I & Campbell PJ (2015) Somatic mutation in cancer and normal cells. *Science* **349**:1483–1489; PMID 26404825.

Martincorena I *et al*. (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**:880–886; PMID 25999502.

McConnell MJ *et al*. (2013) Mosaic copy number variation in human neurons. *Science* **342**:632–637; PMID 24179226.

Weissman IL & Gage FH (2016) A mechanism for somatic brain mosaicism. *Cell* **164**:593–595; PMID 26871622.

Functional variation and natural selection

Boyko AR *et al*. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**:e1000083; PMID 18516229.

Fu W & Akey JM (2013) Selection and adaptation in the human genome. *Annu Rev Genomics Hum Genet* **14**:467–489; PMID 23834317.

Hulse AM & Cai JJ (2013) Genetic variants contribute to gene expression variability in humans. *Genetics* **193**:95–108; PMID 23150607.

Hurst LD (2009) Genetics and the understanding of selection. *Nat Rev Genet* **10**:83–93; PMID 19119264.

Lek M *et al*. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**:285–291; PMID 27535533.

Ward LD & Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**:1675–1678; PMID 22956687.

Genetic variation in the adaptive immune system

Murphy K & Weaver C (2016) *Janeway’s Immunobiology*, 9th edn. Garland Science.

Parham P (2009) *The Immune System*. Garland Science.

Shiina T *et al*. (2009) The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet* **54**:15–39; PMID 19158813.

Spurgin LG & Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc Biol Sci* **277**:979–988; PMID 20071384.

Human population genetics

12

Population genetics is about allele frequencies (often, though not strictly correctly, called gene frequencies). It is concerned to examine the factors that determine allele frequencies, how they may differ between populations or between subsets of individuals within a population, and how they may change or be changed. Information on how often two alleles at a locus are identical, either in a single person or between two people, helps characterize a population, identify population substructure, and quantify the risk a couple faces of having babies with an autosomal recessive disease. As well as individual alleles, we need also to consider **haplotypes**, combinations of alleles at two or more loci on the same chromosomal segment, whose frequencies may not be predictable from the individual allele frequencies. Identifying conserved ancestral haplotype blocks allows tagging SNPs (single nucleotide polymorphisms) to be defined, which are essential tools for the genome-wide association studies of common disease described in Chapter 18. Population genetics, in combination with evidence from archaeology and linguistics, has generated fascinating insights into human prehistory and the origins of populations; these aspects are covered in Chapter 14. Population genetic considerations are also important when considering proposals for population-wide genetic screening; these are covered in Chapter 20. Here we will start by exploring the concepts of gene (allele) and haplotype frequencies and their relation to genotype frequencies.

12.1 ALLELE FREQUENCIES AND GENOTYPE FREQUENCIES: THE HARDY–WEINBERG RELATIONSHIP

Over a whole population there may be many different alleles at a particular locus, although at any autosomal locus, each individual person has just two alleles, which may be identical or different. The **gene pool** for the *A* locus consists of all alleles at that locus across the population. The frequency of allele *A*1 is the proportion of all *A* alleles in the gene pool that are *A*1.

A thought experiment: picking genes from the gene pool

Consider two alleles at the *A* locus, *A*1 and *A*2. Let their frequencies be p and q, respectively (p and q are each between 0 and 1). Imagine a thought experiment:

- Pick an allele at random from the gene pool. There is a chance p that it is *A*1 and a chance q that it is *A*2.
- We put the first allele back into the pool and then make a second pick at random. Again, the chance of picking *A*1 is p and the chance of picking *A*2 is q.

It follows that:

- The chance that both alleles were *A*1 is p^2.
- The chance that both alleles were *A*2 is q^2.
- The chance that the first allele was *A*1 and the second *A*2 is pq. The chance that the first was *A*2 and the second *A*1 is qp. Overall, the chance of picking one *A*1 and one *A*2 allele is $2pq$.

If we pick a person at random from the population, this is equivalent to picking two alleles at random from the gene pool. Staying with our alleles *A*1 and *A*2, the chance that the person is *A*1*A*1 is p^2, the chance that they are *A*1*A*2 is $2pq$, and the chance that they are *A*2*A*2 is q^2. This simple relationship between allele frequencies and genotype frequencies is called the **Hardy–Weinberg distribution**. It holds whenever a person's two alleles are drawn independently and at random from the gene pool. Note that the Hardy–Weinberg relationship is a distribution; it is a common student error to suppose it is the equation $p^2 + 2pq + q^2 = 1$. That equation is only true if *A*1 and *A*2 are the only alleles at the locus, so that $p + q = 1$. There was nothing in our thought experiment to say that should be so. There might have been all manner of other alleles; the frequencies of the *A*1*A*1, *A*1*A*2, and *A*2*A*2 genotypes would still hold; just, there would be other genotypes as well.

The Hardy–Weinberg distribution provides an important check on population genotyping data. If a random sample of unrelated individuals from a population has been genotyped, the first check on the data is to see whether the types follow the Hardy–Weinberg distribution. Consider, for example, the data in Box 12.1.

BOX 12.1 CHECKING FOR HARDY–WEINBERG DISTRIBUTION

Ten thousand unrelated Chinese individuals were typed for the MN blood group. There are three main types, M, N, and MN, determined by co-dominant alleles at the Glycophorin A locus on chromosome 4. The survey results were:

M	MN	N
3156	4997	1847

Are these numbers compatible with the Hardy–Weinberg distribution?

First, calculate the allele frequencies. Individuals of type M have two *M* alleles, those of type N have two *N* alleles, and MN individuals have one of each. Hence there are (2 × 3156) + 4997 = 11,309 *M* alleles and (2 × 1847) + 4997 = 8691 *N* alleles. The frequencies are 11,309/20,000 = 0.5655 (*M*) and 8691/20,000 = 0.4345 (*N*).

Now calculate the expected genotype frequencies, if the Hardy–Weinberg distribution holds, and compare with the observed frequencies:

	MM	MN	NN
Observed	3156	4997	1847
Expected frequency	p^2 $= (0.5655)^2$ $= 0.3197$	$2pq$ $= 2 \times 0.5655 \times 0.4345$ $= 0.4914$	q^2 $= (0.4345)^2$ $= 0.1888$
Expected number	3197	4914	1888
(Obs – Exp)^2/Exp	0.526	1.402	0.890

χ^2 (one degree of freedom) = 2.818 $0.05 < p < 0.1$

Remember that for the chi-squared test you use actual numbers, not proportions

The distribution is not significantly different from the prediction on the null (Hardy–Weinberg) hypothesis. Note that chi-squared has only one degree of freedom since $p + q = 1$; once p is defined, q, $2pq$, and q^2 follow automatically. If there are k alleles, there are $k - 1$ degrees of freedom.

If the observed distribution is significantly different from the Hardy–Weinberg expectation, there are a number of possible explanations:

- Genotyping errors—maybe some heterozygotes are being missed, and misclassified as homozygotes. Check using an independent technique.
- Population stratification—maybe the population is not homogeneous but comprises two or more relatively isolated subpopulations. In the example in Box 12.1, might the sample have been taken in Xinjiang province, where the large ethnic Uighur population interbreed rather little with Han Chinese brought in by mass immigration? Each may within itself be in Hardy–Weinberg equilibrium, but with different allele frequencies.
- Loss of specific genotypes before sampling—for example, if the character being studied were a serious autosomal recessive disease, affected homozygotes might have been present at birth in the expected proportion, but many might have died before the adult population was sampled.
- Inbreeding—if many couples in a population are consanguineous (blood relatives), when we consider the genotypes of their children, one of the prerequisites of the random-picking model we used to establish the Hardy–Weinberg distribution is violated. The second allele is not completely independent of the first because relatives share genes more than expected by chance. The effect is to increase the proportion of both homozygous genotypes and decrease the proportion of heterozygotes, compared to the Hardy–Weinberg expectation. We consider this further in Section 12.4.

Note that when the chi-squared test shows no significant deviation from the Hardy–Weinberg expectation, this does not necessarily mean that none of the above forces is acting; it merely means that if any is acting, it is not sufficiently strong to produce a clear deviation, given the number of individuals tested. Statistically the chi-squared test has quite low power to detect deviations.

Predicting carrier frequencies for Mendelian conditions

The Hardy–Weinberg distribution is also very useful for predicting risks in genetic counseling. The basic pedigree pattern for a simple Mendelian autosomal recessive condition like cystic fibrosis was shown in **Figure 5.4**. We can classify people into healthy and affected, but without a lot of work in the molecular laboratory we cannot tell how many of the healthy people are heterozygote gene carriers. Sometimes it would be very useful to know that figure. Suppose, for example, the healthy sister of a boy with cystic fibrosis is getting married. Her fiancé is an unrelated healthy man with no family history of cystic fibrosis. However, either of them might be a carrier, and if both of them were, the risk of an affected child would be 1 in 4. They might like to know how great that risk is, in order to decide whether to take genetic tests and how to manage any pregnancy.

The man is, for genetic purposes, an individual picked at random from the population, and so his risk of being a carrier is $2pq$. Knowing the incidence of cystic fibrosis in that population, we can use the Hardy–Weinberg distribution to calculate that risk. Call the disease-causing allele a and the corresponding normal allele A. Suppose that, in that population, 1 in 2000 newborn babies have cystic fibrosis (regardless whether they are diagnosed at that time or later):

	Unaffected		**Affected**
Genotype	AA	Aa	aa
H–W frequency	p^2	$2pq$	$q^2 = 1/2000$

$q^2 = 1/2000$, therefore $q = \sqrt{(1/2000)} \approx 1/45$. $p = 1 - q = 44/45$, so $2pq = 2 \times 44/45 \times 1/45 = 1/23$.

The fiancé has a 1 in 23 chance of being a carrier. For the woman, her risk follows from simple Mendelian segregation (**Figure 12.1**).

The chance that both partners are carriers is $1/23 \times 2/3$, and the risk that any child of theirs would be affected is $1/23 \times 2/3 \times 1/4$ = 1 in 138. Should they risk having children? That is their own personal decision. The job of the genetic counselor is to explain the risk, to make sure they fully understand it, to help them arrive at their own decision, and to support them in that decision, whatever it is. The counselor might point out that any pregnancy is at 1–2% risk of producing an abnormal baby, so the extra risk due to cystic fibrosis should be seen in that context, but it is not for the counselor to tell them what to do.

For X-linked loci (see **Figure 5.5**), males, being **hemizygous** (having only one allele), are A or a with frequencies p and q, respectively, whereas females can be AA, Aa, or aa (**Figure 12.2**). So, for example, if a survey shows that one man in 12 has X-linked red–green color blindness, we can see directly that $q = 1/12$, and so we would predict that one woman in 144 would be affected, and 22 in 144 would be carriers. In reality the genetics of red–green color blindness is rather more complicated, but hopefully this simplified example serves present purposes.

A major complicating factor for these simple calculations is inbreeding. This is considered in Section 12.4.

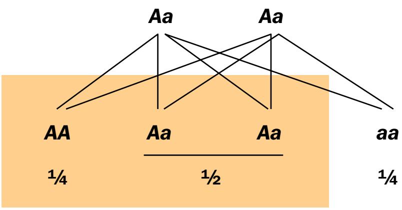


Figure 12.1 Carrier risk for the unaffected sister of a person with cystic fibrosis. She has an affected brother, so both her parents must be carriers (there is a negligible rate of fresh mutation for cystic fibrosis). The genotype probabilities for their children are as shown. The chance the unaffected sister is a carrier is not 1 in 2; it is 2 in 3 because we know she has one of the genotypes in the shaded box.

	Autosomal locus			**X-linked locus**				
				Males		Females		
Genotype	AA	Aa	aa	A	a	AA	Aa	aa
Frequency	p^2	$2pq$	q^2	p	q	p^2	$2pq$	q^2

Figure 12.2 Hardy–Weinberg expectation of genotype frequencies for an X-linked locus with alleles A (frequency p) and a (frequency q).

12.2 HAPLOTYPE FREQUENCIES AND LINKAGE DISEQUILIBRIUM

Allele frequencies at different loci are not always independent of one another. **Figure 12.3** shows five SNPs in the Interleukin 8 (*IL8*) gene. The authors of this study estimated the frequency of each allele by genotyping 190 European chromosomes.

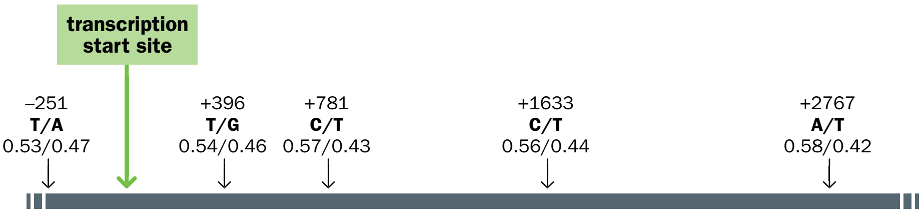


Figure 12.3 Five SNPs in the Interleukin 8 (*IL8*) gene. Positions are shown relative to the transcription start site. (Data from Hull J *et al.* [2001] *Am J Hum Genet* **69**:413–419; PMID 11431705.)

There are 2 × 2 × 2 × 2 × 2 = 32 possible combinations of types at the five SNPs. We can calculate the expected frequency of each combination by multiplying up the allele frequencies. However, the actual frequencies turned out to be very different from those expectations (Table 12.1).

TABLE 12.1 HAPLOTYPES FOR THE FIVE SNPS SHOWN IN FIGURE 12.3 IN 190 EUROPEAN CHROMOSOMES

Combination of alleles (in chromosomal order)	Expected frequency	Observed frequency
T - T - C - C - A	0.053	0.52
A - G - T - T - T	0.017	0.41
A - G - C - C - A	0.040	0.03
A - T - C - C - A	0.047	0.01
A - G - T - T - A	0.024	0.01
T - G - T - T - T	0.019	0.005
A - T - C - C - A	0.047	0.005
A - G - C - T - T	0.023	0.005
A - T - T - T - A	0.028	0.005

There are 32 possible combinations, but none of the other 23 possible haplotypes was observed in this sample. Data from Hull J *et al.* (2001) *Am J Hum Genet* **69**:413–419; PMID 11431705. SNP, single nucleotide polymorphism.

This is an example of **linkage disequilibrium** (LD; Box 12.2). The chromosomal segment exists as a block that has only rarely been broken up by recombination during the common ancestry of the study subjects. Of the 32 possible combinations of alleles, over 90% of the chromosomes studied carried either T-T-C-C-A or A-G-T-T-T. Linkage disequilibrium is a quantitative phenomenon: the relationship between alleles at two loci can lie anywhere on a spectrum from completely unrelated to perfectly correlated. Box 12.2 outlines some of the measures that have been used to express the degree of LD in a dataset.

BOX 12.2 MEASURES OF LINKAGE DISEQUILIBRIUM (LD)

If two loci have alleles A,a and B,b with frequencies P_A, P_a, P_B, and P_b, there are four possible haplotypes: AB, Ab, aB, and ab. Suppose that the frequencies of the four haplotypes are P_{AB}, P_{Ab}, P_{aB}, and P_{ab}. If there is no LD, $P_{AB} = P_A P_B$, and so on, because the haplotype will be constructed just by random assortment of the constituent alleles. The degree of departure, D, from this random association can be measured by $D = P_{AB} - P_A P_B$. If there is no linkage disequilibrium, $D = 0$.

As a measure of LD, D suffers from the property that its maximum absolute value depends on the allele frequencies at the two loci, as well as on the extent of disequilibrium. Among preferred measures are:

- $D' = D/D\text{max}$, where Dmax is the maximum value of D possible with the given allele frequencies. The maximum value of D' is 1.
- $r^2 = D^2/(P_A P_a P_B P_b)$. r^2 is the squared correlation coefficient between the two loci. Again, the maximum value is 1.

Typing 25 SNPs surrounding the Interleukin 8 gene and testing all pairs for linkage disequilibrium produced the result shown in **Figure 12.4**. Across this 550 kb region there are evidently three separated blocks of strong linkage disequilibrium ($r^2 > 0.3$). SNPs in different blocks do not show any correlation, suggesting that there has been free recombination between variants in adjacent blocks. Detailed studies of recombination show that this is a common situation. Although, at first sight, crossovers in meiotic recombination appear randomly distributed along pairs of chromosomes, this is only true on a low-resolution view. Both population genetic studies (Myers *et al.* [2005], PMID 16224025) and direct typing of sperm (Jeffreys *et al.* [2000], PMID 10749979) have shown that meiotic recombination is concentrated in 1–2 kb hotspots. In humans there are maybe 30,000 such hotspots, typically occurring every 50–100 kb across the human genome. Hotspots carry the epigenetic histone methylation mark H3K4me3 (see **Box 10.2**) imposed by sequence-specific binding of the PRDM9 histone methyltransferase enzyme (see Baudat *et al.* [2010], PMID 20044539, in Further Reading).

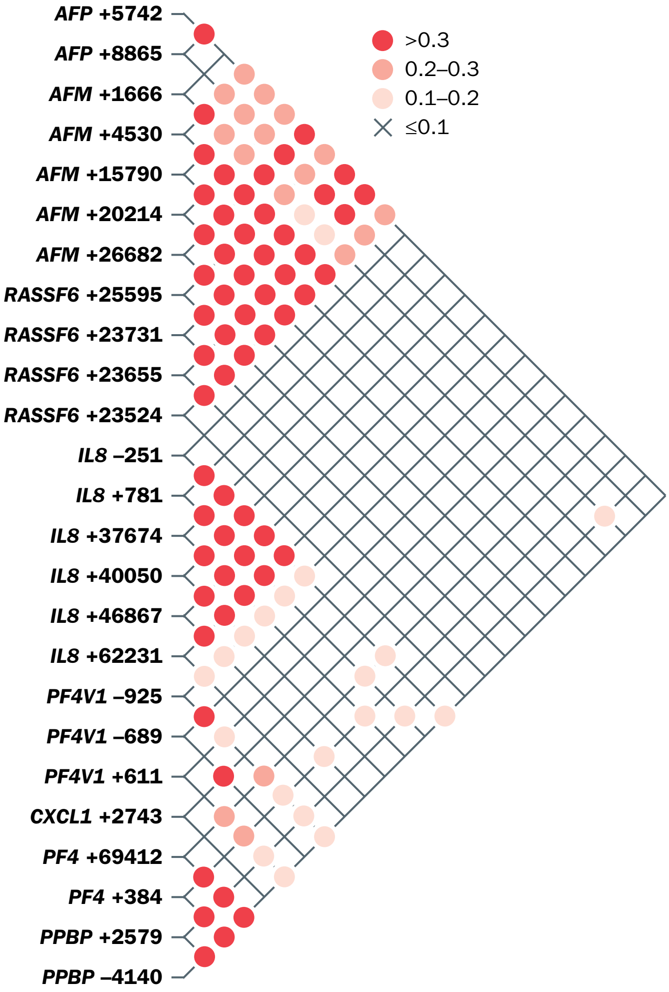


Figure 12.4 Patterns of linkage disequilibrium between 25 SNPs across a 550 kb region surrounding the Interleukin 8 gene. Pairwise correlations are defined by the r^2 statistic (see **Box 12.2**). The region contains three conserved haplotype blocks. (Reprinted from Hull J *et al.* [2004] *Hum Genet* **114**:272–279; PMID 14605870. With permission from Springer International Publishing AG. Copyright © 2004.)

Systematic genome-wide studies of this phenomenon were initiated by the International HapMap Project (the data originally at http://hapmap.ncbi.nlm.nih.gov/ have been relocated to http://www.1000genomes.org/). In Phase I of the project, a consortium of research institutions in America, Asia, and Europe typed one million SNPs in 269 individuals drawn from four human populations: 30 American parent–child trios of white European origin from Utah (CEU); 30 Yoruba parent–child trios from Ibadan, Nigeria (YRI); 45 individual Han Chinese from Beijing (CHB); and 44 individual Japanese from Tokyo (JPT). All were ostensibly healthy and, although not a formal population sample, were hopefully typical of the population from which they were drawn. The Phase I report (International HapMap Consortium [2005], PMID 16255080) summarized results from typing 1,007,329 SNPs (**Table 12.2**, **Figure 12.5**). SNPs were chosen to represent common variants (minor allele frequency [MAF] ≥0.05) and spaced across the whole genome, ideally one per 5 kb. Phase II of the project (International HapMap Consortium [2007], PMID 17943122) added a further 2.1 million SNPs, giving an average of one SNP per kilobase. Phase III (International HapMap 3 Consortium [2010], PMID 20811451) extended the analysis to seven other populations (Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Tuscans in Italy; Gujarati Indians in Houston, Texas; metropolitan Chinese in Denver, Colorado; people of Mexican origin in Los Angeles; and people with African ancestry in the southwestern United States).

TABLE 12.2 HAPLOTYPE BLOCK STRUCTURE OF HUMAN GENOMES

Sample	YRI	CEU	CHB + JPT
Average number of SNPs per block	30.3	70.1	54.4
Average length of block (kb)	7.3	16.3	13.2
Fraction of genome spanned by blocks (%)	67	87	81
Average number of haplotypes (MAF ≥0.05) per block	5.57	4.66	4.01
Fraction of chromosomes due to haplotypes with MAF ≥0.05 (%)	94	93	95

Data from the HapMap Phase I study. YRI, Yorubans from Ibadan, Nigeria; CEU, people of northern European ancestry from Utah, USA; CHB, Han Chinese from Beijing; JPT, Japanese from Tokyo. MAF, minor allele frequency. The precise numbers in the table are a function of the statistical definition of a block; a different definition identifies shorter but more numerous blocks, but the overall picture is similar. Data from International HapMap Consortium (2005) *Nature* **437**:1299–1320; PMID 16255080.

Overall, the results showed similar block structures to the Interleukin 8 data shown in Figure 12.4. Figure 12.5 shows data for one chromosomal region in the CEU panel. Comparing the recombination data in the lower part of the figure with the blocks in the upper part, we see that block boundaries reflect locations with high recombination rates within the sample. Not all recombination hotspots identified in a previous genome-wide analysis (small red triangles) were active in the history of this sample of 30 CEU trios, but a subset marks the boundaries seen here.

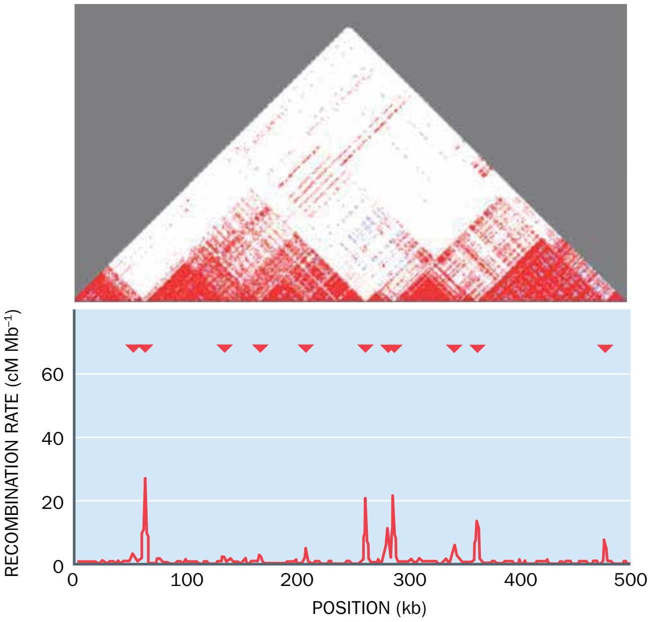


Figure 12.5 Haplotype blocks. HapMap data for a 500 kb region at chromosome 7q31.33 in the CEU sample. Upper part: pairwise correlations of SNPs. Lower part: estimated recombination rates across the region. Note how the boundaries between blocks in the upper part of the figure, marked by the peaks of recombination at the bottom of the figure, coincide with a subset of the small, red, inverted triangles below the upper figure, which indicate recombination hotspots defined in an earlier independent analysis of genome-wide recombination. (Adapted from International HapMap Consortium [2005] *Nature* **437**:1299–1320; PMID 16255080. With permission from Springer Nature. Copyright © 2005.)

Considering just the common variants studied by the HapMap consortium, there is less variability between genomes than would be suggested by the number of SNPs. For example, in Table 12.2 the average CEU block contains 70 SNPs. In principle this would allow 2^{70} different combinations—vastly more than the number of people in the world—but, in practice, an average of less than five common haplotypes per block accounted for

93% of the chromosomes. The lack of variety means that most of the common variation can be captured by typing a small number of “tagging” SNPs. This has turned out to be a key driver of genome-wide association studies, as we will see in Chapter 18. Given genotypes at a few tagging SNPs, genotypes at other SNPs within a block can, if necessary, be guessed (“imputed”) with good though not perfect accuracy.

As the emphasis has moved from chip-based SNP analysis to sequencing, a more complete picture of variation has emerged, encompassing less common variants. SNPs are often categorized as common (MAF ≥0.05), low-frequency (MAF 0.05–0.005), and rare (MAF <0.005). As described in Chapter 11, the 1000 Genomes Project and other population genomics projects (see **Table 11.3**) are providing a detailed account of the full extent of sequence variation between and within populations, which supersedes the HapMap data. One use of these data is to construct population-specific reference sequences that allow missing genotypes to be imputed more accurately from microarray data than when using just the HapMap data (see the discussion in Section 18.3).

Rarer variants are likely to be younger, originating as relatively recent mutations compared to the common SNPs that define the main haplotype blocks in **Table 12.2**. Those haplotype blocks represent ancestral chromosome segments that have been transmitted intact through many generations. There may have been subsequent mutations, but the blocks have not been broken up by recombination. More recent mutations will affect particular examples of a given block, so that there is a degree of heterogeneity layered on top of the basic structure revealed by the common and ancient SNPs. At any particular chromosomal location, most European genomes will have one of only around five alternative ancestral segments as defined by the high-frequency ancient SNPs. That does not mean that we are all descended from just five “cavemen”. At the next location, again there may be just five or so common alternative ancestral segments, but they will be from a different set of ancestors. Our remote ancestry is with populations, not individuals, as described in Chapter 14.

The whole topic of ancestry is rather more complicated than it might appear at first sight. Ancestry enthusiasts are playing a highly artificial game when they go back more than a few generations. You have two parents, four grandparents, eight great-grandparents, and so on; *n* generations ago you had 2^n ancestors. Thirty generations, around 1000 years ago, you had 2^{30} ancestors, which is over 1000 million. These cannot all have been separate individuals: there were not 1000 million people in the world at that date, let alone in the region where the people you may consider as your ancestors lived. Any two present-day Europeans, no matter that they may come from opposite ends of the continent, are reasonably likely to share common ancestors within the last 1000 years, and are certain to share many within the last 2500 years (see Ralph & Coop [2013], PMID 23667324). Looking further back, everybody in the world shares common ancestors, which is why we function as a single species, with genome structures sufficiently similar to one another that matings between any two humans are potentially fertile.

It nevertheless still makes sense to talk of more-specific ancestry. Many of your 1000 million ancestors from 1000 years ago will be the same person over and over again, connected to you by multiple different lines of descent through the generations. It might be that a disproportionate number of the individuals who feature many times in your list of 1000 million ancestors lived in Denmark. In that case it would be meaningful to say your ancestry is mostly Danish.

The 1000 million (or whatever the real number is, after eliminating repeats) are *genealogical* ancestors—individuals who would feature in your genealogy if you were able to trace it that far back. The number of *genetic* ancestors—people from 1000 years ago from whom you inherit significant parts of your genome—though still large, will be much smaller than that. In each successive generation there is a 1 in 2 chance of *not* inheriting any specific piece of parental DNA, so all of the DNA from most of your genealogical ancestors will have been lost over 30 generations. What does survive will be in quite small segments. Consider the way chromosomes are broken up during meiosis. On average there are 50–60 crossovers in male meiosis and maybe 90 in females. Something over 2000 crossovers will have affected the genome of a genetic ancestor living 1000 years ago before you inherit any of his or her genetic material. Most of these will have occurred at recombination hotspots, but there are over 30,000 hotspots, so most of the crossovers will be at different locations when mapped onto your ancestor’s genome. There will be a wide distribution of possible segment sizes that you inherited from each of your 1000-year-old genetic ancestors, but dividing your genome size by the total number of crossovers suggests that the average would be around 1 megabase, or a hundred or so haplotype blocks. Your genome has space for only a few thousand genetic ancestors from 1000 years ago, from whom you will have inherited chromosomal segments varying widely in size but averaging around a megabase. Further back you will have more genetic ancestors, but the chromosomal segments you inherit from each will on average be smaller.

The haplotype blocks, being much smaller than the 1000-year-old segments, reflect much deeper ancestry, and the structures provide important insights into human evolution, as described in Chapter 14. Each generation brings fresh opportunities for recombination, as well as fresh mutations, and so blocks are shorter but more diverse in older populations, and longer in newer. The shorter blocks and higher average number of common haplotypes in the Nigerians (**Table 12.2**), compared to Europeans or East Asians, are consistent with the "Out of Africa" hypothesis of human evolution, while unusually long haplotypes can be evidence of directional selection.

Two special parts of the genome are free of recombination and each is inherited as a single haplotype from a single ancestor, no matter how many generations back one goes:

- Mitochondrial DNA is inherited only from the mother, not from the father. Sperm have mitochondria to power their swimming, but those do not get incorporated into the zygote. A person's complete mitochondrial DNA is inherited in a block from the mother, from the mother's mother, from the mother's mother's mother, and so back as far as one cares to go. Whereas n generations back you have 2^n autosomal ancestors, you have only one mitochondrial ancestor, who is both your genealogical and your genetic ancestor.
- Similarly, the nonrecombining portion of a man's Y chromosome is inherited only through the male line and as an unbroken block, so that a man has a single Y-chromosome ancestor, even many generations back.

The insights into ancestry and population history that can be derived from typing mitochondrial, Y-chromosome, and autosomal DNA are described in Chapter 14.

12.3 CHANGING ALLELE FREQUENCIES

Allele frequencies are subject to change over time because of a number of factors. Principal among these are mutation, drift, and selection. We will consider these in turn here before looking at their combined effects.

Estimating mutation rates

The molecular processes generating mutations were described in Section 11.1. In Section 11.3 we described various ways in which human mutation rates have been estimated. Early efforts considered the amount of DNA sequence divergence between two species—for example, humans and chimpanzees—and related this to their time of separation as estimated from the fossil record. This gave a figure for mutations per base pair per year, but subject to many uncertainties. Not only is the date of separation of the two lineages usually debatable, but one must also make assumptions about the relative rates at which sequence variants accumulated in the two lines once they had diverged. That may require assumptions about generation times in the two lines in the remote past. Fortunately, our new-found ability to perform whole-genome sequencing in parents and offspring has provided unbiased estimates for human mutation rates and has rendered earlier approaches obsolete, at least when considering the present-day population.

As long suspected, the frequency of novel single nucleotide variants in a person depends sharply on the age of their father when they were conceived (see **Box 11.2**). The study by Kong *et al.* (2012) (PMID 22914163; see Further Reading) documented a doubling of the paternal mutation rate every 16.5 years. On average a child carried 60 new single nucleotide mutations, but the actual figure depended strongly on the age of the father. Mothers, regardless of age, transmitted around 15 mutations; for fathers the figure ranged from 25 at age 20 to 65 at age 40. The existence of such a strong effect makes it difficult to give any general figure for the human mutation rate—average paternal age has varied significantly both over recent time in Western societies and between different types of society.

Despite these uncertainties, one can make some useful generalizations about the fate of new mutant alleles. For neutral mutations the important factor is genetic drift, described below, which depends critically on population size. For disease-causing mutations, considering all allelic mutations that cause a given Mendelian disease, there is a relation between the frequency of fresh mutations, the degree of selection against the condition, and the incidence of the disease (see **Box 12.3**). The relationship depends critically on the mode of inheritance. Advantageous mutations may spread through a population in a selective sweep, with nearby neutral variants hitchhiking along with the selected variant.

Neutral variation and genetic drift

The gametes that create the next generation of a population contain a sample of the alleles that were present in the parental population—but it is only a sample. It may not perfectly reflect the composition of the parental population. A given allele might be slightly more or slightly less frequent among the successful gametes that make up the following generation than it was across the whole parental generation. An allele that was present at very low frequency in the parental population might, by chance, not be present in any of the gametes that make up the next generation. In that case, barring fresh mutation or inward migration, that allele would be lost forever from the population. The alternative allele (in a two-allele system) has become **fixed**. Thus allele frequencies vary in a random way between generations, even without mutation, selection, or migration. This is called **genetic drift**.

It is easy to simulate genetic drift on a computer, provided one makes the simplifying assumption that generations are discrete and non-overlapping, like annual plants but unlike real human populations. A random number function is used to select alleles to go into the gametes, and the process can be repeated for as many generations as desired. Such simulations illustrate several features of genetic drift. In a biallelic system with no mutation, selection, or migration, one or other of the alleles will inevitably eventually reach fixation, when its frequency is 100%. If alleles *A*1 and *A*2 have initial frequencies *p* and *q*, when the simulation is repeated many times, *A*1 will end up fixed in a proportion *p* of the runs and *A*2 in *q*. The size of fluctuations, and the time to fixation, depend critically on the number of gametes used to create each successive generation. If the number is large, a very large average number of generations will be needed to reach fixation. In small populations the random variations due to drift can be larger, and fixation can occur more rapidly (**Figure 12.6**).

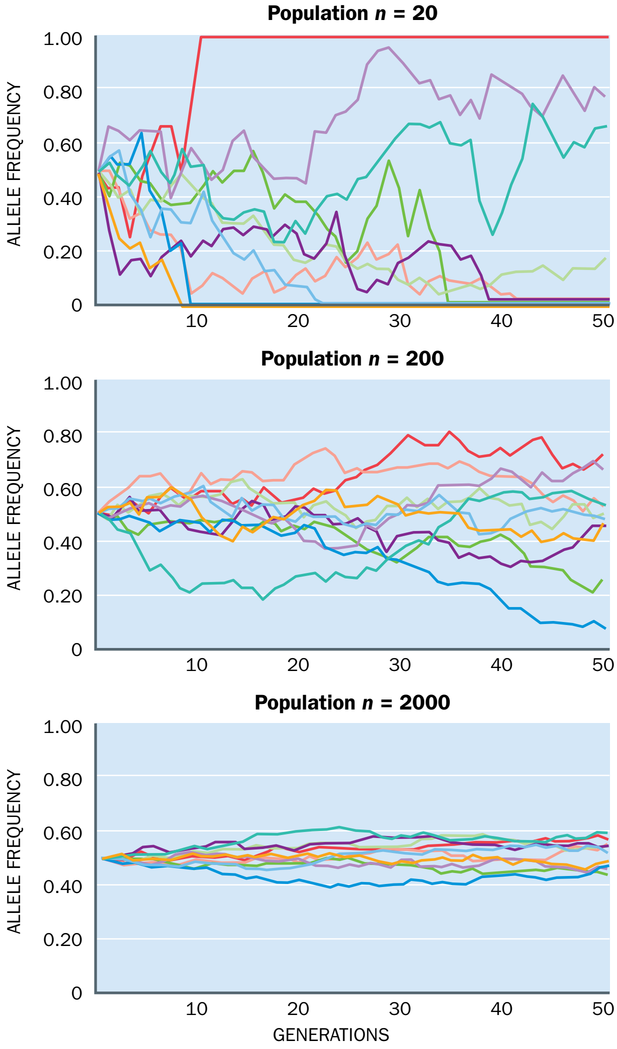


Figure 12.6 Computer simulations of genetic drift. Ten trajectories of allele frequency over 50 generations of random drift, starting from a value of 0.5, in a population of constant size 20 (top), 200 (middle), or 2000 (bottom) individuals. Drift effects are much stronger in the smaller populations. (Source https://biologydictionary.net/genetic-drift.)

Real populations have complications compared to the models used in these simple simulations. Generations overlap, there is always some nonrandom mating, and the population size will vary over time. The **effective population size**, N_e, is the size of an ideal population, as in the simple model of Figure 12.6, that would give the same amount of drift as the actual population. N_e is almost always less than the actual census population, not least because real population censuses include children, parents, and grandparents, of whom only the parents are likely to contribute to the immediately following generation. The concept is further explored in Box 14.2.

The mathematical geneticist WJ Ewens calculated the mean number of generations, T, for one allele in a two-allele system to reach fixation. For a population of size N with initial allele frequencies p_0 and q_0 ($q_0 = 1 - p_0$), $T = 4N[p_0\log_e(p_0) + q_0\log_e(q_0)]$ where $\log_e$ is the natural logarithm. For $p_0 = 0.5$ the expression in brackets is 0.69 and T = 2.8N. If the initial allele frequencies are closer to 0 or 1, T is a lower multiple of N. As Figure 12.6 shows, there is wide individual variation.

Measures of diversity

Various measures are used to summarize the genetic diversity within a population. Nei's gene diversity statistic, h, measures the extent of heterozygosity at a locus; h is 1 minus the summed frequencies of homozygotes. It is the probability that two alleles at a locus, drawn at random from the population, will be different from each other. The **population mutation parameter**, θ, measures the diversity per nucleotide site that is expected due to the balance between mutation creating new alleles and drift to fixation eliminating them, all in the absence of selection. It is defined (for autosomal loci) as $\theta = 4N_e\mu$, where N_e is the effective population size and μ the mutation rate. θ can be compared to π, the actual diversity that includes the effects of mutation, drift, and selection. More detail on these measures, how to calculate them, and what they mean can be found in Chapters 5 and 6 of Jobling *et al.* (2014) (see Further Reading) or in population genetics textbooks.

A long-term small effective population size will reduce the genetic diversity of a population. Two historical events can also contribute substantially to reduced diversity: bottlenecks and founder effects (Figure 12.7). In either case, at one time in its history a population consisted of only a small number of individuals—at the very beginning with founder effects, or as the result of some historical misfortune in the case of a bottleneck. In both cases, diversity in subsequent generations is reduced, and allele frequencies may differ significantly from those in earlier generations (the original source population in the case of founders). Examples of currently quite large populations where diversity reflects founder effects or past bottlenecks include Finns and Ashkenazi Jews. From a medical genetic perspective, they have their own particular spectrum of recessive diseases, with some diseases that are rare elsewhere being relatively common, while other diseases are much less frequent than in many related populations. The overall conclusion is that drift

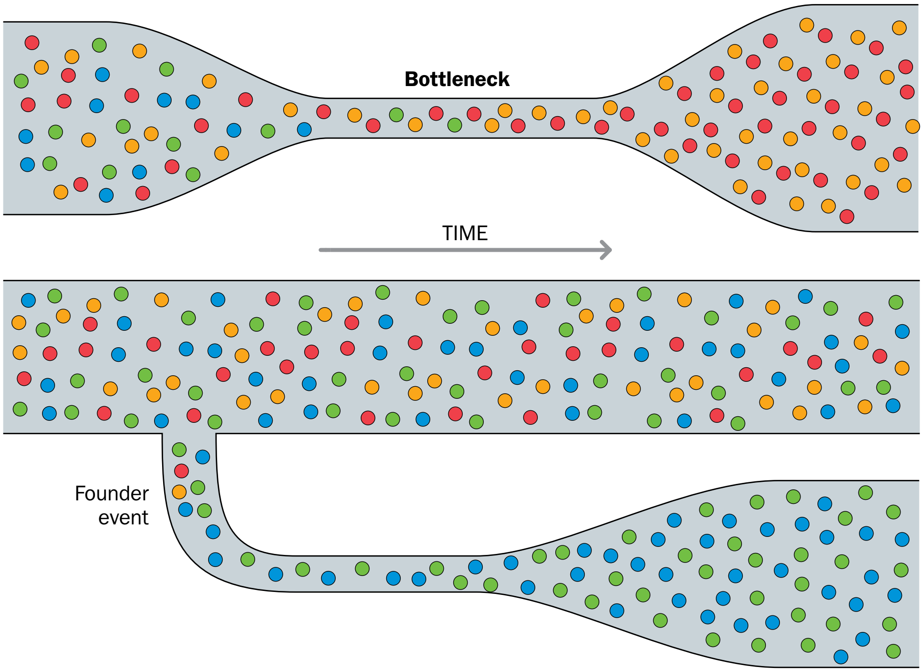


Figure 12.7 Genetic bottlenecks and founder effects. When a population either has been through a time when effective population size was very small (top) or stems from a small number of initial founders (bottom), diversity is reduced compared to the parent population. Circles of different colors represent different alleles. (From Jobling M *et al.* [2014] *Human Evolutionary Genetics*, 2nd edn. Garland Science.)

can be significant in very small populations (including large populations that have been through a bottleneck), but has negligible effects if a population has always been large.

Selection

Mutation and drift operate at the level of genes: they directly affect allele frequencies. Selection is different. It works on the phenotype, and only indirectly on the genotype. Most phenotypes do not depend on a single genotype, but on the combined effects of genotypes at numerous loci, together with epigenetic and environmental factors. **Figure 12.8** shows a typical example. Human birth weight is subject to purifying selection: there is an optimum birth weight, and perinatal mortality is higher among babies that are either lighter or heavier than the optimum. However, a great range of factors affect birth weight, many of which are not genetic at all, so that the graph gives no information on selective effects on any specific genotype.

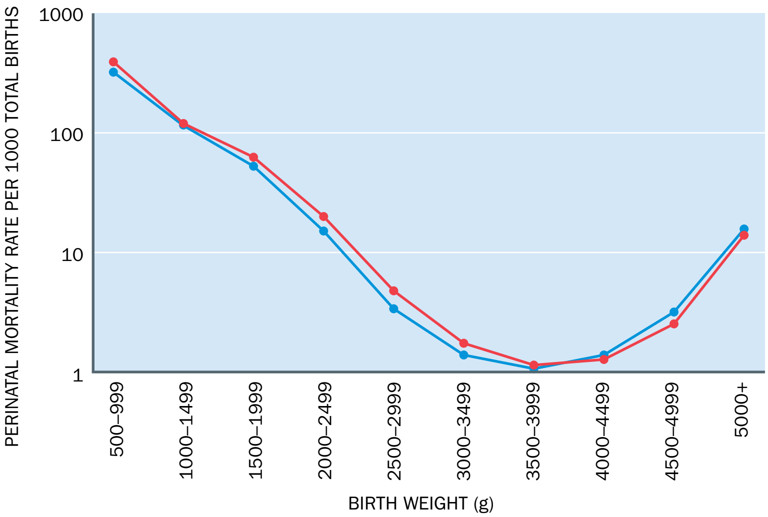


Figure 12.8 Perinatal mortality versus birth weight for 7,628,847 singleton births in the USA in 1997–8. Males, red; females, blue. (From Joseph KS *et al.* [2005] *BMC Pregnancy Childbirth* **5**:3; PMID 15720720. With permission from BioMed Central Ltd.)

Despite the uncertainty, signals of selection may be detectable in genomes. A number of different statistical analyses have been used:

- In coding sequences, the ratio of **nonsynonymous** to **synonymous** substitutions, K_a/K_s, can carry a signal of selection (**Figure 12.9A**). To take account of the structure of the genetic code, the frequency of nonsynonymous substitutions per nonsynonymous site is compared to the frequency of synonymous substitutions per synonymous site (see **Box 13.2** for more detail). In functional sequences, nonsynonymous changes that replace one amino acid with a different one are often deleterious and will be removed by purifying selection, whereas synonymous changes (that replace one codon for an amino acid with another for the same amino acid) are less likely to affect the function of the gene product. Thus in coding sequences subject to purifying selection, K_a/K_s is likely to be lower than the value predicted from random changes or the value found in nonselected coding sequences. The test can be applied to variation within current populations, to identify functional regions, or to comparisons between humans and other species (see, for example, **Figure 11.12**). Note that for some functions there may be selective pressure in favor of generating new amino acid sequences. For example, in the immune system it is important to keep ahead in the arms race between hosts and pathogens, so novelty can be an advantage.

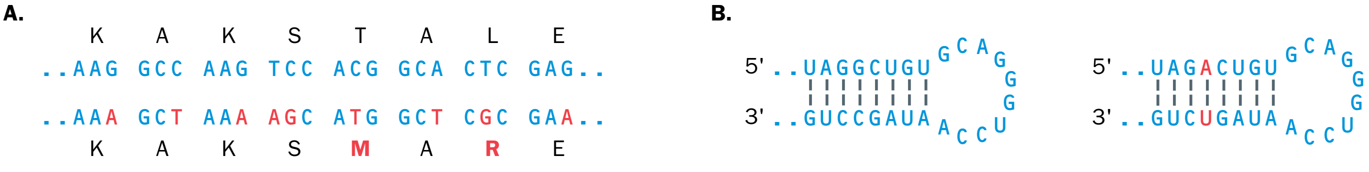


Figure 12.9 Signals of purifying selection. (**A**) There are nine differences (red) in the DNA sequence of these hypothetical orthologous coding sequences, but only two of them cause a change at the amino acid level. (**B**) Compensatory second substitutions in this hypothetical RNA sequence maintain the stem-loop structure (note that G can pair with U in double-stranded RNA).

- For functional noncoding RNAs, secondary structure may be at least as important as the nucleotide sequence. Here a signal of purifying selection could be compensatory double substitutions that preserve stem-loop structures (**Figure 12.9B**).
- A similar argument suggests that functional regions, whether coding or noncoding, should show lower levels of sequence diversity across the population than nonfunctional sequences. Low diversity could be the result of purifying (negative) selection, but it could also reflect a **selective sweep**, where an advantageous variant has spread through the population. After selection, the haplotype that carries the advantageous variant, together with all its associated SNPs, will be present at high frequency, to the detriment of all other haplotypes and alternative SNP alleles. Thus many neutral variants can "hitchhike" on one advantageous variant and rise to high frequencies. This will be true whether the advantageous variant has just arisen *de novo* or whether it is an old variant that has become advantageous in a new environment.
- Haplotypes involved in recent selective sweeps would be expected to be unusually long. There has not been time for recombination to occur at every hotspot in the vicinity, so the selected haplotype can include several adjacent blocks where older haplotypes have been broken up. An extended haplotype test can pick up a signal of selection, but additional data such as functional analysis would be needed to identify the actual beneficial variant.
- If the allele frequencies at a certain SNP are very different in two related human populations where most SNPs show similar allele frequencies, this may signal selection in one of the populations.

Various examples are discussed in Section 14.4. In practice, signals that could be marks of selection often have alternative explanations rooted in demography (drift, migration, or population fluctuations) or random chance. Sophisticated methods are used to try to tease apart the effects. The reader should consult an up-to-date specialist textbook (for example, Jobling *et al*. [2014], see Further Reading) for more detail.

Only when there is a simple direct relationship between the phenotype and the genotype at a single locus—in other words, only for fully penetrant Mendelian characters (see Chapter 5)—can we use simple models of selective effects on a single locus. Such simple population-genetic models use the **biological fitness**, f, and **coefficient of selection**. The biological fitness of a genotype is defined as the average number of offspring of persons with that genotype that survive to reproductive age, compared to the average number from the fittest type in the population. It varies from 0 (no surviving offspring) to 1 (the fittest genotype). The coefficient of selection, **s**, is 1 – f. Biological fitness is not the same as fitness in the everyday sense—a person might be a champion athlete, but if they happen to be infertile their biological fitness is zero.

A new mutant allele causing a deleterious dominant condition will be eliminated by purifying selection at a rate that depends on its degree of deleteriousness. At the extreme, any allele causing a fitness of zero (a genetic lethal, whether or not it is also a physical lethal) will be eliminated immediately, and every case must be due to a new mutation. Thus deleterious dominant conditions have a high turnover of mutant alleles, and a high proportion of cases are due to new mutations. A 1941 study by the Danish physician ET Mørch illustrates this. Mørch studied achondroplasia, an autosomal dominant form of dwarfism (OMIM #100800). People with achondroplasia are fertile, of normal intelligence, and can live full lives, but they are inevitably disadvantaged in finding a partner. Mørch's survey showed that on average they had only one-fifth the number of surviving babies as people of normal stature. Meanwhile his survey of newborns in Copenhagen revealed 10 achondroplastic babies in 94,075 consecutive births. Only two of the ten had an affected parent; the rest were new mutants. Some of Mørch's diagnoses have been questioned, but taken at face value his study is an excellent illustration of the dynamics of mutations and selection. Achondroplastics had a fitness one-fifth of their normal-stature counterparts, and four-fifths of achondroplastic babies were due to new mutations.

The same is true, though to a lesser degree, for deleterious X-linked conditions: one-third of X chromosomes are in males, so an allele causing a deleterious X-linked recessive condition is exposed to selection one-third of the time. For autosomal recessive conditions, on the other hand, selection acts much more slowly because the majority of all disease alleles are in phenotypically normal heterozygotes. Most cases are due to alleles inherited from parents who, unknown to anybody, were carriers. In most cases there is no previous family history of the condition, but nevertheless few cases are due to new mutations. Below, we show how exceedingly slowly selection would act against a lethal recessive condition, even if there were no fresh mutations.

If a deleterious Mendelian condition persists in the population over the generations with roughly the same incidence, there must be a balance between the loss of mutant alleles through selection and the gain through mutation. **Box 12.3** shows how the balance can be quantified.

BOX 12.3 MUTATION–SELECTION EQUILIBRIUM

Consider two alleles, *A* and a at a locus, with frequencies *p* and *q*, in a population of N individuals (2N alleles) that is in Hardy–Weinberg equilibrium. A disease is present at frequency F. Affected people have fitness f; the coefficient of selection against affected genotypes is s (s = 1 – f). Unaffected people have fitness 1. Normal alleles mutate into disease alleles at the rate μ per allele in each generation.

If the disease is autosomal dominant, by convention we use the uppercase letter for the mutant allele. The relation between phenotypes and genotypes is shown below:

Phenotype	**Normal**	**Affected**	
Genotype	*aa*	*aA*	*AA*
Frequency	p^2	$2pq$	q^2
A alleles lost per generation		$2pq$s	$2q^2$s
A alleles gained per generation	$2p^2\mu$ (there are 2 *a* alleles per normal person)		

For equilibrium, $2p^2\mu = 2pq\text{s} + 2q^2\text{s}$.

For a typical rare deleterious dominant condition, *p* will be almost 1 and q^2 negligibly small, so that virtually all affected people will be *Aa* heterozygotes, and our equation simplifies to

$$2\mu = 2pq\text{s},\ \mu \approx q\text{s} \text{ or } \mu = \tfrac{1}{2}\,\text{F}(1 - \text{f})$$

If the disease is autosomal recessive (disease allele *a*), a similar calculation would be:

Phenotype	**Normal**		**Affected**
Genotype	*AA*	*Aa*	*aa*
Frequency	p^2	$2pq$	q^2
a alleles lost per generation			$2q^2$s
a alleles gained per generation	$(2p^2+2pq)\mu$		

The loss is $2q^2$s because two *a* alleles are lost each time an affected person fails to reproduce. For equilibrium, $(2p^2 + 2\,pq)\mu = 2q^2\text{s}$. For a rare recessive condition, *p* will be almost 1 and *q* small, whereupon the equation simplifies to $2\mu \approx 2q^2\text{s}$, $\mu = \text{s}q^2$ or $\mu = \text{F}(1 - \text{f})$.

For an X-linked recessive condition, one-third of mutant alleles are in males and exposed to selection, so $\mu = \tfrac{1}{3}\,\text{F}(1 - \text{f})$.

The equations in **Box 12.3** have been used in the past to estimate mutation rates. However, if we apply them to cystic fibrosis we see a problem. Cystic fibrosis is an autosomal recessive condition. Affected people are homozygotes, *aa*, and in the past would not have lived to reproductive age. Even now there is very strong adverse selection: affected people have a lot to cope with in their daily lives that can get in the way of childbearing, and in any case, affected men are infertile because they have no vas deferens. Thus their biological fitness is still low, and was zero in times past when the current allele frequencies were evolving. The condition has an incidence of 1 in 2000 in Northern Europe, and so the equations of **Box 12.3** would suggest a pathogenic mutation rate of 5×10^{-4} per generation. That figure is not credible. It is orders of magnitude higher than pathogenic mutation rates for other genes, and, in addition, molecular studies virtually never identify new mutant cases. The high frequency is actually maintained by heterozygote advantage, not recurrent mutation.

The case of sickle cell disease is described in Section 14.4. This severe autosomal recessive condition is present at very high frequency in West Africa and in other populations where malaria is endemic, because heterozygotes are more resistant to malaria than normal homozygotes. A similar, though less dramatic, situation must underlie the high frequency of cystic fibrosis in Northern Europeans and some other populations: heterozygotes are fitter than either homozygote, though the coefficient of selection against the normal homozygotes may be very small. For the general case where heterozygotes have fitness 1, *AA* homozygotes have fitness $1 - \text{s}_1$, and *aa* homozygotes $1 - \text{s}_2$, a rather laborious calculation shows that a stable equilibrium is reached (that is, allele frequencies henceforth remain the same through the generations, ignoring new mutations) when $p/q = \text{s}_2/\text{s}_1$.

Applying the Hardy–Weinberg distribution to cystic fibrosis in Northern Europe, $q^2 = 1/2000$, therefore *q* is 0.022, *p* is 0.978, and so p/q is 0.978/0.022 = 44.45. If s_2/s_1 is 44.45 and s_2 is virtually 1, it follows that s_1 is 1/44.45 = 0.022. In other words, a stable gene frequency would have been established without the need for new mutations if heterozygotes on average had 2.2% more surviving children than normal homozygotes.

The nature of the advantage is unknown—possibly heterozygous babies are slightly more resistant to chloride-losing diarrhea, a major cause of infant mortality in the past. Such a slight advantage would be undetectable in historic population records even if we knew who the heterozygotes were. There is a general lesson here. Clinical geneticists are most concerned with conditions that are both common and severe, and this means that among recessive conditions they have a systematic bias toward those where heterozygote advantage may have played some role, and hence where naive calculations of mutation rates are likely to be grossly wrong. As sequencing is used more and more in families with genetic conditions, *de novo* cases can be identified, and eventually we should have true mutation rates.

Manipulating gene frequencies: the dream of eugenics

All our farm and domestic animals, and all our crop and garden plants, are the result of selective breeding that changed allele frequencies, fixed rare mutations, and subverted epigenetic controls. Probably the only items in a typical Western diet that have not been genetically modified in this way are wild-caught sea fish. The various breeds of dogs are a powerful witness to how far both physical and behavioral phenotypes can be modified by selective breeding. No doubt the same could be done with humans.

Eugenicists would like to improve humans by selective breeding. They might not wish to breed humans corresponding to the various dog breeds, but they may be very exercised by data suggesting that in many advanced societies, feckless people produce disproportionate numbers of the next generation. Proponents of negative eugenics would seek to eliminate "undesirable" types by discouraging or preventing them from breeding, while positive eugenicists would encourage those with especially desirable traits (among whom they naturally number themselves) to be especially fecund. Such ideas were part of the political and social mainstream in the USA and many European countries in the early twentieth century. Nowadays, having seen where such thinking led, most people are repelled by eugenic ideas, both on general moral grounds and through considering the practical arrangements that would be necessary to implement eugenic schemes. However, there is one corner of eugenic ideas that may seem more attractive: the idea of eliminating severe genetic diseases. Much of the eugenic literature and practice from 100 years ago was deeply flawed because it addressed undesirable behaviors like drunkenness and criminality as though they were simple genetic traits. That objection would not apply to thinking about true genetic disease.

Many devastating diseases are inherited as simple autosomal dominant, autosomal recessive, or X-linked characters, and surely everybody would feel the world would be a better place without them. There might be an argument from social responsibility as well. For example, in most advanced countries all newborn babies are screened for phenylketonuria, an autosomal recessive condition that, untreated, leads to severe intellectual disability (see Section 20.4). Affected babies are put on a special diet and, provided the parents and child can manage to stick rigidly to the diet, they will grow up intellectually normal and able to lead normal lives. That is likely to include having children. But genetically they are still homozygous for the mutant allele, and so will pass it on to every child. It might be argued that there should be a bargain: society pays for the test and diet, and in return they should agree not to pass on their genes.

Rather than agonizing about the ethics of such a bargain, we should look at the genetics. We have already seen that serious autosomal dominant conditions are largely maintained by recurrent fresh mutations. Affected people have low biological fitness, and preventing them altogether from reproducing would not eliminate the condition, although of course it would somewhat reduce the incidence. A similar argument applies to serious X-linked conditions. For autosomal recessive conditions we need to invoke the Hardy–Weinberg distribution:

Phenotypes	Unaffected		Affected
Genotypes	*AA*	*Aa*	*aa*
Frequency	p^2	$2pq$	q^2

Our proposed bargain involves affected (but successfully treated) people agreeing not to pass on their *a* genes. That would affect the transmission of a proportion $2q^2/(2pq + 2q^2)$ of all the *a* alleles in the population (remember, homozygotes have two copies of the relevant allele). Dividing numerator and denominator by $2q$, and remembering that $p + q = 1$, the expression simplifies to just q. So, if phenylketonuria has an incidence of 1/10,000 ($q = 0.01$), our bargain would prevent transmission of just 1% of the mutant alleles in the population, while having no effect on the 99% that are in healthy heterozygotes. If we

BOX 12.4 SELECTION AGAINST AN AUTOSOMAL RECESSIVE CONDITION

Consider a locus with alleles *A* and *a*, frequencies p_0 and q_0. Homozygous *aa* people have a recessive condition that we propose to eliminate by preventing them from reproducing.

Phenotype	Unaffected	Affected	
Genotype	*AA*	*Aa*	*aa*
Frequency	p_0^2	$2p_0q_0$	q_0^2

After one generation of selection we have eliminated $2q_0^2$ *a* alleles, but retained p_0q_0 (half the alleles contributed by heterozygous individuals are *a*).

In the reduced-diversity gene pool, the new *a* allele frequency, q_1, is $p_0q_0/(p_0^2 + 2p_0q_0)$.

Dividing numerator and denominator by p_0,

$$q_1 = q_0/(p_0 + 2q_0) = q_0/(1 + q_0)$$

Similarly, in generation 2,

$$q_2 = q_1/(1 + q_1)$$

$$= (q_0/[1 + q_0])/(1 + (q_0/[1 + q_0]))$$

Multiplying numerator and denominator by $1 + q_0$,

$$q_2 = q_0/(1 + q_0 + q_0) = q_0/(1 + 2q_0)$$

After *n* generations of complete selection,

$$q_n = q_0/(1 + nq_0)$$

For example, if q_0 was 0.01, q_{10} would be 0.0091 and q_{100} 0.005.

nevertheless persist, the calculation in **Box 12.4** shows what sort of reduction in incidence such a policy could achieve.

The calculation in **Box 12.4** shows that selection against recessive conditions is extremely inefficient. If q_0 is 1/100, as in phenylketonuria, it would take 100 generations of selection—around 3000 years—to halve the frequency. The only way such artificial selection could have a serious impact on the incidence of a recessive disease would be to identify all heterozygotes and prevent them from reproducing, or at least prevent them passing on their mutant allele through a program of compulsory prenatal screening and abortion. The problem there—ethical concerns aside—is that there are thousands of severe recessive conditions. Individually they may all be rare, but we are all carriers for more than one of them. We could not eliminate recessive disease without subjecting every one of us to such a program. Conceivably, in the future, genome editing (see Chapter 8) might provide an alternative method, but simple eugenic measures can never eliminate genetic disease.

12.4 POPULATION STRUCTURE AND INBREEDING

So far, we have used the word "population" as though it is self-evident how we could define a population, and as though mating within a population is perfectly random with respect to the partners' genotypes. Both these assumptions are gross simplifications. The population of the United States, for example, if defined as everybody living within the frontiers, includes people of many diverse ethnicities, and mating is not random with respect to ethnicity. If we try to limit our population to people of one particular ethnicity, we run into problems knowing where to draw the line with people of mixed origin. Thus the definition of a population for genetic purposes is always to some extent arbitrary. Every real population, however defined, has some degree of substructure. The question is how we can measure that and take it into account in our calculations.

All population substructure leads to **assortative mating**, violating the random mating requirement of the Hardy–Weinberg relationship. People within a subgroup are more likely to select partners from within their subgroup than from outsiders, and on average people in the same subgroup are more closely related than unselected people from the wider population. The extreme of assortative mating is **consanguinity**, where people marry blood relatives. Assortative mating can occur even without evident inbreeding and in a homogeneous population if, for example, people tend to pick partners of similar height or similar intelligence. In so far as variation in height or intelligence involves genetic factors, this is assortative mating in the genetic sense.

The Hardy–Weinberg principle can be modified to take account of assortative mating. We can imagine that the population comprises a fraction F who are completely inbred and homozygous at every polymorphic locus, while the remaining 1 – F are completely outbred. The proportions then become:

Genotype	*AA*	*Aa*	*aa*
Frequency	$(1 - \text{F})p^2 + \text{F}p$	$(1 - \text{F})2pq$	$(1 - \text{F})q^2 + \text{F}q$

Thus assortative mating decreases the proportion of heterozygotes and increases the proportions of the two homozygotes. It is also possible to imagine negative assortative mating—for example, a population divided into clans where people only marry outside their own clan. That would result in an increase of heterozygosity over the level predicted by the Hardy–Weinberg relationship.

Sewall Wright's F_{ST} statistic gives a measure of how different two (sub)populations are by considering how much heterozygosity would be gained if the two were a single random-mating population. F_{ST} varies between 0 (the two populations are identical) and 1 (complete genetic separation; different alleles are fixed at each locus in the two). F_{ST} is sometimes called a fixation index. One definition of F_{ST} is based on the variance (σ^2) of allele frequencies across the two populations:

$$F_{ST} = \sigma^2/\bar{p} \times \bar{q}$$

where $\bar{p}$ and $\bar{q}$ are the average frequencies of the two alleles in the total population. For comparisons based on genome sequence data, a more intuitive definition is:

$$F_{ST} = (\pi_{\text{between}} - \pi_{\text{within}})/\pi_{\text{between}}$$

where π_{between} is the average number of nucleotide differences between pairs of individuals from different populations (calculated as the total number across the samples divided by the number of individuals) and π_{within} is the same statistic for pairs of individuals within one population.

Various programs are available that calculate F_{ST} from genomic data using rather more elaborate formulae that avoid bias from unequal sample numbers or small sample sizes. For example, Nelis *et al.* (2009) (PMID 19424496; see Further Reading) calculated pairwise F_{ST} values for the four HapMap I populations (**Table 12.3**).

TABLE 12.3 PAIRWISE FST VALUES FOR THE FOUR HAPMAP I POPULATIONS

Population	CEU	JPT	CHB
YRI	0.153	0.192	0.190
CEU		0.111	0.110
JPT			0.007

The African sample (YRI, Yorubans from Ibadan, Nigeria) is roughly equally different from the Europeans (CEU, white Americans of Northern European origin from Utah) and East Asians (JPT, Japanese from Tokyo; CHB, Han Chinese from Beijing). The two East Asian samples are similar to each other and equally dissimilar to the Europeans. Data from Nelis M, Esko T, Mägi R *et al.* (2009) *PLoS One* **4**:e5472; PMID 19424496.

Identifying ancestry

The HapMap data (**Table 12.3**) show significant differences between the samples in terms of allele frequencies. An unknown individual could be assigned to one of the four populations by typing for the >1 million SNPs used in the study. There is considerable interest in using much smaller panels of markers to make much more detailed estimates of ancestry. A suitable panel of **ancestry informative markers** would comprise markers distributed across the genome that show large allele frequency differences between populations (say, F_{ST} >0.3), negligible linkage disequilibrium with one another, and a Hardy–Weinberg distribution within each of the tested populations. Various such panels have been assembled and tested in different populations (see, for example, Kidd *et al.* [2014], PMID 24508742, in Further Reading). Identifying the ancestry of an individual is a more demanding task than demonstrating differences between two population samples. One analysis (Pardo-Seco *et al.* [2014], PMID 24981136) suggests that the key factor is the number of markers typed, rather than the informativeness of individual markers, and that a panel should include at least 400 markers to be reliable. The GenoChip, a dedicated genotyping platform for genetic anthropology, uses approximately 12,000 Y-chromosomal SNPs, approximately 3300 mtDNA SNPs, and over 130,000 autosomal and X-chromosomal SNPs, all chosen to be free of any known health, medical, or phenotypic relevance (Elhaik *et al.* [2013], PMID 23666864; see Further Reading). As pointed out in Section 14.2, these markers can be used to identify the ancestry of an individual, within limits, but the results do not support the idea that people can be divided into different races.

Identifying substructure and ancestry is valuable in several ways. Genome-wide association studies (described in Chapter 18) use large population samples in case–control studies. Population stratification or differences between the case and control populations can produce false–positive associations. Checks on F_{ST} and ancestry can control for such factors. Law-enforcement agencies would like to be able to determine the ethnicity of a suspect whose DNA was found at a crime scene (see Chapter 20). Many individuals would like to know their ancestry just out of interest, for which purpose various companies offer DNA-based ancestry tests. In that context, it is worth noting that studies have shown a poor correlation between individual ancestry as inferred from a simple test of mitochondrial haplogroups and that determined using large panels of autosomal markers (see Emery *et al.* [2015], PMID 25620206, in Further Reading).

Consanguinity and the coefficient of relationship

When we consider individuals or couples rather than populations, their **coefficient of inbreeding** and **coefficient of relationship** are relevant for thinking about the risk of recessive disease. The coefficient of relationship of two individuals is the proportion of alleles they share that are **identical by descent**. The coefficient of inbreeding of an individual is the probability that, at a given locus, they receive two alleles that are identical by descent. It is one-half the coefficient of relationship of their parents. Note that identity by descent is not the same as **identity by state**. Two alleles are identical by state if they have the same DNA sequence, but they may or may not be identical by descent (Figure 12.10). Identity by state could be established by DNA sequencing. Identity by descent can only be unequivocally established by inspection of the pedigree, but it can be inferred probabilistically when the genotyping of an individual shows a long run of contiguous markers that are all homozygous. Shorter runs of homozygosity could be just coincidental.

For close relationships we can easily estimate the coefficient of relationship:

- First-degree relatives (parent and child, full sibs) share one-half their genes identical by descent—precisely one-half for parent and child, one-half on average for sibs.
- Second-degree relatives (grandparent–grandchild, uncle/aunt–nephew/niece, half-sibs who share just one parent) share one-quarter of their genes by descent, on average.
- Third-degree relatives (for example, first cousins) share on average one-eighth of their genes by descent.

For more distant or complicated relationships, the coefficient can be estimated using Sewall Wright's path coefficient method (Figure 12.11).

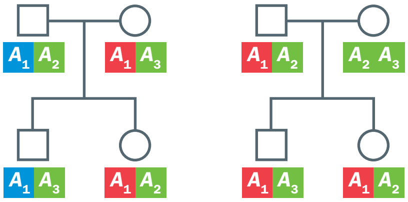
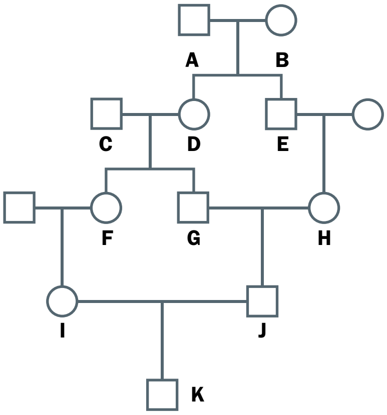


Figure 12.10 Identity by state and identity by descent. Both sib pairs share allele A_1. The first sib pair have two independent copies of A_1 (blue, red), indicating identity by state but not by descent. The second sib pair share copies of the same paternal A_1 allele (both red), showing identity by descent. The difference is only apparent if the parental genotypes are known.



Coefficient of relationship of I and J:

Path	Steps	Contribution
I-F-C-G-J	4	$(½)^4$ = 1/16
I-F-D-G-J	4	$(½)^4$ = 1/16
I-F-D-A-E-H-J	6	$(½)^6$ = 1/64
I-F-D-B-E-H-J	6	$(½)^6$ = 1/64
Total		= 10/64

Figure 12.11 Using Sewall Wright's path coefficient method to calculate a coefficient of relationship. To calculate the coefficient of relationship of individuals I and J, each possible pathway linking the two through a common ancestor is identified. Each path of n steps contributes $(½)^n$ to the coefficient of relationship. The coefficient of inbreeding of their offspring K is 5/64, half the coefficient of relationship of the parents.

Societies differ in the range of consanguineous marriages that they permit. Relationships that are too close are labeled incest and banned. Parent–child and brother–sister unions are almost universally considered incestuous. The only exception seems to have been in ancient Egypt, where the pharaohs practiced brother–sister marriage, presumably to avoid contaminating their royal blood with the blood of commoners. Such matings were reportedly frequent also among Roman Egyptians of less elevated status: according to Scheidel (1997) (PMID 9881142), 37% of all documented marriages in the city of Arsinoe in the second century CE were between brother and sister. However, that is very exceptional and unprecedented elsewhere. Uncle–niece marriages are considered incestuous in many societies, but are almost the favored type in some parts of South India. In many countries of the Muslim Near East and South Asia a high proportion of all marriages are consanguineous, but the closest degree of consanguinity is double first cousins, with normal first cousins making up the bulk of close consanguinity (Figure 12.12). Keeping marriage within the family may be a prudent precaution in societies where arranged marriages with large dowries are a frequent cause of strife. In Europe, different religions take different attitudes to cousin marriage. The Greek Orthodox church prohibits them, the Catholic church requires a dispensation from the Church for first cousin or closer marriages, while most Protestant churches have no objection to marriage between first cousins.

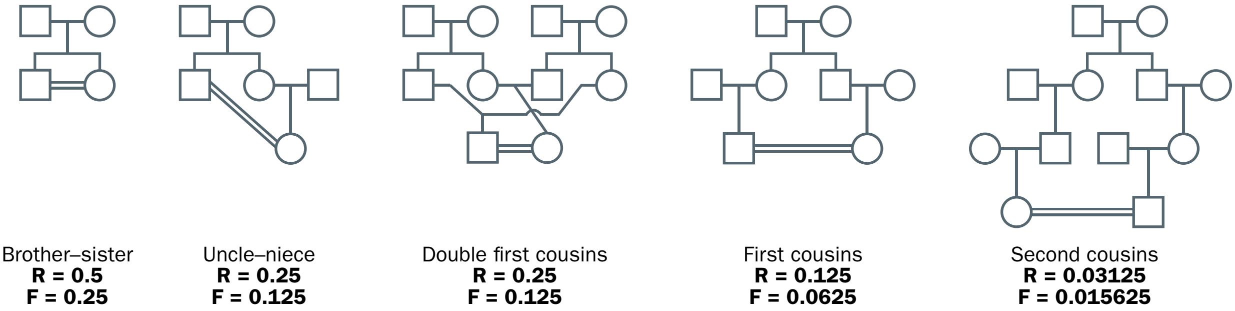


Figure 12.12 Consanguineous marriages. Consanguineous marriages are indicated by a double marriage line. The coefficient of relationship, R, and coefficient of inbreeding, F, of any offspring are shown.

Bittles and Black (2015) (see Further Reading) have collected and tabulated a vast array of information and data relating to types and frequencies of consanguineous marriage across the globe. Their map (**Figure 12.13**) shows the overall frequency of consanguineous marriage (defined as second cousin or closer) by country. **Table 12.4** gives some examples from their data. The population average inbreeding coefficient, symbolized α, is well below 0.01 in the great majority of populations.

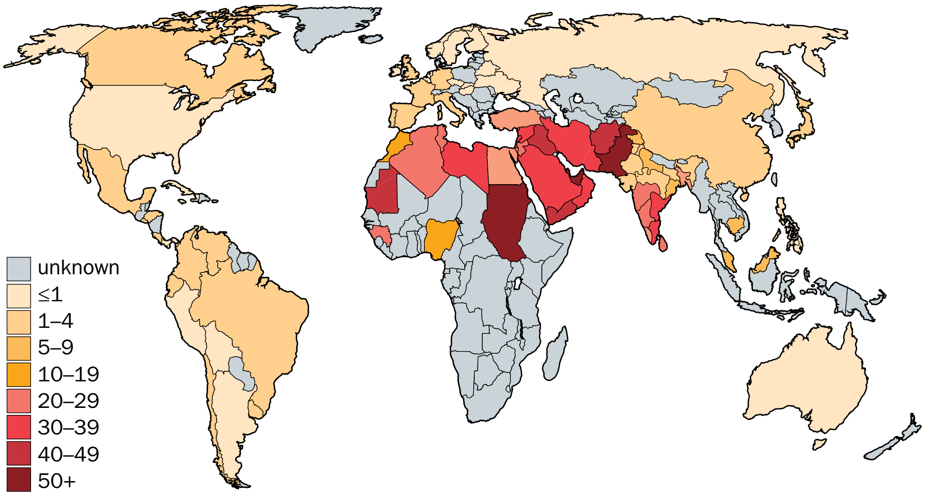


Figure 12.13 Rates of consanguineous marriage (%) by country across the world. (Adapted from Bittles AH & Black ML [2010] *Proc Natl Acad Sci USA* **107**(Suppl 1):1779–1786; PMID 19805052. With permission from National Academy of Sciences.)

Consanguinity and recessive disease

When parents are consanguineous the risk of recessive disease in their offspring is increased. Consider a recessive condition that affects one newborn in 10,000. The risk a randomly selected man is a heterozygous carrier is (very nearly) 1 in 50. If he marries his first cousin, she shares one-eighth of her genes with him, so if he is a carrier, there is a 1 in 8 chance she is also a carrier. The risk of an affected child is $1/50 \times 1/8 \times 1/4 = 1$ in 1600, compared to 1 in 10,000 for an unrelated couple.

Generalizing, if the frequency of the mutant allele is q, a nonconsanguineous couple are at risk q^2 while a couple who are first cousins are at risk $2pq \times 1/8 \times 1/4 = pq/16$. The ratio of risks is $q^2{:}pq/16$. For a rare condition, p is almost 1, so this ratio is very nearly $q^2{:}q/16$ or $1{:}1/16q$ (**Table 12.5**).

The calculation in **Table 12.5** shows that for very rare recessive conditions, the majority of cases will come from consanguineous unions. There are a few exceptions to this, when a rare recessive condition depends on an interaction between two different alleles at a locus, rather than homozygosity for a single allele. These rare exceptions are described in Section 16.1, but in general the association of rare recessive conditions with parental consanguinity holds true. Nevertheless, it is important not to consider consanguineous marriage as a purely medical problem. Customs regarding consanguinity are deeply embedded in the religion and culture of many non-Western societies, and expectations that rates of consanguineous marriage would fall as societies develop economically have often not been realized. In any case, the overall risk of cousin marriage should not be exaggerated. The risk of serious abnormality in a baby is roughly doubled when the parents are first cousins—but that only reduces the chance of a normal baby from 98% to 96%.

TABLE 12.4 CONSANGUINITY ACROSS THE GLOBE

Population or group	Date	% Consanguinity	Closest consanguinity	Average inbreeding coefficient, α
USA (all)	1959/60	0.2	First cousin	0.0001
Amish farmers, Pennsylvania, USA	1950	–	–	0.0012
Argentina (all)	1980/81	0.4	First cousin	0.0002
Amerindians, South Brazil	1961	13	Uncle–niece	0.0052
UK, Birmingham	1986/7	0.2	First cousin	0.0001
UK Birmingham, Pakistanis	1982/3	47	First cousin	0.0293
Japan (all)	1972	5.7	First cousin	0.0018
Israel, Ashkenazi migrants	1955/7	1.4	Uncle–niece	0.0009
Israel, Samaritans	1960	46	First cousin	0.0190
South India, Karnataka (Hindus)	1980/89	33.5	Uncle–niece	0.0333
South India, Karnataka, (Muslims)	1980/89	23.7	Uncle–niece	0.0160
South India, Karnataka (Christians)	1980/89	18.6	Uncle–niece	0.0173
Nyertiti, Fur, Sudan	1985	71	First cousin	0.0415
Tristan da Cunha	1961	–	–	0.0400

Consanguinity is defined here as second cousin or closer. Example data from Bittles AH & Black ML (2015), http://consang.net, which should be consulted for references and more detail.

TABLE 12.5 RISK OF RECESSIVE DISEASE FOR FIRST-COUSIN PARENTS

Frequency of disease allele	Frequency of condition in nonconsanguineous marriages	Frequency of condition in first-cousin marriages	Relative risk to first-cousin couple	Proportion of affected babies born to first-cousin parents
1 in 50	1 in 2500	1 in 800	3.125	0.03 (for C = 0.01)
1 in 100	1 in 10,000	1 in 1600	6.25	0.06 (for C = 0.01)
1 in 200	1 in 40,000	1 in 3200	12.5	0.11 (for C = 0.01)
1 in 500	1 in 250,000	1 in 8000	31.25	0.24 (for C = 0.01)
q	q^2	$q/16$	$1/16q$	$1/(1 - 16q) + (16q/C)$

Relative risk is risk compared to risk for unrelated parents. C, frequency of first-cousin marriages in the population (all others assumed to be unrelated). Note how a disproportionate number of the affected children are born to first-cousin parents, and the rarer the condition is, the greater is the disproportion.

SUMMARY

- The frequency of allele *i* in a population is the proportion of all alleles at that locus that are *i*. It is also the probability that an allele, picked at random, should be *i*.
- The Hardy–Weinberg distribution describes the relationship between allele frequencies and genotype frequencies in an undisturbed, random-mating population.
- Deviations from Hardy–Weinberg proportions of genotypes can result from genotyping errors or nonrandom mating, migration, or selection in the current generation.
- Nonrandom mating occurs when there is inbreeding, assortative mating (people choose their mate based on phenotypic similarity), or population stratification.

- Linkage disequilibrium is seen when the frequency of a multilocus haplotype differs from the value predicted from the frequencies of the individual alleles. Patterns of linkage disequilibrium show that our genome is organized into relatively stable haplotype blocks separated by recombination hotspots.
- Haplotype blocks are statistical constructs: it is not the case that there is complete linkage disequilibrium within a block and complete equilibrium between blocks. The size and number of blocks can change if the statistical definition of a block is changed—but the overall picture of ancestral blocks identifiable by tagging SNPs remains true for most parts of most genomes.
- Allele frequencies in a population are stable over generations unless changed by mutation, migration, selection, or random genetic drift.
- Genetic drift has a significant effect on allele frequencies in isolates where few individuals contribute to each succeeding generation, and historically in populations that have been through a bottleneck or stem from a small number of founders. In large populations its effects are very small.
- Mutation rates can be estimated in various indirect ways, but can now be quantified directly by sequencing parent–child trios. They depend critically on the age of the father.
- Alleles causing deleterious dominant or X-linked conditions are subject to strong negative selection, and such conditions can only be maintained in a population by recurrent mutation. Alleles causing deleterious autosomal recessive conditions are subject to only weak negative selection, and such conditions may be maintained for many generations in a population even without recurrent mutation.
- A high frequency of a deleterious autosomal recessive condition in a population may be due to a founder effect or heterozygote advantage (balancing selection).
- Realistically practicable eugenic interventions can have little effect on the frequency of monogenic diseases in human populations.
- Two alleles are identical by state if they have the same DNA sequence. They are only identical by descent if both are inherited from an individually identifiable common ancestor.
- The coefficient of relationship of two individuals is the proportion of alleles they share that are identical by descent. The coefficient of inbreeding of a person is the chance both alleles at a locus are identical by descent.
- Population-wide coefficients of inbreeding seldom exceed 0.01, and are unlikely to be above 0.04 even in populations with high levels of consanguineous marriage.
- Ultimately everybody is related to everybody else through remote common ancestors, but a marriage is commonly regarded as consanguineous if the couple are related as second cousins or closer.
- Consanguineous marriage increases the risk of having babies affected by an autosomal recessive condition. The rarer the condition, the greater is the proportion of cases that are born to consanguineous parents and the greater is the relative risk to consanguineous couples compared to unrelated couples. The precise figures presented here used the simplifying assumption that all marriages were either between first cousins or fully outbred.
- Population genetics can be very complicated, and we have considered only the simplest population models here. These have generally involved infinitely large populations with nonoverlapping generations and in mutation–selection balance. As soon as any of these simplifying assumptions is removed, the mathematics gets a great deal more complicated, and an advanced textbook should be consulted for details.

FURTHER READING

Recombination & haplotype blocks

Baudat F *et al.* (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**:836–840; PMID 20044539.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**:1299–1320; PMID 16255080.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**:851–861; PMID 17943122.

International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**:52–58; PMID 20811451.

Jeffreys AJ *et al.* (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum Mol Genet* **9**:725–733; PMID 10749979.

Myers S *et al.* (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321–324; PMID 16224025.

Mutation, selection, and population structure

Jobling M *et al.* (2014) *Human Evolutionary Genetics*, 2nd edn. Garland Science.

Kong A *et al.* (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**: 471–475; PMID 22914163.

Ancestry

Elhaik E *et al.* (2013) The GenoChip: a new tool for genetic anthropology. *Genome Biol Evol* **5**:1021–1031; PMID 23666864.

Emery LS *et al.* (2015) Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. *Am J Hum Genet* **96**:183–193; PMID 25620206.

Kidd KK *et al.* (2014) Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet* **10**:23–32; PMID 24508742.

Nelis M *et al.* (2009) Genetic structure of Europeans: a view from the North-East. *PLoS One* **4**:e5472; PMID 19424496.

Pardo-Seco J *et al*. (2014) Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC Genomics* **15**:543; PMID 24981136.

Ralph P & Coop G (2013) The geography of recent genetic ancestry across Europe. *PLoS Biol* **11**:e1001555; PMID: 23667324.

Reich D (2018) *Who We Are and How We Got Here*. Oxford University Press. (A highly readable introduction to DNA studies of population origins.)

Consanguinity

Bittles AH & Black ML (2010) Consanguinity, human evolution and complex diseases. *Proc Natl Acad Sci USA* **107**(Suppl 1):1779–1786; PMID 19805052.

Bittles AH & Black ML (2015) Global patterns and tables of consanguinity. http://consang.net

Scheidel W (1997) Brother-sister marriage in Roman Egypt. *J Biosoc Sci* **29**:361–371; PMID 9881142.

Comparative genomics and genome evolution

13

In Chapter 11 we focused on genetic variation within our own species, before going on to examine population genetics in Chapter 12. In addition to human genomes, the genome sequencing revolution has also been applied to very many animal genomes, and in this chapter, we address aspects of genetic variation between species. Comparing genomes across species has provided unprecedented understanding of how we differ genetically from other animal species, and how genomes evolve.

Understanding genetic variation across species, and how it affects the phenotype, is of interest for both scientific and medical reasons. From the basic science standpoint, there is interest in answering fundamental questions on evolution and evolutionary relationships. How closely are we related to other species at the level of nucleic acid and protein sequences? How did our genome evolve, and how did its component parts including chromosomes, genes, regulatory sequences, and other DNA sequences evolve? What makes us unique?

We also rely heavily on extrapolation from experimental model organisms to understand basic aspects of biology. Certain model organisms are important for basic research, allowing invasive procedures designed to study early development or, more generally, to study cellular mechanisms, gene function, and physiology in ways that illuminate our understanding of the equivalent human processes.

In addition, medical research is advanced by using model organisms in different ways: to investigate or confirm the pathogenic potential of DNA variants suspected of contributing to disease; to model human diseases and understand the underlying molecular basis of disease; and to investigate the efficacy and safety of novel types of treatment. In these cases, it is important to be aware of genetic differences between humans and the model organisms, allowing us to make reasonable inferences when extrapolating from data obtained from the animal models. We describe some of the more important model organisms in Chapter 21.

In Section 13.1 we cover how comparative genomics was established, and how it depends on powerful computer programs that enable sequence comparisons and alignment at the genome level. As whole genome sequences became available, comparative genomics emerged as a powerful tool for identifying novel genes and validating predicted genes, and it was to become crucially important in defining functional noncoding DNA sequences. In Sections 13.2–13.4 we cover different facets of genome evolution, beginning in Section 13.2 with, notably, the evolution of genome size and of gene number in genomes, and the evolutionary importance of gene duplication and exons. We progress to considering the evolution of mammalian chromosomes in Section 13.3. In Section 13.4 we consider how regulatory sequences evolve, and the importance of noncoding DNA, and especially transposon sequences, in providing novel functional sequences. Finally, in Section 13.5, we describe phylogenetic approaches and give an overview of where humans fit in within the Tree of Life.

13.1 COMPARATIVE GENOMICS

Obtaining a complete genome sequence is a starting point for compiling a catalog of genes and functionally important sequences for that organism, and for beginning to work out how the genome functions. The Human Genome Project delivered thousands of novel human genes that were initially predicted by bioinformatic analyses (and required subsequent experimental validation). In those early days of genome analysis, the extensive numbers and roles of noncoding RNAs was not appreciated; initial bioinformatics analyses were skewed toward trying to identify protein-coding genes, using computer programs that were heavily dependent on scanning for long open reading frames, and scanning databases for evolutionary conservation of suspected novel genes.

The need for comparing genome sequences was recognized at the outset of the Human Genome Project: one of its component projects aimed to sequence the genomes of four model organisms—*E. coli, S. cerevisiae, D. melanogaster,* and the mouse. Sequencing these genomes was intended not just as a way of testing the developing genome sequencing technologies and to drive research on these model organisms, but also to assist interpretation of the human genome sequence by making available other genomes for comparison. As other genome sequences began to become available, too, a new field of **comparative genomics** developed. Computer programs were used to perform multiple alignments of sequences from corresponding genome regions in multiple species, and eventually sets of homologous sequences could be grouped with their sequences aligned at the base-pair level in an attempt to define whole-genome sequence alignments.

Although comparing genomes from distantly related species was valuable, it became clear that many genes, sometimes called **orphan genes**, appeared to be restricted to one of the species under study. That seemed unlikely (and we now know that genes specific to a species are extremely rare). It also raised the possibility that additional genes might be found if genomes of more closely related species were available for comparison. To maximize interpretation of the human genome, therefore, a new comparative genomics project was launched in the mid-2000s to sequence the genomes of 22 mammals, and compare them with previously obtained genome sequences from seven other mammals. Comparative analyses on the 29 mammalian genomes, including several primate genomes, were published in the late 2000s; the latest proposal is to add a further 200 mammalian genomes.

Comparative genomics can be applied toward different ends. Predicted genes and suspected pathogenic mutations can be validated or rejected and conserved noncoding DNA elements can be defined. Two important general uses depend on the ability to identify signatures of negative and positive natural selection, as described in the first two sections below.

Identifying functionally important DNA sequences that have been conserved by purifying selection

By comparing our genome with other genomes, it became possible to systematically identify sequences conserved during evolution. The human genome, like those of other complex metazoans, was shown to consist of a sea of poorly conserved sequences with islands of highly conserved sequences (sequences that have not changed so very much over many millions of years of evolution; the first estimates suggested that about 5% of the human genome sequence was highly conserved).

Highly conserved DNA sequences are presumed to be functionally important, and to have been maintained by **purifying (negative) selection** that selects against alleles with deleterious changes at functionally important nucleotide positions. By contributing to loss of function, the deleterious changes reduce the biological fitness of the organism (lowering its capacity for reproductive success and its ability to transmit genotypes to future generations). As a result, alleles with deleterious changes are selectively eliminated from populations. Sequences subject to purifying selection therefore appear to be highly or moderately conserved between species, when compared to sequences where no selection of this kind is operating.

To identify evolutionarily conserved sequences, corresponding sequences need to be aligned from the genomes of different organisms. Comparisons of three or more aligned sequences will reveal regions that probably show sequence conservation as a result of purifying selection (rather than just by chance, which can occur if just two sequences are aligned). Very highly conserved sequences can be identified when the compared sequences are from distantly related organisms; comparisons of sequences from closely related organisms are needed to identify moderately conserved sequences. Because regions in which purifying selection is operating do not tolerate many mutational changes, the sequences are said to be evolutionarily *constrained* (**Figure 13.1**).

Because coding DNA sequences are comparatively easy to detect by bioinformatic analyses, comparative genomics is particularly valuable in identifying functional noncoding DNA sequences, notably regulatory DNA elements (typically very short DNA sequences that are often difficult to identify when analyzing a single genomic sequence). Novel conserved sequences identified by comparative genomics can then be targeted for experimental investigations to identify how they work.

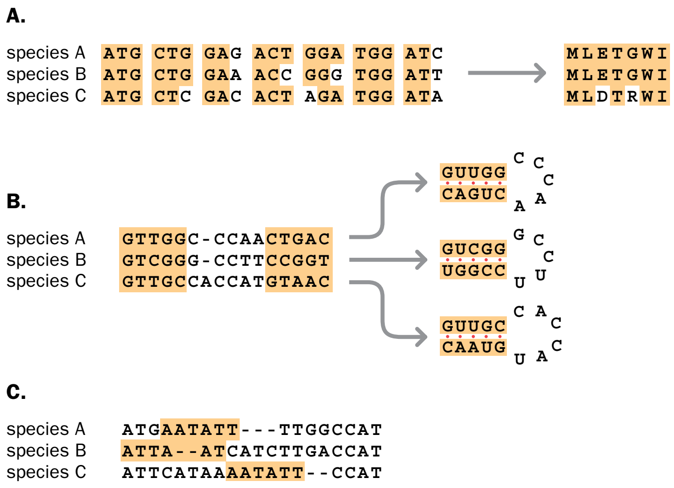

A.
species A ATG CTG GAG ACT GGA TGG ATC
species B ATG CTG GAA ACC GGG TGG ATT
species C ATG CTC GAC ACT AGA TGG ATA
MLETGWI
MLETGWI
MLDTRWI
B.
species A GTTGGC-CCAACTGAC
species B GTCGGG-CCTTCCGGT
species C GTTGCCACCATGTAAC
GUUGG
CAGUC
C C C A A
GUCGG
UGGCC
G C C U U
GUUGC
CAAUG
C A C C A U
C.
species A ATGAATATT---TTGGCCAT
species B ATTA--ATCATCTTGACCAT
species C ATTCATAAAATATT--CCAT

Figure 13.1 Purifying selection results in evolutionarily constrained sequences. (**A**) In a protein-coding gene, the pattern of mutation is constrained because of the need to conserve functionally important amino acids. Base triplets that will specify codons at the RNA level are marked by the boxes; yellow shading indicates the predominant base at each position. Mutations at the third base position of codons are less likely to alter the encoded amino acid and so are more tolerated. The encoded heptapeptide sequence (shown on the right) seems well conserved, but the fifth amino acid position seems to be less constrained and may be less critical for the function of the protein (because of the nonconservative G to R substitution). (**B**) In this hypothetical RNA gene, the shaded sequences base-pair at the RNA level, forming a stem-loop structure, as shown on the right (note that G can base-pair with U as well as with C in double-stranded RNA). Bases in the loops are less well conserved, but base pairing within the 5 bp stem region is conserved and is maintained by compensatory double substitutions (as seen in the first base positions of the stem, immediately adjacent to the loop). (**C**) The sequences of the short elements (highlighted) that form *cis*-acting regulatory sequences are conserved but their positions vary in the aligned sequences from different species.

Identifying the basis of recent lineage-specific evolutionary adaptations

The focus on highly conserved sequences is understandable, but not all functionally important sequences are very highly conserved, and comparing genomes of closely related species can also pinpoint significant differences between functional DNA sequences in the genomes of closely related species. That in turn can help identify DNA variants that played roles in lineage-specific adaptations occurring in the recent evolutionary past. Recall that **positive selection** works to selectively promote a new or existing advantageous allele that benefits the organism in some way so as to increase its biological fitness. Organisms carrying the advantageous allele have a significant reproductive advantage over those that lack it in the same population. The allele then increases in frequency in the population, and positive selection can drive novel lineage-specific adaptations.

Identifying species-specific functional DNA variants might be expected to help us understand why species differ and what makes humans unique. We are so closely related to the great apes that by any objective measure we should be classified in the same genus as them. And yet we have many distinguishing characteristics, notably our unique cognitive abilities, that developed as a result of large-scale expansion of the frontal cortex in the human lineage. Positive selection has been invoked to explain some of our unique characteristics, but it also appears to have been important in human lineages following the divergence from the common ancestor of humans and the great apes. We consider these aspects in Chapter 14.

Positive selection is most readily identified in coding DNA as rapidly evolving codons within a background of evolutionarily constrained sequence. The standard way to do this involves calculating the relative frequencies of nonsynonymous and synonymous substitutions—see **Box 13.1**). Signatures of positive selection may also be found in RNA genes (see **Figure 13.2**).

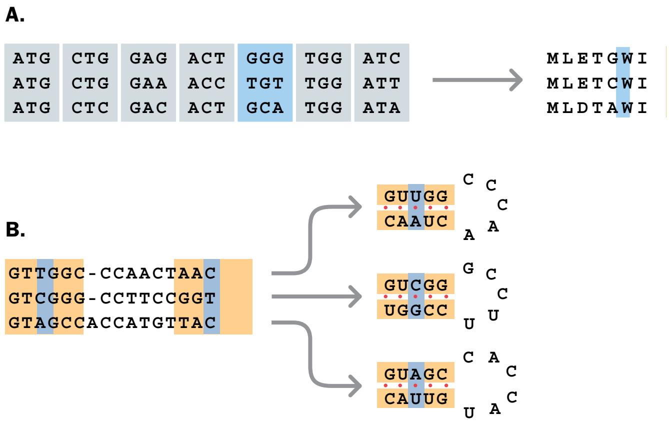


Figure 13.2 Sequence diversification by positive selection. (**A**) Positive selection within a coding sequence. In this example, the fifth base triplet is envisaged to be a target of positive selection and varies in all three species shown, whereas the surrounding sequence is constrained by purifying selection. (**B**) Suggestive positive selection within an RNA gene. The hypothetical sequences shown here have inverted repeats that permit base pairing to form hairpins with a 5 bp stem. The middle nucleotides in the 5 bp stem are highly variable, but compensatory double substitutions at the DNA level maintain the stem. The middle nucleotides of the stem seem to be selected to be diversified within a highly conserved RNA structure.

BOX 13.1 ASSAYING SELECTION PRESSURES ACTING ON CODING DNA SEQUENCES USING dN/dS RATIOS

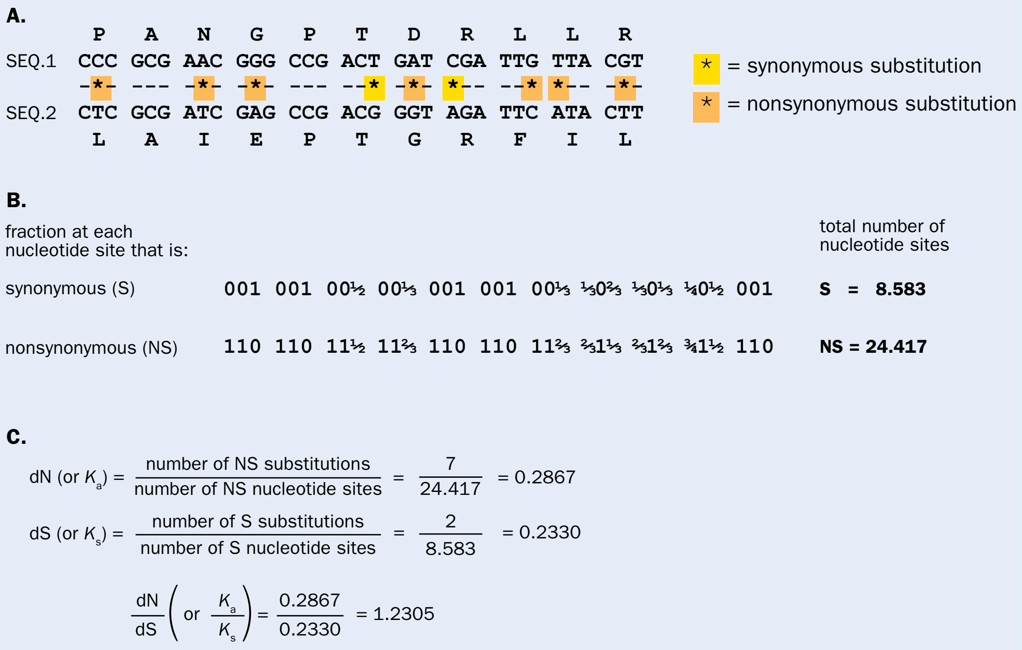
Coding sequences are functionally important and subject to natural selection. Many are subject to a high degree of purifying selection to conserve functionally important amino acid sequences, and they appear to evolve slowly. Segments of some coding sequences, however, may be subject to positive selection in a lineage, and can diverge comparatively rapidly in sequence during evolution.

In order to examine the nature of the natural selection forces working on a coding DNA sequence, orthologous sequences from different species (such as human and mouse insulin coding DNA sequences) are aligned (or sometimes, suitably polymorphic allelic sequences are aligned). The aligned sequences are examined to identify base substitutions, and to count the number of **synonymous mutations** (base substitutions that do not change the amino acid) and **nonsynonymous mutations** (base substitutions that change the amino acid). Recall that synonymous mutations mostly involve changes to the third base position of codons (because of "base wobble"), while nonsynonymous mutations include any substitution at the second base position of a codon, most that occur at the first base position, and a small number at the third base position.

A preponderance of synonymous mutations in the aligned sequences indicates that purifying selection has been at work. In Chapter 11, Figure 11.12 provides a classic example: alignment of coding sequences for aa 1–24 of human and mouse insulin beta chains shows a 10:1 ratio of synonymous substitutions to nonsynonymous substitutions. Evidence for positive selection requires an unexpectedly high proportion of nonsynonymous mutations. But in order to evaluate the likelihood of positive selection, it is not sufficient to simply work out the ratio of nonsynonymous and synonymous substitutions: a more rigorous assay is required that takes into account the possibilities for a synonymous or nonsynonymous change at each of the nucleotide sites being compared.

Take the example of the aligned sequences 1 and 2 in Figure 1A. In just 11 codons (33 nucleotides of coding DNA), there are a remarkable seven nonsynonymous substitutions, but also two synonymous substitutions. Has positive selection been operating on the sequence? To work that out we need to take into account the degree to which each of the 33 nucleotide sites is a synonymous site or a nonsynonymous site. Taking the first codon—CCC in sequence 1 and CTC in sequence 2—we look in turn at each base position. The first C in the CCC codon of sequence 1 is a nonsynonymous site: if it were to be substituted (by A, G, or T), it would change the amino acid, and the same is true of the first C in codon CTC in sequence 2. Therefore, the first nucleotide position is 100% nonsynonymous and 0% synonymous. We can enter a value of 0 for the fraction of this first site that is synonymous, and 1 for the fraction that is nonsynonymous (see Figure 1B).

Similarly, at base position 2, substituting the middle C of CCC, or the T of CTC, always results in a change of an amino acid. The second site is scored 1 for nonsynonymous and



A.

```
        P   A   N   G   P   T   D   R   L   L   R
SEQ.1  CCC GCG AAC GGG CCG ACT GAT CGA TTG TTA CGT
       -*- --- -*- -*- --- --* -*- *-- --* *-- -*-
SEQ.2  CTC GCG ATC GAG CCG ACG GGT AGA TTC ATA CTT
        L   A   I   E   P   T   G   R   F   I   L
```

🟨 * = synonymous substitution

🟧 * = nonsynonymous substitution

B.

fraction at each nucleotide site that is:												total number of nucleotide sites
synonymous (S)	001	001	00½	00⅓	001	001	00⅓	⅓0⅔	⅓0⅓	¼0½	001	**S = 8.583**
nonsynonymous (NS)	110	110	11½	11⅔	110	110	11⅔	⅔1⅓	⅔1⅔	¾1½	110	**NS = 24.417**

C.

$$\text{dN (or } K_a) = \frac{\text{number of NS substitutions}}{\text{number of NS nucleotide sites}} = \frac{7}{24.417} = 0.2867$$

$$\text{dS (or } K_s) = \frac{\text{number of S substitutions}}{\text{number of S nucleotide sites}} = \frac{2}{8.583} = 0.2330$$

$$\frac{\text{dN}}{\text{dS}}\left(\text{or } \frac{K_a}{K_s}\right) = \frac{0.2867}{0.2330} = 1.2305$$

Box 13.1 Figure 1 Calculating dN/dS ratios manually. (**A**) Sequence alignment of two hypothetical coding DNA sequences, 33 nucleotides long, showing encoded amino acids at top and bottom with substitutions shown by the highlighted boxed asterisks. (**B**) Calculation of the fraction (from 0 to 1) that is synonymous and nonsynonymous at each of the 33 nucleotide sites. Some of these are fractions. At the third base position in the third codon, for example, substitution of the C of AAC by a T is silent, but substitution by an A or G is nonsynonymous, while substitution of the C of ATC by either A or T is synonymous, but substitution by a G is nonsynonymous. That means that the fraction of this site that is synonymous is 0.5, as is the fraction that is nonsynonymous. (**C**) Final calculation of the dN/dS ratio.

0 for synonymous. By contrast, the third nucleotide site is scored 1 for the synonymous fraction and 0 for the nonsynonymous fraction (substituting the final C of CCC or CTC never alters the amino acid). We continue until each of the 33 nucleotide positions are evaluated for the fraction that is synonymous and the fraction that is nonsynonymous, finding that some of the nucleotide positions are part synonymous and part nonsynonymous (see **Figure 1B**). Across the 33 nucleotide sites, the total number of nonsynonymous sites is nearly three times as high as the number of synonymous sites.

Then, one calculates the **dN/dS ratio** (alternatively called the ***K_a/K_s* ratio**), where dN (or K_a) is the number of nonsynonymous substitutions per nonsynonymous site, and dS (or K_s) is the number of synonymous substitutions per synonymous site. A dN/dS ratio <1 means that purifying (negative) selection (selection against deleterious nonsynonymous substitutions) has definitely operated. If the ratio is 1, the situation is ambiguous (the amino acid substitutions may be largely neutral, or a degree of positive selection has operated to cancel purifying selection).

In **Figure 1C**, however, the dN/dS (*Ka*/*Ks*) ratio is greater than 1, meaning that positive selection has caused at least some of the amino acid substitutions. Some other substitutions may have been caused by *genetic drift;* although purifying selection may also be operating on some residues, it is not strong enough to overcome the contribution by positive selection. (Note: synonymous mutations are often assumed to be neutral mutations, but they may not always be so: the codon sequence may be part of a regulatory sequence within an exon, such as an exonic splice enhancer sequence).

Although dN/dS ratios can be calculated manually over parts of a protein sequence, computer-based methods help calculate the dN/dS ratio over full lengths of proteins. With pairs of sequences, a maximum likelihood method can be used, such as that described in the 2000 paper by Yang and Nielsen (PMID 10666704). But, wherever possible, however, it is better to begin with a multiple sequence alignment and use a program that is suited to multiple sequence alignment such as the PAML (codeml program) of Ziheng Yang (see http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf).

A variety of computer programs allow automated genome sequence alignments

Underpinning comparative genomics is the need to carry out extensive genome sequence comparisons. Basic sequence alignment tools such as the standard BLAST programs rely on simple queries; using them to compare large genome sequences, such as those of vertebrates, is remarkably inefficient. A significant difficulty is the extent of large-scale genome rearrangements (involving regions >50 kb) that have periodically occurred during evolution. As a result, although many genome regions can be identified that are clearly homologous between species, a linear order of the homologous genome regions in one species, with the individual elements in a certain orientation, is often not conserved between species (**Figure 13.3A**).

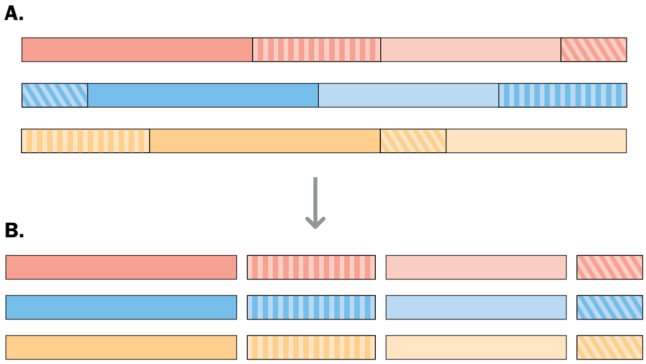


Figure 13.3 Aligning whole genome sequences begins by reconstructing homologous collinearity relationships. (**A**) Homologous genomic sequences (shown by similar shading patterns) from three organisms are located in nonhomologous positions as a result of intrachromosomal rearrangements since they diverged from a common ancestor (interchromosomal rearrangements are not shown for the sake of simplicity). (**B**) The homologous sequences need to be grouped into sets that are then arranged in a linear sequence so as to simplify subsequent whole-genome alignments. (Adapted from Margulies EH & Birney E [2008] *Nat Rev Genet* **9**:303–313; PMID 18347593. With permission from Springer Nature. Copyright © 2008.)

To compare large genome sequences more powerful programs are needed. For a collection of extant (currently existing) genomes, the challenge is to find the minimum set of groups of homologous sequence so that inside each group the sequences to be compared are both homologous and collinear (**Figure 13.3B**). This collinearity reconstruction problem seeks answers to two basic questions. First, which segments of extant genomes arose from the genome of a common ancestor? Secondly, what was the likely evolutionary path of events that generated the different segments? Various computer programs have been developed to reconstruct collinearity for use in large-scale alignment (**Table 13.1**).

Once collinear relationships between homologous genomic regions have been established, computer programs take the unaligned sequences and determine which bases in each sequence are orthologous. Multiple alignment of very large sequences consumes huge computing power and inevitably the programs use *heuristic* methods; that is, they apply workable solutions that are not formally correct in order to reduce

TABLE 13.1 EXAMPLES OF COMPUTER PROGRAMS USED IN ALIGNING WHOLE COMPLEX GENOME SEQUENCES AND DETECTING EVOLUTIONARILY CONSTRAINED SEQUENCES

Task	Program	Characteristics	Website/PMID
Reconstructing homologous collinearity	MERCATOR	Uses only coding exons as initial anchoring points	http://www.biostat.wisc.edu/~cdewey/mercator/
	GRIMM	Analyzes rearrangements in pairs of genomes	http://grimm.bioprojects.org/GRIMM/index.html
	MLAGAN	Multiple global alignment of genomic sequences	PMID 12654723
	PECAN	A consistency-based multiple-alignment program	http://www.ebi.ac.uk/~bjp/pecan
Multiple alignment	TBA/MULTIZ	TBA is designed to be able to align many megabase-sized regions of multiple mammalian genomes. To perform the dynamic-programming step, TBA runs a standalone program called MULTIZ, which has been used to align much of the human genome sequence with the sequences of 27 other vertebrate genomes	PMID 17984227. The 28-way vertebrate alignment can be viewed with the UCSC Genome Browser at http://genome.ucsc.edu, downloaded in bulk by anonymous FTP from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz28way, or analyzed with the Galaxy server at http://g2.trac.bx.psu.edu
Constraint detection	PHASTCONS	Identifies evolutionarily conserved elements in a multiple alignment, given a phylogenetic tree	http://compgen.bscb.cornell.edu/~acs/phastCons-HOWTO.html; PMID 16024819
	SCONE	Gives a sequence conservation score after estimating the rate at which each nucleotide is evolving, and then computing the probability of neutrality given the estimated rate	PMID 18166073

For further information on these and additional programs, see Margulies & Birney (2008) (PMID 18347593) under Further reading.

computational time. Nevertheless, by 2007, megabase-sized chunks of the human genome sequence were able to be aligned with sequences from 27 other vertebrates. Additional programs are used to inspect the aligned sequence to identify sequences that are evolutionarily constrained.

Using cross-species comparisons to validate or reject predicted genes and to identify novel genes

When the sequences of metazoan genomes were first reported, large numbers of novel genes were predicted for which there was little or no supportive experimental evidence. Often gene prediction relied quite heavily on using bioinformatics programs to identify open reading frames (ORFs), which are expected of protein-coding genes. The difficulty here is that some ORFs are not associated with coding DNA: a segment of noncoding DNA might happen to have, just by chance, a significant ORF. Thus, a random sequence 2 kb long and with a 50% GC base composition has a 50% chance of harboring an ORF 400 bases long in one of the six possible translational reading frames (three reading frames, forward and backward). The DNA triplets corresponding to termination codons are AT-rich and so noncoding transcripts, which are often GC-rich, may contain moderately long ORFs simply by chance.

Validating and rejecting predicted genes

Cross-species comparisons provide a simple way to validate a predicted gene. The initial test is to seek sequences in the genomes of other species that are clearly related in sequence to significantly long segments of the putative gene. After an open reading frame has been identified for a putative protein-coding gene, identified sequences in other genomes can be scanned for corresponding ORFs that when translated give protein sequences homologous to the protein sequence predicted for the putative gene. Validating genes encoding a functional RNA is less straightforward, and relies on detecting signatures of purifying selection.

The ability to reject predicted genes has also been important for re-annotating genomes. An early illustration was provided by genome sequencing of multiple different *Saccharomyces* and *Drosophila* species to help annotate the previously studied *S. cerevisiae* and *D. melanogaster* genomes. Many originally predicted genes in the latter two genomes were discounted when equivalent ORFs or similar codon substitution frequencies were not consistently found in orthologous sequences in the other *Saccharomyces* or *Drosophila* species. Significant numbers of predicted human genes have also been discounted: for this and other reasons, the number of human protein-coding genes has been steadily revised downward from about 24,500 in mid-2006 to currently close to 20,000.

Identifying novel genes

When genome sequencing came to be applied to large numbers of species, it became possible to identify novel genes in some genomes by sequencing genomes of very closely related species. For example, 1275 novel *C. elegans* genes were predicted only after the genome of the related nematode *C. briggsae* was sequenced. The novel *C. elegans* genes were so distantly related from genes in other well-studied genomes, such as human, mouse, zebrafish, and *Drosophila* genomes, that they had previously gone undetected. (Truly species-specific orphan genes may be extremely rare but a significant number of genes are taxonomically restricted.) That kind of result prompted efforts to extend sequencing of mammalian genomes and include other primate genomes to maximize the chance of detecting recently evolved genes.

Comparative genomics has been particularly valuable in identifying novel genes that make functional noncoding RNA. Short RNA genes can easily be overlooked in bioinformatics analyses of a single genome sequence, but comparative genome analyses can identify signatures of purifying selection in noncoding sequences as well as in protein-coding DNA. Identifying regions containing functionally important noncoding DNA has then led to discovery of novel RNA genes.

Using comparative genomics to estimate how much of the genome is under purifying selection

The need to define the complete set of functionally important human DNA sequences—including coding DNA, RNA genes, and regulatory sequences—was a principal driving force behind sequencing of the human genome. Coding DNA sequences have been progressively studied for decades and have long been known to constitute a very small fraction of our genome—the most current estimates indicate 1.1%. However, in the pre-genome sequencing era, very little had been known about the functional noncoding DNA elements in our genome. Even the amount of functionally important noncoding DNA was quite unknown until the human genome was compared with the second mammalian genome to be sequenced, the mouse genome.

Purifying selection works to conserve functionally important DNA sequences, both coding and noncoding DNA sequences. By comparing DNA sequences from related species, one can identify the proportion of DNA sequences under purifying selection: they evolve more slowly than functionless DNA sequences—so-called "neutral DNA"—where mutation is not constrained. And because purifying selection is, by a very long distance, the most pervasive form of natural selection, evolutionary geneticists consider that the proportion of a genome under purifying selection closely approximates the proportion of functional DNA (it will be a slight underestimate: a small amount of functional DNA sequence is subject to positive selection and is evolving rapidly).

Human–mouse comparative genomics

The first attempt to estimate the proportion of the human genome under purifying selection began by using genome-alignment programs to sort human and mouse genome sequences into those sequences that can be aligned at the nucleotide level, and those that cannot. Although humans and mice diverged from a common ancestor around 90 million years ago, about 40% of the human genome sequence could be aligned at the nucleotide level to the mouse genome; an average of 69.8% of the aligned bases were matching bases. Thereafter, the aligned sequences were divided into nonoverlapping windows, typically 50 to 100 nucleotides long. The number of base matches per window were counted, and then the score for each window was referenced against the score for neutral DNA sequences or sites.

Although often considered as neutral sites, *fourfold degenerate sites* in coding DNA (**Table 13.2**) were not used as the neutral DNA reference (because of concerns that a low level of purifying selection might apply). Instead, ancestral transposon repeats were chosen that were present in the last common ancestor of humans and mice. Located at orthologous positions in the compared genomes (such as in the same intron of equivalent human and mouse genes), they had been inactivated many millions of years ago. The overall average percent identity in these ancestral repeats was 66.7%, significantly lower than the 67.2% for fourfold degenerate sites.

Using non-overlapping 50-nucleotide windows of aligned human–mouse sequence, and taking a threshold of 40 base matches per window as being significantly more

TABLE 13.2 THE EIGHT TYPES OF FOURFOLD DEGENERATE SITE IN CODING DNA

Site	AC**x**	CC**x**	CG**x**	CT**x**	GC**x**	GG**x**	GT**x**	TC**x**
Amino acid	Threonine	Proline	Arginine	Leucine	Alanine	Glycine	Valine	Serine

The highlighted third base positions are mostly free from mutational constraint with regard to base substitutions (which are always synonymous at these positions). Occasionally, however, purifying selection may operate at these sites. For example, a site might just happen to fall within a conserved exonic regulatory sequence, such as a splice enhancer site.

conserved than freely mutating neutral DNA, 5.19% of the human genome was estimated to be under purifying selection. This was the first indication that the amount of functional noncoding DNA significantly exceeds the amount of coding DNA (now known to be ~1.1% of the genome). The 5.19% figure was the original source for the oft-quoted figure of 5% for the proportion of highly conserved DNA in the human genome, but it was an approximation: changing the window size to 100 or 200 nucleotides, while keeping the proportion of matched nucleotides at 80%, for example, changed the proportion of DNA under purifying selection from 5.19% to, respectively, 6.15% or 7.92%.

Multimammal comparative genomics

The method above is limited in different ways. First, it estimates the proportion of windows, rather than bases, that are under purifying selection. As a result, it tends to overlook functionally important nucleotides that are distributed diffusely, and over-counts neutrally evolving nucleotides within evolutionarily constrained windows. And applying the method to just two species meant that it would not be able to detect lineage-specific functional DNA sequences (including regulatory DNA sequences with primate-specific functions).

Follow-up studies have involved sequence alignments from multiple genomes. Initially, the human genome was compared to three other mammalian genomes (dog and rat, as well as mouse), but subsequently an additional 25 mammalian genomes also became available for comparison, including the chimpanzee, our closest relative, and some other primates. Aligning these 28 mammalian genomes against the human genome allowed a high-resolution map of human evolutionary constraint (see also below), increasing the power to detect lineage-specific functional DNA elements.

In these follow-up studies, different approaches have been used to measure the proportion of the genome under purifying selection. Some, such as the SCONE program listed in **Table 13.1**, have the advantage of being able to estimate evolutionary rates for individually aligned bases rather than sequence windows. Most of the studies focus on nucleotide substitution rates within aligned sequences, but one approach, the neutral indel method, exploits indels rather than substitutions in predicting functional sequences.

The neutral indel method assesses the lengths of “intergap segments” between adjacent indel events. For neutrally evolving genomic sequence, the lengths of the intergap segments follow a geometric distribution. However, when purifying selection purges deleterious indels, a fraction of the intergap segments across the genome becomes longer than expected under the neutral DNA model. The ensuing excess of long intergap segments provides an estimate of the total length of sequence from which indels have been purged, and hence an estimate of the proportion of DNA under purifying selection. The neutral indel method has also shown that functional (mostly noncoding) DNA turns over at a high rate, and as the divergence between mammalian species increases, the predicted amount of shared functional DNA sequence dramatically drops off.

Over the last decade or so, many studies have attempted to quantitate the proportion of the genome that is under purifying selection; most estimates are in a broad range of 3–12%, with a strong likelihood of the value falling within the narrower 7–9% range—see Table 1 of Ponting & Hardison (2011) (PMID 21875934), and also Rands *et al.* (2014) (PMID 25057982). These methods cannot assess additional, very recent, purifying selection in the human lineage; that would require comparisons between the genomes of very large numbers of humans (sequence divergence between humans is very low, making it difficult to get enough data; population genomics studies may provide a solution).

Evolutionary geneticists interpret the comparative genomics data to mean that a small fraction of our genome—likely to be 10% at most—is functionally significant. That estimate is in marked contrast to interpretations of the ENCODE Project data, where 80% of the genome was suggested to be associated with some function, sparking controversy and a lively debate (see **Box 13.2**).

BOX 13.2 A POST-ENCODE CONTROVERSY: JUNK DNA AND THE PROPORTION OF THE GENOME THAT IS FUNCTIONALLY IMPORTANT

A principal conclusion of the ENCODE Project (detailed in Section 9.4) was that biochemical functions could be assigned to 80% of the genome. A *Nature* "News & Views" article described this finding as "dispatching the widely held view that the human genome is mostly junk DNA", and a *Science* "News & Analysis" commentary was entitled "Encode Project Writes Eulogy for Junk DNA". The term *junk* DNA signifies any DNA sequence in a genome that does not play a functional role in development, physiology, or some other organism-level capacity (in use since the 1960s, it was popularized by Susumu Ohno in an influential 1972 paper).

The opposing view, that a small fraction only (most likely <10%) of the human genome has a biological function, is supported by estimates of the amount of DNA under purifying selection (see main text), and by unexpected variability in the genome sizes of diploid organisms that is often unrelated to organism complexity. Even quite closely related diploid species can have rather different genome sizes—pufferfish have a genome less than half the size of that of medaka, for example—and our genome is unexpectedly one fifth the size of that of the onion. (Note: it is not just genome size that does not correlate well with organism complexity, but also gene number—see **Figure 1**.)

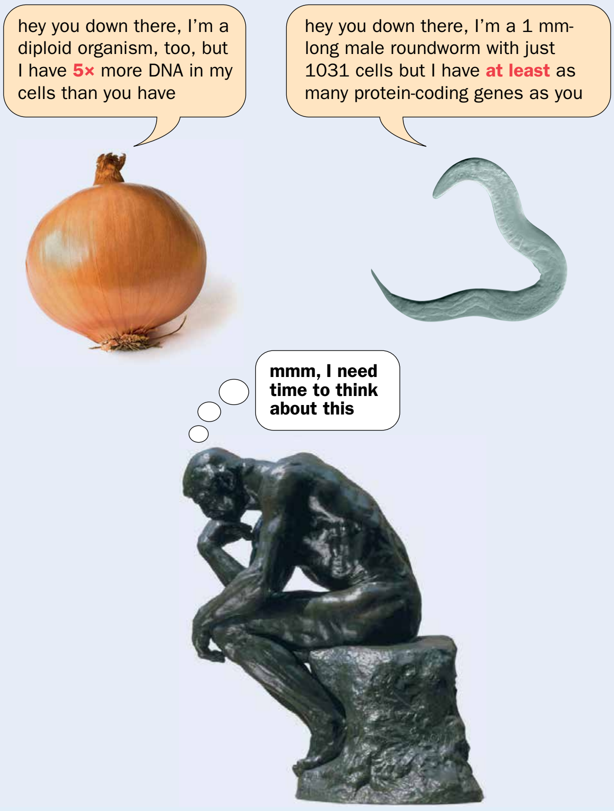
The idea that in the post-ENCODE era it was time to dispense with the concept of junk DNA generated a great deal of controversy, and a series of strong ripostes (see Graur *et al*. [2013], PMID 23431001; and Doolittle [2013], PMID 23479647, for two examples). The interpretation of ENCODE data has been criticized on the grounds that simply showing that a DNA sequence exhibits a property *associated* with function doesn't prove that it is functional, and one of the most contentious issues has been the significance attached to biochemical function. The figure of 80% for the percentage of the genome to which the ENCODE Project attached biochemical functions derives very largely from RNA transcription (74.7% of the genome was transcribed in at least one of 15 human cell lines). But that doesn't mean that 80% of the genome is functional (or that 100% of the genome is functional because every single nucleotide is associated with another biochemical function, DNA replication).

It is now clear from ENCODE and other transcriptome projects that the mammalian transcriptome is dominated by long noncoding RNAs (lncRNAs) and that, although the functional status of many lncRNAs is still uncertain, quite a large number of them have been shown to have a function (see **Table 10.3** for some examples). However, lncRNAs have been poorly conserved during evolution and their sequences may be allowed to drift. Possibly, purifying selection is concentrated on a few short sequence elements, such as those that are important in intramolecular base pairing (the structure of an RNA is often highly conserved, if not the sequence) and those required for recognizing interacting molecules.

Whatever percentage of our genome is biologically functional, it cannot be a very high one on the grounds of the excessive **mutational load** it would impose. Depending on the paternal age, a child is born with from 40 to 100 new mutations. If most of the nucleotides in our genome had a significant functional role, then the number of new mutations in each generation would simply be too high for humans to be viable.

As well as having a core of functional DNA sequences, genomes do seem to have variable and often large amounts of junk DNA, DNA that does not seem to do anything useful and is free from selection pressure. Evolution cannot plan ahead, but sometimes something that is of no current value, and neither selected for or against, may come to have value at some later time. Imagine two closely-related organisms occupying similar ecological niches. One has a lot of nonfunctional junk DNA, the other has a more compact genome with very little junk. It makes no difference to the selective pressures they both face. At some much later time, as a result of a mutation or a series of mutations in a junk DNA sequence, a new sequence is produced with a useful function. That species now has a selective advantage and survives while the species lacking the junk DNA dies out. Thus indirectly, junk DNA might have evolutionary value.

Much of the junk DNA in mammalian genomes is composed of sequences originating from transposable elements. As detailed in Section 13.4, mutation operating on sequences like this can result in a variety of novel functional sequences, including novel exons, regulatory sequences, lncRNAs, and, very occasionally, even genes.


hey you down there, I'm a diploid organism, too, but I have 5× more DNA in my cells than you have
hey you down there, I'm a 1 mm-long male roundworm with just 1031 cells but I have at least as many protein-coding genes as you
mmm, I need time to think about this

Box 13.2 Figure 1 How special is our genome? The diploid onion genome is estimated to contain 15.5 Gb of DNA, five times the size of the 3.1 Gb human genome. The number of *C. elegans* protein-coding genes listed in the most recent WormBase release (version WS265, March 2018) is 20,208, slightly more than the 19,940 human protein-coding genes recorded by GENCODE (version 29, October 2018).

Comparative genomics as a global way of identifying conserved noncoding DNA elements

From the above, comparative genomics gives an estimate for the total amount of functional noncoding DNA: if we take 8% as an average for the proportion of the genome under purifying selection, then a maximum of close to 7% of the genome should consist of functional noncoding DNA (after subtracting the 1.1% that is coding DNA).

But comparative genomics can also be used to *identify* the functional DNA elements. Functional noncoding DNA sequences are the primary objective here because coding DNA sequences have been comparatively easy to find by using a combination of bioinformatics programs (primarily hunting for open reading frames) and evolutionary conservation (coding sequences are generally well conserved across vertebrate genomes). Functional noncoding DNA sequences are a different matter. First, they are often not so well conserved as coding DNA sequences (many RNA genes are poorly conserved, and many enhancers, and some other regulatory sequences, are not so strongly conserved and can be lineage-specific). Add to that the lack of some easily identifiable characteristic for noncoding DNA (in stark contrast to open reading frames, the hallmark of coding DNA).

Happily, many regulatory sequences are quite strongly conserved, and comparative genomics has dramatically changed the landscape of regulatory sequence identification. When genomes of related organisms are aligned, large numbers of conserved noncoding sequence elements can be identified across the genome, including very many regulatory sequences. Depending on the evolutionary distance between the organisms, there is a trade-off between sensitivity of detection and the specificity. The number of conserved noncoding elements detected by comparing the human genome against that of the pufferfish, *Fugu rubripes*, is low (**Figure 13.4**); many genuine regulatory elements are not detected, but the specificity is very high, and virtually all of the detected sequences are worth investigating. For more closely related genomes, such as humans and mice, there is a much higher detection rate, but also a higher false-positive rate.

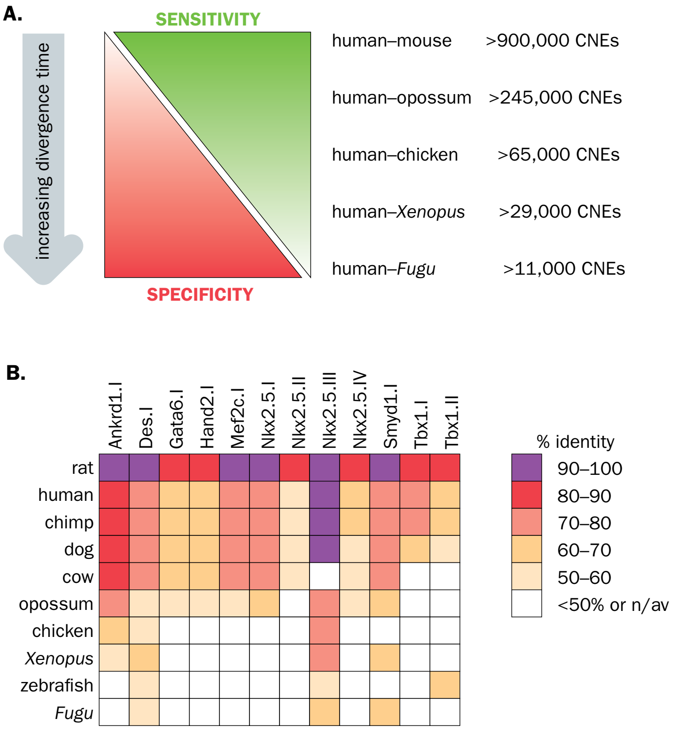


Figure 13.4 The sensitivity–specificity trade-off in detecting conserved noncoding elements by comparative genomics. (**A**) Many apparently conserved noncoding elements (CNEs) are identified by aligning the human and mouse genomes. Few functionally important CNEs are missed because of the high sensitivity, but many apparent CNEs may simply reflect the overall high human–mouse sequence similarity. The human and pufferfish (*Fugu*) genome sequences are much more distantly related. Few CNEs will be identified, but the great majority are very likely to be functionally important. (**B**) Vertebrate conservation of 12 known mouse *cis*-regulatory elements that are important in regulating cardiac gene expression. Most of the elements are minimally conserved beyond mammals and would have been missed by human–fish comparisons. (Adapted from Visel A *et al*. [2007] *Semin Cell Dev Biol* **18**:140–152; PMID 17276707. With permission from Elsevier.)

In order to identify all functional noncoding DNA elements, high-resolution maps of human evolutionary constraint are needed. To achieve that, the human genome must be aligned with the genomes of as many primates and other mammals as possible. When the human genome was compared in a multiple alignment with those of mouse, rat, and dog, there simply was not enough power to detect many evolutionarily constrained DNA elements. Adding the genomes of another 25 mammals, including the chimpanzee and some other primates (**Figure 13.5**), dramatically improved the situation and produced the first high-resolution map of human evolutionary constraint.

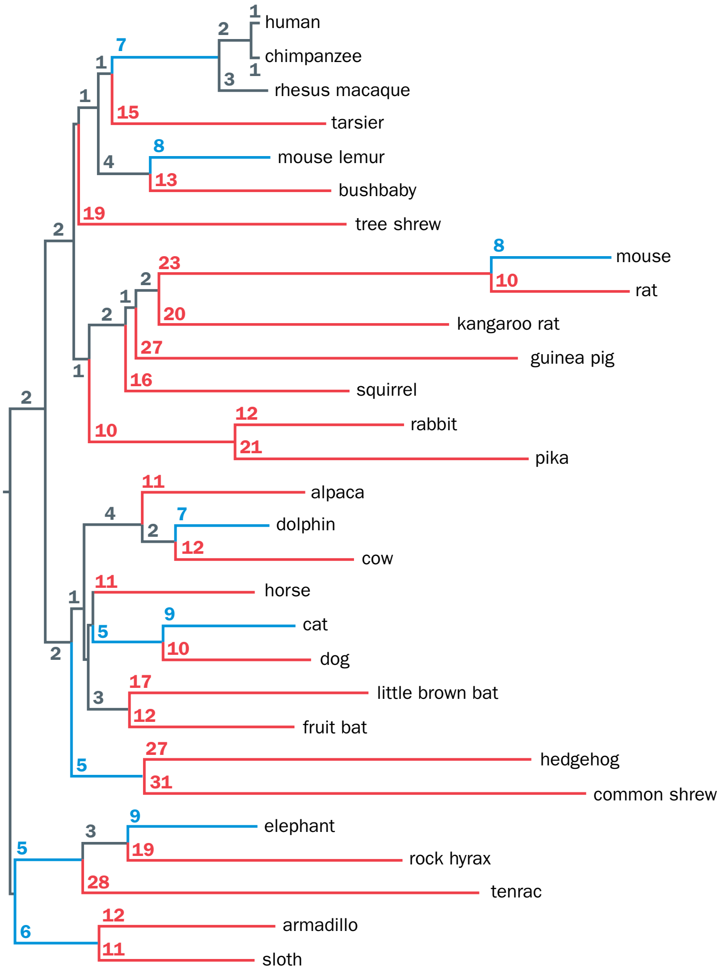


Figure 13.5 Phylogeny of 29 eutherian mammalian genome sequences used to build the first high-resolution map of human evolutionary constraint. The phylogenetic tree shown here was based on the substitution rates identified in MULTIZ alignments. At the time, only two of the genome sequences were finished sequences (human and mouse) and another seven had high-quality draft genome sequences (chimpanzee, rhesus macaque, rat, guinea pig, cow, horse, and dog); the others had genome assemblies. Substitutions per 100 bp are given for each branch; branches with 10 or more substitutions are colored red, and branches with less than 10 substitutions are colored blue. (Adapted from Lindblad-Toh K *et al*. [2011] *Nature* **478**:476–482; PMID 21993624. With permission from Springer Nature. Copyright © 2011.)

The power to detect constrained elements depends largely on the total branch length of the phylogenetic tree connecting the species. Comparing the 29 mammalian genomes gave a total effective branch length of about 4.5 substitutions per site, considerably greater than the 0.68 substitutions per site for the human–mouse–rat–dog comparison. In the latter case, the median size of the detected constrained DNA elements was 123 bp, and many short elements were not detected; in the 29-mammal comparison, the median size of detected constrained elements was 36 bp, with a power to detect all such elements that were 12 or more nucleotides long.

As well as confirming that at least 5.2% of the human genome has undergone purifying selection, the high-resolution map was able to identify 3.6 million constrained DNA elements, accounting for about 4.2% of the genome. To identify all constrained DNA elements, a map of human evolutionary constraint at single-nucleotide resolution will be required; current estimates indicate that 200 or so mammalian genomes will need to be aligned. By 2015, high-quality assemblies were available for about 50 mammalian genomes, and a 200 Mammals Genome Project had been initiated to sequence an additional 150 mammalian genomes, of which more than 90% had been sequenced by mid 2017.

Ultraconserved elements and rapidly evolving sequences

While many functional noncoding sequences are less well conserved than coding DNA sequences, alignment of genomes provided an opportunity to look for ultraconserved and rapidly evolving genomic DNA sequences. As originally defined, *ultraconserved elements* were identified as genomic sequences greater than 200 bp in length that were identical in humans, mice, and rats. Such an extraordinary level of sequence conservation would normally exclude coding DNA sequences (even if an encoded protein were to be 100% conserved in human, mouse, and rat, redundancy in the genetic code and base wobble at the third base of codons means that the respective coding DNA sequences would not be expected to be identical). Additionally, extraordinarily conserved coding sequences tend to be conserved over large evolutionary distances, but like enhancer elements, a typical human ultraconserved element does not have recognizable invertebrate homologs. Over 480 ultraconserved elements have been identified in the human genome, often located within introns, or close to genes that are involved in regulating transcription and development. They seem to mostly function as regulatory elements.

Rapidly evolving DNA sequences are ones that may be generally well conserved but have undergone rapid changes in a lineage, such as the human lineage. Sequences like this in the human genome, so-called *human accelerated regions* (*HAR*), have been of great interest. Current indications are that many of them work as developmental enhancers, and some of them are predicted to be involved in the evolution of unique human characteristics—we will return to this theme in Chapter 14.

Internet resources for comparative genomics

Various databases are dedicated to helping find homologs of genes and proteins of interest across species—see **Table 13.3** for some examples. A selection of genome browsers allows users to explore selected regions of a genome and to compare it with related genomes. Popular genome browsers that allow comparative genome analysis include notably those hosted by the University of California at Santa Cruz (UCSC) and the European Bioinformatics Institute (EBI) within the ENSEMBL suite. See **Table 13.4**

TABLE 13.3 EXAMPLES OF INTERNET RESOURCES FOR IDENTIFYING HOMOLOGOUS AND ORTHOLOGOUS GENES ACROSS SPECIES

Resource	Description	Website
Clusters of orthologous groups	Provides groups of orthologous proteins for various eukaryotes	http://www.ncbi.nlm.nih.gov/COG/
HCOP	Predicts orthology after combining orthology data for many species from different databases	http://www.genenames.org/cgi-bin/hcop
HomoloGene	Automated prediction of homologs for genes of several eukaryotes	http://www.ncbi.nlm.nih.gov/homologene
Inparanoid	Generates pairwise groups of orthologous proteins for various species	http://inparanoid.cgb.ki.se
OrthoDisease	Generates pairwise orthologs between human disease genes and genes from other species	http://orthodisease.cgb.ki.se
OrthoMCL-DB	Predicts orthologous groups of proteins for multiple species simultaneously	http://www.orthomcl.org/orthomcl/

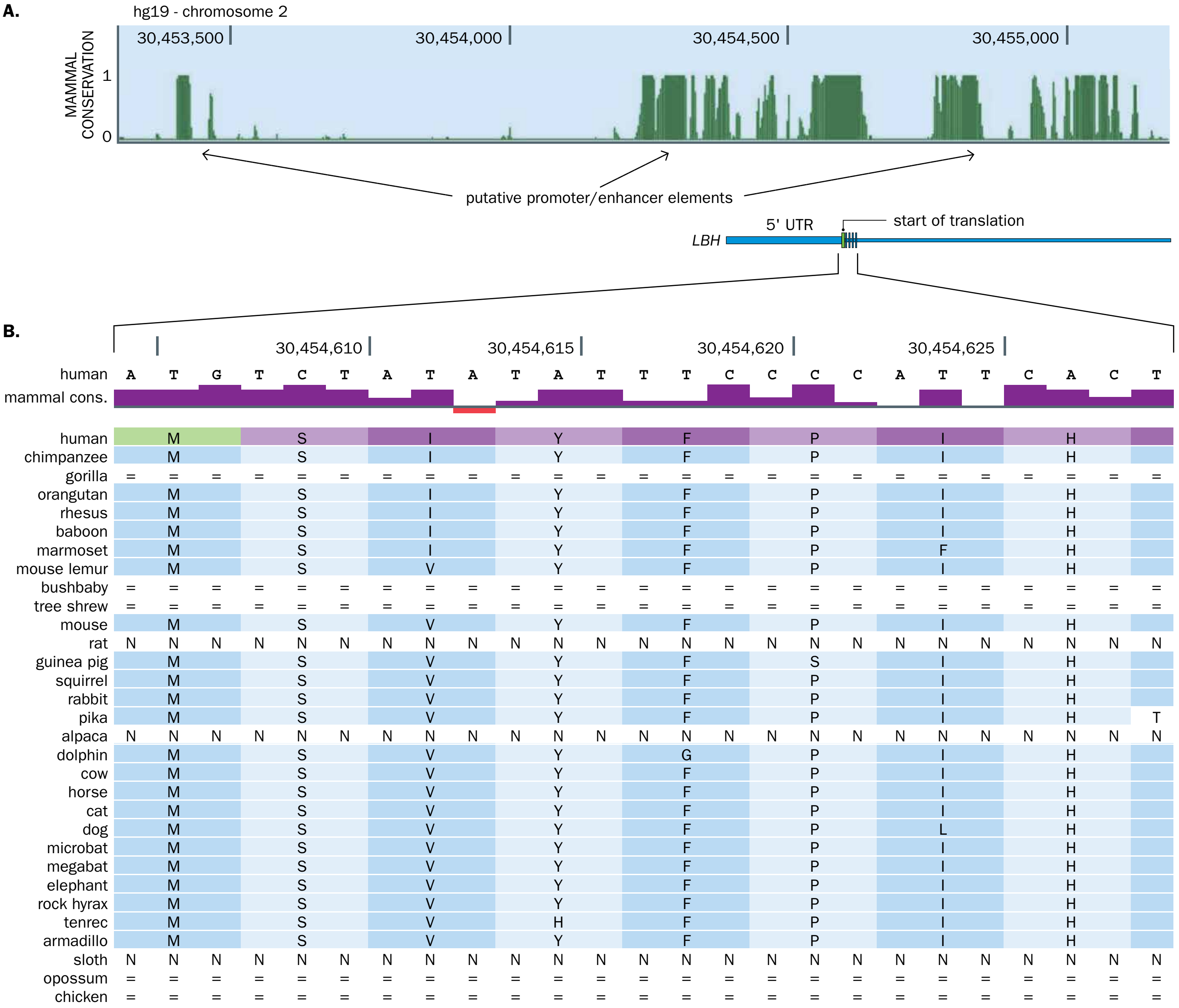
TABLE 13.4 PRINCIPAL GENOME BROWSERS FOR IDENTIFYING CONSERVED NONCODING SEQUENCES IN VERTEBRATES

Genome browser	Conserved elements identified using	Sequence alignment based on	URL
Dcode ECR Browser	PiP (percent identity plot)	BLASTZ	http://ecrbrowser.dcode.org
UCSC Genome Browser	PHASTCONS	MULTIZ (multiple, local)[a]	http://genome.ucsc.edu
VISTA Genome Browser	PiP	SLAGAN (pairwise, glocal)[a]	http://pipeline.lbl.gov
	GUMBY	MLAGAN (pairwise, glocal)[a]	
VISTA Enhancer Browser	GUMBY	MLAGAN (multiple, global)[a]	http://enhancer.lbl.gov

[a] In *global* alignment, one sequence string is transformed into the other; in *local* alignment, all locations of similarity between the sequence strings under comparison are returned. Global alignments are less prone to demonstrating false homology because each letter of one sequence is constrained to being aligned with only one letter of the other. However, local alignments can cope with rearrangements between nonsyntenic, orthologous sequences by identifying similar regions in sequences; this, however, comes at the expense of a higher false-positive rate. *Glocal* alignment is a combination of global and local alignment that allows the user to create a map that transforms one sequence into another while allowing for arrangement events.

for examples of genome browsers that allow users to access details of conserved noncoding sequences.

Figure 13.6 shows how the UCSC Genome Browser can reveal details of mammalian sequence homology to a selected human gene. Shown in **Figure 13.7** is an example of using the Vista Genome Browser to identify highly conserved noncoding DNA sequences within a 15 kb segment of human chromosome 2, enabling follow-up experimental investigations that showed clear evidence for some tissue-specific enhancer sequences.



human	A	T	G	T	C	T	A	T	A	T	A	T	T	T	C	C	C	C	A	T	T	C	A	C	T
human		M			S			I			Y			F			P			I			H		
chimpanzee		M			S			I			Y			F			P			I			H		
gorilla	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
orangutan		M			S			I			Y			F			P			I			H		
rhesus		M			S			I			Y			F			P			I			H		
baboon		M			S			I			Y			F			P			I			H		
marmoset		M			S			I			Y			F			P			F			H		
mouse lemur		M			S			V			Y			F			P			I			H		
bushbaby	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
tree shrew	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
mouse		M			S			V			Y			F			P			I			H		
rat	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
guinea pig		M			S			V			Y			F			S			I			H		
squirrel		M			S			V			Y			F			P			I			H		
rabbit		M			S			V			Y			F			P			I			H		
pika		M			S			V			Y			F			P			I			H		T
alpaca	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
dolphin		M			S			V			Y			G			P			I			H		
cow		M			S			V			Y			F			P			I			H		
horse		M			S			V			Y			F			P			I			H		
cat		M			S			V			Y			F			P			I			H		
dog		M			S			V			Y			F			P			L			H		
microbat		M			S			V			Y			F			P			I			H		
megabat		M			S			V			Y			F			P			I			H		
elephant		M			S			V			Y			F			P			I			H		
rock hyrax		M			S			V			Y			F			P			I			H		
tenrec		M			S			V			H			F			P			I			H		
armadillo		M			S			V			Y			F			P			I			H		
sloth	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
opossum	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
chicken	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=

Figure 13.6 A comparative genomics display derived from the UCSC Genome Browser. (**A**) Genomic region surrounding the 5′ end of the human *LBH* (limb bud and heart development homolog) gene. The top track indicates mammalian conservation as determined using the PHASTCONS program (see **Table 13.1**). Putative promoter and enhancer elements are indicated. The second track shows the 234 bp first exon of the human *LBH* gene, and start of the following intron. The first exon has a 209 bp 5' untranslated region (UTR) plus the first 25 nucleotides of coding sequence. (**B**) A close-up of the first 25 nucleotides of the *LBH* coding sequence. Here, the top track shows the human DNA sequence, and the second track, entitled "mammal cons.," shows the degree of mammalian conservation as determined by PhyloP. Below that are amino acid sequences for listed mammals, with chicken as a reference vertebrate. N indicates gaps in sequence; = indicates unalignable sequence. (Reproduced with permission from Alföldi J & Lindblad-Toh K [2013] *Genome Res* **23**:1063–1068; PMID 23817047.)

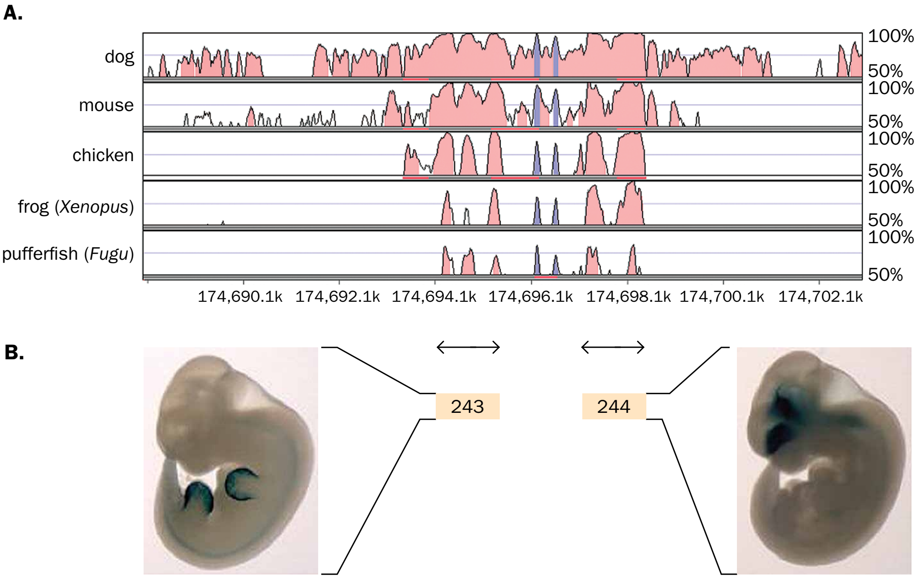


Figure 13.7 Using the VISTA Genome Browser to identify regions of highly conserved noncoding DNA. (**A**) The query here was a 15 kb region of human chromosome 2 (chr2: 174,688,000–174,703,000 in the hg18 human genome reference sequence (released in March 2006); it is located within the human *OLA1* gene (also called *PTD004*). This region was selected to be compared with five other vertebrate genomes and the outputs for individual genome comparisons are shown in the horizontal tracks as graphs that display percentage identity from 50% (bottom) to 100% (top). Horizontal tracks show comparison of the selected human DNA sequence query with individual genomes for the five vertebrate species listed. The VISTA genome browser uses gray-blue coloring to signifiy exons and pink to signify noncoding DNA. Two highly conserved *OLA1* coding exons are shown near the center of the graphs. Note that most of the noncoding DNA is poorly conserved, but two regions of highly conserved noncoding DNA can be identified, spanning the regions shown by lines with double arrowheads. (**B**) The VISTA Enhancer Browser stores experimental data for conserved noncoding DNA sequences identified as enhancers. The enhancer assay involves cloning the test DNA into a reporter vector with a *lacZ* gene driven by a minimal promoter. The construct is injected into a one-cell mouse embryo and after 11.5 days of development the mouse embryo is analyzed to see where the *lacZ* gene is expressed (blue signal). Here, we see that the conserved noncoding DNA regions identified in panel (**A**) are tissue-specific enhancers: enhancer 243 regulates limb-bud expression (left image), whereas enhancer 244 regulates forebrain expression (right image). See **Table 13.4** for Web addresses for the VISTA browsers.

13.2 GENE DUPLICATION, SPECIES DIFFERENCES IN GENE NUMBER, AND EVOLUTIONARY ADVANTAGES OF EXONS

This is the first of three sections where we look at aspects of genome evolution. Here, we primarily concentrate on the evolution of genome size (with the involvement of whole-genome duplication and gene duplication), differences between species in gene number, and the advantages of both gene duplication and exons. Although the primary focus will be on how vertebrate and mammalian genomes evolved, we occasionally consider aspects of earlier evolution and invertebrate lineages. The most closely related invertebrates are included along with vertebrates in the phylum Chordata (chordates; see **Box 13.3**). Other well-studied invertebrates, such as *Drosophila melanogaster* and *Caenorhabditis elegans*, are more distantly related to us, being members of two different phyla, Arthropoda and Nematoda, respectively.

BOX 13.3 CHORDATE CLASSES

Chordates are animals that go through an embryonic stage in which they possess a notochord, nerve cord, and gill slits. There are three major groups, as listed below.

- Craniates (all animals with skulls, including vertebrates and also hagfish, which lack a backbone).
- Cephalochordates (also called lancelets, or amphioxus, they are exclusively marine animals that resemble small, slender fishes but without eyes or a definite head or brain).
- Urochordates (also called tunicates or ascidians, they are underwater saclike filter feeders; they are exemplified by the sea squirt, *Ciona*).

The *C*-value and *G*-value paradoxes: organism complexity is not simply related to genome size or to the number of genes

Cross-species comparisons have demonstrated increasing protein and gene sequence complexity during the evolution of biologically complex organisms. That happened mostly as a result of acquiring extra DNA sequences, either by duplication of expressed intragenic segments, notably exons, or by insertion of additional DNA sequences into genes, including into exons. Whereas the genomes of single-celled organisms are rather compact, genome (and gene) sizes tend to have increased in the lineages leading to metazoans, and vertebrate genomes tolerate many large introns and also large amounts of highly repetitive DNA.

There is an inconstant relationship between the complexity of an organism and the amount of DNA in its cells (the **C-value paradox**), even when we discount polyploidy. The example of the onion in **Box 13.2** suggests a surprising degree of flexibility in the genome size of complex eukaryotes, but in some metazoan species, there appear to be constraints on the amount of genomic DNA that can be tolerated. For some reason, for example, there is a high genome-wide intrinsic rate of DNA loss in *Drosophila*, which appears to have constrained the extent of gene duplication. Rampant DNA deletion in the *D. melanogaster* genome has been thought to explain the comparative lack of both repeated sequences (all but 8% of the genome is composed of unique sequences) and also pseudogenes (only ~100 compared to >12,000 in the human genome).

Another important driver of increased organism complexity was an increase in total gene number. That was particularly important in allowing the evolutionary leap from unicellular organisms to metazoans. Unicellular genomes tend to have a few thousand protein-coding genes (often ~4000–5000 in bacteria; and ~6000 in yeast) whereas vertebrate genomes tend to have ~20,000 protein-coding genes. However, the expansion in gene number has been much less important in the evolution of metazoan lineages, and organism complexity is not simply related to gene number. Humans and the very simple nematode *Caenorhabditis elegans* each have close to 20,000 protein-coding genes (see **Box 13.2**), and *Trichomonas vaginalis*, a protist that causes the common sexually-transmitted trichomoniasis infection, has been reported to have significantly more protein-coding genes than we do (PMID 17218520).

The inconstant relationship between gene number and biological complexity has been called the **G-value paradox** (where *G*-value denotes gene number). It raises two important questions. First, if the total number of genes is not the primary determinant of organism complexity, what is? Secondly, why should some very simple organisms, including some examples of protists, need unusually large numbers of genes.

The answer to the first question is now widely thought to be *cis*-acting regulatory sequences, which are incontestably important in morphological development (we consider this aspect in Section 13.4), but are also required for regulating a wide variety of additional cellular programs throughout life. As for the unexpectedly high gene number in some very simple organisms, various suggestions have been proposed but the issue remains to be clarified.

Two or three major whole-genome duplication events have occurred in vertebrate lineages since the split from tunicates

As noted above, genome size increased in the transition from unicellular to multicellular organisms, and different mechanisms have been involved. An important one is whole-genome duplication (WGD), which offers a powerful way of increasing genetic complexity. At a stroke, the gene number doubles, initially. In most gene pairs, however, one of the genes is eventually lost from the genome, so that WGD is followed by a period of *diploidization*. Over time the net result is that the total gene content is significantly, but not greatly, increased above that pre-WGD. The gene pairs that survive are often genes retained because of dosage constraints (for example, where gene products work as part of a protein complex where precise ratios of the interacting proteins are important). Over long evolutionary timescales the retained duplicate genes can eventually diverge in function (**Figure 13.8**).

An ancient WGD event is difficult to detect because during the ensuing period of diploidization there is not only massive gene loss, but also major genome rearrangements (chromosome inversions, translocations, and so on that naturally occur over long periods of time). An evolutionarily recent WGD event, however, can be readily detected by identifying an organism that has twice as many chromosomes as closely related species. The most convincing example comes from *Paramecium tetraaurelia*: more than 50% of its genes can clearly be seen to be duplicated following a recent WGD. The *Paramecium* studies show that following WGD, gene loss does not occur initially on a massive scale; instead, gene loss occurs over a very long timescale, and highly expressed genes and genes encoding proteins functioning as part of complexes are more likely to be retained after WGD.

Polyploidy resulting from WGD is common in plants, but is also evident in some vertebrates, notably amphibians and reptiles, and in some bony fishes. Constitutional polyploidy in mammals is expected to be extremely rare, however, because of gene dosage difficulties arising from the X-Y sex chromosome system. For most land-dwelling vertebrates, the most recent WGD occurred early in chordate evolution: two rounds of genome duplication appear to have occurred in the vertebrate lineage shortly after the

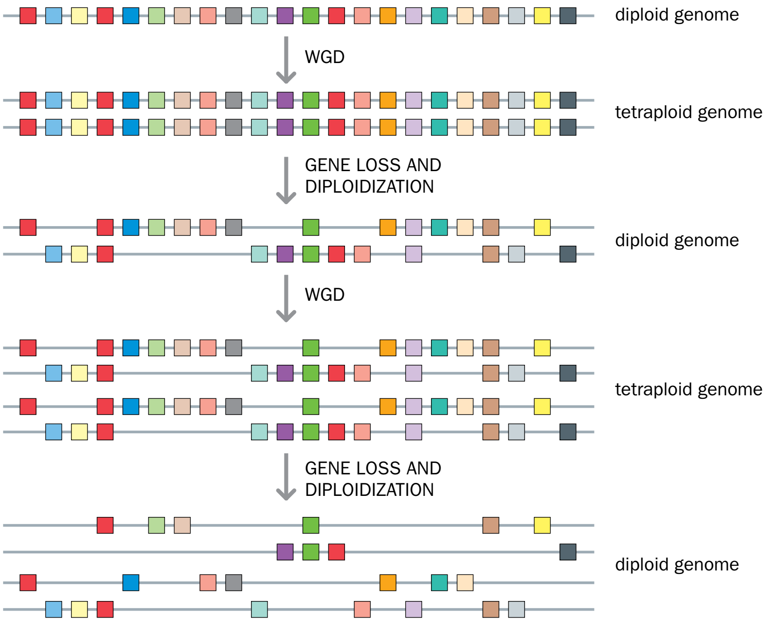


Figure 13.8 Paralogous genes originating from whole-genome duplication followed by large-scale gene loss. This hypothetical genome has 22 different genes; for clarity, only one haploid set is shown, but the genome is diploid. Whole-genome duplication (WGD) results in a complete series of paralogs in identical order and, initially, a tetraploid genome. In most of the paralogous gene pairs, one gene acquires disabling mutations to become a pseudogene and is eventually lost, restoring diploidy. A subsequent round of WGD followed by gene loss would result in paralogous sets, most commonly of two gene copies (as shown here; three or even four gene copies might also occur but would be much rarer). (Adapted from Dehal P & Boore JL [2005] *PLoS Biol* **10**:e314; PMID 16128622.)

split from the lineage giving rise to tunicates (such as *Ciona*). After the divergence from mammals, a major WGD event also occurred subsequently in the lineage leading to teleost (bony) fishes—see **Figure 13.9**.

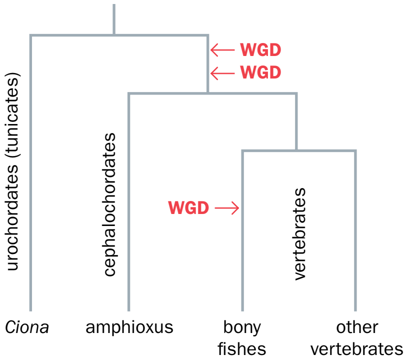


Figure 13.9 Major whole-genome duplication events arising in vertebrate lineages. The evolutionary lineages leading to most of the current vertebrates have undergone two major whole-genome duplication (WGD) events that took place after the split from tunicates some time within the approximate interval of 525–875 million years ago. In addition, a major WGD event occurred about 340–350 million years ago in the lineage leading to bony fishes (teleosts).

Gene duplication: an important evolutionary route toward developing functional complexity

Gene families are a major general feature of metazoan genomes, and emerged as a result of different gene duplication events in germ-line DNA. Genes arising from a duplication event are said to be **paralogs** (as opposed to **orthologs**, which describes sequences transmitted to species from a common ancestor but that diverge because of acquiring different mutations—see **Figure 13.10**). We detailed the different types of gene duplication in Section 9.2; here, we begin by providing a short summary of the major types of duplication, and then describe alternative outcomes of gene duplication.

Duplication giving rise to different functional genes usually occurs by copying a gene sequence at the DNA level. A gene copy may continue to be functional, and subsequently becomes fixed in the population (by then it has usually accumulated mutations, resulting in sequence divergence and the possibility of developing different characteristics from the parent gene). The alternative is that it acquires deleterious mutations and becomes a nonprocessed pseudogene. Different sizes of genomic DNA have been copied,

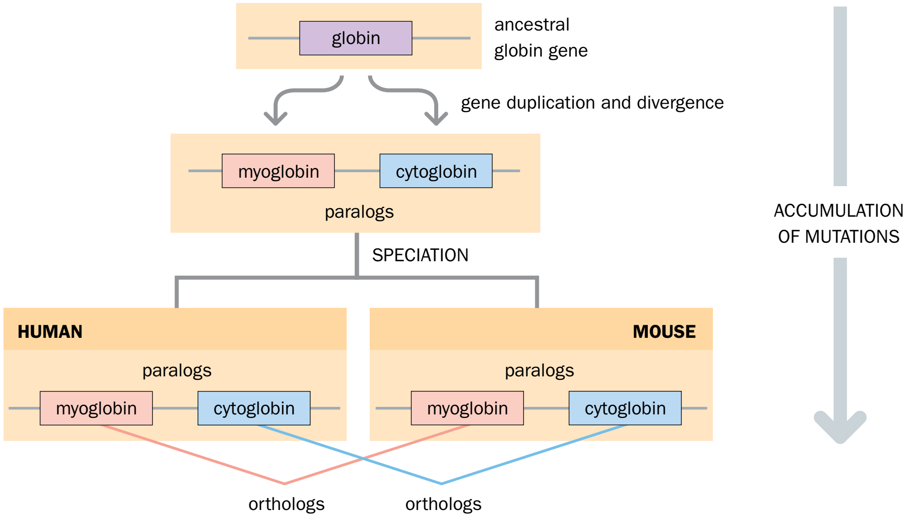


Figure 13.10 Orthologs and paralogs. Homologs, genes that have significant sequence identity suggesting a close evolutionary relationship, can be one of two types. *Paralogs* are closely related genes present in a single genome as a result of prior gene duplication, such as the myoglobin and cytoglobin genes. They may have been identical in sequence immediately after gene duplication, but then gradually accumulated mutations causing them to diverge in sequence. *Orthologs* are genes present in the genomes of different species that are directly related through descent from a common ancestor, such as the myoglobin genes in humans and mice, which originated by descent from a myoglobin gene present in the last common ancestor of mice and humans.

from whole genomes (see below) to single genes. Evidence for **segmental duplication** (recent duplication of large to moderately large DNA segments) is common in mammals, and especially so in primates (where segmental duplications showing >90% sequence identity can be up to 1 Mb in length, but are usually less than 300 kb). As described in Chapter 14, segmental duplications have been important in creating new genes in hominoid lineages.

Alternative gene duplication mechanisms start by copying a gene sequence at the RNA level: a gene transcript is converted into a complementary DNA (retrocopy) that integrates into the genome. A retrocopy of a protein-coding gene lacks the promoter, and also regulatory elements located within introns of the gene or near to the gene. If, however, the retrocopy integrates close to other *cis*-acting regulatory sequences, it may be expressed, and may become a functional retrogene; for an overview of the functions of human retrogenes, see **Box 9.2**. More frequently, the retrocopy acquires deleterious mutations and becomes a processed pseudogene (also called a retropseudogene).

Increased gene dosage

One outcome of gene duplication, but a rare one, is that a duplicated gene is simply retained to carry out the same function as the parent gene, simply allowing more of the same gene product to be made (**Figure 13.11A**). That can be advantageous when there is an important need to make large amounts of gene product: each of our different ribosomal RNAs is made by hundreds of almost identical gene copies. Natural selection can also occasionally drive gene duplication to provide an advantageous increase in gene product in response to an altered environment. For example, selection pressure has resulted in amplification of the *AMY1* gene, which makes the enzyme salivary amylase, as a way of facilitating starch digestion in human populations with a long tradition of high-starch diets. (Human populations with a tradition of low starch diets have lower numbers of *AMY1* gene copies; the chimpanzee has a single *AMY1* gene.)

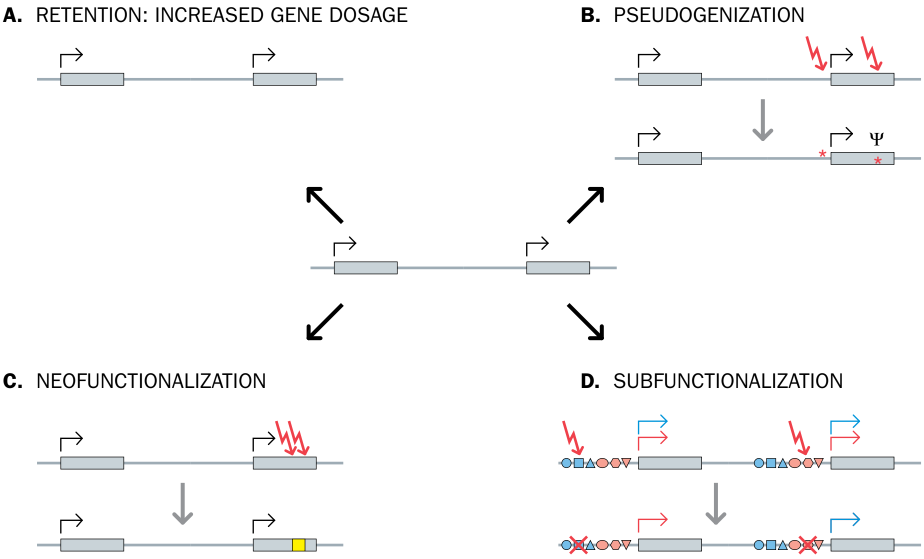


Figure 13.11 Fate of duplicated genes. (**A**) Retention. Natural selection occasionally works to increase gene dosage when there is a selective advantage for more gene product. Usually, however, the two gene copies undergo some type of sequence divergence. (**B**) Pseudogenization. The most common fate for duplicated genes is that natural (purifying) selection maintains one functional gene copy, but after purifying selection is relaxed on the second gene copy it acquires deleterious mutations (red asterisks) and becomes a pseudogene (Ψ). (**C**) Neofunctionalization. Here, one of the duplicated genes retains the function of the ancestral gene. The other gene undergoes mutations (red arrows); as a result, it develops altered characteristics and acquires a new function. Here we show mutations producing an altered structure (yellow box), but the mutations might instead have resulted in altered regulatory control leading to a significantly altered or new function. (**D**) Subfunctionalization. The duplicated gene copies undergo complementary deleterious mutations, often at the level of regulatory elements, as shown here. In this example, the ancestral gene is imagined to be regulated by two sets of upstream control elements (blue and orange symbols), so that it is expressed in two different cell types, or different tissue types, or at different developmental stages. Mutations (red arrows) inactivate one set of control elements in one paralog and the other set in the second paralog so that the expression patterns of the ancestral gene are partitioned between the two daughter genes.

Divergent function of duplicated genes

The most common outcome after gene duplication is that the duplicated genes accumulate different mutations and diverge in sequence. If purifying selection acted on one of the duplicated genes to maintain a functional sequence, the other gene might be expected to mutate freely. In that case, a common outcome is that the gene copy acquires deleterious mutations and becomes a pseudogene (**Figure 13.11B**). Sometimes, however, two diverged functional gene copies are retained. The diverged gene copies may acquire different properties and often they come to be expressed in different ways that are functionally advantageous. In this way, gene duplication is thought to be a major motor that drives increasing functional complexity.

Duplicated genes may acquire different functions by different means. One possibility is that one of the duplicated genes retains the function of the parent gene and is subject to purifying selection, while the other acquires a distinctive new function (*neofunctionalization*—**Figure 13.11C**). Pure neofunctionalization is rare, however. Instead, duplicated genes have been thought to develop expression or functional differences as a result of acquiring complementary mutations that remove different subsets of the overall expression characteristics or functional characteristics of the original gene. Imagine, for example, complementary mutations that inactivate different *cis*-acting regulatory sequences: the expression domains of the ancestral gene could be maintained but partitioned between the duplicated daughter genes (*subfunctionalization*—**Figure 13.11D**). Over evolutionary time periods, the duplicated genes may acquire slightly different functions, simply by being expressed in different tissues (where they may interact with different proteins). Subfunctionalization is a slow process, however.

In the *escape from adaptive conflict* model, adaptive mutations are involved in partitioning of gene function between duplicated genes. In this model adaptive conflict is envisaged between an evolutionary old function and an emerging new function *within a single gene*. The adaptive conflict within the single gene is then imagined to drive gene duplication and fixation, so that one of the duplicated genes assumes sole responsibility for the old function, and the other gene has sole responsibility for the new function. That is, divergent functions may develop in a protein made by a single gene and selection drives gene duplication to resolve the adaptive conflict. The evolution of a type III antifreeze protein gene in certain fish living in Antarctic waters illustrates this mechanism. The ancestral gene, which made products with both sialic acid synthetase activity and rudimentary ice-binding properties, duplicated so as to partition the two functions. That is, one of the duplicated genes became exclusively responsible for making the sialic acid synthetase, and the other gene evolved to make an efficient antifreeze glycoprotein (see Deng C *et al*. [2010]; PMID 21115821).

Gene duplication and divergence: the example of the globin gene family

Globin genes are of ancient evolutionary origin, being present in all three domains of life, and different globin chains are adapted to different cellular and developmental environments. Bacterial globins are known to work in responses to nitric oxide and nitrosative stress, but vertebrate hemoglobins and myoglobin, the first globins to be studied, were investigated because of their role as transport proteins in blood and muscle, respectively, and because of interest in their O_2- and CO_2-binding properties. Myoglobin was subsequently shown to also have a critical role as an intrinsic nitrite reductase that regulates responses to cellular hypoxia and re-oxygenation; other vertebrate globins may have similar functions.

Whereas hemoglobin operates as a tetramer with two copies each of two different globin chains, myoglobin works as a monomer. In jawed vertebrates three further types of globin are found in all, or almost all, lineages: neuroglobin (as a monomer in neurons and in some endocrine tissues); cytoglobin (a homodimer, found in fibroblasts and related cell types, and in distinct nerve cells); and androglobin (a chimeric globin, having an N-terminal calpain domain and an internal globin domain, and expressed preferentially in testis). Three additional globins—globins X, Y, and E—are restricted to certain classes of vertebrate; see the legend to **Figure 13.12**. The functions of the recently identified vertebrate globin genes are currently unclear.

The current globin superfamily originated by a series of gene duplications. Early gene duplications arose as a result of whole-genome duplication; more recent duplications originated by tandem gene duplication. Some of the recent duplications were clearly unproductive, giving rise to pseudogenes (see **Figure 13.12**), but the overall value of gene duplication has been to produce different variant globin chains that are adapted to different cellular and developmental environments. By partitioning regulatory control sequences, subfunctionalization could have resulted in some globins being restricted to specialized tissues with the possibility of acquiring modified functions, maybe even additional or divergent functions in the case of neuroglobin.

The blood globins have subsequently undergone further specialization, but this time differences in gene regulation result in differences in developmental timing. Thus, in early human development ζ-globin is expressed instead of α-globin, which will take its place at later stages, and ε-globin (embryonic period) and γ-globins (fetal period) are used instead of β-globin that begins to be synthesized at three months' gestation and gradually accumulates while γ-globin production declines. One rationale is that the globins used in hemoglobin during embryonic and fetal periods are better suited to the more hypoxic environment at these stages.

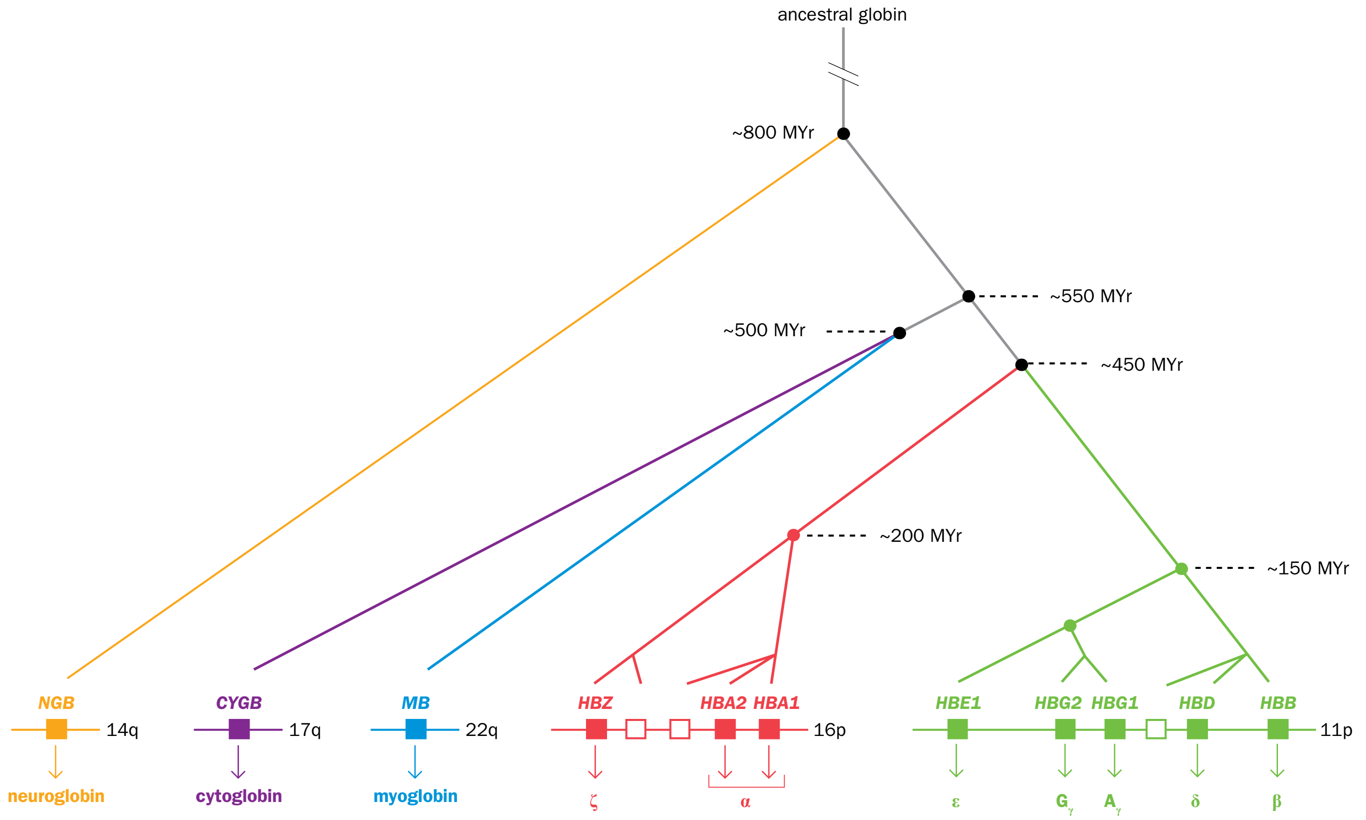


Figure 13.12 The globin superfamily evolved by a series of gene duplications. The best-characterized human globin genes, shown at bottom, are located on five chromosomes (not shown is the androglobin gene *ADGB*, located at 6q24, which makes an enigmatic testis-expressed chimeric globin with an N-terminal calpain domain). Globin genes arising from evolutionarily ancient duplication have subsequently been separated by chromosomal rearrangements. More recent tandem gene duplications produced gene clusters at 16p (the α-globin gene family) and 11p (the β-globin gene family). Here, the globins encoded by genes within a cluster show a greater degree of sequence identity than for globins encoded by genes in different clusters. Some duplications have been very recent: *HBA1* and *HBA2* encode identical α-globins, and *HBG1* and *HBG2* produce γ-globins differing by a single amino acid. Other vertebrates show differences in copy number (for example, mice have a single α-globin and two β-globin genes, goats have three copies of the full β-globin gene cluster, and many fish have two cytoglobin genes). Additional globins not found in humans are globin X (present in many bony vertebrates but not in mammals), globin Y (in many lobe-finned fishes and some amphibians), and globin E (notably found in birds). MYr, millions of years.

The evolutionary importance of lineage-specific gene copy number changes

Within lineages, some genes can be duplicated and others can be lost or inactivated, often with functional consequences that make species from different lineages different from each other. As a result, the copy number of individual gene families can vary very significantly between organisms, and is the most striking difference between closely related species. The great majority of human genes have counterparts in mice, and close to 14,000 human and mouse genes can be seen to be simple 1:1 orthologs, having retained these genes from the common ancestor of humans and mice. But because the human and mouse genomes have undergone extensive remodeling in the ~90 million years following human–mouse divergence from the common ancestor, there can be major gene copy number differences in the two species for the remaining genes (which are usually members of gene families).

Comparison of the completed human, chimpanzee, mouse, and rat genomes reveals a few gene families that are present in high copy number in rodent genomes but absent from primate genomes, for example. Many such gene families have roles in reproduction. In mouse, they include over 100 *Speer* genes that specify a type of spermatogenesis-associated glutamate (E)-rich protein, and more than 200 genes encoding type 1 or 2 vomeronasal receptors (thought to act as pheromone receptors). Similar major copy number differences are evident in invertebrate genomes: different *Caenorhabditis* species have many hundreds of seven-transmembrane chemoreceptor and nuclear hormone receptor genes whereas *Drosophila* species have a few tens of each.

Extended comparative genomics analyses show that the differences in gene copy number do not simply indicate rapid gene family expansion on certain lineages, but a rather complicated process of gene duplication and gene loss. Vomeronasal receptors, for example, are found in a range of vertebrates; they are expressed in the vomeronasal organ, an auxiliary olfactory sense organ found in many animals but absent in primates, presumably as a result of evolutionarily recent gene loss.

Lineage-specific gene family expansions often involve what have been called environmental genes; that is, genes involved in responding to the external environment. Their products may be chemoreceptors that can be involved in various functions such as sensing odors, pheromones, and so on (**Table 13.5**) or they may be involved in immunity, in response to infection, and in degrading toxins (such as the cytochrome P450 family). The example given above of the rapid, very recent expansion of α_1-amylase genes in some human populations constitutes another example of genes responding to change in the external environment.

TABLE 13.5 HIGHLY VARIABLE COPY NUMBER FOR CHEMOSENSORY RECEPTOR GENE FAMILIES IN DIFFERENT VERTEBRATES

Gene family	Number of functional gene copies									
	Human	Chimp	Mouse	Rat	Dog	Cow	Opossum	Platypus	Chicken	*Xenopus*	Zebrafish
OR	388	399	1063	1259	822	1152	1198	348	300	1024	155
V1R	-	-	187	106	8	40	98	270	-	21	2
V2R	-	-	121	79	-	-	86	15	-	249	44
TAAR	6	3	15	17	2	17	22	4	3	6	109

OR, olfactory receptor; *V1R*, vomeronasal type 1 receptor; *V2R*, vomeronasal type 2 receptor; *TAAR*, trace amine-associated receptor. (Data abstracted from Nei M *et al*. [2008] *Nat Rev Genet* **9**:951–963; PMID 19002141.)

A rationale for some gene family expansions in some species may be straightforward. It is not surprising that rodents have many more olfactory receptor genes than we do. In mammals, olfactory neurons are monospecific: they express just one type of olfactory receptor. Mice and rats rely much more on the sense of smell than we do, and so by having about three times as many olfactory receptor genes they are much better than us at discriminating between a multitude of smells.

It has recently become clear that gene loss is also common during evolution. Take, for example, the 12 vertebrate *Wnt* gene families, which have counterparts in invertebrates. The jellyfish, an invertebrate that diverged very early from the lineage leading to vertebrates, has counterparts of 11 of the 12 *Wnt* genes, but in more recently diverged invertebrate lineages some of the *Wnt* gene counterparts inherited from a common ancestor have been lost (*D. melanogaster* and *C. elegans*, for example, have counterparts, respectively, of only seven and four of the 12 vertebrate *Wnt* families).

Although orthologous genes may make extremely similar proteins in closely related species, therefore, some genes present in one species are absent in another as a result of lineage-specific gene loss. And, as described below, occasionally novel genes arise *de novo* from noncoding DNA. The gene catalogs can therefore differ a little between closely related species, and some of these differences may make significant contributions to regulatory sequence divergence. We consider this in Section 13.4.

Gene copy number changes in primate lineages

In the case of primates, there has been differential gene loss in the lineages leading to modern humans and chimpanzees: during the same time period, 729 genes have been lost in the chimpanzee lineage and 86 genes in the human lineage. We give examples in Chapter 14 where gene loss/gene inactivation in human lineages may have been important in the development of human-specific characteristics.

As for gene expansion, the primate KRAB (Kruppel-associated box)-zinc finger gene family appears to be one of the fastest-growing gene families. KRAB-zinc finger (KRAB-ZNF) proteins account for almost one-half of the ~700 human C2H2 (Kruppel) zinc finger transcription factors (by far the largest human transcription factor family), and get their name because they contain a potent KRAB repression domain and tandem arrays of usually three zinc fingers that are involved in DNA binding, each recognizing a specific three-nucleotide motif.

Two characteristics make the KRAB-ZNF transcription factors exceptional. First, they are able to bind to long stretches of DNA by combinatorial binding of up to several dozen different KRAB-ZNF proteins. Second, a large fraction of KRAB-ZNF proteins have DNA-binding domains that are rapidly evolving. A large proportion of evolutionarily young KRAB-ZNF proteins form part of a surveillance system that protects mammalian genomes by repressing transcription of infectious retroviruses plus endogenous retroviruses and nonviral retrotransposons. The rapid evolution of the DNA-binding domains seems to be a consequence of an evolutionary arms race between these proteins and potentially harmful retroviruses and retrotransposons.

Exon duplication and exon shuffling result in increased gene complexity

Gene duplication and gene loss/inactivation have underpinned important functional differences in the evolution of different species. An evolutionarily earlier catalyst toward developing greater functional complexity was the evolution of spliceosomal introns, which require dedicated spliceosomes to carry out the splicing reaction; they are found in most eukaryotic genes. (The other major class of introns, autocatalytic introns, are self-splicing; present in both bacteria and eukaryotes, they are comparatively rare, being mostly limited to certain rRNA and tRNA genes and to a few protein-coding genes in some types of mitochondria and chloroplasts.)

Spliceosomal introns have long been thought to have evolved from group II introns, a subclass of autocatalytic introns that subsequently lost their capacity for self-splicing. Believed to have arisen in eukaryotes, they likely inserted into genes at an early stage in eukaryote evolution, but there is evidence for some later integrations too. However, the great majority of introns pre-date chordate evolution. Although humans and the primitive chordate *Amphioxus* diverged from a common ancestor approximately 550 million years ago, comparison of the human and *Amphioxus* genomes reveals that 85% of the introns are conserved at precisely analogous locations.

During the evolution of complex genomes, spliceosomal introns expanded in size by insertion of different types of DNA sequence, notably repetitive DNA sequences originating by retrotransposition (where RNA transcripts are naturally converted into cDNAs by reverse transcriptases and then integrate into genomic DNA). As a result, exons came to be well-separated in the genomic DNA. By splitting functionally important genomic sequences into separate exons, introns made it possible for the separated DNA segments (exons) to be duplicated, or for exons from different genes to be combined in evolutionarily advantageous ways, as described below.

Exon duplication

Tandem duplication of exons is evident in about 10% of genes in humans, *D. melanogaster*, and *C. elegans*, and is often responsible for generating repeated protein domains (see **Figure 9.8** for an extraordinary example). Unequal crossover (or unequal sister chromatid exchange) can result in intragenic duplication; that is, a DNA segment spanning one or more exons of a gene is duplicated.

Exon duplication offers several types of advantage. It can result in extension of a structural domain, as in the case of collagens. For example, 41 exons of the *COL1A1* gene encode the part of α1(I) collagen that forms a triple helix; each exon encodes essentially one to three copies of an 18-amino-acid motif composed of six tandemly repeated (Gly-X-Y) tripeptides, where X is often hydroxyproline and Y is often either hydroxyproline or hydroxylysine. Alternative splicing can also produce different isoforms by selecting one exon sequence from a group of duplicated exons that have diverged in sequence.

Exon shuffling

Protein domains are rarely restricted to one type of protein. For example, the protein-binding kringle domains of lipoprotein A, Lp (a), are widely found in blood clotting factors and fibrinolytic proteins, and the type II domains of fibronectin are found in many cell surface receptors and extracellular matrix proteins. Different exon-shuffling mechanisms can give rise to spreading of protein domains to different proteins. Nonallelic recombination is one possibility, but duplicative transposition is likely to have made by far the most important contribution: retrotransposons offer the possibility of a copy-and-paste mechanism that means domains are retained in the donor gene and copied into an acceptor gene (**Figure 13.13**).

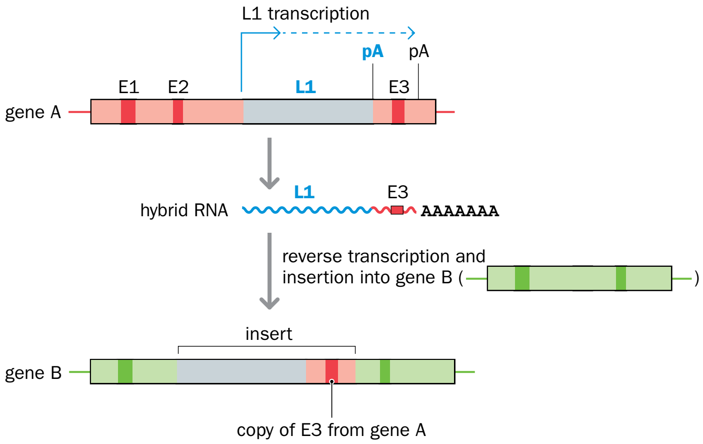


Figure 13.13 Exon shuffling between genes can be mediated by transposable elements. Exon shuffling can be carried out using retrotransposons such as actively transposing members of the LINE-1 (L1) sequence family as shown here. LINE-1 elements have weak poly(A) signals and so transcription often continues past such a signal until another nearby poly(A) signal is reached (for example, after exon 3 [E3] in gene A). The resulting RNA copy contains a transcript not just of LINE-1 sequences but also of a downstream exon (in this case E3). The LINE-1 reverse transcriptase machinery can then act on the extended poly(A) sequence to produce a hybrid cDNA copy that contains both LINE-1 and E3 sequences. Subsequent transposition into a new chromosomal location may lead to insertion of exon 3 into a different gene (gene B)—see **Box 9.2 Figure 1** for the mechanism of retrotransposition.

13.3 EVOLUTION OF MAMMALIAN CHROMOSOMES

The genomes of prokaryotes are generally small circular genomes. Linear chromosomes became prominent in eukaryote evolution. The nuclear genomes became much larger and needed to be highly structured by being wound around histones, allowing a huge degree of compaction so that cell division could occur without the DNA getting entangled. The size and structuring of nuclear DNA would have imposed an unbearable torque strain if the DNA were circular; but the very small mtDNA and chloroplast DNA remained circular because there was less torque strain. The evolutionary transition to linear DNAs imposed additional requirements. To protect the ends of the linear DNA molecules, specialized telomere structures were needed at the ends of chromosomes; otherwise the emergency machinery that responds to double-strand breaks within cells would be activated. And telomerases evolved to compensate for the limitations of DNA-directed DNA polymerases that synthesize new DNA strands in the 5′-to-3′ direction only and so cannot extend the lagging strand at the end of a linear DNA molecule (the end-replication problem, described in **Figure 2.24**).

Here we focus primarily on how mammalian chromosomes evolved. We show how the chromosome number and form can be surprisingly different between some closely related species. We explain how the very different mammalian sex chromosomes, the X and Y chromosomes, were formed from a pair of ancestral autosomal homologs, and the origins of pseudoautosomal regions and X-inactivation. We also illustrate how the mammalian X-Y sex determination system differs from other sex determination systems in animals.

Major chromosome rearrangements have occurred during mammalian genome evolution

Large-scale chromosome rearrangements have been frequent during mammalian genome evolution, and sometimes chromosome evolution can be seen to be uncoupled from phenotype evolution. A classic example is provided by the Chinese and Indian species of muntjac, a type of small deer. The two species are so closely related that they can mate to produce viable hybrids (but the males are sterile). The Chinese muntjac has 46 chromosomes, but as a result of various chromosome fusion events, the Indian muntjac has far fewer, but much larger, chromosomes (**Figure 13.14**).

This example of karyotype divergence in closely related muntjacs is an extraordinary one, but various types of chromosome rearrangements occur regularly during mammalian evolution. Inversions appear to be particularly frequent; translocations somewhat less so. Centromeres (which have rapidly evolving DNA sequences) can also change positions. During evolution, new centromeres periodically form in euchromatin. Although the underlying mechanisms are poorly understood, studies of such constitutional *neocentromeres* in humans suggest that they tend to form at certain chromosomal hotspots, and that chromosome rearrangements can induce neocentromere formation.

Comparison of mammalian chromosome sets shows that, in general, conservation of gene order is limited to small chromosome segments. Genes on one human chromosome usually have orthologs on several different mouse chromosomes, and genes on a mouse chromosome tend to have orthologs on different human chromosomes. The human–mouse comparisons show that the size of conserved blocks of orthologous

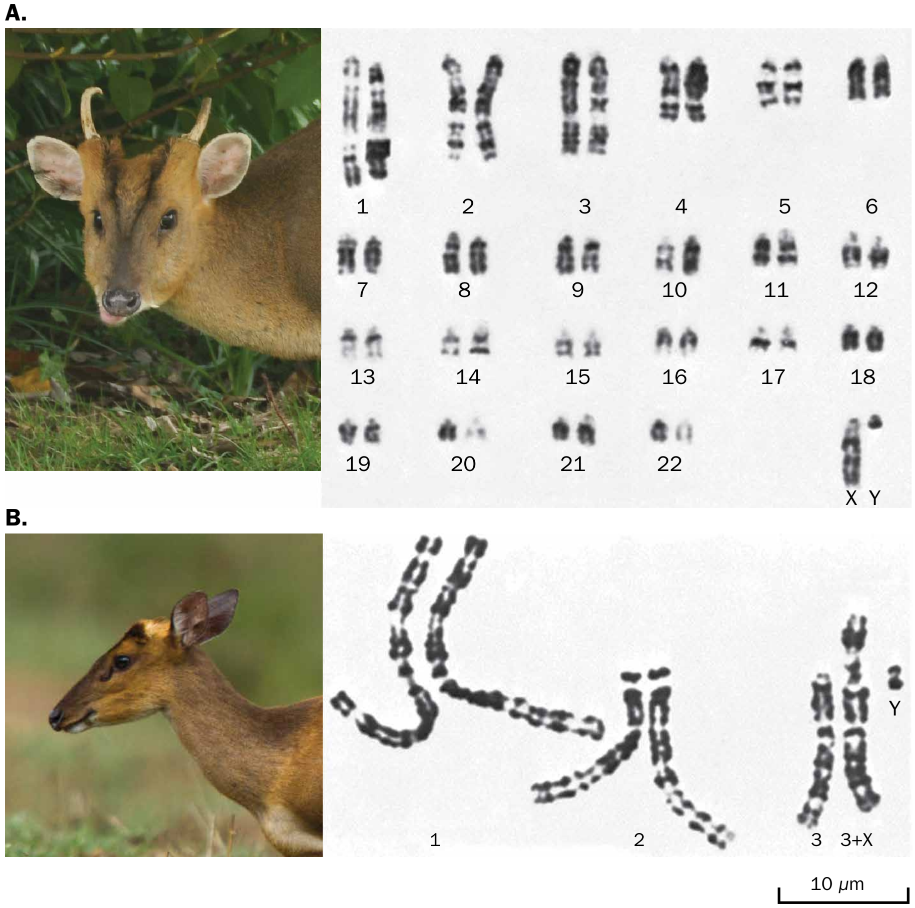


Figure 13.14 Uncoupling of chromosome evolution from phenotype evolution. (**A**) The Chinese muntjac (*Muntiacus reevesi*) has 2n = 46 chromosomes, all acrocentric. (**B**) The very closely related Indian muntjac (*Muntiacus muntjak*) has 6 chromosomes in females and 7 chromosomes in males (largely because of multiple chromosome fusions that occurred in its evolutionary lineage). As a result, the Indian muntjac has a pair of very large metacentric autosomes labeled 1, and two pairs of acrocentric autosomes (labeled 2, 3). The X chromosome is fused to the centromeric end of autosome 3. The very small Y chromosome remains separate. (Karyotype images courtesy of R Church. A, courtesy of Nilfanion/CC BY-SA 3.0; B, courtesy of NA Naseer/www.nilgirimarten.com/naseerart@gmail.com/CC BY-SA 2.5 India.)

genes that are on the same chromosome for both species (called *conserved synteny segments*) are, on average, somewhat less than 10 Mb in size. The X chromosome is a notable exception, however (see **Figure 13.15**). X-chromosome inactivation evolved to ensure an effective 2:1 gene dosage ratio for autosomal:X-linked genes, and so genes on one mammalian X chromosome typically have orthologs on the X chromosome of other mammalian species. Although synteny may be largely conserved in the case of mammalian X chromosomes, gene order is not. For example, the linear order of orthologs on the human and mouse X chromosomes has been scrambled because of numerous lineage-specific inversions. Analysis of the lengths and numbers of conserved synteny segments on the autosomes fits with a process of random chromosome breakage.

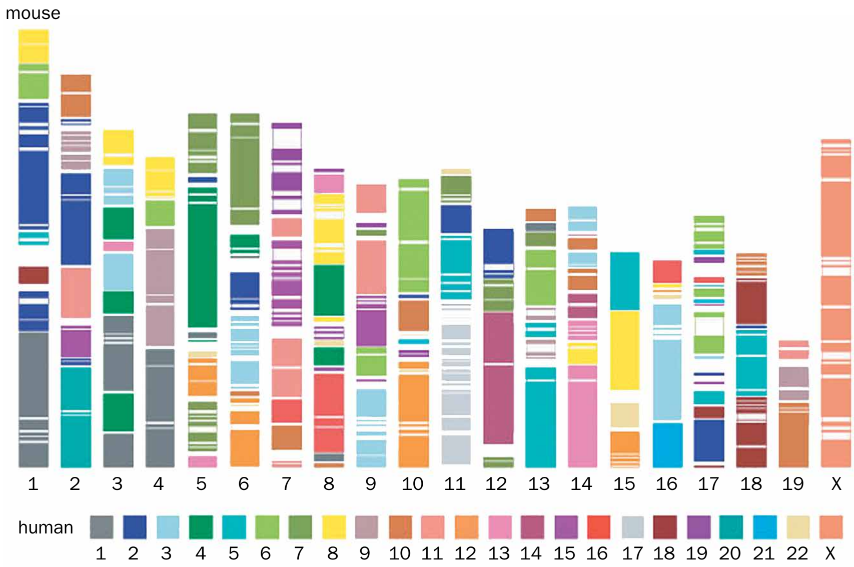


Figure 13.15 Human–mouse conservation of synteny is generally limited to small chromosome segments. Segments and blocks >300 kb in size with conserved synteny in humans are superimposed on the 20 mouse chromosomes. Each color corresponds to a particular human chromosome. The 342 segments are separated from each other by thin, white lines within the 217 blocks of consistent colour. Human X-chromosome segments are all found on the mouse X chromosome as single, reciprocal syntenic blocks. Human chromosome 20 corresponds entirely to a portion of mouse chromosome 2, and human chromosome 17 corresponds to mouse chromosome 11, but with extensive rearrangements into at least 16 segments. Other chromosomes show evidence of much more extensive interchromosomal rearrangement. (Reprinted from the Mouse Genome Sequencing Consortium [2002] *Nature* **420**:520–562; PMID 12466850. With permission from Nature Publications Ltd. Copyright © 2002.)

When human chromosomes are compared with those of our nearest living relatives, the great apes, there are very strong similarities in chromosome banding patterns. The most frequent rearrangements have been inversions, and some translocations have also occurred, but the most obvious difference is that the haploid chromosome number is 24 in great apes as opposed to 23 in humans. After the human lineage diverged from those of the great apes, two chromosomes fused to give human chromosome 2.

In heteromorphic sex chromosomes the smaller chromosome is limited to one gender and is mostly nonrecombining with few genes

Different sex-determining systems have evolved in vertebrates. For some fish and reptiles, environmental factors—notably, the temperature at which an egg is incubated—are the primary determinants of the sex of the developing organism. However, genetic sex determination is prevalent in mammals, birds, and amphibians, and also occurs in some fish, reptiles, and invertebrates (see **Table 13.6**).

TABLE 13.6 EXAMPLES OF DIFFERENT GENETIC SEX DETERMINATION SYSTEMS

System	Sex chromosomes	Distribution/examples	Comments
XY	XY (male, heterogametic); XX (female)	Prevalent in mammals; also found in some amphibians and some fish[b]	Y chromosome carries male-determining factor; sperm are X or Y, and so determine the sex
ZW	ZW (female, heterogametic); ZZ (male)	Prevalent in birds and reptiles, but also in some amphibians and some fish[b]	Eggs are Z or W, and so determine the sex
X:autosome ratio	XX (female/hermaphrodite); X0[a] / XY (male)	In many types of insect and nematodes	*C. elegans* is an example in which XX individuals are hermaphrodites; *D. melanogaster* has a Y chromosome but it does not confer maleness
Haplo-diploid	Relies on ploidy only	In many types of insect	Unfertilized eggs produce haploid fertile males; fertilized eggs are diploid and give rise to females or sterile males

[a] The 0 denotes zero, and so X0 means a single X. [b] There can be different systems in some kinds of organism. For example, some medaka species have the XY system; others have the ZW system.
Note: the XY and ZW nomenclature simply indicate whether the heterogametic sex is male (XY system) or female (ZW system).

Genetic sex determination involves the development of an unmatched pair of chromosomes. In some cases of genetic sex determination, the sex chromosomes appear to be *homomorphic*; that is, virtually identical at the cytological level (like autosomal homologs) but presumed to differ at the gene level. In other cases, the sex chromosomes can clearly be seen to be structurally different, and are said to be *heteromorphic*.

In organisms that have heteromorphic sex chromosomes, one gender has two different sex chromosomes and is said to be *heterogametic*, being able to produce gametes with either one of the two sex chromosomes. The other gender is *homogametic* and produces gametes with the same sex chromosome. For some organisms, such as mammals, the male is heterogametic, and the sex chromosomes are labeled as X and Y. For other organisms, such as birds and reptiles, the female is heterogametic, and here the sex chromosomes are labeled Z and W.

In both the X-Y and Z-W sex chromosome systems, the chromosome that is exclusively associated with the heterogametic sex (Y and W) is comparatively small and poorly conserved, with very few genes and with high amounts of repetitive DNA. For example, the human X chromosome contains many genes (including over 800 protein-coding genes) within 155 Mb of DNA. The human Y chromosome has just 59 Mb of DNA and much of it is genetically inert constitutive heterochromatin (effectively the Y chromosome encodes just ~40 different types of protein, although several genes are present in multiple copies).

In female meiosis the two X chromosomes pair-up to form bivalents, much like any pair of autosomal homologs. However, in male meiosis X-Y pairing is more problematic: the X and Y chromosomes differ greatly in size and sequence composition, and so X-Y recombination occurs exclusively within short terminal regions. As a result of engaging in frequent X-Y crossovers in male meiosis, the terminal regions on the X and

Y chromosomes have identical sequences (the sequences here regularly switch positions between the X and Y chromosomes). Because, therefore, they are neither X-linked nor Y-linked, these regions are known as **pseudoautosomal regions**.

The remaining regions of the sex chromosomes, containing the great majority of their sequences, do not recombine in male meiosis and so are *sex-specific regions*. Because the Y chromosome is exclusively male, the nonrecombining region on the Y chromosome never engages in recombination and is sometimes referred to as the *male-specific region*. The equivalent region on the X chromosome does, however, recombine with its partner X chromosome in female meiosis.

The Xp/Yp and Xq/Yq pseudoautosomal regions are effectively allelic sequences that contain several functional genes, as described below. Outside these regions, however, the X-specific and Y-specific components are rather different in sequence, but nevertheless exhibit multiple regions of X-Y homology that indicate a common evolutionary origin for the X and Y chromosomes. They include several gene pairs where both the X and Y sequences are functional genes (**gametologs**; see **Figure 13.16**).

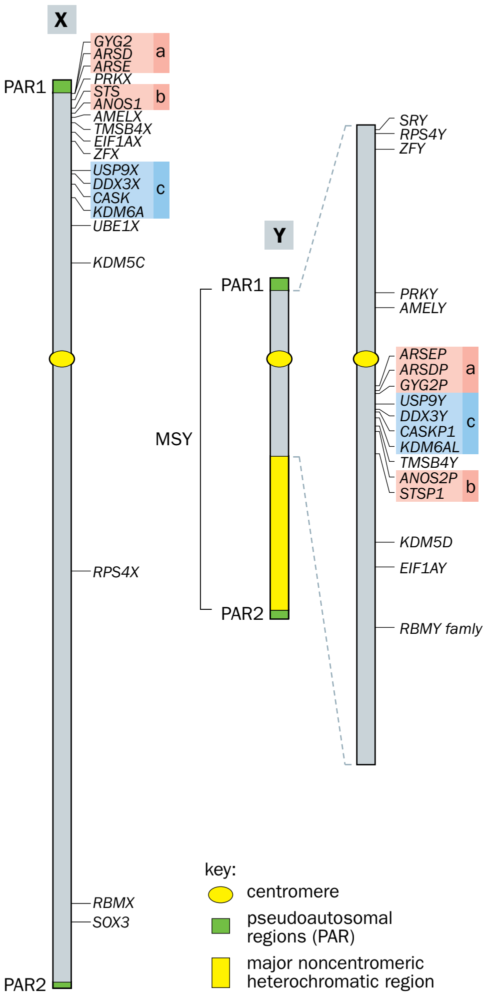


Figure 13.16 The human X and Y chromosomes show several regions of homology indicative of a common evolutionary origin. The human PAR1 pseudoautosomal regions at the tips of Xp and Yp are identical (illustrated in **Figure 13.17**), as are the PAR2 regions at the tips of Xq and Yq. The remaining nonrecombining regions show several clearly homologous X–Y gene pairs (**gametologs**), which usually have similar gene symbols. The gene symbols sometimes end in an X or Y to denote the chromosome (*PRKX/PRKY*, *AMELX/AMELY,* and so on) but some of the Y-chromosome homologs have degenerated into pseudogenes, with symbols terminating in a P, or in a P followed by a number (such as *ANOS2P*, *STSP1*). Clustering of X-Y gene pairs in blocks is evident in some regions (colored boxes labeled a–c); other gametologs are more dispersed. As a result of positive selection, the sequence of the male-determining gene, *SRY*, located very close to PAR1, is now rather different from that of its original gene partner on the X, the *SOX3* gene, which is located on distal Xq. MSY, male-specific region of the Y chromosome. (Adapted, with updated gene symbols, from Lahn BT & Page DC [1999] *Science* **286**:964–967; PMID 10542153. Reprinted with permission from the AAAS.)

The mammalian pseudoautosomal regions are not well conserved and are evolutionarily unstable

Humans have two pseudoautosomal regions. The major one, PAR1, contains at least 16 protein-coding genes and nine RNA genes, and spans 2.77 Mb of DNA just internal to the telomere sequences at the extreme tips of Xp and Yp. The PAR1 region is thought to have evolved by repeated addition of autosomal segments onto the pseudoautosomal region of one of the sex chromosomes, before being recombined onto the other sex chromosome. This region is the site of an *obligate* crossover during male meiosis that ensures correct meiotic segregation. As a result, the sex-averaged recombination frequency is 28%, about 10 times the normal recombination frequency for a region of this size. The genes in the PAR1 region of the X chromosome escape X-inactivation, and so both alleles are expressed at individual loci on the two X chromosomes in females, just like the homologous genes in the PAR1 region of the X and Y chromosomes in males (which behave effectively as alleles).

The human PAR2 region spans 330 kb just internal to the telomere sequences at the extreme tips of Xq and Yq. It emerged in the human lineage after divergence from the great apes (possibly by an L1-mediated ectopic recombination event that copied the tip of the ancestral Xq and transferred the copy onto the tip of the Yq). Containing just three functional protein-coding genes and an RNA gene plus seven pseudogenes, it has been formed very recently in the human lineage. X-Y crossovers also occur in PAR2, but they are not so frequent as in PAR1, and they are neither necessary, nor sufficient, for successful male meiosis.

Mammalian pseudoautosomal regions are not well conserved. Because the human PAR2 is an evolutionarily very recent development, orthologs of PAR2 genes in nonprimate mammals are typically autosomal. Primates have a similar PAR1 region to humans, and the PAR1 region is significantly extended in several other mammals.

However, the mouse PAR1 region is much smaller (only 0.7 Mb long), is located at the tip of the *long* arm of the X, and is quite different in sequence. A very few human PAR1 genes have well-conserved autosomal orthologs in the mouse, and some others have a diverged autosomal ortholog, but for most there is no recognizable mouse ortholog (**Figure 13.17**).

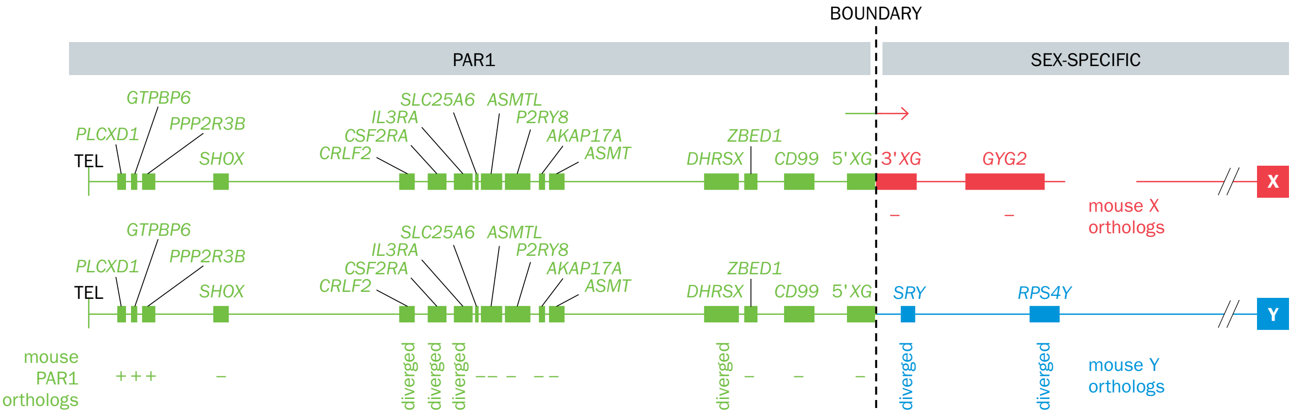


Figure 13.17 Organization and evolutionary instability of the major pseudoautosomal region. The 2.77 Mb human PAR1 region is common to the tips of Xp and Yp (chromosome coordinates = 10,001–2,781,479) and contains 16 protein-coding genes as shown here plus various RNA genes and some pseudogenes (not shown). Adjacent to PAR1 are the very long sex-specific parts of the X and Y; only the first few of the genes of the sex-specific regions are shown here. The PAR1 boundary occurs within the sixteenth protein-coding gene, the *XG* blood group gene, so that the 5′ end of *XG* is in the pseudoautosomal region while the 3′ end of *XG* is X-specific. On the Y chromosome, there is a truncated 5′ *XG* gene homolog containing just the promoter and the first few exons. The adjacent Y-specific DNA contains unrelated sequences including *SRY* and *RPS4Y*. Genes within PAR1 and in the neighboring sex-specific regions (which were part of a former pseudoautosomal region) have been poorly conserved during evolution, and mouse orthologs are often either undetectable (–) or have very diverged sequences when compared with their human counterparts. TEL, telomere.

PAR1 and neighboring regions are comparatively unstable regions. Frequent DNA exchanges result in a high incidence of gene fusions, exon duplications, and exon shuffling. Some of the genes in the adjacent part of the sex-specific regions (which were pseudoautosomal regions in the recent evolutionary past) also do not appear to have detectable orthologs in the mouse; those that do have orthologs appear to have undergone rapid evolution, as in the case of the *SRY* gene which is just 5 kb from the pseudoautosomal boundary (see **Figure 13.17**).

Mammalian sex chromosomes evolved after a sex-determining locus developed on one autosome, causing it to diverge from its homolog

The X chromosome is highly conserved in mammals, as is the Z chromosome in birds. But the X-Y pair of sex chromosomes is not homologous to the Z-W sex chromosomes of birds. Mammalian X-linked genes have autosomal orthologs in birds; avian Z-linked genes have autosomal counterparts in mammals. In each case, the different sex chromosomes (X and Y; Z and W) are believed to have evolved from a pair of autosomal homologs that diverged after one of them happened to evolve a major sex-determining locus, such as a testis-determining factor.

The major pseudoautosomal region is known to be of recent autosomal origin (see above) and, in addition, much of human Xp is also known to be of recent autosomal origin. Many X-linked genes in placental mammals are found to be also X-linked in marsupials, but genes mapping distal to Xp11.3 have orthologs on autosomes of both marsupials and monotremes. The simplest explanation is that at least one large autosomal region was translocated to the X chromosome early in the eutherian lineage that led to placental mammals.

The emergence of a sex-determining locus happens by modification of a pre-existing gene. For example, the male determinant gene *SRY* (Sox-related gene on the Y chromosome) and the X-linked *SOX3* gene are thought to be an X-Y gene pair that evolved from a common autosomal gene. *SRY* is found in both placental mammals and marsupials such as the kangaroo, but it is absent in monotremes (egg-laying mammals, such as the platypus), and the monotreme *SOX3* gene is autosomal. It is likely that *SRY* developed from a modification of what was an autosomal *SOX3* gene some time after monotremes split off from the main mammalian lineage, but before the divergence of marsupials and placental mammals (**Figure 13.18**). *SRY* has been subject to positive selection and so over millions of years its sequence has diverged very significantly from the *SOX3* sequence.

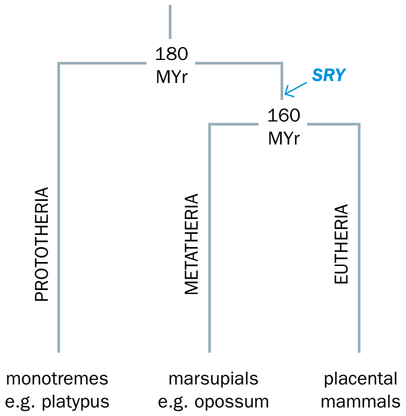


Figure 13.18 Divergence of the three major mammalian subclasses and the emergence of *SRY*. The *SRY* gene, the major male determinant in placental mammals and marsupials, is thought to have evolved by modification of an ancestral *SOX3*-like gene on what was an autosome to give a new sex-determining locus that then underwent evolutionarily rapid sequence changes. Monotremes lack an *SRY* gene and have autosomal *SOX3* genes. *SRY* appears to have developed after the divergence of the prototheria (the ancestors of modern-day monotremes) from the therian lineage, and shortly before the divergence, approximately 160 million years (Myr) ago, of the metatheria and eutheria (ancestors of marsupials and placental mammals, respectively).

Once a sex-determining locus is established, recombination needs to be suppressed in the region containing it to ensure that the sex-determining locus remains on just the one chromosome. This can be achieved through chromosomal inversions which are known to be able to suppress recombination over broad regions in mammals, and which appear to have occurred on the Y chromosome. For example, a Y-specific inversion would explain why the PAR1 boundary crosses a gene, *XG*, that is intact on the X chromosome but disrupted on the Y (see **Figure 13.17**).

In the case of the X-Y system, natural selection will favor the emergence of other alleles on the Y that confer an advantage to males, and as described below, many genes on the Y are involved in testis functions. As more alleles are selected with a male advantage, the area in which recombination is suppressed is extended (**Figure 13.19**).

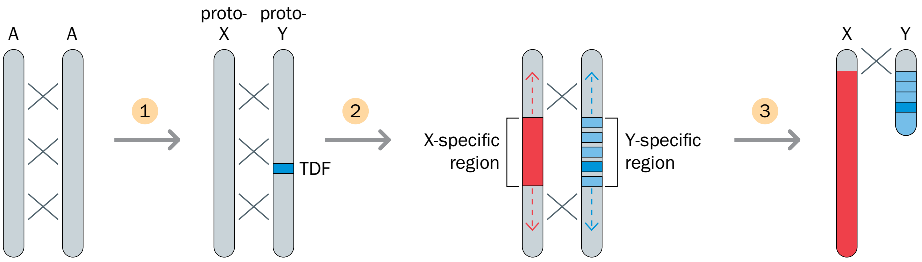


Figure 13.19 The X and Y chromosomes evolved from a pair of autosomes. The progression from a homomorphic pair of ancestral autosomes to heteromorphic sex chromosomes is shown here in three steps. (1) The ancestral autosomes (A) diverge into proto-X and proto-Y after a gene on the proto-Y chromosome is modified so as to become a sex-determining locus, such as a testis-determining factor (TDF). (2) Recombination is suppressed in the neighborhood of the male sex-determining locus to ensure that it remains on just the one chromosome, and natural selection promotes acquisition of additional male-specific alleles, causing further suppression of recombination. The sex-specific regions (where recombination is suppressed) extend outward along the chromosomes. (3) The X-specific region can engage in recombination with its partner X chromosome in female meiosis, but the Y-specific region can never recombine with another chromosome. It will gradually pick up harmful mutations causing genes to become inactive pseudogenes that are eventually lost by deletions, because there is no selective constraint to maintain inactive genes. As a result of periodic large deletions, there is progressive loss of DNA and the Y will contract in size and lose many genes. Only a few genes remain, including many that confer a male advantage. Eventually, recombination between the X and the small Y is limited to a small pseudoautosomal region (PAR).

Y chromosome degeneration

Unlike the X chromosome, which can recombine throughout its length with a partner X chromosome in female meiosis, the Y chromosome can be viewed as a mostly asexual (nonrecombining) chromosome. Population genetics predicts that a nonrecombining chromosome will acquire harmful mutations. Natural selection acts to eliminate deleterious mutations (which are lost after recombination), but the lack of recombination over most of the Y chromosome means that this option for eliminating deleterious mutations is not available. Once harmful mutations accumulate in the nonrecombining Y and cause loss of gene function, inactive pseudogenes are formed. There is no selective pressure to retain the relevant DNA segment, and so genes will be progressively lost. Normal DNA turnover mechanisms would ensure gradual contraction of the Y through a series of periodic deletions.

The X chromosome is thought to have retained something like 98% of the estimated 640 genes on the ancestral chromosome (and gained additional genes mostly by transposition of sequences from autosomes). The Y chromosome, however, has lost much of the original DNA content and the great majority of the original genes; and most Y-linked genes are predominantly expressed in the testis. Given that the human Y chromosome had the same number of genes as the X several hundred million years ago, one calculation has suggested that the Y chromosome would be driven to extinction within only 10 million years from now. There is a precedent: at least one mammal, the Ryukyu spiny rat, which is indigenous to a single island in Japan, has no Y chromosome.

The idea that contemporary mammalian Y chromosomes are merely male switches on their way to extinction has been rebuffed by recent detailed comparative genomics and comparative transcriptomics studies. The studies show that after an early evolutionary period of Y chromosome degeneration with massive gene loss, gene content on the Y has subsequently stabilized. In some cases there have been no gene losses over many millions of years, and some individual genes have been retained on the Y of placental mammal and marsupial lineages that diverged from a common ancestor 180 million years ago.

The genes retained on the mammalian Y fall into two main classes: testis-expressed genes (mostly present in multiple copies, and thought to be important in spermatogenesis); and widely expressed dosage-sensitive genes. The latter include members of X-Y gene pairs that play regulatory roles in basic cell functions, such as in translation (*EIF1AY*), transcription (*ZFY*), protein degradation (*USP9Y*), and chromatin modification (the histone demethylase genes *KDM5D* and *UTY*).

During evolution, the loss of particularly important genes from the Y has either been managed by compensatory mechanisms or actively resisted by a form of intrachromosome recombination. In the former case, genes lost from the Y have been maintained by transposition to other chromosomes. In the Japanese Ryukyu spiny rat (which lacks a Y chromosome), four important ancestral Y-chromosome genes were maintained in the genome by transposition to an autosome or to the X, and rescue of Y-linked genes via transposition is widespread among mammals. The intrachromosome recombination mechanism to resist gene loss is described in the section that follows below.

Abundant testis-expressed genes on the Y chromosome are mostly maintained by intrachromosomal gene conversion

As described above, during differentiation of the sex chromosomes the Y chromosome has lost numerous genes and much of the original DNA sequence. Many of the surviving genes are predominantly expressed in the testis, and are located in the euchromatic portion of the male-specific region on the Y chromosome. This part of the Y has three major sequence classes, as listed below and depicted in **Figure 13.20A**.

- ***X-transposed sequences***. Accounting for a total of 3.4 Mb in distal Yp, they exhibit 99% sequence identity to sequences in Xq21. They originated from a massive X-Y transposition that occurred only ~3–4 million years ago, following the divergence of the human and chimpanzee lineages. Only two protein-coding genes have been identified.
- ***X-degenerate sequences***. Thought to be surviving relics of the ancient autosomes from which the X and Y evolved, they contain several protein-coding genes (and also many pseudogenes that are more distantly related to X-linked genes). Most of the functional genes are housekeeping genes, and they make proteins closely related to those made by partner genes on the X (which all escape X-inactivation).
- ***Ampliconic segments***. Largely composed of intrachromosomal repeated DNA segments that can be large, they contain most of the protein-coding genes on the Y chromosome. Unlike the genes in the other two regions, the genes here are uniformly predominantly expressed in the testis, and they are members of gene families, such as the *DAZ* gene family (four genes, two pseudogenes), the *RBMY* gene family (five genes, one pseudogene), and the *TSPY* gene family (with many members and showing copy number variation). The amplicons include eight very large inverse repeats, or palindromes (so called because they have the same 5′→ 3′ sequence on both DNA strands—see **Figure 13.20A**). The palindromes are, however, imperfect: they consist of two paired arms with the same sequence, separated by a short stretch of unique sequence. The sequence identity between the paired arms of individual palindromes is extraordinarily high. For example, the P1 palindrome consists of two arms that are 1.45 Mb long but show 99.97% sequence identity over their lengths.

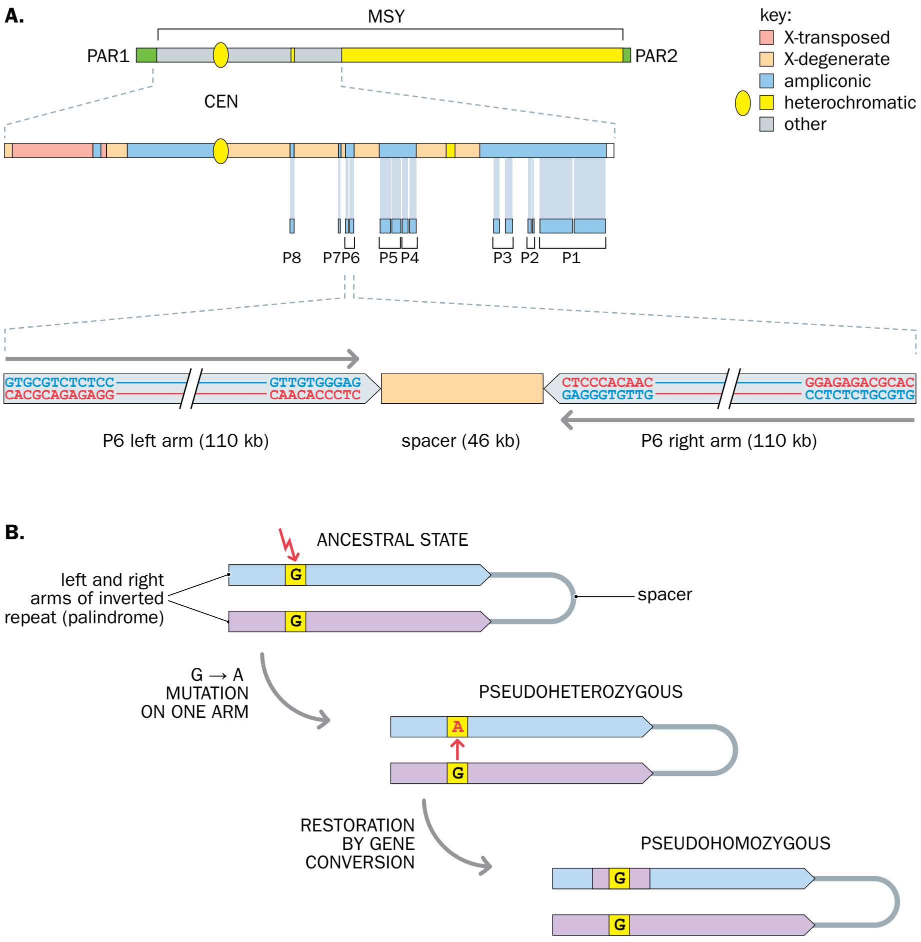


Figure 13.20 Sequence classes and intrachromatid recombination in the male-specific region of the human Y chromosome (MSY). (**A**) Internal to PAR1 and PAR2, the nonrecombining male-specific region of the Y (MSY) has a large distal heterochromatic region and a ~23 Mb region spanning Yp and proximal Yq that is mostly euchromatic. The latter region contains three major sequence classes: X-transposed sequences (transposed from Xq21 only 3–4 million years ago); X-degenerate sequences (surviving relics of ancient autosomes from which the X and Y co-evolved); and ampliconic segments (sequences that show marked similarity to sequences elsewhere on the MSY). The ampliconic sequences include eight inverted repeats or palindromes (P1 to P8) that have arms with essentially identical sequences separated by shorter unique sequences. The structure of the P6 palindrome is shown as an example. CEN, centromere. (**B**) Gene conversion between the arms of palindromes can suppress mutations. The arms of the individual palindromes maintain their extremely high sequence identity by gene conversion. After one arm picks up a mutation to make the sequences different (pseudoheterozygous), gene conversion can replace a segment of DNA on one arm with a copy of the equivalent sequence on the other arm to restore the pseudohomozygous state. The most frequent outcome is that a segment of DNA containing the new mutation is replaced by a copy of the original sequence from the other arm. (A, adapted from Skaletsky H *et al*. [2003] *Nature* **423**: 825–837; PMID 12815422. With permission from Springer Nature. Copyright © 2003.)

The very high degree of sequence identity between palindrome arms in ampliconic segments does not reflect very recent duplication (the palindromes mostly originated before the chimpanzee–human split 5 million years ago). Instead, the palindrome arms appear to have undergone concerted evolution by a mechanism that resembles **gene conversion**, a recombination-associated mechanism involving nonreciprocal transfer of DNA sequence from a donor DNA sequence to a highly homologous acceptor DNA sequence. (In gene conversion, a copy is made of the donor DNA sequence that is then used to replace the original acceptor sequence; for detail on the mechanism, see Section 15.3 and **Figures 15.22** and **15.23**.) In the context here, the outcome is that one arm of a palindrome acts as a donor sequence and the other arm as an acceptor so that the sequence of one arm is periodically replaced by a perfect copy of the sequence of the other arm (**Figure 13.20B**). Essentially, this is a type of intrachromosomal recombination and may act so as to preserve the testis-expressed genes from the extinction that they might otherwise have faced in the absence of recombination.

X-chromosome inactivation developed in response to gene depletion from the Y chromosome

As a result of large-scale destruction of Y chromosome sequences, there is an imbalance in copy number of X-linked genes between the two sexes. Females have two gene copies but males mostly have just one copy of X-linked genes because outside the pseudoautosomal regions, there is usually no partner gene on the Y (but see **Figure 13.16** for some exceptions). For many genes on the X that lack a functional gene partner on the Y, dosage needs to be tightly controlled (X-linked gene products interact with those made by autosomal genes where dosage is maintained). To achieve this, a form of gene dosage compensation evolved called **X-chromosome inactivation** whereby a single X chromosome is selected to be inactivated in female cells (see Section 10.4 for the mechanism).

About 85% of the genes on the inactivated human X chromosome are silenced. Genes on the X that have a functional homolog on the Y escape X-inactivation. Thus, all PAR1 genes seem to escape X-inactivation. PAR2 genes are different: the distally located *IL9R* gene escapes inactivation and there are two functional copies in males and females. However, the two proximal genes, *VAMP7* and *SPRY3*, are subject to X-inactivation in females, and also when residing on the Y chromosome they are epigenetically silenced (by methylation); as a result, males and females each have one functioning copy of both these genes.

Other genes that escape X-inactivation tend to be concentrated in clusters mainly on Xp, and many of them appear to derive from recent autosomal additions to the sex chromosomes. Genes that do not have a functional homolog on the Y but yet escape inactivation may be ones where 2:1 dosage differences are not a problem.

13.4 REGULATORY SEQUENCE EVOLUTION AND TRANSPOSON ORIGINS OF FUNCTIONAL SEQUENCES

If increased gene number is not the principal explanation for organism complexity, what is? Well, alternative splicing can be discounted. On average, human protein-coding genes make three to four different isoforms as a result of alternative splicing, but the capacity for extending the proteome is not vastly greater than in other metazoans. We have seen in Section 13.2 that lineage-specific gene expansion and loss is important. But the greatest single contributor to organism complexity, and the principal explanation for what makes us different from other animals, is expected to lie in gene regulation.

Although humans and mice have almost identical sets of protein-coding genes (sharing 99% of them, with generally high sequence conservation), the sequences that control gene expression are often significantly less conserved and generally evolve more rapidly. Transposable elements, which also evolve rapidly and can show striking lineage-specific expansion and loss, are an important source of novel regulatory sequences, and evolve to produce other functional sequences, too.

The noncoding DNA of complex metazoans is dominated by sequences arising from transposable elements and is evolutionarily unstable

The genomes of metazoans evolved by a process of net DNA gain, accumulating additional DNA by duplications and insertions. However, genome evolution does not simply mean an inexorable increase in size: additions of DNA sequence to the genome are offset by periodic deletions that can involve significant amounts of DNA. For example, after the lineages leading to modern humans and mice diverged from a common ancestor about 90 million years ago, there was a net loss of DNA sequences from the mouse lineages. As a result, the mouse genome became significantly smaller than the human genome (which has retained much the same genome size as in the common ancestor).

In total, about 22% of the DNA in the genome of the last common ancestor of humans and mice appears to have been deleted in the lineage leading to modern humans. However, that loss has been offset by gain of a similar amount of DNA sequences that originally derived from transposable elements, notably retrotransposons and retroviruses (**Figure 13.21**). About 40% of the ancestral genome sequences have been lost from the *Mus musculus* lineage and replaced by a smaller amount of DNA copies originating from transposable elements. The DNA that has been retained from the ancestor includes coding DNA plus some ancestral transposon repeats, but other sequences originating from the common ancestor have diverged markedly in sequence.

Standard methods clearly show that close to one-half of the sequences in the human genome are derived from sequences that were able to transpose (but only a very few of these sequences are currently capable of transposition). Some additional sequences originating in the same way are so diverged that they are not readily detected, and more-sensitive methods suggest that more than two-thirds of our genome is derived from these elements (see the legend to **Figure 13.21**). Much of the “unique” sequence in our genome, therefore, is not unique but appears to be composed of extensively diverged repeats.

As described below, having large amounts of noncoding DNA in the genome can be useful over evolutionary time periods because noncoding DNA, and especially transposable elements, can give rise to all types of new functional DNA sequences: exons, genes, regulatory sequences, and long noncoding RNAs. Applying the term “junk DNA” to transposon repeats and endogenous retroviruses may not be so inaccurate after all: “junk” is how we describe any items for which we have no immediate use, but keep for long periods in the hope they may be useful one day.

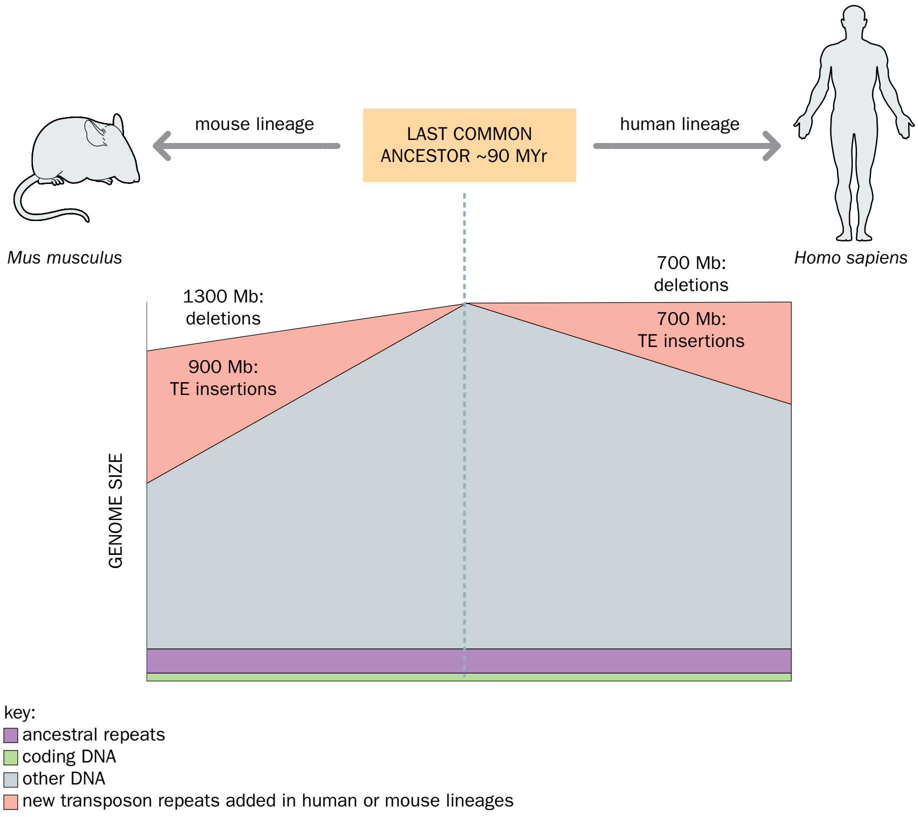


Figure 13.21 Rapid turnover of noncoding DNA sequences in human and rodent lineages since the last common ancestor, approximately 90 million years (Myr) ago. The 3200 Mb human genome is expected to be much the same size as that of the last common ancestor. It has been formed by loss of 700 Mb of DNA from the ancestral genome (through intermittent deletions of sequences present in the last common ancestor) and a compensatory gain of 700 Mb of novel transposon repeats after periodic duplicative insertions by transposable elements (TE). The 2800 Mb mouse genome is thought to have lost, through deletion, 1300 Mb of the original genome in the last common ancestor and gained 900 Mb of TE-derived sequence. Both species have retained about 173 Mb of ancestral transposon repeat sequence and 33–34 Mb of coding DNA present in the last common ancestor. For simplicity, the TE insertions are shown here as being acquired evenly over time, but the human lineage acquired novel TEs at an elevated rate approximately 40 million years ago. Although sequence elements derived from TEs can readily be identified as accounting for nearly 50% of the current human genome, more-sensitive methods suggest that the proportion could be as high as two-thirds of the genome (see de Koning *et al*. [2011], PMID 22144907). (Adapted from Ponting CP *et al*. [2011] *Annu Rev Genomics Hum Genet* **12**:275–299; PMID 21721940. With permission from Annual Reviews. Permission conveyed through Copyright Clearance Center, Inc.)

Regulatory sequence divergence as a way of explaining morphological divergence

How can morphological divergence between very closely related species be explained when their coding sequences are so similar? Genome-wide comparisons reveal an average 99% identity between orthologous human and chimpanzee protein sequences, for example. One idea, first proposed in 1975 by Mary-Clair King and Allan Wilson, was that phenotype evolution may be more attributable to changing gene regulation than changing protein sequences. Orthologous genes in extremely closely related species that have diverged from a common ancestor very recently show much greater divergence in expression than in coding DNA sequence. And transcription factor binding sites generally evolve rapidly—in the 90 million years following human–mouse divergence, over one-third of human transcription factor binding sites are nonfunctional in rodents, and vice versa.

Analyses in various organisms have confirmed the importance of *cis*-regulatory elements (CREs) in dictating morphological evolution. In many case studies, precise changes in CREs that control single genes have been shown to underlie morphological differences. In each case, the divergence in traits has arisen between closely related populations or species, and CRE evolution is considered sufficient to account for the changes in gene regulation within and between closely related species. The cumulative data led Sean Carroll to propose a new genetic theory of morphological evolution in 2008.

Mosaic pleiotropism and modular CREs

Proteins that work as key developmental regulators typically exhibit *mosaic pleiotropism*: they are pleiotropic (serve multiple roles), but they also function independently in different cell types, germ layers, body parts, and developmental stages. Because the same protein can shape the development of many different body parts, the coding sequences of regulatory proteins like this are under very considerable evolutionary constraint. They are, therefore, extremely highly conserved, and sometimes their functions can be maintained after being artificially substituted by very distantly related orthologs. As shown in Figure 13.22, for example, the human *OTX2* gene, an ortholog of the *Drosophila apterous* gene can regulate the formation of fly wings. And it is not just the master gene regulators that are highly conserved, but also the proteins that they interact with to carry out their functions; in some cases the conservation has been maintained over as much as the roughly 800 million years of evolution that separate humans and flies.

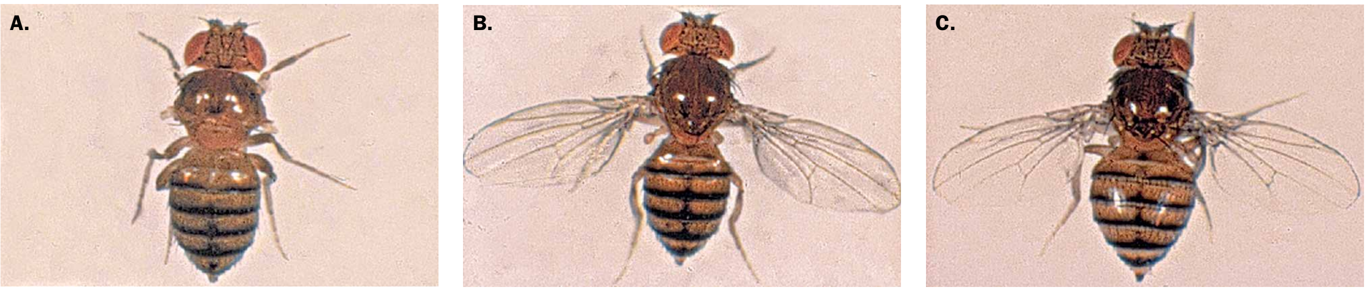


Figure 13.22 Humans have a gene that makes flies grow wings. (**A**) Flies with a mutation in the *apterous* gene lack wings. This mutation can be corrected either (**B**) by the wild-type fly gene or (**C**) by the human *OTX2* (formerly *hLHx2*) gene. (Adapted from Rincón-Limas DE *et al.* [1999] *Proc Natl Acad Sci USA* **96**:2165–2170; PMID 10051612. With permission from the National Academy of Sciences. Copyright [1999] National Academy of Sciences, USA.)

Although there was a rapid expansion in the number of fundamentally different genes encoding transcription factors, signaling molecules, and receptors at the beginning of metazoan evolution, there has been comparatively little expansion since the period just before the origin of bilaterians, animals that are bilaterally symmetrical. For example, 11 out of the 12 vertebrate *Wnt* gene families, which encode proteins that regulate cell–cell interactions in embryogenesis, have counterparts in cnidarians such as jellyfish and, remarkably, six of the major bilaterian signaling pathways are represented in sponges that diverged from other metazoans at the very beginning of metazoan evolution. The Hox genes, which specify the anterior–posterior axis and segment identity in embryogenesis, have also undergone very few changes over many hundreds of millions of years. Coelacanths, cartilaginous fish that are very distantly related to us, have much the same classical Hox genes as we do. More accurately, they have four more Hox genes than us because of loss of Hox genes in vertebrate lineages leading to mammals (**Figure 13.23**).

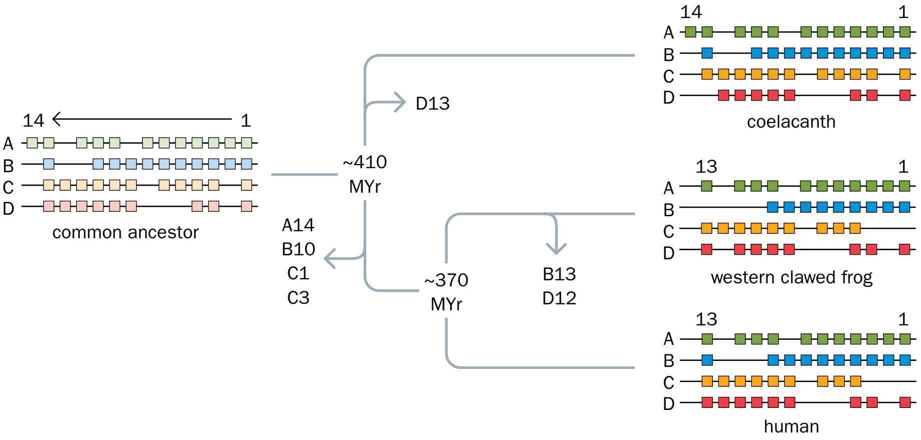


Figure 13.23 Gene duplication has not occurred over hundreds of millions of years of Hox cluster evolution. The coelacanth is a cartilaginous fish that descended from the earliest diverging lineage of jawed vertebrates. Coelacanths, humans, and frogs have remarkably similar Hox gene organizations and there does not appear to have been any gene duplication in the 400 million years (MYr) or so since they diverged from a last common ancestor. Rather, there have been instances of occasional Hox gene loss during this time. For simplicity, only the classical Hox genes are shown within the Hox cluster. (Adapted from Hoegg S & Meyer A [2005] *Trends Genet* **21**:421–424; PMID 15967537. With permission from Elsevier.)

Thus, while many other protein families were diversifying over the last several hundred million years, there has been comparatively little diversification of key proteins involved in developmental regulation, possibly because of constraints imposed by gene dosage. Instead, expansion of gene function without gene duplication has often been achieved by altering how the gene is regulated to shape several entirely different traits. By appropriately regulating gene expression, the same gene product can be sent to different, but specific, tissues (*heterotopy*) and at different developmental stages (*heterochrony*) where, according to its environment, it can interact with different sets of molecules. For example, the *Drosophila* Decapentaplegic signaling protein shapes embryonic dorsoventral axis polarity, epidermal patterning, gut morphogenesis, and the patterning of wings, legs, and other appendages. To achieve this type of highly specific regulation, pleiotropic gene regulators typically have several large, modular CREs that *independently* regulate a specific pattern of expression (see **Figure 13.24**).

Independent regulation of a developmental regulator by different CREs allows multifunctionality, and harmful mutations in one CRE will not affect the functions of other CREs that regulate the expression of the same protein or the protein itself. Each CRE consists of a series of modules with short binding sites (often around 6–12 nucleotides) for *trans*-acting regulators, notably transcription factors. In some regulatory elements, individual binding sequences recognized by different transcription factors may be very tightly packed; in others, such as diffuse enhancers, they may be much more spread out along the genomic DNA.

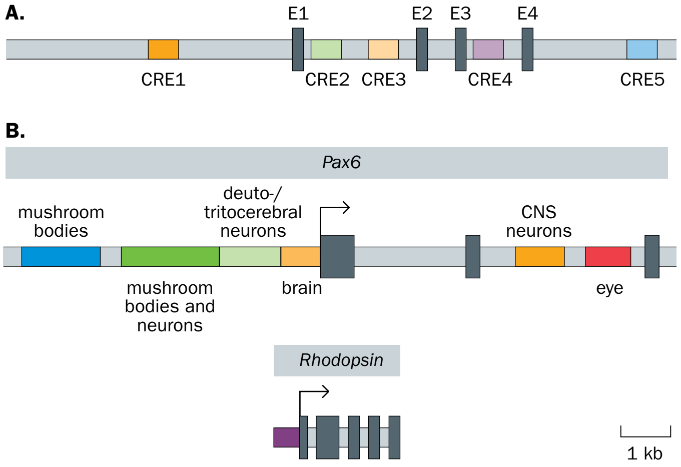
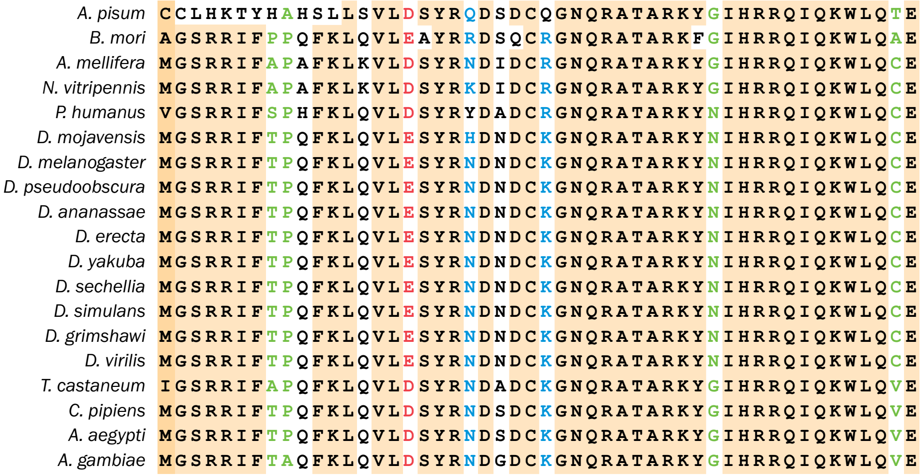


Figure 13.24 Complex *cis*-regulatory regions in pleiotropic developmental regulators. (**A**) In this generalized scheme, a gene with four exons (E1 to E4) is controlled by five *cis*-regulatory elements (CRE1 to CRE5). Each regulatory element consists of a variety of modules that contain binding sites for *trans*-acting proteins. Clustered binding sites have extended sequence conservation, which makes them easier to detect by comparative genomics. (**B**) The *Drosophila Pax6* gene, also called *eyeless* (*ey*), has three exons (black bars) and encodes a transcription factor that is expressed in specific parts of the developing brain, central nervous system (CNS), and eyes. Expression is regulated by six CREs (colored blocks), most being 1 kb in length or longer. The *Rhodopsin* gene at the bottom is one of the target genes of *Pax6/ey*. It has a single function that involves expression in the photoreceptor cells of the eye, and has a single CRE, as is commonly found in genes that are restricted in expression. (Adapted from Carroll SB [2008] *Cell* **134**:25–36; PMID 18614008. With permission from Elsevier.)

Novel functional sequences—exons, regulatory elements, lncRNAs, sometimes even genes—often originate from transposable elements

Complex metazoans have complex genetic regulatory networks, where pleiotropic transcription factors control hundreds of target genes. *Cis*-regulatory elements are at the heart of these networks and although some such sequences seem remarkably conserved within a taxon such as a class or an order, the sequence conservation often does not extend over broader evolutionary taxa. Note, for example, the restricted range of evolutionary conservation of many enhancers that bind heart-specific transcription factors (see **Figure 13.4B**). New lineage-specific regulatory elements evolve from time to time and co-evolution may result in *trans*-acting factors that are lineage-specific (**Figure 13.25**).



```
A. pisum         CCLHKTYHAHSLLSVLDSYRQDSDCQGNQRATARKYGIHRRQIQKWLQTE
B. mori          AGSRRIFPPQFKLQVLEAYRRDSQCRGNQRATARKFGIHRRQIQKWLQAE
A. mellifera     MGSRRIFAPAFKLKVLDSYRNDIDCRGNQRATARKYGIHRRQIQKWLQCE
N. vitripennis   MGSRRIFAPAFKLKVLDSYRKDIDCRGNQRATARKYGIHRRQIQKWLQCE
P. humanus       VGSRRIFSPHFKLQVLDSYRYDADCRGNQRATARKYNIHRRQIQKWLQCE
D. mojavensis    MGSRRIFTPQFKLQVLESYRHDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. melanogaster  MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. pseudoobscura MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. ananassae     MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. erecta        MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. yakuba        MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. sechellia     MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. simulans      MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. grimshawi     MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
D. virilis       MGSRRIFTPQFKLQVLESYRNDNDCKGNQRATARKYNIHRRQIQKWLQCE
T. castaneum     IGSRRIFAPQFKLQVLDSYRNDADCKGNQRATARKYGIHRRQIQKWLQVE
C. pipiens       MGSRRIFTPQFKLQVLDSYRNDSDCKGNQRATARKYGIHRRQIQKWLQVE
A. aegypti       MGSRRIFTPQFKLQVLDSYRNDSDCKGNQRATARKYGIHRRQIQKWLQVE
A. gambiae       MGSRRIFTAQFKLQVLDSYRNDGDCKGNQRATARKYGIHRRQIQKWLQVE
```

Figure 13.25 The DNA-binding domain of the brinker transcriptional repressor is very highly conserved between different inspect species but restricted to insect lineages. Brinker is a transcriptional repressor that is important in dorsoventral patterning in insects. The alignment shows the conserved core from a selection of insect species, including: the pea aphid, *Acyrthosiphon pisum*, the silkworm *Bombyx mori*, the honeybee *Apis mellifera*, the wasp *Nasonia vitripennis*, the head louse *Pediculus humanus*, 10 *Drosophila* species, the flour beetle *Tribolium castaneum*, and three mosquito species, *Culex pipiens*, *Aedes aegypti*, and *Anopheles gambiae*. The very strong sequence conservation within insects supports the functional importance of this sequence to insect lineages. However, sequences with significant similarity cannot be identified in other organisms. (Adapted from Copley RR [2008] *Philos Trans R Soc Lond B Biol Sci* **363**:1453–1461; PMID 18192189. With permission from the Royal Society. Permission conveyed through Copyright Clearance Center, Inc.)

If CRE evolution is central to morphological evolution and phenotype divergence, how do CREs evolve? Existing CREs can be modified in different ways. Mutations in an existing CRE can result in new binding sites for transcription factors, for example, or loss or modification of existing binding sites. But new CREs may evolve by random mutation and also as a result of sequence insertion events involving transposable elements.

Transposable elements have been viewed as genome parasites that owe their survival to their ability to replicate faster than the host that carries them; mechanisms such as targeted RNA interference, DNA methylation, and histone modification (such as by KRAB-ZNF protein repression) are needed to limit their spread within germ-line cells. Transposons and transposon repeats, mostly derived from retrotransposons and endogenous retroviruses, account for maybe up to two-thirds of our genome (if we take into account very highly diverged sequences). By integrating into new locations in the DNA, they can influence expression of nearby genes by a variety of possible mechanisms (**Figure 13.26**).

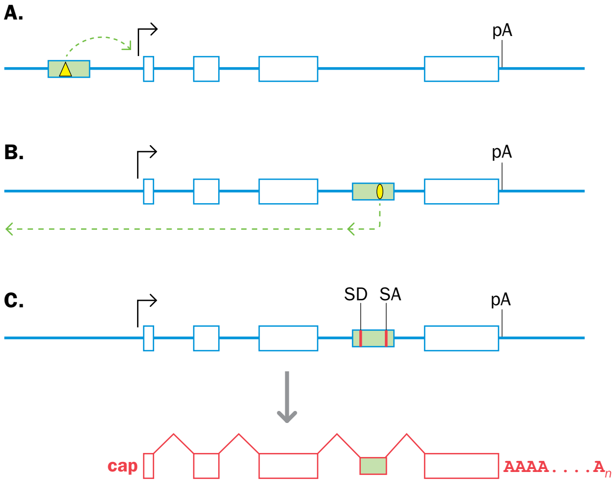


Figure 13.26 Transposable elements can influence gene expression in several ways. (**A**) When inserted upstream of a gene, a transposable element (green box) may provide new regulatory sequences, such as an enhancer sequence (yellow triangle) that alters the expression of a neighboring gene. (**B**) A transposable element inserted within an intron may contain promoter elements (yellow oval) that drive antisense transcription, sometimes making a long noncoding RNA. The antisense transcript may interfere with sense transcription and potentially regulate transcription. (**C**) An inserted transposable element may carry cryptic splice donor (SD) and splice acceptor (SA) sequences that allow it to be incorporated as an alternative exon (*exonization*). For various other possibilities, see Feschotte C (2008) *Nat Rev Genet* **9**:397–405; PMID 18368054. See **Figure 13.27** for specific examples of (**A**) and (**C**). pA, polyadenylation site.

By disrupting the normal patterns of gene expression, transposable elements are known to cause disease. In addition to being a potential threat at the level of the genome and the individual, however, transposable elements can also be beneficial. When the genomes of 29 mammals were compared, 200,000 conserved DNA elements were identified to have evolved under strong purifying selection and to have derived from ancient transposable elements. Many of these elements are located close to genes encoding developmental regulators, and at least some have been strongly conserved over hundreds of millions of years.

Evolutionary value and exaptation of transposable elements

The more common transposable element repeat families, such as Alu repeats, are present at very high copy numbers and may facilitate unequal pairing of chromosomes and chromatids to generate gene duplication. In addition, as described in **Figure 13.13**, transposable elements are thought to contribute to exon shuffling by transporting copies of exon sequences from one gene to another. Transposable elements can also be functionally important in other ways, notably through **exaptation,** the process where they donate sequences that then give the host genome a new function. One extraordinary exaptation is when new genes originate in large part from transposon sequence. Such **neogenes** include *TERT* (telomerase reverse transcriptase), *CENPB* (encoding a centromere protein), and the recombination activating genes *RAG1* and *RAG2*. Genes can also arise *de novo* by other means, too (**Box 13.4**).

BOX 13.4 MOST NEW GENES ARISE BY COPYING OLD GENES OR FUNCTIONAL GENE SEGMENTS, BUT SOME ARISE *DE NOVO* FROM NONCODING DNA

During evolution, new genes usually arise by duplication of a segment of DNA containing a gene, or by making a DNA copy of an RNA transcript that then integrates into germ-line DNA; the duplicated gene copy may acquire mutations (distinguishing it from the parent gene) and be retained because it has acquired some property that is advantageous. The initiating events happen in a single germ-cell genome but the new gene copy spreads through the population and becomes fixed.

Occasionally, a new gene evolves by combining segments from different pre-existing genes. Segments of two unrelated genes can be directly joined together to form a chimeric gene that may have quite novel characteristics. That happens quite frequently by chromosome translocation in cancer cells, but is rare in germ-line cells. An early example involved retrotransposition in *Drosophila*: a cDNA copy of the *Adh* alcohol dehydrogenase mRNA integrated into the middle of the *yande* gene to form the chimeric *Jingwei* gene that has been retained in some African *Drosophila* species. Various chimeric germ-line genes have been found in other species, too, including genes that are specific for hominoid lineages, as described in Chapter 14.

Sometimes new protein structures arise from coding sequences that evolved partly or wholly from noncoding DNA components. An early example identified in vertebrates was an entirely novel protein, the AFGP antifreeze glycoprotein that evolved to protect species of fish in very cold Antarctic waters (as low as −2° C because of dissolved salts). A short sequence spanning intron 1 and exon 2 of a trypsinogen-like protease gene, ACAGCGGCA, was amplified many times to give a novel coding DNA. After translation and processing, AFGP proteins were produced consisting of a very simple repeating Thr-Ala-Ala sequence that was effective in binding ice crystals. Natural selection drove expansion of the nine-nucleotide sequence (possibly by replication slippage followed by unequal crossover or unequal sister chromatid exchange), and the downstream exons of the original protease gene underwent degeneration; see Cheng & Chen (1999) (PMID 10519545).

Other genes have arisen *de novo* entirely from noncoding DNA. They include the mouse-specific *Pldi* gene, the first proven example of a *de novo* RNA gene. It emerged about 3 million years ago by cryptic splice site mutation to generate an entirely novel testis-expressed transcript that may be involved in chromatin organization. Knocking out the gene results in a lowered sperm-cell mobility. The *CLLU1* gene is an example of a human-specific gene created from noncoding DNA. Homologs are present in the genomes of other primates, but do not produce a protein. Activation of the *CLLU1* gene occurred by deletion of an ancestral single nucleotide, creating a large open reading frame that is translated; see Knowles & McLysaght (2009) (PMID 19726446).

More commonly, transposable element sequences have contributed smaller gene components. One common exaptation is donation of sequences with appropriate splice sites that can be included as novel exons (**exonization**—see Figure 13.26C). Such exons have contributed to both functional noncoding RNA as well as to various proteins. Approximately 2000 exons in the human genome are derived from Alu repeats, and 12% of them have been incorporated into coding sequences without introducing a premature termination codon in the normal reading frame. Scanning human peptide databases (such as the large-scale mass spectrometry PRIDE dataset) suggests that about one-third of the putative coding Alu exons are translated.

Transposable element sequences that contain binding sites for *trans*-acting factors have also been exapted to provide novel *cis*-acting regulatory elements. As an example, the LF-SINE family of retrotransposons are known to have contributed sequences to generate both a novel enhancer and a novel exon. The starting point for this discovery was uc.338, a 223-bp-long ultraconserved element located at 12q13 whose sequence is 100% conserved in human, mouse, and rat. The uc.338 sequence lies within the *PCBP2* gene that encodes an RNA-binding regulatory protein, and spans an alternatively spliced exon encoding part of the PCBP2 protein. Comparative genomics also revealed that the sequence of uc.338 exhibited a high degree of sequence identity to members of the LF-SINE retrotransposon family in the coelacanth, a fish that has descended from one of the most evolutionarily ancient lineages of jawed fishes (Figure 13.27).

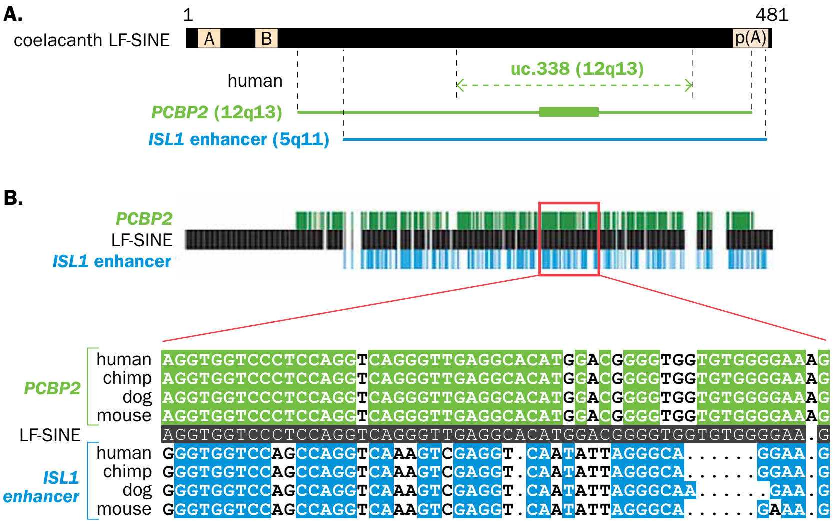

A.
1
481
coelacanth LF-SINE
A
B
p(A)
human
uc.338 (12q13)
PCBP2 (12q13)
ISL1 enhancer (5q11)
B.
PCBP2
LF-SINE
ISL1 enhancer
PCBP2
human AGGTGGTCCCTCCAGGTCAGGGTTGAGGCACATGGACGGGGTGGTGTGGGGAAAG
chimp AGGTGGTCCCTCCAGGTCAGGGTTGAGGCACATGGACGGGGTGGTGTGGGGAAAG
dog AGGTGGTCCCTCCAGGTCAGGGTTGAGGCACATGGACGGGGTGGTGTGGGGAAAG
mouse AGGTGGTCCCTCCAGGTCAGGGTTGAGGCACATGGACGGGGTGGTGTGGGGAAAG
LF-SINE AGGTGGTCCCTCCAGGTCAGGGTTGAGGCACATGGACGGGGTGGTGTGGGGAA.G
ISL1 enhancer
human GGGTGGTCCAGCCAGGTCAAAGTCGAGGT.CAATATTAGGGCA......GGAA.G
chimp GGGTGGTCCAGCCAGGTCAAAGTCGAGGT.CAATATTAGGGCA......GGAA.G
dog GGGTGGTCCAGCCAGGTCAAAGTCGAGGT.CAATATTAGGGCAA......GAA.G
mouse GGGTGGTCCAGCCAGGTCAAAGTCGAGGT.CAATATTAGGGCA......GAAA.G

Figure 13.27 A transposable element can give rise to a novel enhancer and a novel exon. (**A**) The consensus 481 bp coelacanth LF-SINE repeat has tRNA-like A- and B-box sequences and ends in a poly(A) tail [p(A)]. It shows 80% sequence identity over a 360 bp region that spans the human uc.338 ultraconserved element, and significant homology (shown by colored bars) to components of the *PCBP2* gene (including an alternatively spliced exon) and an enhancer sequence that regulates the *ISL1* neurodevelopmental gene. (**B**) Alignment of the coelacanth LF-SINE consensus with some mammalian sequences (human, chimp, dog, and mouse) representing the *PCBP2* exonized element and a proximal enhancer that regulates *ISL1*, with detailed sequence homology for a conserved region spanning nucleotides 296–349 (enclosed within red box at top) of the LF-SINE consensus sequence. Not shown here are related *ISL* proximal enhancer sequences from other vertebrates including chick, frog, and so on. (Adapted from Bejerano G *et al.* [2006] *Science* **441**:87–90; PMID 16625209. With permission from Springer Nature. Copyright © 2006.)

There are about 100,000 LF-SINEs in the coelacanth but, like other SINE families, evolutionary conservation of LF-SINEs is limited. Nevertheless, a few diverged nontransposing copies are found in other vertebrates, including 244 human copies in addition to the sequence within *PCBP2*. Most have evolved more slowly than expected if they were neutral sequences, suggesting that they could have been exapted into cellular roles that benefited the host and then became subject to purifying selection. They also appear to be found preferentially near genes involved in transcriptional regulation and neuronal development, suggesting that they may function as enhancers. Testing of one such candidate showed that it acted as an enhancer that regulated the *ISL1* gene, which encodes a transcription factor needed for motor neuron differentiation. A conserved sequence in this enhancer is clearly related to the LF-SINE consensus sequence (see Figure 13.27).

The LF-SINE family members had been active in transposition in a common ancestor hundreds of millions of years ago (and long before the appearance of mammals), and so the alternatively spliced *PCBP2* exon and the proximal *ISL1* enhancer sequence were generated by evolutionarily ancient retrotransposition events.

13.5 PHYLOGENETICS AND OUR PLACE IN THE TREE OF LIFE

Phylogenetics seeks to assess evolutionary relationships (**phylogenies**) between organisms or populations. Classical phylogenetic approaches have been based on anatomical and morphological features of living organisms and information gleaned from the fossil record. Using such approaches it has been possible to classify organisms at different hierarchical levels known as taxonomic units or *taxa* (see **Figure 13.28** and the Tree of Life Website at http://tolweb.org/tree/phylogeny.html).

More recently, molecular genetics has made a huge contribution to phylogenetics. Here, we first consider the basis of molecular phylogenetics and how molecular genetics is transforming our understanding of the evolutionary relationships of metazoans. Then we sketch out where humans fit in within the Tree of Life. Our closest relatives have long been known to be the great apes, and in Chapter 14 we will examine in some detail how we are related to the great apes, and the molecular basis of what makes us unique.

Molecular phylogenetics uses sequence alignments to construct evolutionary trees

Molecular phylogenetics uses nucleic acid or protein comparisons as the basis of classifying evolutionary relationships between organisms. Until recently, sequences from at most a very few genomic locations were used. Phylogenies based on very small sequence datasets can, however, be misleading and give differing results. Modern molecular phylogenetics can take advantage of whole-genome comparisons and the discipline is transforming into *phylogenomics*.

To construct an evolutionary tree it is necessary to compare sequences from different species; nucleic acid sequences are often used, but if the sequences are from distantly related organisms they may be difficult to align and, where available, protein sequences are used instead. If two or more sequences show a sufficient degree of similarity (*sequence homology*) they can be assumed to be derived from a common ancestral sequence. Sequence alignments can then be used to derive quantitative scores describing the extent of relationship between the sequences.

Comparing sequences of equal fixed length is usually straightforward if there is a reasonably high sequence homology. Often, however, the nucleic acid sequences that are compared will have previously undergone deletions or insertions, and so rigorous mathematical approaches to sequence alignment are needed.

Once the sequences have been aligned, **evolutionary trees** can be constructed (**Figure 13.29**). They are most commonly represented as diagrams that use combinations of lines (*branches*) and nodes. The different organisms (sequences) under comparison are placed at external nodes, connected via branches to interior nodes (intersections between the branches) that represent ancestral forms for two or more organisms.

A *rooted tree* (or *cladogram*) infers the existence of a common ancestor (represented as the trunk or *root* of the tree) and indicates the direction of the evolutionary process (**Figure 13.29B**). The root of the tree may be determined by comparing sequences against an *outgroup* sequence (one that is clearly but distantly related to the sequences under study; for example, a marsupial mammal can serve as an outgroup when making evolutionary trees of placental mammals). An *unrooted tree* (**Figure 13.29A**) does not infer a common ancestor and shows only the evolutionary relationships between the organisms. Note that the number of possible rooted trees is usually much higher than the number of unrooted trees.

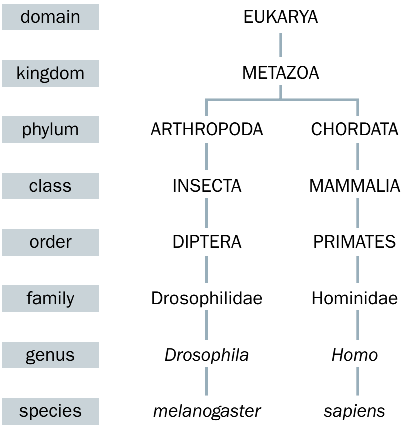


Figure 13.28 Eight major taxonomic ranks are traditionally used to classify living organisms. A taxonomic unit or *taxon* (plural, taxa) is a name that either identifies a specific type of organism (*species*) or a group of related organisms (*genus*, *family*, *order*, and so on). There are seven major taxonomic ranks, ranging from domain (the most basic division of life, comprising bacteria, archaea, and eukarya) to genus. More detailed taxonomy information can be accessed at the US National Center for Biotechnology Information's Taxonomy Browser at http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi.

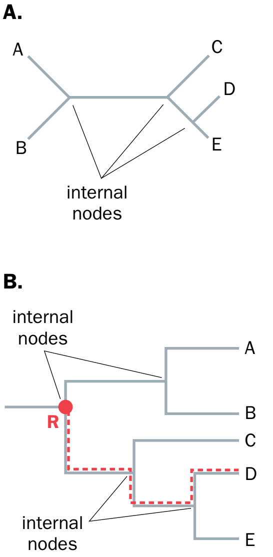


Figure 13.29 Unrooted and rooted evolutionary trees. (**A**) This unrooted tree has five external nodes (A, B, C, D, and E) that are linked by lines (branches) that intersect at internal nodes. Such a tree specifies only the relationships between the sequences (or other characters) under study and does not define the evolutionary path. (**B**) In this rooted tree (cladogram), from the internal node known as the root (R), there is a unique evolutionary path that leads to any other node, such as the path to D (dashed red line). With both trees, a close evolutionary relationship is indicated when the cumulative length of the lines that join two external nodes is short. Thus, both trees suggest that D and E are much more closely related to each other than either is to A.

Evolutionary trees can be constructed in different ways and their reliability is tested by statistical methods

Different approaches are used for constructing evolutionary trees, but many employ a *distance matrix* method. The first step calculates the evolutionary distance between all pairs of sequences in the dataset and arranges them in a table (matrix). This may be expressed as the number of nucleotide/amino acid differences between the two sequences, or the number of nucleotide/amino acid differences per nucleotide/amino acid site (see **Figure 13.30**).

A.

```
     human MSLTKTERTIIVSMWAKISTQADTIGTETIERLFLSHPQTKTYFPHFDLH
     chimp .......G..........................................
     mouse ...M.N..A..M...E.MAA..EP..........C.Y.............
       rat ...M.N..A..M...D.MAPH.EP..........S.Y.............
     chick .A..QA.KAAVTTI...VA..IES..L.S.....A.Y...........VS
zebrafish ....AKDKAV.KGF.G..AS...S..Q.AMG.MLTVY....I..A.WPD.
```

B.

	human	chimp	mouse	rat	chick	zebrafish
human		0.02	0.24	0.26	0.42	0.54
chimp			0.26	0.28	0.42	0.54
mouse				0.08	0.40	0.54
rat					0.42	0.54
chick						0.56

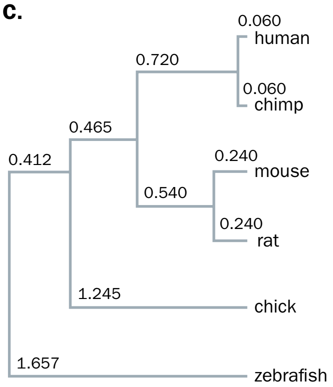


Figure 13.30 Constructing a phylogenetic tree using the UPGMA hierarchical clustering method. In this example, the N-terminal 50 amino acids of zeta globins from human and six other vertebrate species were aligned, and a rooted tree was constructed using the UPGMA (unweighted pair group method with arithmetic mean) method that first calculates a *distance matrix*. (**A**) The human zeta globin sequence is used as a reference for the alignment. Sequence identity is shown by a dot. (**B**) A distance matrix is constructed in which pairwise comparisons are used to establish the fraction of amino acid sites that are different between two sequences. For example, the human and chimp sequences differ at one amino acid out of 50 and so the fraction is 0.02. (**C**) The distance matrix values are used to construct the phylogenetic tree. Note that the lengths of the horizontal branches are proportional to the distances between the nodes. Phylogeny software packages like PHYLIP are available at various web servers – e.g. at the Institut Pasteur in Paris (http://bioweb2.pasteur.fr/phylogeny/intro-en.html) – and typically offer neighbor joining (unrooted trees) and UPGMA (rooted trees) as alternatives.

Having calculated the matrix of pairwise differences, the next step is to link the sequences according to the evolutionary distance between them. For example, in one approach the pair of sequences that have the smallest distance score are connected with a root in between them. The average of the distances from each member of this pair to a third node is used for the next step of the distance matrix, and the process repeated until all sequences have been placed in the tree. This always results in a rooted tree but variant methods such as the *neighbor relation* method can create unrooted trees.

The methods used to construct evolutionary trees usually make some assumptions that may not always be true. Thus, they usually assume that all changes are independent, which can cause false interpretations if a single event led to multiple changes. The mutation rate is also usually assumed to be constant in different lineages, which may not be correct. However, the *neighbor joining* method is a variant that does not require that all lineages have diverged by equal amounts per unit time. It is especially suited for datasets comprising lineages with largely varying rates of evolution.

Alternatives to distance matrix methods include maximum parsimony and maximum likelihood methods. *Maximum parsimony* methods seek to use the minimum number of evolutionary steps. They consider all the possible evolutionary trees that could explain the observed sequence relationships but then select those that require the fewest changes. *Maximum likelihood* methods create all possible trees and then use statistics to evaluate which tree is likely. This may be possible for a small number of sequences. For a large number of sequences, the number of generated trees becomes extremely large and the computational time needed to identify the best evolutionary tree becomes impossibly long. In that case, a subset of possible trees is created using *heuristics*: methods are used that will produce an answer in a computable length of time, even if it may not be the optimal one.

Once an evolutionary tree has been derived, statistical methods can be used to gain a measure of its reliability. A popular method is **bootstrapping**, a form of Monte Carlo simulation. Typically, a subsample of the data is removed and replaced by a randomly generated equivalent dataset and the resulting *pseudosequence* is analyzed to see if the suggested evolutionary pattern is still favored. If there is a clear relationship between two sequences, the randomization introduced by the resampling will not erase it. If, on the other hand, a node connecting the two sequences in the original tree is spurious, it may disappear upon randomization because the randomization process changes the frequencies of the individual sites.

Bootstrapping often involves re-sampling subsets of data 1000 times. The *bootstrap value* is typically given as a percentage. A value of 100 means that the simulations fully support the original interpretation; values of 95–100 indicate a high level of confidence in a predicted node; values less than 95 do not mean that the original grouping of sequences is wrong, but that the available data do not provide convincing support.

As a result of a combination of classical and molecular phylogenetic approaches, humans have been classified as belonging to a range of different groups of organisms (see **Figure 13.31**).

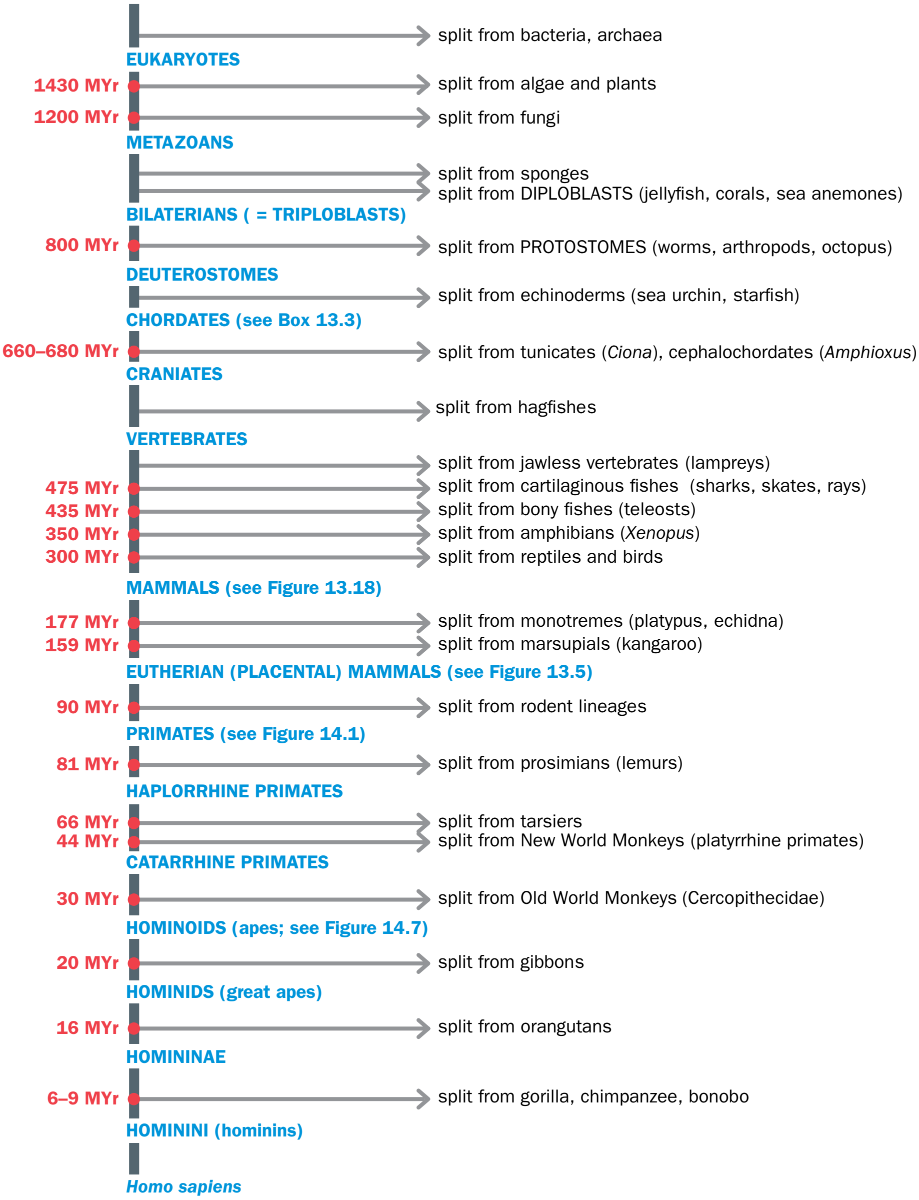

split from bacteria, archaea
EUKARYOTES
1430 MYr
split from algae and plants
1200 MYr
split from fungi
METAZOANS
split from sponges
split from DIPLOBLASTS (jellyfish, corals, sea anemones)
BILATERIANS (= TRIPLOBLASTS)
800 MYr
split from PROTOSTOMES (worms, arthropods, octopus)
DEUTEROSTOMES
split from echinoderms (sea urchin, starfish)
CHORDATES (see Box 13.3)
660–680 MYr
split from tunicates (Ciona), cephalochordates (Amphioxus)
CRANIATES
split from hagfishes
VERTEBRATES
split from jawless vertebrates (lampreys)
475 MYr
split from cartilaginous fishes (sharks, skates, rays)
435 MYr
split from bony fishes (teleosts)
350 MYr
split from amphibians (Xenopus)
300 MYr
split from reptiles and birds
MAMMALS (see Figure 13.18)
177 MYr
split from monotremes (platypus, echidna)
159 MYr
split from marsupials (kangaroo)
EUTHERIAN (PLACENTAL) MAMMALS (see Figure 13.5)
90 MYr
split from rodent lineages
PRIMATES (see Figure 14.1)
81 MYr
split from prosimians (lemurs)
HAPLORRHINE PRIMATES
66 MYr
split from tarsiers
44 MYr
split from New World Monkeys (platyrrhine primates)
CATARRHINE PRIMATES
30 MYr
split from Old World Monkeys (Cercopithecidae)
HOMINOIDS (apes; see Figure 14.7)
20 MYr
split from gibbons
HOMINIDS (great apes)
16 MYr
split from orangutans
HOMININAE
6–9 MYr
split from gorilla, chimpanzee, bonobo
HOMININI (hominins)
Homo sapiens

Figure 13.31 A simplified phylogeny showing our position in the Tree of Life. The vertical line indicates the lineage leading down to modern humans. Numbers in red followed by MYr (millions of years) indicate *approximate* times for selected splits from this lineage (the times given here are from TIMETREE at www.timetree.org and are average times from different studies). Craniates are animals with skulls. Diploblasts have two germ layers and bilaterians (= triploblasts) have three germ layers; both are bilaterally symmetrical. Deuterostomes are animals in which the blastopore (the opening of the primitive digestive cavity to the exterior of the embryo) is located to the posterior of the embryo and becomes the anus; later on, the mouth opens at the opposing end. Protostomes are organisms in which the mouth originates from the blastopore; later on, the anus will open at the opposing end. Note that the fundamental protostome–deuterostome split, which occurred about 800 Myr ago, means that *C. elegans* and *D. melanogaster* are more distantly related to humans than are some other invertebrates, such as the sea urchin *Strongylocentrotus purpuratus*, the tunicate *Ciona intestinalis*, and the cephalochordate *Amphioxus*. We are the only surviving members of the Hominins, which included other species of *Homo* and some other genera, notably *Australopithecus*, see **Figure 14.2**.

SUMMARY

- Cross-species comparisons of nucleotide and inferred protein sequences are important in gene prediction and in validating putative genes.

- Comparative genomics relies on computer programs to identify homologous sequences from the genomes of different species, group them into sets, and arrange the sets in a linear sequence. The resulting genome-wide sequence alignments provide information on the relatedness of genome sequences and insights into the processes shaping genome evolution.

- Comparative genomics is valuable for identifying regions of the genome that are subject to purifying or positive selection. Quantifying the percentage of a genome subject to purifying selection allows estimates of the proportion of the genome that is functionally important.

- Comparative genomics has been especially valuable in identifying functionally important noncoding DNA.

- During evolution, the genomes of complex metazoans have undergone periodic gain and loss of DNA sequences, including both gain and loss of genes and regulatory sequences.

- Orphan (species-specific) genes are extremely rare, but lineage-specific gene losses are common, and gene families often show lineage-specific rapid expansion in copy number in response to selection.

- Coding DNA sequences have been highly conserved during evolution, but account for a small—often very small—proportion of the genomes of complex metazoans.

- Functionally important RNAs show less evolutionary sequence conservation than coding DNA, but the RNA structure may be strongly conserved. Regulatory sequences are also generally less broadly conserved, but some may be much more highly conserved than coding sequences when comparing closely related species.

- Evolutionary novelty often arises through periodic duplication of genes or gene segments in germ-line DNA, either at the DNA level or by using reverse transcriptase to make a DNA copy of an RNA transcript that then integrates elsewhere in the genome (a retrocopy).

- Genes originating by duplication of a gene within a single genome are known as paralogs (as opposed to orthologs, homologous genes in different species that descended from the same gene in a common ancestor).

- Duplicated genes may offer the possibility of increased gene dosage, but the major advantage is to enable genes to be produced with different characteristics and eventually different functions.

- Chromosome evolution can be uncoupled from phenotype evolution: very closely related species may sometimes differ markedly in the number and average size of their chromosomes.

- Sex chromosomes can originate from a pair of autosomes after one of the pair develops a sex-determining locus that must be maintained on that chromosome (by inversions that suppress recombination between the two chromosomes).

- In mammals, males are heterogametic (with two different sex chromosomes, X and Y); females usually have two X chromosome copies. Suppression of recombination between the proto-X and proto-Y led to degeneration of the Y chromosome with loss of much of the original DNA and most of the genes (the X escaped this fate: unlike the Y, it can recombine with another X in female meiosis).

- The male-specific region on the Y is a very large central region that never recombines. X-Y recombination is confined to short terminal pseudoautosomal regions (so called because the sequences are neither X-linked nor Y-linked).

- The Y chromosome has multiple-copy genes that work in spermatogenesis and are maintained by a form of intrachromatid recombination between inverted repeats, plus some widely expressed genes that work in basic cell functions and have functional homologs on the X.

- Mammalian X-inactivation developed because of the imbalance between genes on the X and Y, but the minority of genes on the X that have a functional gene homolog on the Y escape X-inactivation.

- Morphological evolution is largely driven by mutations in *cis*-regulatory elements rather than differences in protein sequences. Whereas genes are generally highly conserved, regulatory sequences often evolve rapidly, but some ultraconserved sequences are important in gene regulation.

- Lineage-specific regulatory elements are common (the sequences may be highly conserved within different species in a common lineage, but be absent from more distantly related sequences). Their emergence may drive selection for transcription factors with lineage-specific sequences that can bind to the novel elements.

- In addition to facilitating tandem gene duplication and shuffling of exons between genes, transposable elements can also be co-opted to make new functionally important sequences (exaptation); they can give rise to new exons (exonization), regulatory sequences, long noncoding RNAs, and sometimes even new genes.

- Modern molecular phylogenetics compares nucleotide and protein sequences from different organisms in order to deduce the evolutionary history of DNA sequences and proteins and to infer evolutionary relationships between species and groups of organisms.

- The reliability of an evolutionary tree is often assessed by bootstrapping, a reiterative statistical procedure: subsamples of the data are repeatedly removed and replaced by randomly generated equivalent datasets and the resulting pseudosequences are tested to see to what extent the predicted evolutionary relationships are upheld.

FURTHER READING

Comparative genomics: general reviews and methodology

Alföldi J & Lindblad-Toh K (2013) Comparative genomics as a tool to understand evolution and disease. *Genome Res* **23**:1063–1068; PMID 23817047.

Blanchette M (2007) Computation and analysis of genomic multi-sequence alignments. *Annu Rev Genomics Hum Genet* **8**:193–213; PMID 17489682.

Copley RR (2008) The animal in the genome: comparative genomics and evolution. *Philos Trans R Soc Lond B Biol Sci* **363**:1453–1461; PMID 18192189.

Dolinski K & Botstein D (2007) Orthology and functional conservation in eukaryotes. *Annu Rev Genet* **41**:465–507; PMID 17678444.

Margulies EH & Birney E (2008) Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet* **9**:303–313; PMID 18347593.

Visel A *et al*. (2007) Enhancer identification through comparative genomics. *Semin Cell Dev Biol* **18**:140–152; PMID 17276707.

General and computational molecular evolution

Bromham L (2008) *Reading the Story in DNA: A Beginner's Guide to Molecular Evolution*. Oxford University Press.

Graur D (2016) *Molecular and Genome Evolution*. Sinauer Associates Inc.

Yang Z (2006) *Computational Molecular Evolution*. Oxford University Press.

General genome evolution, gene duplication, gene loss, and intron evolution

Albalat R & Cañestro C (2016) Evolution by gene loss. *Nat Rev Genet* **17**:379–391; PMID 27087500.

Conant GC & Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**:938–950; PMID 19015656.

Dunn CW & Ryan JF (2015) The evolution of animal genomes. *Curr Opin Genet Dev* **35**:25–32; PMID 26363125.

Ponting CP (2008) The functional repertoires of metazoan genomes. *Nat Rev Genet* **9**:689–698; PMID 18663365.

Rodríguez-Trelles F *et al*. (2006) Origins and evolution of spliceosomal introns. *Annu Rev Genet* **40**:47–76; PMID 17094737.

Sémon M & Wolfe KH (2007) Consequences of genome duplication. *Curr Opin Genet Dev* **17**:505–512; PMID 18006297.

Storz JF (2016) Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. *Physiology* **31**:223–232; PMID 27053736.

Wang X *et al*. (2006) Gene losses during human origins. *PLoS Biol* **4**:e52; PMID 16464126.

Chromosome evolution

Bellott DW *et al*. (2014) Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**:494–499; PMID 24759411.

Connallon T & Clark AG (2010) Gene duplication, gene conversion and the evolution of the Y chromosome. *Genetics* **186**:277–286; PMID 20551442.

Ferguson-Smith MA & Trifonov V (2007) Mammalian karyotype evolution. *Nat Rev Genet* **8**:950–962; PMID 18007651.

Graves JA (2016) Evolution of vertebrate sex chromosomes and dosage compensation. *Nat Rev Genet* **17**:33–46; PMID 26616198.

Hughes JF & Page DC (2015) The biology and evolution of mammalian Y chromosomes. *Annu Rev Genet* **49**:507–527; PMID 26442847.

Raudsepp T & Chowdhary BP (2015) The eutherian pseudoautosomal region. *Cytogenet Genome Res* **147**:81–94; PMID 26730606.

Evolution of mammalian and primate genomes

Lindblad-Toh K *et al*. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482; PMID 21993624.

Marques AC *et al*. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3**:e357; PMID 16201836.

Marques-Bonet T *et al*. (2009) The origins and impact of primate segmental duplications. *Trends Genet* **25**:443–454; PMID 19796838.

Meunier J *et al*. (2013) Birth and expression evolution of mammalian microRNA genes. *Genome Res* **23**:34–45; PMID 23034410.

Navarro FCP & Galante PAF (2015) A genome-wide landscape of retrocopies in primate genomes. *Genome Biol Evol* **7**:2265–2275; PMID 26224704.

Ponting CP & Goodstadt L (2009) Separating derived from ancestral features of mouse and human genomes. *Biochem Soc Trans* **37**:734–739; PMID 19614585.

Simonti CN & Capra JA (2015) The evolution of the human genome. *Curr Opin Genet Dev* **35**:9–15; PMID 26338498.

Evolution of noncoding DNA and regulatory sequences

Carroll SB (2008) Evo-Devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**:25–36; PMID 18614008.

Visel A *et al*. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**:158–160; PMID 18176564.

Wray GA (2007) The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet* **8**:206–216; PMID 17304246.

Evolutionary importance of transposable elements and new genes from noncoding DNA

Andersson DI *et al*. (2015) Evolution of new functions *de novo* and from preexisting genes. *Cold Spring Harb Perspect Biol* **7**:a017996; PMID 26032716.

Cordaux R & Batzer M (2009) The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**:691–703; PMID 19763152.

Eisenberg E (2016) Proteome diversification by genomic parasites. *Genome Biol* **17**:17; PMID 26832153.

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**:397–405; PMID 18368054.

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**:1313–1326; PMID 20651121.

Long M *et al*. (2013) New gene evolution: little did we know. *Annu Rev Genet* **47**:307–333; PMID 24050177.

McLysaght A & Guerzoni D (2015) New genes from non-coding sequence: the role of *de novo* protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci* **370**:20140332; PMID 26323763.

Rayan NA *et al*. (2016) Massive contribution of transposable elements to mammalian regulatory sequences. *Semin Cell Dev Biol* **57**:51–56; PMID 27174439.

Tautz D & Domazet-Lošo T (2011) The evolutionary origin of orphan genes. *Nat Rev Genet* **12**:692–702; PMID 21878963.

Molecular phylogenetics, taxonomy and the Tree of Life Project

Delsuc F *et al*. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* **6**:361–375; PMID 15861208.

DeSalle R, Rosenfeld JA (2013) *Phylogenomics. A Primer*. Garland Science.

Hall BG (2007) *Phylogenetic Trees Made Easy: A How-To Manual*, 3rd edn. Sinauer Associates.

Hedges SB *et al*. (2015) Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol* **32**:835–845; PMID 25739733.

NCBI Taxonomy Browser. http://www.ncbi.nlm.nih.gov/taxonomy

Nei M & Kumar S (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press.

Phylogeny Programs. http://evolution.genetics.washington.edu/phylip/software.serv.html (Phylogeny programs available through the Web have been collated at various sites such as this one.)

TIMETREE. http://www.timetree.org/ (A public resource for knowledge on the timescale and evolutionary history of life.)

Tree of Life web project. http://tolweb.org/tree/

14 Human evolution

"Nothing in biology makes sense except in the light of evolution." (Dobzhansky, 1973)

In this chapter, we consider our current understanding of the evolutionary past of humans. We also show how it helps us understand aspects of our biology that would otherwise be puzzling or even paradoxical.

Humans belong to the great apes, and shared the same evolutionary history until a few million years ago. However, humans differ from all the other great apes in both phenotype and population genetics. Why are humans ubiquitous and with a growing population size, while other great apes are confined to small areas of the tropics and critically endangered? When did this difference start to develop, and did genetic changes underlie it? We will see in Section 14.1 that some of these questions can now be answered, while others remain unclear.

Our understanding of human evolution has been transformed by new DNA sequencing technologies, which have provided access to whole genome sequences from large numbers of contemporary humans, from great ape species, and even from the remains of extinct ancient humans such as Neanderthals. In Section 14.2 we describe the insights these have provided into human origins, and the genetic changes responsible for a few of the phenotypic changes that have made humans unique. Differences between the behaviors of men and women in the past have strongly influenced patterns of genetic variation in the parts of the genome that are inherited from only one parent—the male-specific Y chromosome, and maternally inherited mitochondrial DNA (Section 14.3).

Finally, in Section 14.4 we turn to the medical implications of our evolutionary past, describing the role of natural selection in shaping our genomes. Alleles in particular genes have been positively selected in some human populations—these include sequence variants linked to the lactase gene, which confer the ability to digest fresh milk into adulthood and are associated with a history of dairying. Some variants provide a selective advantage when present in a single allelic copy but cause disease when present in two copies—the phenomenon of balancing selection.

14.1 HUMAN ORIGINS

Humans have always been interested in their origins, and scientific thinking has been applied to this question over the last few centuries. Long before genetic data were available, humans were recognized as mammals, within the Primate order, and related most closely to other **great apes**: chimpanzees, bonobos, gorillas, and orangutans. Yet we are distinct from these great apes in our ecology, morphology, and behavior and in many other ways, and intensive studies over the last 150 years of the fossils and artifacts left by earlier humans (**paleontology** and archeology, respectively) have documented the development of these differences. Over the last few decades, advances in technology have led to the possibility of extracting DNA fragments from some fossils, creating the field of studying **ancient DNA**, often abbreviated **aDNA**, which merges paleontology with genetics. We introduce these topics in this section, before moving on to a more thorough consideration of the evidence from genetics in the next section.

The place of humans within the mammalian class

Humans are mammals because we are warm-blooded and have hair and (in females) mammary glands, as well as less obvious characteristics such as a neocortex (part of the brain) and three middle ear bones. We have large brains and flexible behavior, binocular vision, and well-developed hands, characteristics that classify us as Primates. Within the Primates order, we are apes related to the other great apes (Figure 14.1).

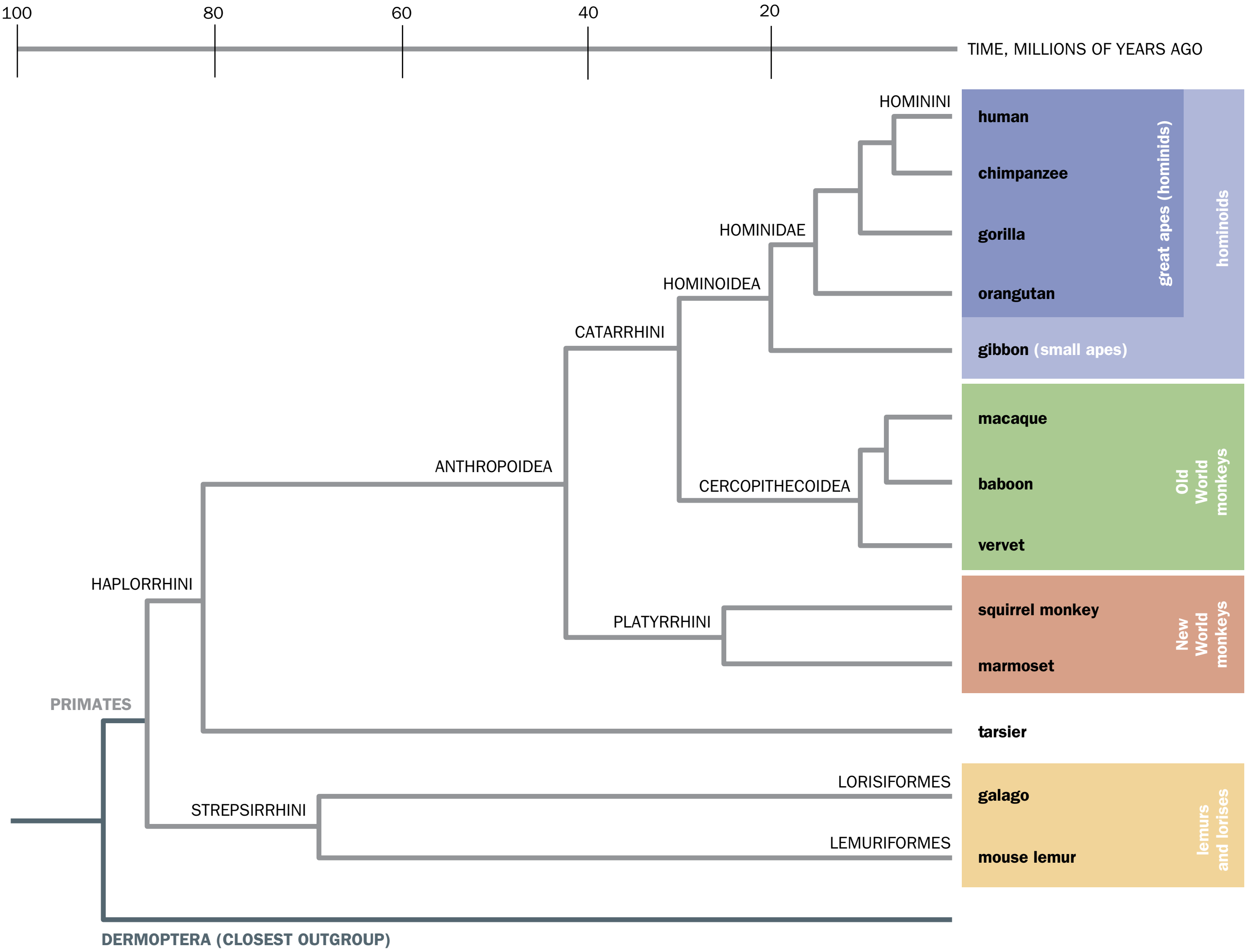


Figure 14.1 A phylogenetic tree of the primates, based on genome sequences. (From Jobling M *et al.* [2014] *Human Evolutionary Genetics*, 2nd edn. With permission from Garland Science.)

Genetic analyses, now including genome sequences, show that humans are most closely related to chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*), then gorillas (*Gorilla gorilla* and *G. beringei*), with orangutans (*Pongo pygmaeus*, *P. abelii, and P. tapanuliensis*) as our most distantly-related great apes (see Figure 14.1).

While the evolutionary relationships within the great apes are now established, the species **divergence times** are less clear. There are very few informative fossil remains, and the classifications of the available fossils are often contentious, so dating has depended more on molecular data, especially DNA sequences. Yet even with the benefit of genome sequences of all the relevant species, it is not simple to calculate divergence times between them. Such calculations require that we know the nucleotide **substitution rate** over the relevant time period. The **mutation rate** of single nucleotides has been measured in present-day human families by counting the number of new mutations arising between parents and their children (see Section 11.3) and can be used if we assume a constant generation time and a constant mutation rate over millions of years (a "**molecular clock**"). More recently, the mutation rate has been estimated using ancient human DNA sequences from remains up to 45,000 years old. These estimates have the advantage that the rate is calibrated in years rather than generations, so assumptions about the human generation time in the past are not needed. Using all the available evidence, times of 5–7 million years for the human/chimpanzee divergence, 8–10 million years for the human/gorilla divergence, and 12–16 million years for the human/orangutan divergence have been estimated.

Humans stand out among the great apes in several ways. First, there is currently only one species of human, while there are multiple species and subspecies among all the other genera. Second, humans inhabit every latitude and continent, while other great apes are confined to small areas of the tropics in Africa or Southeast Asia. Third, humans, with a census population size of over 7 billion, are more than 10,000 times more numerous than any other great ape. We will see that there is a simple evolutionary explanation for all these differences. In addition to these population characteristics, humans differ from the other great apes in a number of phenotypic traits, ranging from obvious ones like upright walking and minimal bodily hair, to less obvious ones like T-cell hyper-reactivity (**Table 14.1**).

TABLE 14.1 SOME PHENOTYPIC DIFFERENCES BETWEEN HUMANS AND OTHER GREAT APES

Phenotype/characteristic	In humans	In other great apes
Upright walking	Yes	No
Endurance running	Yes	No
Brain size	Large (1000–1500 cm^3)	Small (275–750 cm^3)
Opposable thumb	Yes	No
Body hair	Most visible on head and around genitalia (after adolescence)	Covers most of the body
Eccrine sweat gland density	High	Low
Descended larynx	Yes	No
T-cell hyper-reactivity	Yes	No
Group size	Large	Small (<50)
Emotional crying	Yes	No
Syntactical language	Yes	No
Capacity for full artistic expression	Yes	No

The fossil and archeological records of human origins

For most great apes, very few fossils are known: only a handful of chimpanzee teeth half a million years old from Kenya, the 8 million-year-old early gorilla *Chororapithecus abyssinicus* from Ethiopia, and more abundant 8–12 million-year-old *Sivapithecus* fossils from South Asia, which were probably related to orangutans. Ancient great apes would have been rare, and their hot, humid environments would usually have prevented fossilization.

In contrast, more fossils of **hominins**—extinct species that are more closely related to humans than to other great apes—are known (**Figure 14.2**), perhaps because they inhabited more varied environments, some of which were more favorable for fossilization.

The oldest of these, at around 7 million years, is *Sahelanthropus tchadensis*; if truly a hominin, this would demonstrate that the human/chimpanzee divergence was earlier than this, but its hominin status is disputed. The later species, especially those after 4 million years ago, are more widely accepted as hominins. From them, we can draw several general conclusions:

- Before 2 million years ago, hominins are known only from Africa, providing strong evidence for an origin of our lineage in Africa;

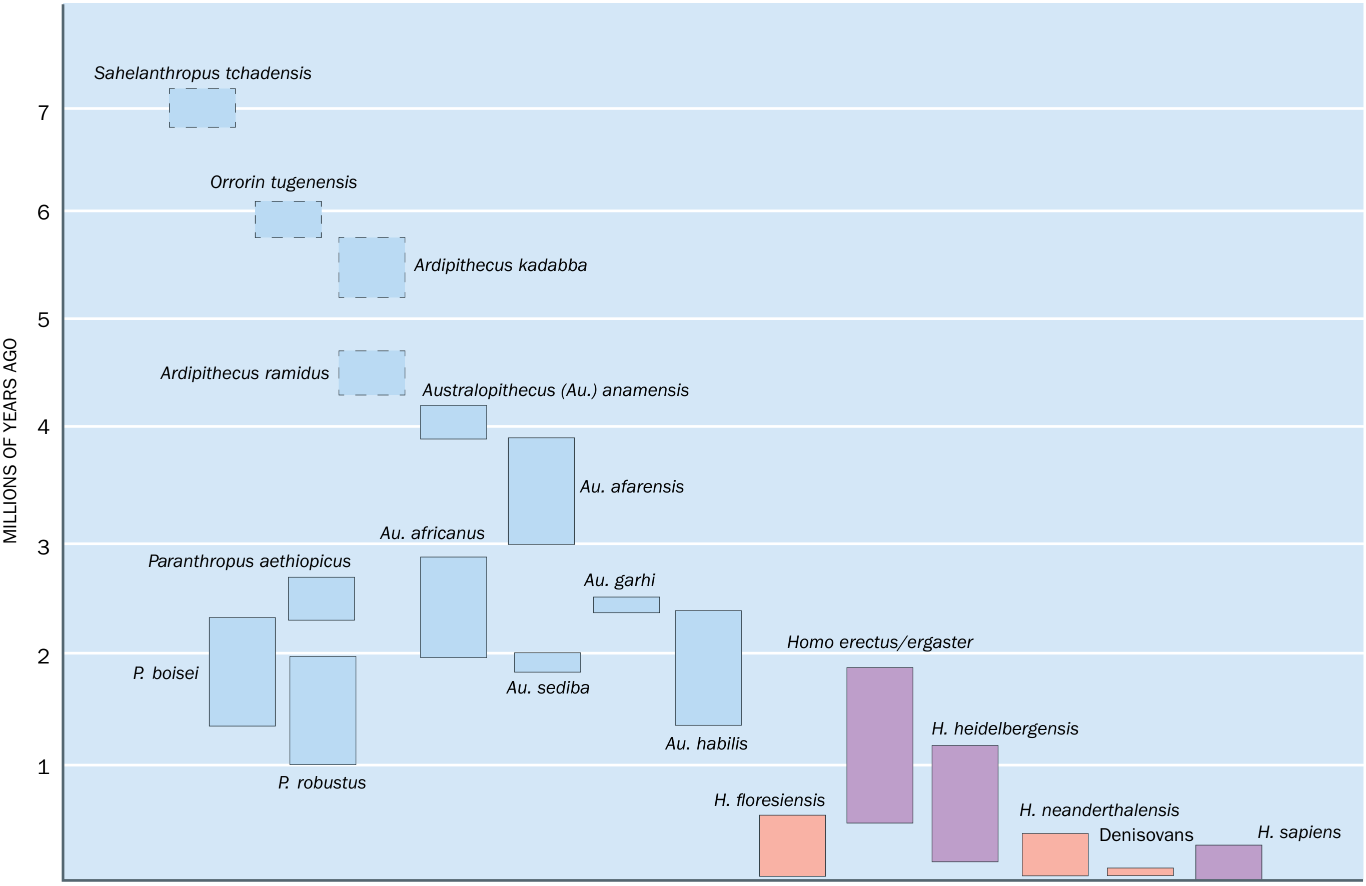


Figure 14.2 Candidate hominin fossils. Border indicates hominin status (dashed: unclear hominin status; solid: hominin). Color indicates geographical location: light blue, found only inside Africa; purple, both inside and outside Africa; pink, only outside Africa. (Updated from Jobling M *et al.* [2014] *Human Evolutionary Genetics*, 2nd edn. With permission from Garland Science.)

- When the fossil record is reasonably rich (that is, after 3 million years ago), multiple species were always present, until recently;
- Our genus *Homo* appears 1.9–2.5 million years ago (depending on whether the species *habilis* is assigned to the genus *Homo* or to *Australopithecus*);
- Several *Homo* species, including *H. erectus*, *H. heidelbergensis*, and *H. sapiens*, expanded out of Africa at different times;
- Our species *sapiens* is first recognized only ~300,000 years ago in individuals who have clearly modern anatomy;
- After 30,000 years ago, only *H. sapiens* is found.

Thus one of the striking human characteristics—one species rather than many—has a recent evolutionary origin.

Some chimpanzees use stone tools, so it seems possible that the common ancestor of humans and chimpanzees may have also done so, but these tools are not sufficiently different from natural stones to be readily recognized in archeological excavations. The earliest recognizable tools (from Kenya) date to about 3.3 million years ago. In contrast to hominin fossils, archeological remains after this date are relatively common and so can provide rich additional evidence for the presence and activities of hominins.

Archeologists classify assemblages into broad groups:

- Oldowan (Mode 1), starting 3.3 million years ago;
- Acheulean (Mode 2), starting 1.8 million years ago;
- Middle Paleolithic/Middle Stone Age (Mode 3), starting 300,000 years ago;
- Upper Paleolithic/Later Stone Age (Mode 4), starting 50,000–70,000 years ago;
- Neolithic, Bronze Age, and Iron Age, after 10,000 years ago.

These technologies may have local forms, and might be used at different times in different places; the hominin species that produced them often cannot be identified from the technology alone. For example, both Neanderthals and early modern humans produced Middle Paleolithic tools, which in some contexts can be indistinguishable.

The transition to Neolithic technology might occur 10,000 years ago in one area, 5000 years ago in another, and not at all in a third.

An important finding from archeology is that the early *H. sapiens* ~300,000 years ago used Mode 3 tools and continued to do so for over 200,000 years. The first known Mode 4 tools, along with evidence for regular longer-distance transport of materials, the use of more varied materials including pigments like ochre, and the production of ornaments such as ostrich eggshell beads, occur in Africa between 100,000 and 60,000 years ago. These changes have been interpreted as marking a transition (sometimes called the "human revolution") from anatomically modern but behaviorally pre-modern *H. sapiens* to fully modern *H. sapiens*. It is these fully modern humans who expanded throughout the whole world and caused the extinction of the other *Homo* species. Nevertheless, this process was not entirely a simple replacement: Mode 4 tools were not used everywhere. Neanderthals show evidence of some tools and behaviors otherwise associated with modern humans, and genetics reveals several episodes of interbreeding, as discussed below. Despite such complexities, modern humans essentially similar to Africans 70,000 years ago became the only surviving hominin species, leading to the other two striking characteristics listed above by which humans differ from other great apes: wide dispersal and enormous numbers.

A second striking transition in the archeological record occurred much later, less than 12,000 years ago. In the Middle East, a new stone tool technology was adopted, and the change was therefore named the "Neolithic" transition. But the key aspect was the change from a hunter-gatherer subsistence to a food-producing subsistence such as agriculture, horticulture, or pastoralism. Such transitions originated independently in several different parts of the world at different times after this date, including in Africa, the Near East, East Asia, New Guinea, and the Americas, and were mostly not associated with a particular stone tool technology, so the changes may be more appropriately referred to as the "food-production" transitions. Their underlying basis was a worldwide change in climate from one that was both generally cold and sometimes fluctuating rapidly in temperature to one that was warmer and more stable. The consequences were increased food availability leading to massive population growth, domestication of the food sources, and often a more sedentary lifestyle. Longer-term consequences included further technical innovations such as metalworking, larger settlements, specialization of roles, and the suite of characteristics that we call "civilization". These changes in settlement patterns and lifestyle also introduced new health challenges in the form of new infectious diseases and requirements to adapt to new food resources.

Language, like DNA, is universal among humans and generally transmitted from parents to children with only minor changes; consequently, the relationships between language, genetics, and human history have attracted much attention. These parallels, and some examples of the insights from combining genetic and linguistic data, are discussed further in **Box 14.1**.

BOX 14.1 LANGUAGE, GENETICS, AND HISTORY

Related languages can be recognized and grouped into families, for example, English, Greek, and Hindi which, along with many others, belong to the Indo-European family and all descend from a common ancestral language, Proto-Indo-European. In contrast, other languages like Swahili, Basque, or Mandarin Chinese are very distinct and are not derived from Proto-Indo-European. This grouping of languages was one of the first to be recognized, and is widely accepted, but how should such grouping best be done, and can all 7000 or so present-day languages be arranged into a single phylogeny that summarizes their historical relationships?

One way of grouping languages uses the similarities of words, initially focusing on a list of 100 or 200 that, like "I," "you," "woman," or "sun," are universal: a **Swadesh list**, named after the US linguist Morris Swadesh. Critics question how universal and equivalent such words really are and point out that anyway words can readily be borrowed from a different language. An alternative is to use the **grammatical structure** of languages: rules such as the order in which words are combined into sentences, which are more stable. Recently, language phylogenies have been constructed using large-scale approaches that borrow methodologies from evolutionary biology, in the expectation that scale will overcome the noise in word lists and allow the confidence in the tree structure to be assessed.

Most linguists consider that languages change so rapidly that they retain little information about their deep relationships beyond 5000 or at most 10,000 years ago. Nevertheless, there are many events of great interest within this time period. One is the time and place of the origin of Indo-European languages, spoken in most of Europe and parts of Asia even before the expansions of the last few centuries. A large-scale analysis of 2449 words incorporating calibration information from extinct languages inferred an origin in Anatolia 7800–9800 years ago and an expansion associated with farming. An alternative **Kurgan hypothesis** proposes an origin in the West Asian **Steppes** around 4500 years ago and a spread during the Bronze Age by populations such as the Yamnaya; as described later in this chapter, there is indeed strong genetic evidence for such population expansions. The two hypotheses are not mutually exclusive if an early origin in Anatolia was followed by a later re-expansion from the Steppes.

Exceptions to the general pattern are also of particular interest. For example, in Europe the Hungarians speak a non-Indo-European language belonging to the Uralic family that is more closely related to Finnish than to the German or Romanian spoken in neighboring countries, yet genetically the Hungarians resemble their geographical neighbors. This is because one thousand years ago the country was conquered by a Uralic-speaking elite that imposed its language on the local people, but had little impact on the gene pool.

Analyzing DNA from fossils—the study of ancient DNA

Some fossils and nonfossilized bones or teeth contain traces of DNA, and analysis of these is currently transforming our understanding of the human past: a technological revolution that has been compared in its impact to the introduction of radiocarbon dating in the mid-twentieth century. The challenges of applying this technology are:

- aDNA fragments are invariably degraded, usually to less than 100 bp;
- The fragments are often damaged and chemically modified;
- aDNA is present in very small amounts, effectively zero in many samples;
- DNA from other sources may be more abundant than the endogenous DNA. These sources include bacteria and fungi from the environment, the investigators themselves, and the laboratories in which they work.

Consequently, aDNA studies need to be very sensitive, and as a result have been beset by problems of contamination.

Early aDNA work was based on PCR amplification of specific regions of genomic DNA chosen in advance. This approach achieved some spectacular successes, such as the determination of the sequence of Neanderthal mitochondrial DNA (mtDNA) and the demonstration that it was no more closely related to present-day Europeans (the Neanderthal samples came from Europe) than to people from other regions, but instead formed an outgroup to all living humans (**Figure 14.3**). But this approach also had severe limitations: its scale was limited, only rare long fragments (the size of two PCR primers plus an intervening sequence long enough to provide usable information) could be analyzed, and contaminants were not only frequently encountered but could also be difficult to recognize. A standard requirement was replication of results in an independent laboratory to minimize the chances of a DNA sequence being the result of contamination.

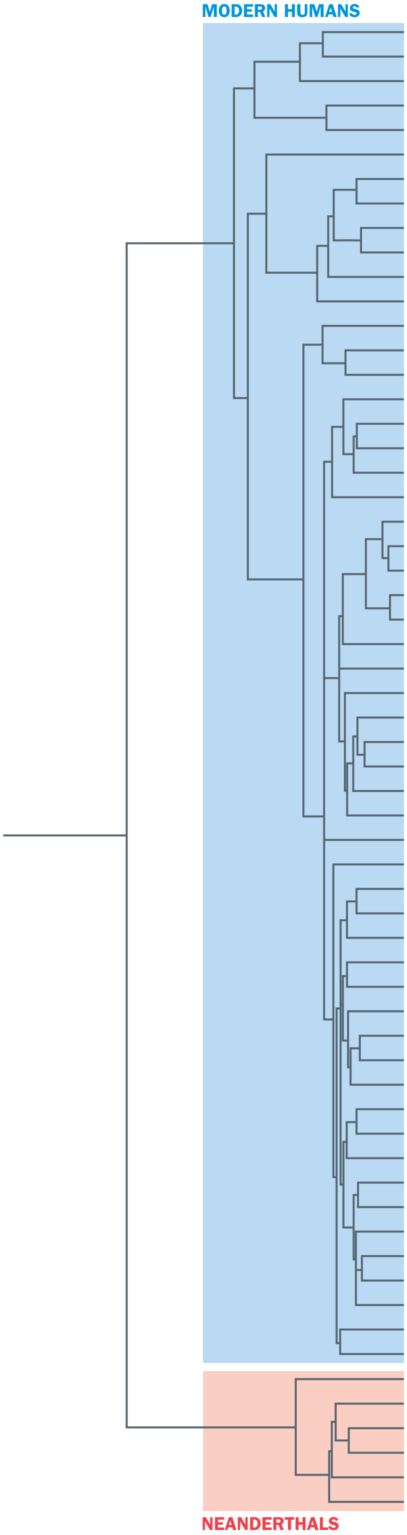


Figure 14.3 Neanderthal mtDNA forms an outgroup to all modern human mtDNAs. (Reprinted from Krause J *et al*. [2010] *Nature* **464**:894–897; PMID 20336068. With permission from Springer Nature. Copyright © 2010.)

A number of approaches have more recently been used to address the challenges listed above, and have transformed the study of aDNA:

- Next-generation sequencing technologies are well suited to sequencing very short DNA fragments in their entirety;
- Some chemical modifications can be recognized and removed by specific enzymes, such as uracil DNA glycosylase, which removes uracil residues;
- Next-generation sequencing generates data on a vast scale, allowing many samples to be screened and informative data to be generated even when the endogenous DNA content is low;
- Large amounts of data allow internal tests for contamination to be performed, such as measuring the level of Y chromosome sequence reads in a female sample (should be zero) or the number of distinct mtDNA lineages in any sample (should be one). aDNA damage introduces a characteristic pattern of C-to-T or G-to-A changes at the ends of reads, so even in a contaminated sample, a subset of authentic (that is, damaged) aDNA reads can be identified and used (**Figure 14.4**);
- The petrous bone (part of the skull) has been identified as having the highest endogenous DNA content, and sometimes yields aDNA when other bones or teeth fail;
- Enrichment by hybridization can be used for DNA sequences of interest, which may be short stretches, entire chromosomes, or potentially the whole human genome.

As a result, aDNA sequences can now be generated from increasing numbers of samples, sometimes even sequences for the whole genome, and the data are being applied to more and more questions of evolutionary interest.

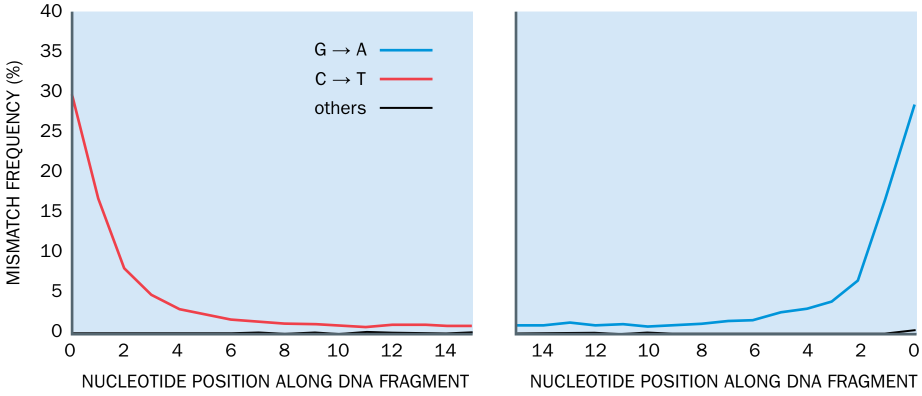


Figure 14.4 Ancient DNA fragments show a characteristic pattern of damage at their ends. (Reprinted from Krause J *et al.* [2010] *Nature* **464**:894–897; PMID 20336068. With permission from Springer Nature. Copyright © 2010.)

Ancient hominin DNA sequences

The oldest hominin fossils to have yielded DNA sequence data thus far are 430,000-year-old early Neanderthals from Sima de los Huesos ("pit of bones") in northern Spain. Despite extreme care during excavation, storage, and extraction, DNA of this age is highly fragmented and its analysis pushed current technology to its limits, so only a very low-coverage genome sequence could be generated. Nevertheless, this was sufficient to identify the bones as early Neanderthals, an important finding since their morphology could be interpreted as either *H. heidelbergensis* or *H. neanderthalensis*.

Several later (40,000 to 80,000-year-old) Neanderthals from Europe and Asia have been sequenced to low coverage, and two—one from the toe bone of a woman from Denisova Cave in Siberia, at the eastern limit of the Neanderthal range and known as the "Altai Neanderthal" and the other from a woman from Vindija cave in Croatia—to high coverage. As discussed in the next section, these sequences have allowed the genetic relationship of Neanderthals and modern humans to be clarified, and genetic differences between them to be identified.

Perhaps the most surprising outcome of ancient hominin genetic studies emerged when a finger bone, also from Denisova Cave, was analyzed. This bone was 74,000–82,000 years old and contained exceptionally well-preserved DNA, which allowed a high-coverage genome sequence to be generated. This genome was not from a Neanderthal or a modern human as expected, but formed a distinct lineage, known as "Denisovan" after its place of origin, which has not yet received a formal species name. A few additional teeth have been shown by sequencing of mtDNA to belong to the same lineage, but its overall morphology and status remain unknown. Nevertheless, as with the Neanderthal sequences, the Denisovan genetic data have been informative about hominin relationships, as also discussed below.

The oldest modern human DNA sequence currently available is from a 45,000-year-old man from Ust'-Ishim, also in Siberia, and also of high coverage. A handful of other Paleolithic human sequences have been reported, and increasingly larger numbers of sequences from more recent times. Most of these are from northern parts of Europe, Asia, or the Americas.

It is no coincidence that the high-quality Neanderthal, Denisovan, and early modern human sequences all originate in Siberia: the cold environment favors DNA preservation. Nevertheless, other regions of the world, especially Africa, are of great importance for understanding our evolutionary history. Hot and wet conditions promote DNA degradation; the oldest African sequence currently available is from Malawi, and dates back only to 8100 years ago. aDNA sequences from other parts of the world are eagerly awaited and promise many surprises.

14.2 HUMAN EVOLUTIONARY HISTORY FROM GENOME SEQUENCES

Human origins and uniqueness: genetic insights

In the previous section we learned from the fossil evidence that human ancestors are likely to have split from the ancestors of chimpanzees and bonobos 5–7 million years ago. Genome sequences of all great ape species inform us about the extent and nature of the genetic differentiation among them, and provide us with better estimates of the species divergence times. The recognized great ape species and subspecies are genetically distinct, although there is an element of circular reasoning in this conclusion because genetic data are sometimes used to inform taxonomic divisions.

By comparing genome sequences we can in principle identify genetic differences responsible for the phenotypic changes that have made humans unique. However, between the human and chimpanzee genomes, for example, there are tens of millions of differences—on average, twelve single nucleotide variants per kilobase. These include ~40,000 nonsynonymous nucleotide differences, resulting in an average of two amino acid differences in each protein-coding gene. If nucleotide substitution rates are the same in each species' lineage, half of these differences would be unique to humans and the other half unique to chimpanzees. Using the genome sequence of a third species, such as gorilla or orangutan, as an outgroup allows us to tell which are which. However, the significance of the changes is unclear since the vast majority of nonsynonymous changes do not affect the function of the proteins. The regulatory regions of genes are less well understood than their coding regions, yet changes here can be functionally more important than amino acid changes. Many human-specific phenotypes could be caused by alteration of the level or timing of gene expression in key stages of development.

One of the distinct changes in the hominin fossil record is the gradual increase in the size of the brain. In relation to the size of the face and overall body size, the braincase has grown since the origins of the genus *Homo* approximately 2 million years ago (**Figure 14.5**). As a result of this process, human adults are **neotenic**: the adult's cranium is more similar in shape to that of an infant, rather than an adult, chimpanzee. Brain size is determined by many genes, and a number of these have been proposed to be responsible for humans' large brains. The peak time of expression of many prefrontal cortex synaptic genes is shifted toward a later developmental stage in humans compared to chimpanzee and macaque. One example is the *SRGAP2A* gene, which in early stages of mammalian development promotes the maturation of the spines of nerve-cell extensions known as dendrites. A gene duplication around 2.4 million years ago created a human-specific gene copy, *SRGAP2C*. The protein encoded by this gene competes with the protein encoded by the *SRGAP2A* gene, inhibiting its function, and therefore delaying the maturation of dendrite spine development. This results in higher dendrite density and more connections between the neurons. Increasing copy number of brain-expressed genes via repeated rounds of gene duplication is another possible mechanism for generating a larger brain (see also Chapter 13 for the general background on gene birth and death). A further list of examples from a range of molecular mechanisms and types of genetic changes that may have contributed to the evolution of increased brain volumes in hominins is presented in **Table 14.2**.

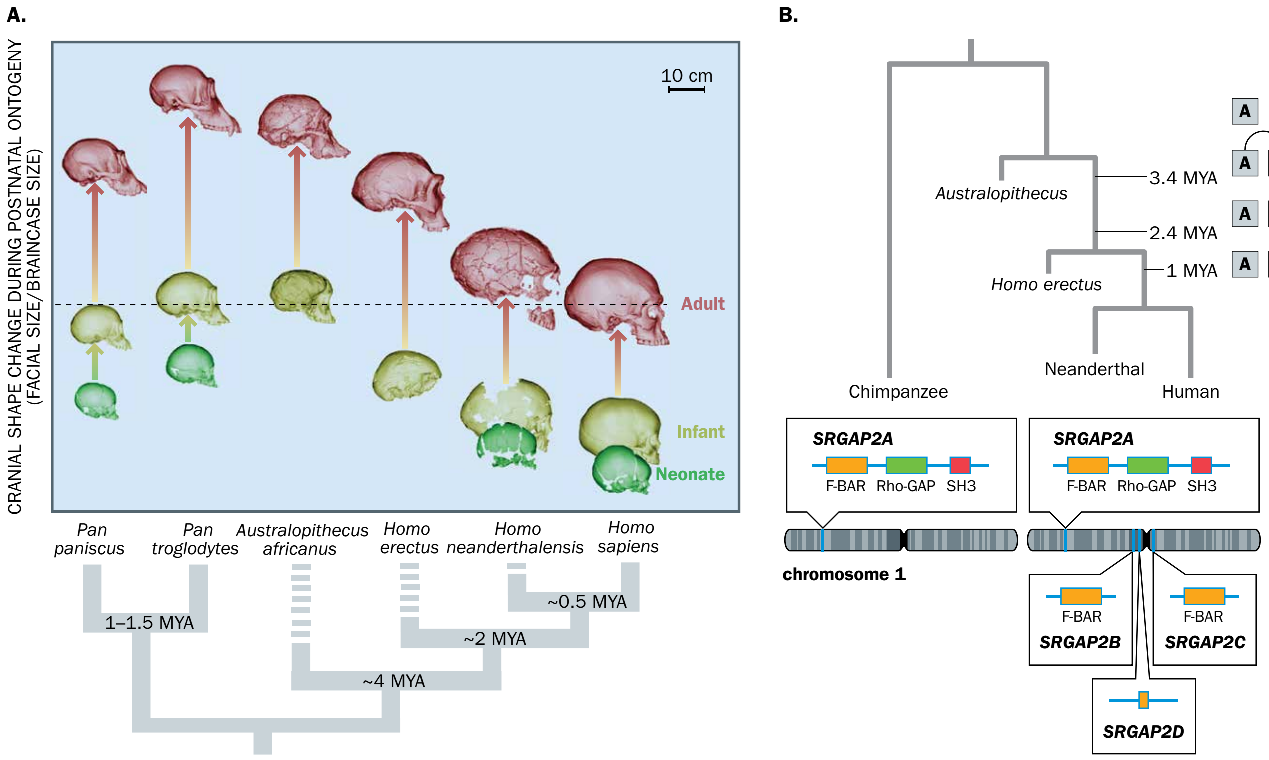


Figure 14.5 Neoteny in human evolution. (**A**) Comparison of the development of cranial shape in chimpanzee (*Pan troglodytes*), bonobo (*P. paniscus*), fossil hominins, and living humans. Extant species and Neanderthals are represented by neonates, infants (before eruption of the first permanent teeth), and adults; earlier hominins are represented by those fossil specimens that best correspond to these developmental stages. Human adult cranial shape resembles that of an infant chimpanzee. Note the evolutionary trend toward short developmental trajectories, especially of the early phase (neonate to infant) in fossil hominins. (**B**) Series of duplication events of the *SRGAP* gene. The location of *SRGAP2* paralogs on human chromosome 1 is shown with putative protein products and their functional domains based on cDNA sequencing. Arrows show the reconstructed evolutionary history of *SRGAP2* duplication events. The dates of the duplication events are inferred from sequence data assuming a "molecular clock". A and C copies encode functional proteins. MYA, million years ago. (A, from Zollikofer CP & Ponce de León MS [2010] *Semin Cell Dev Biol* **21**:441–452; PMID 19900572. With permission from Elsevier; B, from Dennis MY *et al.* [2012] *Cell* **149**:912–922; PMID 22559943. With permission from Elsevier.)

TABLE 14.2 EXAMPLES OF GENES IN WHICH THE WILD TYPE HAS HUMAN-SPECIFIC CHANGES AND WHERE THE PHENOTYPE OF HUMAN PATHOGENIC VARIANTS SUGGESTS A BRAIN-RELATED FUNCTION

Gene	Mechanism of change	Possible gene-associated disease(s)
ASPM	Novel substitutions	Microcephaly
BOLA2	Copy number variation	Autism
CDK5RAP2	Novel substitutions	Microcephaly
DUF1220 family	Protein domain copy number increase	Microcephaly; macrocephaly
GADD45G	Deletion of regulatory sequence	Thyroid carcinoma
HAR1F	Novel substitutions	Unknown
HYDIN	Copy number variation	Autism
MCPH1	Novel substitutions	Microcephaly
MYH16	Loss of function	Unknown
NOTCH2NL	Copy number increase	Microcephaly; macrocephaly
PAK2	Copy number increase	3q29 microdeletion syndrome
PDE4DIP	Copy number increase	Myeloproliferative disorder associated with eosinophilia
SLC6A13	Copy number increase	Unknown
SRGAP2	Copy number increase	Early infantile epileptic encephalopathy

Adapted and updated from O'Bleness M *et al*. (2012) *Nat Rev Genet* **13**:853–866; PMID 23154808.

Humans have low genetic diversity compared to most other mammals

Humans are one of the most widely spread mammalian species in the world. There are more than 7.6 billion of us dispersed across five different continents today, but human census sizes have not been always so high—humanity passed the 1 billion mark only two centuries ago. Historical sources can be used, with lower and lower accuracy as we look backward in time, to inform us about human census sizes in the last few thousand years. We cannot learn directly from historical sources what human population sizes were tens or hundreds of thousands of years ago, and the archeological inferences based on site densities associated with human occupation can be quite approximate. However, we can make increasingly accurate inferences about the changes in **effective population sizes** (N_e; **Box 14.2**) from genetic data as more and more genome-scale data become available. In a theoretical population of a constant size that gains through mutation each generation as much variation as it loses by **drift** (see Section 12.3) and selection, the effective population size remains the same, and is proportional to the genetic variation it contains. Larger populations generally have higher genetic diversity than small ones, and even large populations that go through

BOX 14.2 EFFECTIVE POPULATION SIZE

Effective population size, N_e, is a concept introduced to population genetics by American geneticist Sewall Wright in 1931. It is commonly used as an estimate derived from genetic data to serve as a proxy of the number of individuals in a population that contribute their genes to the next generation. In an idealized world the effective population size would match the census size if all individuals in the population were breeding adults. In the real world the effective population size is always smaller than census size because of the age structure of the population and because not all adult individuals contribute their genes equally to the next generation. Hence, N_e represents the size of an idealized population that experiences over time the same amount of drift as the population under study. Methods to infer N_e from genetic data can be based either on estimates of nucleotide diversity at different loci, the extent of linkage disequilibrium, or the extent of allele sharing and inbreeding in a population. When genetic diversity is estimated in the form of a single summary statistic, it typically reflects the long-term harmonic mean of effective population size changes in the past. Bottlenecks and major fluctuations of population sizes may dramatically reduce long-term N_e, while the existence of population structure and barriers to gene flow can increase it.

temporary size reductions (bottlenecks) will rapidly lose genetic diversity before slowly regaining it. However, real populations are hardly ever in **mutation–drift equilibrium** and the genetic diversity reflects long-term changes in effective population size. Different measures of genetic diversity can be used to assess variation and to compare the genetic histories of species. **Heterozygosity** is one such simple measure that can be estimated even if only one representative member of a species has been sampled and its genome sequenced. Compared to other great ape species, humans are characterized by relatively low levels of genome-wide average heterozygosity (Figure 14.6). Heterozygosity in humans is lower than in some endangered mammals, such as the giant panda and Sumatran orangutan, and our levels of genetic diversity make us comparable to species with a quite restricted geographical spread.

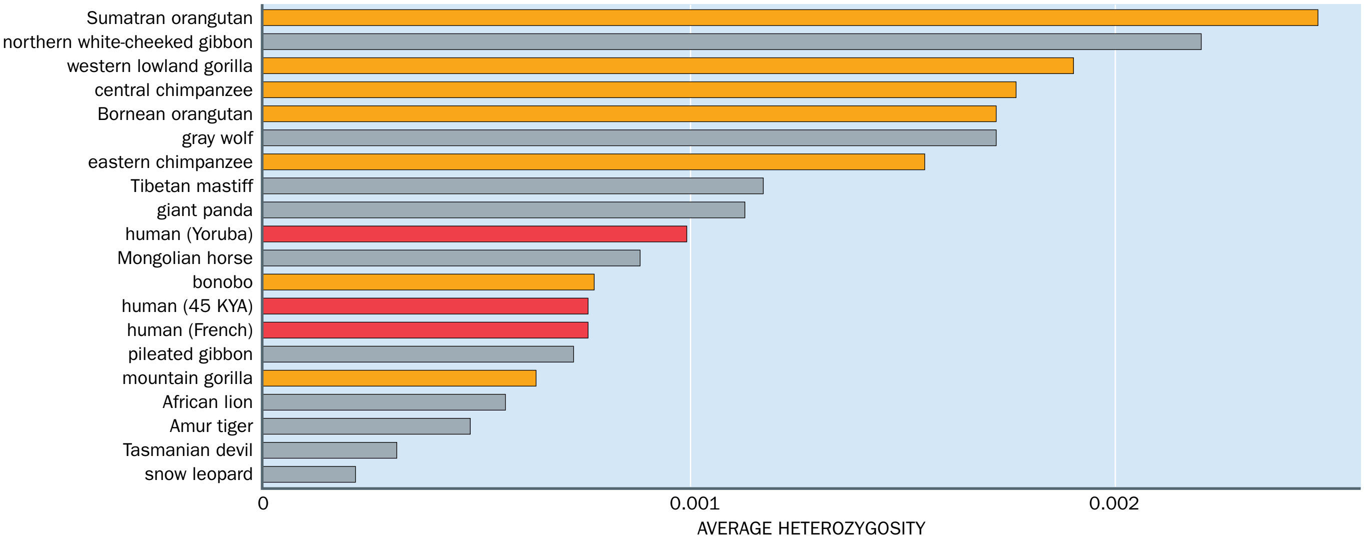


Figure 14.6 Genome-wide average heterozygosity in humans (red bars), other great apes (orange bars), and other mammals. KYA, thousand years ago.

The estimates of human genetic diversity suggest that, relative to great ape species, the uniquely high population size and wide geographic spread of our species are both likely to be recent phenomena. Major growth of human populations worldwide has occurred in the last few thousand years. Comparisons of genome-wide sequence data reveal that the N_e of human ancestral lineages in the past has been relatively low compared, for example, to the ancestral population size of the chimpanzee and bonobo lineage (Figure 14.7), and

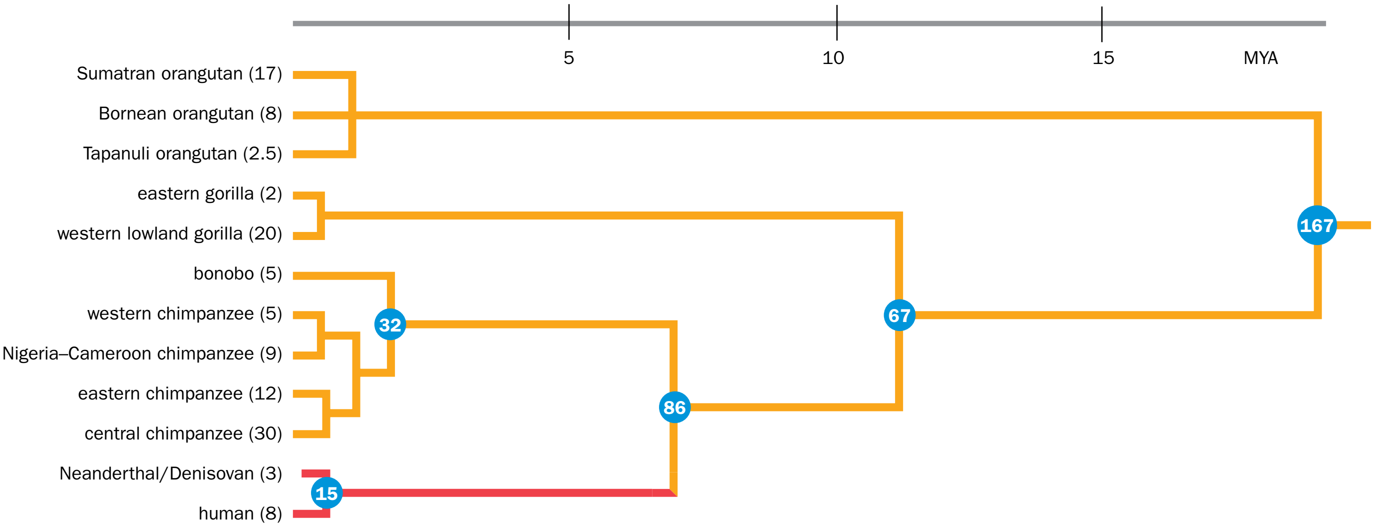


Figure 14.7 Species tree of hominoids inferred from whole-genome sequence data. The hominin branch including modern and archaic humans is highlighted in red. Effective population sizes, in thousands of individuals, are shown in parentheses after species labels. The split times assume a mutation rate of 0.5×10^{-9}/base pair/year. Ancestral effective population sizes, in thousands of individuals, are presented at coalescent nodes. (Tree adapted from Prado-Martinez J *et al.* [2013] *Nature* **499**:471–475; PMID 23823723. Ancestral effective population sizes at coalescent nodes from Mailund T *et al.* [2014] *Annu Rev Genet* **48**:519–535; PMID 25251849; Nater A *et al.* [2017] *Curr Biol* **27**:3487–3498; PMID: 29103940.)

this low N_e has been retained since at least the split of modern and archaic human lineages 550,000–750,000 years ago. The extent of genetic differentiation between humans and Neanderthals is comparable to differences among chimpanzee subspecies. The N_e of Neanderthals and Denisovans has been estimated as <3000, suggesting that even though their territory was many times larger than that of the living great ape species, their population densities were very low. If we look backward in time for several millions of years we find that the N_e of our ancestral species was much higher than that of any of its descendants today, in the range of 60,000–90,000 in the ancestor to humans and chimpanzees, or the ancestor of all African great apes.

Extensive sharing of common variants among human populations and the Out-of-Africa model

The distribution of genetic diversity among human populations is not uniform. The highest number of single nucleotide variants per genome in the 1000 Genomes Project data is observed in sub-Saharan African populations, while individuals from other continents are characterized by ~20% lower values, the exception being individuals with recently mixed ancestry such as African Americans (**Figure 14.8**). Maximum values of F_{ST} (a measure of genetic differentiation; see Chapter 12) within each continental region are within the range of a few percent, suggesting that between-population differences are minor, and that most variation is within populations. Only a small proportion of the variants in an individual genome are restricted to one continental region, and these tend to be rare in the populations in which they are found. In contrast, most genetic variants found in an individual genome are common and shared by all major continental groups. In fact, there are only a handful of variants that show the opposite pattern, and are so highly differentiated that one allele is near to fixation (100% frequency) in one geographic area and the other allele is near to fixation in another, and none that are completely fixed in this way. These extremely rare examples of highly differentiated alleles will be further discussed in the final section of this chapter in the context of their phenotypic relevance and evidence for natural selection.

The rarity of functional genes with highly differentiated alleles in human populations, combined with the high levels of sharing of common variants among continental groups, contrasts with the predictions of the **anthropological concept of race**. Anthropological races were assumed to have separate ancient origins and to be associated with specific biological, behavioral, and cultural differences. If these assumptions were supported by genetic data, then "racial" affiliation should predict individuals' genotypes. Although genetic information, when harvested from a large enough number of genetic loci, can be used to predict quite accurately an individual's birthplace (**Figure 14.9**), the reverse—predicting genotype from birthplace—is not possible. This is because alleles that have a frequency of more than a few percent in any given population tend to be shared by many populations across the world (see **Figure 14.8B**). On the other hand, those genetic markers that are confined to specific continental regions tend to have very low frequencies, which means that although there are many of them, each individual has a low probability of carrying any particular one. This, in turn, means that typically no single genetic variant, whether associated with a particular phenotype or not, contains much information about individual ancestry. The pattern of distribution of genetic diversity across continental regions supports a model that implies a recent common African origin for non-African genetic diversity, the **Out-of-Africa** (**OoA**) **model**, rather than being compatible with the model of ancient origins and long-term separation of human "races".

Two main observations about the distribution of human genetic diversity form the pillars of the OoA model: (1) genetic diversity is higher in Africans than in other continental groups; and (2) common genetic variants are mostly shared among continents. These two observations suggest that genetic variation outside Africa is a subset of African variation. At its core, the OoA model proposes that a recent **founder event**, within the last 100,000 years, was the source for most of the existing diversity outside Africa. Furthermore, human genetic variation outside Africa follows a **cline**, or gradient, where the level of heterozygosity is inversely correlated with distance from East Africa (**Figure 14.10**). This means that knowing how many heterozygous positions there are in your genome can provide you with an approximate estimate as to how far from East Africa you were born (ignoring recent migrations). This would, of course, apply only if both of your parents came from approximately the same geographic region; in practice, far more accurate predictions can be made from allele frequency data, as presented in **Figure 14.9**. One possible explanation for the observation of the decline of genome-wide heterozygosity with distance from East Africa is that, in addition to the founder event at the dawn of the OoA dispersal, the peopling of Eurasia, Sahul, and the Americas involved

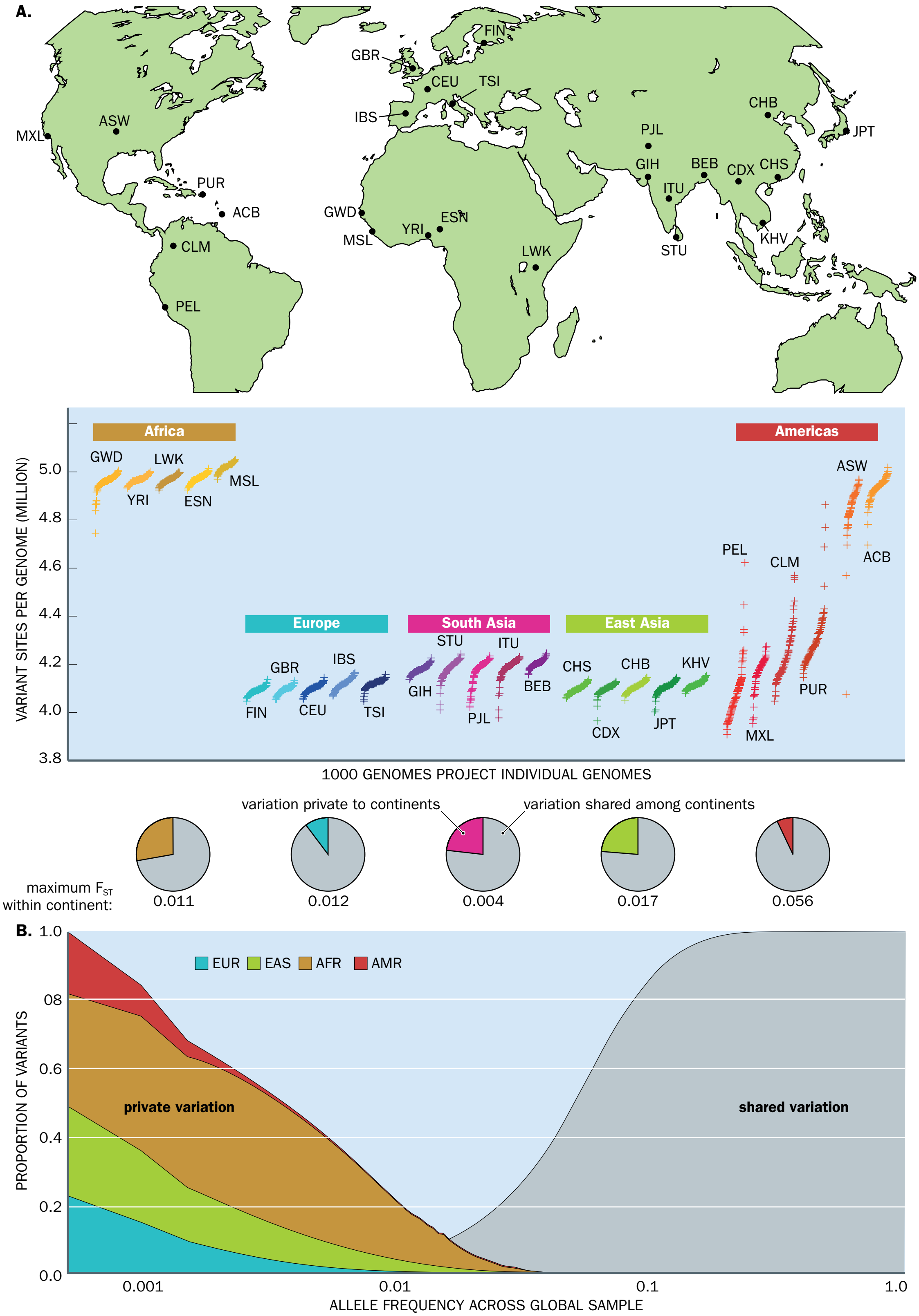


Figure 14.8 Distribution of genetic diversity among populations of the 1000 Genomes Project. (**A**) Map showing locations of 1000 Genomes Project populations (see **Figure 11.10** for definition of population abbreviations), and graph showing variant sites per genome by population, organized by continent. Note that the scale starts at 3.8 million, not zero. Below are pie-charts indicating the proportion of variation private to each continent. (**B**) Fraction of variants restricted to a single continental group (indicated by colors as shown in the key), and found in all groups (gray). Variants shared by two or three continental groups are shown in light blue. Note that the choice of the African, European, and Asian samples analyzed by the project excluded very divergent populations, so does not represent the full range of F_{ST} values. EUR, Europe; EAS, East Asia; AFR, Africa; AMR, Americas. (A, adapted from 1000 Genomes Project Consortium [2015] *Nature* **526**:68–74; PMID 26432245; B, adapted from 1000 Genomes Project Consortium [2012] *Nature* **491**:56–65; PMID 23128226. Both with permission from Springer Nature ©.)

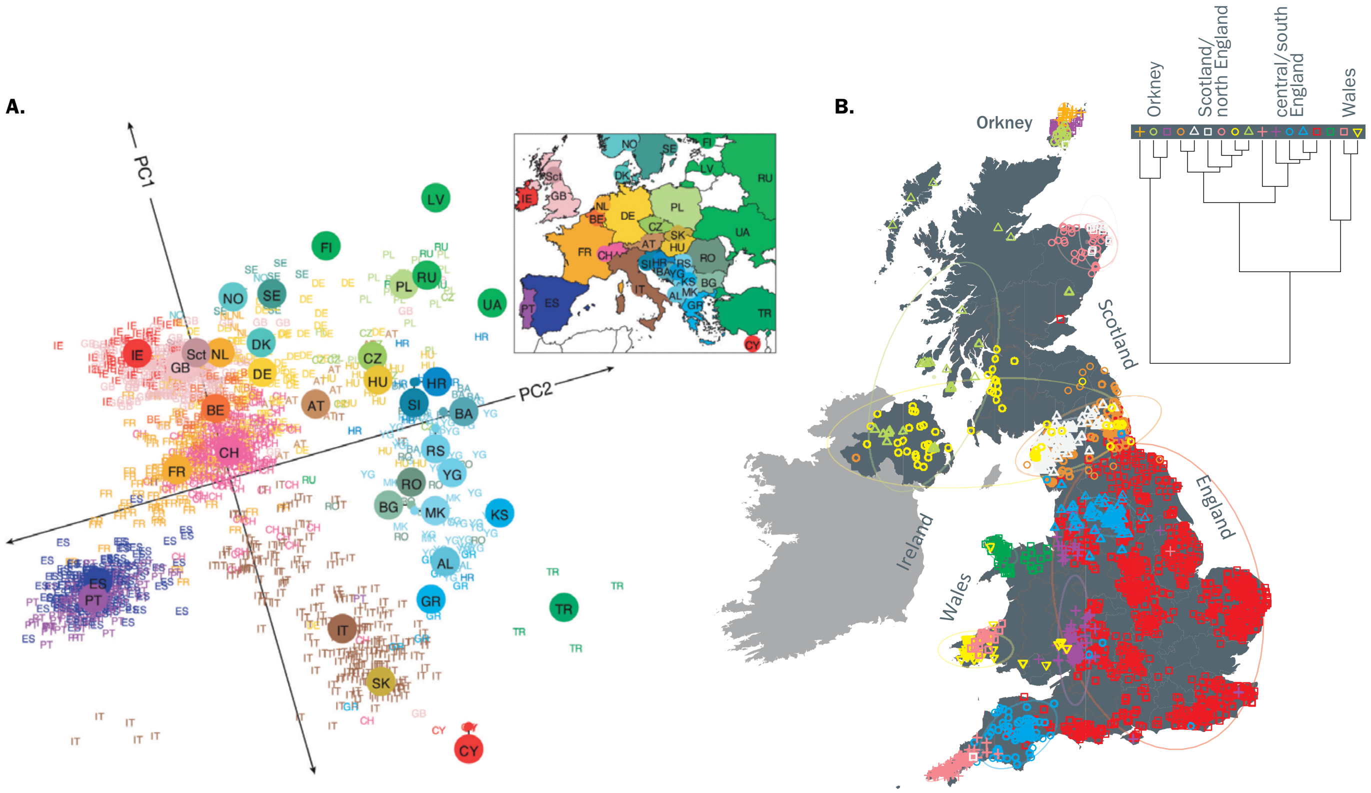


Figure 14.9 Population structure within our species, geographic regions, and populations. (**A**) Principal component (PC; a type of multivariate analysis allowing multidimensional information to be displayed graphically with minimum loss of information) map of Europe based on the analyses of 3000 individuals for ~0.5 million SNP markers. Birthplace of individuals can be predicted with the accuracy of a few hundred kilometers. PC1 and PC2 together account for 0.45% of genetic variation in Europe. (**B**) Clustering of the 2039 UK individuals of the People of the British Isles (PoBI) project into 17 clusters based on genotype data from ~0.5 million SNP markers. For each individual shown on the map of the UK, the colored symbol represents the genetic cluster to which the individual is assigned. The tree (top right) depicts the order of the hierarchical merging of the genetic clusters. (A, adapted from Novembre J *et al*. [2008] *Nature* **456**:98–101; PMID 18758442; B, adapted from Leslie S *et al*. [2015] *Nature* **519**:309–314; PMID 25788095. Both with permission from Springer Nature ©.)

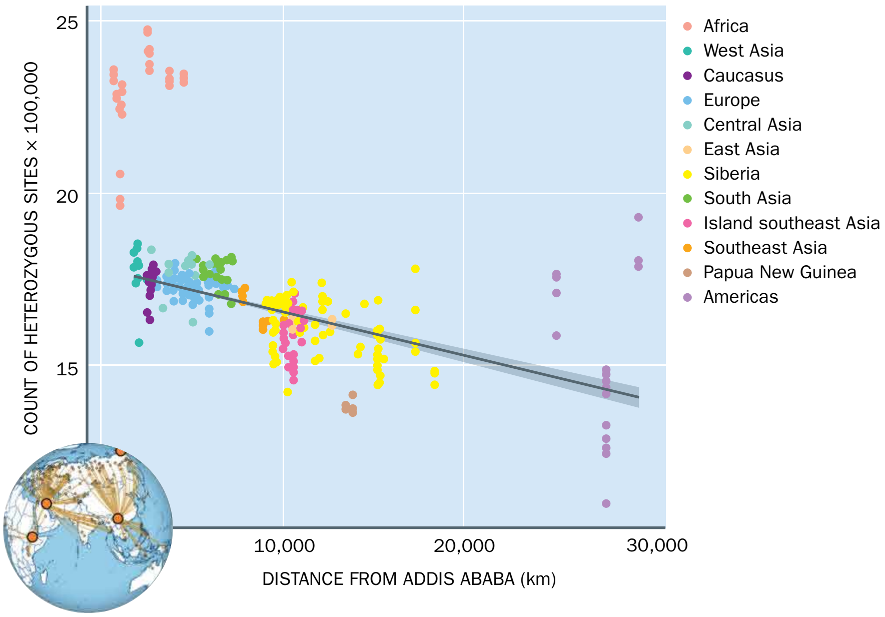


Figure 14.10 Cline of decreasing heterozygosity as a function of geographic distance from Addis Ababa (horizontal axis). The plot is based on high-coverage whole-genome sequence data. Addis Ababa is conventionally taken as a starting point in East Africa for the Out-of-Africa dispersal. The wide range of heterozygosities in the American samples can be explained by recent admixture patterns. (Adapted from Pagani L *et al*. [2016] *Nature* **538**:238–242; PMID 27654910. With permission from Springer Nature. Copyright © 2016.)

further founder events, each diminishing genetic diversity in the newly founded populations. Following this logic (the **serial founder model**), we find that regions with the lowest level of genetic diversity, the Americas and Oceania, are those furthest away from Africa. Some individuals from these remote areas, however, have higher or lower heterozygosities than expected given their birthplace, including, for example, seven

Native American individuals represented in **Figure 14.10**. Such individuals have mixed ancestries, whereas individuals with unusually low heterozygosity could represent either cases of **inbreeding** or sampling from populations that have had historically very small effective population sizes.

The relationship of heterozygosity with distance from East Africa shown in **Figure 14.10** was calculated using distances that took into account geographical barriers such as seas and mountain ranges. The effect of geographical barriers on shaping patterns of human genetic diversity can be explored further by comparing geographical distance and genetic distance between populations. Genetic differentiation between populations is commonly measured using the F_{ST} statistic (see Section 12.4), a value calculated from allele frequency differences between populations, with higher values reflecting a larger genetic differentiation between populations. Human populations from the same geographic neighborhood have usually differentiated only recently, continue to exchange migrants, and show low levels of genetic differentiation. Most population pairs in Europe, for example, show a genome-wide average F_{ST} <0.02. Over larger geographic ranges, the genetic differentiation between populations is shaped by the geographic barriers. For example, genetic differentiation increases more rapidly across the Himalaya mountain range than across the plains of Europe and the Middle East (**Figure 14.11**).

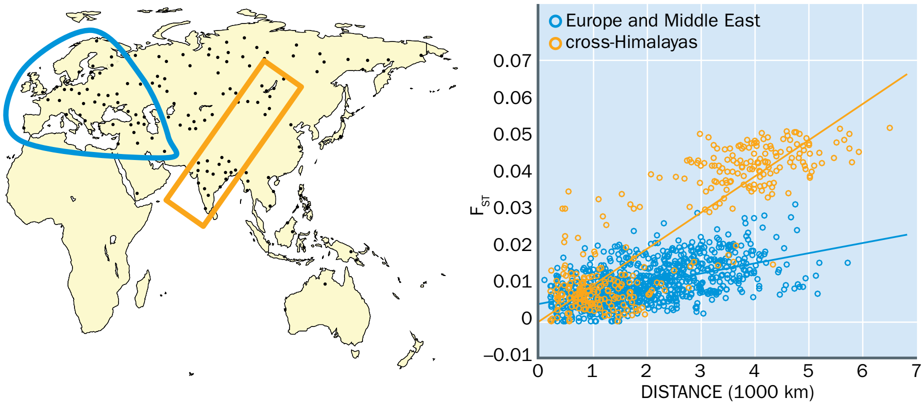


Figure 14.11 Increase of genetic distances in the presence of geographic barriers. The left panel shows the sampling points in Eurasia for genomes that were sequenced and between which the genetic distances were calculated. The right panel shows the plot of genetic distances estimated with the F_{ST} statistic against geographic distances between the sampling points. (Adapted from Pagani L *et al*. [2016] *Nature* **538**: 238–242; PMID 27654910. With permission from Springer Nature. Copyright © 2016.)

Evidence of archaic introgression

While the patterns of genetic diversity among human populations can be largely explained by the recent OoA dispersal model, several lines of evidence point to an additional source of genetic diversity: admixture with archaic humans. Populations living outside Africa show consistently higher sharing of derived alleles (alleles resulting from mutations in the human lineage) with the Neanderthal genome than Africans do. This suggests admixture outside Africa with Neanderthals. Estimates based on the pattern of decay of **linkage disequilibrium** (Section 12.2) around the shared alleles suggest that the admixture occurred ~55,000 years ago and that all populations living outside Africa obtained Neanderthal alleles from this single event of admixture, which occurred before the split of European and East Asian populations (**Figure 14.12**). Ancient DNA studies of modern human remains from the last 50,000 years have shown a decrease of Neanderthal allele-sharing over time, which suggests that, at a large number of functionally important sites, the segments of DNA our ancestors gained from archaic hominins were not fully compatible with the existing genomes, and that negative selection removed these incompatibilities over time. In particular, the X chromosome shows the lowest proportion of sharing of Neanderthal alleles.

Our current evidence for admixture with archaic humans is not limited to the one described above. The remains of a 40,000-year-old Oase individual from Romania, for example, have revealed an additional case of genetic admixture with Neanderthals: one of the direct ancestors of this individual, four to six generations before him, was inferred to have been Neanderthal. However, the Oase individual came from a population that did not directly contribute to the genetic diversity of modern Europeans.

Also, Neanderthals were not the only archaic hominins to contribute to modern human genetic diversity. Papuans and Aboriginal Australians show evidence of gene flow from Denisovans (**Figure 14.13**). Furthermore, analyses of Neanderthal remains from the Denisova cave have pointed to gene flow from modern humans to Neanderthals before the time of the Out-of-Africa expansion discussed above. It is still debated whether this earlier modern human dispersal, sometimes called the xOoA event, was restricted

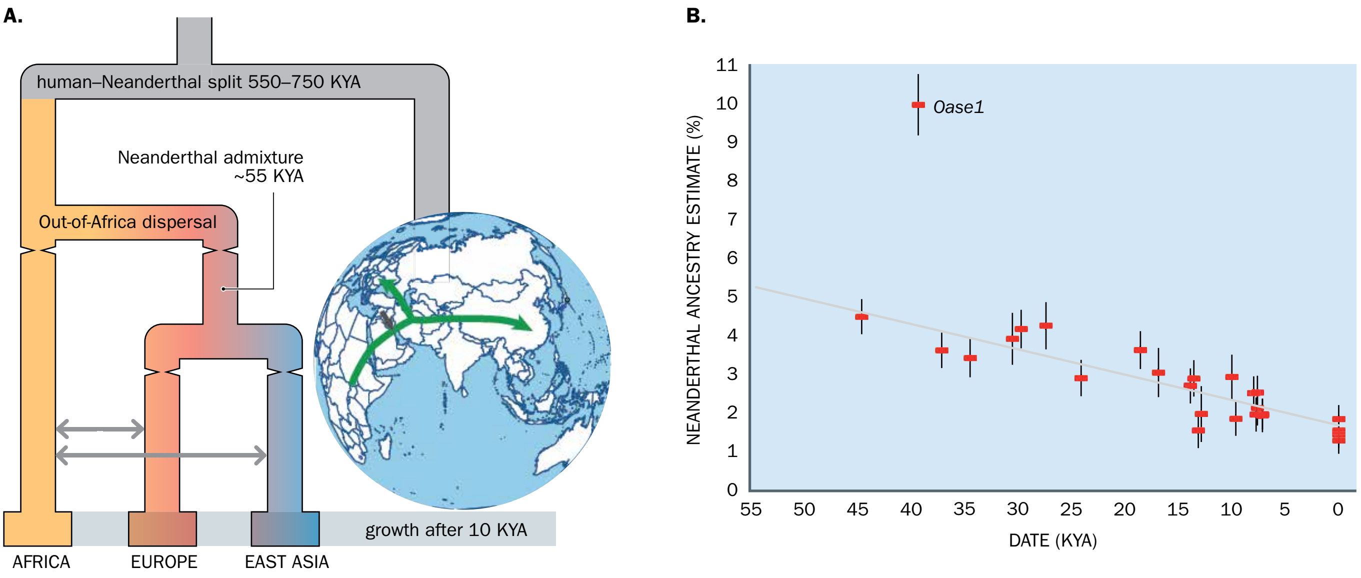


Figure 14.12 Simple Out-of-Africa model with Neanderthal admixture. (**A**) Tree relating the African, European, and East Asian population histories. Population bottlenecks and growth are shown with variable branch widths. Gene flow between populations is indicated by arrows. (**B**) Estimates of Neanderthal ancestry in Eurasian skeletal remains from the last 45,000 years. The outlying sample *Oase1* has additional Neanderthal admixture. KYA, thousand years ago. (B, adapted from Fu Q *et al*. [2016] *Nature* **534**:200–205; PMID 2713593. With permission from Springer Nature. Copyright © 2016.)

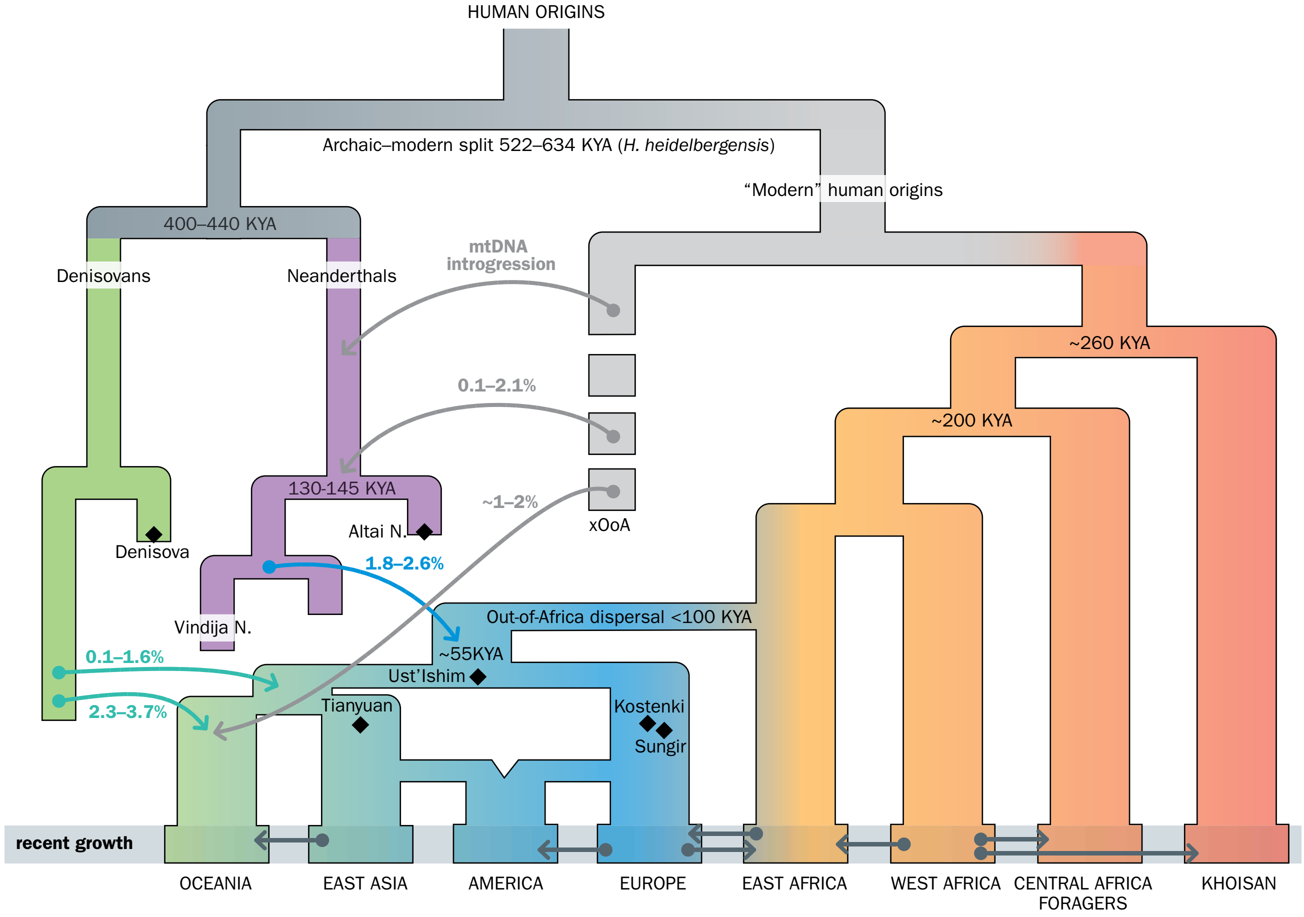


Figure 14.13 One model of the complex admixture patterns among *Homo* species, with two Out-of-Africa dispersals and low levels of gene flow between archaic and modern humans. Arrows indicate the direction of gene flow in cases of admixture. Blue arrow indicates Neanderthal admixture, which in non-African populations has been dated to 55,000 years ago. Green arrows highlight Denisovan contributions to East Asian and Oceanian populations and gray arrow the genetic contributions from an extinct African lineage (xOoA) into the Neanderthals from the Altai region and to Oceanians. KYA, thousand years ago; N., Neanderthal. (Adapted from Kuhlwilm M *et al*. [2016] *Nature* **530**:429–433; PMID 26886800; Lazaridis I *et al*. [2016] *Nature* **536**:419–424; PMID 27459054; and Pagani L *et al*. [2016] *Nature* **538**:238–242; PMID 27654910.)

only to the contribution to the Neanderthals from the Altai region or whether genetic variation derived from this population has also survived in **Oceanic** populations. These additional admixture events contribute to a complex picture of the genetic history of our species where admixture between distant lineages is the norm. Nevertheless, the vast majority of the genetic variation seen in present-day human populations is consistent with the simple Out-of-Africa origins model.

14.3 INFERRING FEMALE AND MALE HISTORIES USING MITOCHONDRIAL DNA AND THE Y CHROMOSOME

The preceding section presented evidence on human origins and pre-history based on a consideration of the 98% of our genome that is inherited from both parents—the autosomes and X chromosome. However, there are two additional segments that are uniparentally inherited: mitochondrial DNA (mtDNA) is passed only from mothers to their children, and the male-specific region of the Y chromosome (MSY) is passed only from fathers to their sons.

Despite their relatively small size, these two segments have been examined particularly intensively in human evolutionary studies. This is partly because of technical aspects that made them relatively easy to study and understand, but also because their sex-specific inheritance can provide unique insights into the different behaviors of women and men in past events such as migrations and colonizations.

There are many differences between these two segments of DNA aside from their modes of inheritance, and these shape how they can be used, and how their patterns of diversity are interpreted (**Table 14.3**). In general, sequence variants are analyzed either by DNA sequencing or targeted genotyping, and result in haplotypes (variants linked on the same DNA molecule). Sets of related haplotypes are often considered together as **haplogroups**.

TABLE 14.3 COMPARISON OF mtDNA AND THE MALE-SPECIFIC REGION OF THE Y CHROMOSOME

Property	mtDNA	MSY
Copy number/cell	1000s	1 in males, 0 in females
Analyzable length (kb)	16.5	~10,000
Gene number	37	~80
Type and known numbers of variants	SNPs (~10,000)	SNPs (>60,000), STRs, (>4500), indels (~1500), CNVs (>100)
Relative mutation rate	10–100 × faster than the nuclear genome average	2× faster than autosomes, due to male bias in mutation

SNP, single nucleotide polymorphism; STR, short tandem repeat; CNV, copy number variant.

mtDNA and MSY are each a single genetic locus. Each is passed intact from generation to generation, without the re-shuffling process of crossing over, so (unlike autosomal DNA) the diversity that we observe in populations results only from mutations. This simplifies the process of interpretation, but it is important to recognize that analyzing mtDNA or MSY in human populations tells us only about a single ancestral lineage (maternal or paternal), and ignores the many other ancestors that each of us has. As was pointed out in Chapter 12, as we move back into the past, each additional generation doubles the potential number of our ancestors (two parents, four grandparents, eight great-grandparents, and so on). By about 1000 years ago, each of us has a theoretical ~8.5 billion ancestors, assuming a 30-year generation time—more than the entire human population today. In fact, many of these ancestors were the same people, so the true number of ancestors is much smaller, but this serves to illustrate what a small proportion of our deep ancestry is captured by mtDNA or MSY.

If we consider a single mating couple as a microcosm of the human species, they carry between them four versions of each autosome, three versions of the X chromosome, two versions of mtDNA (only one of which can be inherited by the next generation), and a single MSY. This increases the chance that the frequencies of mtDNA and

MSY haplotypes will change from generation to generation thanks to sampling effects (genetic drift).

In principle, haplotype frequencies can also be affected by natural selection. Negative selection is certainly at work on mtDNA and MSY, since all mtDNA-coded genes are essential for ATP generation, and mutations in MSY can cause reduced reproductive success via lowered fertility. However, negative selection that removes new deleterious variants is not likely to have a significant effect on the frequencies of specific haplotypes in different populations. Positive selection would be potentially more significant. For example, if a variant arose on a particular mtDNA or MSY sequence that increased the probability of offspring surviving and reproducing, then the entire haplotype carrying that variant would increase in frequency through a selective sweep. Some studies have suggested that particular mtDNA haplogroups have been positively selected in the past for resistance to sepsis, or adaptation to cold climates, but these observations could also be explained by other, nonselective effects. For MSY there are no known examples of positive biological selection. Generally, population studies assume that patterns of diversity can be explained by population history, and that mtDNA and MSY are behaving in a selectively neutral way.

Mitochondrial DNA and MSY trees are rooted in Africa

Mitochondrial DNA molecules from thousands of humans across the world have now been studied, and none of these are more closely related to the distinct mtDNA sequences found in Neanderthals or Denisovans than they are to any modern human sequence. Similarly, MSY sequences in thousands of modern humans can be compared with the Neanderthal MSY (no Denisovan sequence is available, because the finger bone from which DNA was extracted belonged to a female). Here again, the Neanderthal sequence is distinct from that of all modern humans. These findings suggest that, following the admixture between humans and archaic hominins, archaic mtDNA and MSY sequences have been lost and become extinct through genetic drift or negative selection.

Sequencing whole mtDNA molecules in different human populations defines mtDNA haplogroups that can be arranged into a unique phylogenetic tree (**Figure 14.14**). The tree is rooted by comparison with the chimpanzee mtDNA sequence. The strong genetic drift experienced by mtDNA means that haplogroups in the tree tend to show high levels of geographical differentiation, so that African branches can generally be distinguished from non-African ones. The root of the tree and its deepest-rooting branches lie in sub-Saharan Africa, suggesting an African origin for modern mtDNA sequences. Because variants are ascertained in an unbiased way by sequencing, the lengths of branches are roughly proportional to time. With the help of a measure of the mtDNA mutation rate, we can estimate that the most recent common ancestor (MRCA) of all modern mtDNAs (sometimes known, by Biblical analogy, as “mitochondrial Eve”) lived about 190,000 years ago. A similar approach can be taken to the MSY. The tree (see **Figure 14.14A**) again has a sub-Saharan African root, implying an African origin for modern MSY sequences, though here the age of the MRCA (sometimes “Y-chromosomal Adam”) is older at ~250,000 years.

Absolute estimates of ages in phylogenetic trees are inherently imprecise due to uncertainty about rates and modes of mutation, and, often, generation times. Nonetheless, the differences in the times to most recent common ancestor (TMRCAs) of the mtDNA and MSY trees should not be surprising—these are independent loci, with independent histories that are susceptible to sex-specific effects. In western lowland gorillas, for example, the TMRCA of the mtDNA tree is ~290,000 years, but that of the MSY tree only ~60,000 years, reflecting the alpha-male mating behavior of gorilla groups.

As well as their African roots, the mtDNA and MSY trees each contain a single node from which all non-African lineages descend, which supports a single source for all non-Africans. The ages of these nodes are subject to the same uncertainty as that of the root, but both are around 60,000–70,000 years, compatible with other evidence for the timing of the Out-of-Africa migration. Both mtDNA and MSY also show evidence of expansions in non-African lineages 40,000–50,000 years ago (see **Figures 14.14C** and **D**).

Insights from comparisons between mtDNA and MSY

As the section above describes, the two uniparentally inherited segments of the genome show broad similarities in their phylogenies that reflect and support the Out-of-Africa model for human origins. However, there are also interesting differences that tell us about contrasts in female and male behaviors. These can be shown by demographic reconstructions that trace the female and male effective population sizes back through time in the form of graphs. **Figures 14.14C** and **D** show these reconstructions for the mtDNA and MSY trees; each is based on sequencing in the same 320 male individuals.

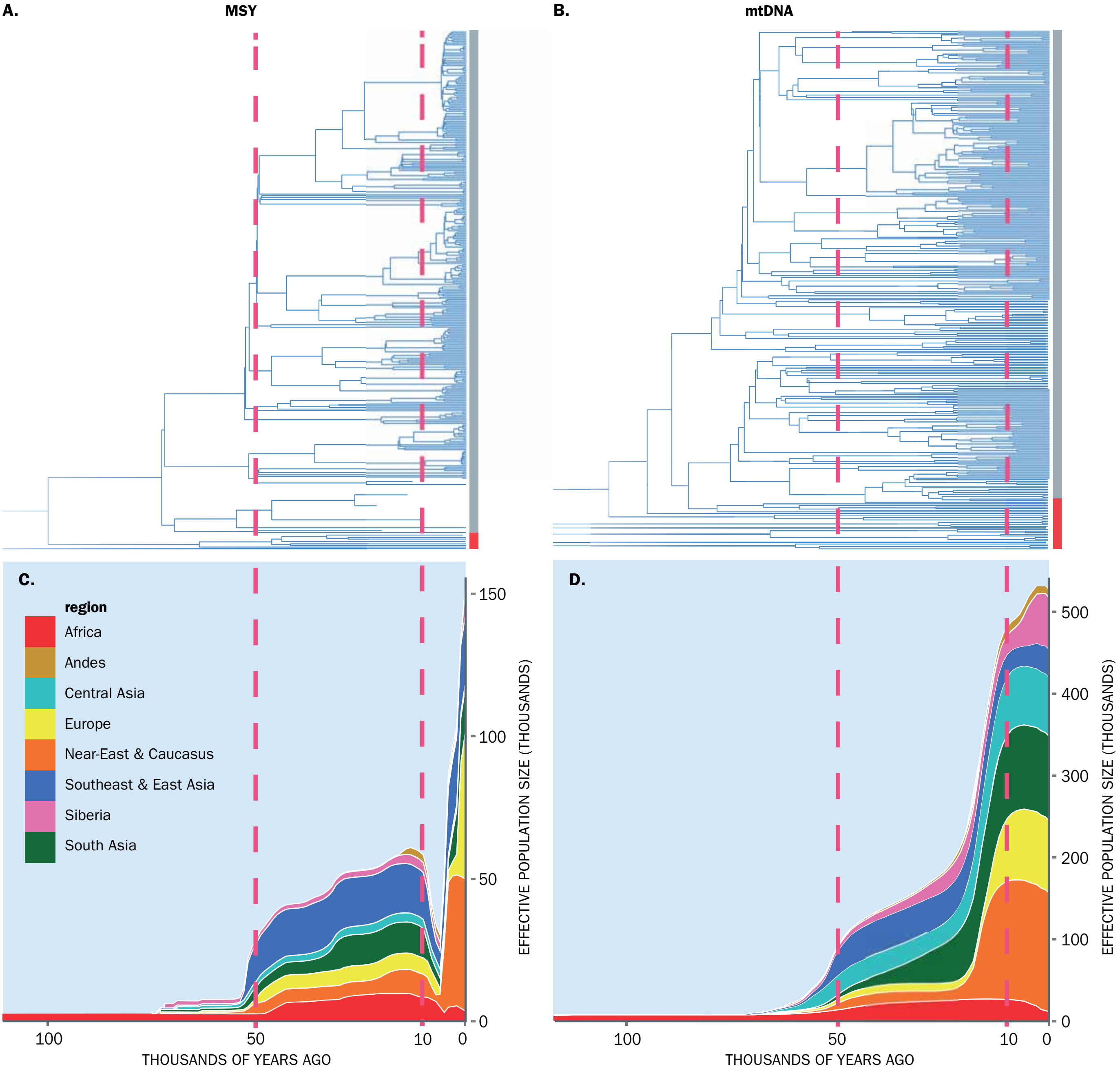


Figure 14.14 MSY and mtDNA trees for global samples, and the underlying demographic histories. Phylogenetic trees for (**A**) MSY and (**B**) mtDNA, showing sub-Saharan African-specific lineages at the roots (red bars). Note that these trees are truncated and the deepest branches are not shown. (**C, D**) Plots of effective population size against time, with different world regions indicated by colors. Note the different vertical scales in parts **C** and **D**. (Adapted from Karmin M *et al*. [2015] *Genome Res* **25**:459–466; PMID 25770088.)

Both show increases in N_e at ~40,000–60,000 years ago, reflecting the Out-of-Africa expansions. However, they differ in two important features. First, the N_e estimates for mtDNA are consistently more than twice as high as those for MSY; this could be interpreted as evidence for polygyny—a system in which men have more than one female partner. Second, the MSY graph shows a reduction at around 8000–4000 years ago, when female N_e is up to 17-fold higher than male N_e, followed by an increase. This is best explained by cultural changes that led to sex-specific changes in variance in offspring number. The drop in male N_e corresponds to a change in the archeological record characterized by the spread of new cultures, demographic changes, and shifts in social behavior. Other studies, including analysis of the 1244 Y chromosomes in the 1000 Genomes Project samples, document abundant and extreme male-specific expansions in all continents in the last 15,000 years. The earliest correspond to the peopling of the Americas, but the others are in the same 8000–4000-years-ago time period and followed innovations that may have led to increased variance in male reproductive success; for example, developments in metalworking, invention of the wheel, horse-riding, and organized warfare. Social stratification associated with these changes could have allowed privileged male lineages to undergo preferential amplification for many generations. A historical example of this could be the expansion of the Mongol empire in the last millennium—a frequent MSY haplotype is today spread across the territory of Genghis Khan's empire, and may reflect his numerous descendants.

Admixture, the mixing together of populations from different sources, has been a universal aspect of human history and pre-history. During the last 600 years, following the so-called age of exploration, admixture has involved populations that were previously widely separated. Following a past admixture event between two populations (for example, an African and a European population), descendants carry autosomal DNA from each, and the proportion of each contribution can be estimated based on a knowledge of the allele frequencies in modern proxies from Africa and Europe.

Often, however, the contribution of men and women from each of the mixing populations is unequal. This can result from one population being of predominantly one sex only. Many cases originate in the colonial era: for example, the European soldiers, explorers, missionaries, and traders who traveled the world over the last few centuries were mostly men, and so descendants of admixed colonial populations tend to carry a substantial proportion of European Y chromosomes, but little or no European mtDNA. Sex-biased admixture can also result when the admixing populations are not themselves sex-biased, but social rules act to bias mating behaviors. The colonial situation again illustrates this, because the mixing populations rarely had equal status, and males from the colonizing population tended to mate with females from the colonized. Figure 14.15 illustrates a number of examples from the Americas showing the results of sex-biased admixture among populations with at least three different continental origins—indigenous Native Americans, colonizing Europeans, and transported African slaves.

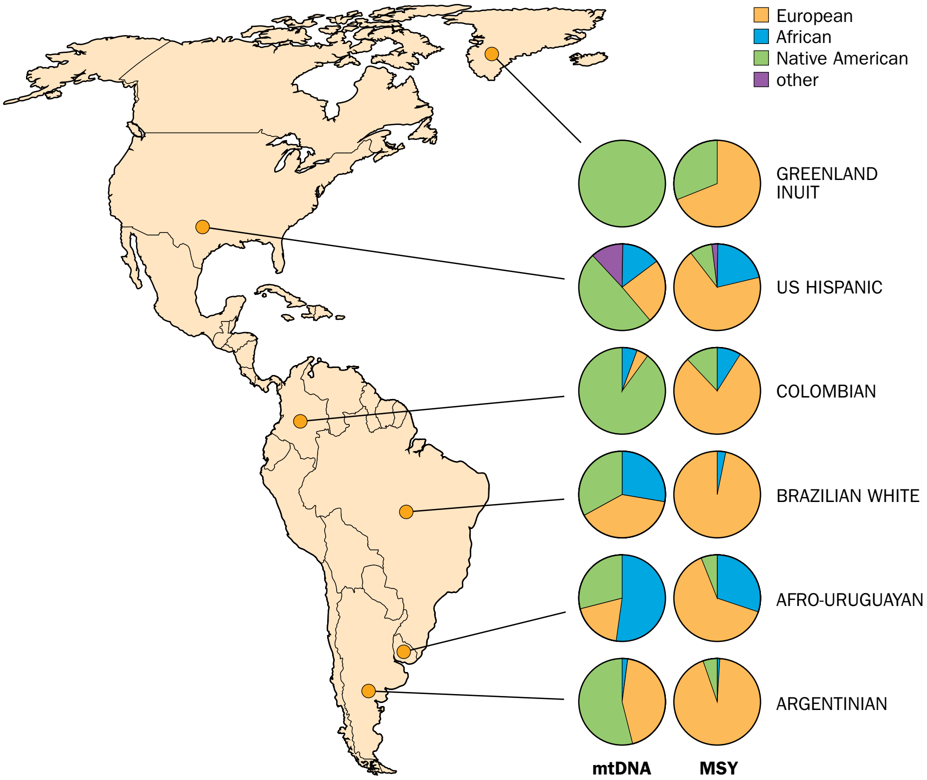


Figure 14.15 Sex-biased admixture in six populations of the Americas, shown by Y-chromosomal and mtDNA ancestry proportions. (Data from Sans M *et al.* [2002] *Am J Phys Anthropol* **118**:33–44; PMID 11953943; Alves-Silva J *et al.* [2000] *Am J Hum Genet* **67**:444–461; PMID 10873790; Carvalho-Silva DR *et al.* [2001] *Am J Hum Genet* **68**:281–286; PMID 11090340; Rojas W *et al.* [2010] *Am J Phys Anthropol* **143**:13–20; PMID 20734436; Lao O *et al.* [2010] *Hum Mutat* **31**:E1875–E1893; PMID 20886636; Corach D *et al.* [2010] *Ann Hum Genet* **74**: 65–76; PMID 20059473; Saillard J *et al.* [2000] *Am J Hum Genet* **67**:718–726; PMID 10924403; Bosch E *et al.* [2003] *Hum Genet* **112**:353–363; PMID 12594533.)

Sex-biased admixture is most clearly seen via analysis of the uniparentally inherited mtDNA or MSY. However, these are not the only parts of the genome that show sex-biased inheritance. The X chromosome is inherited twice as often from mothers as from fathers, so past male-biased admixture is detectable from an increased X-chromosomal contribution from the female-biased population. An example is seen in islands of the Caribbean, which show skews toward increased proportions of Native American and African X-chromosomal contributions compared to autosomes.

14.4 HEALTH CONSEQUENCES OF OUR EVOLUTIONARY HISTORY

Understanding the role of natural selection in shaping humans provides evidence of how the environment has influenced our species, and may explain some of the diversity of human phenotypes observed across the world. It also informs medical genetics, since it can explain why some variants that are associated with disease have reached high frequency: in the past they had some beneficial effect in the context of a different environment or lifestyle.

As discussed above, humans have had a small population size throughout almost all of our history. Given this, we would expect human genomic variation to be shaped primarily by genetic drift and selection against severely deleterious alleles, rather than mild natural selection, because population-genetic theory tells us that in small populations, natural selection needs to be very strong to prevail over genetic drift. To identify regions that have experienced natural selection, we can analyze patterns of variation at particular genes that are very different from the genome-wide average, and also very different from the population-genetic theoretical expectation in the absence of selection. An example was given in a previous chapter (see Section 11.4 and **Box 11.3**), where a variant that causes pale skin has undergone positive selection in Europeans, resulting in a very high frequency of a variant in Europeans that is rare or absent elsewhere, by a process called a **selective sweep**.

Ubiquitous negative selection affects human genomic variation

Negative selection removes alleles that decrease the **fitness** of an individual from the population, a process that Darwin called “rejection of injurious variations”. It is the evolutionary force that keeps the frequency of most severe genetic diseases low, as new disease-causing variants generated by *de novo* mutation are rapidly removed, since the disease caused by the variant reduces the chance of viable offspring from that individual, either due to early death or sterility. Such deleterious alleles are in mutation–selection balance, a process described in an earlier chapter (see Chapter 12), and are of course a major focus of clinical genetics. The extent of linkage disequilibrium in the genome means that adjacent neutral variants are also removed by negative selection, and this form of selection can therefore influence much of the genome.

Because selection needs to be strong to overcome genetic drift and alter allele frequencies, we would expect that in small isolated populations, where genetic drift is strongest, there will be reduced negative selection removing mildly deleterious variants from the population. Analysis of the genomes of individuals from several small isolated populations, such as those of the islands of Orkney (off the north coast of Scotland) and Crete (in the eastern Mediterranean), has shown that they have a greater proportion of very rare missense variants (which change an amino acid) compared to closely-related larger, nonisolated populations. By analyzing variants in genes that were observed only once (singletons), the ratio of missense variants to synonymous variants (which do not change an amino acid) was higher in the small isolated populations compared to their nonisolated neighbors. The ratio ranged from an average of ~1% higher, when comparing the inhabitants of Orkney and mainland Britain, up to almost 50% higher when comparing population isolates from the valleys of the Friuli Venezia Giulia region to other Italians. This shows that negative selection struggles to remove such variants from these isolated populations.

When populations experience other demographic changes, such as strong population bottlenecks, genetic drift can increase the frequency of some disease-causing variants so that the disease is rather frequent in the population (**Table 14.4**; see Chapter 12 and **Figure 12.7**). A much-studied example is the “Finnish disease heritage,” where the Finnish population has a unique constellation of genetic diseases that have risen to high frequency because of demographic events in the past.

Positive selection of advantageous variants

Identifying selective sweeps by comparing genomes from different populations can help to identify genes and individual variants that have been under selection in the past. Selective sweeps may result in unusually long haplotypes, as discussed previously, or in unusually high allele-frequency differences between populations, or may be recognized by more complex statistics. Nevertheless, such approaches will not identify all cases of selection, as many, perhaps most, selective events will act on multiple variants influencing the selected phenotype. These selective events will leave a signature on genomic variation that is more difficult to recognize using current approaches. It is also important to remember that apparent signatures of selection can be generated by chance, in

TABLE 14.4 SOME POPULATION ISOLATES USED IN GENETIC STUDIES

		Example of a mapped disease gene			
Genetic isolate	**Country**	**Disease**	**OMIM #**	**Gene symbol**	**Reference**
Finnish	Finland	Congenital chloride diarrhea	214700	*SLC26A3*	Hööglund P *et al.* (1996) *Nat Genet* **14**:316–319; PMID 8896562
Jewish communities	Various	Familial dysautonomia	223900	*IKBKAP*	Slaugenhaupt SA *et al.* (2001) *Am J Hum Genet* **68**:598–605; PMID 11179008
Old Order Amish	USA	Ellis–van Creveld syndrome	225500	*EVC*	McKusick VA (2000) *Nat Genet* **24**:203–204; PMID 10700162
Hutterite	USA	Bowen–Conradi syndrome	211180	*EMG1*	Armistead J *et al.* (2009) *Am J Hum Genet* **84**:728–739; PMID 19463982
Sardinian	Italy	Early-onset Parkinson disease	605909	*PINK1*	Valente EM *et al.* (2004) *Science* **304**: 1158–1160; PMID 15087508
Margarita Island	Venezuela	Cleft lip/palate-ectodermal dysplasia	119530	*CLPED1*	Suzuki K *et al.* (2000) *Nat Genet* **25**:427–430; PMID 10932188
Costa Rica central valley	Costa Rica	Inherited deafness	124900	*DIAPH1*	Lynch ED *et al.* (1997) *Science* **278**: 1315–1318; PMID 9360932

the absence of any selection. Therefore methods designed to identify selection using genomic signatures will also identify some regions as false positives.

For most selective sweeps, the actual reason for the selection remains unclear. This section will focus on two well-characterized examples where both the genetic and functional evidence for a selective sweep are particularly clear—the lactase (*LCT*) gene in populations that drink fresh milk and the *EPAS1* gene in high-altitude Himalayan populations living in conditions of permanent hypoxia.

Lactase persistence and the consumption of fresh milk

In Europeans, the strongest signal of a selective sweep is in the genomic region containing the lactase gene (*LCT*). Lactase is an enzyme, produced by cells lining the small intestine, that digests the disaccharide lactose into its two constituent monosaccharides, glucose and galactose. These two sugars are then directly absorbed by the intestine. In almost all mammals, lactose is present in the mother's milk and is an important source of energy for the infant. Lactase is present at high levels in the infant small intestine, but is lost after weaning when transcription of the gene is repressed. This occurs in all mammals and most humans; but in some humans, variants in an enhancer 14 kb upstream of *LCT* disrupt this repression. This causes the lactase enzyme to be present in the small intestine throughout adulthood. This phenotype is called lactase persistence, and the more common phenotype of absence of lactase in adulthood is called lactase nonpersistence.

Different human populations have different frequencies of lactase-persistent individuals. The phenotype is common in Europeans, particularly in north-western Europeans where >90% of people are lactase persistent, and in specific populations of Africa and Arabia (**Figure 14.16**). These populations are not especially closely related to each other, but they do share a culture of drinking fresh milk from cattle, camels, or goats. Lactase persistence is therefore a genetic adaptation to the culture of drinking fresh milk as adults, enabling those adults to digest milk effectively, avoiding unpleasant symptoms and gaining an important source of water and nutrition.

Several different variants in the lactase enhancer are known to cause lactase persistence, but each one is present only in a restricted number of populations (see **Figure 14.16**). This suggests that lactase persistence arose several times independently, and is an example of convergent evolution. It also suggests that the variants have a relatively recent origin. However, it is not clear whether these variants evolved before or after animal domestication and adoption of a milk-drinking culture.

Evidence for natural selection acting on lactase persistence has accumulated over several decades. The phenotype was first discovered because lactase nonpersistent individuals showed symptoms of lactose intolerance—including stomach pain, diarrhea, and flatulence—after drinking large quantities of fresh milk. Because lactase persistence

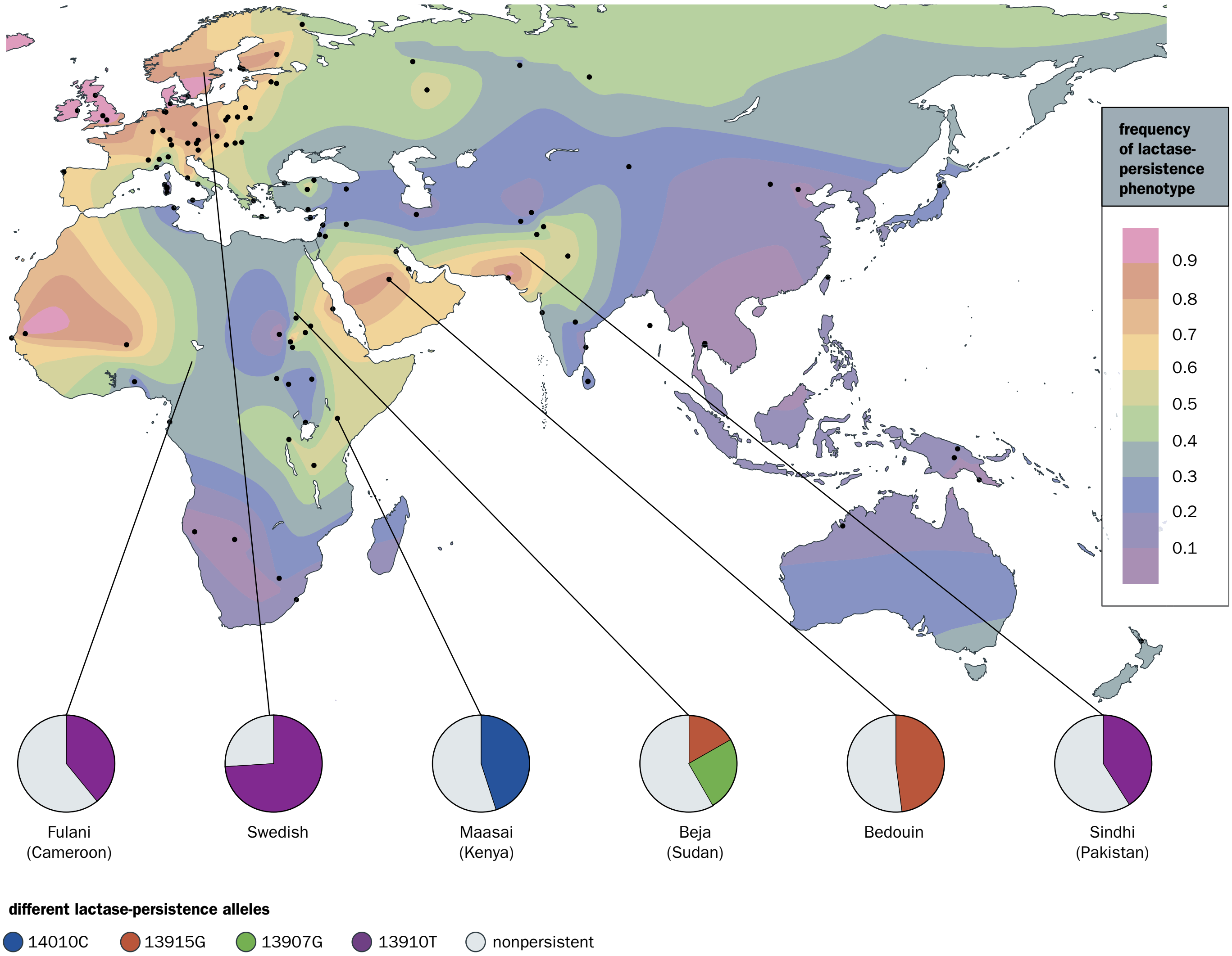


Figure 14.16 Lactase persistence in different populations. The map, together with the key on the right, shows the frequency of the lactase-persistence phenotype, with sampling locations shown as black dots. The pie charts, together with the key below, show the frequencies of different lactase-persistence alleles in different populations. (Adapted from Itan Y *et al*. [2010] *BMC Evol Biol* **10**:36; PMID 20144208. Published under CC BY 2.0)

is the common trait in Europeans, it was regarded as the medically "normal" phenotype, and lactase non-persistence as the dysfunctional one. However, the unpleasant symptoms of lactose intolerance only show themselves in lactase nonpersistent homozygotes living in a culture where drinking fresh milk is common. This is a simple example of how a medically important trait with a simple genetic basis can still be influenced by the environment.

Adaptation to high altitude

Comparison of the genomic variation between two populations that are genetically closely related yet live in very different environments can identify alleles that have changed in frequency due to a particular selection pressure. An example is the comparison between Han Chinese living at low altitudes and Tibetans living at high altitudes on the Tibetan Plateau. Both a comparison of allele frequency differences (Figure 14.17A) and haplotype structures showed that a haplotype including the *EPAS1* gene had been driven to high frequency in Tibetans. The frequency of this haplotype correlates with the altitude of diverse populations in the Himalayas (Figure 14.17B). *EPAS1* encodes a transcription factor that acts on the **hypoxia-inducible factor** pathway, but exactly how the Denisovan haplotype favors high-altitude living is unclear. Furthermore, the functional variant is not known with certainty, so more work is needed to establish a direct functional link between *EPAS1* genotype and high-altitude adaptation.

The *EPAS1* haplotype favored in high altitudes differs substantially from other human *EPAS1* haplotypes, but closely resembles a haplotype present in the genome of the

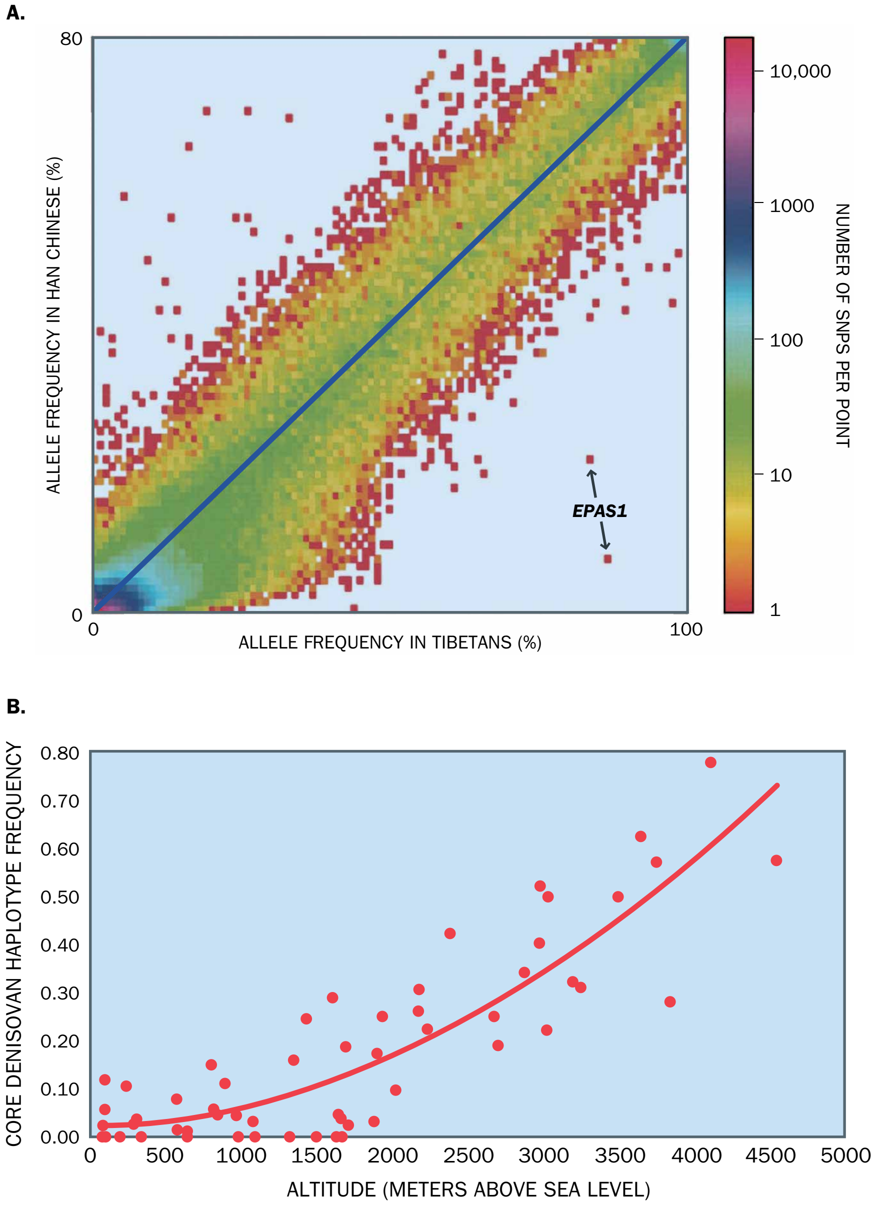


Figure 14.17 Selection of the *EPAS1* gene at high altitude. (**A**) Each point on the graph corresponds to a derived allele frequency in Tibetans (x axis) and Han Chinese (y axis), with the number of SNPs showing those particular two allele frequencies indicated by the color of the point (legend on the right of the graph). Derived alleles of two SNPs in *EPAS1* highlighted on the graph are at unusually high frequency in Tibetans and low frequency in Han Chinese. (**B**) Correlation between altitude and frequency of the adaptive *EPAS1* haplotype, which is similar to that seen in Denisovan. Each circle represents a population sample genotyped in this study. (A, from Yi X *et al.* [2010] *Science* **329**:75–78; PMID 20595611. Reprinted with permission from the AAAS; B, from Hackinger S *et al.* [2016] *Hum Genet* **135**:393–402; PMID 26883865. Published under CC BY 4.0.)

archaic Denisovan. This implies that the adaptive haplotype evolved in Denisovans and spread into modern humans by interbreeding. This is a mechanism of adaptation ("adaptive introgression") whose importance is only beginning to be appreciated.

The importance of infectious disease in human evolution

Until the last one hundred years, infectious diseases were a major cause of human mortality everywhere, and particularly in children. Indeed, in many parts of the world this remains the case. It is likely that infective agents have influenced genomic variation particularly at immunity genes, as variants that provided resistance would have been strongly favored.

In Chapter 11 the extraordinary diversity of the human major histocompatibility complex, which encodes proteins involved in the response to pathogens, was explained by balancing selection. In this form of selection, heterozygotes are favored because they confer protection against a wider range of pathogens. However, there are only a few well-characterized examples of the influence of infectious diseases on the human genome. This may be because most bacterial and viral pathogens evolve rapidly, so that the pathogen that was an agent of natural selection in the past is not present today, or because different pathogens exert different selective pressures, obscuring the genomic signal. There are exceptions, including eukaryotic protist pathogens, particularly those that cause malaria, perhaps because they evolve more slowly than prokaryotes and viruses and have killed many people over a long period.

Sickle cell anemia and malaria

Malaria, caused in Africa predominantly by the protist *Plasmodium falciparum*, is a febrile disease spread by mosquitoes that can have neurological symptoms and severe complications, and causes high levels of mortality, particularly in children. There are several examples of genetic variants that have been under strong selection by malaria in malaria-endemic regions, but we will focus here on just one, which was discovered in 1954 yet remains an excellent example of the power of malaria as a selective agent, and its tangible effect on medical genetics.

Sickle cell anemia is a severe anemia with characteristic **sickling** of red blood cells, and has a high mortality rate in children, particularly due to septicemia. It is caused by homozygosity of a variant allele at the beta-globin gene (*HBB*), which changes a glutamic acid to a valine at the sixth amino acid of the protein, which is known as HbS. Hemoglobin is a tetramer of two alpha-globin and two beta-globin molecules, which transports oxygen in the blood; HbS molecules tend to aggregate when deoxygenated, resulting in fibers composed of multiple long strands of HbS tetramers, and in the sickle cell phenotype. Because the deformed sickle cells have a much shorter life span than normal red blood cells, the body cannot replace dead red blood cells fast enough, and anemia results. The long HbS fibers also block small blood vessels, causing hypoxic tissue damage.

Despite the morbidity of sickle cell anemia, the HbS allele frequency is remarkably high in many African populations where malaria is endemic, reaching frequencies as high as 15% (**Figure 14.18**). This is because heterozygotes for the HbS allele, who do not develop sickle cell anemia, are protected against the most severe malarial symptoms.

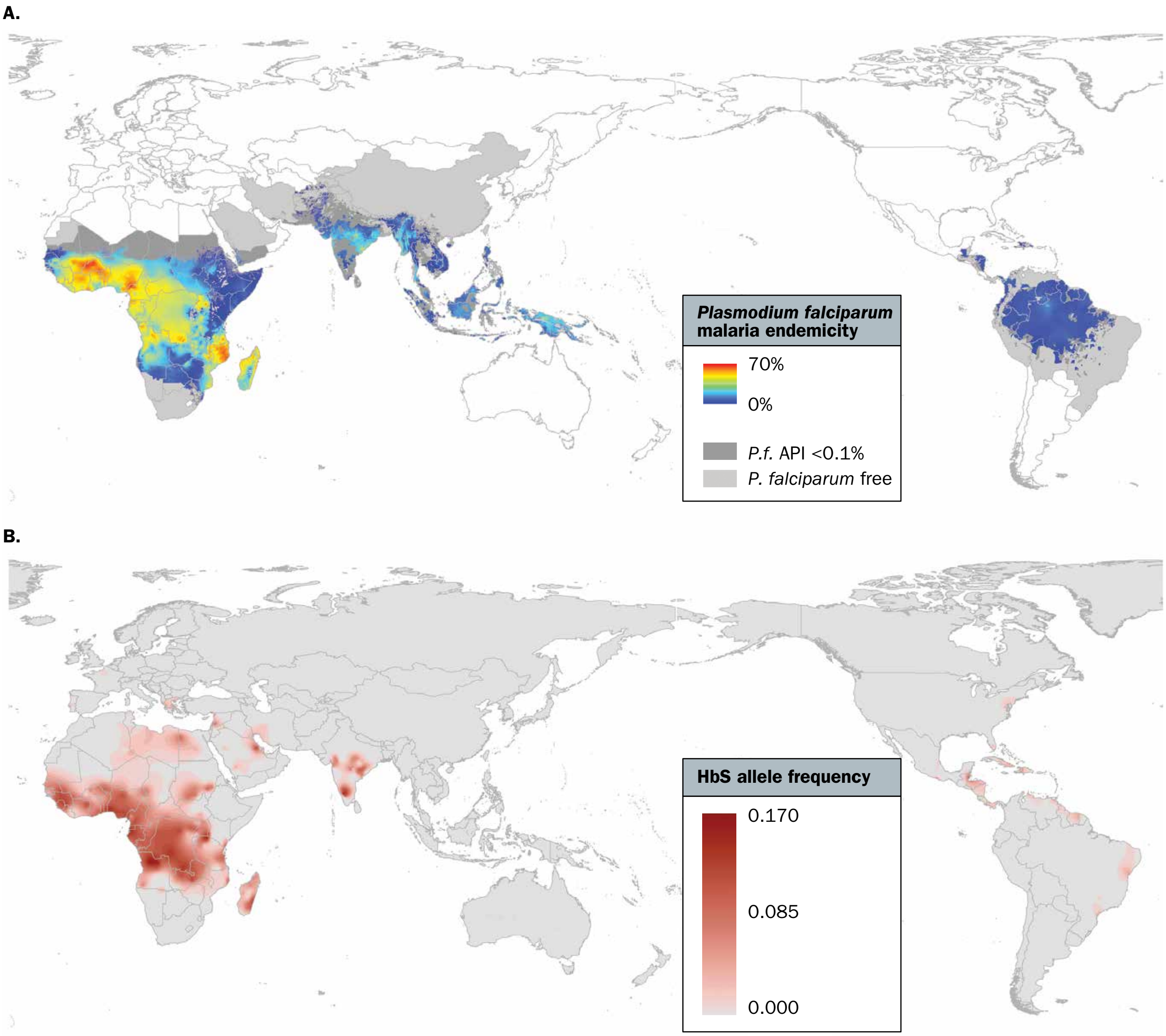


Figure 14.18 The overlapping distributions of malaria and sickle cell anemia. (**A**) Current geographical distribution of *Plasmodium falciparum* (*P.f.*) malaria endemicity. API, annual parasite incidence. (**B**) Frequency distribution of the sickle cell hemoglobin allele, HbS. (Courtesy of Fred Piel and Simon Hay of the Malaria Atlas Project.)

The HbS allele is therefore maintained by a particular type of balancing selection called heterozygote advantage. In the absence of malaria, the heterozygotes lose their advantage and the allele is therefore gradually removed from the population by negative selection acting against the deleterious sickle cell anemia phenotype.

APOL1 and kidney disease

Kidney failure is about four times more frequent in African Americans than in other Americans. Part of this increased susceptibility is due to the increased frequency of two risk haplotypes of the apolipoprotein L1 (*APOL1*) gene. An individual carrying two risk haplotypes is over 10 times more likely to develop kidney disease than an individual carrying one or no risk haplotype. One risk haplotype carries a glycine at position 342 and a methionine at position 384, while the other risk haplotype has a 6 bp deletion that removes two amino acids at positions 388 and 389 (**Figure 14.19**). How these two risk haplotypes (known as G1 and G2, respectively) increase the risk of kidney disease remains unclear.

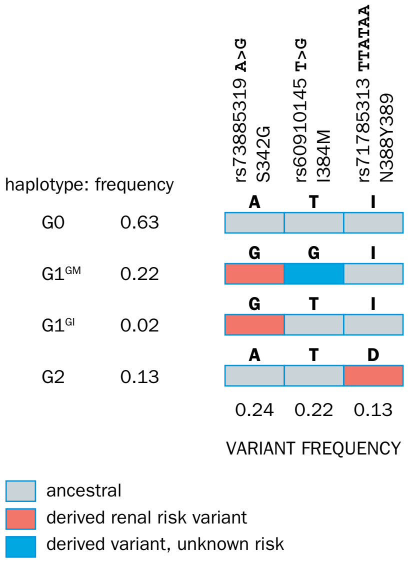


Figure 14.19 *APOL1* haplotypes in African Americans. Haplotypes G1^{GM} and G1^{GI} are alternative versions of the G1 haplotype. (Figure adapted from Limou S *et al.* [2014] *Adv Chronic Kidney Dis* **21**:426–433; PMID 25168832. With permission from Elsevier.)

Before the discovery of the association with kidney disease, ApoL1 was known as an important part of the immune response against trypanosomes, protist pathogens that cause sleeping sickness. ApoL1 molecules carrying the risk alleles for kidney disease are able to lyze a particular strain of trypanosome, *Trypanosoma brucei rhodesiense,* but not another strain, *T. brucei gambiense*. The variants carried by the G1 and G2 haplotypes are within a domain of ApoL1 that binds the serum resistance-associated protein (SRA). SRA is a protein produced by the *rhodesiense* strain that inhibits the antitrypanosome activity of ApoL1, so it is likely that these variants disrupt the binding of trypanosome SRA to host ApoL1. *T. brucei rhodesiense* causes sleeping sickness in East Africa but not in West Africa, where *T. brucei gambiense* is the causative pathogen. The *APOL1* G1 and G2 haplotypes are present throughout Africa but surprisingly are at highest frequencies in West Africa (**Figure 14.20**). One theory to explain this is that *T. brucei rhodesiense* may have existed in West Africa in the past but high frequencies of G1 and G2 have been effective at eliminating this strain, leaving only the *gambiense* strain today. Alternatively, distribution of *T. brucei* strains may have been altered by changes in the distribution of its vectors, different species of tsetse fly.

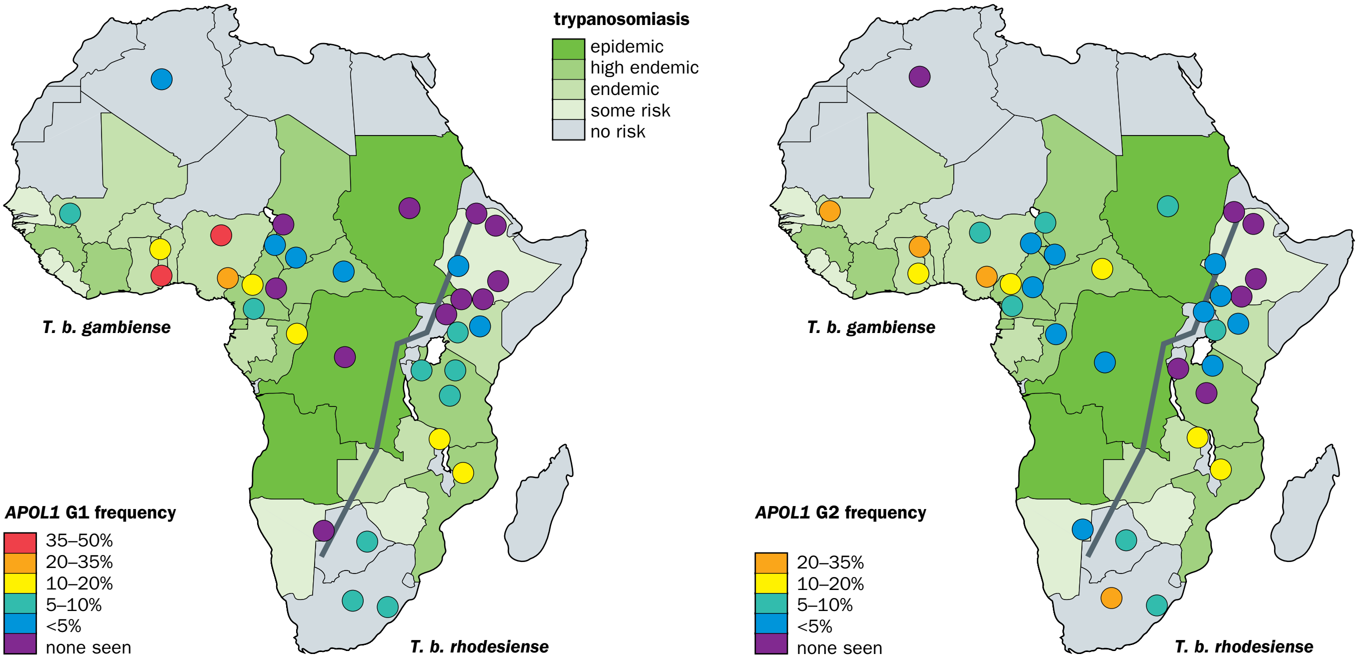


Figure 14.20 Geographical distribution of *APOL1* haplotypes and *Trypanosoma brucei* subspecies. The frequencies of the *APOL1* haplotypes G1 and G2 are shown, together with the population endemicity of trypanosomiasis. Each circle represents a human population sampling location. The Great Rift Valley is shown as a line running from southwest to northeast, with *T. brucei (T. b.) rhodesiense* to the east and *T. brucei gambiense* to the west. (Figure from Limou S *et al.* [2014] *Adv Chronic Kidney Dis* **21**:426–433; PMID 25168832. With permission from Elsevier.)

Taking these data together, this example is potentially another instance of balancing selection, where homozygotes for the *APOL1* G1 or G2 haplotypes are at a greatly increased risk of kidney disease in later life, yet these deleterious haplotypes are maintained in populations endemic for sleeping sickness by their protective effect against trypanosomes.

Evolutionary genetics of inflammatory disease

We have seen that increased resistance to infectious disease provides a strong selective force that can maintain at high frequencies variants that can cause other diseases. There is also evidence that selective pressure by infectious diseases has caused an increase in alleles that increase the risk of common inflammatory diseases such as Crohn disease. Like all common complex diseases, individual risk alleles each have a small effect on an individual's risk of disease but, because there are many risk alleles, together they contribute significantly to the overall risk of disease.

An example is a variant in the *FUT2* gene that generates a premature stop codon leading to a loss of function. *FUT2* encodes fucosyltransferase 2, a glycosylation enzyme that regulates the expression of ABO blood group antigens on gut mucosal surfaces and nonblood body fluids. The variation encodes the secretor phenotype, where individuals who have the allele that generates an active *FUT2* gene have ABO antigens on nonblood body fluids and gut mucosal surfaces. In contrast, individuals homozygous for the nonfunctioning *FUT2* allele do not produce an active fucosyltransferase 2 enzyme and are called nonsecretors. Homozygote nonsecretors are at twice the risk of developing Crohn disease compared to carriers of other genotypes.

Homozygote nonsecretors are also more resistant to some bacterial and viral infectious diseases. They are completely resistant to the most common strains of norovirus, known in the UK as the "winter vomiting bug," which causes viral gastroenteritis with symptoms including nausea, vomiting, and diarrhea. Nonsecretors are also more resistant to infection by *Helicobacter pylori*, a gut bacterium that can cause gastritis, ulceration, and gastric cancer. It appears likely that resistance to infection was responsible for increasing the frequency of the nonsecretor allele to around 30% in Europeans. Although we cannot be certain that infection by noroviruses and *H. pylori* was the selective force in the past, it is highly likely that it was gastrointestinal infections of some kind.

There is increasing general evidence that selection for alleles that increase resistance to infection has increased the frequency of inflammatory diseases. Crohn disease is often grouped together with ulcerative colitis as inflammatory bowel disease (IBD). Large-scale genetic epidemiological studies find many variants (163 at the latest count) across the genome where an allele subtly increases the risk of IBD (the rationale and design of such studies are discussed in detail in Chapter 18). These IBD-associated variants are very often within or near to genes that are known to be involved in the inflammatory response, suggesting that different alleles modify the function of these genes in some way. Furthermore, the IBD-associated variants are likely to show evidence of positive or balancing selection, compared to randomly-ascertained variants across the genome. This strongly suggests that infectious disease has maintained the frequencies of alleles that in modern environments increase susceptibility to IBD.

SUMMARY

- Humans are a species of great ape, most closely related to chimpanzees and bonobos.
- A number of phenotypic and cultural features, including upright walking and syntactical language, distinguish us from other great apes.
- The appearance of our genus *Homo* around 2 million years ago is associated with findings of increasingly complex stone tools.
- Ancient DNA studies have enabled us to reconstruct the evolutionary relationships with our closest extinct relatives—the Neanderthals and Denisovans.
- The evolution of phenotypic traits unique to our species, such as large brain size, has involved changes in the amino acid sequences of proteins and the copy number and regulatory regions of genes, often leading to the delay of peak gene expression times in development.
- Compared to other great apes, and considering our current census size, humans are characterized by unusually low genetic diversity and low long-term effective population size. This means that all human populations descend from the same small ancestral population.
- The genetic differences among human populations are minor compared to the extent of diversity within groups. Most common DNA variants are shared among continental populations

and the differences among groups are manifested by minor frequency differences or in the distribution of rare variants. This means that genetic data do not support the concept of races.

- The majority of the genetic variation outside Africa derives from a single Out-of-Africa dispersal with minor (<3%) additional contribution from admixture with Neanderthals. Papuan and Australian populations have received a further genetic contribution from Denisovans on top of this.
- Phylogenetic trees drawn from mtDNA and Y chromosome data are rooted in Africa and show a clear distinction between the lineages in modern and archaic humans.
- Comparative analyses of mtDNA and Y chromosome data among human populations show that male effective population sizes have been generally smaller than female effective population sizes and that admixture between human populations has often been sex-specific.
- Analyses of functionally important genes across global populations reveal that purifying selection has been ubiquitous in removal of deleterious variants from our genome.
- As human populations colonized new environments and changed their diet and lifestyle over the last few tens of thousands of years, new local genetic adaptations have occurred. Examples include mutations in lactase-persistence and hypoxia genes.

FURTHER READING

Jobling MA *et al.* (2014) *Human Evolutionary Genetics*, 2nd edn. Garland Science. (Covers in greater detail the material throughout this chapter.)

Human origins

Groucutt HS *et al.* (2015) Rethinking the dispersal of *Homo sapiens* out of Africa. *Evol Anthropol* **24**:149–164; PMID 26267436.

Pakendorf B (2014) Coevolution of languages and genes. *Curr Opin Genet Dev* **29**:39–44; PMID 25170984.

Stringer C (2016) The origin and evolution of *Homo sapiens*. *Philos Trans R Soc Lond B Biol Sci* **371**:20150237; PMID 27298468.

Human evolutionary history from genome sequences

Der Sarkissian C *et al.* (2015) Ancient genomics. *Philos Trans R Soc Lond B Biol Sci* **370**:20130387; PMID 25487338.

Haber M *et al.* (2016) Ancient DNA and the rewriting of human history: be sparing with Occam's razor. *Genome Biol* **17**:1; PMID 26753840.

Nielsen R, *et al.* (2017) Tracing the peopling of the world through genomics. *Nature* **541**:302–310; PMID: 28102248.

O'Bleness M *et al.* (2012) Evolution of genetic and genomic features unique to the human lineage. *Nat Rev Genet* **13**:853–866; PMID 23154808.

Wall JD (2013) Great ape genomics. *ILAR J* **54**:82–90; PMID 24174434.

Inferring female and male histories using mitochondrial DNA and the Y chromosome

Jobling MA & Tyler-Smith C (2017) Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet* **18**:485–497; PMID 28555659.

Kivisild T (2015) Maternal ancestry and population history from whole mitochondrial genomes. *Investig Genet* **6**:3; PMID 25798216.

Health consequences of our evolutionary history

Gerbault P (2013) The onset of lactase persistence in Europe. *Hum Hered* **76**:154–161; PMID 24861860.

Karlsson EK *et al.* (2014) Natural selection and infectious disease in human populations. *Nat Rev Genet* **15**:379–393; PMID 24776769.

Labrie V *et al.* (2016) Lactase nonpersistence is directed by DNA-variation-dependent epigenetic aging. *Nat Struct Mol Biol* **23**:566–573; PMID 27159559.

Pritchard JK & Di Rienzo A (2010) Adaptation – not by sweeps alone. *Nat Rev Genet* **11**:665–667; PMID 20838407.

Scheinfeldt LB & Tishkoff SA (2013) Recent human adaptation: genomic approaches, interpretation and insights. *Nat Rev Genet* **14**:692–702; PMID 24052086.

HUMAN GENETIC DISEASE

PART FOUR

Chromosomal abnormalities and structural variants

15

If we compare the genomes of two unrelated people we would typically find around four million differences. Chapter 11 gave an overview of the different types of variants. Most would be single nucleotide variants. In addition there would be numerous small insertions or deletions and some larger copy number variants; some microsatellites might have different numbers of repeat units, and there might be some variable insertions or deletions of transposons. A very few of all those variants might directly determine an observable difference between our two people—different blood groups or eye color, perhaps. Several thousand would contribute to a difference without being the sole cause. They might contribute toward making one of the people taller than the other, or more likely to put on weight. The vast majority would have no observable effect.

If one of our two people had a genetic or part-genetic condition, we would expect to find the cause among the variants. If the condition was monogenic (Mendelian), then somewhere among the four million variants there should be a single variant that caused the condition, most likely in a protein-coding sequence. Chapter 16 explains how such variants can cause a disease or other phenotype. If the condition was one of the complex or multifactorial conditions described in Section 5.4, we would expect to find a number of variants, each of which contributed to increasing his or her risk of developing the condition without being either necessary or sufficient for it to develop. Those variants are considered in Chapter 18. In a few cases we might find an extra or missing copy or a major rearrangement of a significant portion of a chromosome, or maybe even a whole chromosome. In this chapter we will consider those chromosomal changes, starting with the way they can be detected and studied, and going on to look at the causes and consequences of abnormalities ranging from changes in chromosome number through major structural changes visible under the microscope and concluding with smaller structural variants such as microdeletions and microduplications.

15.1 STUDYING HUMAN CHROMOSOMES

Human chromosomes have been analyzed for research and diagnostic purposes for over 50 years. A succession of technological advances has permitted chromosome analyses with ever-increasing resolution and structural discrimination. Until recently, the standard method was karyotyping under the microscope, but in the past few years diagnostic laboratories have increasingly moved to molecular methods, primarily array-comparative genomic hybridization, as described below. Whole-genome sequencing will probably eventually replace all these methods, but it is currently too expensive for that.

Karyotyping under the microscope

Traditional karyotyping still has a place, particularly for analyzing cases where the DNA has been wrongly packaged into chromosomes without any being extra or missing (a **balanced abnormality**). Karyotypes also have a considerable educational value. Abnormalities that involve the large-scale structure of chromosomes or their behavior during cell division are best understood by looking at karyotypes rather than at molecular analyses, and so we still illustrate chromosomes and karyotypes here, even though in the laboratory the diagnosis may nowadays rest on molecular data.

Chromosome structure and behavior are relevant in both mitosis and meiosis. However, when human chromosomes are studied under the microscope it is almost always during mitosis. Studying meiosis in humans is difficult. Male meiosis can be studied in a testicular biopsy from any post-pubertal male who is willing to give one, but the key stages of female meiosis take place during fetal life before a woman is born, so they can only be studied in aborted female fetuses. In clinical service, meiotic analysis is used for some investigations of male infertility, but in general it is easier to study the results of meiosis rather than attempting to visualize the process directly in a biopsy. Genotyping markers in individual sperm, in small pooled samples of sperm, or in a set of individuals from a pedigree can allow recombinants to be identified and recombination rates estimated. The recombination hotspots shown in **Figure 12.5** were identified from analyzing pedigrees and populations, not gonadal biopsies.

Chromosomes can only be studied under the microscope in dividing cells, but the options for obtaining these directly from a patient are limited. Bone marrow is a possible source, but it is more convenient to obtain nondividing cells and then propagate them in cell culture in the laboratory. Circulating blood cells are the most commonly used sources of human cells for cytogenetic analyses. People rarely mind giving a small blood sample, and the T lymphocytes in blood can be readily induced to divide by treatment with lectins such as phytohemagglutinin. Alternatively, fibroblasts can be cultured from a skin biopsy. In addition, prenatal diagnosis of chromosome abnormalities traditionally involves analysis of fetal cells shed into the amniotic fluid or removed from chorionic villi (but see Section 20.4).

Although chromosomes were described accurately in some organisms as early as the 1880s, for many decades all attempts to prepare spreads of human chromosomes produced a tangle that defied analysis. It was not until 1956 that Tjio and Levan devised a method that produced good-quality, analyzable preparations. The key to their method was treating cells in culture with hypotonic saline to make them swell. This teased the chromosomes apart. White blood cells are put into a rich culture medium laced with phytohemagglutinin and allowed to grow for 48–72 hours, by which time they should be dividing freely. Nevertheless, because M phase occupies only a small part of the cell cycle, few cells will be actually dividing at any one time. The proportion of cells in mitosis (the mitotic index) can be increased by treating the culture with a spindle-disrupting agent such as colcemid. Cells enter metaphase but are then unable to progress through the rest of M phase, so they accumulate at the metaphase stage of mitosis. The fixed cell suspension is then dropped from a height on to a microscope slide. The impact and subsequent drying spreads and flattens the cells.

Chromosomes are identified by size and banding pattern

Until the 1970s, human chromosomes were identified on the basis of their size and the position of the centromeres, as in the insert in **Figure 15.1**. This allowed chromosomes to be classified into groups (**Table 15.1**) but not individually identified. The introduction of techniques that revealed chromosome banding patterns allowed each individual chromosome to be identified, and permitted more accurate definition of chromosomal abnormalities. The banding techniques require the chromosomes to undergo denaturation or enzymic digestion, followed by exposure to a DNA-specific dye. The procedures produce alternating light and dark bands in mitotic chromosomes (**Figures 15.1** and **15.2**).

A number of different banding techniques have been developed. G-banding is the default karyotyping procedure in cytogenetic laboratories; the other techniques are now seldom employed.

- Q-banding. This was the first banding technique. The chromosomes are stained with a fluorescent dye that binds preferentially to AT-rich DNA, and viewed by ultraviolet fluorescence. The main dyes used were quinacrine, DAPI (4′,6-diamidino-2-phenylindole), and Hoechst 33258.
- G-banding. The chromosomes are subjected to controlled digestion with trypsin before being stained with Giemsa stain. Positively-staining dark bands are known as G-bands. Pale bands are G-negative. G-banding reveals the same patterns as Q-banding, without the complications of fluorescence microscopy, and unlike Q-bands, G-bands do not fade away.
- R-banding. This produces the reverse of the G-banding pattern. The chromosomes are heat-denatured in saline before being stained with Giemsa. The heat treatment denatures AT-rich DNA, and dark R-bands correspond to pale G-bands. R-banding is useful for studying the telomeres of chromosomes, which are pale and hard to make out on G-banded preparations. The same pattern can be produced by using GC-specific dyes such as chromomycin A3, olivomycin, or mithramycin.

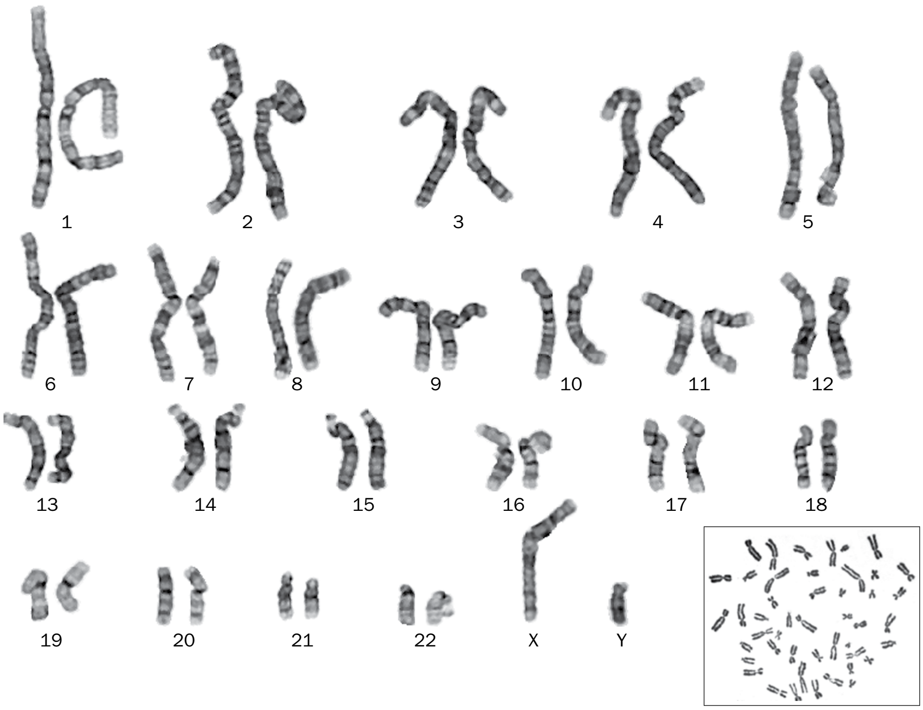


Figure 15.1 G-banded prometaphase mitotic chromosomes from lymphocytes of a normal human male, 46,XY. The randomly scattered chromosomes seen under the microscope have been identified and paired-up by manipulation of a digital image (in the old days, by cutting-up a photograph). It is instructive to compare this figure with the idealized ideograms in the inside back cover of this book. The insert shows chromosomes as prepared in pre-banding days—at metaphase, with sister chromatids held together only at the centromeres, and uniformly stained. (Main image from Rooney DE [ed.] [2001] *Human Cytogenetics: Constitutional Analysis*, 3rd edn. With permission from Oxford University Press.)

TABLE 15.1 HUMAN CHROMOSOME GROUPS

Group	Chromosomes	Description[a]
A	1–3	Largest; 1 and 3 are metacentric but 2 is submetacentric
B	4, 5	Large; submetacentric with two arms very different in size
C	6–12, X	Of medium size; submetacentric
D	13–15	Of medium size; acrocentric with satellites
E	16–18	Small; 16 is metacentric but 17 and 18 are submetacentric
F	19, 20	Small; metacentric
G	21, 22, Y	Small; acrocentric, with satellites on 21 and 22 but not on Y

[a] Autosomes are numbered from largest to smallest, except that chromosome 21 is slightly smaller than chromosome 22. A metacentric chromosome has its centromere at or near the middle. A submetacentric chromosome has its centromere placed so that the two arms are of clearly unequal length. An acrocentric chromosome has its centromere at or near one end. A satellite, in this context, is a small segment separated by a noncentromeric constriction from the rest of a chromosome; these occur on the short arms of most acrocentric human chromosomes.

- C-banding. This is thought to demonstrate constitutive heterochromatin, mainly at the centromeres. The chromosomes are denatured with a saturated solution of barium hydroxide before Giemsa staining.

Cytogeneticists prefer to analyze chromosomes from a slightly earlier part of M phase of the cell cycle (prometaphase) when they are less contracted and show more detail. To obtain substantial cell numbers, the cells are synchronized by temporarily preventing them from progressing through the cell cycle. This can be achieved by adding excess thymidine, bromodeoxyuridine, or amethopterin to the cell culture. Resuspending the cells in fresh medium releases the block, and the cells progress through the cycle synchronously. By trial and error, an optimum interval between release of the block and harvesting the cells can be determined, when a good proportion of cells are in the desired prometaphase stage.

Figure 15.1 shows a typical normal **karyotype**. Strictly, the image is a karyogram, while the karyotype is a verbal description of chromosome number and any abnormalities, but such images are commonly described as karyotypes. See **Box 15.1** for the nomenclature used to describe individual chromosome bands. Banding resolution can be increased by studying the chromosomes at an earlier time in prometaphase when they are more elongated. High-resolution procedures for human chromosomes can identify 400, 550, or 850 bands. As the resolution is increased, main bands split into sub-bands and sub-sub bands (**Figure 15.2**). Nowadays, analyses using highly extended chromosomes under the microscope have largely been abandoned in favor of molecular methods.

BOX 15.1 NOMENCLATURE OF CHROMOSOME BANDS

Inside the back cover of this book is a set of ideograms showing ideal G-banding patterns for all the human chromosomes, and the nomenclature (International System for Human Cytogenetic Nomenclature, ISCN) for referring to each band. These perfect patterns are never observed down the microscope in any single cell, but are used as a reference against which cell after cell is examined until every part of every chromosome has been checked. The system was agreed at an international meeting in Paris in 1971, and hence is known as the Paris nomenclature. Short-arm locations are labeled **p** (*petit*) and long arms **q** (*queue*). Each chromosome arm is divided into regions labeled p1, p2, p3, and so on, and q1, q2, q3, and so on, counting outward from the centromere. Regions are delimited by landmarks such as the ends of the chromosome arms, the centromere, and prominent bands. Regions are divided into bands labeled p11 (one-one, not eleven!), p12, p13, and so on, sub-bands labeled p11.1, p11.2, and so on, and sub-sub-bands, for example p11.21 ("p one-one point two-one"), p11.22, and so on, in each case counting outward from the centromere. Locations close to the centromere are labeled **proximal**, and those far from it are **distal**. Thus, proximal Xq means the segment of the long arm of the X that is near the centromere, and distal 2p means the portion of the short arm of chromosome 2 that is distant from the centromere, and therefore closest to the telomere.

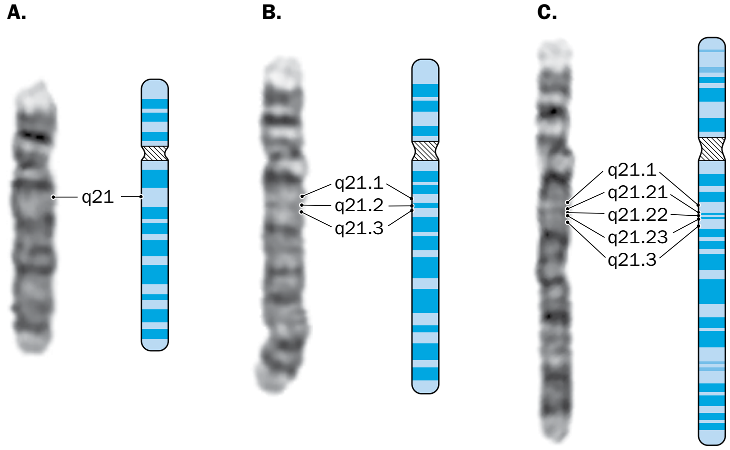


Figure 15.2 Different chromosome banding resolutions can resolve bands, sub-bands, and sub-sub-bands. G-banding patterns for human chromosome 4 (with accompanying ideogram at the right) are shown at increasing levels of resolution. The levels correspond approximately to (**A**) 400, (**B**) 550, and (**C**) 850 bands per haploid set. (Adapted from Rooney DE [ed.] [2001] *Human Cytogenetics: Constitutional Analysis*, 3rd edn. With permission from Oxford University Press.)

Banding patterns correlate with functional elements of chromosome structure. The DNA of the dark G-bands replicates late in S phase and is relatively condensed, whereas the DNA of the pale bands generally replicates early in S phase and is less condensed. The dark G-bands contain relatively few genes and the DNA is less transcriptionally active. Although the AT content of human G-band DNA is only slightly higher than that of R-band DNA, individual G-bands consistently have lower GC content than their immediate flanking sequences. This may correlate with how the DNA becomes organized as it condenses.

Fluorescence *in situ* hybridization represents a fusion of microscopic and molecular techniques

Classical cytogenetic techniques allow the number and gross structure of chromosomes to be analyzed, but their resolution is limited. Changes involving less than 3–5 Mb of DNA are too small to be identified on a standard karyotype. Fluorescence *in situ* hybridization (FISH) allows DNA sequences down to a few kilobases to be detected and localized to a region of a chromosome. The basic technology was described in Section 6.3. To check for the presence and location of a given DNA sequence, a matching fluorescently-labeled single-stranded probe is prepared. Under very carefully controlled conditions, the DNA of a spread of chromosomes on a microscope slide, as prepared for traditional cytogenetic analysis, can be denatured and rendered single-stranded without destroying the

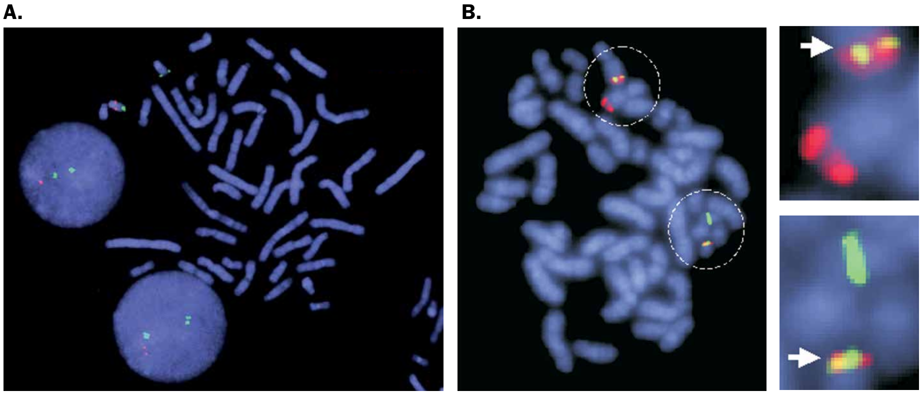


Figure 15.3 Using fluorescence *in situ* hybridization to identify a chromosomal deletion and a translocation. (**A**) On the basis of her phenotype, a patient is suspected of being heterozygous for a submicroscopic deletion of sequences at 22q11. The green probe marks the centromere of chromosome 22, the red probe targets the suspect region. Only one of the two copies of chromosome 22 hybridizes to the red probe, thus confirming the diagnosis. (**B**) Patients with chronic myeloid leukemia usually have a characteristic 9:22 translocation, where those two chromosomes have exchanged segments (see Chapter 19 for an explanation of why this causes leukemia). Chromosomes from a patient are hybridized with probes for chromosome 9 (red) and 22 (green). One normal copy (single color) of each chromosome is seen, but the remaining two copies have closely apposed red and green signals. The panels to the right are blow-ups of the areas in the dashed circles, with the translocated chromosomes marked by arrows. (Courtesy of Fiona Harding, Northern Regional Genetics Service, Newcastle upon Tyne.)

outline of the chromosome. The slide can then be bathed in a solution containing the labeled probe and matching sequences allowed to hybridize. After careful washing, the hybridized probe can be seen as a pair of fluorescent spots at the relevant chromosomal location. The spots are paired because the chromosomes were prepared at prometaphase or metaphase and so consist of paired sister chromatids. By using probes carrying different fluorescent dyes, several probes can be hybridized simultaneously so that the location of specific sequences can be identified in relation to each other (**Figure 15.3**).

FISH can be used on interphase cell nuclei as well as on chromosome spreads. **Figure 15.4A** shows how it could be used to show the number of copies of a chromosome in an interphase cell, while **Figure 15.4B** shows the same translocation as in **Figure 15.3B**, but this time observed on uncultured lymphocytes. Hybridization to highly extended DNA fibers from interphase nuclei (fiber-FISH) has sometimes been used to explore repetitive structures, although nowadays one would probably prefer to use DNA sequencing.

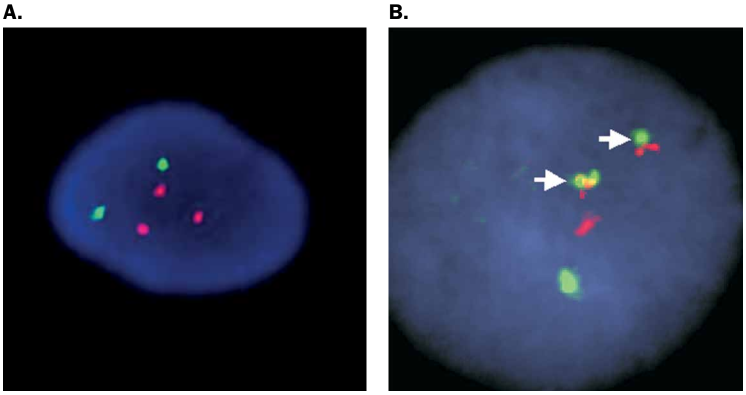


Figure 15.4 Interphase FISH. (**A**) An uncultured cell from a patient with Down syndrome (trisomy 21) hybridized to probes for chromosomes 18 (green) and 21 (red). The result shows the two copies of chromosome 18 but three copies of chromosome 21. (**B**) An interphase cell from the same patient as in **Figure 15.3B** tested with the same probes. Again, the rearranged chromosomes can be identified through their combination of labels (arrows). (B, courtesy of Fiona Harding, Northern Regional Genetics Service, Newcastle upon Tyne.)

Rather than using a probe for a single sequence on a chromosome, one could use a large cocktail of probes that hybridize to sequences along the whole length of one particular chromosome. Such "chromosome paints" light up the whole chromosome and can highlight structural variants (**Figure 15.5**). **Figures 2.20** and **10.1** show how multiple chromosome paints can be used to explore the positioning of chromosomes within interphase nuclei. In general, all these FISH techniques have been rather superseded by sequencing for exploratory investigations, but they remain useful as targeted confirmatory techniques, with their unique ability to relate findings to chromosome structure.

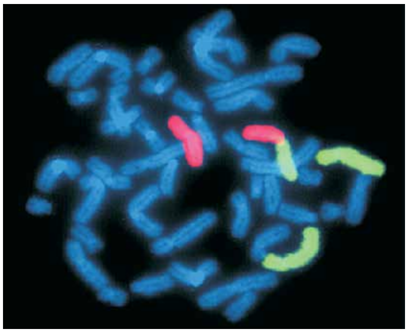


Figure 15.5 Defining a chromosome rearrangement by chromosome painting. By karyotyping a peripheral blood sample, an abnormal X chromosome was identified with extra chromosomal material present on the short arm. Follow-up chromosome painting investigations showed that the additional material present on the short arm of the abnormal X chromosome originated from chromosome 4, as revealed here with a chromosome X paint (red signal) and a chromosome 4 paint (green signal). The background stain for the chromosomes is the blue DAPI stain. (Courtesy of Gareth Breese, Northern Regional Genetics Service, Newcastle upon Tyne.)

Comparative genomic hybridization allows imbalances anywhere in a genome to be detected

Fluorescence *in situ* hybridization is a powerful technique, but it requires one to know in advance which chromosomal location is to be tested, so that an appropriate specific probe can be used. Comparative genomic hybridization (CGH) removes that limitation. CGH uses three sets of DNA sequences: the test DNA, a normal control DNA, and a collection of known and characterized DNA fragments immobilized on a microarray (**Figure 15.6**). Used in this way, the technique is termed array-CGH (aCGH).

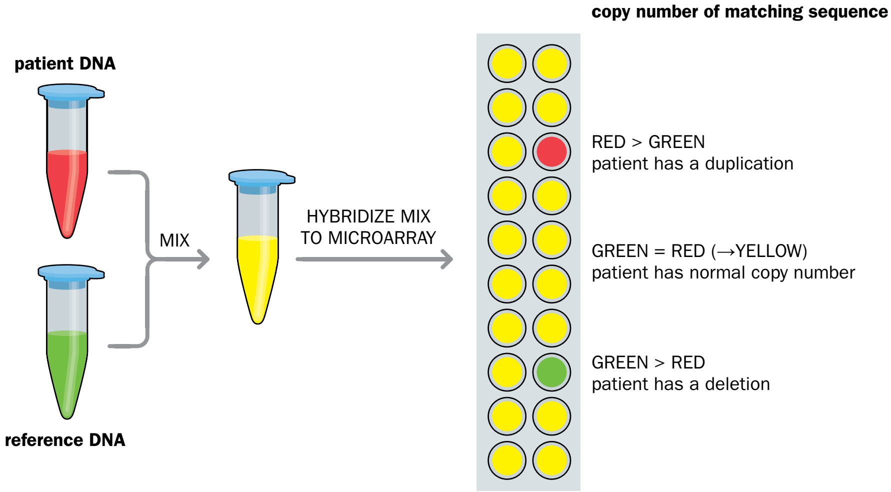


Figure 15.6 Principle of array-comparative genomic hybridization (array-CGH). Test (patient) and reference DNA samples are labeled with different colored fluorophores, fragmented, and made single-stranded. They are mixed in equal genomic amounts and allowed to hybridize to the microarray. Each cell of the array contains a large number of identical single-stranded DNA molecules from a known genomic location. Corresponding fragments of the differently colored test and reference DNA compete to hybridize to the molecules on the array. The average color of a cell of the array after hybridization is a measure of the relative amounts of the corresponding fragments in the test and reference samples. (From Read A & Donnai D [2015] *New Clinical Genetics*, 3rd edn. With permission from Scion Publishing.)

The beauty of array-CGH is that the resolution, and the choice of whether to check the whole genome or some particular part of it, depends purely on the choice of molecules that were used to construct the microarray. To look for imbalances anywhere in the genome, 50,000 long probes fairly regularly spaced across the genome could be used. The average probe spacing of 60 kb would determine the resolution. For a high-resolution examination of, say, just chromosome 9, the same number of probes might be used, but they would be oligonucleotides closely spaced across just that chromosome. Various companies sell standard arrays or will make custom arrays for special purposes.

The array scanner software will present the results in a way that allows one to immediately spot deletions or duplications and see which genes might be involved (**Figure 15.7**). Mosaicism might sometimes be apparent if the color ratios do not fit a simple deletion or duplication, although deep sequencing would be a more sensitive tool for that purpose. Note, however, that array-CGH cannot detect balanced abnormalities such as inversions or balanced translocations. Array-CGH is now the default technique for routine cytogenetics in most diagnostic laboratories.

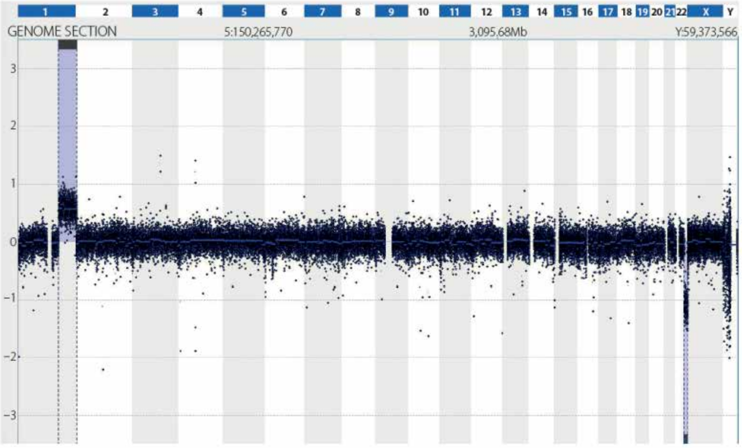


Figure 15.7 Typical result of an array-CGH analysis. Relative intensity of the two fluorescent dyes is plotted on the vertical axis. The reading from each cell of the array is shown as a black dot. Results have been sorted along the horizontal axis by chromosomal location (shown at the top). The result shows that this person has three copies of part of chromosome 1, and only a single copy of part of chromosome 22. (From Read A & Donnai D [2015] *New Clinical Genetics*, 3rd edn. With permission from Scion Publishing. Data produced using an Oxford Gene Technology 8*60 array; courtesy of Lorraine Gaunt and Ronnie Wilson, St Mary's Hospital, Manchester.)

SNP chips can provide similar information to array-CGH

SNP chips (**Figure 15.8**; see **Box 20.1**) use noncompetitive hybridization of just the test DNA. Deletions and duplications are identified by the differing intensity of hybridization, compared to probes from normal diploid sequences. SNP chips have the advantage that they can detect copy-neutral uniparental disomy (UPD). As described in Chapter 10, occasionally, although a sequence is present in the correct two copies per genome, both copies are inherited from just one of the parents. For most chromosomal regions this is only important to the extent that it can create homozygosity for recessive conditions, but for some regions it directly causes developmental abnormalities because of the presence of imprinted genes. When the mechanism producing UPD is trisomy rescue (see Section 15.2), there would be homozygosity for an entire chromosome. Alternatively, mitotic recombination (an abnormal event: normally recombination occurs only in prophase I of meiosis) followed by segregation at cell division can produce segmental UPD, affecting a terminal portion of a chromosome. On a SNP-chip analysis as in **Figure 15.8**, UPD would be seen as a segment having only two genotypes (1-1 or 2-2, depending on the individual SNP) as in Track 3 of the figure, but without any deletion in Tracks 1 or 2. Runs of homozygosity are also seen without UPD when a person inherits both copies of a chromosomal segment from a common ancestor through different lines of descent (**autozygosity**), but except in a closely inbred person the runs are short. If necessary, UPD could be proved unambiguously by genotyping the parents.

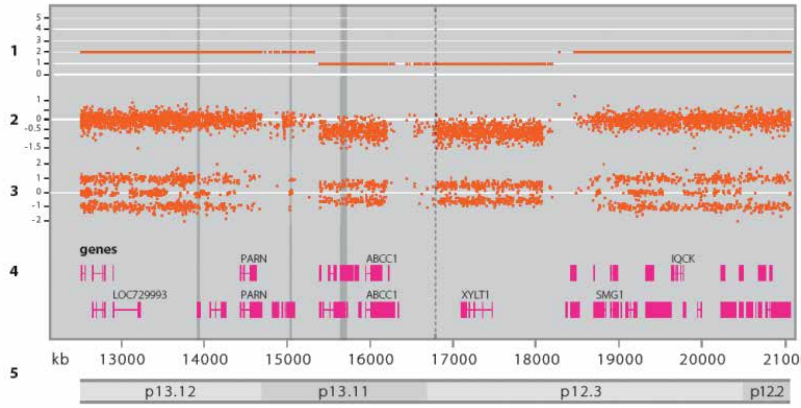


Figure 15.8 SNP-chip data showing a microdeletion at 16p13.11. Across the bottom is an ideogram of chromosome 16, showing the bands and physical distance from 16pter. As with the CGH data in **Figure 15.7**, dots in Track 2 represent the data from each cell, with hybridization intensity, summed across both alleles of each SNP, plotted vertically and chromosomal location horizontally. Track 1 shows the interpretation: there is only a single copy of the central part of the sequence, between positions 15,400 and 18,200. Track 3 shows the genotype at each SNP. In the nondeleted regions there are three possible genotypes, 1-1, 2-1, or 2-2 (arbitrary numbering of alleles), while in the deleted region there are only two, 1 or 2. In summary, there is a 2.8 Mb deletion encompassing the genes shown in Track 4. (From Read A & Donnai D [2015] *New Clinical Genetics*, 3rd edn. With permission from Scion Publishing. Data generated using an Affymetrix SNP 6® microarray, courtesy of Lorraine Gaunt, St Mary's Hospital, Manchester.)

15.2 GROSS CHROMOSOME ABNORMALITIES

Nowadays the widespread use of molecular cytogenetic techniques and whole-genome sequencing has removed any clear dividing line between changes traditionally described as chromosomal and changes thought of as molecular or DNA defects. Nevertheless, in this section we will describe the large-scale abnormalities that are visible under the microscope and traditionally described as chromosomal. We will deal with microdeletions and microduplications in Section 15.3. One might consider an alternative definition of a chromosomal abnormality as an abnormality produced by a specifically chromosomal mechanism, such as incorrect segregation of chromosomes during mitosis or meiosis, improper recombination events, or misrepair of broken chromosomes.

Like all other genetic abnormalities, chromosomal abnormalities can be constitutional or mosaic. Constitutional abnormalities are present in all cells of the body and most likely were present in the original fertilized egg, the result of an abnormal sperm, an abnormal ovum, or a mishap during fertilization. Mosaic abnormalities result when something goes wrong with a single cell in a post-zygotic embryo—most likely, nondisjunction during mitosis. Chromosomal abnormalities, whether constitutional or mosaic, can be classified into numerical and structural abnormalities (**Table 15.2**).

Numerical chromosomal abnormalities include polyploidy and aneuploidy

Polyploidy

Normal somatic cells are diploid, having two genomes. Gametes (sperm or egg) are haploid, with a single genome (23 chromosomes). Polyploid cells have more than two complete genomes. Out of all recognized human pregnancies, 1–3% involve a triploid

TABLE 15.2 TYPES AND NOMENCLATURE OF CHROMOSOME ABNORMALITIES

Type of abnormality	Example karyotypes	Explanation/notes
NUMERICAL		
Triploidy	69,XXX; 69,XXY; 69,XYY	Three complete genomes per cell
Trisomy	47,XXY; 47,XX,+21	Gain of an autosome is indicated by +
Monosomy	45,X; 45,XY,–21	Loss of an autosome is indicated by –
STRUCTURAL		
Deletion	46,XY,del(4)(p16.3)	Terminal deletion (breakpoint at 4p16.3)
	46,XX,del(5)(q13q33)	Interstitial deletion of 5q13–q33
Inversion	46,XY,inv(11)(p11p15)	Paracentric inversion (breakpoints on the same arm)
Duplication	46,XX,dup(1)(q22q25)	Duplication of region spanning 1q22 to 1q25
Insertion	46,XX,ins(2)(p13q21q31)	A rearrangement of one copy of chromosome 2 by insertion of segment 2q21–q31 into a breakpoint at 2p13
Ring chromosome	46,XY,r(7)(p22q36)	Joining of broken ends at 7p22 and 7q36 to form a ring
Marker	47,XX,+mar	The cell contains an extra unidentified chromosome
Reciprocal translocation	46,XX,t(2;6)(q35;p21.3)	A balanced reciprocal translocation with breakpoints at 2q35 and 6p21.3
Robertsonian translocation (gives rise to one derivative chromosome)	45,XY,der(14;21)(q10;q10)	A balanced carrier of a 14;21 Robertsonian translocation. q10 is not really a chromosome band, but indicates the centromere; der (derivative) is used when one chromosome from a translocation is present
	46,XX,der(14;21)(q10;q10),+21	An individual with Down syndrome possessing one normal chromosome 14, a Robertsonian translocation 14;21 chromosome, and two normal copies of chromosome 21

For a more complicated nomenclature that allows complete description of any chromosome abnormality see Shaffer LG, Slovak ML & Campbell LJ (eds) (2009) ISCN 2005: *An International System for Human Cytogenetic Nomenclature*. S Karger AG.

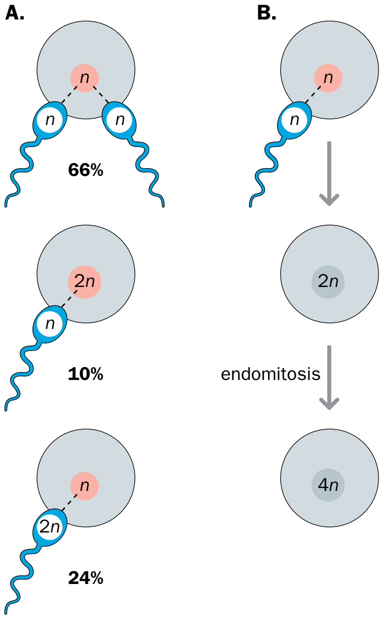


Figure 15.9 Origins of triploidy and tetraploidy. (**A**) Origins of human triploidy. Dispermy is the principal cause, accounting for 66% of cases. Triploidy can also be caused by diploid gametes that arise by occasional faults in meiosis; fertilization of a diploid ovum and fertilization by a diploid sperm account for 10% and 24% of cases, respectively. (**B**) Tetraploidy involves normal fertilization and fusion of gametes to give a normal zygote. Subsequently, however, tetraploidy arises by endomitosis when DNA replicates without subsequent cell division.

embryo (**Figure 15.9A**). The usual cause is two sperm fertilizing a single egg (dispermy), but triploidy is sometimes attributable to fertilization involving a diploid gamete. Triploids very seldom survive to term, and the condition is not compatible with continued life. Constitutional tetraploidy (**Figure 15.9B**) is much rarer and always lethal. It is usually due to failure to complete the first zygotic division: the DNA has replicated to give a content of 4C but cell division has not then taken place, as it should. Although constitutional polyploidy is rare and lethal, all normal people have some tetraploid or higher polyploid cells and so are formally mosaic. Mosaic triploidy (46/69 chromosomes) has been reported but is rare. It probably arises through fusion of the second polar body (which is haploid) with one of the cleavage nuclei of a normal diploid zygote.

Aneuploidy

Euploidy means having complete chromosome sets. **Aneuploidy** is the opposite: one or more individual chromosomes are present as an extra copy or are missing. In trisomy there are three copies of a particular chromosome in an otherwise diploid cell, for example trisomy 21 (47,XX,+21 or 47,XY,+21) in Down syndrome. In monosomy a chromosome is lacking from an otherwise diploid state, as in monosomy X (45,X) in Turner syndrome. Autosomal monosomies are always lethal in constitutional form. Cancer cells often show extreme aneuploidy, with many chromosomal abnormalities.

Aneuploid cells arise through **nondisjunction** or **anaphase lag**. In nondisjunction, paired chromosomes fail to separate (*disjoin*) during anaphase of meiosis I, or, alternatively, sister chromatids fail to disjoin at either meiosis II or mitosis. Nondisjunction during meiosis produces gametes with 22 or 24 chromosomes, which after fertilization with a normal gamete produce a monosomic or trisomic zygote. For reasons that are not very clear, most cases of nondisjunction arise during maternal meiosis (but men should not feel smug—the great majority of point mutations are of paternal origin; see Section 12.3). Nondisjunction during mitosis produces one monosomic and one trisomic daughter cell. The monosomic cell will probably die, but its trisomic partner may survive to establish mosaic trisomy. Anaphase lag is when a chromosome or chromatid is delayed in its movement during anaphase, lags behind the others, and fails to be incorporated into a daughter nucleus. Chromosomes that do not enter the nucleus of a daughter cell are eventually degraded (though sometimes they persist in **micronuclei**—small membrane-bound vesicles). Loss of the Y chromosome through anaphase lag is a frequent cause of Turner syndrome, and is also seen in individual cells of healthy older men.

Clinical consequences of numerical chromosome abnormalities

Having the wrong number of chromosomes has serious, usually lethal, consequences (**Table 15.3**). Even though the extra chromosome 21 in a person with trisomy 21 (Down syndrome) is a perfectly normal chromosome, inherited from a normal parent, its presence causes multiple abnormalities that are present at birth (congenital). Embryos with trisomy 13 or trisomy 18 can also survive to term, but have severe developmental malformations that are incompatible with long-term survival. Other autosomal trisomies are not compatible with survival to term except sometimes in mosaic form. Autosomal monosomies have even more catastrophic consequences, and are invariably lethal at the earliest stages of embryonic life.

TABLE 15.3 CLINICAL CONSEQUENCES OF NUMERICAL CHROMOSOME ABNORMALITIES IN HUMANS

Abnormality	Clinical consequences
Triploidy (69,XXX or 69,XYY)	1–3% of all conceptions; rare liveborn babies do not survive long
AUTOSOMAL ANEUPLOIDY	
Nullisomy (lacking both homologs of a pair)	Lethal at pre-implantation stage
Monosomy (one chromosome missing)	Lethal during embryonic development
Trisomy (one extra chromosome)	Usually lethal during embryonic or fetal[a] stages, but fetuses with trisomy 13 (Patau syndrome) and trisomy 18 (Edwards syndrome) may survive to term; those with trisomy 21 (Down syndrome) can survive beyond age 40
SEX CHROMOSOME ANEUPLOIDY	
Additional sex chromosomes	Individuals with 47,XXX, 47,XXY, or 47,XYY all have relatively minor problems and a normal life span
Lacking a sex chromosome	45,Y is never viable. 99% of cases of 45,X (Turner syndrome) abort spontaneously; survivors are of normal intelligence but are infertile and show minor but characteristic abnormalities

[a] In humans, the embryonic period spans fertilization through to the end of the eighth week of development. Fetal development then begins and lasts until birth.

The developmental abnormalities associated with monosomies and trisomies must be the consequence of an imbalance between the numbers (dosage) of different chromosomes. Normal development and function depend on innumerable interactions between proteins, RNA molecules, and DNA sequences that are often encoded or located on different chromosomes. For at least some of these interactions, balancing the levels of the interacting partners is critically important. This might be particularly the case for transcription factors, which are likely to be encoded on a different chromosome from their target sequences. Altering the relative numbers of chromosomes will affect these interactions. Monosomies have a more profound effect than trisomies, since reducing the copy number of one partner by 50% could be expected to be more disruptive than providing three copies in place of two.

Having extra sex chromosomes has far fewer ill effects than having an extra autosome. People with 47,XXX or 47,XYY karyotypes often function within the normal range, and 47,XXY men have relatively minor problems compared to people with any autosomal trisomy. Even monosomy (in 45,X women) can have remarkably few major consequences—although 45,Y is always lethal. As explained in Section 10.4, special mechanisms allow normal development to proceed in individuals with different numbers of sex chromosomes. The Y chromosome carries relatively few genes, while X-inactivation ensures that each cell has just one functional X chromosome, regardless of the number on the karyotype.

It is not so obvious why triploidy is lethal in humans and other animals. With three copies of every autosome, the dosage of autosomal genes is balanced and should not cause problems. Triploids are always sterile because triplets of chromosomes cannot pair and segregate correctly in meiosis, but many triploid plants are in all other respects healthy and vigorous. The lethality in animals is probably due to an imbalance between products encoded on the X chromosome and autosomes, which cannot be compensated by X-inactivation.

Embryos with autosomal trisomies, other than trisomies 13, 18, or 21, cannot survive. However, they might be rescued if a chance mitotic nondisjunction very early in embryonic development produced a disomic and a tetrasomic daughter cell. The disomic cell would have a growth and survival advantage, and the resulting baby might be a mosaic with only a minor trisomic line. If the nondisjunction event happened sufficiently early in pregnancy, progeny of that single disomic cell might come to make up the whole later fetus. This **trisomy rescue** is one mechanism by which uniparental disomy (UPD) can arise (in constitutional or mosaic form). If we label the three copies of the trisomic chromosome by their parental origin, M for maternal and P for paternal, one-third of random nondisjunctions of the trisomic chromosome in an MMP cell would produce an MM cell, with UPD. The other two-thirds would produce an entirely conventional MP cell, and if the revertant cells are constitutional, no examination of the resulting baby would hint at its lucky escape.

A variety of structural chromosome abnormalities result from misrepair of damage, inappropriate recombination, or errors in replication

The immensely long DNA molecules in chromosomes frequently break. To survive, cells rely on a number of independent repair mechanisms (see Section 11.2). Cellular enzyme systems recognize broken chromosomes and try to repair them by joining broken ends. The telomeres on normal chromosome ends protect them from being treated as breaks. Normally the system works so well that we are unaware of its operation; however, if there are more than two broken ends, the repair machinery may join them to the wrong partners. This is one of the origins of **reciprocal translocations**, where two nonhomologous chromosomes swap segments (**Figure 15.10A**). Provided the swap is such that each partner ends up with just one centromere, the translocated chromosomes can go through mitosis without problems. Exchanges that produce dicentric and acentric chromosomes no doubt occur, but will not found a clone of similar abnormal cells. The acentric fragment would be lost at the next mitosis, while a dicentric chromosome would be trapped in unstable

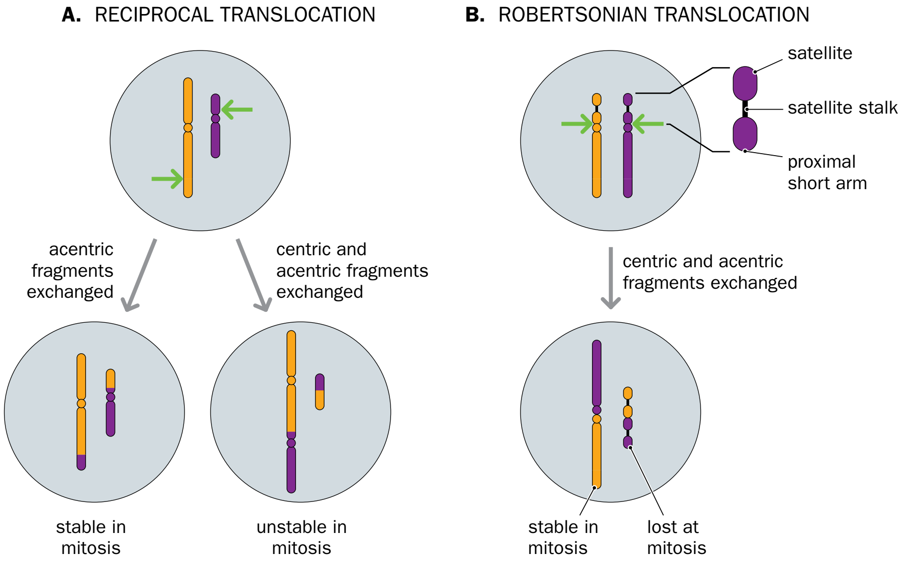

A. RECIPROCAL TRANSLOCATION
B. ROBERTSONIAN TRANSLOCATION
satellite
satellite stalk
proximal short arm
acentric fragments exchanged
centric and acentric fragments exchanged
centric and acentric fragments exchanged
stable in mitosis
unstable in mitosis
stable in mitosis
lost at mitosis

Figure 15.10 Reciprocal and Robertsonian translocations. (**A**) Reciprocal translocation. The derivative chromosomes are stable in mitosis when each has a single centromere. (**B**) Robertsonian translocation. Recombination between the proximal short arms of two acrocentric chromosomes produces dicentric and acentric products. The two centromeres of the dicentric chromosome are so close together that they act as a single centromere, and the chromosome is stable in mitosis. The acentric fragment is lost, but this has no phenotypic effect because the short arms of all five pairs of acrocentric chromosomes (13, 14, 15, 21, and 22) contain similar ribosomal RNA genes, so the loss of two of the ten is not critical.

breakage–fusion–bridge cycles, which underlie some of the gross chromosomal abnormalities seen in cancer cells, but not in healthy individuals. If there are two breaks on the same chromosome, misrepair could produce a **deletion** or an **inversion** (Figure 15.11).

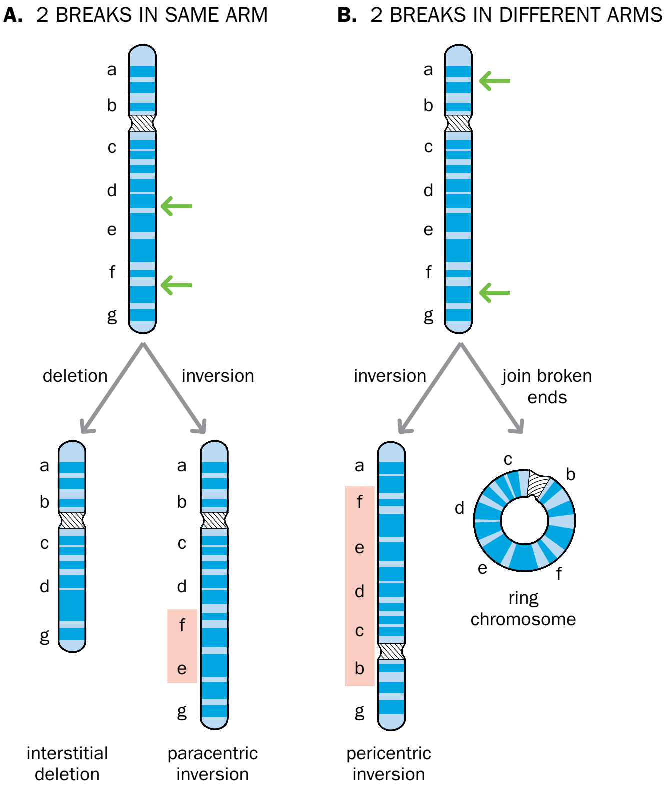


Figure 15.11 Stable outcomes after incorrect repair of two breaks on a single chromosome. Incorrect repair of two breaks (green arrows) with no loss of material can produce paracentric or pericentric inversions. If the repair produces an acentric fragment that would be lost at the next mitosis, the stable result is an interstitial deletion or ring chromosome.

Structural chromosome abnormalities can be produced by errors in recombination, in addition to faulty DNA repair. Recombination involves paired homologous sequences and is initiated by a double-strand break. Normally the paired sequences are allelic as well as homologous but repeated sequences can allow **non-allelic homologous recombination** (NAHR, see Figure 15.17). Structural variants can also arise by template switching during DNA replication. If a replication fork encounters damaged DNA it may stall, and one possible outcome is a switch to another template, usually helped by microhomology (two or a few nucleotides) between the old and new templates. These mechanisms are described in more detail below.

A **Robertsonian translocation** is a special type of translocation that joins two acrocentric chromosomes (numbers 13, 14, 15, 21, and 22). The short arm of each of these chromosomes is very small and contains very similar DNA: 1–2 Mb arrays of tandemly repeated ribosomal RNA genes (see Section 9.2). Recombination between homologous sequences in the short arms of two different acrocentric chromosomes can result in acentric and dicentric products (Figure 15.10B). Unusually, the two centromeres of the dicentric chromosome are sufficiently close together that they function as a single large centromere, so that the fusion chromosome segregates regularly and is stable in mitosis. The acentric fragment is lost, but this has no phenotypic effect because it contains only highly repetitive sequences that are also present at high copy number on all the other acrocentric chromosomes.

Many more complex structural variants have been reported, including insertional or multiway translocations and isochromosomes (symmetrical metacentric chromosomes consisting of two long arms or two short arms of some particular chromosome, presumably the result of a U-turn at the centromere during DNA replication). An extreme case is **chromothripsis** (see Section 19.4). Originally discovered in cancer cells, chromothripsis has now also been recorded in noncancer cases. It appears as though one chromosome, or maybe one part of a chromosome, has been pulverized into fragments and the fragments reassembled randomly. It is thought the mechanism involves anaphase lag, but instead of being lost, the lagged chromosome is incorporated into a micronucleus where it undergoes extensive rearrangement. At some later time, the micronucleus gets incorporated back into the main nucleus.

Sequencing the DNA around breakpoints has revealed that breaking and joining is not always the neat cut-and-paste process implied by our diagrams. Often there are small deletions, insertions, or duplications at the junction. These can provide clues to the mechanisms involved, as described in Section 15.3.

Consequences of structural chromosome abnormalities

Structural chromosome abnormalities are **balanced** if there is no net gain or loss of chromosomal material, and **unbalanced** if there is a net gain or loss. In general, balanced rearrangements have no phenotypic effect, while unbalanced abnormalities may have an effect, depending on what material is gained or lost. Robertsonian translocations are regarded as balanced even though some material is lost (**Figure 15.10B**) because, as explained above, there is no phenotypic effect from the loss.

A significant proportion of *de novo*, apparently balanced rearrangements are associated with phenotypic abnormalities, but often sequencing the breakpoints shows that the rearrangement is not in fact truly balanced. Truly balanced abnormalities can still affect the phenotype under some circumstances:

- A chromosome break may disrupt an important gene;
- A chromosome break may affect the expression of a gene without disrupting the coding sequence by, for example, translocating an active gene into heterochromatin, or separating a gene from an essential enhancer. This is obvious if a rearrangement puts a *cis*-acting control element on a different chromosome from its target gene. But much smaller rearrangements can also have large effects. Developmental genes can be dysregulated by small rearrangements that move the boundaries of topologically-associated domains (TADs, see Section 10.1). For an enhancer that normally regulates expression of gene A in the same TAD, moving the boundary could place it in a different TAD, where it can no longer regulate gene A but instead inappropriately regulates gene B (see **Figure 16.9**). Thus, when considering the possible effects of a rearrangement, it is useful to consider not just whether it disrupts a gene but also whether it moves a gene or enhancer from one TAD into another;
- Balanced X-autosome translocations are a special case. They cause nonrandom X-inactivation. As discussed in Section 10.4, X-inactivation spreads physically along the chromosome from the X-inactivation center. If the inactivated X is disrupted by a translocation, the detached portion cannot be inactivated. This is likely to produce an imbalance that will be fatal to the cell. Thus the whole body of a female who carries an X-autosome translocation is made of the descendants of cells that happened to inactivate the intact X. That would not matter, unless the translocated X chromosome harbors a significant mutation—most likely if the translocation break disrupted an important gene. In that case, every cell of the woman's body will show the effect of the mutation, in the same way as every cell of a 46,XY male carrying the mutation would. A striking example is the two dozen or so women worldwide who have severe Duchenne muscular dystrophy. They are not homozygous for a mutant dystrophin gene; they all have X-autosome translocations, with different autosomal partners in each case, but all with the X-chromosome breakpoint disrupting the huge dystrophin gene.

Unbalanced structural chromosome abnormalities can arise directly through deletion or, less frequently, by duplication. They also frequently arise indirectly by mal-segregation of chromosomes during meiosis in a carrier of a balanced abnormality. Provided each chromosome has one and only one centromere, a chromosomal abnormality can be transmitted through mitosis without problems, and if there is no extra or missing material, a person carrying it will most usually be entirely normal phenotypically, with the exceptions noted above. However, problems arise in meiosis. When homologous chromosomes pair in prophase of meiosis I, the rearranged chromosomes do not match their normal counterparts. Homologous sequences within chromosomes will still pair, as long as that does not involve impossible chromosomal contortions. The resulting structures are shown in **Figure 15.12**.

- In carriers of a reciprocal translocation, the two rearranged chromosomes form a cross-shaped quadrivalent with their normal counterparts (**Figure 15.12A**).
- In carriers of a Robertsonian translocation, the fusion chromosome forms a trivalent with the two normal chromosomes (**Figure 15.12B**).
- In carriers of an inversion, the inverted and normal chromosomes form a loop, unless the inverted segment is small, in which case there may be just an unpaired section of the normal bivalent. In a pericentric inversion (**Figure 15.12C**) the loop includes the centromeres but in a paracentric inversion (**Figure 15.12D**) it does not.

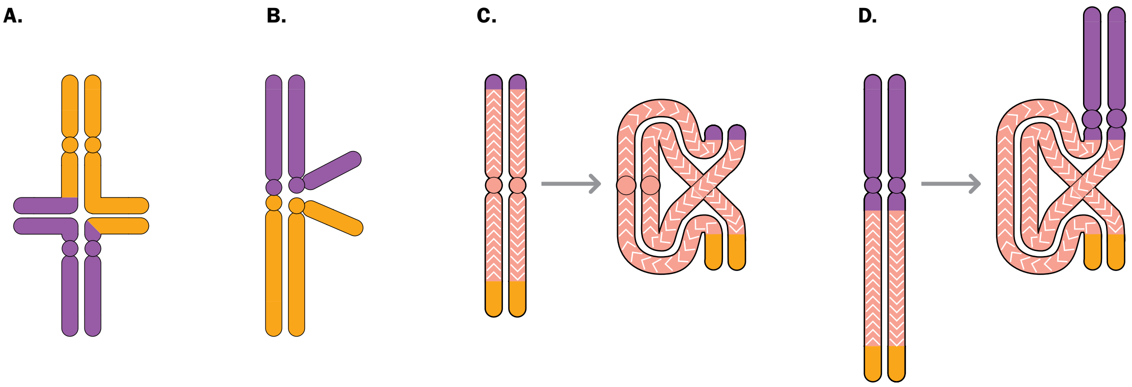


Figure 15.12 Chromosome pairing in prophase of meiosis I in carriers of balanced chromosomal rearrangements. (**A**) A reciprocal translocation. (**B**) A Robertsonian translocation. (**C**) A pericentric inversion. (**D**) A paracentric inversion. Inverted segments in (**C**) and (**D**) are colored pink, noninverted segments are purple and yellow.

For translocation carriers (reciprocal or Robertsonian), the outcome of meiosis depends on the way the centromeres segregate in anaphase I. The quadrivalent in a reciprocal translocation can segregate so as to give an entirely normal gamete, one carrying the same balanced translocation, or various combinations of partial trisomy for one of the translocation partners combined with partial monosomy of the other (Figure 15.13). In addition to the 2:2 segregations shown in the figure, 3:1 segregations are frequent with some translocations. It is not possible to predict with accuracy the relative likelihood of each outcome, although the book by Gardner, Sutherland, and Shaffer (see Further Reading) gives some guidelines and examples. For genetic counseling, moderate imbalances that could produce a liveborn, abnormal baby are more to be feared than gross imbalances that would be incompatible with survival to term.

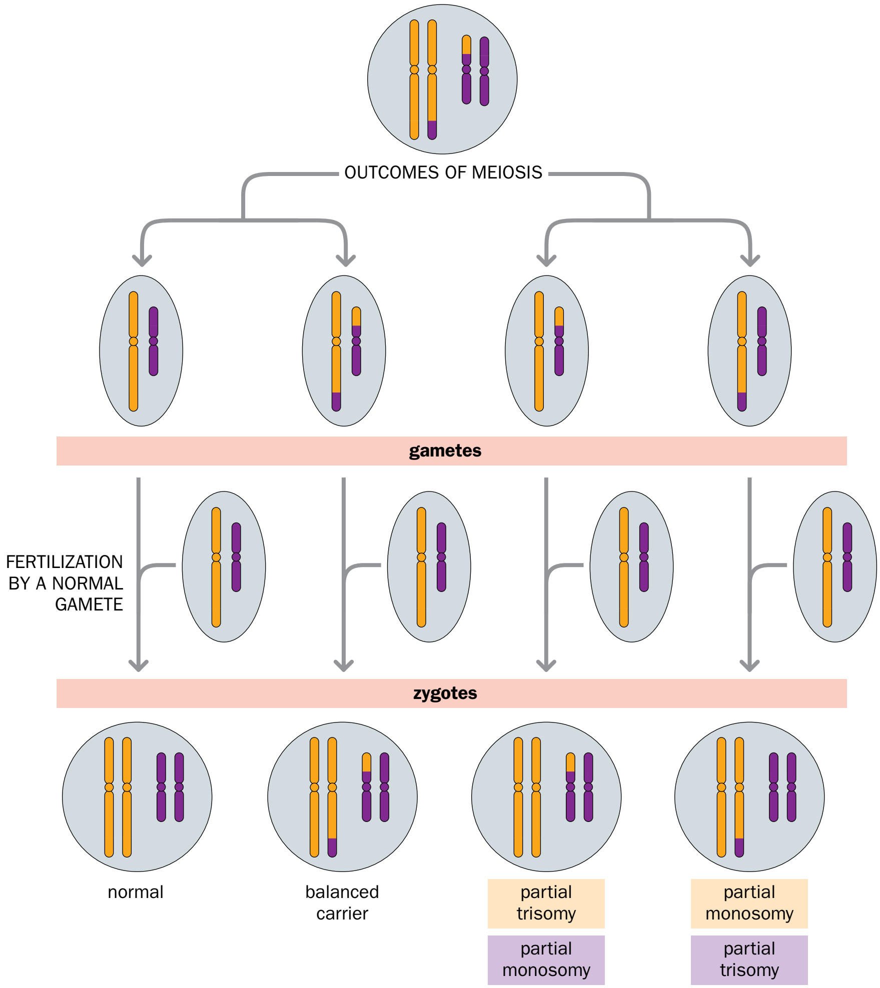


Figure 15.13 Possible outcomes of meiosis in a carrier of a balanced reciprocal translocation. This and the next two diagrams (**Figures 15.14** and **15.15**) are slightly simplified because in meiosis I each chromosome consists of two sister chromatids that only become separated at anaphase II. This does not alter the consequences as shown here. Other modes of segregation are also possible, for example 3:1 segregation. The relative frequency of each possible gamete is not readily predicted. The risk of a carrier having a child with each of the possible outcomes depends on the frequency of an outcome in the gametes, but also on the likelihood of a conceptus with that abnormality surviving to term. See the book by Gardner, Sutherland, and Shaffer (in Further Reading) for discussion.

For a carrier of a Robertsonian translocation, Figure 15.14 shows the possible modes of segregation. The possible outcomes are an entirely normal conceptus, one carrying the same balanced translocation, or full trisomy or full monosomy for one of the translocation partners.

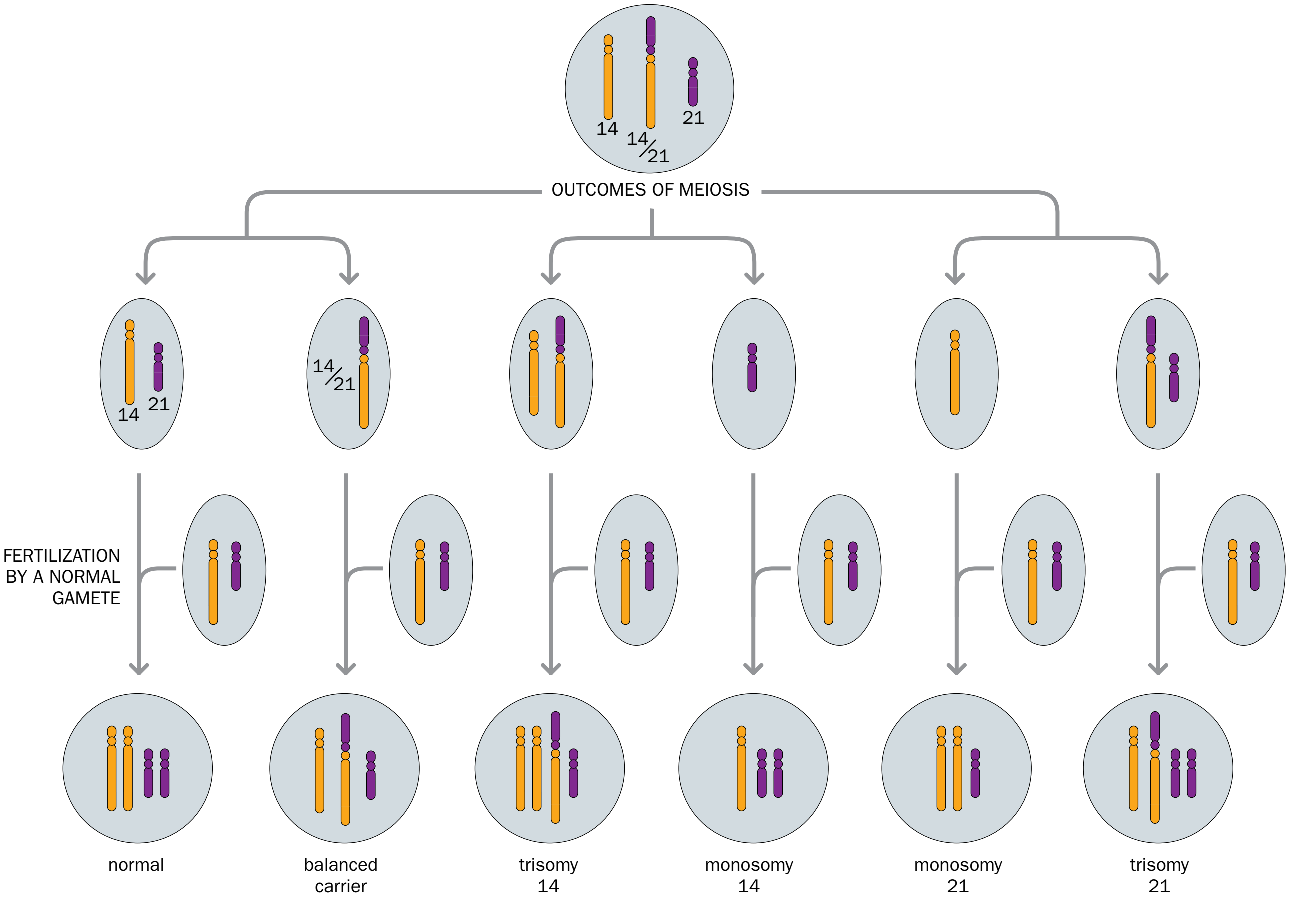


Figure 15.14 Possible outcomes of meiosis in a carrier of a Robertsonian 14:21 translocation. The two monosomic zygotes and the trisomy 14 zygote in this example would not develop to term. Around 5% of Down syndrome births are the result of this process rather than simple nondisjunction. The clinical phenotype is the same regardless of the mechanism of origin.

In the cases with inversions, the potential problems do not arise from anaphase segregation but from prophase recombination. Normally there is at least one crossover per chromosome arm. For translocation carriers these make no difference to the main outcomes. However, for an inversion heterozygote, a crossover within the inversion loop has serious consequences (Figure 15.15). In a pericentric inversion loop, the result is a duplication on one chromatid and a deletion on the other. For a paracentric inversion, the result is an acentric and a dicentric chromatid that would not be stable at anaphase.

15.3 STRUCTURAL VARIANTS, MICRODELETIONS, AND MICRODUPLICATIONS

Between gross chromosomal aberrations and single nucleotide changes lies a whole series of structural variants. These are variants ranging in size from a few kilobases to a few megabases that are too small to be seen under the microscope, but too large to be contained within a single PCR product. Chromosomal abnormalities too small to be seen under the microscope were long suspected to be the cause of a number of unexplained recurrent syndromes and also of many individual cases in which a patient had a unique pattern of abnormalities. A deletion or duplication of 3 Mb of DNA would be invisible under the microscope on a standard karyotype, but could involve wrong dosage of dozens of genes. However, before the widespread use of array-CGH there was no systematic way of identifying submicroscopic variants. Progress depended on chance lucky breaks.

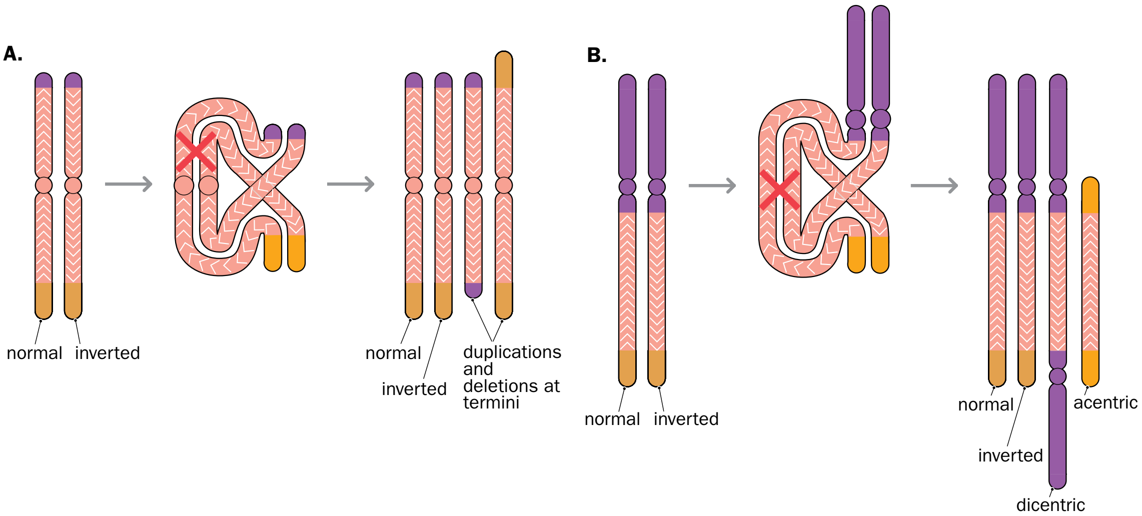


Figure 15.15 The result of recombination in a carrier of an inversion. (**A**) In a pericentric inversion, the recombinant chromatids have a deletion and a duplication. (**B**) In a paracentric inversion, the recombinant chromatids are acentric and dicentric.

One example is Williams–Beuren syndrome (WBS; OMIM #194050). Affected individuals have a unique and well-recognizable combination of a distinctive facial appearance, mild to moderate intellectual disability, specific cognitive and behavioral patterns, and a heart problem, supravalvular aortic stenosis (SVAS). Isolated SVAS also occurs as an autosomal dominant character (OMIM #185500). In 1993 a family was reported where SVAS co-segregated with a chromosomal translocation t(6;7)(p21.1;q11.23) that disrupted the elastin gene on chromosome 7. It seemed plausible that an abnormality of a blood vessel wall should be caused by mutation in a component of connective tissue. Thus the elastin gene became a strong candidate for SVAS, and this pointed to the elastin region on chromosome 7 as the likely location of the abnormality causing WBS. Now that a candidate location was suggested, Southern blotting and FISH could be used, and these confirmed that WBS was caused by heterozygosity for a recurrent 1.5–1.8 Mb microdeletion at 7q11.23.

While array-CGH revealed many variants in patients, it also did so in healthy controls (**Figure 15.16**). This was unexpected: it had been commonly assumed that structural variants involving significant loss or gain of material would normally be pathogenic. It thus became a major challenge to decide whether a variant found in a patient might be responsible for their condition. A first step is to search databases of variants found in normal healthy subjects, such as the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). Data from patients, including clinical detail, can be checked in dbVAR (http://www.ncbi.nlm.nih.gov/dbvar/) and the Decipher database (https://decipher.sanger.ac.uk/), among other sources. One would hope to find that the variant under investigation had been reported from patients with similar phenotypes, but not in healthy controls. However, many variants seen in patients are novel, and in that case there are some widely used rules of thumb that can provide some guidance:

- Larger imbalances are more likely than smaller ones to be pathogenic. However, this is not universally true. Many pathogenic variants are smaller than some of the nonpathogenic variants shown in **Figure 15.16**;
- Deletions are more likely than duplications to be pathogenic, although there are many individual exceptions;
- *De novo* imbalances are more likely to be pathogenic than inherited ones. Nevertheless, most normal individuals have *de novo* variants while, as described below, some inherited variants, even though sometimes inherited from an apparently normal parent, may be pathogenic.

The software used to identify a variant will normally list the genes involved (as in **Figure 15.8**), and these can be checked to see if any have functions that seem relevant to the patient's phenotype, or have reported mutations that cause a related condition. The boundaries of TADs (see Section 10.1) should also be considered. As Lupiáñez and

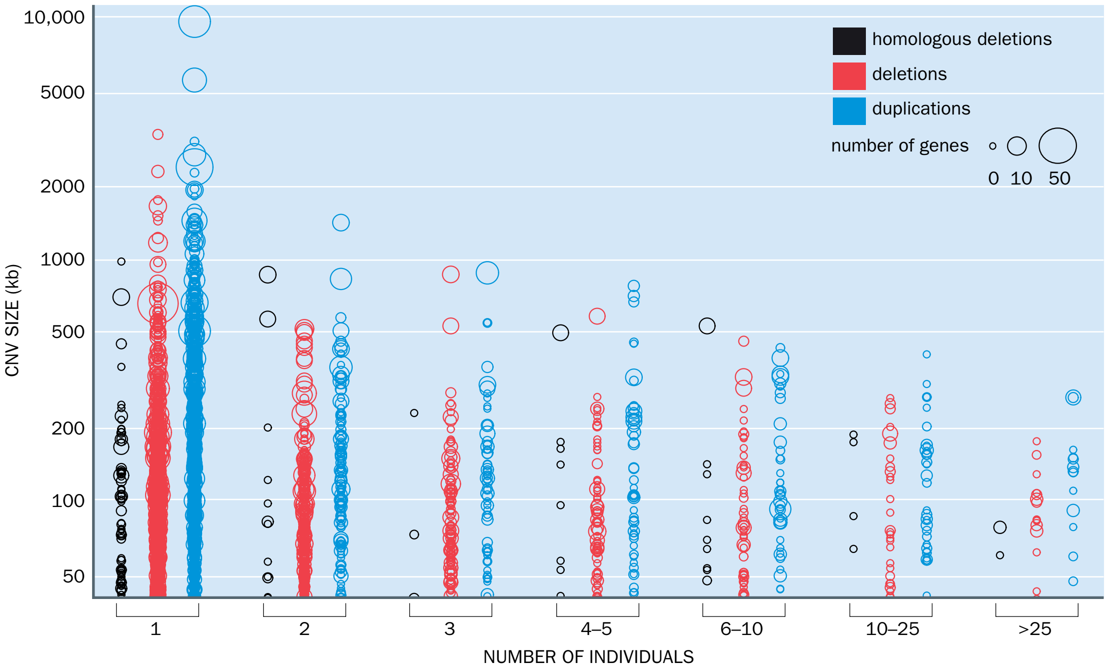


Figure 15.16 Structural variants in 2493 healthy individuals. Variants are classified according to type (duplications, heterozygous deletions, homozygous deletions), size (y axis), frequency (x axis), and number of genes involved (size of circles). Common variants were less than 500 kb in size, but much larger variants were seen in rare individuals. CNV, copy number variants. (Adapted from Itsara A *et al*. [2009] *Am J Hum Genet* **84**:148–161; PMID 19166990. With permission from Elsevier.)

colleagues have shown (PMID 25959774; see Further Reading), structural variants that change the boundaries of TADs can cause pathogenic disruption of regulatory processes.

Recurrent and nonrecurrent structural variants

Structural variants can be divided into recurrent and nonrecurrent. Recurrent variants are likely to be the result of recurrent nonallelic homologous recombination (NAHR) between repeated sequences (**Figure 15.17**). NAHR can produce deletions, duplications, or inversions. Different mechanisms are responsible for nonrecurrent variants, as described below. **Table 15.4** shows some examples of syndromes caused by microdeletions. Whether the corresponding duplication will produce a clinical syndrome depends on whether a 50% increase in dosage of the region interferes with normal development. WBS, Smith-Magenis syndrome (SMS), and hereditary neuropathy with liability to pressure palsies (HNPP) all have corresponding microduplication syndromes: an unnamed syndrome (OMIM #609757) for WBS, Potocki-Lupski syndrome (OMIM #610883) for SMS, and Charcot-Marie-Tooth disease type 1A (CMT1A; OMIM #118220) for HNPP.

NAHR between repeats on the same chromosome that are in opposite orientations produces an inversion (**Figure 15.17B**). Such inversions are normally nonpathogenic unless they disrupt a gene or separate it from a regulatory element. Many are present as polymorphic variants in the normal population. It turns out that for many *de novo* NAHR-mediated microdeletions the affected chromosome was inherited from a parent who was heterozygous for an inversion of the region in question (**Table 15.5**). The inversion will be too small to form the sort of loop shown in **Figure 15.12**, but it would prevent normal meiotic pairing; meanwhile, repeats within the inverted segment, that would normally be in opposite orientations in the two chromosomes, are now in the same orientation, allowing a further act of NAHR to produce a deletion.

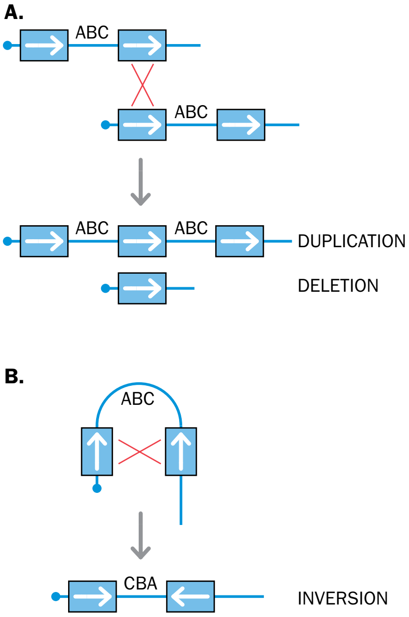


Figure 15.17 Nonallelic homologous recombination (NAHR). The blue boxes represent low-copy repeats that are highly homologous (they have closely similar sequences) but are not allelic (they are at different locations, though close together on the same chromosome). (**A**) NAHR between repeats in the same orientation produces duplication or deletion of the sequence between the repeats. (**B**) Intrachromosomal NAHR between repeats in opposite orientations inverts the sequence between them.

TABLE 15.4 EXAMPLES OF RECURRENT SYNDROMES CAUSED BY MICRODELETIONS

Syndrome	OMIM #	Location	Type[a]	Main mechanism
Wolf–Hirschhorn	194190	4pter	CGS	Terminal deletions
Cri-du-chat	123450	5pter	CGS	Terminal deletions
Williams–Beuren	194050	7q11.23	CGS	NAHR
Langer–Giedion	150230	8q24	CGS (*TRPS1*, *EXT1*)	Interstitial deletions
WAGR	194072	11p13	CGS (*PAX6*, *WT1*)	Interstitial deletions
Prader–Willi	176270	15q11q13	SGS (*SNORD116*)	NAHR
Angelman	105830	15q11q13	SGS (*UBE3A*)	NAHR
Miller–Dieker	247200	17pter	CGS (*PAFAH1B1* etc.)	Terminal deletions
Smith–Magenis	182290	17p11.2	SGS (*RAI1*)	NAHR
HNPP	162500	17p12	SGS (*PMP22*)	NAHR
Alagille Type 1	118450	20p12	SGS (*JAG1*)	Interstitial deletions
DiGeorge/VCFS	188400/192430	22q11.21	SGS (*TBX1*)	NAHR

[a] CGS, contiguous gene syndrome; SGS, single gene syndrome (but other deleted genes may contribute minor features). Terminal deletions are associated with variable proximal breakpoints, and are often the result of unbalanced segregation of a reciprocal translocation in a parent. NAHR, nonallelic homologous recombination; WAGR, Wilms tumor, aniridia, genitourinary abnormalities, and mental retardation; HNPP, hereditary neuropathy with liability to pressure palsies; VCFS, velocardiofacial syndrome.

TABLE 15.5 EXAMPLES OF POPULATION INVERSION POLYMORPHISMS THAT PREDISPOSE OFFSPRING OF HETEROZYGOUS CARRIERS TO NAHR-MEDIATED DISEASE

Location	Size
3q29	1.9 Mb
8p23	4.7 Mb
15q13.3	2.0 Mb
15q24	1.2 Mb
17q12	1.5 Mb
17q21.31	900 kb

The inversions are flanked by low-copy repeats and arise by the mechanism shown in **Figure 15.17**. Data from Antonacci F, Kidd JM, Marques-Bonet T *et al*. (2009) *Hum Mol Genet* **18**:2555–2566; PMID 19383631.

As indicated in **Table 15.4**, syndromes can also be divided into single gene and **contiguous gene syndromes**.

- SMS is an example of a single gene syndrome. In 90% of cases there is a standard 3.7 Mb deletion at 17p11.2 caused by NAHR between flanking repeats. However, some patients have no deletion but just point mutations in the *RAI1* gene that maps in the normally deleted region. Thus SMS is primarily the result of having only a single functional copy of *RAI1*, although loss of other genes in the commonly deleted region probably contributes to the variable phenotype and overall severity of the condition.
- WBS, on the other hand, is a contiguous gene syndrome. As described above, patients are heterozygous for a 1.5–1.8 Mb deletion at 7q11.23 caused by NAHR between complex flanking repeats. All patients have a deletion of the elastin gene, which explains the supravalvular aortic stenosis that is part of the syndrome, but people with mutations affecting just the elastin gene do not have any of the other features of WBS. The deletion includes at least 25 genes, and deletion of specific individual genes is assumed to cause the different individual features of WBS.

Single gene syndromes can result from either microdeletions/duplications or point mutations in the target gene. Which is the more important cause depends on the DNA flanking the causative gene. The Smith–Magenis gene *RAI1* is flanked by repeats that predispose to NAHR, whereas the Alagille gene *JAG1* lacks these. Thus most patients with SMS have microdeletions, while most Alagille patients have point mutations.

Sometimes a microdeletion can produce a pathogenic effect by unmasking a recessive abnormality on the nondeleted homolog. For example, TAR syndrome (thrombocytopenia-absent radius; OMIM #274000) is associated with a recurrent microdeletion on chromosome 1 (1q21), yet many people with the deletion are entirely normal, and affected patients often inherit the microdeletion from an unaffected parent. It turns out that the deletion only causes TAR when, on the homolog, one gene in the deleted region, *RBM8A*, carries one of two low-frequency SNPs (found in 3.05% and 0.42% of a Caucasian population sample) that reduce expression of that gene.

The RBM8A protein has important functions in mRNA processing. Complete absence of the protein would probably be lethal, but the combination of absence from the deleted chromosome and reduced expression from the nondeleted homolog causes TAR syndrome. As a general lesson, if a particular case of a deletion results in an unexpected phenotype, it is always worth checking genes on the nondeleted homolog for possible mutations.

The strange case of neurosusceptibility variants

Investigations of a number of common neurodevelopmental conditions including intellectual disability, schizophrenia, and autism spectrum disorders (ASDs) revealed an unexpected situation. Several NAHR-mediated recurrent variants were identified in different patients; the same variants could also be found in healthy controls, but at significantly lower frequency (**Table 15.6**). Patients often inherited their variant from a parent who was either completely healthy or only borderline affected. Unlike with TAR syndrome (see above), there is no evidence that the deletions acted by unmasking mutated genes on the homolog. Moreover, the same variants were found with significantly raised frequency in more than one of the conditions. Evidently these variants contribute some sort of general susceptibility to a range of neurodevelopmental problems, implying that these supposedly different conditions share common causes.

TABLE 15.6 NEUROSUSCEPTIBILITY VARIANTS

Variant	ID	ASD	SCZ	SCZ cases vs WTCCC[a]	Penetrance %[b]
Del 1q21.1	+		+	20/11,392 vs 1/10,259, $p = 3.2 \times 10^{-5}$	36.9
Dup 1q21.1	+	+			29.1
Del 15q11.2	+	+	+	68/11,863 vs 40/10,259, $p = 0.05$	10.4
Del 15q13.3	+		+	21/10,887 vs 4/10,259, $p = 0.001$	
Del 16p11.2	+	+	+		46.8
Dup 16p11.2	+	+	+	26/8590 vs 4/10,259, $p = 3.9 \times 10^{-6}$	27.2
Del 16p12.1	+	+	+		12.3
Dup 16p13.11	+	+	+		
Del 22q11.2			+	35/11,400 vs 0/10,259, $p = 3.4 \times 10^{-10}$	

Examples of structural variants that have been reported in neurodevelopmental conditions at frequencies above those in neurologically normal controls. Most predispose to more than one neurodevelopmental condition. ID, intellectual disability; ASD, autism spectrum disorder; SCZ schizophrenia; WTCCC, Wellcome Trust Case–Control Consortium (a collection of individuals with various non-neurologic conditions, so for this purpose can be seen as a population of healthy controls).

[a] Data from Grozeva D, Conrad DF, Barnes CP *et al*. (2012) *Schizophr Res* **135**:1–7; PMID: 22130109.

[b] Probability of somebody with the variant having intellectual disability, developmental delay, or congenital anomalies; data from Rosenfeld JA, Coe BP, Eichler EE *et al*. (2013) *Genet Med* **15**:478–481; PMID: 23258348. Other data collated from references cited by Watson *et al*. (2014) (PMID 24773319; see Further Reading).

The poor genotype–phenotype correlation is well illustrated by the report in 2011 by Sahoo and colleagues (PMID 21792059; see Further Reading). A USA commercial testing laboratory identified 1035 cases where array-CGH showed that individuals had a copy number variant of one of six recurrent loci associated with susceptibility to schizophrenia, namely 1q21.1, 15q11.2, 15q13.3, 16p11.2, 16p13.11, and 22q11.2 (see **Table 15.6**). Reviewing the indications for which those cases were referred for testing showed a very diverse set, including developmental delay, intellectual disability, autism spectrum disorder, and multiple congenital anomalies. De Wolf and colleagues (PMID 24123946; see Further Reading) give examples of how a genetic counselor might handle these uncertainties. A further complication is that there is evidence that the severity of the phenotype may depend on the total burden of structural variants across the genome. Several studies have shown that patients with ASD or schizophrenia carry on average a significantly higher number of structural variants compared to healthy controls, and the excess is not due simply to known susceptibility variants.

Mechanisms producing nonrecurrent structural variants

We have seen that recurrent structural variants are usually the result of nonallelic homologous recombination. Nonrecurrent structural variants, on the other hand, probably all have their origins in DNA strand breaks. Our DNA is subject to constant damage that must be repaired to ensure the integrity of the genome. As explained in Section 11.2, cells have a number of different DNA repair mechanisms adapted to different types of lesion. Double-strand breaks (DSB) are the most challenging to repair. Misrepair of DSB can produce structural variants. DSB can be produced by ionizing radiation or as part of the normal recombination process, but they can also arise, as described below, when a replication fork encounters an unrepaired single-strand break or other lesion.

Double-strand breaks can be repaired precisely by the recombination-like process illustrated in **Figure 11.6** if a sister chromatid is available to act as a homologous template. This is only possible in the S/G_2 phase of the cell cycle. DSB near the end of a chromosome may not be repaired but just capped by a telomere, producing a terminal deletion. Otherwise DSB can be repaired by nonhomologous end-joining (NHEJ, **Figure 15.18**). NHEJ usually requires the broken ends to be stripped back and polished to remove single-strand overhangs, so a hallmark of NHEJ-repaired sequences is the presence of small deletions (but sometimes also insertions) at the breakpoint. Thus NHEJ causes small-scale deletions or insertions, but not large structural variants. However, if a replication fork is stalled by encountering an unrepaired single-strand break or a DNA adduct that prevents progress of the polymerase, a series of events can lead to a single-ended DSB (**Figure 15.19**). The cell may then attempt to repair it by NHEJ using any other available double-stranded broken end, and this can produce translocations or the whole range of other major structural variants. A series of articles in the May 2014 issue of *DNA Repair* reviews various aspects of this flexible and complex process.

If no second DSB is available, a complicated sequence of events unrolls. The 3′ end of the broken strand may invade any sequence where there is microhomology (2–6 matching base pairs) and preferably exposed single-stranded DNA, as in a replication fork, a region undergoing repair, or a region with a cruciform or other unusual DNA

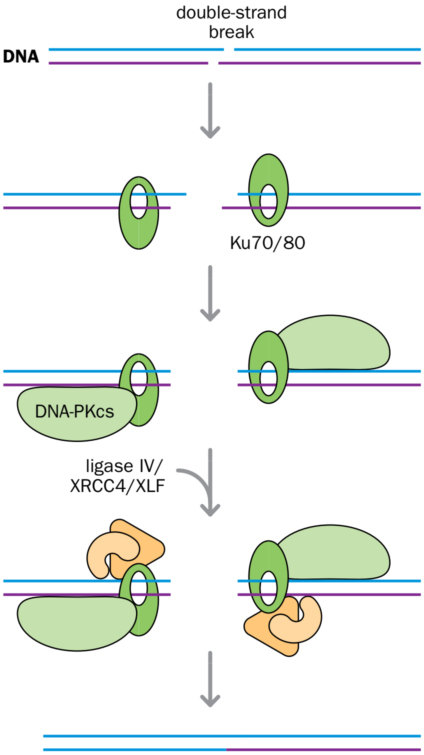


Figure 15.18 Repair of a DNA double-strand break by nonhomologous end-joining (NHEJ). The Ku70/80 heterodimer recognizes and stabilizes the broken ends and recruits various enzymes including DNA protein kinase (DNA-PKcs [catalytic subunit]) to process them. DNA is resynthesized as necessary by polymerases μ and λ, then the ends are ligated by the DNA ligase IV complex (comprising ligase IV, XRCC4, and XLF).

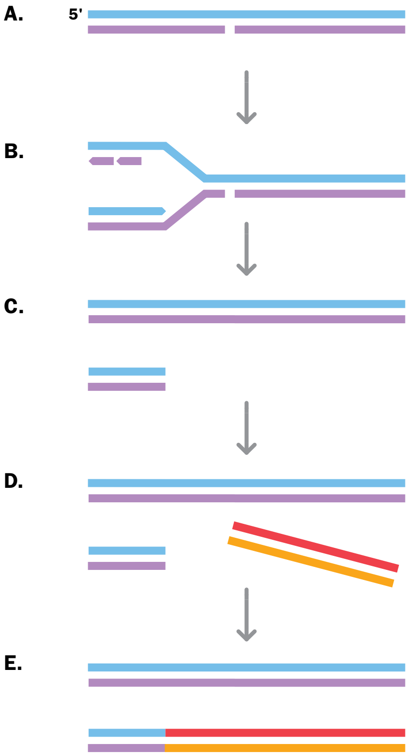


Figure 15.19 Origin of structural variants when a replication fork is stalled. (**A**) The DNA contains an unrepaired single-strand break or base adduct. (**B**) Leading-strand synthesis is stalled, while lagging-strand synthesis can continue. (**C**) The result is a single double-strand break. (**D**) NHEJ joins this to any available broken end from anywhere else in the genome. (**E**) The result is a translocation or other major structural variant. (Adapted from Helleday T *et al.* [2007] *DNA Repair (Amst)* **6**:923–935; PMID 17363343. With permission from Elsevier.)

structure (**Figure 15.20**). A low-processivity replication fork is established, using DNA polymerase θ, but it is unstable, leading to repeated rounds of separation and invasion of novel sequences, before a stable return to the original template is achieved. The result may be overall a duplication or deletion, but including short segments from other genomic regions. The mechanisms are termed microhomology-mediated break-induced replication (MMBIR) or fork stalling and template switching (FoSTeS).

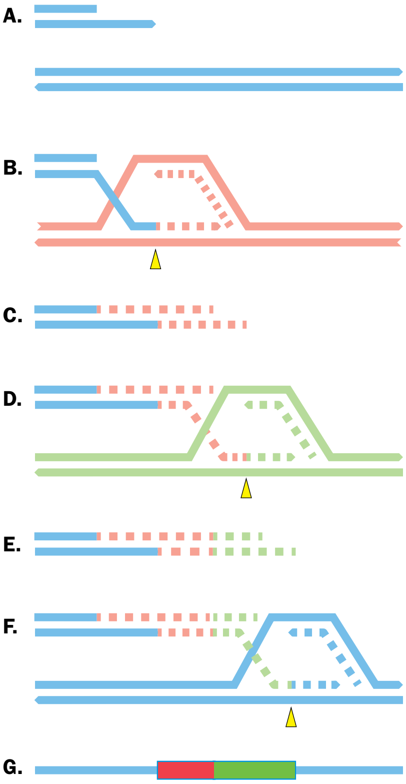


Figure 15.20 Microhomology-mediated break-induced replication. (**A**) A collapsed replication fork forms as in **Figure 15.19C**. (**B**) In the absence of a second double-strand break for NHEJ, the 3′ end invades another sequence, guided by microhomology (yellow arrow), and establishes a replication fork using a poorly processive polymerase. (**C–F**) Repeated episodes of separation and invasion of sequences elsewhere in the genome (shown as different colors) eventually produce (**G**) a stable product incorporating short sequences from elsewhere. (Adapted from Hastings PJ *et al*. [2009] *PLoS Genet* **5**:e1000327; PMID 19180184. With permission from PLoS.)

Detailed examination of nonrecurrent microdeletions and microduplications by Lupski's group and others has confirmed the presence of complex sequence rearrangements that are best explained by episodes of replicative template switching as proposed in the MMBIR/FoSTeS mechanism. **Figure 15.21** shows an example. The papers by Zhang and by Hastings and their colleagues (PMID 19543269 and 19180184; see Further Reading) should be consulted for details of this and other cases.

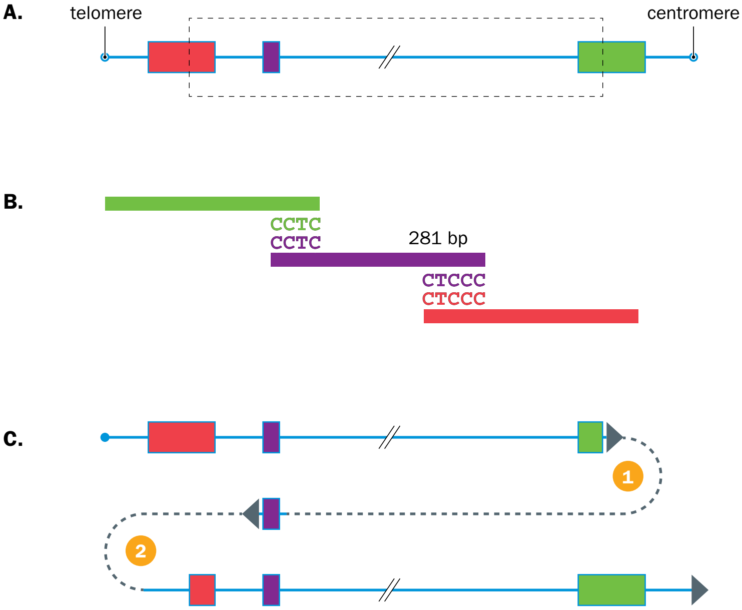


Figure 15.21 An apparently simple nonrecurrent duplication shows hidden complexity. (**A**) Array-CGH showed a patient had an apparently simple duplication of the region on chromosome 17p enclosed in the dotted box. (**B**) Sequencing showed that a 281 bp region within the duplication (too small to show on the CGH array) is triplicated, with one copy in reverse orientation. The boundaries of the inverted triplicated segment are marked by microhomologies (4 bp and 5 bp) with its flanking sequences. (**C**) Interpretation in terms of two template switches. (Adapted from Zhang F *et al*. [2009] *Nat Genet* **41**:849–853; PMID 19543269. With permission from Springer Nature. Copyright © 2009.)

Strand invasion can result in gene conversion

The strand invasion shown in **Figure 15.20** also forms part of the normal recombination process. Normal meiotic recombination is initiated by a double-strand break made by the Spo11 nuclease at sites where the PRDM9 histone methyltransferase has deposited the H3K4me3 mark. Events then follow as depicted in **Figure 11.6**. After strand invasion, synthesis, and ligation, we arrive at a recombination intermediate that has two junctions (called Holliday junctions after Dr Robin Holliday who first described them) linking the two double helices (**Figure 15.22A**). These are resolved by cutting both strands and ligating the ends to the opposite partner. There are two ways the second cut can be made (**Figure 15.22C**). One way produces two recombinant strands, the other way leaves the strands nonrecombinant.

Any sequence differences between the two original double helices will result in mismatches (in **Figure 15.22D** these will be in sequences where a blue strand is paired with a purple strand). Mismatch repair enzymes will correct these by stripping back and resynthesizing one of the strands, chosen at random. The result can be to patch short (typically 1 kb) stretches of sequence derived from one parental double helix into the other, without making any reciprocal replacement. This nonreciprocal transfer of sequence is known as **gene conversion**.

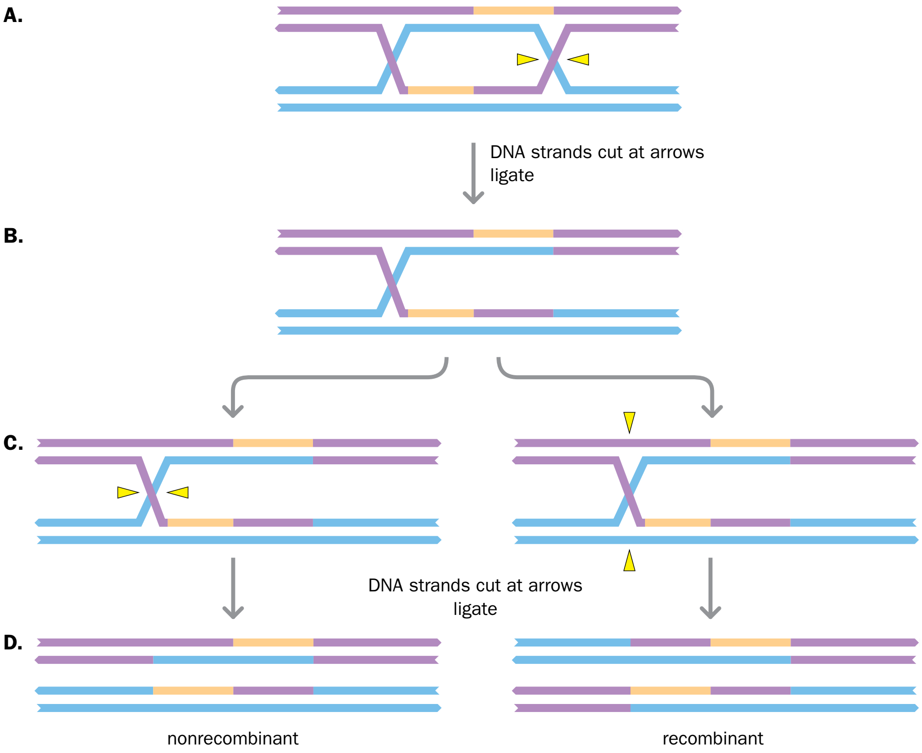


Figure 15.22 Two ways of resolving recombination intermediates. (**A**) The structure with a double Holliday junction that is formed after an initial double-strand cut followed by stripping back, strand invasion, DNA synthesis, and ligation, as shown in **Figure 11.6**. (**B–D**) The Holliday junctions are resolved by cutting two strands and religating the cut ends to the opposite partner. It can be done in either of two ways, one of which results in recombination, the other in gene conversion (see text).

When the two double helices involved are truly homologous, then the only effect of gene conversion is to make affected alleles segregate 3:1 instead of 2:2. In some fungi, where all four products of meiosis are lined up together in an ascus, the 3:1 segregation can be observed, but in humans one would not notice anything. Gene conversion may serve to homogenize tandemly repeated sequences, as for example the arrays of ribosomal RNA genes on the short arms of the acrocentric chromosomes. However, when gene conversion occurs in the context of nonallelic homologous recombination involving a functional gene and a nearby pseudogene, the result can be to replace part of the functional gene with corresponding pseudogene sequence. **Figure 15.23** shows an example. The *CYP21A2* gene at 6p21.3 encodes a steroid 21-hydroxylase enzyme that has an essential role in synthesis of adrenal hormones. A nearby pseudogene *CYP21A1P* has a closely similar sequence but with frameshifts and a stop codon (see Section 16.1) that render it nonfunctional. Three-quarters of patients with 21-hydroxylase deficiency (OMIM #201910) have variants of *CYP21A2* that have incorporated nonfunctional sequence from the pseudogene by gene conversion.

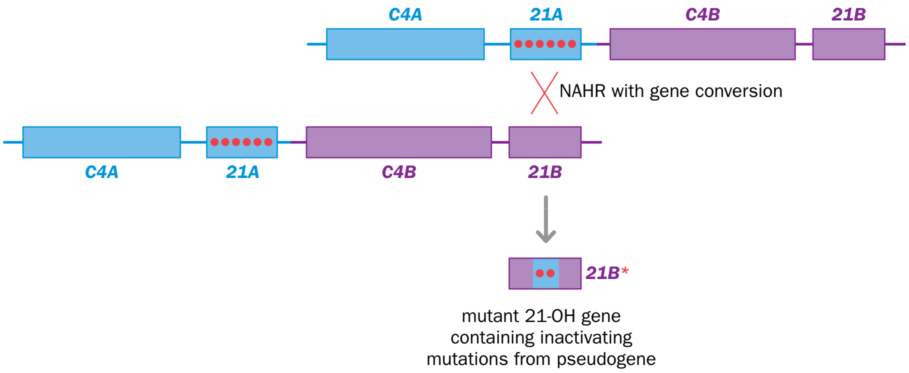

C4A
21A
C4B
21B
NAHR with gene conversion
C4A
21A
C4B
21B
21B*
mutant 21-OH gene containing inactivating mutations from pseudogene

Figure 15.23 Gene conversion inserts sequence from the *CYP21A1P* pseudogene (labeled 21A here) into the functional *CYP21A2* gene (labeled 21B). The pseudogene sequence contains inactivating mutations (red dots); the result is to make the 21B gene nonfunctional. 21-OH, 21-hydroxylase.

Conclusions

This chapter has described a range of abnormalities involving wrong numbers or structures of chromosomes or that, like gene conversion, are produced by essentially chromosomal mechanisms. Next-generation sequencing has removed the traditional boundary between chromosomal abnormalities, studied by cytogeneticists, and abnormalities identified by sequencing that are studied by molecular geneticists. Whole-genome sequencing can identify all abnormalities, from gross chromosomal aberrations all the way down to single nucleotide substitutions. However, cytogenetic insights are still needed to understand how abnormalities can arise through nondisjunction, segregation of structural variants, nonallelic homologous recombination, and other chromosomal mechanisms. The major question, now that we can have a unified view of all genetic abnormalities, is why some are pathogenic and how they cause disease. This is the subject of the following chapter.

SUMMARY

- This chapter describes genetic abnormalities that affect the phenotype of a person. There is an unbroken spectrum, ranging from complete extra or missing chromosomes all the way to changes involving just a single nucleotide.

- In times past, different laboratory techniques were used to study different-sized abnormalities. While a range of methods are still in current use, increasingly DNA sequencing is being used to study every type and size of abnormality.

- Chromosomes can be seen under the microscope only in dividing cells. Usually these are cells undergoing mitotic cell division in laboratory culture. Human meiotic chromosomes can only be studied in testicular biopsies, or in the ovaries of female fetuses.

- Chromosomes can be made to stain in a reproducible and recognizable pattern of dark and light bands; chromosomal locations are described by reference to these bands.

- For clinical purposes, chromosome analysis under the microscope has been largely superseded by array-based molecular analyses, but traditional karyotypes are still the best tool for understanding the origin and effects of chromosome abnormalities.

- Chromosome abnormalities can be numerical or structural, and can be constitutional (present in every cell of the body) or mosaic (present only in some cells).

- Numerical abnormalities, most commonly trisomy (one extra chromosome), usually arise from errors in meiosis or, for mosaic abnormalities, in mitosis. Constitutional numerical abnormalities are incompatible with survival of embryos to term, with the exception of trisomies 13, 18, and 21 and sex chromosome abnormalities.

- Structural abnormalities such as deletions, insertions, translocations, and inversions are often the result of errors in DNA replication or repair. When they involve no loss or gain of material, they are termed balanced. Balanced abnormalities usually have no phenotypic consequences, but will cause problems in meiosis. Unbalanced abnormalities may or may not cause phenotypic effects, depending on the number and identity of the genes affected.

- Microdeletions and microduplications are structural variants that are too small to be seen under the microscope. They normally involve less than about 5 Mb of DNA, but can still include numerous genes.

- Recurrent microdeletions and microduplications are usually the result of nonallelic homologous recombination between low-copy repeats that flank the affected sequence. Nonrecurrent variants may be produced by various mechanisms that usually involve failure to repair double-strand DNA breaks correctly.

- In some cases, the phenotypes caused by microdeletions or microduplications represent the combined effects of deletion or duplication of several genes in the affected region (contiguous gene syndromes); in others, the main features can be ascribed to a missing or extra copy of a single gene.

FURTHER READING

Gross chromosome abnormalities

Gardner RM, Sutherland GR & Shaffer LG (2011) *Chromosome Abnormalities and Genetic Counselling*, 4th edn. Oxford University Press.

Shaffer LG, Slovak ML & Campbell LJ (eds) (2009) ISCN 2005: *An International System for Human Cytogenetic Nomenclature*. S Karger AG.

Microdeletions

Lupiáñez DG *et al*. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**:1012–1025; PMID 25959774.

Radhakrishnan SK *et al*. (2014) Non-homologous end joining: emerging themes and unanswered questions. *DNA Repair (Amst)* **17**:2–8; PMID 24582502. (See also subsequent reviews in the same issue of the journal.)

Watson CT *et al*. (2014) The genetics of microdeletion and microduplication syndromes: an update. *Annu Rev Genomics Hum Genet* **15**:215–244; PMID 24773319.

Copy number variants

Hastings PJ (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5**:e1000327; PMID 19180184.

Itsara A *et al*. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**:148–161; PMID 19166990.

Zhang F *et al*. (2009) The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**: 849–853; PMID 19543269.

Neurosusceptibility variants

De Wolf V *et al*. (2013) Genetic counseling for susceptibility loci and neurodevelopmental disorders: the del15q11.2 as an example. *Am J Med Genet* **161A**:2846–2854; PMID 24123946.

Grozeva D *et al*. (2012) Independent estimation of the frequency of rare CNVs in the UK population confirms their role in schizophrenia. *Schizophr Res* **135**:1–7; PMID 22130109.

Rosenfeld JA *et al*. (2013) Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet Med* **15**:478–481; PMID 23258348.

Sahoo T *et al*. (2011) Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. *Genet Med* **13**:868–880; PMID 21792059.

16

Molecular pathology: connecting phenotypes to genotypes

Molecular pathology has moved center stage in our attempts to understand genetic effects on human characters. This has not always been the case. Until recently the main focus was on identifying variants. Clinicians and scientists pondered hard to think what candidate gene might harbor mutations responsible for a syndrome; they invested huge efforts into identifying informative families or special patients who might provide a vital clue, and they used their laboratory skills to pinpoint the chromosomal location of those mutations. The story finished in triumph with the identification of mutations in a candidate gene.

Next-generation sequencing changed all that. Now that it is routine to sequence the entire exome or, increasingly, the entire genome of a patient, identifying variants is trivial. The problem lies in sorting through the list of variants to pinpoint the crucial one. An exome sequence will typically identify 20,000 differences between the exome of that individual and the Reference Human Genome. A whole genome sequence would identify 4–5 million variants. Earlier approaches narrowed down the search space by first identifying a small candidate chromosomal region or even a single candidate gene. Now that the search space covers the entire exome, or even the whole genome, for any one of the variants the prior probability that it would be the cause of the subject's condition is extremely low. It follows that the evidence that a variant is causal needs to be correspondingly strong in order to convince a skeptical world that the cause really has been identified. Thus molecular pathology moves center stage.

The fundamental distinction in molecular pathology is between loss of function and gain of function. A variant may cause a phenotype because it fails to do something that its normal counterpart does, or it may cause it because it does something that the normal version does not do. A loss of function may be total or partial, and if a gene product has several functions the loss may affect only one of them or all of them. A total loss of function is often the result of failure to make any of the gene product. A gain of function is rarely the acquisition of a totally novel function; more usually it is a failure of regulation, so that the gene product functions inappropriately. It may be expressed at the wrong time, in the wrong place, at the wrong level. It may misinterpret signals or respond to the wrong signal. The distinction between loss and gain of function can sometimes be blurred. A regulatory change can cause a gene product to lose some functions and acquire others. Loss of function of an inhibitory molecule can cause an effect through gain of function of its target. Nevertheless, the first question to ask about a variant is whether it causes a loss or a gain of function.

In all of this it is important to distinguish what a variant does at the biochemical level—production of the gene product and its biochemical function—from what the variant does to the person carrying it. Large-scale sequencing surveys have shown that normal healthy people often carry variants that inactivate one or another gene or cause significant changes in a gene product. Changing or abolishing the function of a gene is not necessarily pathogenic. Thus there are always two questions in molecular pathology: what a variant does to a gene, and what it does to the person. The first of these questions is the easier one to answer, and we will start by looking at how variants can cause a loss or gain of function of a gene product.

16.1 LOSS OF FUNCTION

The many ways a gene product can lose function are summarized in **Table 16.1**.

TABLE 16.1 THE MAIN TYPES OF CHANGE THAT CAN CAUSE A GENE PRODUCT TO LOSE FUNCTION

Changes that apply whether the gene product is a protein or a functional RNA
Delete all or part of the gene
Disrupt the gene by a chromosomal rearrangement
Prevent or reduce transcription of the gene by deletion or alteration of the promoter
Remove access to a necessary enhancer by a chromosomal rearrangement or by activating an insulator element
Abolish correct splicing of the primary transcript, or of specific splice isoforms
Changes that apply when the gene product is a protein
Delete or change the AUG translation initiation codon
Insert or delete nucleotides so as to cause a frameshift
Change a codon for an amino acid into a UAG, UAA, or UGA stop codon (a nonsense change)
Change a codon for one amino acid into one for a different amino acid (a missense change)

Deletion or disruption of a gene would normally lead to a loss of function

Complete deletion of a gene will necessarily mean absence of product from that allele. Partial deletions will usually have the same effect. Because most exons are small compared to most introns or the stretches of DNA between genes, most random breakpoints lie in intergenic DNA or in introns. Thus partial deletions often involve loss of one or more complete exons. Loss of exons can lead to loss of gene function for several reasons. A deletion including the first exon of a gene would often also include the promoter and other crucial upstream elements. For a protein-coding gene, deletion of the exon containing the ATG start codon would normally prevent translation, although some genes have secondary internal ribosome entry sites downstream of the normal start codon. The final exon of a protein-coding gene might or might not contain coding sequence, but the 3′ untranslated regions of mRNAs contain sequences important for their stability. Thus deletion of the final exon of a gene is likely to make the transcript unstable. Deletions of internal exons can have varying effects. As explained below, in two-thirds of cases one would expect deletion of a coding exon to introduce a frameshift, with the result that no functional protein would be produced. Even an in-frame deletion may remove a vital part of the protein, or affect its three-dimensional structure so as to make the protein nonfunctional. Duplications of internal exons might have similar effects. A similar argument applies to small deletions of sequences lying entirely within a single exon: they can cause a frameshift or alter the structure of the encoded protein. However, it is important to remember that most protein-coding genes have multiple splice isoforms. If an isoform does not use a certain exon, it would be unaffected by deletion of that exon or any other change in it.

Chromosomal rearrangements, even if balanced, can affect function by disrupting a gene. A classic example is hemophilia A, an X-linked condition (OMIM #306700) where blood fails to clot because of a deficiency of clotting Factor VIII. Affected patients may have a variety of loss-of-function changes in the *F8A* gene that encodes Factor VIII, but around half of all cases of severe disease are caused by an inversion that disrupts the gene. This is the result of intrachromosomal recombination between low-copy repeats in the distal part of the X chromosome (**Figure 16.1**). Exons 1–22 are separated and in opposite orientation to the remaining exons, causing a complete loss of function. Each of the 26 *F8A* exons is present with its correct sequence, and so no abnormality would be detected by exome sequencing.

For functional (noncoding) RNAs the effects of a partial deletion or disruption are harder to predict. These RNAs seem in general to be less sensitive to sequence changes

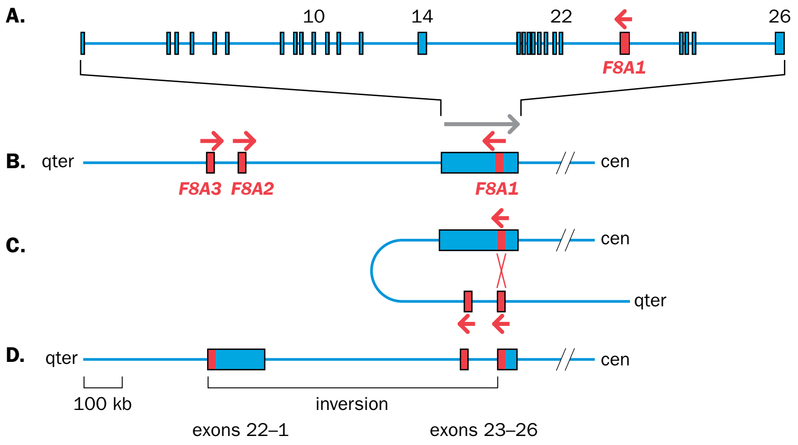


Figure 16.1 Hemophilia A can be caused by an inversion that disrupts the *F8A* gene. (**A**) There is a repetitive sequence in intron 22 of the *F8A* gene (*F8A1*, red bar); (**B**) two additional copies are located 360 kb and 435 kb upstream of the *F8A* gene. Arrows indicate the relative orientations of the three copies. (**C**) During male meiosis, this part of the X chromosome has no homologous pairing partner. The *F8A* repeats may pair, forming a loop. (**D**) A crossover between paired *F8A* repeats causes inversion of a 500 kb segment. Although the *F8A* gene is disrupted and nonfunctional, each individual exon and its flanking intronic sequence is still intact.

than are protein-coding sequences. However, many long noncoding RNAs are spliced and polyadenylated just like protein-coding transcripts, and so might be sensitive to structural rearrangements. For some antisense RNAs, the function seems to lie in the act of transcription, rather than any properties of the transcript itself. Transcribing the RNA prevents transcription of an overlapping gene on the other DNA strand. A deletion could abolish that effect, but it would likely also delete some or all of the overlapping gene, and thus have a more direct effect on the cell.

Loss of function could be due to deletion or alteration of the promoter

Many cases have been documented of patients with loss-of-function conditions having a nucleotide substitution in a transcription factor binding site upstream of the transcription start site of the relevant gene (Table 16.2). However, such reports should be treated with a degree of caution. Without detailed functional studies it is difficult to know how often such variants are truly causative. To establish causation, first of all it would be necessary to show that the variant does indeed affect the level of transcription, by using a transient transfection assay in a relevant cell type or quantifying the mRNA level. Electrophoretic mobility shift assays (EMSA) could then be used to identify the binding partner. The electrophoretic mobility of a wild-type promoter fragment should be reduced when run together with a nuclear extract, showing that the extract contained protein(s) that bound the fragment, increased its weight, and thus slowed its migration. The variant sequence should show altered behavior. Analysis of the wild-type promoter sequence for known binding sites could suggest candidate binding proteins. These could be checked by using purified protein in the EMSA. Alternatively, adding an antibody

TABLE 16.2 EXAMPLES OF REPORTED PATHOGENIC MUTATIONS IN PROMOTER SEQUENCES

Condition	OMIM #	Gene	Variant	Disrupted regulatory element
β-thalassemia	141900	*HBB*	c.–101C>T c.–30T>A	CACCC box (binds Sp1 and EKLF) TATA box
Hemophilia B	300746	*F9*	c.–26G>C c.–6G>A	HNF4 binding site HNF4 + other factors
Familial hypercholesterolemia	143890	*LDLR*	c.–139C>G c.–60C>T	Sp1 site Sterol regulatory element 2
Pyruvate kinase deficiency	266200	*PKLR*	c.–83G>C c.–72A>G	PKR-RE1 GATA1
[Many cancers]	187270	*TERT*	c.–91C>T, c.–69C>T	Activating mutations, create ETS binding sites
Cowden syndrome	158350	*PTEN*	c.–930G>A c.–920G>T	Sp1 binding site Sp1 binding site
Congenital erythropoietic porphyria	263700	*UROS*	c.–90C>A c.–70T>C	CP2 binding site GATA1 binding site

to that protein to a nuclear extract should produce a supershift in the electrophoretic mobility, as the large protein–antibody complex binds to the promoter fragment. Such functional studies are the province of research laboratories, and so diagnostic laboratories tend to be cautious about either seeking or reporting promoter variants.

The most frequent changes that cause loss of function of a promoter are epigenetic. As described in Chapter 10, epigenetic marks define promoters and regulate their normal activity. Thus it is not surprising that abnormal epigenetic marks can silence promoters. Cancer cells often silence crucial growth-suppressing genes by methylation of the promoter (see Chapter 19). Epigenetic silencing of a promoter can also cause other genetic diseases. Fragile X syndrome (OMIM #300624) is caused by lack of the FMR1 RNA-binding protein. The usual cause is silencing of the *FMR1* gene promoter by methylation, triggered by an expanded trinucleotide repeat (see **Table 16.7**).

Removing access to an enhancer can cause a tissue-specific loss of function

As described in Chapter 10, expression of many genes depends on enhancers. The enhancers bind transcription factors and other proteins, and loop round to come into close proximity to the promoter that they control (see **Figure 10.24**). Deleting or mutating an enhancer can abolish or change expression of the gene. **Table 10.5** showed examples of multisystem conditions caused by loss of function of a protein-coding gene that is controlled by multiple tissue-specific enhancers. The table showed how inactivating one of the enhancers, while leaving the coding sequence intact, can cause just one component of the full syndrome. Each enhancer is responsible for one facet of the tissue-specific gene expression; loss of an enhancer causes a tissue-specific loss of function (although important developmental genes often have multiple redundant enhancers).

Enhancer–promoter interactions are limited by the domain structure of chromosomes. As described in Chapter 10, chromosomes are partitioned into megabase-size topologically-associated domains (TADs) separated by insulator or barrier elements. An enhancer may control expression of a gene that lies within the same TAD, but not genes that lie outside the TAD. If an insulator element is placed between an enhancer and its target gene, that enhancer would no longer have any influence on gene expression. The insulator might be placed because of a chromosome structural rearrangement (see **Figure 16.9**, below, and Lupiáñrez *et al*., [2015] [PMID 25959774; see Further Reading] for examples), or a potential insulator might be regulated epigenetically. **Figure 10.17** showed how activity of a potential insulator located between an enhancer and the *IGF2* gene at 11p15 depended on its state of methylation. When the sequence was methylated it had no insulator function, allowing enhancer-driven *IGF2* expression. When not methylated it acted as an insulator and *IGF2* was not expressed. Chromosome structural variants may allow an enhancer to control expression of a different gene from its normal target ("enhancer capture"). That gene would gain function. Such changes are discussed below when we consider gain-of-function variants.

Aberrant splicing is a frequent cause of loss of function

Mutations that alter splice sites are one of the most frequent causes of loss of function of a gene. The consequences of loss of a splice site can vary (**Figure 16.2**). Often the affected exon is skipped, but sometimes a cell will use an alternative sequence nearby. Splice sites are not clearly defined all-or-nothing sequences. The context of the invariant GU or AG is

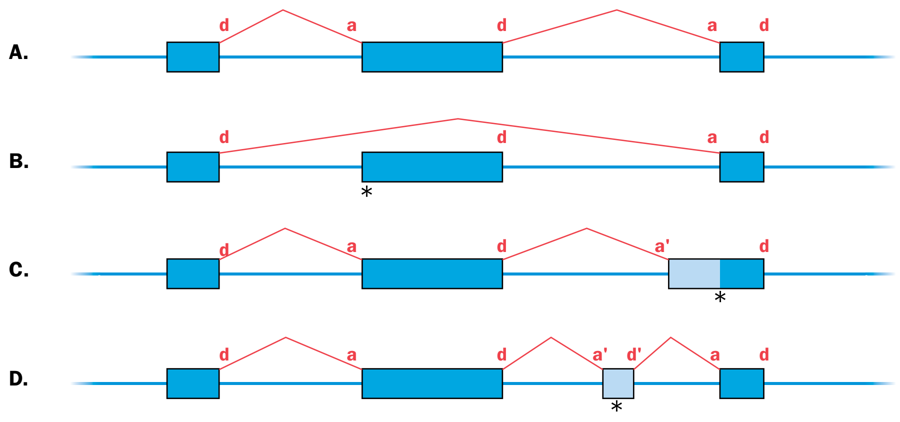


Figure 16.2 Possible consequences of a splice-site mutation. (**A**) In the wild-type sequence, the three exons are spliced as shown. Donor sites (GU) are labeled d, acceptor sites (AG) as a. (**B**) A sequence change (asterisk) has abolished the acceptor site at the 3′ end of intron 1. Exon 2 is skipped. (**C**) A sequence change has abolished the acceptor site at the normal end of intron 2. Instead of skipping exon 3 (similar to [**B**]), the spliceosome has found a nearby alternative acceptor-like sequence, a′ in intron 2, to use. The spliced transcript includes some intronic sequence (pale color). (**D**) A sequence change has activated a cryptic acceptor site, a′ in intron 2. A downstream sequence d′ is used as a new donor site, creating a novel extra exon.

crucial, and there are strong and weak splice sites. Spliceosomes assemble on the nascent RNA transcript as it is being synthesized. Generally they will choose the strongest available site in the emerging transcript, but if a mutation has weakened or inactivated a strong site that is normally used, a nearby weaker site (a "cryptic splice site") may be used instead.

Changes to the (almost) invariant GU…AG sites at the ends of introns will always prevent the spliceosome from using that sequence. Other changes nearby also often affect splicing, but the effect is harder to predict. Various computer programs are used to predict the effect of changes in the sequences immediately flanking a splice site (**Box 16.1**). Generally they have been trained with the experimentally verified effects of a large panel of variants, and they may also incorporate data on binding sites of SR and other splice-modifying proteins. A study of prediction programs by the UK National Genetics Reference Laboratory in 2009 showed that predictions were typically 60–85% correct for changes outside the invariant GU…AG dinucleotides. Different programs have different strengths and weaknesses, and laboratories often use a consensus approach in assessing variants. A recent, very comprehensive analysis based on over 650,000 intronic and exonic variants by Xiong and colleagues (PMID 25525159; see Further Reading) may point the way to more reliable predictions.

BOX 16.1 PREDICTING SPLICING EFFECTS

A number of freely available Web-based programs try to predict splice sites in the transcript of a DNA sequence. Examples include:

- GeneSplicer (www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml);
- Human Splicing Finder (www.umd.be/HSF3/);
- MaxEntScan (genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html);
- NNSplice (www.fruitfly.org/seq_tools/splice.html) (includes an option for human sequences).

The query sequence can be pasted in, and the program will report the positions of possible splice sites, with a score to indicate how strong each site is. The Human Splicing Finder program will also report predicted motifs for intronic and exonic splicing enhancers and suppressors, which can be helpful when assessing the likely effect of a sequence change. See Xiong *et al.* (2015) (PMID 25525159; in Further Reading) for a more systematic effort to predict splicing.

Experimentally, splicing can be checked by sequencing the mRNA or by using a **minigene assay**. The exon of interest, together with flanking intronic sequence, is inserted into a vector that has a strong promoter. The vector is transfected into splicing-competent cells. RNA is extracted and the transcripts from the vector are amplified by RT-PCR and sized on a gel. **Figure 16.3** shows an example.

When the effect of a change in a coding sequence is assessed, normally the question is whether it will change the amino acid sequence of the encoded protein or not, and if it does, what effect that change might have on the function of the protein. It is important to remember that the change might also affect splicing. **Figure 16.4** shows two examples of apparent missense or synonymous mutations in exons that prevent correct splicing.

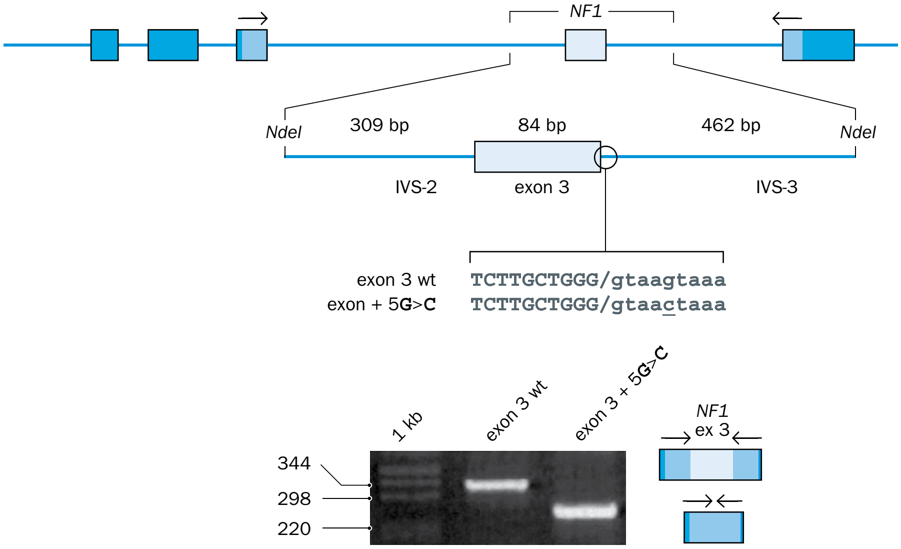


Figure 16.3 A minigene assay to assess splicing. A patient had a G>C substitution in intron 3 of the *NF1* gene, five nucleotides downstream of the donor splice site. To check its effect on splicing of the transcript, the wild-type (wt) and mutant sequences of exon 3, together with flanking intron (IVS-2 and -3), were cloned into NdeI restriction sites of a vector driven by a strong promoter, shown at the top. The constructs were transfected into splicing-competent HepB3 cells and RNA was extracted. RT-PCR based on flanking vector exons (arrows) amplified the spliced products. The electrophoretic gel at the bottom shows that the mutation prevented inclusion of exon 3 in the spliced product. (Adapted from Baralle M *et al.* [2003] *J Med Genet* **40**:220–222; PMID 12624144. With permission from BMJ Publishing Group Ltd.)

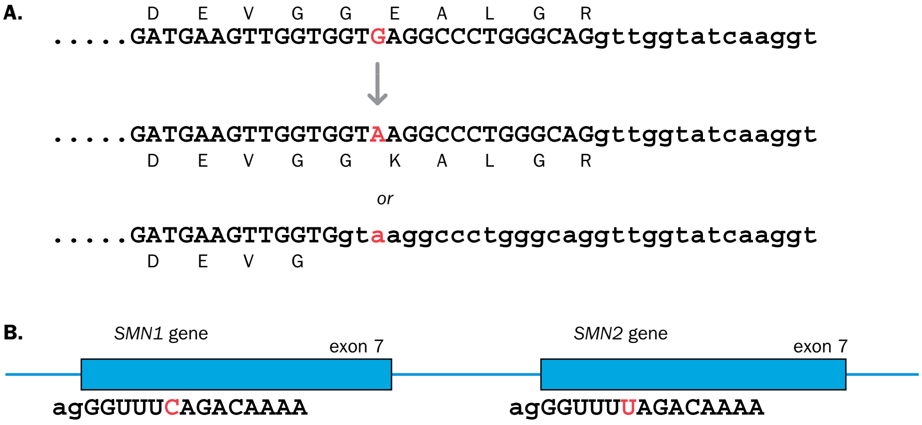

A.
D E V G G E A L G R
.....GATGAAGTTGGTGGTGAGGCCCTGGGCAGgttggtatcaaggt
.....GATGAAGTTGGTGGTAAGGCCCTGGGCAGgttggtatcaaggt
D E V G G K A L G R
or
.....GATGAAGTTGGTGgtaaggccctgggcaggttggtatcaaggt
D E V G
B.
SMN1 gene
exon 7
agGGUUUCAGACAAAA
SMN2 gene
exon 7
agGGUUUUAGACAAAA

Figure 16.4 Nucleotide substitutions in exons that affect splicing. (**A**) The G>A change in exon 1 of the β-globin gene would cause the protein change p.Glu26Lys, as shown. But it also creates an additional weak donor splice site. Any transcript that is spliced using this novel site is nonfunctional. Only some transcripts are spliced in this way; the others use the normal site as shown (exons in uppercase, introns in lowercase letters). The effect is a mild (β^+) thalassemia. (**B**) Spinal muscular atrophy (OMIM #253300) is caused by lack of SMN protein. A repetitive structure on chromosome 5q13 contains two copies of the *SMN* gene. The *SMN1* transcript is efficiently spliced, but 90% of the spliced *SMN2* transcripts skip exon 7 and are nonfunctional. The difference is a C>U change six nucleotides from the start of the exon. This would be predicted to be a silent mutation, replacing one phenylalanine codon (UUC) by another (UUU), but in fact it creates an exonic splice suppressor motif. Individuals with no functional *SMN1* gene have spinal muscular atrophy, despite having intact *SMN2* genes.

Such effects may be underreported since diagnostic laboratories do not routinely use minigene assays or RNA sequencing to check splicing. Di Giacomo and colleagues (PMID 23983145; see Further Reading) used minigene assays to check for possible splicing effects of 36 mutations scattered across exon 7 of the *BRCA2* gene. Eleven of the 36 mutants affected inclusion of the exon.

Mutations that affect splicing sometimes act by creating novel splice sites

Sequence changes that lead to loss of a splice site are common causes of loss of gene function, but so are changes that create novel splice sites. **Figures 16.2D** and **16.4A** showed examples. Random sequences in exons or introns of genes may by chance bear some resemblance to a splice site. As long as the cell has a better alternative to use, those "cryptic splice sites" will be ignored. Sometimes a nucleotide change in a cryptic site will convert it into an effective site that may be used in competition with, or in preference to, the normal site. In the case of hemoglobin E (**Figure 16.4A**) the variant is in an exon and would be picked up by normal mutation screening, although only RNA studies would reveal its effect. In other cases the cryptic site may lie deep within an intron and would not be detected by exome sequencing. For example, one cause of cystic fibrosis is a single nucleotide change that activates a cryptic splice site deep within the very large intron 22 of the *CFTR* gene, c.3849+12191C>T. As with hemoglobin E, only a proportion of transcripts are spliced using this novel site; some correctly spliced transcript is still produced, and the resulting disease (in homozygotes or compound heterozygotes) is mild.

Table 16.3 shows the five β-globin mutations that collectively account for over 98% of all β-thalassemia in Greek Cypriots. Between them they sum up the possible ways that splicing can be affected.

TABLE 16.3 β-GLOBIN MUTATIONS RESPONSIBLE FOR THALASSEMIA IN GREEK CYPRIOTS

Mutation	Location	% of all cases	Sequence change	Effect
c.92+1G>A	Intron 1	5.1	AG**g**ttggtat AG**a**ttggtat	Abolishes a canonical gt donor site
c.92+6T>C	Intron 1	5.5	AGgttgg**t**at AGgttgg**c**at	Weakens a normal splice site
c.93–21G>A	Intron 1	79.8	ctatt**g**gtctattttccc ctatt**a**gTCTATTTTCCC	Activates a cryptic splice acceptor site in intron 1
c.316–106C>G	Intron 2	5.1	cag**c**taccat CAG**g**taccat	Activates a cryptic splice donor site in intron 2
p.Gln39*	Exon 2	2.9	TGGACC**CAG**AGGTTC TGGACC**TAG**AGGTTC	Creates a premature termination codon

Normal and mutant sequences are shown, with the change highlighted in bold. Exonic sequences are in uppercase, intronic in lowercase. Only one of the five most frequent mutations is a conventional coding-sequence change. One other inactivates a canonical splice site. Both these changes completely abolish gene function, and in homozygotes cause a severe β^0 disease (complete absence of β-globin). The other three variants have more subtle effects on splicing; some transcripts are still normally spliced, and homozygotes have a milder β^+-thalassemia (some β-globin is present). (Data from HbVar database, http://globin.cse.psu.edu/globin/hbvar)

Changes affecting the AUG initiation codon will affect translation

Changing the initiator AUG into another sequence should prevent initiation at that position. However, in some cases this can be bypassed. Normally the ribosome attaches to the 5′ cap of an mRNA and then scans in the 5′ → 3′ direction, as described in Section 1.5, until it encounters an AUG codon in the right context to initiate translation. If a sequence change prevents the normal start site from being recognized, the ribosome may continue scanning and initiate translation at a downstream AUG. The change may or may not have an important effect on the function of the protein product. It has also long been known that some genes have cap-independent ribosome entry sites further downstream, and a study by Weingarten-Gabbay and colleagues (PMID 26816383; see Further Reading) has suggested that these are much more widespread than previously thought. Again, they provide the possibility of initiating translation at a downstream AUG codon when the normal initiator codon is nonfunctional. Strong RNA structures (hairpins and stem-loops, see **Figure 10.32**) sometimes trigger initiation at a non-AUG codon. The strange case of repeat-associated non-ATG (RAN) translation is discussed in Section 16.2 below.

Changes near the normal initiation codon might also affect its use. To be recognized as an initiator codon, the AUG needs to be embedded in a Kozak sequence with the consensus 5′GCCPuCC**AUG**G3′ (Pu = purine; see Section 1.5). Other changes in the 5′ untranslated region of an mRNA may affect initiation. Formation of strong RNA secondary structures may inhibit the progress of the ribosome. Moreover, up to half of all mammalian mRNAs have upstream open reading frames—that is, potentially translatable sequences upstream of the main coding sequence (see **Figure 10.31**). These can divert ribosomes from reaching the normal start codon. Mutations that prevent ribosomes recognizing the upstream open reading frame are likely to enhance expression of the main gene product (see the discussion of gain-of-function mechanisms below).

Insertions or deletions often cause frameshifts

Successive triplet codons in a messenger RNA are not separated by any form of punctuation. As described in Section 1.5, the reading frame is set by the AUG initiation codon. Any insertion or deletion of nucleotides downstream has the potential to create a frameshift (**Figure 16.5**). Since 3 of the 64 possible codons are stop codons, a frameshifted message will usually fairly soon include a stop codon. **Figure 16.6** shows an example.

raw sequence	ISAWTHEBIGBADDOGEATTHECAT
WITH READING FRAME	I SAW THE BIG BAD DOG EAT THE CAT
INSERT 1 LETTER	I SAW XTH EBI GBA DDO GEA TTH ECAT
DELETE 1 LETTER	I SAT HEB IGB ADD OGE ATT HEC AT
INSERT 2 LETTERS	I SAW XYT HEB IGB ADD OGE ATT HEC AT
INSERT 3 LETTERS	I SAW THE BIG BAD RED DOG EAT THE CAT
DELETE 3 LETTERS	I SAW THE BIG DOG EAT THE CAT

Figure 16.5 The reading frame. Where the number of inserted or deleted letters is not an exact multiple of 3, a frameshift is generated and the downstream message is wrecked. Thus, two-thirds of random deletions or insertions are likely to produce a frameshift.

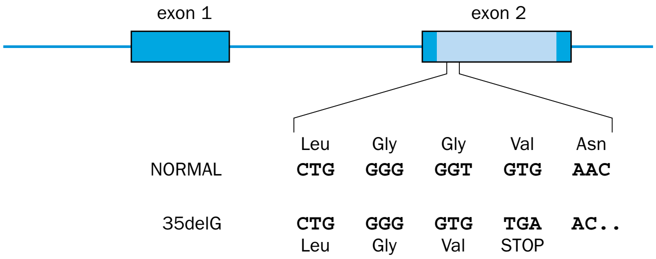


Figure 16.6 Deletion of a single nucleotide in the *GJB2* gene produces a frameshift, leading to an early stop codon. The *GJB2* gene has a run of six consecutive G nucleotides in exon 2. Such homopolymer runs are hotspots for mutation by slippage during DNA replication. The protein product, connexin 26, has essential functions in the inner ear. Homozygosity for a variant, c.35delG, that omits one G, is the cause of almost half of all congenital deafness in many Western countries.

Deletions of whole exons will similarly frequently produce frameshifts, and this explains the counterintuitive molecular pathology of dystrophin deletions. Around 65% of cases of the severe Duchenne or the milder Becker muscular dystrophy (both OMIM #310200) are due to deletion of one or more exons of the huge dystrophin gene at Xp21. In the central part of the gene, deletion of any one of exons 43–46 causes the severe disease, but deletion of all four, or of pairs of adjacent exons (such as exons 43 + 44, 44 + 45, or 45 + 46), causes the milder condition. The outcome depends not on the size of the deletion but on whether or not it produces a frameshift (Table 16.4). This result should not be generalized too readily to other proteins. Even an in-frame deletion will often result in production of an inactive or unstable protein, by deleting amino acids that were essential to the activity or three-dimensional structure of the protein. Dystrophin is an unusual protein. It is like a rope with hooks on each end. It links the contractile apparatus of muscle cells to the outer membrane. The two hooks are essential to its function, but the rope will still function, albeit less efficiently, if a gene deletion makes it a bit shorter.

TABLE 16.4 EXON DELETIONS IN THE DYSTROPHIN GENE CAUSE THE SEVERE DUCHENNE OR MILD BECKER MUSCULAR DYSTROPHY DEPENDING ON WHETHER OR NOT THEY PRODUCE A FRAMESHIFT

Exon	Size (bp)	Frame	Effect of single-exon deletion	Multiexon deletions causing BMD		
42	195	0	BMD			
43	173	–1	DMD	Deletion of exons 43 + 44		Deletions of exons 43–46
44	148	+1	DMD		Deletion of exons 44 + 45	
45	176	–1	DMD	Deletion of exons 45 + 46		
46	148	+1	DMD			
47	150	0	BMD			
48	186	0	BMD			

Multiexon deletions produce Becker muscular dystrophy if the frameshifts due to the individual exons cancel out. DMD, Duchenne muscular dystrophy; BMD, Becker muscular dystrophy.

Premature termination codons usually act as null mutations

A premature termination codon, whether the result of a frameshift as in Figure 16.6 or of a nucleotide substitution in the codon for an amino acid, as in the p.Gln39* mutation in Table 16.3, might be expected to lead to production of a truncated protein. When ribosomes encounter a stop codon they dissociate from the mRNA, and the nascent polypeptide is released. In fact, the predicted truncated protein is seldom produced. Cells have a mechanism, **nonsense-mediated decay** (**NMD**), that detects mRNAs containing premature termination codons and degrades them. Thus the usual result of a nonsense mutation is to prevent any production of protein.

NMD is thought to work because the spliced mRNA that travels from the nucleus to the ribosomes retains a memory of the positions of the introns. The splicing mechanism leaves proteins of the **exon junction complex** (**EJC**) attached next to splice sites. During a first ("pioneer") round of translation, as the ribosome passes each splice site it clears the EJC proteins attached to that site. If there is a premature termination codon, the ribosome will not have traversed every splice site before it detaches. Some EJC proteins will remain attached to the mRNA, and this marks the mRNA for destruction (Figure 16.7A).

In heterozygotes, truncated proteins are potentially more pathogenic than simple nonproduction of the protein from the mutated allele (Figure 16.7B) because they have the potential to interfere with the function of the normal product. Such dominant-negative effects will be discussed later in this chapter. It is assumed that NMD has arisen

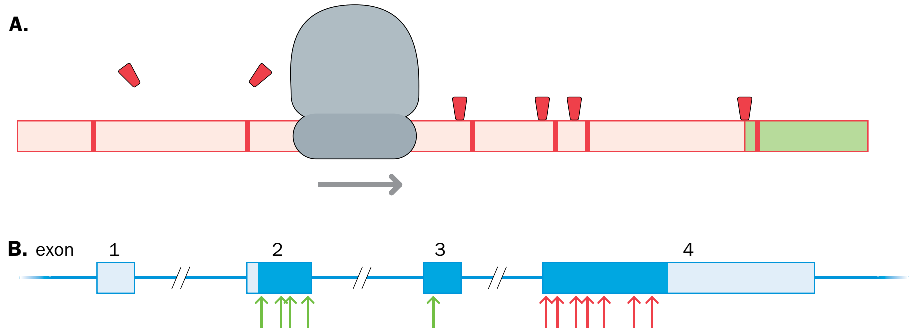


Figure 16.7 The mechanism of nonsense-mediated decay. (**A**) On its first pass along a new mRNA the ribosome displaces proteins of the exon junction complex (EJC, red shapes) that were left near splice sites (red lines) by the splicing machinery in the nucleus. When the ribosome detaches in response to a stop codon, if there are still EJC molecules attached to the mRNA, this marks it for destruction. The EJC marks positions about 50 nucleotides upstream of the actual splice junction. Thus premature termination codons in the region of the gene colored pink will trigger NMD, but any in the part colored green will allow production of a truncated protein. (**B**) Depending on whether or not a premature stop codon triggers NMD, the consequences of a nonsense mutation can be very different. Mutations in the *SOX10* gene that trigger decay (green arrows) result in Waardenburg syndrome type 4 (hearing loss, pigmentary abnormalities, Hirschsprung disease; OMIM #277580). Nonsense mutations in the 3′ region of the mRNA that escape NMD (red arrows) cause a much more severe neurological phenotype. Darker color marks coding sequence, pale areas are the 5′ and 3′ untranslated sequences.

to protect against this problem. However, the NMD mechanism does not apply to all individuals or all genes. In a study of nonsense mutations found in healthy participants in the 1000 Genomes project, MacArthur and colleagues (PMID 22344438; see Further Reading) observed reduced mRNA levels in only 7 of 28 cases where NMD would have been predicted. Probably the rules apply better to small genes than large ones.

Missense changes may or may not affect protein function

A single nucleotide substitution within the coding sequence of a gene may or may not alter the sequence of the encoded protein. The genetic code is degenerate, with 64 codons encoding only 20 different amino acids (plus three stop codons). Thus, some codon changes do not alter the amino acid—they are **silent** or **synonymous**. When a codon change does result in a changed amino acid (a nonsynonymous change), several general factors can underlie the effect:

- As explained in Chapter 1 (see **Figure 1.4**), the 20 amino acids can be classified into nonpolar, uncharged polar, and charged polar types. Replacing an amino acid with one in the same class (a **conservative substitution**) would be expected to have less effect on the protein structure than a nonconservative substitution;
- The unique chemical structure of proline has implications for the way polypeptide chains fold. Prolines disfavor α-helical and β-strand structures, thus missense changes that introduce or remove prolines often affect protein structure;
- Adding or removing cysteine alters the potential for forming disulfide bridges, and so can cause major structural changes to the protein;
- The size of the amino acid side chain may be important. Only glycine, the smallest amino acid, can fit into some protein structures;
- Remember that not all functional missense changes cause a loss of function. Some specific changes can cause a gain of function, as described in Section 16.2.

Similarity matrices have been constructed that give a quantitative score for the likely disruptive effect of any substitution. However, effects often depend on the particular protein. In the aqueous environment of the cell, nonpolar molecules tend to stick together and exclude water, like droplets of oil. Thus globular proteins tend to have uncharged nonpolar amino acids in the interior and charged ones on the outside. Putting a charged residue into the interior may disrupt the three-dimensional folding, while putting a nonpolar residue on the outside may make the protein unnaturally sticky. The sickle cell mutation is pathogenic because it replaces a polar glutamic acid on the outside of the globin molecule with a nonpolar valine. This makes the molecules tend to stick together. Aggregation of hemoglobin S molecules causes the sickle cell phenotype and leads to all the pathological consequences of the disease.

Given the difficulty of making reliable predictions from general principles, multiprotein alignments are widely used to assess the likely effect of a novel missense change. If related proteins (paralogs within a species and orthologs between species) all have a certain amino acid at corresponding positions in their sequence, that amino acid probably has an important role, and replacing it is likely to affect protein function. If, on the other hand, a variety of amino acids can be found at the corresponding position in related proteins, it is unlikely that a change will be deleterious. The two most widely used programs for assessing the effect of missense changes (**Box 16.2**) rely on such multiprotein alignments, supplemented in PolyPhen-2 by information on protein structures.

BOX 16.2 COMPUTER PREDICTIONS OF THE EFFECT OF MISSENSE CHANGES

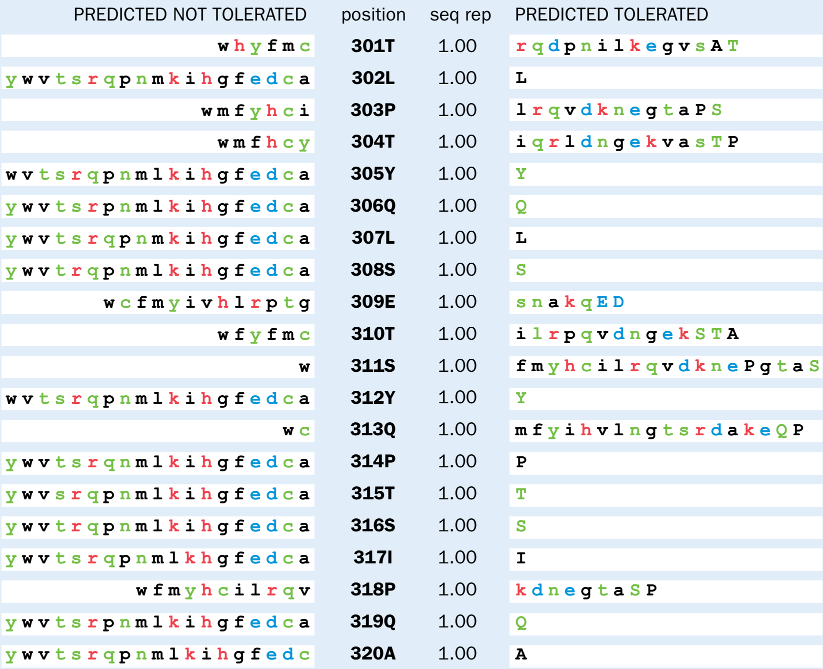
The two most widely used programs are SIFT and PolyPhen-2. Both are Web-based and freely accessible.

SIFT

SIFT (Sorting Intolerant From Tolerant) (http://sift.jcvi.org/) accepts an input protein sequence and uses a PSI-BLAST search to build a multiple sequence alignment. It uses this to predict tolerated and deleterious substitutions for each amino acid in the query sequence. For known proteins, precomputed alignments are available through the SIFT BLink option (http://sift.jcvi.org/www/SIFT_BLink_submit.html), which runs much faster. The output (**Figure 1**) shows each amino acid in the query sequence, with tolerated alternatives on one side and deleterious ones on the other. If you submitted a list of specific missense changes, you also get a specific prediction for each variant, with a score between 0 and 1. A score below 0.05 suggests the change would be deleterious.

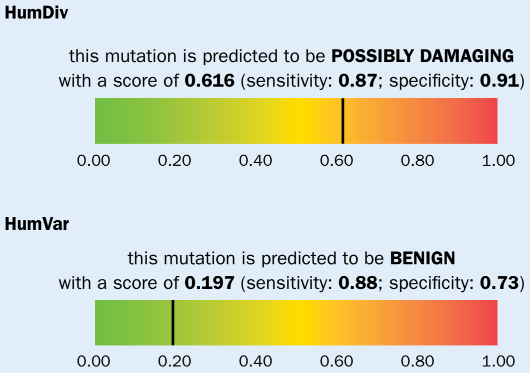
POLYPHEN-2

PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2) takes as input the wild-type and variant amino acids at a stated position in a named protein or a pasted-in polypeptide sequence. The prediction is based on a multiple sequence alignment, as in SIFT, supplemented by any available information about the domains and three-dimensional structure of the named protein. Where possible, PolyPhen-2 notes



PREDICTED NOT TOLERATED	position	seq rep	PREDICTED TOLERATED
whyfmc	301T	1.00	rqdpnilkegvsAT
ywvtsrqpnmkihgfedca	302L	1.00	L
wmfyhci	303P	1.00	lrqvdknegtaPS
wmfhcy	304T	1.00	iqrldngekvasTP
wvtsrqpnmlkihgfedca	305Y	1.00	Y
ywvtsrpnmlkihgfedca	306Q	1.00	Q
ywvtsrqpnmkihgfedca	307L	1.00	L
ywvtrqpnmlkihgfedca	308S	1.00	S
wcfmyivhlrptg	309E	1.00	snakqED
wfyfmc	310T	1.00	ilrpqvdngekSTA
w	311S	1.00	fmyhcilrqvdknePgtaS
wvtsrqpnmlkihgfedca	312Y	1.00	Y
wc	313Q	1.00	mfyihvlngtsrdakeQP
ywvtsrqnmlkihgfedca	314P	1.00	P
ywvsrqpnmlkihgfedca	315T	1.00	T
ywvtrqpnmlkihgfedca	316S	1.00	S
ywvtsrqpnmlkhgfedca	317I	1.00	I
wfmyhcilrqv	318P	1.00	kdnegtaSP
ywvtsrpnmlkihgfedca	319Q	1.00	Q
ywvtsrqpnmlkihgfedc	320A	1.00	A

Box 16.2 Figure 1 Example of SIFT output. Analysis of positions 301–320 of the PAX3 protein. Single-letter amino acid codes are color-coded for class (black nonpolar, green uncharged polar, red basic, blue acidic). Uppercase amino acids were found within the multiple sequence alignment, lowercase are inferred. The "seq rep" figure refers to the depth of the multiple sequence alignment. A figure below 0.25 means it contained few sequences, and so predictions would be unreliable.



HumDiv

this mutation is predicted to be **POSSIBLY DAMAGING**
with a score of **0.616** (sensitivity: **0.87**; specificity: **0.91**)

HumVar

this mutation is predicted to be **BENIGN**
with a score of **0.197** (sensitivity: **0.88**; specificity: **0.73**)

Box 16.2 Figure 2 PolyPhen-2 analysis of the variant p.T315K in the PAX3 protein. The two analyses reflect different datasets used when training the program. HumDiv used known damaging variants causing Mendelian disease, compared to differences between human proteins and their closely related mammalian homologs, assumed to be nondamaging. HumVar used common human nonsynonymous SNPs with no reported disease association as the nondeleterious set. For diagnostic work, checking if a variant is a likely cause of a patient's Mendelian disease, HumVar is the appropriate set. The nonsynonymous variants used in HumVar might actually be mildly deleterious and subject to selection, so for studies of complex disease susceptibility or of selection, HumDiv is better. The strong difference between predictions in this case is because p.T315K occurs as a low frequency population variant, as well as being seen in some patients with Waardenburg syndrome, the normal result of loss of function of PAX3. Note that SIFT labeled this change as not tolerated.

whether the change is in a region carrying a specific annotation in the Swiss-Prot database, such as an active site, transmembrane domain, metal-binding element, and so on, and checks contacts in the PDB structural database. **Figure 2** shows an example of the final output.

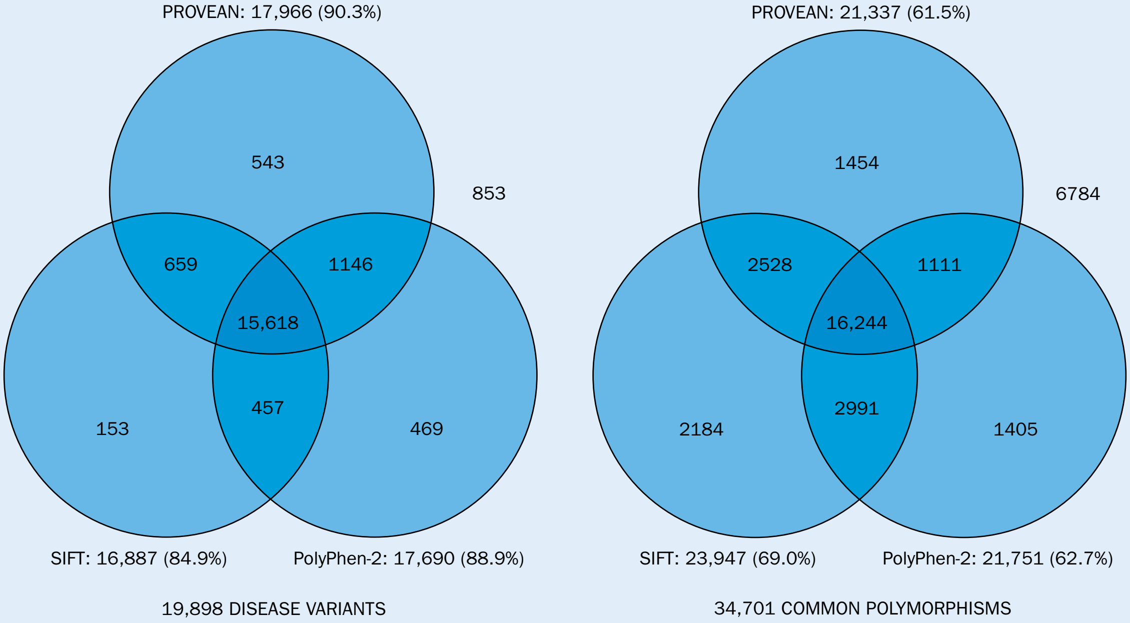
OTHER PROGRAMS

Several other programs are available for these analyses. Examples include:

- Align-GVGD (http://agvgd.iarc.fr/);
- Hansa (http://www.cdfd.org.in/HANSA/);
- MAPP (http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html) (to download program and run it locally);
- MutPred (http://mutpred.mutdb.org/);
- PROVEAN (http://provean.jcvi.org/index.php);
- SNPs&GO (http://snps.uib.es/snps-and-go/).

For a fuller list and discussion, see the National Genetics Reference Laboratory missense prediction tools catalog (http://www.ngrl.org.uk/Manchester/page/missense-prediction-tool-catalogue).

None of these programs is perfect, and different ones perform better or worse with different genes and variants. Typically, they are 70–80% accurate (**Figure 3**). Laboratories often adopt a consensus approach. It is important to remember that none of these protein-level tools can take RNA-level effects such as splicing into consideration (see **Figure 16.4**). Thus a comprehensive analysis of an apparent missense change must include analysis of the nucleotide sequence for splice-site changes.



Box 16.2 Figure 3 Venn diagrams showing predictions from PROVEAN, SIFT, and PolyPhen-2 for the UniProt human protein variant dataset. Score thresholds used: PROVEAN, –1.3; SIFT, 0.05; PolyPhen-2, 0.432. The number and percentage of correctly predicted variants are shown next to each tool. (From Choi Y & Chan AP [2015] *Bioinformatics* **31**:2745–2747; PMID 25851949. With permission from Oxford University Press.)

In this chapter we are concerned only with looking at the ways a particular genotype may lead to a certain phenotype. When a clinical diagnostic laboratory is trying to decide whether a certain sequence variant might explain a patient's condition, they will put this mechanistic information in a wider context, look at some additional information, and make a judgement. We defer consideration of that overall process until Chapter 20.

Sometimes a recessive loss-of-function condition depends on a specific low level of residual function

Some gene products are essential for normal development. Any conceptus that was homozygous for a null allele of such a gene, lacking all function, would never develop to term. Some level of gene function, but maybe less than the full 100%, would be necessary for normal development. There might be a level somewhat below that threshold that was compatible with survival, but not with fully normal development. It might be, to invent an example, that 20% of full function (summed across the two alleles at the locus) was sufficient for normal development, 10% allowed survival but with congenital abnormalities, and any level below 10% was lethal.

If there were a range of rare loss-of-function alleles in the population, some causing a total loss and others a 30% loss, what would one see? Individuals homozygous for the null allele would never be born, while those homozygous for the 30% allele would not be ascertained because they would be phenotypically normal. Heterozygotes with one fully functional allele and either a null or a 30% allele would also be phenotypically normal. But compound heterozygotes for the null and 30% alleles would have overall a 15% level of function, and would present with abnormalities. Thus one would see a rare recessive condition but without the usual parental consanguinity.

This situation was briefly mentioned as a theoretical possibility at the end of Chapter 12. Several concrete cases are known, for example:

- The case of TAR (thrombocytopenia-absent radius) syndrome was mentioned in Section 15.3. Affected people are compound heterozygotes for a deletion (the null allele) and one or other of two low-frequency polymorphisms in the *RBM8A* gene that reduce but do not abolish expression.
- Jenkinson and colleagues (PMID 27571260; see Further Reading) showed a similar situation in the *SNORD118* gene that encodes a small nucleolar RNA. A number of hypomorphic (low function) variants in the gene were described. Patients with leukoencephalopathy, brain calcifications, and cysts (LCC; OMIM #614561) are mostly compound heterozygotes for one severe and one mild variant. Despite the rarity of this recessive condition, very few cases were born to consanguineous parents.

Dominant-negative effects occur in a heterozygous person when the mutated gene product interferes with the function of the normal product

Sometimes in a person heterozygous for a missense protein variant, the abnormal protein not only fails to function correctly, but actively interferes with the function of the normal form. This is a dominant-negative effect. The mechanism of nonsense-mediated decay (see above) probably evolved as a protection against possible dominant-negative effects of abnormal truncated proteins: it may be better to have no product from a mutant gene than to have an abnormal product. Proteins that build multimeric structures are particularly vulnerable to these effects. Collagens provide a classical example.

Fibrillar collagens, the major structural proteins of connective tissue, are built of triple helices of polypeptide chains—sometimes homotrimers, sometimes heterotrimers—that are assembled into close-packed cross-linked arrays to form rigid fibrils. In newly synthesized polypeptide chains (preprocollagen), N- and C-terminal propeptides flank a regular repeating sequence (Gly-X-Y)$_n$, where X and Y are variable amino acids, at least one of which is often proline. Three preprocollagen chains associate and wind into a triple helix under control of the C-terminal propeptide. After formation of the triple helix, the N- and C-terminal propeptides are cleaved off. Mutations that replace glycine with any other amino acid usually have strong dominant-negative effects because they disrupt the tight packing of the triple helix.

Missense mutations in type I collagen are responsible for the most severe forms of brittle bone disease (osteogenesis imperfecta type IIA; OMIM #166210) because of these dominant-negative effects. In heterozygotes the mutant collagen polypeptides associate with normal chains, but then disrupt formation of the triple helix. This can reduce the yield of functional collagen to well below 50%. Null mutations in the same gene might be expected to produce more severe effects, but in fact the disease is milder. The simple absence of some collagen is less disruptive than the presence of abnormal chains (**Figure 16.8**).

The ion channels in cell membranes provide another case of multimeric structures that are susceptible to dominant-negative effects. Connexin 26 provides an example. Six molecules of the connexin 26 protein associate to form a connexon, one half of a gap junction (see **Figure 3.11**) that allows small ions to move between cells. **Figure 16.6** showed an example of a null mutation in the gene that encodes connexin 26. People homozygous for this mutation cannot make connexin 26; they lack functioning gap junctions in their inner ears, potassium ions cannot recirculate as they should, and the patients are deaf. Heterozygotes are entirely phenotypically normal (which is a problem if you wish to identify couples at risk of having deaf children). However, certain missense mutations produce structurally abnormal connexin 26 molecules. These disrupt the function of connexons, even if some of the six connexin molecules are normal; heterozygotes for those variants have hearing loss and the phenotype is dominant.

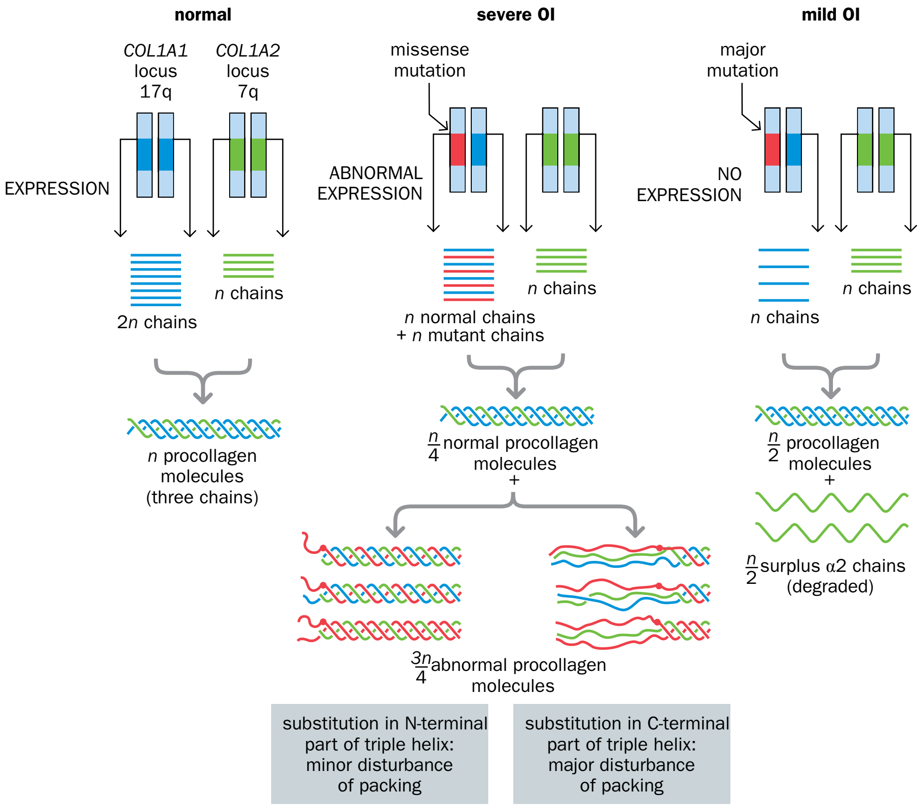

normal
COL1A1 locus 17q
COL1A2 locus 7q
EXPRESSION
2*n* chains
n chains
n procollagen molecules (three chains)
severe OI
missense mutation
ABNORMAL EXPRESSION
n chains
n normal chains + *n* mutant chains
$\frac{n}{4}$ normal procollagen molecules
+
$\frac{3n}{4}$ abnormal procollagen molecules
substitution in N-terminal part of triple helix: minor disturbance of packing
substitution in C-terminal part of triple helix: major disturbance of packing
mild OI
major mutation
NO EXPRESSION
n chains
n chains
$\frac{n}{2}$ procollagen molecules
+
$\frac{n}{2}$ surplus α2 chains (degraded)

Figure 16.8 Dominant-negative effects of collagen gene mutations. Collagen fibrils are built of cross-linked arrays of triple-helical procollagen units. The type I procollagen comprises two chains encoded by the *COL1A1* gene and one encoded by *COL1A2*. In the triple helix, each polypeptide chain consists of repeating units, $(\text{Gly-X-Y})_n$. In osteogenesis imperfecta type IIA (OI; OMIM #166210), mutations that replace glycine with any other amino acid usually have strong dominant-negative effects because they disrupt the packing. The helix is assembled starting at the C-terminus, and substitutions of glycines close to that end have a more severe effect than substitutions nearer the N-terminus. Null mutations in either gene result in fewer but otherwise normal triple helices forming, and produce a less severe clinical phenotype.

16.2 GAIN OF FUNCTION

Unlike with loss of function, only very specific changes are likely to produce a gain of function. There are innumerable ways of destroying the function of a gene product, but only a few ways of making it gain function. Most obviously, a gain of function requires that the altered product should actually exist. Many of the mechanisms producing loss of function result in the total absence of product, and these cannot in principle produce a gain of function. Gain-of-function changes are likely to be either missense changes in a protein or changes in regulatory sequences.

Very occasionally, a missense change causes an enzyme to catalyze a novel reaction

It is unusual for a gain of function to involve acquisition of a completely novel function. More usually the gain involves the protein doing its normal thing, but in an inappropriate way—for example, a receptor signaling even in the absence of its ligand. However, a few examples are known where a missense change causes an enzyme to change its substrate specificity or reaction mechanism.

Alpha-1 antitrypsin (OMIM #107400) protects the body against elastase released into the bloodstream. Methionine 358 of the protein acts as a "bait" to catch elastin molecules for inactivation. In the Pittsburgh variant, methionine is replaced by arginine (p.Met358Arg). This now acts as a bait for thrombin instead of elastin. The variant enzyme is now an antithrombin; the resultant loss of thrombin causes a severe bleeding disorder.

Another example concerns the two isocitrate dehydrogenase enzymes IDH1 and IDH2. These frequently carry specific mutations in cancer. More than 70% of grade II and III astrocytomas and oligodendrogliomas, and the glioblastomas that develop from these lower-grade lesions, have a missense change to amino acid 132 of IDH1 (p.R132H or p.R132S). Tumors without mutations in IDH1 often have mutations affecting the corresponding amino acid (R172) of the IDH2 protein. These changes in the active site change the activity of the enzyme. Normally it converts isocitrate into α-ketoglutarate;

the mutant forms instead reduce α-ketoglutarate to 2-hydroxyglutarate. This abnormal metabolite has a number of effects on cell metabolism; in particular it inhibits histone demethylation, leading to changes in gene expression (see OMIM #147700 and **Figure 19.24**).

A gene product can gain function through a variety of mechanisms

Gain-of-function changes are especially a feature of cancer, where a variety of mechanisms cause overactivity of growth-promoting genes (oncogenes). These are covered in detail in Chapter 19, but briefly, mechanisms include:

- Making extra copies of an active gene so as to produce a quantitative increase in the amount of product (gene amplification);
- Chromosomal rearrangements that place an oncogene under the influence of a powerful enhancer, so as to increase the level of expression;
- Chromosomal rearrangements that create novel, highly active chimeric genes by combining exons of two separate genes;
- Missense changes that alter the properties of a protein.

Similar mechanisms are seen in noncancer conditions, although large chromosomal rearrangements are not normally seen in regularly inherited conditions because they are not stably transmitted through meiosis (see Section 15.2). Thus large rearrangements are causes of cancer and other mosaic conditions that depend only on mitosis, but not of conditions inherited through the generations. Quantitative changes in gene expression (through extra copies of dosage-sensitive genes) must cause the pathology of chromosomal trisomies. Microduplication syndromes show the effect of increased dosage of just one or a very few genes (Section 15.3). Examples include Charcot–Marie–Tooth disease type 1A (CMT1A; OMIM #118220), caused by duplication of the peripheral myelin protein 22 (*PMP22*) gene on chromosome 17p12, and Potocki–Lupski syndrome (OMIM #610883) caused by a microduplication of a nearby sequence at 17p11.2.

A mechanism of enhancer capture underlies some congenital malformations, and has probably also been important in evolution. **Figure 16.9** shows how small chromosomal deletions or inversions that alter the boundaries of TADs can bring a gene under the influence of a novel enhancer and cause a gain of function. An important paper by Lupiáñez and colleagues in 2015 (PMID 25959774; see Further Reading) showed examples of these effects in practice. They showed how distinct human limb malformations are caused by small deletions, inversions, or duplications that affect the relationships of enhancers and insulators to four adjacent genes at 2q35, *WNT6*, *IHH*, *EPHA4*, and *PAX3*.

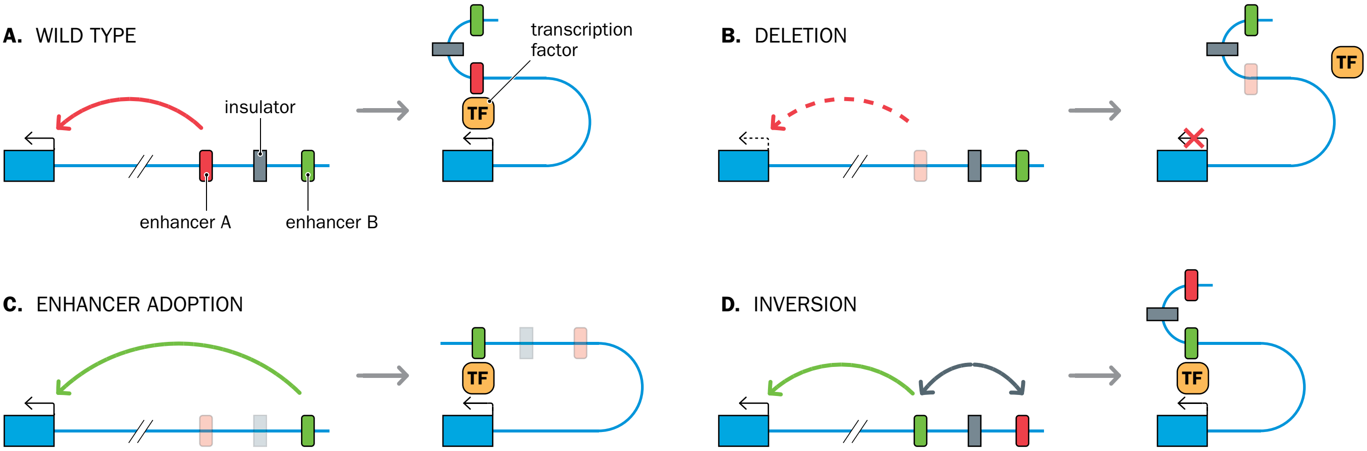


Figure 16.9 Chromosome deletions or inversions can change the relationship between genes and enhancers. (**A**) In the wild-type arrangement, expression of the gene is driven by the red enhancer bringing transcription factors to the gene promoter. An insulator prevents access by the green enhancer. (**B**) Deletion of the red enhancer prevents expression of the gene, a loss of function. (**C**) Deletion of both the red enhancer and the insulator brings the gene under control of the green enhancer, probably causing a gain of function. (**D**) An inversion has the same effect. (Adapted from Spielmann M & Klopocki E [2013] *Curr Opin Genet Dev* **23**:249–256; PMID 23601627. With permission from Elsevier.)

Qualitative changes in proteins, due to missense mutations, underlie many gain-of-function phenotypes. A frequent mechanism is improper activation of signaling pathways. Through a gain of function, a cell surface receptor that would normally only send a signal to the cell interior in response to its ligand may become constitutionally active. In Section 3.1 we described how binding of the appropriate ligand often causes receptors to dimerize (see **Figures 3.3** and **3.4**). The dimerization then triggers a cascade of changes in the cell, leading ultimately to changes in gene expression. Missense changes in the receptor protein may make it liable to dimerize even in the absence of ligand. For example, a missense change, p.C342Y, in the FGFR2 (fibroblast growth factor receptor 2) protein removes a cysteine residue that is normally involved in an intramolecular disulfide bridge. Its normal partner cysteine is now free to form an *inter*molecular disulfide bridge, leading to receptor dimerization and constitutive signaling. The result is Crouzon syndrome (OMIM #123500). Different missense changes in the same protein, p.S252W or p.P253R, cause the related Apert syndrome (OMIM #101200). These changes again cause a gain of function, in this case changing or enhancing the binding affinity of the receptor for members of the nine-strong fibroblast growth factor family. The specificity of gain-of-function missense changes often leads to tight correlations between genotype and phenotype, as discussed in Section 16.5.

The concept of gain of function can be applied to multicomponent cellular pathways, as well as to single gene products. The Ras–MAP kinase pathway provides an example. This multistep intracellular signaling pathway was introduced in Chapter 3 (see **Figure 3.8**). **Figure 16.10** shows more detail. The pathway transmits growth-promoting signals from various cell surface receptors to transcription factors in the cell nucleus. The ultimate targets of the pathway are the ERK1/2 mitogen-activated protein kinases (MAPKs) that turn on transcription of growth-promoting genes. Abnormal activation of the pathway can be caused by mild gain-of-function mutations in any one of a number of the genes involved (*PTPN11*, *SOS1*, *SHOC2*, *HRAS*, *KRAS*, *NRAS*, *RAF1*, *BRAF*, *MEK1*, *MEK2*). On the other hand, the *NF1* and *SPRED1* genes encode inhibitors of Ras–MAPK signaling, and loss-of-function mutations in these genes equally activate the pathway. Loss of function of an inhibitory protein causes gain of function of the pathway. Mutations in the individual genes cause a set of overlapping syndromes (neurofibromatosis 1, OMIM #162200; Noonan syndrome, OMIM #163950; Costello syndrome, OMIM #218040; cardiofacio-cutaneous syndrome, OMIM #115150; and Legius syndrome, OMIM #611431). Mutations in several different genes can cause the same clinical condition, while different mutations in some genes can cause different syndromes (reviewed by Rauen [2013], PMID 23875798; see Further Reading). The genotype–phenotype correlations were deeply confusing until they were all related to activation of the same signaling pathway.

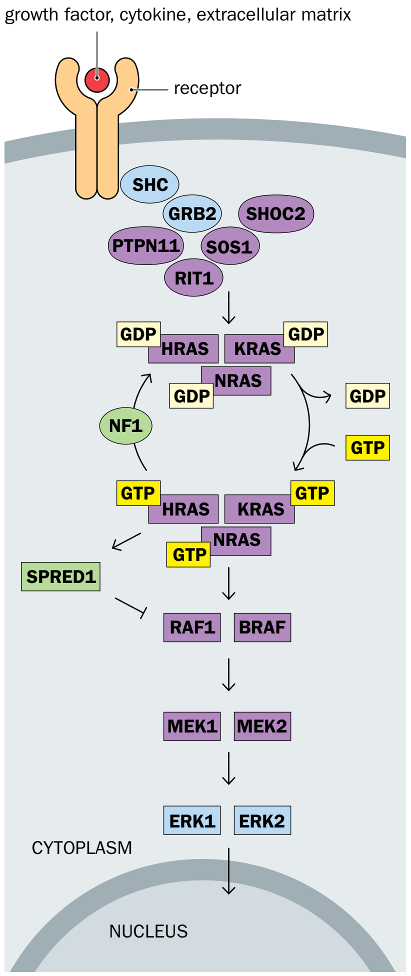


Figure 16.10 The Ras–MAPK intracellular signaling pathway. The cascade of gene products transmits growth-promoting signals from cell surface receptors to the ERK1/2 kinases that stimulate transcription of growth-promoting genes in the nucleus. NF1 and SPRED1 encode inhibitory proteins; other components of the pathway stimulate signaling. Gain-of-function mutations in effector components (purple), or loss-of-function mutations in inhibitory components (green) of the pathway, all result in overactivation of the pathway and produce a set of overlapping clinical syndromes (see text and Rauen [2013], PMID 23875798, for details).

Some mutations produce an RNA with novel toxic properties

One type of gain of function is seen when the product of a mutant allele is toxic to the host cell. The toxicity may be the result of an abnormal protein, as described below, but sometimes it may be a property of an abnormal messenger RNA that causes the effect. Over the past few decades we have increasingly appreciated the complexity of the ways that RNA molecules function in the life of cells. At every step of gene expression, from the primary transcript emerging from the RNA polymerase in a nuclear transcription factory through to a mature cytoplasmic messenger RNA engaged with actively synthesizing ribosomes, gene transcripts do not exist or function as naked RNA strands. They exist in a series of functional RNA–protein complexes, and some RNA–RNA complexes. Sometimes a mutant mRNA can malfunction in one of the complexes in a way that is toxic to the host cell.

RNA toxicity is particularly a feature of dynamic mutations. As described in the following section, in these mutations a short microsatellite-like repeat unit somewhere in a gene sequence is unstable and has a tendency to expand, creating a much increased number of tandem repeats. The repeats may or may not be translated, depending on their position within the gene, but when they are transcribed the result is an RNA molecule containing an extended tandem repeat. The repeat unit may be three, four, five, or six nucleotides and can have varying sequences. In some cases these RNAs are toxic to the host cell (**Table 16.5**).

TABLE 16.5 TOXIC RNA EFFECTS IN DISEASES CAUSED BY DYNAMIC MUTATIONS

Condition	OMIM #	Gene	RNA repeat	Location in gene	Affected proteins
Myotonic dystrophy 1	160900	*DMPK*	$(CUG)_n$	3′ UTR	MBNL1, insulin receptor, cardiac troponin T
Myotonic dystrophy 2	602668	*ZNF9*	$(CCUG)_n$	Intron 1	MBNL1
Huntington disease-like 2	606438	*JPH3*	$(CUG)_n$	3′ UTR	MBNL1
Spinocerebellar ataxia 8	608768	*ATXN8OS*	$(CUG)_n$	[Noncoding]	MBNL1
Spinocerebellar ataxia 10	603516	*ATXN10*	$(AUUCU)_n$	Intron 9	hnRNP K
FTD/ALS	105550	*C9ORF72*	$(GGGGCC)_n$	Intron 1	hnRNPA3, ASF/S2, ADARB2

In each case the repeat is transcribed but not translated. The list of affected proteins is very incomplete. See Section 16.3 for more discussion of dynamic mutations. The *C9ORF72* expansion is the most frequent single cause of frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS), but many cases have other causes. UTR, untranslated region; MBNL1, Muscleblind-like protein.

It is thought these RNAs are pathogenic because they sequester necessary RNA-binding proteins into RNA foci in cells. The best-studied example is myotonic dystrophy (DM1, OMIM #160900). This autosomal dominant, multisystem disease is caused by a mutant version of the *DMPK* protein kinase gene. The mutation, an expanded run of CTG triplets, is in noncoding sequence, the 3′ untranslated region of the mRNA (see **Figure 16.13B**, below). The protein product appears to be unaffected, either qualitatively or quantitatively, but the $(CUG)_n$ in the mutant mRNA forms stable hairpins. These accumulate in RNA foci and sequester CUG-binding proteins, among them the Muscleblind-like (MBNL1) protein, the insulin receptor, and cardiac troponin T. Deficiency of these proteins explains the many features of myotonic dystrophy. MBNL1 is required for correct splicing of other muscle gene transcripts such as the CLCN1 muscle chloride channel. Thus the pathology of the disease is indirect, and unrelated to the function of the mutated gene.

Some mutations produce a mutant protein with novel toxic properties

Cells have elaborate mechanisms to ensure that newly synthesized proteins are correctly folded. Hydrophobic residues need to be confined in the interior of a globular protein, or covered up by interfacing with another protein subunit if on the outside. A variety of chaperone molecules recognize misfolded proteins that have hydrophobic patches on the exterior, and provide a contained space where the proteins can try to fold correctly. Misfolded proteins are often intrinsically unstable, but if not they are targeted for degradation in the proteasome.

Some misfolded proteins have a propensity to form aggregates. Aggregation may be pathogenic simply by sequestering a necessary protein away from its normal site of action, or by trapping other important proteins within insoluble and protease-resistant aggregates. Protein aggregation turns out to be a common feature of a range of neurodegenerative diseases. The diseases may be inherited, like Huntington disease, sporadic, like the great majority of cases of Alzheimer and Parkinson diseases, or even transmissible between individuals of the same or different species, as in some cases of Creutzfeldt–Jakob disease and bovine spongiform encephalopathy ("mad cow disease"). Some mutant proteins may have an inherent tendency to form aggregates because of hydrophobic patches on the exterior; in other cases the pathogenic process may start with conversion of a few molecules of a protein into an abnormally folded state. In inherited conditions, the effect of the causative mutation is to make the protein more likely to adopt the abnormally folded state. In noninherited conditions, the first abnormally folded molecule arises as a rare chance misfolding of the normal molecule. In infectious prion diseases, the initial abnormally folded protein molecules arrive from outside.

A single molecule with the abnormal conformation is probably not pathogenic, but once small aggregates arise they act as seeds (**Figure 16.11**). The seed converts more molecules of the normally folded protein into the abnormal form by a process akin to crystallization, but producing one-dimensional amyloid fibrils characterized by stacked β-sheets rather than three-dimensional crystals (**Figure 16.12**). The fibrils are resistant to proteolytic degradation and accumulate inside cells. They may overwhelm the normal protein quality-control mechanisms of a cell and sequester other important molecules within the aggregate. Mature fibrils are probably not themselves pathogenic, but they can fragment, producing smaller seed aggregates and hence propagating themselves.

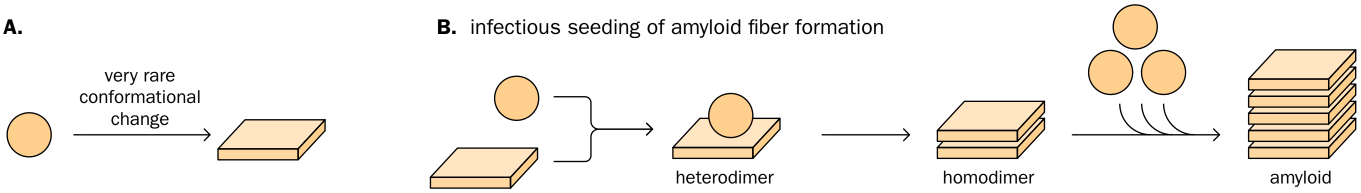

A.
very rare conformational change
B. infectious seeding of amyloid fiber formation
heterodimer
homodimer
amyloid

Figure 16.11 Amyloid fibrils grow by a process akin to crystallization. (**A**) Very occasionally a protein molecule undergoes a change that produces a novel, amyloidogenic conformation. (**B**) A molecule of the abnormally folded amyloidogenic form of the protein can induce other molecules of the same protein to adopt the abnormal conformation and aggregate together, producing an amyloid fibril.

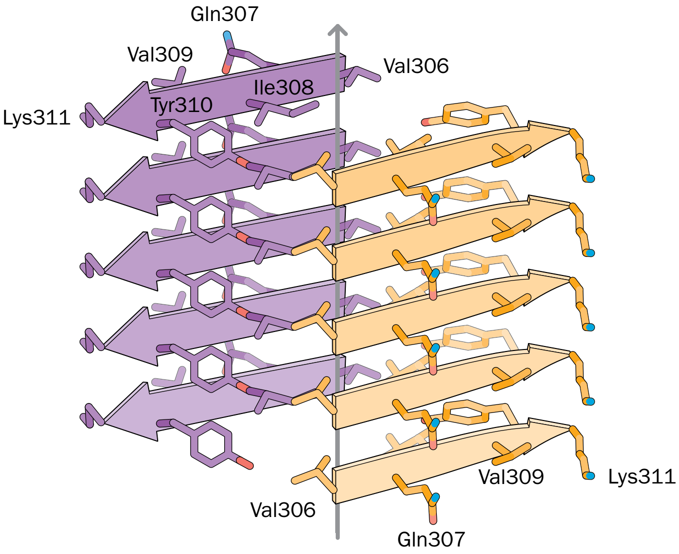


Figure 16.12 Structure of an amyloid fiber. Amino acids 306–311, VQIVYK, of the Tau protein form β-sheets, with the backbones parallel within each sheet and antiparallel between stacks. The vertical arrow shows the axis of the fiber, built up of thousands of stacked molecules. (From Goedert M [2015] *Science* **349**:1255555; PMID 26250687. Reprinted with permission from the AAAS.)

The term "prion" is often applied only to such abnormal proteins when they arise through external infection, but the same intracellular mechanism operates in non-infectious prion diseases. Prion diseases are typically neurodegenerative, in part at least because the aggregate seeds are able to move across neural connections and so spread the infection through the brain. The long fibrils accumulate, producing the characteristic pathology of the disease—amyloid plaques and neurofibrillary tangles in Alzheimer's disease, Lewy bodies in Parkinson's disease, and so on.

Complex pathogenic mechanisms of dynamic mutations

Protein-level gain-of-function effects of the dynamic mutations described in Section 16.3 depend on a mixture of orthodox and unorthodox mechanisms. The orthodox mechanisms are seen in the many conditions where there is an expanded run of $(CAG)_n$ triplets in coding sequence (see **Table 16.7**). CAG is the codon for glutamine (Q), and so the encoded protein has a correspondingly expanded run of glutamine residues. These cause the protein to form abnormal aggregates.

In addition to these orthodox effects, there is also evidence of a different, unorthodox set of gain-of-function effects. While the usual translation product may be present, and may (for example, Huntington disease) or may not (for example, myotonic dystrophy 1) be pathogenic, it appears that the repeat RNA structure allows abnormal translation products to be formed that do not depend on an AUG start codon. Repeat-associated non-ATG translation (RAN translation; see the review by Cleary & Ranum [2013], PMID 23918658, in Further Reading) potentially produces homopolymeric proteins in all three reading frames. Moreover, there is evidence that at least some of the expanded repeats are also transcribed in an antisense direction. RAN translation from both sense and antisense RNA in all three reading frames has the potential to produce six different proteins (**Table 16.6**). Exactly which RAN proteins are actually produced and where, and what role they play in the pathology of these conditions, are still active areas of research, but there is evidence that at least some of them are indeed produced and are toxic to a host cell.

TABLE 16.6 REPEAT-ASSOCIATED NON-ATG (RAN) TRANSLATION

Repeat	Reading frame	RAN protein
$(CAG)_n$	CAG.CAG.CAG…	Poly(Gln)
	AGC.AGC.AGC…	Poly(Ser)
	GCA.GCA.GCA…	Poly(Ala)
$(CUG)_n$	CUG.CUG.CUG…	Poly(Leu)
	UGC.UGC.UGC…	Poly(Cys)
	GCU.GCU.GCU…	Poly(Ala)
$(GGGGCC)_n$	GGG.GCC.GGG.GCC…	Poly(Gly.Ala)
	GGG.CCG.GGG.CCG…	Poly(Gly.Pro)
	GGC.CGG.GGC.CGG…	Poly(Gly.Arg)
$(CCCCGG)_n$	CCC.CGG.CCC.CGG…	Poly(Pro.Arg)
	CCC.GGC.CCC.GGC…	Poly(Pro.Gly)
	CCG.GCC.CCG.GCC…	Poly(Pro.Ala)

Some expanded DNA repeats can be transcribed in both directions, and the resulting RNA translated in all three reading frames without needing the usual AUG start codon. The result is a homopolymeric or simple repeat protein. Data are shown for the (CAG)/(CTG) repeats produced by many dynamic mutations, and the (GGGGCC) repeats and their antisense counterparts in the *C9ORF72* gene. How many of these possible proteins are actually produced is usually uncertain.

16.3 DYNAMIC MUTATIONS: UNSTABLE REPEAT EXPANSIONS

These variants are treated separately here because they do not fit easily into sections on loss of function or gain of function. At the DNA level they may all share the same mechanism of expansion, but the pathogenic consequences of the expansions differ between different conditions.

Microsatellite repeats—tandem repeats of 2–6 nucleotide units—are prone to losing or gaining repeats due to slippage of the polymerase when the DNA is replicated (see **Figure 11.1**). The repeat number often varies between individuals, making them useful polymorphic markers for following a chromosomal segment through a pedigree (see Section 17.1). When an individual microsatellite is PCR-amplified and the product is run out on a gel, "stutter" bands are often seen because of replication slippage during the amplification. A typical mutation rate for a microsatellite would be 10^{-4}–10^{-3} per locus per generation—much higher than the overall rate of base substitution, which is of the order of 2×10^{-8} per nucleotide per generation. However, certain microsatellites suddenly become much more unstable once some threshold repeat number is exceeded, and when this occurs within a gene, the result is a dynamic mutation.

Dynamic mutations occur in many different genes and diseases

Over 20 different diseases are caused by dynamic mutations in different genes (**Table 16.7**). The CAG expansions usually involve modest expansions (dozens to a hundred or so repeat units) in coding sequence, resulting in proteins with expanded runs of glutamines that have a tendency to aggregate, as described in the previous section. Other expansions are usually in noncoding sequence, and can involve very large expansions (thousands of repeat units).

TABLE 16.7 EXAMPLES OF DISEASES CAUSED BY UNSTABLE EXPANDED REPEATS (DYNAMIC MUTATIONS)

Sequence	Condition	OMIM #	Gene	Repeat location	Normal repeat number	Pathogenic repeat number
$(CAG)_n$	Huntington disease	143100	*HTT*	Exon	9–35	36–121
	Dentatorubral-pallidoluysian atrophy	125370	*ATN1*	Exon	7–35	49–93
	Spinal and bulbar muscular atrophy	313200	*AR*	Exon	10–36	38–62
	Spinocerebellar ataxia 1	164400	*ATXN*	Exon	6–39	40–83
	Spinocerebellar ataxia 2	183090	*ATXN2*	Exon	14–31	32–200
	Spinocerebellar ataxia 3	109150	*ATXN3*	Exon	12–44	52–86
	Spinocerebellar ataxia 6	183086	*CACNA1A*	Exon	4–18	19–33
	Spinocerebellar ataxia 7	164500	*ATXN7*	Exon	4–35	37–306
	Spinocerebellar ataxia 17	607136	*TBP*	Exon	25–42	45–66
$(CTG)_n$	Myotonic dystrophy 1	160900	*DMPK*	3′ UTR	5–38	50–>1000
	Spinocerebellar ataxia 8	608768	*ATXN8OS*	[Noncoding]	15–50	71–1300
	Huntington disease-like 2	606438	*JPH3*	3′ UTR	6–28	41–59
$(CGG)_n$	Fragile X A	300624	*FMR1*	5′ UTR	<55	>200
$(GCC)_n$	Fragile X E	309548	*FMR2*	5′ UTR	6–25	>200
$(GAA)_n$	Friedreich ataxia	229300	*FXN*	Intron	5–33	70–>1000
$(CCTG)_n$	Myotonic dystrophy 2	602668	*ZNF9*	Intron	<30	75–11,000
$(ATTCT)_n$	Spinocerebellar ataxia 10	603516	*ATXN10*	Intron	9–32	800–4500
$(GGGGCC)_n$	Frontotemporal dementia/amyotrophic lateral sclerosis	105550	*C9ORF72*	Intron	2–22	700–1600
$(C_4GC_4GCG)_n$	Progressive myoclonus epilepsy	254800	*CSTB*	Promoter	2–3	30–78

Fragile X A, fragile X E, Friedreich ataxia, and progressive myoclonus epilepsy are caused by loss of function of the relevant gene; the expanded repeat represses transcription, and occasional patients have deletions or other conventional loss-of-function mutations. The other conditions are specifically caused by the repeat expansions and represent pathogenic gains of function. As described in Section 16.2, these can involve effects at the RNA and/or protein levels, depending on the condition. See OMIM for further details and references.

Dynamic mutations show a range of molecular pathologies

At the DNA level the mechanisms may be similar, but at the cellular level the molecular pathology of these conditions varies widely. Some prevent transcription of the host gene, causing a loss of function. Others show a gain of function due to producing altered mRNAs and/or proteins that have toxic effects (**Figure 16.13**).

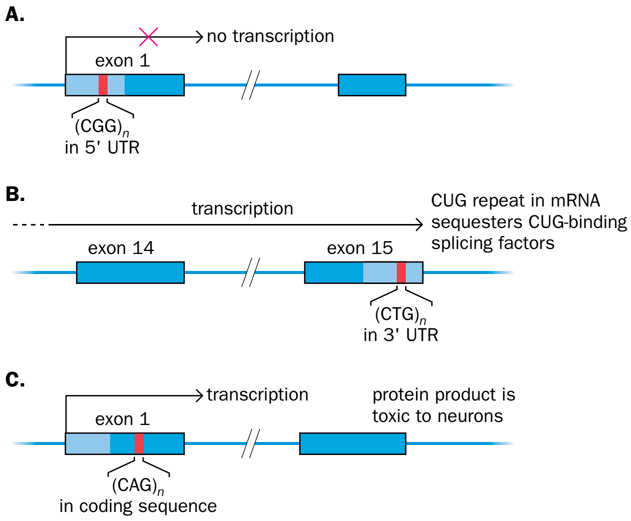


Figure 16.13 Three mechanisms by which dynamic mutations may be pathogenic. (**A**) In fragile X syndrome, the expanded repeat in the 5′ untranslated region (UTR) of the gene triggers methylation of the promoter and prevents transcription. (**B**) In myotonic dystrophy, the expanded repeat in the 3′ untranslated region causes the mRNA transcript to sequester splicing factors in the cell nucleus, preventing the correct splicing of several unrelated genes. (**C**) In Huntington disease, the gene containing the expanded repeat is transcribed and translated as normal, but the protein product has an expanded polyglutamine tract that renders it toxic. Expanded repeats are shown in red, and coding regions in dark blue.

Fragile X syndrome has some special features. Unaffected people have fewer than 55 repeats in the 5′ untranslated region of the *FMR1* gene. These sequences are stable and nonpathogenic. Sequences with 55–200 repeats are described as **premutations**. They do not cause the classical fragile X syndrome, but they are unstable and have a tendency to progress to full mutations (>200 repeats) in subsequent generations. The full repeat changes the chromatin structure such that the promoter is methylated and the *FMR1* gene is not expressed. The clinical features of fragile X syndrome are due to lack of FMR1 protein, an important RNA-binding protein. Occasional affected individuals have conventional loss-of-function mutations in the *FMR1* gene rather than expanded repeats. However, females with premutation alleles are at risk of premature ovarian failure, and premutation males often develop a condition, FXTAS, of tremor and ataxia. It is believed that these premutation phenotypes are caused by a toxic RNA gain of function.

A characteristic of repeat expansion disorders is **anticipation**—that is, the symptoms tend to get more severe or occur at earlier ages down the generations. For example, a person with myotonic dystrophy 1 but no previous family history may show no features except cataracts. Their daughter who inherited the mutation might show the classical disease, with muscle weakness and other signs, while her child might show the very severe congenital form. In Huntington disease the age of onset may become younger down the generations, while in fragile X syndrome the risk of a premutation carrier having full-mutation offspring increases down the generations. All these features are a consequence of the tendency of the repeats to expand in each successive generation. However, it is important to treat claims of anticipation with great caution. Suppose an autosomal dominant condition is very variable, as many are. Mildly affected parents who have a severely affected child will bring it to the clinic. On the other hand, severely affected people may never become parents, or if they do and have a mildly affected child, they would probably not see any reason to bring it to clinical attention. Thus the clinician's experience is usually of mildly affected parents having severely affected children, and not the reverse. There is a systematic **bias of ascertainment** that mimics true anticipation.

The mechanism of expansion remains uncertain

Whatever causes these large expansions, it appears to be independent of the replication slippage that causes minor microsatellite instability because expansions are seen in nonreplicating cells such as neurons and oocytes. The potential for expansion may depend on the DNA adopting an abnormal structure (B- or Z-DNA, hairpins, or G-quadruplexes, and so on—see **Figures 1.10** and **1.13**) that interferes with synthesis or repair. In many diseases the repeats are sometimes interrupted. Such interruptions reduce the stability of the predicted abnormal secondary structures, and also reduce the risk of repeat expansion. For example, 5% of families with myotonic dystrophy 1 have the CTG repeats interrupted by CAG units. In these families the disease is less severe and more stable through the generations. In mouse models of several different conditions due to expanded repeats, manipulation of genes involved in DNA repair affects the tendency of repeats to expand. Abolition of the *Msh2* mismatch repair activity (see **Figure 11.5**) prevents the expansion. The general tentative conclusion is that the expansions occur when DNA sequences that can form abnormal secondary structures require repair.

Some genes show pathogenic expansions of polyalanine tracts

Short polyalanine tracts are present in over 400 human proteins, particularly proteins localized to the cell nucleus. Many are transcription factors. In nine of the proteins an expansion of the polyalanine tract is associated with disease (**Table 16.8**). Unlike the polyglutamine expansions described above, polyalanine expansions are not dynamic; they are stably inherited and show no meiotic or mitotic instability. In some populations a polyalanine expansion can be tracked back to ancestors hundreds of years ago. They are also very modest in size. The expansions probably mostly arise through misaligned crossing over, although some cases may result from replication slippage.

Like the polyglutamine proteins, proteins with polyalanine expansions are misfolded. They form aggregates that in some heterozygous cases sequester the wild-type protein; thus the pathogenic mechanism is loss of function, maybe sometimes with a dominant-negative element. In most cases, other loss-of-function changes in the same gene produce the same or similar phenotypes. All the proteins are transcription factors with developmental roles, except for the PABPN1 protein. This stands a little apart. The *PABPN1* gene encodes a poly(A) tract binding protein involved in nucleocytoplasmic transport of messenger RNAs. Oculopharyngeal muscular dystrophy (OPMD; OMIM #164300) is a late-onset, slowly progressive condition. No other *PABPN1* mutations have been reported in OPMD patients, and the disease-associated expansions are very modest (from 6–7 GCG codons in wild-type alleles to 8–13 in disease-associated alleles).

TABLE 16.8 CONDITIONS ASSOCIATED WITH POLYALANINE EXPANSIONS

Condition	OMIM #	Gene	Expansion
Synpolydactyly type 1	186000	*HOXD13*	$A_{15} \rightarrow A_{22}$–$A_{29}$
Hand-foot-genital syndrome	140000	*HOXA13*	(1) $A_{14} \rightarrow A_{22}$–$A_{24}$* (2) $A_{12} \rightarrow A_{18}$* (3) $A_{18} \rightarrow A_{24}$–$A_{30}$*
Holoprosencephaly	609637	*ZIC2*	$A_{15} \rightarrow A_{25}$
Blepharophimosis, ptosis, and epicanthus inversus (BPES)	110100	*FOXL2*	$A_{14} \rightarrow A_{19}$–$A_{24}$
Cleidocranial dysplasia	119600	*RUNX2*	$A_{17} \rightarrow A_{27}$
X-linked mental retardation with hypopituitarism	300123	*SOX3*	$A_{15} \rightarrow A_{22}$–$A_{26}$
X-linked mental retardation	300419	*ARX*	(1) $A_{16} \rightarrow A_{18}$–$A_{23}$* (2) $A_{12} \rightarrow A_{20}$*
Congenital central hypoventilation syndrome	209880	*PHOX2B*	$A_{20} \rightarrow A_{25}$–$A_{33}$
Oculopharyngeal muscular dystrophy (OPMD)	164300	*PABPN1*	$A_{6} \rightarrow A_{8}$–A_{13}

* The *HOXA13* gene has three polyalanine tracts, and the *ARX* gene has two, each of which is sometimes expanded.

16.4 MOLECULAR PATHOLOGY OF MITOCHONDRIAL DISORDERS

In the remote evolutionary past, mitochondria may have been complete microorganisms, living symbiotically within a larger cell (see **Figure 2.6**). Over time, however, the mitochondrion transferred more and more of its functions to the host's genome, until today the small mitochondrial genome contains only 37 genes (see Section 9.1). The great majority of all mitochondrial functions depend on nuclear-encoded gene products, and the great majority of genetic conditions where there is mitochondrial dysfunction are caused by mutations in nuclear genes. Nevertheless, the 37 mitochondrial genes are essential to life and health, and defects in them cause or contribute to something like 200 different genetic conditions (reviewed by Tuppen *et al.* [2010], PMID 19761752; see Further Reading). Because of the unique features of mitochondria, those conditions show some unique properties.

Mitochondrial gene defects show matrilineal inheritance and possible heteroplasmy

All a person's mitochondria are inherited from the mother and none from the father. Mitochondria of the sperm do enter the egg at fertilization, but they are systematically broken down. Thus conditions caused by variants in mtDNA show the matrilineal inheritance pattern illustrated in Chapter 5 (see **Figure 5.8** and **Box 5.1**). There is no recombination between mitochondrial DNA molecules, thus the mitochondrial genome is transmitted intact down the female line over the generations, changing only because of new mutations. As described in Chapter 14, this makes nonpathogenic mtDNA variants uniquely useful for tracing remote ancestry, at least of the female line.

Cells contain many mitochondria. The liver has been reported to have between 500 and 2500 mitochondria per cell, with 8–10 copies of the genome per mitochondrion. Note, however, that mitochondria are not fixed individual organelles, like microorganisms. They undergo cycles of fusion and fission according to the state of the host cell, and so it is better to think in terms of mitochondrial genomes rather than counts of organelles. The egg cell in particular contains over 100,000 mitochondrial genomes. A person with a mitochondrial mutation may have entirely mutant genomes (**homoplasmy**), but on the other hand they might have a mix of normal and mutant genomes (**heteroplasmy**). Just as most healthy individuals are low-level mosaics for a variety of potentially pathogenic nuclear variants, so most individuals also have a few mutant mitochondrial genomes.

Importantly, and in complete contrast to mosaicism for a nuclear mutation, a heteroplasmic mother can transmit her heteroplasmic condition to her child.

In a heteroplasmic person the proportion of mutant mitochondrial genomes can vary between tissues and over time. The arguments about genetic drift in Chapter 12 also apply to the mitochondria in a developing heteroplasmic embryo. Mitochondria are not precisely partitioned between daughter cells at mitosis, unlike the way nuclear chromosomes are. When a heteroplasmic cell divides, the daughter cells may not happen to receive exactly the same proportion of normal and mutant mitochondrial genomes. Thus a tissue may end up by chance having a higher or lower proportion of mutant genomes than the overall proportion in the embryo. There is also a role for selection. An unknown quality-control mechanism selects against overtly defective mitochondria. Maybe they replicate less efficiently than normal ones, or cells with a high proportion of defective mitochondria may be out-competed by those with a lower proportion.

During maturation of the germ line, mitochondria pass through a bottleneck. Some stage in the process involves cells having a particularly small number of mitochondrial genomes, or at least selecting a particularly small number for onward transmission. In humans the statistics of mother-child levels of heteroplasmy imply a critical bottleneck of only 30–35 genomes—see the paper by Rebolledo-Jaramillo *et al.* (2014) (PMID 25313049) in Further Reading. As explained in Chapter 12 this allows strong genetic drift, so that the proportion of mutant genomes in post-bottleneck cells may differ considerably from the proportion in cells beforehand. This means that when a heteroplasmic mother has a heteroplasmic child, there is often little relation between the proportion of mutant genomes in the two (**Figure 16.14**). The degree of heteroplasmy is usually estimated from a blood sample, rather than from the germ line or zygote, which adds an extra layer of uncertainty to the relationship.

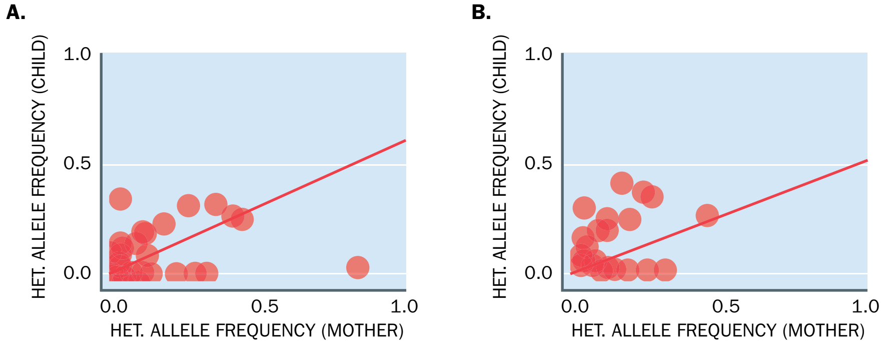


Figure 16.14 Consequences of the mitochondrial bottleneck. There is little relation between the levels of heteroplasmy (het.) in a mother and in her child. (**A**) Levels of various heteroplasmic single nucleotide changes in buccal tissue in mother-child pairs. Correlation $R^2 = 0.13$. (**B**) Similar data for blood; $R^2 = 0.29$. (Adapted from Rebolledo-Jaramillo B *et al.* [2014] *Proc Natl Acad Sci USA* **111**:15474–15479; PMID 25313049. With permission from National Academy of Sciences.)

Conditions caused by mtDNA mutations are extremely variable

Mutations in the mitochondrial DNA have rather unpredictable effects, both in terms of whether or not they will make somebody ill, and in terms of what disease they will cause. A given mtDNA sequence change may be seen in patients with several different diseases, and patients with the same mitochondrial disease may have different mutations in their mtDNA. The MITOMAP database of mitochondrial mutations (http://www.mitomap.org) has extensive data tables showing just how great is the challenge of relating mutations to phenotypes. For example, an A>G change at position 3243 in the tRNA^{Leu} gene has been reported in individuals with the following diseases:

- Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS);
- Chronic progressive external ophthalmoplegia (CPEO);
- Cardiac plus multiorgan dysfunction;
- Mitochondrial myopathy;
- Diabetes mellitus plus deafness;
- Sensorineural hearing loss.

On the other hand, a single mitochondrial phenotype, Leber hereditary optic atrophy (LHON, sudden irreversible loss of vision; OMIM #535000), has been associated with at least 17 different mitochondrial point mutations (see MITOMAP for details). Fifty percent of affected people have an m.11778G>A substitution, causing a missense change, p.R340H, in the *ND4* gene. Most of these patients are homoplasmic, but about 14%, no less severely affected, are heteroplasmic. Even in homoplasmic families the condition is highly variable; penetrance overall is 33–60%, and 82% of affected individuals are male. Two other missense changes, p.A52T (m.3460G>A) in the *ND1* gene and p.M64V (m.14484T>C) in the *ND6* gene, account for all but 5% of other cases, but the remaining few include individuals with various missense mutations in the *ND1*, *ND3,* or *ND6* genes. All these genes encode subunits of the mitochondrial NADH dehydrogenase, which is an essential part of the oxidative phosphorylation complex I. The sudden and late onset implies a response to some external stress, and probably eyes are only vulnerable if some critical cells exceed some threshold proportion of defective mitochondria.

Partial deletions or quantitative depletion of mitochondria are other causes of mitochondrial disease

As well as point mutations, large deletions of the mtDNA are an important cause of disease (**Figure 16.15**). Deletions varying from 1.3 to 8 kb have been found in patients with a number of conditions—CPEO, Kearns–Sayre syndrome, and Pearson syndrome. Often the deletions are found at high frequency in muscle cells, but not in other tissues of the patient. The deletions may arise sporadically, or may be caused by inherited mutations in nuclear genes that are responsible for the replication and maintenance of the mitochondrial genome, such as the gene for mitochondrial-specific DNA polymerase, *POLG*. The smaller, partially deleted mtDNA molecules may have a replicative advantage over full-length molecules, explaining the high frequencies of deletions in muscle cells of patients. In addition, some mitochondrial diseases are caused by a simple quantitative deficiency of mitochondria.

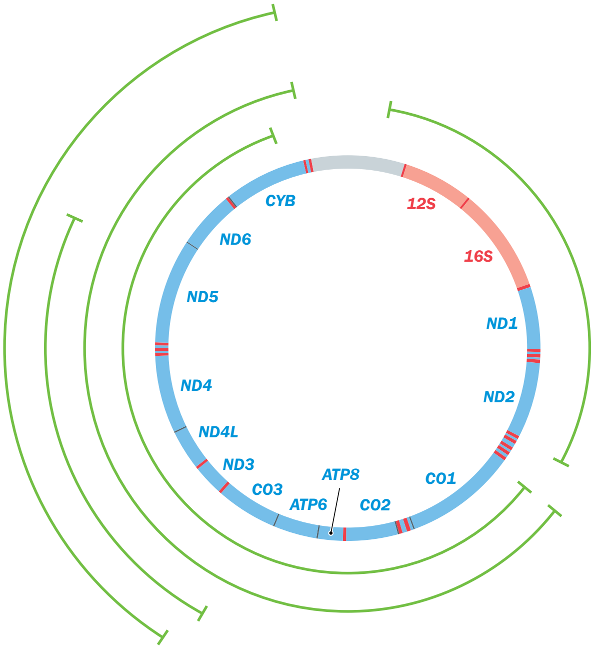


Figure 16.15 Mitochondrial DNA deletions. The curves show examples of large deletions of the mitochondrial genome observed in individuals with sporadic syndromes or inherited defects in nuclear genes necessary for replication or maintenance of the mitochondrial genome.

There are some analogies between the molecular pathology of mitochondrial variants and of cancer. In each case there is a single basic phenotype—unregulated cell proliferation in cancer, deficient energy production in mitochondrial diseases. In both mitochondrial malfunction and cancer, there are different ways, often tissue-specific and context-dependent, of arriving at the common basic phenotype. Mitochondrial DNA is more variable than nuclear DNA, and some syndromes may depend on a combination of the reported mutation with other, unidentified variants. Many mitochondrial functions are encoded by nuclear genes, so that nuclear variation can be an important cause or modifier of mitochondrial phenotypes. There will be some threshold level of mutant mtDNA before oxidative phosphorylation and energy generation are affected. Biochemical studies suggest that most mtDNA mutations need to accumulate to 60–90% of total mtDNA in a cell before energy generation is compromised. The most sensitive tissues are the most metabolically active: skeletal and heart muscle, nerves, brain, and the hair cells of the inner ear.

mtDNA mutations, both point mutations and deletions, have been shown to occur at low levels in different tissues of normal healthy people. Their frequency increases with increasing age of the person. According to the mitochondrial theory of aging, a gradual accumulation of mitochondria deficient in energy production is *the* cause of aging.

However, although there is abundant correlative evidence documenting this increase with age, there is still no clear proof that the changes are a cause of aging rather than one of many features of the aging process.

16.5 GENOTYPE–PHENOTYPE CORRELATIONS

As discussed in the introduction to this chapter, there are always two aspects to molecular pathology: what a variant does to a gene or its product, and what it does to the whole person. So far we have mainly focused on the gene-level aspects. Now it is time to consider phenotypes.

Loss-of-function and gain-of-function mutations in the same gene will cause different phenotypes

We have emphasized the distinction between loss-of-function and gain-of-function mutations in this chapter. Sometimes it is possible to see both types of mutation in the same gene (**Table 16.9**).

TABLE 16.9 EXAMPLES OF DIFFERENT CONDITIONS CAUSED BY LOSS OF FUNCTION OR GAIN OF FUNCTION IN THE SAME GENE

Gene	Location	Loss or gain of function	Condition	Symbol	OMIM #
PMP22	17p11.2	LoF	Hereditary neuropathy with liability to pressure palsies	HNPP	162500
		GoF	Charcot–Marie–Tooth neuropathy type 1A	CMT1A	118220
LHCGR	2p16.3	LoF	46,XY pseudohermaphroditism	LCH	238320
		GoF	Male precocious puberty		176410
GNAS1	20q13.2	LoF	Albright hereditary osteodystrophy	PHP1A	103580
		GoF	McCune–Albright syndrome	MAS	174800
ROR2	9q22.31	LoF	Robinow syndrome	RRS	268310
		GoF	Brachydactyly type B1	BDB1	113000
RET	10q11	LoF	Hirschsprung disease	HSCR	142623
		GoF	Multiple endocrine neoplasia type IIA/IIB	MEN2A/B	171400 162300
		GoF	Familial medullary thyroid cancer	FMTC	155240

See the text for details. GoF, gain of function; LoF, loss of function.

- Nonallelic homologous recombination between low-copy repeats produces microdeletions and microduplications in exactly equal numbers (see **Figure 15.17**). Sometimes the reciprocal dosage changes produce equally reciprocal phenotypic effects—for example, microdeletions at 16p11.2 are associated with obesity while the reciprocal microduplication is associated with being underweight. More often there is no such neat reciprocal relationship, as illustrated by Charcot–Marie–Tooth neuropathy type 1A and hereditary neuropathy with liability to pressure palsies, or by Williams–Beuren syndrome and the corresponding microduplication (see Section 15.3).
- Systems that relay an external signal to the cell interior can have loss-of-function mutations that result in no signal being transmitted, or gain-of-function mutations that cause them to signal even in the absence of ligand. Loss of function of the receptor for luteinizing hormone and gonadotrophin renders tissues unable to respond to those hormones, causing 46,XY embryos to revert to the default female body development, while a gain of function leads to precocious puberty in early childhood. In the case of the G-protein-coupled receptors (see **Figure 3.5A**), constitutional activation would probably be lethal to a developing embryo. Thus McCune–Albright syndrome, caused by gain-of-function mutations at the *GNAS1*

locus that encodes the Gsα subunit, is a mosaic condition with activation in specific cell lineages. This gene locus has a complicated pattern of imprinting (see OMIM #139320) and the loss-of-function phenotype, Albright hereditary osteodystrophy, is seen only when the maternal allele is inactivated. Loss of function of the paternal allele has no effect because that allele is not expressed.

- Loss and gain of function of the ROR2 orphan receptor each result in disturbances of skeletal development: Robinow syndrome and brachydactyly type B1, respectively.
- The *RET* gene encodes a transmembrane receptor tyrosine kinase that responds to Wnt signaling. A variety of loss-of-function mutations are one cause of Hirschsprung disease (OMIM #142623; absence of enteric ganglia in the bowel). Certain very specific missense mutations are seen in a totally different set of diseases: familial medullary thyroid carcinoma (OMIM #155240) and the related but more extensive multiple endocrine neoplasia type II (OMIM #162300 and #171400). These are gain-of-function mutations, producing receptor molecules that react excessively to ligand or are constitutively active and dimerize even in the absence of ligand. Curiously, some people with missense mutations affecting cysteines 618 or 620, which are important for receptor dimerization, suffer from both thyroid cancer and Hirschsprung disease—simultaneous gain and loss of function. This reminds us that loss of function and gain of function are not always simple scalar quantities. No gene product acts in isolation. If a gene product functions in a number of different cellular contexts, mutations may have different effects in the different cell types in which the gene is expressed. One hypothesis to explain the RET effect suggests that the variant proteins may fail to respond to the ligand but may have a constant low level of constitutional activity, which may be insufficient for positive RET function in enteric ganglia, but sufficient to cause constitutional activation in the thyroid.

Loss-of-function and gain-of-function conditions have different distributions of mutations

There are many ways of reducing or abolishing the function of a gene product. When a clinical phenotype results from loss of function of a gene, we would expect any change that inactivates the gene product to produce the same clinical result. We should be able to find point mutations that have the same effect as deletion or disruption of the gene. Gain of function, in contrast, is a rather specific phenomenon. Probably only a very specific change in a gene can cause a gain of function. Thus, the degree of allelic heterogeneity is a strong, although not infallible, pointer to the underlying molecular pathology. For example, among diseases caused by unstable trinucleotide repeats (see **Table 16.7**), fragile X syndrome and Friedreich ataxia are occasionally caused by other types of mutation in their respective genes, pointing to a loss of function, whereas Huntington disease and myotonic dystrophy are never seen with any other type of mutation, suggesting a gain of function. **Figure 16.16** shows the contrasting mutational spectra of a loss-of-function and a gain-of-function condition.

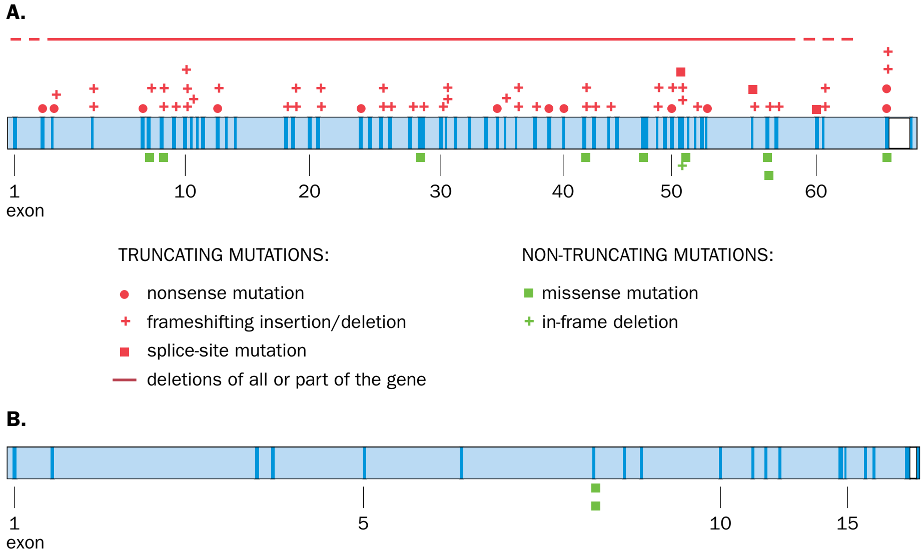


Figure 16.16 The contrasting mutational spectra of a loss-of-function and a gain-of-function condition. (**A**) The spectrum of mutations in the *ATM* gene in a series of unrelated patients with ataxia-telangiectasia (OMIM #208900). The great variety of different mutations, most of which would truncate the gene product, show that this recessive disease is caused by loss of function of the *ATM* gene. (**B**) Only two very specific missense mutations, p.S252F and p.P253R, in the *FGFR2* gene cause all cases of Apert syndrome (OMIM #101200). This dominant syndrome is caused by a specific gain of function of the fibroblast growth factor receptor 2 protein.

Allelic homogeneity is not always due to a gain of function

Allelic heterogeneity is normally a hallmark of loss-of-function phenotypes. However, it is not safe to assume that any condition that shows allelic homogeneity is caused by a gain of function. There are other possible explanations:

- In some diseases, the phenotype is very directly related to the gene product itself, rather than being a more remote consequence of the genetic change. The disease may then be defined in terms of a particular variant product, as in sickle cell anemia;
- Some specific molecular mechanism may make a certain sequence change in a gene much more likely than any other change—for example, the $(CGG)_n$ expansion is seen in the overwhelming majority, though not every single case, of fragile X syndrome patients;
- There may be a founder effect (see **Figure 12.7**). For example, certain disease mutations are common among Ashkenazi Jews. Present-day Ashkenazi Jews are descended from a fairly small number of founders. If one of those founders carried a recessive allele, it may be found at high frequency in the present Ashkenazi population;
- Selection favoring heterozygotes enhances founder effects and often results in one or a few specific mutations being common in a population.

Different degrees of loss of function may produce an allelic series of phenotypes

Loss of function is not necessarily an all-or-nothing affair. Different mutations in a gene may result in a partial or a total loss of function, producing an allelic series. Sometimes the effect is quantitative: a reduction in the amount of a structurally normal protein versus complete absence. Mutations that reduce but do not abolish correct splicing of a primary transcript are a common cause of mild disease. They may weaken a normal splice site, like the c.92+6T>C β-globin mutation that causes a β^{+}-thalassemia (see **Table 16.3**). Activating a weak cryptic splice site can also cause inefficient splicing, as seen with the c.93–21G>A and c.316–106C>G β-globin mutations (see **Table 16.3**) or the mild cystic fibrosis mutation c.3849+12191C>T.

Alternatively, a mutated gene may encode a protein with reduced but not zero function. Complete absence of dystrophin causes the severe Duchenne muscular dystrophy; the milder Becker form is seen when a variant dystrophin is partly functional (see **Table 16.4**). Mutations in the X-linked hypoxanthine guanine phosphoribosyltransferase (HPRT) gene that cause increasing degrees of loss of function produce increasingly severe results (**Table 16.10**).

TABLE 16.10 CONSEQUENCES OF DECREASING FUNCTION OF HYPOXANTHINE GUANINE PHOSPHORIBOSYLTRANSFERASE (HPRT)

HPRT activity (% of normal)	Phenotype
>60	No disease
8–60	Gout; no neurological problems (Kelley–Seegmiller syndrome)
1.6–8	Gout plus variable neurological signs (clumsiness, choreoathetosis); normal intelligence
1.4–1.6	Lesch–Nyhan syndrome, but with normal intelligence
<1.4	Lesch–Nyhan syndrome (OMIM #300322): spasticity, choreoathetosis, self-mutilation, intellectual disability

Loss-of-function changes may produce either dominant or recessive conditions

Loss-of-function phenotypes might be inherited as either dominant or recessive traits. Where a variant interferes with the function of its normal counterpart, the result will be a dominant (dominant-negative) condition, as illustrated in **Figure 16.8**. If a variant causes a simple loss of function, the inheritance will depend on how tolerant particular cells or tissues are to a 50% reduction in function (**Figure 16.17**).

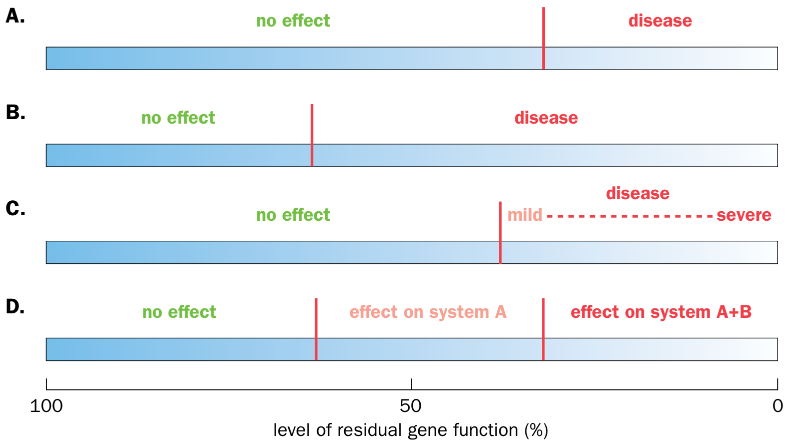


Figure 16.17 Four possible relationships between loss of gene function and clinical phenotype. The bars show the overall level of function produced by the combined effects of the two alleles of the gene. (**A**) This condition, with 50% of residual gene function causing no effect, will be a simple recessive. (**B**) This condition will be dominant because of haploinsufficiency; loss of 50% of gene function causes the disease. (**C**) This condition is recessive, but the severity depends on the level of residual function, so that there is a genotype–phenotype correlation. (**D**) If the clinical consequences are very different, depending on the degree of residual function, the results may be described as different syndromes (A and A + B), and they may have different modes of inheritance, as here. Specific examples are discussed in the text.

Figure 16.17A shows the most common situation. Cystic fibrosis would be an example. Heterozygous carriers of a loss-of-function mutation are entirely healthy and normal. Figure 16.17B shows **haploinsufficiency**. A 50% overall level is not sufficient for normal function. A single loss-of-function variant in a heterozygous person produces a phenotype, which is therefore inherited as a dominant condition. In clinical diagnostics it is often important to decide whether a patient's heterozygosity for a deletion or a loss-of-function mutation in a particular gene could be responsible for their condition. There is therefore great interest in developing predictors of haploinsufficiency. When panels of known haploinsufficient and haplosufficient genes (the latter identified from deletions or mutations found when sequencing healthy controls) are compared, a number of generalizations emerge:

- The coding sequences and promoters of haploinsufficient genes tend to be more highly conserved in evolutionary comparisons;
- Paralogs of haploinsufficient genes tend to have lower sequence similarity (so that a haplosufficient gene is more likely to have a closely similar paralog that can provide a backup for a loss-of-function mutation);
- Haploinsufficient genes tend to have higher expression levels in early embryonic development;
- Haploinsufficient genes tend to have more interaction partners in both protein–protein and gene interaction networks;
- Haploinsufficient genes have higher chances of interacting with other haploinsufficient genes.

Predictive haploinsufficiency scores have been developed by Huang and colleagues and by others (see the paper by Steinberg and colleagues [PMID 26001969] in Further Reading for discussion). In the haploinsufficiency score of Huang *et al.*, proximity to known haploinsufficient genes in the probabilistic gene network was the single most heavily weighted predictor. As more and more sequences of normal genomes become available, the predictions should steadily become more reliable, but the focus on single genes ignores the important roles of context and genetic background.

One might reasonably ask why there should be haploinsufficiency for any gene product. Why has natural selection not managed things better? If a gene is expressed so that two copies make only a barely sufficient amount of product, selection for variants with higher levels of expression should lead to the evolution of a more robust organism, with no obvious price to be paid. The answer is that in many cases this has indeed happened—which is why relatively few genes are dosage-sensitive. There are a few cases in which a cell with only one working copy of a gene just cannot meet the demand for a gene product that is needed in large quantities. An example may be elastin (see Section 15.3). In people heterozygous for a deletion or loss-of-function mutation of the elastin gene, tissues that require only modest quantities of elastin (skin and lung, for example) are unaffected, but the aorta, where much more elastin is required, often shows an abnormality, supravalvular aortic stenosis (OMIM #185500). Hemoglobin and type 1 collagen are other examples. However, certain gene functions are inherently dosage-sensitive. These include:

- Gene products that are part of a quantitative signaling system whose function depends on partial or variable occupancy of a receptor or a DNA-binding site, for example;
- Gene products that compete with each other to determine a developmental or metabolic switch.

In each case, the gene product is titrated against something else in the cell. What matters is not the correct absolute level of product, but the correct relative levels of interacting products. The effects are sensitive to changes in all the interacting partners; thus, these dominant conditions often show highly variable expression. Genes whose products act essentially alone, such as many soluble enzymes of metabolism, seldom show dosage effects.

Examples of allelic series of differing severity (the case in Figure 16.17C) have been discussed above. Sometimes, different amounts of residual gene function can give rise to phenotypes sufficiently different that they are labeled as separate clinical conditions (Figure 16.17D). Mutations in the *DTDST* sulfate transporter gene cause autosomal recessive skeletal dysplasias that have been given different names depending on their severity: diastrophic dysplasia (OMIM #222600), multiple epiphyseal dysplasia 4 (OMIM #226900), atelosteogenesis II (OMIM #256050), and achondrogenesis type 1B (OMIM #600972), in order of increasing severity. Karniski (2001) showed that the severity depended on the overall level of residual DTDST function (PMID 11448940; see Further Reading). Extracellular matrix is rich in sulfated proteoglycans such as heparan sulfate and chondroitin sulfate, and defects in sulfate transport interfere with skeletal development.

Loss of individual functions of a multifunctional protein can cause different phenotypes

If a protein has several functions, different variants may cause loss or gain of different functions and have different phenotypic effects. The complicated molecular pathology of p63 mutations illustrates some of the complexities.

p63 (OMIM #603273) is a transcription factor that has several functional domains (Figure 16.18). As always with transcription factors, there are DNA-binding and transactivation domains. The active form is a tetramer, formed through the ISO domain. The SAM (sterile alpha motif) domain mediates additional protein–protein interactions, and there is also a transactivation-inhibitory (TID) domain that modulates the activity of the main TA domain. The *TP63* gene encodes a variety of p63 isoforms as a result of at least four alternative transcription start sites and extensive alternative splicing. Isoforms lacking exons 1–3 encode proteins that lack the transactivation domain. These probably act as dominant-negative inhibitors of signalling by p63 and maybe also p53.

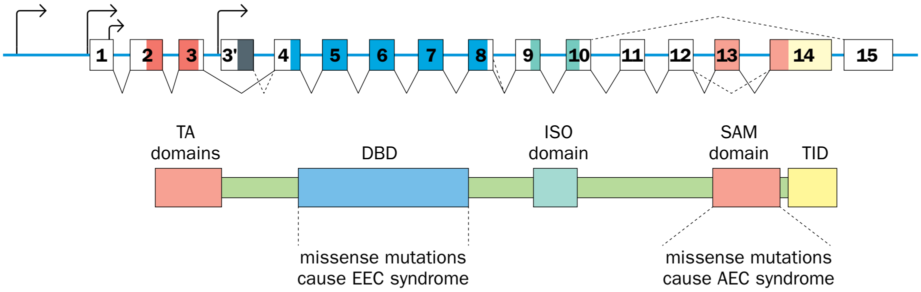

1
2
3
3'
4
5
6
7
8
9
10
11
12
13
14
15
TA domains
DBD
ISO domain
SAM domain
TID
missense mutations cause EEC syndrome
missense mutations cause AEC syndrome

Figure 16.18 Isoforms and domains of the p63 protein. TA, transactivation domain; DBD, DNA-binding domain; ISO, tetramerization domain; SAM, sterile alpha motif (protein interaction) domain; TID, transactivation-inhibitory domain; EEC, ectrodactyly–ectodermal dysplasia–clefting syndrome; AEC, ankyloblepharon–ectodermal dysplasia–clefting syndrome. Arrows show alternative transcription start sites; bent lines show various splicing patterns. There are additional splice isoforms, not shown. (Adapted from van Bokhoven H & Brunner HG [2002] *Am J Hum Genet* **71**:1–13; PMID 12037717. With permission from Elsevier.)

Heterozygous mutations have been reported in patients with six different conditions (Table 16.11). A number of clear genotype–phenotype correlations emerge. Almost all EEC patients have missense mutations in the DNA-binding domain, and particular variants correlate with particular subphenotypes of the syndrome. Almost all AEC patients, who lack the limb abnormalities of EEC syndrome, have missense mutations in the SAM domain. Truncating mutations are not seen in either condition, suggesting that the effects may be due to gains of function. The rare ADULT, LMS, and SHFM4 conditions each have their own distinctive pattern of mutations, some of which are predicted to truncate the protein. Van Bokhoven and Brunner (PMID 12037717; see Further Reading) discuss the molecular pathology in detail.

Genotype–phenotype correlations are often sought but seldom found

Given the complexity of genetic interactions, it is not surprising that molecular pathology is a very imperfect science. The greatest successes so far have been in understanding cancer and hemoglobinopathies: for cancer, the phenotype to be explained—uncontrolled cell proliferation—is relatively simple, while hemoglobinopathies are a very direct result of abnormalities in an abundant and readily accessible protein. For most genetic

TABLE 16.11 CONDITIONS WHERE P63 MUTATIONS HAVE BEEN REPORTED

Condition	OMIM #
Ectrodactyly–ectodermal dysplasia–clefting (EEC) syndrome	604292
Ankyloblepharon–ectodermal dysplasia–clefting (AEC or Hay–Wells) syndrome	106260
Limb–mammary syndrome (LMS)	603543
Acro–dermato–ungual–lacrimal–tooth (ADULT) syndrome	103285
Rapp–Hodgkin syndrome	129400
Nonsyndromic split-hand/split-foot malformation 4 (SHFM4)	605289

EEC syndrome is the prototype; the other syndromes share overlapping features with EEC. Patients with AEC syndrome lack the limb abnormalities; LMS patients have mammary gland aplasia or hypoplasia but fewer ectodermal defects compared to EEC patients. ADULT cases lack orofacial clefting. Rapp–Hodgkin syndrome resembles AEC with milder skin manifestations. Each individual syndrome is variable in itself.

diseases, the clinical features are the end result of a long chain of causation, and the holy grail of molecular pathology, genotype–phenotype correlation, will always be elusive. In reality, even simple Mendelian diseases are not simple at all. The old reviews on this topic by Scriver & Waters (1999) (PMID 10390625) and by Weatherall (2001) (PMID 11283697) are still strongly recommended further reading.

Gain-of-function conditions, with their requirement for very specific sequence changes, are the most likely to provide good genotype–phenotype correlations. A single copy of the variant in a heterozygous person will still have its abnormal function, and so we would expect the phenotype (if any) to be dominant. Whether homozygosity for the variant would have a stronger effect is a secondary question. In humans, most pathogenic gain-of-function variants are rare, and homozygous individuals are doubly rare. They may be seen if there is assortative mating, as in achondroplasia, or as a result of incest. In some cases homozygotes have a more extreme phenotype, in others they resemble heterozygotes. For most human dominant conditions we simply don't know—the offspring of a relevant mating have never been observed. As we argued in Section 5.2, regardless whether homozygotes have a more extreme phenotype or not, these conditions are correctly described as dominant (and not co-dominant or semi-dominant) because the label attaches to the phenotype, not the variant.

Very occasionally, different specific sequence changes can cause a gene to gain different aspects of its function. In such cases one may see very nice genotype–phenotype correlations. The classical case is fibroblast growth factor receptor mutations. As already mentioned, specific mutations in *FGFR2* produce Crouzon or Apert syndrome—but other specific changes cause yet other abnormalities (**Figure 16.19**).

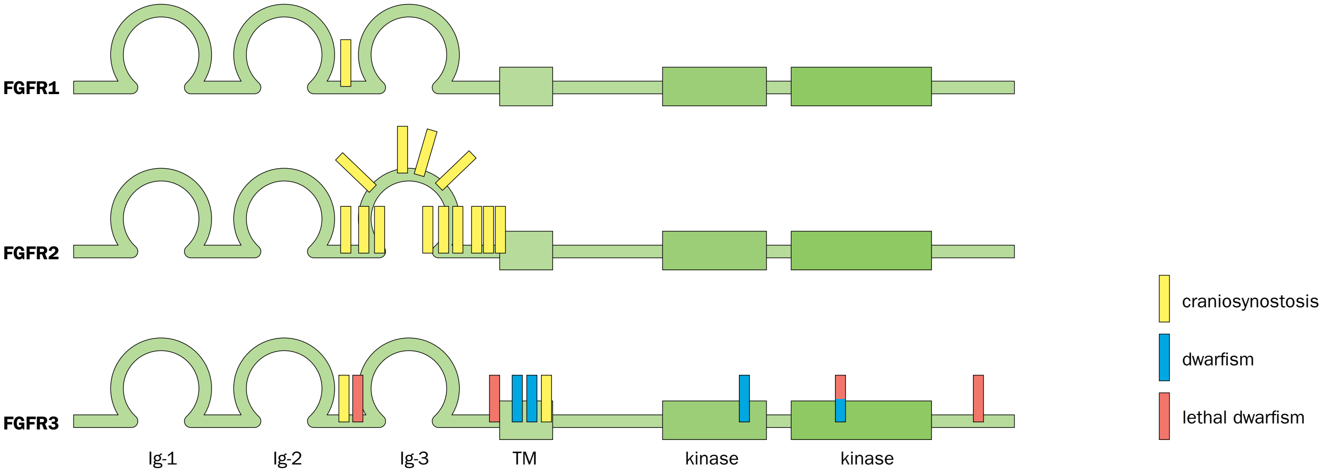


Figure 16.19 Genotype–phenotype correlations among *FGFR1–3* mutations. The fibroblast growth factor receptors (FGFR) 1, 2, and 3 have three extracellular immunoglobulin (Ig)-like domains involved in ligand binding, a transmembrane domain (TM), and two intracellular tyrosine kinase domains that transmit the signal into the cell. Specific gain-of-function mutations cause specific conditions with strong genotype–phenotype correlations (as shown by short vertical bars). (Adapted from Robertson SC *et al*. [2000] *Trends Genet* **16**:265–271; PMID 10827454. With permission from Elsevier.)

Traditional studies of genotype–phenotype correlations were based on highly selected samples

Until recently, researchers interested in a particular disease would collect individuals or families with that disease, sequence one or more candidate genes, and report the variants found. A limited number of healthy controls might also be checked to make sure a variant was not a common population polymorphism. This produced a phenotype-centric view of variants. Pathogenic variants are typically rare (frequency <0.1%) and the study design could not identify rare healthy individuals carrying the variant, or cases where a variant in the same gene caused a different phenotype. Moreover, in the past, when finding variants was difficult, laboratories were sometimes a bit too ready to label a variant found in a patient as pathogenic, and journals were sometimes a bit too uncritical in publishing such findings.

The consequence of these limitations is that classical studies produced far too narrow a view of genotype–phenotype correlations. They overestimated the penetrance of many variants, and they underestimated the range of phenotypes that could result from variants in a given gene. The current era of large-scale sequencing of exomes and genomes has allowed a genotype-centric view of variants. Now that we have exome or genome sequences of tens of thousands of healthy individuals (or individuals who have conditions irrelevant to a disease under study) we can look to see whether any of them have a variant that we were supposing is pathogenic. The 1000 Genomes data were widely used for this, but the ExAC (Exome Aggregation Consortium) has far greater power, since it unites data on over 60,000 exomes (exac.broadinstitute.org). In late 2016 this expanded to the GnomAD (Genome Aggregation database; gnomad.broadinstitute.org) that is more than twice the size, and is likely to grow even larger.

The new-found ability to check such a large sample of unaffected individuals for a supposedly pathogenic variant has led to three conclusions:

- Many supposed pathogenic variants in databases of disease mutations are in fact harmless and present at similar frequencies in healthy individuals;
- The penetrance of truly pathogenic variants is often lower than previously supposed;
- For a number of genes, loss-of-function mutations in both alleles (homozygous or compound heterozygous) can be seen in normal healthy individuals.

Regarding the first of these points, the following examples illustrate the need for a careful review of the evidence for pathogenicity of any variant that is to be included in a clinical report:

- In a 2011 study of the feasibility of highly multiplexed carrier screening for autosomal recessive conditions by Bell and colleagues (PMID 21228398), 460 variants were identified that were cited in published literature as disease-causing. The authors rejected 122 of these as being common polymorphisms or sequencing errors, or because of a lack of evidence of pathogenicity;
- In 2012 Hunt and colleagues (PMID 22200769) looked at nine variants in the *SIAE* (sialic acid acetylesterase) gene that had been shown to affect gene function and that had been reported to occur at increased frequency in people with various autoimmune diseases for which *SIAE* would be a plausible candidate gene. In a very large series of cases and controls (up to 66,000 in total) they found no excess of any of the nine variants among the cases. Thus although they affected gene function, the variants did not apparently cause the claimed disease;
- Andreasen and colleagues in 2013 (PMID 23299917) investigated the prevalence of previously reported cardiomyopathy-associated variants in 6500 exomes reported in the NHLBI GO Exome Sequencing Project (ESP) from individuals without any known relevant disease. Of 1233 variants reported as autosomal-dominant causes of monogenic disease, 190 (15.4%) were found in the ESP cohort, most of them at frequencies far too high to be plausible causes of these rare conditions, even allowing for some degree of reduced penetrance.

The tendency of disease studies to overestimate the penetrance of truly pathogenic variants is well illustrated by a study in 2013 by Flannick and colleagues (PMID 24097065) of MODY (maturity-onset diabetes of the young; OMIM #606391). This is a Mendelian dominant form of type 2 diabetes that is unrelated to diet or lifestyle, and that normally manifests before age 25. It can be caused by heterozygous mutations in any of at least seven genes (*HNF1A*, *GCK*, *HNF4A*, *HNF1B*, *PDX1*, *INS,* and *NEUROD1*). Flannick and colleagues identified variants in these genes in 4003 individuals from well-studied

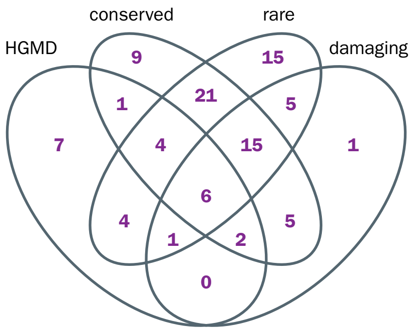


Figure 16.20 Classification of 96 uncommon (minor allele frequency [MAF] <1%) nonsynonymous variants in genes associated with maturity-onset diabetes of the young identified in a random population survey. "Rare" means private to one study individual out of the 4003 cases and not observed in the 1000 Genomes project. "Conserved" means located at an evolutionarily conserved site of the relevant protein. "Damaging" means predicted as damaging by the SIFT and PolyPhen-2 programs. "HGMD" means listed in the Human Gene Mutation Database as pathogenic. (Reprinted from Flannick J *et al*. [2013] *Nat Genet* **45**:1380–1385; PMID 24097065. With permission from Springer Nature. Copyright © 2013.)

population cohorts. **Figure 16.20** shows how these variants could be classified. Twenty-one of the variants are rare, affect conserved sites, and are predicted to be damaging; six of those are listed in the Human Gene Mutation Database as causative of MODY. At least these 21 variants, and maybe some of the others, would normally be listed in a clinical analysis of a MODY patient. Yet none of these well-phenotyped individuals had overt MODY, and the vast majority remained euglycemic in middle age.

A number of other studies have sought rare "resilient" individuals who seem perfectly healthy despite apparently having genotypes normally considered pathogenic. It is difficult to know how much credence to put on some of the more dramatic examples. The initial list of candidates is often large; it is then whittled down by seeking alternative explanations—sequencing errors, annotation errors, individuals mistakenly said to be unaffected, a sample attributed to the wrong individual, and so on. Eventually there remains a very small core of cases where no alternative explanation could be found. One has to wonder how often there might nevertheless actually be an alternative explanation.

It is clear that loss of function of a gene is not necessarily pathogenic. For example, MacArthur and colleagues (2012; PMID 22344438; see Further Reading) checked the whole-genome sequences of 185 supposedly healthy individuals from the pilot phase of the 1000 Genomes project for apparent loss-of-function variants in protein-coding genes. These were defined as nonsense or splice-site mutations, insertion/deletion (indel) variants predicted to disrupt the reading frame, or larger deletions removing either the first exon or more than 50% of the protein-coding sequence of the affected transcript. Missense variants were not considered, and the sequencing technology used would not pick up large deletions, so the survey addressed only some causes of a loss of function.

As expected, the initial list included many false positives: errors in the raw sequence data or in the annotation, or variants unlikely to cause a real loss of function. There were cases of **compensated pathogenic deviation**. This is the phenomenon described by Jordan and colleagues (see Section 17.5 and PMID 26123021 in Further Reading) where a real loss of function due to one variant is rescued by a second nearby change—for example, a frameshifting insertion might be balanced by a nearby deletion, a pathogenic amino acid substitution might be rendered harmless by substitution of a nearby interacting amino acid, or loss of a splice site might be rescued by creation of an alternative. After rigorous filtering, 1285 high-confidence variants remained: 415 of them affected only a subset of the known transcripts of the affected gene; the rest caused a complete loss of function.

It was estimated that the average healthy individual carries around 100 loss-of-function variants (plus those due to missense changes or large deletions, which were not considered in this work). Some of these would be genuine recessive alleles in a phenotypically normal heterozygote, but on average around 20 genes were homozygously inactivated. Evidently some genes are not essential for normal development or health. For example, the many people with blood group O are homozygous for inactivating mutations in the ABO blood group gene. The list was enriched for genes that have closely related paralogs that might be able to provide the missing function, and genes involved in smell and taste sensation, where variations would not be expected to affect general health.

It is clear that many healthy individuals do carry loss-of-function variants of one or another gene, and that many variants described in databases as pathogenic are either nonpathogenic or show reduced penetrance. These studies provide a cautionary tale for diagnostic laboratories: finding a loss-of-function variant is only the beginning of the investigation of a patient, not the end of the search. MacArthur and colleagues (2014) provided a set of guidelines for assessing the pathogenicity of a variant (see PMID 24759409 in Further Reading).

Development of large-scale phenotype databases will be an important tool for the future of molecular pathology

Genotype–phenotype correlations depend on accurate descriptions. For genotypes this is simply the DNA sequence (though a future requirement for epigenetic description might complicate the issue). For phenotypes the answer is less obvious. Most genotype–phenotype correlations describe phenotypes in terms of entries in the OMIM database. Applying the correct OMIM label to a patient is a crucial part of clinical genetics. For the patient, or the parent of a disabled child, a diagnosis means that, at last, somebody understands their condition. It puts an end to a diagnostic odyssey that can involve years of different clinics, clinicians, and tests. It provides a label that can be used when filling in the forms necessary to access social or educational support, and it can allow access to a support group for families with the same syndrome. For the clinician it can suggest the prognosis and help guide management. And by grouping the patient together with others with the same syndrome it can help illuminate the molecular pathology. Clinicians use traditional clinical judgment to arrive at a diagnosis. They will of course consider each separate feature of the patient, but the end judgment relies on an overall impression. The leading international experts are constantly being contacted by colleagues elsewhere about difficult cases. The colleague will send a list of features, but the expert would not make a judgment without seeing photographs, because the judgment relies on a feeling for the overall "Gestalt."

Clinical geneticists are rightly proud of their impressive track record of delineating rare and subtle syndromes—but their way of working does not well fit the era of Big Data. Thus there are calls for a Human Phenome Project. The idea is to record phenotypes in a standardized systematic way using a controlled vocabulary, to produce data suitable for computer analysis. Just as the ability to ask questions across the whole range of DNA sequences has led to many advances in biology and medicine, so an ability to do the same across the whole range of human phenotypes might lead to new insights into disease mechanisms and the relationships between diseases.

A step in this direction is the Human Phenotype Ontology (described by Robinson & Mundlos [2010]; PMID 20412080; see Further Reading). The HPO allows a systematic machine-searchable description of phenotypes. This does not replace clinical judgment but provides an additional and complementary tool. This may be helpful for diagnosis, but, in the context of molecular pathology, a major application is in allowing computer algorithms to make searches across the total phenotype space, paralleling the way DNA sequences can be compared across the totality of recorded sequences. There is still a long way to go before the potential of such approaches is fully realized, but the paper by Oti *et al*. (2009) (PMID 20004759; see Further Reading) shows how this might work. The future of molecular pathology may well lie in this direction.

SUMMARY

- Molecular pathology seeks to explain why a certain genetic change causes a particular phenotype. The task can be split into explaining the effect of a sequence change on gene function, and explaining the effect of a change in gene function on a person's phenotype.
- Pathogenic effects can be mediated by either a loss of function or a gain of function of a gene product.
- Loss of function of a protein or RNA can be due to deletion (total or partial) of the gene, disruption of the sequence, a failure to splice exons correctly, or a change in a regulatory sequence. For a protein, loss can additionally be due to a frameshifting insertion or deletion, introduction of a premature termination codon, or replacement of an important amino acid.
- A change can affect one or all isoforms of a gene product. A loss of function can be partial or total and may affect one or all functions of a multifunctional protein.
- Changes in protein-coding sequence are often readily interpretable, but splicing effects may need experimental confirmation, and functional studies may be needed to confirm the predicted effect of missense changes.
- Gain of function can occur in various ways. Often a gene product functions excessively and inappropriately. Rarely, it may acquire a novel function. Sometimes the mutant mRNA or protein is toxic, for example by sequestering factors needed for RNA processing or by forming toxic protein aggregates.
- Dynamic mutations, where a microsatellite becomes unstable and prone to large expansions, probably all depend on similar DNA-level mechanisms, but they differ greatly in whether their pathogenic effects are exerted at the RNA or protein level, and whether they involve a loss or gain of function.
- Explaining the effect of a variant on a person's phenotype is much more difficult than explaining its effect on

a gene product. Loss of function of a gene product is not necessarily pathogenic. Some gene products are entirely dispensable.

- Pathogenic gain-of-function changes usually lead to dominant phenotypes. Pathogenic loss-of-function changes can lead to either recessive or dominant phenotypes. Dominant phenotypes are seen when a 50% level of the normal gene product is not sufficient to produce a normal phenotype (haploinsufficiency), or if the protein product of the mutant allele interferes with the function of the normal product (a dominant-negative effect).
- Understanding the effects of changes in mitochondrial DNA is complicated by heteroplasmy. The proportion of mutant mitochondrial genomes can vary between tissues of an individual, can change over time, and can differ markedly between mother and child.
- Traditional phenotype-centric views of genotype–phenotype relationships are being challenged by data from large-scale sequencing projects.
- Well-defined genotype–phenotype correlations are rare in humans because we differ greatly from one another in our genetic makeup and environments.
- This chapter has focused on the mechanisms by which a genotype can lead to a phenotype. When a clinical diagnostic laboratory is trying to decide whether a variant in a patient is the cause of their condition, they will use additional information to make their judgment, as described in Chapter 20.

FURTHER READING

Loss-of-function changes

Di Giacomo D *et al.* (2013) Functional analysis of a large set of *BRCA2* exon 7 variants highlights the predictive value of hexamer scores in detecting alterations of exonic splicing regulatory elements. *Hum Mutat* **34**:1547–1557; PMID 23983145.

Huang N *et al.* (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**:e1001154; PMID 20976243.

MacArthur DG *et al.* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**:823–828; PMID 22344438.

Steinberg J *et al.* (2015) Haploinsufficiency predictions without study bias. *Nucleic Acids Res* **43**:e101; PMID 26001969.

Other pathogenic mechanisms

Albrecht A & Mundlos S (2005) The other trinucleotide repeat: polyalanine expansion disorders. *Curr Opin Genet Dev* **15**:285–293; PMID 15917204.

Cleary JD & Ranum LPW (2013) Repeat-associated non-ATG (RAN) translation in neurological disease. *Hum Mol Genet* **22(R1)**:R45–R51; PMID 23918658.

Jenkinson EM *et al.* (2016) Mutations in *SNORD118* cause the cerebral microangiopathy leukoencephalopathy with calcifications and cysts. *Nat Genet* **48**:1185–1192; PMID 27571260.

Jordan DM *et al.* (2015) Identification of *cis*-suppression of human disease mutations by comparative genomics. *Nature* **524**:225–229; PMID 26123021.

Lupiáñez DG *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**:1012–1025; PMID 25959774.

Starck SR *et al.* (2016) Translation from the 5′ untranslated region shapes the integrated stress response. *Science* **351**:aad3867; PMID 26823435.

Weingarten-Gabbay S *et al.* (2016) Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**:aad4939; PMID 26816383. (See also the comment by Ivan Shatsky in PubMed Commons, below the PubMed entry.)

Zhao X-N & Usdin K (2015) The repeat expansion diseases: the dark side of DNA repair. *DNA Repair* **32**:96–105; PMID 26002199.

Genotype–phenotype correlations

Jacquemont S *et al.* (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478**:97–102; PMID 21881559.

Karniski LP (2001) Mutations in the diastrophic dysplasia sulfate transporter (DTDST) gene: correlation between sulfate transport activity and chondrodysplasia phenotype. *Hum Mol Genet* **10**:1485–1490; PMID 11448940.

Rauen KA (2013) The RASopathies. *Annu Rev Genomics Hum Genet* **14**:355–369; PMID 23875798.

Rebolledo-Jaramillo B *et al.* (2014) Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc Natl Acad Sci USA* **111**: 15474–15479; PMID 25313049.

Scriver CR & Waters PJ (1999) Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet* **15**:267–272; PMID 10390625.

Tuppen H *et al.* (2010) Mitochondrial DNA mutations and human disease. *Biochem Biophys Acta* **1797**:113–128; PMID 19761752.

van Bokhoven H & Brunner HG (2002) Splitting p63. *Am J Hum Genet* **71**:1–13; PMID 12037717.

Weatherall DJ (2001) Phenotype–genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat Rev Genet* **2**:245–255; PMID 11283697.

Wilkie AOM & Morriss-Kay GM (2001) Genetics of craniofacial development and malformation. *Nat Rev Genet* **2**:458–468; PMID 11389462.

Xiong HY *et al.* (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**:1254806; PMID 25525159.

Pathogenic and nonpathogenic variants

Andreasen C *et al.* (2013) New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur J Hum Genet* **21**:918–928; PMID 23299917.

Bell CJ *et al.* (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**:65ra4; PMID: 21228398

Flannick J *et al*. (2013) Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat Genet* **45**:1380–1385; PMID 24097065.

Hunt KA *et al*. (2012) Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. *Nat Genet* **44**:3–5; PMID 22200769.

MacArthur DG *et al*. (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**:469–476; PMID 24759409.

Human Phenotype Ontology

Köhler S *et al*. (2016) The Human Phenotype Ontology in 2017. *Nucleic Acids Res* **45(D1)**:D865–D876; PMID 27899602.

Oti M *et al*. (2009) The biological coherence of human phenome databases. *Am J Hum Genet* **85**:801–808; PMID 20004759.

Robinson PN & Mundlos S (2010) The Human Phenotype Ontology. *Clin Genet* **77**:525–534; PMID 20412080.

17 Mapping and identifying genes for monogenic disorders

In itself a genome sequence gives no information about what (if any) phenotype a variant might cause, or which variant might be responsible for a patient's condition. The researcher has to make that connection. There are two main approaches to doing this. Nowadays the usual approach is to sequence the patient's exome (the totality of all protein-coding exons) or their whole genome. This would produce a list of maybe 20,000 variants in the exome, or 4 million in the whole genome. The list must then somehow be narrowed down to identify the one causative variant. This process is described in Sections 17.3–17.5.

In times past, before next-generation sequencing, finding variants was much more difficult. The strategy then was to define as closely as possible the chromosomal location of the unknown causative variant, so as to minimize the amount of Sanger sequencing needed to identify variants—a process known as **positional cloning**. Sometimes this could be done by finding a patient with the condition in question who had a chromosomal deletion or rearrangement. This might be purely coincidental, but alternatively the chromosomal abnormality might have affected expression of the causative gene in this particular patient. That would be especially plausible if the patient had a *de novo* case of a condition that was usually dominantly inherited, together with a *de novo* chromosome abnormality. That would suggest that the relevant gene might very well lie at the affected location. Then variants at that location could be sought in patients with the same condition but without any visible chromosomal abnormality.

Such lucky cases were important in the early days of gene identification, but they were exceptional. More usually the location had to be defined by linkage analysis, using a panel of multicase families. Such studies are less central to genetics research today—not because they are obsolete in any technical sense, but because almost all the monogenic conditions where suitable families can be found have already been investigated and the causative gene identified. There are plenty of multicase families with non-Mendelian conditions like diabetes or schizophrenia that are still under active investigation, but these need a different approach, which will be described in Chapter 18. A linkage approach is also frequently used in autozygosity mapping, described in Section 17.2. In Section 17.1 we will describe the principles of linkage analysis and positional cloning because linkage analysis is an important area of genetic thinking that today's students should appreciate, at least in general terms. Also, as explained by Ott and colleagues (2015) (PMID 25824869; see Further Reading), linkage still has a place in the world of whole-genome sequencing because it can be a powerful tool for prioritizing individuals for sequencing and for reducing the list of variants that need filtering.

This chapter is about research, identifying the genetic variants underlying monogenic conditions. In the past this activity was clearly distinct from clinical genetic testing for diagnosis or prediction, as summarized in **Figure 17.1**. Now that exome sequencing is used not only for research but also for clinical diagnosis, there is a substantial overlap. In both cases there is a need for clarity about what the patient or research subject has consented to, particularly in terms of what will be reported back. These issues will be considered in Chapter 20 when we look at testing and screening for genetic variants.

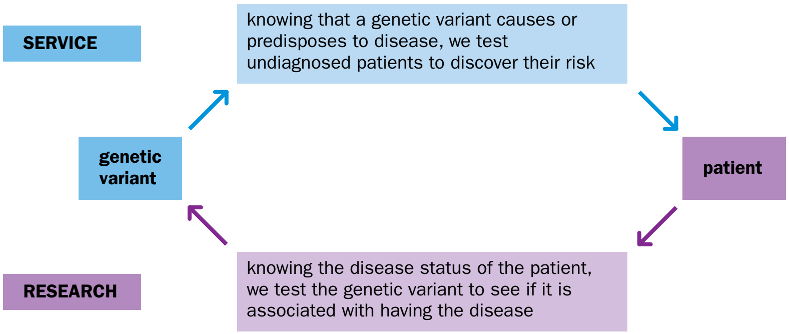

SERVICE
knowing that a genetic variant causes or predisposes to disease, we test undiagnosed patients to discover their risk
genetic variant
patient
RESEARCH
knowing the disease status of the patient, we test the genetic variant to see if it is associated with having the disease

Figure 17.1 Genetic testing for clinical service or for research. The figure summarizes the difference when testing involves a single variant. In the era of exome sequencing, the distinction has become blurred, raising important issues about consent and reporting.

17.1 POSITIONAL CLONING SEEKS TO IDENTIFY DISEASE GENES BY FIRST MAPPING THEM TO A PRECISE CHROMOSOMAL LOCATION

During the 1980s and 1990s the genes responsible for most of the more frequent monogenic diseases were identified by positional cloning. **Figure 17.2** shows the principle. In linkage analysis, a panel of known variants (genetic markers) scattered across the genome is tested with the aim of finding one that reliably tracks with the unknown disease-causing variant through a collection of pedigrees. If recombination during meiosis seldom or never separates the two, they must reside close together at the same chromosomal location.

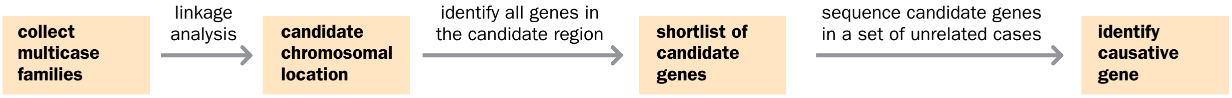

collect multicase families
linkage analysis
candidate chromosomal location
identify all genes in the candidate region
shortlist of candidate genes
sequence candidate genes in a set of unrelated cases
identify causative gene

Figure 17.2 Identifying disease genes by positional cloning. The procedure depends on being able to collect sufficient multicase families for successful linkage analysis. Locus heterogeneity (different genes causing the condition in different families) or irregular inheritance patterns are the main obstacles to success. The causative gene is identified by demonstrating that it has mutations in a panel of unrelated affected individuals.

Recombination is a normal part of every meiotic cell division. During prophase of meiosis I, pairs of homologous chromosomes synapse, and individual chromatids exchange segments at crossovers (see **Figures 2.14–2.16**). To a first approximation, crossovers are randomly distributed across the genome, although as shown in Chapter 12 (see **Figure 12.5**), at the molecular level they are concentrated in about 30,000 hotspots.

Recombinants are identified by genotyping parents and offspring for pairs of loci

Recombination is normally observed by genotyping offspring rather than gametes, although some researchers have typed individual sperm by ultrasensitive PCR. Thus in the context of human genetic mapping it is normal to speak of a *person* being **recombinant** or **nonrecombinant**. It is understood that we are really talking about one of the parental gametes that made the person. If there is any ambiguity about which parent's gamete is involved, it would be necessary to specify this. The proportion of gametes that are recombinant for two loci is the **recombination fraction** between the two loci.

In **Figure 17.3** individual II_1 is heterozygous at two loci (A and B). He has the genotype A_1A_2 B_1B_2. His alleles A_1 and B_1 came from his mother, and A_2 and B_2 from his father. Any of his sperm that carries one of these parental combinations (A_1B_1 or A_2B_2) is nonrecombinant for those two loci, whereas any that carries A_1B_2 or A_2B_1 is recombinant. We see that of his seven children, two were produced by recombinant sperm and five by nonrecombinant sperm. The recombination fraction is 0.28.

The recombination fraction is a measure of the genetic distance between two loci

If two loci are on different chromosomes, they will assort independently. On average, 50% of gametes will be recombinant and 50% nonrecombinant for any pair of loci that are on different chromosomes. The expected recombination fraction is 0.5. When two loci are **syntenic**—that is, they lie on the same chromosome—they will always travel together

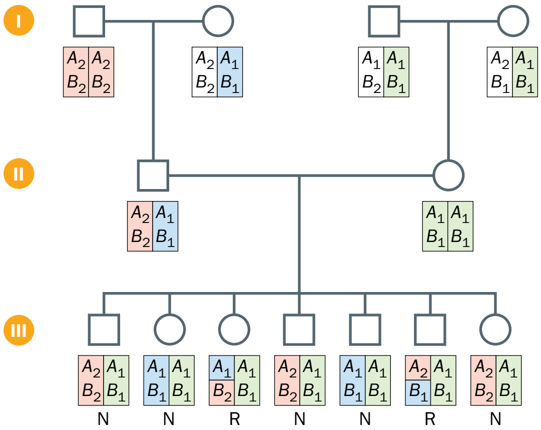


Figure 17.3 Recombinants and nonrecombinants. In this family, there are two loci (*A* and *B*) at which alleles (A_1 and A_2, B_1 and B_2) are segregating. Colored boxes mark combinations of alleles that can be tracked through the pedigree. In generation III, we can distinguish people who received nonrecombinant (N; A_1B_1 or A_2B_2) or recombinant (R; A_1B_2 or A_2B_1) sperm from their father (II_1). Their mother (II_2) is homozygous at these two loci, and so we cannot identify which individuals in generation III developed from nonrecombinant or recombinant oocytes.

unless separated by recombination. Only a crossover located in the region between the two loci will separate them. There will be very few recombinants between loci that lie close together on the chromosome, whereas loci far apart will recombine freely, since there are normally crossovers on every chromosome arm. Thus the recombination fraction between two loci is a measure of their distance apart on the chromosome. This is the **genetic distance**. It is measured in centiMorgans (cM, the uppercase M honors the pioneering geneticist TH Morgan), as distinct from the **physical distance**, measured in kilobases or megabases of DNA. Pairs of loci that recombine 1% of the time are said to be 1 cM apart. Genetic distances and physical distances do not precisely correspond. The order of loci on genetic and physical maps should be the same, but there are recombination hotspots, where a small physical distance can translate to a large genetic distance, and vice versa.

Recombination fractions never exceed 0.5. If loci are far apart on a chromosome there may be two or more crossovers separating them, but as Figure 17.4 shows, not more than 50% (on average) of the gametes produced are recombinant for those two loci. If there are ten loci *A*, *B*, *C*...*J* in that order along a chromosome, and each pair are 10 cM apart, loci *A* and *J* will still show no more than 50% recombination. If we made a genetic map showing the 10 loci, the overall distance between *A* and *J* would be 100 cM, but actually if we assessed the distance between *A* and *J* directly in pedigrees, it would be 50 cM.

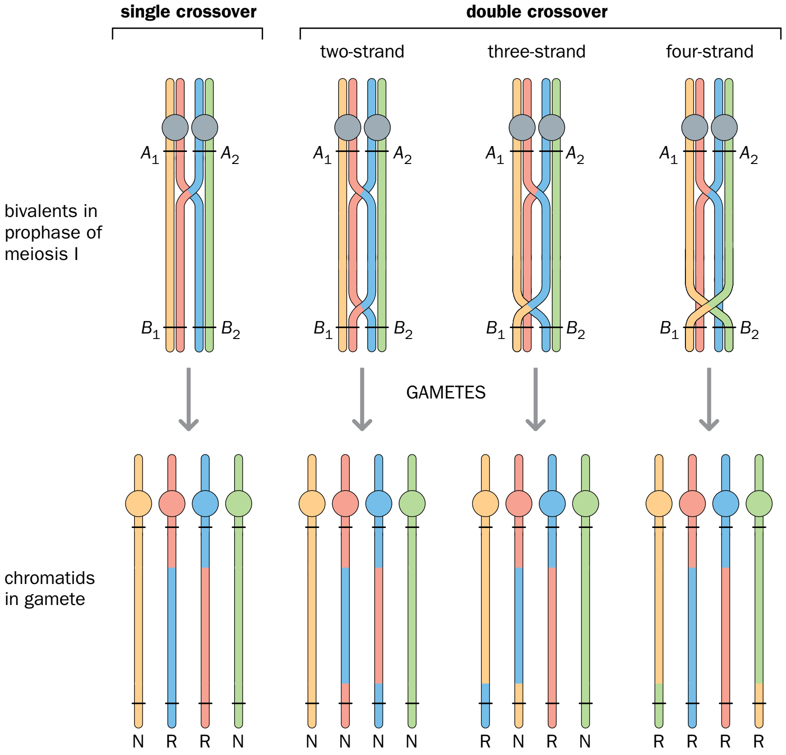


Figure 17.4 Single and double crossovers. The figure shows a pair of homologous chromosomes, one carrying alleles A_1 and B_1, the other alleles A_2 and B_2 at two loci. Each chromosome consists of two sister chromatids. Only two of the four chromatids are involved in any particular crossover. Chromatids in the gametes labeled N carry a parental combination of alleles (A_1B_1 or A_2B_2). Chromatids labeled R carry a recombinant combination (A_1B_2 or A_2B_1). Note that recombinant and nonrecombinant are defined only in relation to these two loci. For example, in the result of the three-strand double crossover, the second chromatid from the left has been involved in crossovers—but it is nonrecombinant for loci *A* and *B* because it carries alleles A_1 and B_1, a parental combination. A single crossover generates two recombinant and two nonrecombinant chromatids (50% recombinants). The three types of double crossover occur in random proportions, so the average effect of a double crossover is to give 50% recombinants.

Thus genetic distances are not additive but are related by a **map function**. Readers interested in that topic should consult the book by Ott and colleagues (see Further Reading).

Recognizing recombinants in human pedigrees is not always straightforward

It was simple to see in **Figure 17.3** that individual II_1 had two recombinant and five nonrecombinant children, but it is not always so simple. Only pedigrees with ideal structures (three or more generations and appropriate samples and clinical data available from everybody) allow direct interpretation. But researchers seeking to map a rare disease must take families as they find them. Consider the three pedigrees in **Figure 17.5**. Here we are supposing the two loci are the locus determining a rare autosomal dominant disease and a microsatellite having six alleles A_1–A_6.

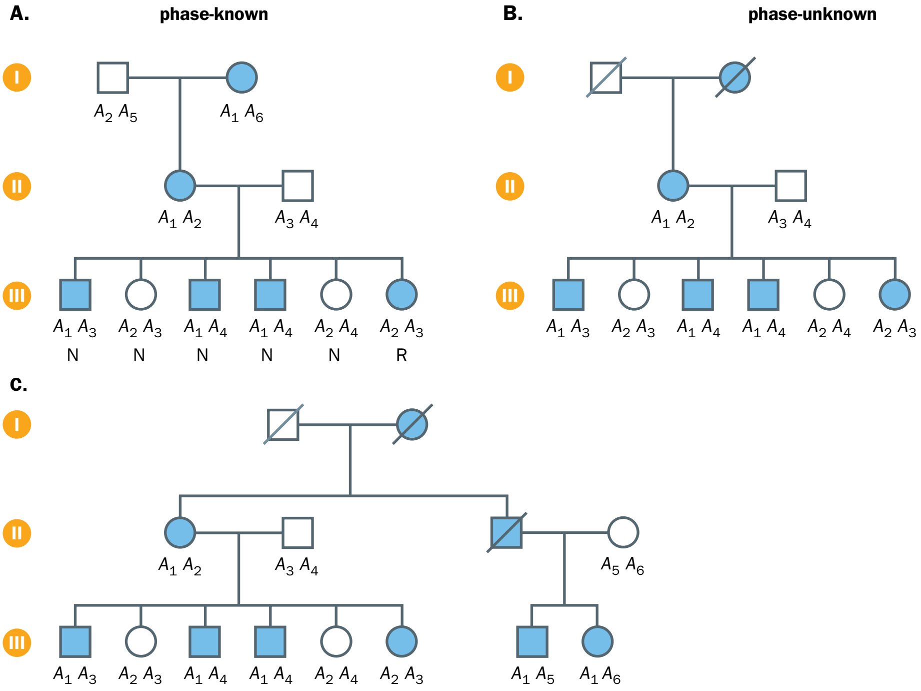


Figure 17.5 Recognizing recombinants. Three versions of a family with an autosomal dominant disease, typed for a microsatellite marker with alleles A_1–A_6. (**A**) All meioses are phase-known. We can identify III_1–III_5 unambiguously as nonrecombinant (N) and III_6 as recombinant (R). (**B**) The same family, but phase-unknown. The mother, II_1, could have inherited either marker allele A_1 or A_2 with the disease; thus her phase is unknown. Either III_1–III_5 are nonrecombinant and III_6 is recombinant, *or* III_1–III_5 are recombinant and III_6 is nonrecombinant. (**C**) The same family after further tracing of relatives. III_7 and III_8 have also inherited marker allele A_1 along with the disease from their father—but we cannot be sure whether their father's allele A_1 is identical by descent to the allele A_1 in his sister II_1. Maybe there are two copies of allele A_1 among the four grandparental marker alleles. The likelihood of this depends on the gene frequency of allele A_1. Thus, although this pedigree contains extra linkage information compared to pedigree B, extracting it is problematic.

- In Pedigree A the interpretation is straightforward, following the logic of **Figure 17.3**.
- In Pedigree B the doubly heterozygous woman II_1 is **phase-unknown**. Among her children, either there are five nonrecombinants and one recombinant, or else there are five recombinants and one nonrecombinant. We can no longer identify recombinants unambiguously, even if the first alternative seems much more likely than the second.
- Pedigree C adds yet more complications. Further relatives have been traced who have also inherited marker allele A_1 along with the disease from their father II_3—but we cannot be sure whether this A_1 allele is identical by descent to the A_1 allele in their aunt II_1. (See **Figure 12.10** for the distinction between alleles identical by descent and identical by state.)

Some method is needed to extract the linkage information from families like this with only incomplete or ambiguous identification of recombinants and nonrecombinants. The answer lies in computer-generated lod scores, as described below.

Mapping human diseases relies on genome-wide panels of genetic markers

Genetic mapping in humans is in principle identical to mapping in *Drosophila* flies or any other sexually reproducing diploid organism. However, there are two practical differences. First, as we have just seen, we must rely on pedigrees whose nonideal structure often makes interpretation difficult, rather than on nice, clean breeding experiments. And second, whereas in *Drosophila* we might map an eye-color variant against a wing-shape variant, in humans we cannot generally map diseases against other diseases. Human monogenic diseases are rare, and families in which two such diseases are segregating are doubly rare. Even if we could find one, it would probably be too small to generate a significant amount of data.

To map a human disease (or the locus determining any other uncommon phenotype), we need to collect families in which the character of interest is segregating, and then find some other Mendelian character that is also segregating in these same families, so that individuals can be scored as recombinant or nonrecombinant. A suitable **genetic marker** would have the following characteristics:

- It should show a clean pattern of Mendelian inheritance, preferably co-dominant so that the genotype can always be inferred from the phenotype;
- It should be scored easily and cheaply using readily available material (for example, a mouthwash rather than a brain biopsy);
- The locus that determines the character should be highly polymorphic, so that a randomly selected person has a good chance of being heterozygous;
- It should be one of a panel of hundreds of such markers spread at known chromosomal locations across the entire genome.

Early attempts at human mapping used protein variants such as blood groups and tissue types. A landmark paper by Botstein and colleagues in 1980 (PMID 6247908; see Further Reading) pointed out the potential of DNA variants for human gene mapping. The original DNA variants were **restriction fragment length polymorphisms** (RFLPs): variation in the size of fragment produced by digestion of a person's genomic DNA with a restriction enzyme because of a sequence variant that creates or abolishes a restriction site, see **Figure 7.4A**. From the mid-1990s onward, human linkage analysis used microsatellites PCR-amplified in multiplexes using fluorescently labeled primers, so that the products could be genotyped on a capillary electrophoresis machine. More recently, single nucleotide polymorphism (SNP) arrays have been used to genotype each family member for 500,000 SNPs in a single operation. Whole-genome sequencing data could also be used.

The raw data are interpreted by computer programs that produce a lod score

Having collected families where a Mendelian disease is segregating, and genotyped them with a suitable marker, how do we know when we have found linkage? There are two aspects to this question:

- How can we work out the recombination fraction?
- What statistical test should we use to see whether the recombination fraction is significantly different from 0.5, the value expected on the null hypothesis of no linkage?

In the pedigree shown in **Figure 17.5C**, it was not possible to identify recombinants unambiguously and count them. It is possible, however, to calculate the overall likelihood of the pedigree, on the alternative assumptions that the loci are linked (recombination fraction = θ) or not linked (recombination fraction = 0.5). The ratio of these two likelihoods gives the odds of linkage, and the logarithm of the odds is the **lod score**. Lod scores are symbolized as ***Z***. In 1955 Newton Morton demonstrated that lod scores represent the most efficient statistic for evaluating pedigrees for linkage, and he derived formulae to give the lod score (as a function of the recombination fraction θ) for various standard pedigree structures. **Box 17.1** shows how this is done for simple structures such as those in **Figure 17.5A** and **B**. Except in such simple cases, human linkage analysis is entirely dependent on computer programs that implement algorithms for handling branching trees of genotype probabilities, given the pedigree data and a table of gene frequencies. Ott *et al*. (2015) (PMID 25824869; see Further Reading) describe the various linkage programs that can be used, and give Web addresses for them.

BOX 17.1 CALCULATION OF LOD SCORES (Z)

The overall likelihood of a pedigree is calculated on two alternative hypotheses:

- Given that two loci are truly linked, with recombination fraction θ, the likelihood of a meiosis being recombinant is θ and the likelihood of its being nonrecombinant is $1 - \theta$.
- If the loci are, in fact, unlinked, the likelihood of a meiosis being either recombinant or nonrecombinant is 0.5.

For the family in Figure 17.5A there are five nonrecombinants $(1 - \theta)$ and one recombinant (θ).

- The overall likelihood, given linkage, is $(1 - \theta)^5 \times \theta$.
- The likelihood, given no linkage, is $(0.5)^6$.
- The likelihood ratio is $(1 - \theta)^5 \times \theta/(0.5)^6$.
- The lod score is the logarithm (base 10) of this likelihood ratio.

θ	0	0.1	0.2	0.3	0.4	0.5
Z	$-\infty$	0.577	0.623	0.509	0.299	0

For the family in Figure 17.5B, the mother (II_1) is phase-unknown. If she inherited A_1 with the disease, there are five nonrecombinants and one recombinant. If she inherited A_2 with the disease, there are five recombinants and one nonrecombinant.

The overall likelihood ratio is ½ $[(1 - \theta)^5 \times \theta/(0.5)^6]$ + ½ $[\theta^5 \times (1 - \theta)/(0.5)^6]$.

This allows for either possible phase, with equal prior probability.

θ	0	0.1	0.2	0.3	0.4	0.5
Z	$-\infty$	0.276	0.323	0.222	0.076	0

For the family in Figure 17.5C, to calculate the likelihood that III_7 and III_8 are recombinant or nonrecombinant, we must take likelihoods calculated for each possible genotype of I_1, I_2, and II_3, weighted by the probability of that genotype. For I_1 and I_2, the genotype probabilities depend on both the gene frequencies and the observed genotypes of II_1, III_7, and III_8. Genotype probabilities for II_3 are then calculated by simple Mendelian rules. This is all impossibly complicated for a hand calculation. For any but the simplest families, nonmasochists turn to the computer.

Being a function of the recombination fraction, the lod score is calculated for a range of θ values. The results can be plotted to give curves of lod score versus recombination fraction. The most likely recombination fraction is the one at which the lod score is highest. If there are no recombinants, the lod score will be maximum at $\theta = 0$. If there are recombinants, Z will peak at the most likely recombination fraction (0.167 = 1/6 for family A in **Figure 17.5**, but harder to predict for family B without more detailed calculations).

In a set of families, the overall probability of linkage is the product of the probabilities in each individual family. Lod scores, being logarithms, can be added up across families. Thus the result of linkage analysis is a table of lod scores at various recombination fractions, maybe with the scores summed across a whole collection of families, and with separate rows for a whole series of markers. Positive lod scores give evidence in favor of linkage, and negative lods give evidence against linkage. Note that only recombination fractions between 0 and 0.5 are meaningful, and that all lod scores are zero at $\theta = 0.5$ because they are then measuring the ratio of two identical probabilities, and $\log_{10}(1) = 0$.

Lod scores of +3 and –2 are the criteria for linkage and exclusion (for a single test)

The second question we posed concerned the threshold of statistical significance. Here the answer is at first sight surprising: $Z = 3.0$ is the threshold for accepting linkage, with a 5% chance of a type 1 error (falsely rejecting the null hypothesis). For most statistics, $p < 0.05$ is used as the threshold of significance, but $Z = 3.0$ corresponds to 1000:1 odds $[\log_{10}(1000) = 3.0]$. The reason why such a stringent threshold is chosen lies in the inherent improbability that two loci, chosen at random, should be linked. With 22 pairs of autosomes to choose from, it is not likely the two loci would be located on the same chromosome, and even if they were, loci well separated on a chromosome segregate independently. Common sense tells us that if something is inherently improbable, we require strong evidence to convince us that it is true. This common sense can be quantified in a Bayesian calculation (**Box 17.2**), which shows that odds of 1000:1 in fact correspond precisely to the conventional $p = 0.05$ threshold of significance. Thus, $Z = 3.0$ is the threshold for accepting linkage with a 5% chance of falsely rejecting the null hypothesis. Linkage can be rejected if $Z <–2.0$. Values of Z between –2 and +3 are inconclusive.

For whole-genome searches, a genome-wide threshold of significance must be used

In disease studies, families are usually genotyped for hundreds or thousands of markers, each of which is checked for evidence of linkage. The appropriate threshold for

BOX 17.2 BAYESIAN CALCULATION OF THE THRESHOLD OF SIGNIFICANCE FOR A TWO-POINT LOD SCORE

Bayesian statistics provide a way of combining probabilities derived from independent pieces of information about a problem, so as to arrive at an overall probability that takes all the information into account. The basic idea was first proposed by the Reverend Thomas Bayes, an eighteenth-century English clergyman, hence the name Bayesian. Students are often frightened by Bayesian calculations, but actually they are just quantitative common sense. If somebody asks you to believe something that has a very low prior probability (for example, that your friend was late for a lecture, not because he overslept but because he had been abducted by aliens) you will require very strong evidence to convince you. Bayesian statistics allow you to combine that prior probability with other relevant information.

The likelihood that two randomly selected loci should be linked (the prior probability of linkage) has been argued over, but estimates of about 1 in 50 are widely accepted. The Bayesian calculation combines this prior probability with **conditional likelihoods**. There are two competing hypotheses: that the loci are linked or not linked. The conditional likelihoods for the two hypotheses take the form:

- Given that the loci are linked, what is the likelihood of the observation that the lod score was 3?
- Given that the loci are not linked, what is the likelihood of the observation that the lod score was 3?

A lod score of 3 is telling us that the ratio of these two likelihoods is 1000:1. To combine the prior probability with this conditional likelihood in a Bayesian calculation we set up a little table as below and multiply down each column. The joint probability, taking into account both the 1 in 50 prior probability and the 1000:1 likelihood ratio, gives us 20:1 odds in favor of linkage.

BOX 17.2 TABLE 1

Hypothesis	Loci are linked (recombination fraction = θ)	Loci are not linked (recombination fraction = 0.5)
Prior probability	1/50	49/50
Conditional likelihood: 1000:1 odds of linkage [lod score Z(θ) = 3.0]	1000	1
Joint probability (prior × conditional)	20	1

We see that a lod score of 3 corresponds to 20:1 overall odds that the loci are linked; that is, the conventional $p = 0.05$ threshold of significance.

significance is a lod score such that there is only a 0.05 chance of a false positive result occurring anywhere during a search of the whole genome. As shown above, a lod score of 3.0 corresponds to a significance of 0.05 at a single point. But if 500 markers have been used, the chance of a spurious positive result is greater than if only one marker has been used. A stringent procedure (Bonferroni correction) would multiply the *p* value by the number of markers used before testing its significance. The threshold lod score for a study using *n* markers would be 3 + log(*n*); that is, a lod score of 4 for 10 markers, 5 for 100, and so on. However, this is overstringent. Linkage data are not independent. Given a Mendelian condition, the causative variant must be *somewhere* in the genome. If one location is excluded, then the prior probability that the character maps to another location is raised. The threshold for a genome-wide significance level of 0.05 has been argued over, but a widely accepted answer for Mendelian characters is 3.3. In practice, claims of linkage based on lod scores below 5, whether with one marker or many, should be regarded as provisional. In these days of whole-genome sequencing, lod scores well below the +3.0 threshold in small families can still be useful as they serve to point out genomic regions to prioritize when analyzing exome sequencing data.

Once a disease locus has been mapped further work is needed to identify the actual causative variant

Most of the more frequent autosomal dominant and X-linked conditions were successfully mapped during the period 1985–2000 using family studies as described above. Also, for some common autosomal recessive conditions such as cystic fibrosis, it was possible to find enough families with more than one affected child to allow the same process. Once a small-enough candidate chromosomal region was defined, genes in the region could be prioritized for Sanger sequencing according to their relevance to the phenotype under study. Having found a sequence variant within the candidate region, one needs some extra confirmation that it really is the cause of the family condition. Part of this can come from checking that the variant is in a gene that can be plausibly linked to the phenotype, and that the variant would be expected to affect expression or function of the gene. The fact that all affected people within the family carry the same mutation adds nothing—it merely confirms the result of the linkage analysis. Powerful confirmation would come from sequencing the candidate gene in a panel of unrelated affected individuals. Hopefully this would show that unrelated affected people all had mutations in that gene, but healthy controls did not.

This procedure worked best for loss-of-function conditions—as explained in Chapter 16, these conditions are usually marked by extensive allelic heterogeneity, with unrelated affected people having different loss-of-function variants in the same gene. Hopefully these would include some truncating variants (deletions, splice-site, nonsense, or frameshift mutations) where the loss of function is unambiguous, compared to missense changes that might just be benign polymorphisms. Conditions caused by a gain of function often show little or no allelic heterogeneity—compare, for example, the mutational spectra of ataxia-telangiectasia and Apert syndrome in **Figure 16.16**. It would be necessary to show a very strong association of the variant with the disease in a large panel of cases and controls. Additional support would come from the sorts of functional evidence described in Section 17.5. Nevertheless, as mentioned in Section 16.5, it has become apparent that these checks were often insufficiently rigorous. Now that we have exome sequences on vast numbers of healthy controls in the ExAC and GnomAD databases, it turns out that many variants labeled as pathogenic from these early studies are in fact present at similar frequencies in healthy controls. Probably in most cases the correct gene was identified, since many unrelated people all had variants in the same gene, but some fraction of the variants were wrongly supposed to be pathogenic—the real causative variant was somewhere else in the gene or even in another gene.

The classical linkage strategy worked well for many Mendelain conditions, but could not be applied to characters where large families with clear Mendelian patterns of inheritance could not be found

Positional cloning worked well when the family collection provided 20–30 meioses that could be used for mapping, and where all the families in the collection had mutations in the same gene. Frequent non-penetrance (see **Figure 5.11**) reduces the statistical power, because if an unaffected person has inherited the disease-associated marker allele this could be due to either recombination or non-penetrance. Other irregularities in the pattern of inheritance similarly complicate the research, while with complex, non-Mendelian conditions like diabetes or schizophrenia the whole procedure proved inapplicable and alternative approaches were necessary, as described in the following chapter. Locus heterogeneity is a serious problem—for example, it took years of work to show that tuberous sclerosis could be caused by mutations at either of two loci, *TSC1* (OMIM #191100) at 9q34 or *TSC2* (OMIM #191092) at 16p13. However, the major limitation with monogenic conditions was the requirement for a good collection of families. Some conditions were just too rare to make such a collection possible. For rare recessive conditions, autozygosity mapping as described in the next section offers a possible solution. But for rare dominant conditions, including many severe sporadic conditions that were suspected to be dominant and due to new mutations, no mapping approach is possible. Unless a patient with a chromosomal rearrangement provided a clue, those conditions remained completely intractable to research until the advent of next-generation sequencing, as described in Section 17.3.

17.2 HAPLOTYPE SHARING AND AUTOZYGOSITY

A search for recurrent ancestral haplotypes is the basis of several approaches to mapping diseases. As described in Chapter 18, this principle underlies the genome-wide association studies (GWAS) that have been widely used to identify genetic susceptibility factors for complex, non-Mendelian diseases. Those studies looked for very short haplotypes, the haplotype blocks defined by the HapMap project and described in Section 12.2. These may reflect shared ancestry in the remote past, thousands of years ago. Less remote shared ancestors can result in people sharing much longer haplotypes. Finding these offers a way to map rare monogenic recessive conditions.

Autozygosity is homozygosity for sequences identical by descent

If a person affected by a rare recessive condition is the product of a consanguineous marriage, it is likely that both copies of the causative variant are derived from a recent common ancestor of his or her parents. **Table 12.5** showed that the rarer a recessive condition is, the more likely it is that both alleles are identical by descent, inherited from the same common ancestor (the distinction between identity by descent [IBD] and identity by state [IBS] was made in Chapter 12; see **Figure 12.10**). If the two alleles of the causative gene are indeed identical by descent, the person will probably be homozygous not just for the causative variant but for all markers in a chromosomal segment surrounding it. Thus identifying segments in a patient's genome that are homozygous and identical by descent (**autozygous**) identifies candidate locations for the disease gene. This is the basis of **autozygosity mapping**.

To prove conclusively that two copies of a sequence are identical by descent it would be necessary to sequence or genotype every connecting person linking the two copies to the presumed common ancestor. However, the rarer a sequence is in the population, the less likely it is that two identical copies would have originated in two separate individuals. Extending the analysis in Section 12.4, suppose a person who is the product of a consanguineous marriage is homozygous for an allele or haplotype with population frequency q. His coefficient of inbreeding (the chance that at any given locus he shares two alleles identical by descent) is F. The odds his homozygosity is IBD rather than just IBS are F:q. Likely values of F are 1/16, 1/64, and 1/256 for the offspring of a first-, second-, or third-cousin marriage, respectively. For alleles at a single locus, q may never be low enough to make the odds ratio conclusive, but shared long, multilocus haplotypes allow much greater confidence. Although we saw in Section 12.2 that the haplotype blocks identified by the HapMap project are quite likely to be shared by ostensibly unrelated individuals, those blocks are only a few kilobases long. Megabase-size blocks are extremely unlikely to be shared except through a recent common ancestor.

Mapping a rare recessive condition using just three affected individuals

A single inbred, affected individual would probably not allow unambiguous localization of a disease gene, but when several members of an extended inbred kinship are all affected by the same rare recessive disease, the power of the analysis is much greater. The potential of autozygosity for disease mapping was first demonstrated by Houwen and colleagues in 1994 (PMID 7894490; see Further Reading). In an isolated Dutch fishing village, three individuals were affected by a rare recessive condition, benign recurrent intrahepatic cholestasis (OMIM #243300). The three were not known to be related, but given the history and degree of isolation of the village, it was likely that they shared a common ancestor perhaps six generations ago.

The three cases and their parents were typed for 256 microsatellite markers. At fourteen chromosomal locations spread across the genome there was homozygosity for two adjacent markers, but typing additional markers within each of those segments showed that there was only one region where a full, detailed haplotype was shared: the other regions showed IBS for the two initial microsatellites but not for all the additional markers, and therefore were not IBD. The only region truly IBD was a 19 cM segment of chromosome 18q21 (**Figure 17.6**). The authors calculated that the probability of finding at least five of six segments of that length showing IBD by chance, rather than because they carried a disease allele and were found in affected individuals, was 5×10^{-7}. At the time this work was done, the markers had been mapped against each other, so their genetic distances were known, but their physical positions on the human genome sequence were not known. Hence the size of the shared haplotype was given in centiMorgans, not megabases. In subsequent work consanguineous families from other parts of the world with the same disease were studied. Within each family there was a shared 18q21 haplotype, but it was different between the different unrelated families. The overlap between the shared haplotypes in the different families defined a candidate region of only 1 cM. This work eventually led to identification of the gene responsible for the condition.

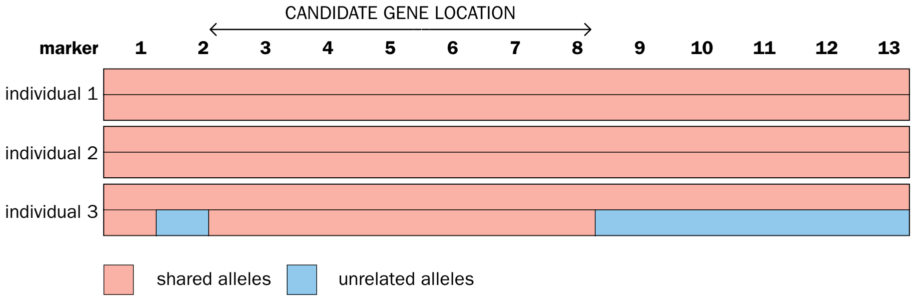


Figure 17.6 Autozygosity mapping of the locus for benign recurrent intrahepatic cholestasis. Three distantly related individuals with the same autosomal recessive condition and from the same isolated Dutch village, together with their parents, were genotyped for a panel of microsatellite markers. Between them, the three affected individuals have six copies of the chromosomal segment that carries the pathogenic variant. The figure summarizes the results from 13 markers spanning 19 cM on chromosome 18. Five of the six haplotypes were identical across this region, strongly suggesting identity by descent. In the sixth haplotype, markers 3–8 had genotypes identical to the shared haplotype, but the alleles at markers 1, 2, and 9–13 evidently had an independent origin. The pathogenic variant most likely lies in the region of chromosome 18 that includes markers 3–8. (Data from Houwen RH *et al*. [1994] *Nat Genet* **8**:380–386; PMID 7894490.)

Demonstrating the common ancestral origin of mutations

A bolder application of the same principle was used a few years later to fine-map the mutation causing the autosomal recessive DNA repair defect, Nijmegen breakage syndrome (NBS; OMIM #251260). Standard linkage analysis had localized the gene to an 8 Mb region on chromosome 8q21, but the available collection of patients had no recombinants across this region that would have allowed finer mapping. With the techniques available at the time, sequencing every gene in that 8 Mb region would have been a daunting undertaking. Acting on the hypothesis that many unrelated patients might share an ancestral mutation, Varon and colleagues (PMID 9590180; see Further Reading) genotyped 51 apparently unrelated patients and their parents for a series of microsatellite markers spaced across the 8 Mb candidate region. This generated 102 haplotypes of chromosomes that carried an NBS disease mutation. Of these, 74 looked like derivatives of a common ancestral haplotype, most probably of Slav origin. The most highly conserved region lay between markers 11 and 12 (**Figure 17.7**), which therefore marked the likely location of the *NBS* gene. The common ancestor of these patients must have lived many generations ago. Over the generations, meiotic recombination had chopped down the ancestral haplotype, so the shared haplotype was much smaller than in the more closely related patients studied by Houwen and colleagues (see **Figure 17.6**). Subsequently, a gene encoding a novel protein was cloned from this location and was shown to carry mutations in NBS patients. As predicted, patients with the common haplotype all had the same mutation, whereas those with independent haplotypes had independent mutations.

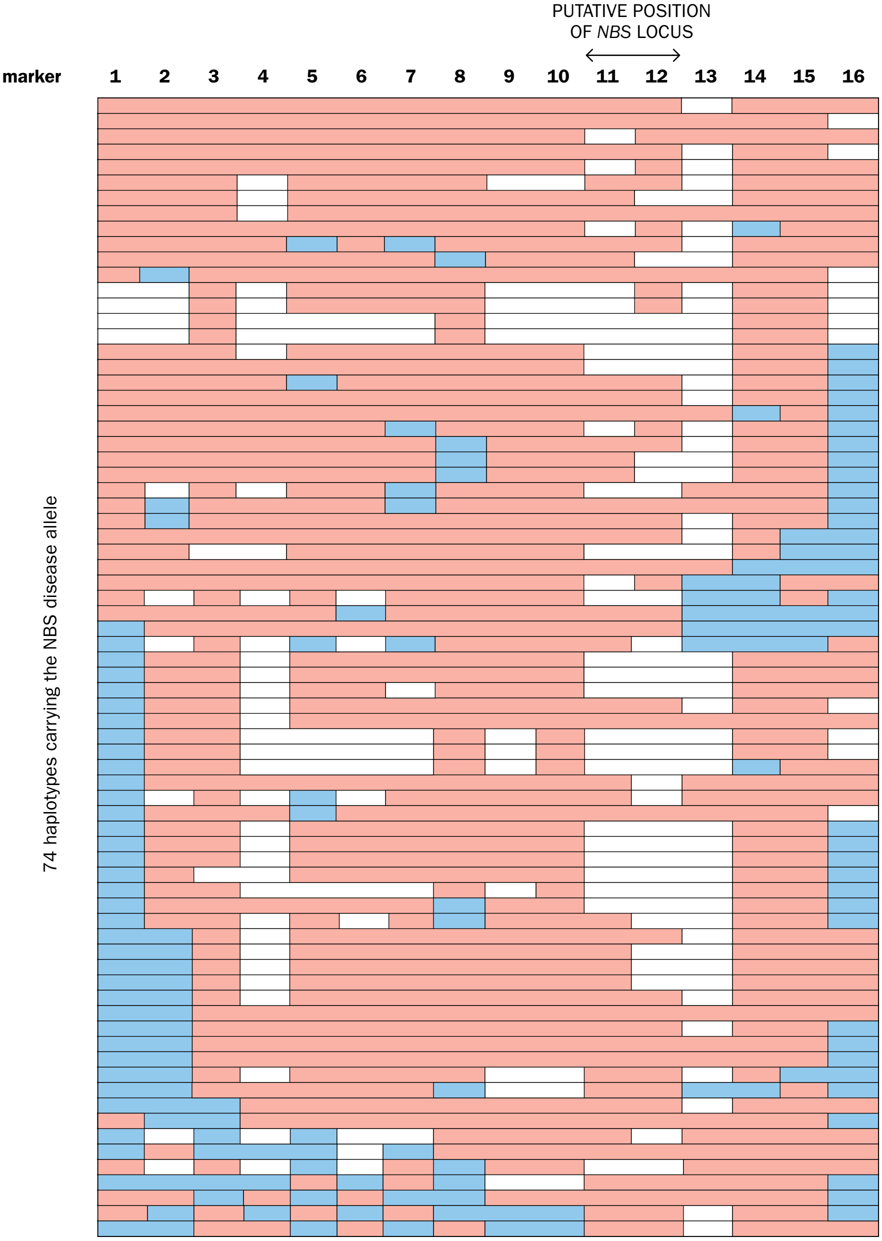


Figure 17.7 An ancestral haplotype in European patients with Nijmegen breakage syndrome. In a set of 51 apparently unrelated patients with this recessive condition, 74 out of 102 haplotypes around the disease location on chromosome 8q21 seemed to be derived from a common ancestor. The 74 haplotypes (rows) were defined using 16 markers, shown in chromosomal order across the top of the table. As in **Figure 17.6**, the pink color marks locations with alleles identical to those of the inferred ancestral haplotype. Alleles colored blue differ from the putative ancestral version. Blanks mark loci for which there are no data. Only at loci 11 and 12 are there no recombinant (blue) alleles, suggesting that the *NBS* gene maps to this position. (Data from Varon R *et al*. [1998] *Cell* **93**:467–476; PMID 9590180.)

Turning this work round, haplotype sharing can be used to investigate the history of a mutation. If two individuals, maybe living far apart, have the identical sequence variant, one might ask whether this is because they share a common ancestor or whether this is a case of two independent mutations. Checking for a shared haplotype can answer the question. In population genetics, shared haplotypes can be used to follow the spread of a mutation. The example of lactose tolerance was described in Chapter 14.

Autozygosity mapping using SNP arrays

Modern autozygosity mapping uses SNP arrays or whole-genome sequence data. The question asked is not usually whether a specific haplotype is shared, but whether different inbred individuals with the same recessive condition share a region of homozygosity. Unrelated individuals might be homozygous for different alleles, but the region where they are homozygous should be the same. The resolution is highest when unrelated individuals are studied, though this carries the risk that different genes might be mutated in the different families, even though they all have the same clinical condition. Figure 17.8 shows a typical example. Six individuals from four unrelated, consanguineous families were studied. Within a family, affected people share similar large stretches of homozygosity. In the figure, tracks 2 and 3 are from two people from the same family, as are tracks 5 and 6. But between families there is only a small overlap (compare tracks 1, 2, 4, and 6), which provided a precise localization of the causative gene, and allowed it to be identified.

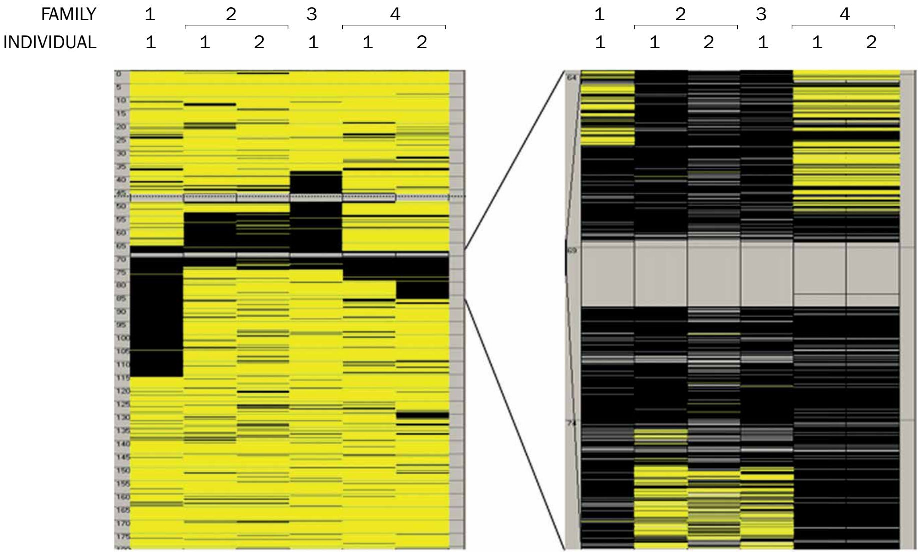


Figure 17.8 Autozygosity mapping of an autosomal recessive condition using SNP arrays. SNPs from across the genome were genotyped in six affected individuals from four unrelated, consanguineous families with the rare autosomal recessive neurological disorder, band-like calcification with simplified gyration and polymicrogyria. The six vertical tracks in the figure show the results for chromosome 5 in the indicated six individuals. Heterozygous genotypes are shown in yellow, homozygous in black, at the appropriate chromosomal location. The scale on the left shows the distance in megabases from the tip of the short arm. No SNPs were available in the grayed-out areas. A 6.5 Mb region of shared homozygosity was identified at position 5q13. (From O'Driscoll MC *et al.* [2010] *Am J Hum Genet* **87**:354–364; PMID 20727516. With permission from Elsevier.)

17.3 WHOLE-EXOME AND WHOLE-GENOME SEQUENCING ALLOW AN UNBIASED AND HYPOTHESIS-FREE APPROACH TO IDENTIFYING THE CAUSE OF A MONOGENIC CONDITION

The first massively-parallel "next-generation" sequencing machine was released onto the market in 2005 (see Section 6.5). Soon the first few individuals had their genomes sequenced using the new technology—James Watson in 2008, followed later in the same year by anonymous West African and Han Chinese individuals. However, these were expensive prestige projects and beyond the means of most researchers or laboratories. It was the development of efficient exome-capture methods that allowed next-generation sequencing to move into smaller laboratories. Whole exome sequences could be generated for costs, in time and money, that were acceptable for a relatively modest laboratory, and the subsequent bioinformatic analysis, though still far from trivial, was somewhat less daunting than searching a whole genome sequence for a relevant variant. With the continuing fall in the cost of sequencing, whole-genome sequencing is beginning to look feasible for routine gene identification, but most applications to date have used exomes rather than genomes.

In 2009 Ng and colleagues (PMID 19684571) reported how they had captured and sequenced twelve human exomes and had been able to detect previously identified pathogenic variants in them. This provided the proof of principle that exome sequencing could be used to identify disease genes. Not long afterward the same group (PMID 19915526) used exome sequencing to identify the gene mutated in Miller syndrome (OMIM #263750). Figure 17.9 shows how rapidly exome sequencing took off as the

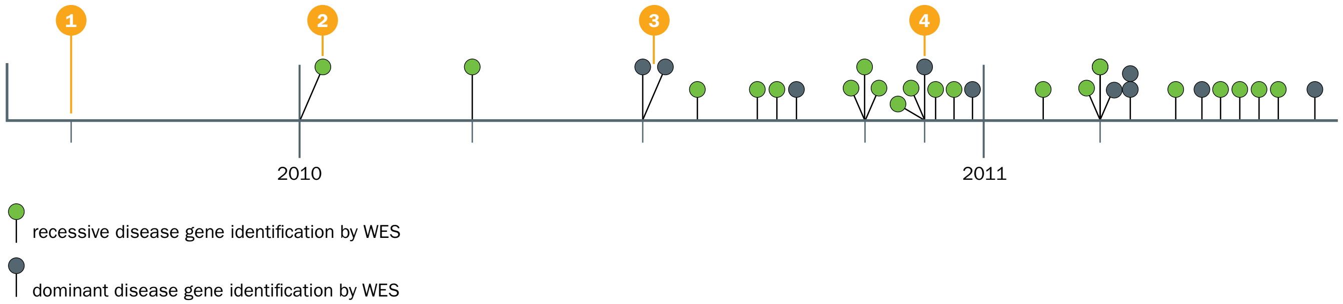


Figure 17.9 The rapid early development of exome sequencing for gene identification. Highlighted landmarks along the timeline are: (1) Ng SB *et al*. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**:272–276; PMID 19684571 (proof of principle that exomes can be captured and sequenced to identify pathogenic variants). (2) Ng SB *et al*. (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**:30–35; PMID 19915526 (identification of the cause of the recessive condition Miller syndrome). (3) Hoischen A *et al*. (2010) *De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat Genet* **42**:483–485; PMID 20436468 (identification of *de novo* gain-of-function mutations causing a sporadic dominant condition). (4) Vissers LE *et al*. (2010) A *de novo* paradigm for mental retardation. *Nat Genet* **42**:1109–1112; PMID 21076407 (evidence that sporadic nonsyndromic intellectual disability is often the result of a highly heterogeneous set of *de novo* mutations). WES, whole-exome sequencing. (Adapted from Gilissen C *et al*. [2011] *Genome Biol* **12**:228; PMID 21920049. With permission from BioMed Central Ltd.)

default way of identifying disease genes. The following few years saw a flood of gene identifications. All over the world, laboratories were retrieving their archived DNA samples from the freezer, sequencing exomes, and identifying causative variants. In the year 2012 alone over 130 novel, rare, disease-causing genes were reported. Progress particularly focused on the rare recessive and sporadic dominant conditions that had been refractory to previous gene-identification methods.

Exome sequencing relies on commercial exon-capture kits

Next-generation sequencing machines sequence the whole of whatever collection of DNA fragments is loaded into the machine. Exome sequencing therefore requires a way of producing a collection of fragments that between them cover just the 180,000 or so protein-coding exons in the human genome. These total about 33 Mb. Allowing for some flanking intronic sequence, and maybe some noncoding RNAs and other sequences known to harbor pathogenic mutations, an exome sequencing library might comprise 40–60 Mb of DNA, around 1–2% of the human genome. These sequences are selected by hybridization to a library of oligonucleotide probes.

Early exome-capture systems used oligonucleotides anchored on microarrays, but these have been superseded by solution capture, which is more efficient and requires less input DNA. Various companies such as Agilent, Illumina, and Nimblegen market solution capture kits, and each announce frequent upgrades. The probes are typically around 100 nucleotides long, and may be either DNA or RNA. The general procedure is shown in **Figure 17.10**.

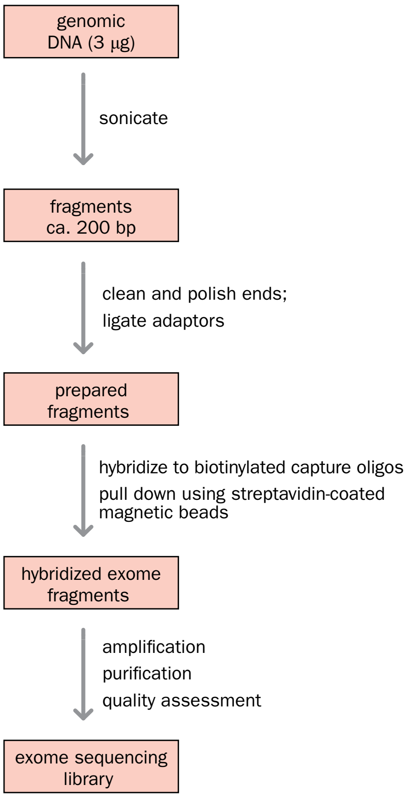


Figure 17.10 Workflow for exome sequencing. Genomic DNA is fragmented to an average size of 200 bp (human exons average 145 bp). After processing the ends to produce fragments with a 5′ phosphate and 3′ dA overhang, adaptor oligonucleotides are ligated to allow PCR amplification with a single primer pair (see **Figure 6.10**). The adaptors may also include a barcode (an "index") to allow samples to be multiplexed for sequencing. Before the capture step, the fragments may be size-selected and subjected to a few rounds of PCR amplification to enrich for fragments correctly tagged with adaptors at each end. The capture probes are biotinylated, so that after hybridization they can be pulled down using streptavidin-coated magnetic beads. After release from the capture probes, the exome fragments are purified and amplified ready to be loaded into the sequencing machine.

Identifying true variants from next-generation sequencing data is far from trivial

In Section 6.5 we described the main next-generation sequencing (NGS) technologies, and illustrated how their raw output consisted of millions of short reads that needed to be aligned against the human reference sequence (see **Figure 6.20**). This process and all subsequent stages in the analysis are extremely computationally intensive and require substantial bioinformatic expertise to perform them correctly. **Figure 17.11** gives an overview of the steps necessary; interested readers should consult the paper by DePristo and colleagues (PMID 21478889; see Further Reading) for technical detail.

All the different high-throughput DNA sequencing machines have error rates that are high compared to Sanger sequencing. Random errors can be controlled by sequencing a given sample to a sufficient depth. Each system also generates its own pattern of systematic errors, which are much more difficult to correct (that is the "base quality recalibration" step in the workflow of **Figure 17.11**). Sequence alignment is particularly difficult for repetitive sequences and insertions or deletions (indels). Reads containing indels align poorly and risk being rejected by quality filters, thus indels are typically underestimated in NGS data that use short reads.

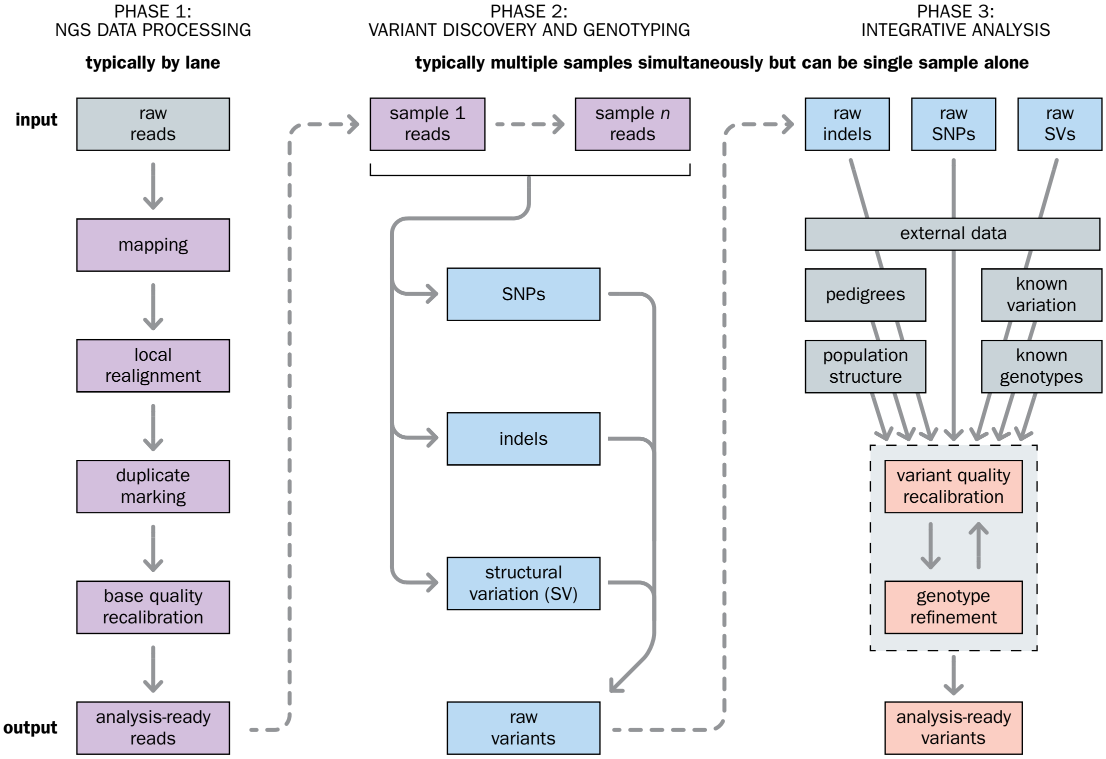


Figure 17.11 Overview of the bioinformatic work necessary to identify true variants in next-generation sequencing data. (Reprinted from DePristo MA *et al*. [2011] *Nat Genet* **43**:491–498; PMID 21478889. With permission from Springer Nature. Copyright © 2011.)

Exome capture introduces extra pitfalls in identifying variants

All exons are not equally represented in sequencing libraries prepared by exon capture. Some exons are captured and/or amplified more efficiently than others. Exons having a particularly high or particularly low GC content are usually underrepresented. The first exon of a gene is often poorly represented because first exons tend to have a high GC content. **Figure 17.12** shows an example: the average coverage of exons of this gene is over 40×, but coverage is far from uniform and some areas have only 10× or less coverage. In order to search across a patient's whole exome for an unknown pathogenic variant, the overall coverage must be set at a level that gives adequate coverage of poorly covered areas. For example, in one reported set of exomes the median coverage was 42×, but 18% of sequences were covered at less than 10×. In another series with average coverage of 100×, 97–98% of sequences were covered at 10× or better and 93–96% at 20× or more. For clinical exome screening, an overall coverage of 80× is often recommended. In some cases, for example where there is a suspicion of mosaicism, much deeper coverage is required, such as 1000×.

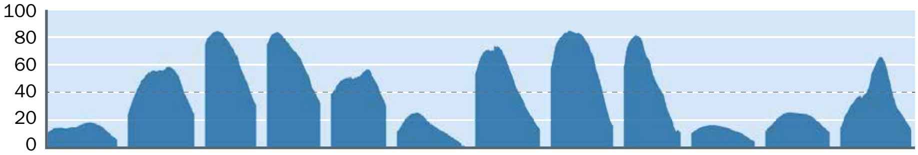


Figure 17.12 Variable coverage of exons of a gene in exome sequencing. Data for the 12 exons of the *PCSK9* gene downloaded from the ExAC server (exac.broadinstitute.org/) for transcript ENST00000302118. The vertical axis shows the coverage. Average coverage is 40.83× (dashed line), but some whole exons, and parts of others, are much more poorly represented in the data.

Much technical development is directed at reducing this inequality. Even so, there will probably always remain some sequences that fail the quality filters for analysis.

One radical answer to the problem of unequal coverage is to abandon exon capture and opt for sequencing the whole genome. The bioinformatic analysis can still be limited to the exome, because the aim is to detect exonic variants more efficiently, not to see the many, largely uninterpretable noncoding variants. A study by Gilissen and colleagues in 2014 (PMID 24896178; see Further Reading) illustrated the value of this approach. The Nijmegen group had undertaken a large study of patients with severe intellectual disability. Testing for copy number variants had identified the cause in some cases, and exome sequencing had answered the question in others. But for some cases no answer had been found. The whole genomes of 50 such patients and their unaffected parents were sequenced. Sequencing was performed to 80× depth, twice the depth usually considered for clinical whole-genome screening. Eighty-four *de novo* coding mutations were identified, of which 65 had been missed in the preceding exome sequencing because of poor coverage. Positive diagnoses were achieved for 21 of the 50: 13 had *de novo* dominant single nucleotide variants, 7 had *de novo* copy number variants, and one was a compound heterozygote for a recessive condition. All of these might in principle have been detected by the previous microarray and exome analysis, but despite the extensive experience of this group of workers, they had been missed. This study highlights the superior power of genome sequencing to detect exonic variants. As genome sequencing becomes cheaper this may become a more frequent approach.

17.4 STRATEGIES FOR EXOME-BASED DISEASE-GENE IDENTIFICATION

Sequencing a patient's exome will typically reveal around 20,000 variants relative to the Reference Human Sequence. Whole-genome sequencing would reveal 4 million. If the patient has a Mendelian condition, just one of those variants should be responsible for their condition. How can we get from a long list of variants to the one causative one? Experience suggests that most cases of Mendelian conditions are caused by missense, nonsense, frameshift, or splice-site variants within the coding sequence of protein-coding genes. There are always exceptions, especially among variants causing missplicing, which may be in unexpected places, but the initial search would focus on variants predicted to alter the amino acid sequence of a protein. Thus, once the various quality-control filters have produced a list of true variants from the raw sequence data, further filtering is likely to involve the following steps, though not necessarily in this order:

1. Select only variants in protein-coding sequence, rejecting variants in untranslated regions and in introns unless they immediately flank a splice site;
2. Select only missense, nonsense, and predicted splicing variants or small indels;
3. Variants causing rare conditions should themselves be rare. Therefore, for dominant conditions, reject variants present in exome databases in individuals who do not have any phenotype related to that of the patient. These would include the dbSNP, 1000 Genomes, ExAC, or GnomAD databases, plus in-house databases. For recessive conditions, reject variants in those databases that occur at frequencies too high to be plausible causes of the condition under investigation, given its rarity;
4. Reject variants predicted to be nonpathogenic by programs such as PolyPhen-2 or SIFT (see **Box 16.2**);
5. For recessive conditions, select genes where affected people have two likely pathogenic changes; for dominant conditions, select genes where affected people have at least one likely pathogenic change;
6. For sporadic conditions thought likely to be dominant, select *de novo* changes whenever DNA from both parents is available for sequencing.

In clinical service work, applying these filters to the exome sequence of a single patient may suffice to establish a diagnosis, if among a small number of remaining candidate genes there is one that is known to cause the relevant condition. For identifying novel disease genes, a single exome is unlikely to provide a definitive answer. It will be necessary to combine data from more than one person. Depending on the condition and mode of inheritance, various strategies are available to select suitable people to sequence (**Figure 17.13**).

The linkage approach in **Figure 17.13A** would point to a candidate region that might contain a dozen or more genes, which must then be searched for candidate variants. The other strategies shown in **Figure 17.13** would directly point to a particular gene. To show how it works out in practice, we will describe some real examples from pioneering applications of exome sequencing.

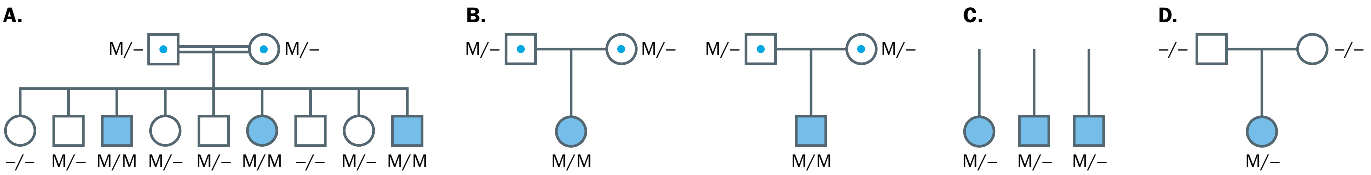


Figure 17.13 Possible strategies for using exome or genome sequencing to identify a disease gene. (**A**) Identifying a candidate gene for a recessive condition by homozygosity mapping in a large family. (**B**) Seeking a gene where unrelated patients with the same recessive condition all have homozygous or compound heterozygous mutations. (**C**) Seeking a gene where unrelated patients with the same dominant condition all have heterozygous mutations. (**D**) Seeking a gene where a patient with a *de novo* dominant condition has a *de novo* mutation.

Identifying the gene mutated in Miller syndrome, an autosomal recessive condition

In this pioneering study Ng and colleagues (PMID 19915526) sequenced the exomes of four individuals with the rare condition Miller syndrome (OMIM #263750). Affected individuals have a recognizable pattern of congenital malformations, including severe micrognathia, cleft lip and/or palate, hypoplasia or aplasia of the postaxial elements of the limbs, coloboma of the eyelids, and supernumerary nipples. Two of the four cases were sibs, the other two were from separate families. There was some uncertainty whether Miller syndrome was dominant or recessive: of the 30 well-characterized reported cases there were only three multiplex families, each consisting of two affected siblings born to unaffected, nonconsanguineous parents. Thus, although on balance of probabilities the condition was thought likely to be recessive, it might be dominant, with the familial cases being due to gonadal mosaicism in one parent. The data were therefore analyzed on both hypotheses.

An average coverage of 40× was achieved over 26.6 Mb of sequence captured using a microarray. Table 17.1 shows how the list of variants was filtered. Note how the filters work much more strongly for the hypothesized recessive condition because of the requirement that each individual should have two mutations in the causative gene.

TABLE 17.1 FILTERING THE LIST OF VARIANTS FOUND ON SEQUENCING EXOMES OF FOUR INDIVIDUALS WITH MILLER SYNDROME

	Kindred 1A	Kindred 1B	Kindred 1 A+B	Kindreds 1+2	Kindreds 1+2+3
HYPOTHESIS 1: MILLER SYNDROME IS DOMINANT					
NS/SS/I	4670	4687	3940	3099	2654
Not in dbSNP 129	641	647	369	105	63
Not in HapMap 8	898	923	506	117	38
Not in either	456	464	228	26	8
Predicted damaging	204	204	83	5	2
HYPOTHESIS 2: MILLER SYNDROME IS RECESSIVE					
NS/SS/I	2863	2859	2362	1810	1525
Not in dbSNP 129	102	114	53	25	21
Not in HapMap 8	123	128	46	7	4
Not in either	31	33	9	1	1
Predicted damaging	6	12	1	0	0

Two individuals were sibs (Kindred 1 A and B), the other two were from separate families (Kindreds 2 and 3). On the dominant hypothesis, each individual should have a single damaging mutation in the causative gene; on the recessive hypothesis, each should have two (but could be compound heterozygotes). NS/SS/I, nonsynonymous, splice-site, or indel variants; HapMap 8, exomes of eight individuals from the HapMap project; Predicted damaging, prediction using PolyPhen. See text for discussion. (Data from Ng SB *et al*. [2010] *Nat Genet* **42**:30–35; PMID 19915526.)

There were two surprises in the analysis.

- First, no gene carried two predicted damaging mutations in all four individuals, apparently ruling out the recessive hypothesis. One gene, *DHODH* encoding dihydroorotate dehydrogenase, did carry two mutations in each case. All cases were compound heterozygotes; the sibs from Kindred 1 had the same two mutations. Five of the six variants were predicted to be damaging. The suspicion was that PolyPhen had misclassified the sixth mutation, p.G202A, as benign when it was in fact damaging. Later studies showed that although both SIFT and PolyPhen predict this mutation as tolerated or benign, other prediction programs (LRT, MutationTaster, PhyloP, and GERP++) correctly predict it as damaging. The identity of *DHODH* as the Miller syndrome gene was confirmed by Sanger sequencing of the gene in three further kindreds and finding two predicted damaging mutations in each case.
- The second surprise was finding that both sibs in Kindred 1 had mutations in both alleles of the *DNAH5* gene, encoding the dynein heavy chain. *DNAH5* mutations are a well-documented cause of primary ciliary dyskinesia (OMIM #608644), a condition characterized by recurrent lung infections and chronic lung disease. Reviewing the two sibs, it was apparent that they had this condition on top of their Miller syndrome.

Compared to more recent studies, the exomes sequenced here were less inclusive than most current examples (26.6 versus 40–60 Mb) and the available normal exomes for filtering were much more limited.

Identifying the gene mutated in Schinzel–Giedion syndrome, a sporadic condition

Schinzel–Giedion syndrome (OMIM #269150) is a highly recognizable syndrome characterized by severe intellectual disability, distinctive facial features, and multiple congenital malformations. As with Miller syndrome, most cases are sporadic but at least two familial cases have been reported, again raising the question whether this is a recessive condition or a dominant condition with rare parental mosaicism. Hoischen and colleagues (PMID 20436468) sequenced exomes of four sporadic cases and analyzed their data on the assumption that Schinzel–Giedion syndrome is dominant. The severity of the condition would rule out affected people becoming parents, so all cases are expected to be due to new mutations.

In this study, 37 Mb of sequence was selected by exome capture and sequenced to an average depth of 43×. On average 85% of the exome was covered at least to 10×. The analysis proceeded as in **Table 17.2**.

TABLE 17.2 VARIANTS FOUND BY SEQUENCING EXOMES OF FOUR UNRELATED INDIVIDUALS WITH SCHINZEL–GIEDION SYNDROME

	Case 1	Case 2	Case 3	Case 4	Variants in all 4
Total variants	22,916	22,602	22,152	19,528	4735
NS/SS	5556	5618	5427	4802	1634
Not in dbSNP 130	405	401	390	387	35
Not in in-house database	299	289	275	288	12

NS/SS, nonsynonymous or splice-site (± 2 nucleotides) variants. (Data from Hoischen A *et al.* [2010] *Nat Genet* **42**:483–485; PMID 20436468.)

Compared to a recessive condition, the filtering is much less stringent because cases are required to have only a single variant in a candidate gene. Thus 12 genes survived the process (note that the reported filtering scheme did not attempt to assess the potential pathogenicity of any of the variants at this stage). For 10 of the 12 genes, all four individuals had the same change, suggesting that these 10 changes were in fact SNPs that had not at that time been added to dbSNP. Nowadays variant databases are much better developed and this would be less of a problem, except perhaps in some very obscure ethnic group. One of the remaining two genes, *CTBP2*, was known from previous exome-sequencing experiments to show numerous variants, maybe due to highly homologous sequences from other genomic loci, and so was excluded from further consideration. This left just the *SETBP1* gene. As with Miller syndrome, confirmation came from identifying *SETBP1* mutations in a further eight individuals with Schinzel–Giedion syndrome. For 10 of the 12 cases, parental DNA was available, and in each case the mutation was shown to be *de novo*.

Identifying the gene mutated in Kabuki syndrome, a heterogeneous dominant condition

Kabuki syndrome (OMIM #147920) is characterized by a distinctive facial appearance, cardiac and skeletal abnormalities, immunological defects, and mild to moderate intellectual disability. The condition is rare but recognizable, and several hundred cases have been reported worldwide, mostly sporadic but with a few cases of parent–child transmission. It was assumed to be dominant, occasionally transmitted by a mildly affected parent but mostly maintained by recurrent mutation.

Ng and colleagues (PMID 20711175) sequenced exomes (31 Mb) to an average coverage of 40× from 10 unrelated affected individuals. Data were analyzed on the dominant model by seeking a gene where all 10 individuals carried a rare and presumed damaging variant. No such gene was found. One gene, *MUC16*, showed rare variants in all 10 cases, but the encoded protein, mucin 16, is very large with frequent variants, and was an unlikely candidate for a developmental disorder. Suspecting locus heterogeneity, the researchers looked for variants present in some subset of the 10 patients (**Table 17.3**).

TABLE 17.3 FILTERING EXOME DATA FOR 10 PATIENTS WITH KABUKI SYNDROME

	1 of 10	2 of 10	3 of 10	4 of 10	5 of 10	6 of 10	7 of 10	8 of 10	9 of 10	All 10
NS/SS/I	12,042	8722	7084	6049	5289	4581	3940	3244	2486	1459
Rare*	6935	2227	701	242	104	44	16	6	3	1
LoF**	753	49	7	3	2	2	1	0	0	0

The figures show the number of genes harboring variants of each type. No gene harbored a likely rare, pathogenic variant in all 10 cases; therefore the researchers looked for genes containing a suitable variant in some proportion of cases (any 1 of the 10, any 2 of the 10, and so on). NS/SS/I, nonsynonymous, splice-site, or indel variants. * Rare meant not found in dbSNP 129, 1000 Genomes, or 26 in-house exomes. ** LoF indicates nonsense, frameshifting indel, or splice-site variants that, unlike missense variants, are confidently predicted to cause a loss of function. See text for discussion. (Data from Ng SB *et al*. [2010] *Nat Genet* **42**:790–793; PMID 20711175.)

One gene, *KMT2D* (also known as *MLL2*), had likely loss-of-function variants in seven of the ten cases. It was a good candidate for a developmental disorder because it encodes a lysine methyltransferase that modifies histones as part of the epigenetic control mechanisms described in Chapter 10. Sanger sequencing confirmed the seven mutations and identified mutations in two more of the ten cases that had been missed by the next-generation sequencing. No mutation was found in the tenth case. The conclusion is that Kabuki syndrome is heterogeneous, with the majority of cases caused by mutations in *KMT2D*, but some cases having other causes. Significant locus heterogeneity is an obstacle to gene identification by the strategies described so far. The researchers were lucky because later analysis of 116 patients identified *KMT2D* mutations in only 74. A few patients have mutations in the related *KDM6A* gene. The next case, where extensive locus heterogeneity is expected, shows an alternative approach.

Gene identification in highly heterogeneous conditions relies on detecting *de novo* changes

Severe intellectual disability (ID) is common and almost always sporadic. A few percent of cases are caused by gross chromosomal abnormalities (for example, Down syndrome), while large-scale *de novo* structural variants identified by comparative genomic hybridization explain a further 10–20% of cases. In countries with high levels of consanguineous marriage, autosomal recessive causes are frequent, and the causative variants may be identified by autozygosity mapping. In countries with low levels of inbreeding, recessive cases are likely to be infrequent; the condition would more likely be dominant and maintained by recurrent mutation since the biological fitness of affected persons is negligible. Probably the high frequency of new mutations reflects the fact that the mutational target is very large—that is, mutations in any of a large number of genes can cause severe ID. This makes intuitive sense: our brain is the most complex part of our body, it must rely on the correct functioning of many different components, and therefore there are many different ways it can go wrong.

The expected high degree of locus heterogeneity means that the approaches used in Schinzel–Giedion and Kabuki syndromes are unlikely to work. The key to identifying causative variants is to focus on *de novo* changes. As mentioned above, some of these are structural variants, but the majority are likely to be point mutations. An exome would typically contain 20,000 variants, and a genome 3–4 million, but only 0–4 and 50–100, respectively, are likely to be *de novo*. These can be identified by sequencing parent–child trios. Many apparent *de novo* changes will turn out to be sequencing errors, so true *de novo* changes must be confirmed by Sanger sequencing of the relevant exon in the affected individual and both parents. After the usual filters have been applied to eliminate variants unlikely to affect protein function, at most only one or two genes usually remain in the frame. In an early proof-of-principle experiment, Vissers and colleagues (PMID 26503795) sequenced 10 parent–child trios where the proband had moderate to severe ID and a negative family history. After filtering, nine *de novo* nonsynonymous variants in seven individuals were confirmed. Functional data (see below) suggested that three of these were unlikely to be causative of ID, but the other six remained strong candidates.

The study of Vissers and colleagues showed the power of trio analysis for conditions like ID where a high proportion of cases are likely due to *de novo* mutations. Conditions that are common in the population but associated with low biological fitness, such as schizophrenia, autism-spectrum disorders, and male infertility would be other candidates for this approach. However, once a candidate gene has been identified, it remains necessary to prove that the variant really does cause the condition. This is a particular problem for conditions that are genetically highly heterogeneous but clinically homogeneous, like many cases of severe ID. For recognizable syndromes like Miller, Schinzel–Giedion, or Kabuki syndromes, showing that all or most of a good number of affected individuals have mutations in the candidate gene is powerful evidence. For conditions like ID this route is not open. In the next section we look at alternative approaches.

17.5 CONFIRMING THAT THE CANDIDATE GENE IS THE CORRECT ONE

With clinically distinctive syndromes, showing that a panel of affected cases have mutations in the same gene provides powerful confirmation that the correct gene has been identified. That approach is not available when the condition in question is exceedingly rare or, like intellectual disability, likely to be extremely heterogeneous. In those cases, functional evidence becomes crucial. Having defined a likely candidate gene, biochemical, functional, and phenotypic evidence must be adduced to show why mutations in that gene should produce the phenotype in question. That second stage is always important, regardless how a candidate gene was identified, but it is particularly important for genes defined by identifying *de novo* variants in patients with conditions that are not clinically distinctive. The available lines of evidence can be divided into using existing datasets and generating new data.

Existing datasets can give a great deal of information about a candidate gene

It needs no more than an internet-enabled computer to access a remarkable amount of information to help judge whether a gene is a promising candidate for a disease under study. Topics on which one might hope to find information include:

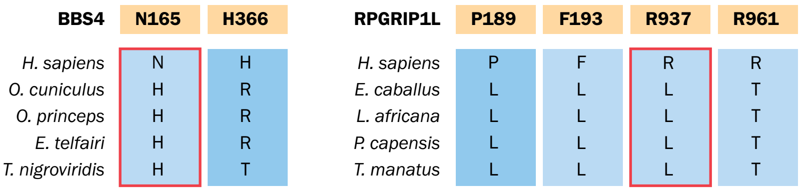
- Whether the gene has already been implicated in any disease or phenotype;
- Whether structural variants encompassing the gene have been described, and if so, whether that was in healthy subjects or persons with some pathological condition;
- What variants have been reported in the gene, and at what frequencies;
- Where in the body and when during development the gene is expressed;
- What splice isoforms of the gene occur;
- The likely biochemical function of the gene product;
- Other proteins with which the gene product interacts;
- Aspects of the way expression of the gene is regulated;
- What orthologs and paralogs of the gene are known;

- The extent of sequence conservation among orthologs and paralogs, in particular conservation of amino acids affected by the candidate disease variant(s);
- Possibly data on the effect of mutation, knock-down, or overexpression of orthologs in other species.

Precedent, rarity, and conservation are probably the three most useful initial pointers to variants causing Mendelian disease.

- The ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) collects information on possibly pathogenic variants. Data are collected from individual laboratories or extracted from other databases. As of July 2018 it contained 676,579 records. ClinVar can be searched for a specific variant, for all variants in a given gene, for variants at a given genomic location, or for variants associated with a disease or phenotype. As well as listing variants and any claimed associated phenotype, it includes an assessment of the likely pathogenicity and links to relevant publications and many other database features. Where the same variant has been reported by more than one submitter, ClinVar puts the two entries side-by-side so that users can check for agreement or review alternative interpretations.
- The Genome Aggregation Database (http://gnomad.broadinstitute.org/), the successor to the widely used ExAC database, can be used to check the frequency of a putatively pathogenic variant in healthy controls. GnomAD contains so many exomes (126,216 exomes plus 15,136 whole-genome sequences as of October 2017) and so many rare variants that it would be unrealistic to require that a candidate pathogenic variant should be totally absent from the database. Rarity should be defined in relation to the estimated frequency of the disease. A candidate variant for a dominant condition, if present at all in GnomAD, should be present at a much lower frequency than the population frequency of the condition. A candidate variant for a recessive condition should be present at lower frequency than the frequency of carriers, as estimated by the Hardy–Weinberg relationship (with allowance for inbreeding for a rare recessive condition; see Section 12.4).
- In the past, one would have checked a few hundred healthy individuals from the relevant population to make sure that the variant was not a common nonpathogenic polymorphism. A calculation in Chapter 5 (Section 5.3) can be used to put an upper limit on the population frequency of a variant if a panel of controls has been checked without finding any healthy person carrying the variant. Unless one is looking at a very unusual population, however, this approach has now been completely superseded by use of the ExAC and GnomAD databases. The main limitation on these immensely powerful resources is that subjects are not necessarily free of disease—the databases include data generated from schizophrenia cohorts, for example. Realistically, one might ask what is meant by a "healthy" subject. None of us is perfect. GnomAD output can be taken at face value for severe pediatric disease, because such cases have been excluded from the database, but the lists might need filtering if you are looking at late-onset or less severe conditions.
- Conservation is the basis of the PolyPhen and SIFT programs that are used to assess pathogenicity of missense variants, as well as the basis of a number of other programs (see **Box 16.2**). Nucleotide-level alternative programs include PhyloP, available as a track on the UCSC browser, and GERP or GERP++ (Genomic Evolutionary Rate Profiling). CADD (Combined Annotation-Dependent Depletion) and VEP (Variant Effect Predictor, available through Ensembl) integrate multiple data types.

Results from these tools should be treated as indicative only. One reason why conservation-based predictions give false negatives, labeling a damaging variant as benign, is **compensated pathogenic deviation**. **Figure 17.14** shows examples. Functional studies might confirm that a certain amino acid variant is pathogenic in the context of the human protein. But in certain animals the same variant is the normal, wild-type amino acid in the orthologous protein. Because of this, programs like PolyPhen and SIFT classify it as benign. But it is only benign in the context of proteins that contain particular amino acid changes at other positions. Polypeptide chains fold through interactions between different amino acid residues, and often function through such interactions, so there is nothing surprising in principle about this. The study by Jordan and colleagues (PMID 26123021; see Further Reading) estimated that at least 3% of human amino acid substitutions are subject to this effect in some other species. **Figure 12.9** showed a similar phenomenon in a noncoding RNA.



BBS4	N165	H366
H. sapiens	N	H
O. cuniculus	H	R
O. princeps	H	R
E. telfairi	H	R
T. nigroviridis	H	T

RPGRIP1L	P189	F193	R937	R961
H. sapiens	P	F	R	R
E. caballus	L	L	L	T
L. africana	L	L	L	T
P. capensis	L	L	L	T
T. manatus	L	L	L	T

Figure 17.14 Compensated pathogenic deviation. Functional assays confirmed that two human amino acid substitutions, p.N165H in the *BBS4* protein and p.R937L in the *RPGRIP1L* protein, are pathogenic, even though the replacement amino acid is the normal, wild-type amino acid in some other species. In those animals, amino acid substitutions elsewhere in the protein compensate for the pathogenic effect: R366 or T366 in BBS4; L189, L193, or T961 in RPGRIP1L. Programs like PolyPhen and SIFT wrongly classify p.N165H and p.R937L as benign. (Adapted from Jordan DM *et al.* [2015] *Nature* **524**:225–229; PMID 26123021. With permission from Springer Nature. Copyright © 2015.)

Information about the expression and interactions of a gene product can be very helpful. Most obviously, a candidate gene should be expressed in the cells, tissues, or organs affected by the disease. Thus, for example, in the Vissers study of intellectual disability, described above, two genes showing confirmed *de novo* mutations were discounted because they were not expressed in the central nervous system. In addition they had known functions that made it seem unlikely that any mutation should cause intellectual disability. Even if the specific functions of a gene product are not well described, information about genetic interactions can be informative. Its interactions can locate a gene in networks or pathways that suggest the likely sort of phenotype that variants might produce.

Relevant variants can be created and studied in cells in culture or in whole animals

In addition to these uses of existing data, targeted experiments may be necessary to explore the effect of a variant. As described in Section 16.1, transient transfection experiments can be used to check the effect of a variant in a promoter region, and minigene splicing assays can test for effects on splicing (see **Figure 16.3**). Where the protein has an observable biochemical role in cells, the functioning of the wild-type and variant proteins can be compared in cultured cells. CRISPR/Cas or other gene-editing technologies (Section 8.4) make it possible to create the variant in a cell, or induced pluripotent stem cells from a patient with the variant can be used, differentiated as necessary into the appropriate cell type.

If these cell-based approaches do not provide a clear answer, it may be necessary to re-create the variant in a model organism and observe its effect. Suitable model organisms include mice, zebrafish, and *Drosophila*. As an example, **Box 17.3** shows

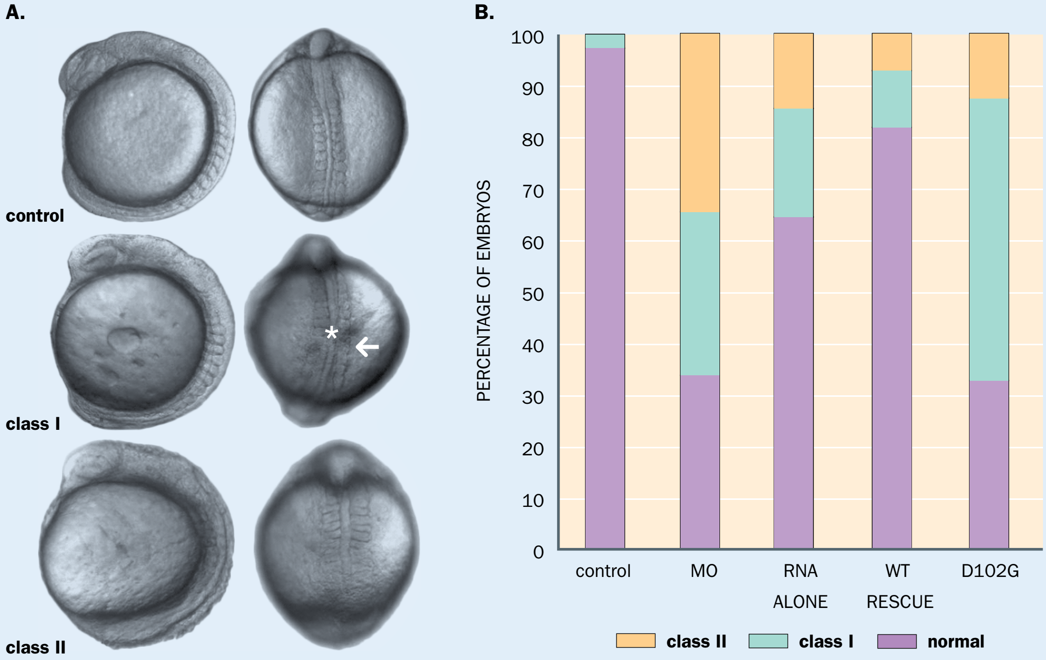
BOX 17.3 USING ZEBRAFISH FOR FUNCTIONAL ANALYSIS OF CILIOPATHY-ASSOCIATED VARIANTS

Ciliopathies are diseases caused by dysfunction of cilia. Cilia are present on almost all mammalian cells and have a variety of functions, not just in moving extracellular fluid but also in signaling (reviewed by Gerdes and colleagues, PMID 19345185; see Further Reading). Up to 1000 proteins may be involved in ciliary function, and there are many different Mendelian ciliopathies. One such is Bardet–Biedl syndrome (BBS; see OMIM #209900). BBS is an autosomal recessive condition characterized by retinitis pigmentosa, obesity, kidney dysfunction, polydactyly, behavioral dysfunction, and hypogonadism. Homozygous loss of function of any of 20 different genes has been seen in BBS patients. There has been some controversy whether BBS is often triallelic, with variants at more than one locus contributing to pathogenicity through an overall burden of reduced ciliary function, or whether it is a simple recessive and the extra variants found at secondary loci are merely coincidental and irrelevant. Thus there has been an interest in devising functional tests for pathogenicity of variants.

Here we describe a system using zebrafish that has been developed by the group of Nicholas Katsanis at Duke University (**Figure 1**). The test goes in three stages:

- First, an antisense morpholino oligonucleotide (see Section 8.5) is used to knock down expression of the relevant orthologous gene in 1–4-cell wild-type zebrafish embryos. The assay depends on the knockdown causing gastrulation defects that can include shortened body axes, longer somites, and broad and kinked notochords. These phenotypes are consistent with abnormal planar cell polarity (PCP) signaling, which likely underlies several clinical phenotypes in BBS patients. Provided the knockdown does produce these effects, one can move to the next stage;
- Next, wild-type embryos are co-injected with the morpholino oligonucleotide and the wild-type human mRNA of the relevant gene. Provided this rescues the phenotype, the next step can follow;
- Finally, embryos are co-injected with the morpholino oligonucleotide and human mRNA carrying the variant under investigation. If the phenotype is still rescued, the variant is judged nonpathogenic, but if the gastrulation defects remain, the variant lacks the function of the wild-type mRNA.

In some cases, injection of the mutant mRNA alone, without the morpholino oligonucleotide, causes relevant abnormalities. This suggests a dominant effect. In cases where that effect can be titrated out by co-injection of the wild-type mRNA, the variant is having a dominant-negative effect; if wild-type mRNA is unable to compensate, the variant has a pathogenic gain of function.



Box 17.3 Figure 1 A functional test for variants found in patients with Bardet–Biedl syndrome. (**A**) Images of 9-somite zebrafish embryos showing normal development and class I and class II abnormalities. The asterisk marks a kinked notochord, the arrow marks broadened somites. (**B**) Injection of a *BBS4*-blocking morpholino oligonucleotide (MO) causes a high frequency of abnormal embryos (second bar). The effect can be largely rescued by co-injecting wild-type (WT) human *BBS4* mRNA (fourth bar) but not mRNA carrying the p.D102G mutation (fifth bar). The p.D102G RNA therefore lacks the function of the wild-type mRNA. (Courtesy of Erica E Davies, Duke University.)

how zebrafish have been used to investigate whether variants seen in patients with Bardet–Biedl syndrome are pathogenic. The data shown in **Figure 17.14** were obtained using this system.

Of the most widely used model organisms, mice are the most likely to show phenotypes directly related to the corresponding human mutation. Nevertheless, despite the many similarities between mice and humans, one should not expect a perfect correspondence between orthologous mutant phenotypes in the two species. It is not uncommon when a human pathogenic mutation is reproduced in mice for there to be either no discernible phenotypic effect, or for it to be lethal. For example, in one project, knock-out mouse mutations were created in 37 genes where OMIM reports human recessive phenotypes. Seventeen of the homozygous mouse knock-outs were lethal. Mouse geneticists are well aware that phenotypes depend a great deal on genetic background. A given variant is often expressed differently in different strains of laboratory mice (see **Box 21.4**). The massive differences in background between mice and humans will have much greater effects. Within humans, the effects of differences in genetic background show up as variable penetrance and expressivity.

Systematic generation and phenotyping of a null mutant for every mouse gene is well under way through the International Mouse Phenotyping Consortium (IMPC; www.mousephenotype.org). Knock-outs may have more extreme phenotypic effects than loss-of-function point mutations, partly because the latter may not always produce a complete loss of function. Thus mouse knock-outs are useful pointers to the totality of possible phenotypic effects of loss of function, at least in mice, but only a subset may be seen in mice homozygous for point mutations. For gain-of-function variants it will usually be necessary to construct the exactly corresponding mouse mutant. In former times this was a lengthy and demanding job; some of the newer genome-editing tools, notably CRISPR/Cas genome editing (Section 8.4), make it considerably easier.

SUMMARY

- Genetic mapping is used to identify the chromosomal location of the variant responsible for a monogenic phenotype, in order to provide a shortlist of candidate genes.
- In traditional (lod score) mapping, a collection of pedigrees is checked for co-segregation of the relevant phenotype with the alleles of genetic markers. The main markers used are short tandem repeat polymorphisms (STRPs or microsatellites) or single nucleotide polymorphisms (SNPs). The statistical criterion for significance is the lod score.
- Traditional lod score mapping is now seldom used as a primary route to gene identification because almost all the Mendelian phenotypes for which suitably large pedigrees exist have already been mapped. It still has a place in helping reduce the extensive list of variants identified by exome or genome sequencing.
- Traditional lod score mapping has been very successful for mapping monogenic conditions, but mapping loci conferring susceptibility to common complex diseases requires analysis using different approaches, which are described in Chapter 18.
- Autozygosity mapping can be used for recessive conditions. SNPs are used to identify chromosome segments that are homozygous and identical by descent in a set of affected people from one or more consanguineous families.
- Whole-exome or whole-genome sequencing is widely used to search for mutated genes in individuals with sporadic or very rare conditions.
- Sporadic conditions are often the result of *de novo* dominant mutations, and these can be found by whole-exome sequencing of an affected individual and both parents.
- The raw data produced by next-generation sequencing machines need extensive bioinformatic processing, first to produce a list of variants that are truly present in the sample analyzed, and then to filter the long list of such variants to identify a shortlist of candidates for the condition in question.
- Variants identified by gene sequencing in affected individuals must be assessed for their likely pathogenicity and relevance to the phenotype.
- For variants in coding sequence, their likely effect on gene function can be predicted by various computer programs, for example PolyPhen-2 and SIFT at the protein level and VEP or splice-site prediction programs at the DNA level.
- A candidate variant should not be present in healthy controls at too high a frequency. For a recessive condition this would be the frequency of predicted carriers; for a dominant condition it should ideally be zero (if the disorder were highly penetrant) but in any case substantially less than the population frequency of the condition.
- The relevance of a candidate gene to a phenotype can be assessed using knowledge of the biochemical function, the timing and location of expression, and evidence of the phenotypes of persons or animals having mutations in related genes.
- A more definitive assessment of pathogenicity requires functional studies, in cell extracts, intact cells, or animal models.

FURTHER READING

Linkage mapping

Botstein D *et al.* (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**:314–331; PMID 6247908.

Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd edn. Johns Hopkins University Press.

Ott J *et al.* (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* **16**:275–284; PMID 25824869.

Autozygosity mapping

Houwen RHJ *et al.* (1994) Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet* **8**:380–386; PMID 7894490.

Varon R *et al.* (1998) Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. *Cell* **93**:467–476; PMID 9590180.

Gene identification by exome sequencing

DePristo MA *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**:491–498; PMID 21478889.

Gilissen C *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**:344–347; PMID 24896178.

Hoischen A *et al.* (2010) *De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat Genet* **42**:483–485; PMID 20436468.

Lelieveld SH *et al.* (2015) Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum Mutat* **36**:815–822; PMID 25973577.

Ng SB *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**:30–35; PMID 19915526.

Ng SB *et al.* (2010) Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat Genet* **42**:790–793; PMID 20711175.

Vissers LELM *et al.* (2016) Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* **17**:9–18; PMID 26503795.

Assessing conservation and pathogenicity of a variant

Adzhubei IA *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat Methods* **7**:248–249; PMID 20354512. (PolyPhen-2; http://genetics.bwh.harvard.edu/pph2)

Cooper GM *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**:901–913; PMID 15965027. (GERP)

Gerdes JM *et al.* (2009) The vertebrate primary cilium in development, homeostasis, and disease. Cell **137**:39–45; PMID 19345185.

Jordan DM *et al.* (2015) Identification of *cis*-suppression of human disease mutations by comparative genomics. *Nature* **524**:225–229; PMID 26123021.

Kircher M *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**:310–315; PMID 24487276. (CADD)

Kumar P *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**:1073–1082; PMID 19561590. (SIFT; http://sift.jcvi.org/)

Siepel A *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**:1034–1050; PMID 16024819. (PhyloP)

Functional tests of pathogenicity

White JK *et al.* (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154**:452–464; PMID 23870131.

Zaghloul NA *et al.* (2010) Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. *Proc Natl Acad Sci USA* **107**:10602–10607; PMID 20498079.

Complex disease: identifying susceptibility factors and understanding pathogenesis

18

INTRODUCTION

In Chapter 5 we contrasted the pedigree patterns shown by monogenic and multifactorial or complex characters and reviewed the polygenic-threshold theory that explains why non-Mendelian conditions can still run in families. The arguments about molecular pathology, set out in Chapter 16, apply to all genetic determinants, whether of Mendelian or complex conditions, but identifying those determinants requires different approaches in the two categories. Linkage analysis, as described in Section 17.1, achieved major successes with Mendelian conditions but has for the most part not been successful at identifying susceptibility factors for complex disease. A variant approach, model-free linkage, is described in Section 18.2. Until recently the cost of next-generation sequencing has precluded its use on a scale sufficient to identify susceptibility factors. This may be changing as costs continue to fall, but the main success story to date with complex or multifactorial conditions has been that of the genome-wide association studies described in Sections 18.3–4.

Students in elementary genetics courses risk ending up supposing that Mendel's patterns of inheritance describe the norm for how genes affect phenotypes. From this perspective, some characters might show irregularities like nonpenetrance, and some like diabetes simply don't follow the rules at all, but these are exceptions and not part of mainstream genetics. This is a complete inversion of the truth. In reality it is Mendelian characters that are the exceptions. It is true that most *actionable* variants, in a clinical context, are determinants of Mendelian or near-Mendelian characters, but it is absolutely not true that most genetic determinants are of this type. To show a Mendelian pattern of inheritance a character must depend entirely on the genotype at a single locus, regardless of genotypes at every other locus, and regardless of somebody's environment, history, or lifestyle. Not surprisingly, the great majority of genetically-influenced characters are not Mendelian. DNA sequence variants themselves are inherited in a clean Mendelian manner, but this is rarely true of their downstream consequences. The more steps there are on the pathway between a DNA sequence and an observable trait, the less likely is it that the trait will show simple Mendelian inheritance.

The counterpart of the belief that well-behaved phenotypes should be Mendelian is the popular fallacy that there is "the gene for" every character. Given that some inherited characters are entirely nongenetic (one's surname or mother tongue, for example), claims of genetic influence on a non-Mendelian character need supporting by evidence. This is not necessary for characters that show a clear Mendelian pattern of inheritance or a consistent chromosome structural variant, but it is necessary for non-Mendelian characters. A major thrust of genetic research is towards identifying the genetic variants that underlie common non-Mendelian differences between people, including health-related conditions like diabetes, obesity, heart disease, and mental health problems. Hopefully, knowledge of such factors will shed light on the pathogenesis of these conditions and suggest approaches for prevention or treatment. Genotyping a person for a set of variants might possibly allow a prediction of individual risk. But before we dive in to try and identify susceptibility factors, we need to look at the epidemiology of a condition to check that there is evidence for any involvement of genetic factors. This is the subject of Section 18.1.

18.1 INVESTIGATION OF COMPLEX DISEASE: EPIDEMIOLOGICAL APPROACHES

The first hurdle in an epidemiological approach is establishing the criteria by which people are to be labeled as affected. Monogenic conditions due to mutations in nuclear DNA are self-defined as phenotypes that show a Mendelian pedigree pattern, but for non-Mendelian phenotypes this is not an option. Modern studies typically involve large consortia of researchers, and it is particularly important to make sure everybody is using precisely the same diagnostic criteria—failure to do so can lead to spurious results in any subsequent genome-wide association study. Many conditions can present with widely differing degrees of severity, from nearly normal through to catastrophically severe. If people are nevertheless to be divided into normal and affected it is essential to define exactly where the border is to be drawn. Alternatively, it may be better to treat the condition as a continuous variable. This would be straightforward for a quantitative character like hypertension, but would require careful construction and validation of severity scores for a condition like diabetes.

Agreed diagnostic criteria are important for making different studies comparable, but they do not guarantee that a useful genetic distinction is being addressed. A diagnostic label can be valid, in the sense that independent investigators will agree whether or not it applies to a given patient, without being biologically meaningful. The problem is especially acute for psychiatric and behavioral conditions, which rely on labels and diagnostic criteria codified in the successive versions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) published by the American Psychiatric Association. As far as possible, phenotypes for genetic studies should be defined to emphasize gene action rather than an overall clinical appearance (‘endophenotypes’).

Once we have agreed clear diagnostic criteria for a condition, we then need to show whether or not genetics has a role in its etiology. If there is a genetic susceptibility, people who share more of their DNA should be more likely to share the phenotype under investigation. The obvious way to approach this is to show that the character runs in families. The examples below assume a condition is defined as dichotomous, and people are classified as affected or unaffected. Sometimes, as mentioned above, it may be more realistic to see a condition as a continuous quantitative character, with people having some continuously variable measure of affection status. In such cases the epidemiological evidence would be in the form of correlations.

The risk ratio (λ) is a measure of familial clustering

The degree of family clustering of a disease can be expressed by the **risk ratio** (λ_R), the risk to a relative (R) of an affected proband compared with the risk in the general population. A risk ratio of 1 implies no additional risk above that of the general population. Separate values can be calculated for each type of relative, for example λ_S for sibs. The mathematical properties of λ_R are derived in the 1990 papers by Risch (PMID 2301392–4, see Further Reading). As an example, **Table 18.1** shows pooled data from several studies of schizophrenia. Family clustering is evident from the raised λ values, for example a sevenfold increased risk for somebody, one of whose parents is schizophrenic. As expected, λ values drop back toward 1 for more distant relationships such as nephews, nieces, or cousins.

TABLE 18.1 RISK OF SCHIZOPHRENIA AMONG RELATIVES OF SCHIZOPHRENICS: POOLED RESULTS OF SEVERAL STUDIES

Affected relative	No. at risk	Risk (%)	λ
One parent	8020	5.6	7
One sib	9920.7	10.1	12.6
One sib and one parent	623.5	16.7	20.8
Half-sib	499.5	4.2	5.2
Uncle, aunt, nephew, niece	6386.5	2.8	3.5
Grandchild	739.5	3.7	4.6
First cousin	1600.5	2.4	3

Numbers at risk are corrected to allow for the fact that some at-risk relatives were below or only just within the age of risk for schizophrenia (say, 15–35 years), hence the noninteger values. λ values are calculated assuming a population incidence of 0.8%. (Data from McGuffin P, Shanks MF, Hodgson RJ (eds.) [1984] *The Scientific Principles of Psychopathology*. Grune & Stratton.)

Shared family environment is an alternative explanation for familial clustering

Geneticists must never forget that humans give their children their environment as well as their genes. Many characters run in families because of the shared family environment—whether one's native language is English or Chinese, for example. One always has to ask whether shared environment might be the explanation for familial clustering of a character. This is especially important for behavioral attributes such as IQ or schizophrenia, where the influence of upbringing might be significant. Even for physical characters or birth defects it cannot be totally ignored: the shared family lifestyle might include an unusual diet or some traditional medicine that could cause developmental defects. The difficulty of distinguishing the effects of shared family environment from those of heredity has often made studies controversial, especially for psychiatric conditions. Something more than a familial tendency is necessary to prove genetic susceptibility. Twin or adoption studies are the usual solution.

Twin studies suffer from many limitations

Francis Galton, the brilliant but eccentric cousin of Charles Darwin, who laid so much of the foundation of quantitative genetics, pointed out the value of twins for human genetics. Monozygotic (MZ) twins are genetically identical clones and should always be **concordant** (both the same) for any genetically-determined character. This is true regardless of the mode of inheritance or number of genes involved; the only exceptions are for characters dependent on post-zygotic genetic changes (the pattern of X-inactivation in females, the repertoire of functional immunoglobulin and T-cell receptor genes, and random post-zygotic somatic mutations leading to mosaicism). Dizygotic (DZ) twins share half their genes on average, the same as any pair of sibs. The heritability of a trait can be estimated as $h^2 = 2(r_{MZ} - r_{DZ})$ where r is the concordance.

Genetic characters should show a higher concordance in MZ than DZ twins, and many characters do. Polderman and colleagues presented a massive meta-analysis of 2,748 studies covering 17,804 traits from publications including 14,558,903 partially overlapping twin pairs (see Further Reading, PMID 25985137). The overall conclusion was that for most traits the MZ concordance was indeed higher than the DZ concordance. Such a higher concordance is necessary but not sufficient to prove a genetic effect. For a start, half of DZ twins are of different sexes, whereas all MZ twins are the same sex. Even if the comparison is restricted to same-sex DZ twins (as most studies are), at least for behavioral traits the argument can be made that MZ twins are more likely than DZ twins to look very similar, to be dressed and treated the same, and perhaps to create their own private shared environment.

Separated monozygotic twins appear to provide the ideal experimental design

Monozygotic twins separated at birth and brought up in entirely separate environments seem to provide the ideal solution to the problem. Francis Crick once made the tongue-in-cheek suggestion that one of each pair of twins born should be donated to science for this purpose. Such separations happened in the past more often than one might expect—the birth of twins was sometimes the last straw for overburdened parents. Fascinating television programs can be made about twins reunited after 40 years of separation, who discover they have similar jobs, wear similar clothes, and like the same music. As research material, however, separated twins have many drawbacks:

- Any conclusion is necessarily based on small numbers of arguably exceptional people;
- The separation was often not total—often the twins were separated some time after birth and were brought up by relatives;
- There is a bias of ascertainment—everybody wants to know about strikingly similar separated twins, but separated twins who are very different are not newsworthy;
- Research on separated twins cannot distinguish intrauterine environmental causes from genetic causes. This may be important, for example in studies of sexual orientation (the 'gay gene'), where it has been suggested that maternal hormones may affect a fetus *in utero* so as to influence its future sexual orientation.

Thus, for all their anecdotal fascination, separated twins have contributed relatively little to human genetic research.

Adoption studies are the gold standard for disentangling genetic and environmental factors

If separating twins is an impractical way of disentangling heredity from family environment, studying adopted people is much more promising. Two study designs are possible:

- Find adopted people who suffer from a particular disease known to run in families and ask whether it runs in their biological family or their adoptive family;
- Find affected parents whose children have been adopted away from the family and ask whether being adopted saved the children from the family disease.

A celebrated but controversial study by Kety & Rosenthal (PMID 5570994, see Further Reading) used the first of these designs to test for genetic factors in schizophrenia. The diagnostic criteria used in this study have been criticized, and there have also been claims (disputed) that not all diagnoses were made truly blind. However, an independent re-analysis using DSM-III diagnostic criteria reached substantially the same conclusion: it was the genes rather than the family environment that increased the risk for the offspring. **Table 18.2** shows the results of a later extension of this study.

TABLE 18.2 AN ADOPTION STUDY IN SCHIZOPHRENIA

	Schizophrenia cases among biological relatives	Schizophrenia cases among adoptive relatives
Cases (47 chronic schizophrenic adoptees)	44/279 (15.8%)	2/111 (1.8%)
Controls (47 nonschizophrenic adoptees)	5/234 (2.1%)	2/117 (1.7%)

The study involved 14,427 adopted persons aged 20–40 years in Denmark; 47 of them were diagnosed as chronic schizophrenic. The 47 were matched for age, sex, social status of adoptive family, and number of years in institutional care before adoption with 47 nonschizophrenic control subjects from the same set of adoptees. (Data from Kety SS, Wender PH, Jacobsen B *et al.* [1994] *Arch Gen Psychiatry* **51**:442–455; PMID 8192547.)

The main obstacle in adoption studies is lack of information about the biological family, frequently made worse by the undesirability of approaching them with questions. Efficient adoption registers exist in only a few countries. A secondary problem is selective placement, in which the adoption agency, in the interests of the child, chooses a family likely to resemble the biological family.

Adoption studies are unquestionably the gold standard for checking how far a character is genetically determined, but because they are so difficult, they have in the main been performed only for psychiatric conditions, for which the nature–nurture arguments are particularly contentious.

Figure 18.1 summarizes the various approaches that can be used to assess the role of genetic factors in causation of a condition, and the problems with each.

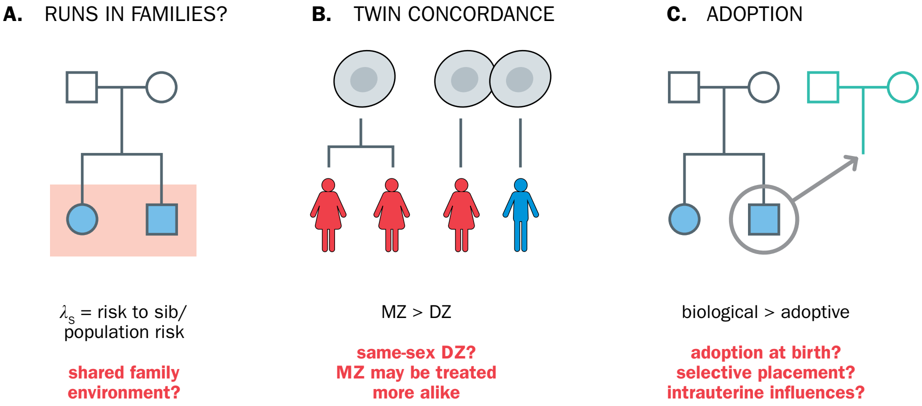


Figure 18.1 Ways of deciding whether genetic factors play a role in a condition. Red text shows the problems in each approach.

Segregation analysis attempts to use epidemiological data to understand the genetic architecture of multifactorial characters

Many factors can contribute to the epidemiology of a complex condition. There could be both genetic and environmental factors at work; the genetic factors could be polygenic, oligogenic, or monogenic (Mendelian) with any mode of inheritance, or any mixture of these, and the environmental factors may include both familial and nonfamilial variables. Given survey data on the relatives of a large collection of affected people, the statistical tool of segregation analysis can be used to explore the possible mix of factors. The analysis can provide evidence for or against the existence of a major susceptibility locus and can at least partly define its properties, for example whether the susceptibility is mainly dominant or recessive. In complex segregation analysis a computer is allowed to consider all possible explanatory factors, including a whole range of possible inheritance patterns, allele frequencies, penetrances, and so on in an unconstrained way to find the mix of scenarios that gives the best overall fit to the observed data. The likelihoods are then re-calculated omitting or constraining individual factors, so as to arrive at the minimum mix that must be included to avoid a significant loss of explanatory power. This minimum set of factors is taken as the most likely representation of the true genetic architecture of the condition.

Various computer statistical packages have been used for segregation analysis (for example PAP, SAGE, FISHER, MENDEL; see Konigsberg *et al*. [1989; PMID 2606342] for discussion and references). The papers by Lalouel *et al*. (1983; PMID 6614001), Lange, Weeks & Boehnke (1988; PMID 3061869), and Rao & Province (2000; PMID 10545756) (see Further Reading) give a flavor of the issues and provide numerous references. However, given the likely genetic heterogeneity of most if not all complex phenotypes, the value of these top-down views can be questioned. Provided there is evidence that genetic factors are involved somewhere in the etiology, it may be more productive to dive in and use the tools of molecular genetics to hunt for the factors directly, rather than worry about their overall statistical properties. Thus, the focus has moved on from genetic epidemiology, first to linkage and association analysis and more recently to large-scale sequencing.

18.2 INVESTIGATION OF COMPLEX DISEASE USING LINKAGE

The theoretical framework set out in Section 5.4 has proven helpful in making non-Mendelian conditions more comprehensible, but it provides no help in analyzing the genetic architecture of any particular condition. Family, twin, and adoption studies can confirm the relevance of genetic factors in a complex condition, but they do nothing to identify those factors. Inspired by the success of linkage analysis in mapping the genes responsible for Mendelian diseases, researchers in the 1990s attempted to apply similar tools to complex diseases. Special computer programs were needed to do this. The standard programs used to calculate lod scores from family data (see Section 17.1) require the frequency and penetrance of each allele at the postulated disease locus to be specified. While reasonable guesses for these parameters can be made for Mendelian characters, this is clearly not possible for complex conditions. Thus, linkage analysis for complex conditions must be model-free (often called **nonparametric linkage**, NPL, although statisticians might object to that term). **Model-free linkage** analysis compares the extent to which relatives share alleles or haplotypes identical by descent (IBD, see **Figure 12.10**) with the extent to which they share phenotypes. If alleles or haplotypes identical by descent are shared by affected relatives more often than would be expected under simple Mendelian principles, that is evidence of linkage. Model-free linkage analysis takes two forms:

- **Relative pair linkage methods** are used for dichotomous (e.g. affected vs. unaffected) characters. The genomes of pairs of related affected individuals are searched for chromosomal areas in which the IBD sharing for these pairs differs significantly from what is expected by chance;
- **Variance component methods** are used for quantitative traits. The variance of quantitative trait loci shared IBD between relatives is compared to their phenotypic covariance (see Almasy & Blangero, 1998, PMID 9545414 under Further Reading).

Affected sib pairs provide the main material for relative-pair linkage analysis

Suppose we had a collection of pairs of sibs, all affected by a genetic or part-genetic condition. Call the four parental haplotypes at some particular genomic location *A*, *B*, *C*, *D*, with the haplotypes inherited by the first affected sib in a family being labeled *A* and *C*.

Figure 18.2 shows the expected distribution of haplotypes in the second affected sib, as a function of the mode of inheritance of the condition and whether or not the haplotypes include a locus relevant to the condition.

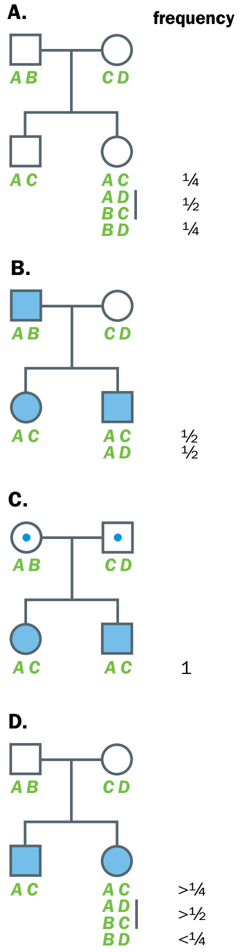


Figure 18.2 Affected sib pair analysis. (**A**) By random segregation, sib pairs share 2 (both *AC*), 1 (*AC* and either *AD* or *BC*), or 0 (*AC* and *BD*) parental haplotypes, ¼, ½, and ¼ of the time, respectively. (**B**) Pairs of sibs who are both affected by a Mendelian dominant condition must share the segment that carries the disease allele, and they may or may not (a 50:50 chance) share a haplotype from the unaffected parent. (**C**) Pairs of sibs who are both affected by a Mendelian recessive condition necessarily share the same two parental haplotypes for the relevant chromosomal segment. (**D**) For complex conditions, haplotype sharing above that expected to occur by chance (as in panel A) identifies chromosomal segments containing susceptibility genes.

In this form of model-free linkage analysis, a collection of a hundred or so affected sib pairs and (preferably) their parents are typed for markers spaced across the genome. Chromosomal regions are sought where the sharing is above the random 1:2:1 ratios of sharing 2, 1, or 0 haplotypes identical by descent. If the sib pairs are tested only for identity by state (see **Figure 12.10** for an explanation of the difference), the expected sharing on the null hypothesis must be calculated as a function of the allele frequencies.

The experience from many such studies in the 1990s was disappointing

Despite much effort in the 1990s, only a few candidate regions were identified, and different studies of the same disease often produced conflicting results. One of the few unambiguously positive findings was the discovery in 1991 of linkage between late-onset Alzheimer disease and chromosome 19 (later shown to be linkage to the apolipoprotein E locus), using model-free linkage but also standard lod score analysis (see Pericak-Vance *et al.* [1991] PMID 2035524 in Further Reading).

Affected sib pair analysis is a very robust procedure; there are few methodological problems or theoretical pitfalls. The problem is its low statistical power. Unfeasibly large numbers of sib pairs would be required to detect anything other than very strong effects. The power can be increased in various ways, for example by focusing on strongly discordant pairs (one affected, the other unaffected, or for quantitative traits, pairs with one sib in the top 10% and the other in the bottom 10% of the distribution). Provided paternity is confirmed, linkage would be indicated by IBD sharing less than the Mendelian expectation. Also, there is no special reason to focus on sib pairs. Affected individuals can be analyzed across extended pedigrees—indeed most of the relatively few successes of model-free linkage have been achieved by analyzing quantitative characters across extended pedigrees.

The disappointing results of linkage studies prompted a move to association studies

An important paper by Risch and Merikangas (1996, PMID 8801636; see Further Reading) calculated the number of pairs needed to have 80% power to detect significant linkage (**Table 18.3**). The number depends on the frequency of the susceptibility allele and on the degree of extra risk that it confers, but the overall message is clear: only uncommon alleles conferring a relative risk of 4 or more are reliably detectable with feasible numbers of sib pairs. The disappointing results from 10 years of effort show that such strong risk factors are few and far between.

TABLE 18.3 THE NUMBER OF AFFECTED SIB PAIRS REQUIRED TO GIVE 80% POWER TO DETECT LINKAGE, AS A FUNCTION OF THE FREQUENCY OF THE RISK ALLELE AND THE RELATIVE RISK IT CONFERS

Relative risk	4				2				1.5			
Allele frequency	0.01	0.1	0.5	0.8	0.01	0.1	0.5	0.8	0.01	0.1	0.5	0.8
Number of sib pairs	4260	185	297	2013	296,710	5382	2498	11,917	4,620,807	67,816	17,997	67,816

Data from Risch N & Merikangas K (1996) Science **273**:1516–1517; PMID 8801636.

In addition to pointing out the impracticality of using model-free linkage to identify weak susceptibility factors, the Risch and Merikangas paper showed that, given certain assumptions, tests of association had much greater power. Their paper was one catalyst for a general move of complex disease research over the next few years away from linkage and into association studies.

18.3 INVESTIGATION OF COMPLEX DISEASE USING ASSOCIATION

Linkage and association are sometimes confused, but they are different in principle (**Box 18.1**). A population association can have many possible causes, not all of which are genetic.

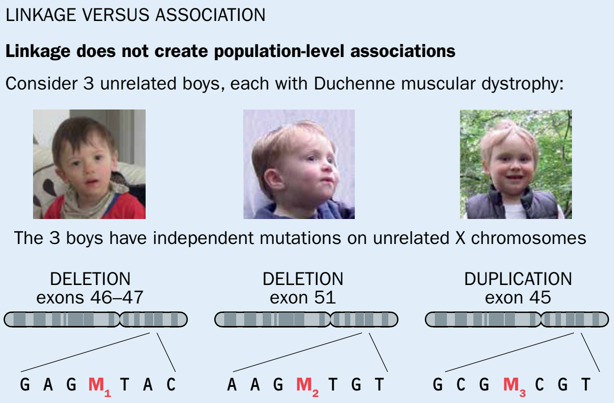
- **Direct causation:** having allele *A* makes you more susceptible to disease D. Possession of *A* may be neither necessary nor sufficient for somebody to develop D, but it increases the likelihood.
- **An epistatic effect:** people who have disease D might be more likely to survive and have children if they also have allele *A*.
- **Population stratification:** the population contains several genetically distinct subsets, and both the disease and allele *A* happen to be particularly frequent in one subset. Lander & Schork gave the example of the association in the population of the San Francisco Bay area between carrying the *A1* allele at the HLA locus and being able to eat with chopsticks. *HLA*A1* is more frequent among Chinese than among Caucasians.
- **Linkage disequilibrium (LD):** the disease-associated allele *A* has no direct role in susceptibility, but it marks an ancestral chromosome segment that also carries a susceptibility variant, as described in Section 12.2. The overwhelming majority of the valid associations identified by current studies are assumed to be of this type. It is then an additional (large) step to identify the actual causative sequence variant.

BOX 18.1 LINKAGE AND ASSOCIATION

Linkage is a relationship between *loci*. A disease *locus* is linked to a marker *locus*. The linkage relationship is the same whether the disease locus is occupied by a disease-causing allele or its normal counterpart, and regardless what allele is present at the marker locus. Association, by contrast, is a relationship between *alleles* or *phenotypes*. Having a certain disease or phenotype is associated with having a certain *allele* at a marker locus.

Linkage is a specifically genetic phenomenon—two loci are linked because they lie close together on a chromosome—but association is simply a statement of co-occurrence, with no features specific to genetics. In genetics, association most frequently takes the form of noting that people with a certain marker allele or haplotype are significantly more (or maybe significantly less) likely than on random expectation to have a certain phenotype or disease. For example, the HLA-DR4 antigen is found in about 36% of the general UK population, but in about 80% of people with rheumatoid arthritis. HLA-DR4 and rheumatoid arthritis are associated in this population.

Alleles at linked loci are not necessarily associated. Imagine, for example, three unrelated boys, each affected with Duchenne muscular dystrophy (**Figure 1**). There is no overall association between the disease and any allele at any of the six SNPs that are *linked* to the nearby dystrophin *locus*. However, if one of the boys had an extensive family history with several other affected males, within the family all affected cases would probably have the same SNP alleles. Thus, linkage creates associations within a family, but not among unrelated individuals.


LINKAGE VERSUS ASSOCIATION
Linkage does not create population-level associations
Consider 3 unrelated boys, each with Duchenne muscular dystrophy:
The 3 boys have independent mutations on unrelated X chromosomes
DELETION exons 46–47
DELETION exon 51
DUPLICATION exon 45
G A G M1 T A C
A A G M2 T G T
G C G M3 C G T

Box 18.1 Figure 1 Linkage does not create an association. Each of these three imaginary unrelated boys has Duchenne muscular dystrophy because of a frameshifting deletion or duplication in the dystrophin gene. They have different mutations (M_1, M_2, M_3) that arose independently on different X chromosomes. Genotypes are also shown for six SNPs that are *linked* to the dystrophin locus (the first SNP G/A, the second A/C, etc.). Overall there is no association between Duchenne muscular dystrophy and any particular allele at any of the linked SNPs.

Early association studies sought causative variants in candidate genes

Before 1980, when DNA-based markers suitable for linkage analysis became available, association studies were widely used in genetic analysis. The aim was to find variants that directly contributed to disease susceptibility. HLA-disease association studies were a major element of genetic research in the 1960s and 1970s. It was plausible that variation in tissue types should be a factor in disease, and their very high polymorphism fitted them well for association studies. These early studies detected the associations of HLA-DR4 with rheumatoid arthritis, HLA-DR3 and DR4 with type 1 diabetes, and HLA-B27 with ankylosing spondylitis. Later studies looking at non-HLA markers also had some successes. A 2003 meta-analysis of 301 publications on 25 frequently studied associations in 11 different diseases by Lohmueller and colleagues (PMID 12524541, see Further Reading) concluded that at least 8 of the 25 associations had been adequately replicated. However, in general, these early association studies were plagued by the same problems of poor reproducibility as the model-free linkage studies described above. Several causes can be identified:

- **Inadequate matching of controls.** These were all case–control studies, and often insufficient attention was paid to matching cases and controls. This is in fact a major concern, even in the most carefully designed studies;
- **Insufficient correction for multiple testing.** In any set of tests 5% of random observations will be significant at the $p = 0.05$ level and 1% at the $p = 0.01$ level. Each time the data are checked for another possible association, there is another chance of a type I error. The overall threshold of significance must be adjusted for the number of independent questions asked. A full (Bonferroni) correction divides the threshold *p* value by *N*, the total number of questions asked. To maintain an overall 5% chance of a false-positive result, the threshold *p* value for a single question is 0.05, for 10 questions 0.005, and for 1,000,000 questions (typical of studies using high-density SNP arrays) 5×10^{-8}. In Section 17.1 we argued that such a full Bonferroni correction was unnecessarily stringent in multilocus mapping of a Mendelian condition. The locus for a Mendelian character must be somewhere in the genome, and so excluding it from one chromosomal location increases the chance that it must be somewhere else. However, this argument does not apply to tests of association for a non-Mendelian condition. There is no necessity for *any* of the markers tested to show a true association;
- **Striking lucky in underpowered studies.** Even a true association may not be replicated in an independent sample. Many of these early studies were small and underpowered. An underpowered study may occasionally get lucky, but this luck is unlikely to be repeated in a similarly powered replication study. Targeted replication studies need much more power than the initial trawl. **Figure 18.3** illustrates the problem.

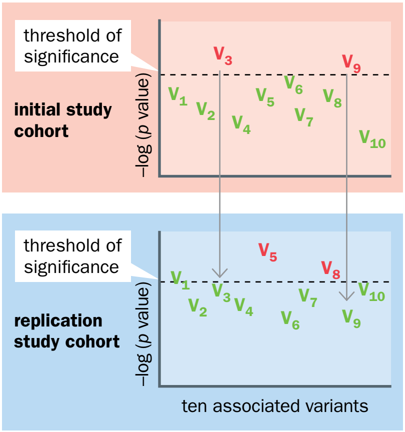


Figure 18.3 Striking lucky. (**A**) A study lacks the statistical power to reliably detect any of these 10 truly associated susceptibility factors. However, chance favorable combinations of genotypes in the cases and controls result in p values for two of the factors lying beyond the threshold of significance (dashed line). (**B**) In a similarly powered replication study, association with those two factors is not confirmed; instead two other factors show marginally significant associations.

The modern era of genome-wide association studies started with the Wellcome Trust Case–Control Consortium in 2007

After years of frustration and irreproducible results, genome-wide association studies (GWAS) moved on to a surer footing with the Wellcome Trust Case–Control Consortium (WTCCC). Three factors were essential contributors to the success of this study, which has served as a template for modern GWAS:

- The development of high-density SNP genotyping chips allowed a sample to be genotyped for 500,000 or more SNPs spaced across the genome in a single operation. See **Box 20.1** for a description of the technology;
- The HapMap project provided data that permitted a rational choice of tag-SNPs (**Figure 18.4**) to capture a significant proportion of all the common genetic variation in a population;

- Researchers appreciated that adequate statistical power required thousands of cases and controls and agreed to merge their individual efforts into consortia able to recruit such numbers; meanwhile funding agencies agreed to support such large-scale studies.

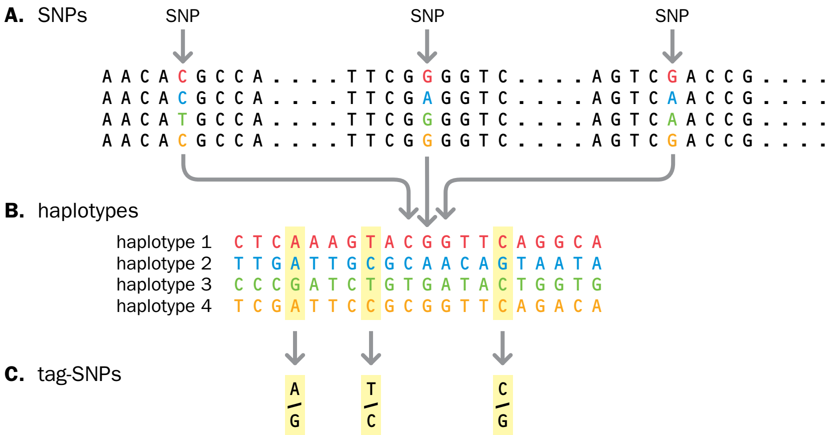

A. SNPs
SNP
SNP
SNP
A A C A C G C C A . . . . T T C G G G G T C . . . . A G T C G A C C G . . . .
A A C A C G C C A . . . . T T C G A G G T C . . . . A G T C A A C C G . . . .
A A C A T G C C A . . . . T T C G G G G T C . . . . A G T C A A C C G . . . .
A A C A C G C C A . . . . T T C G G G G T C . . . . A G T C G A C C G . . . .
B. haplotypes
haplotype 1 C T C A A A G T A C G G T T C A G G C A
haplotype 2 T T G A T T G C G C A A C A G T A A T A
haplotype 3 C C C G A T C T G T G A T A C T G G T G
haplotype 4 T C G A T T C C G C G G T T C A G A C A
C. tag-SNPs
A/G
T/C
C/G

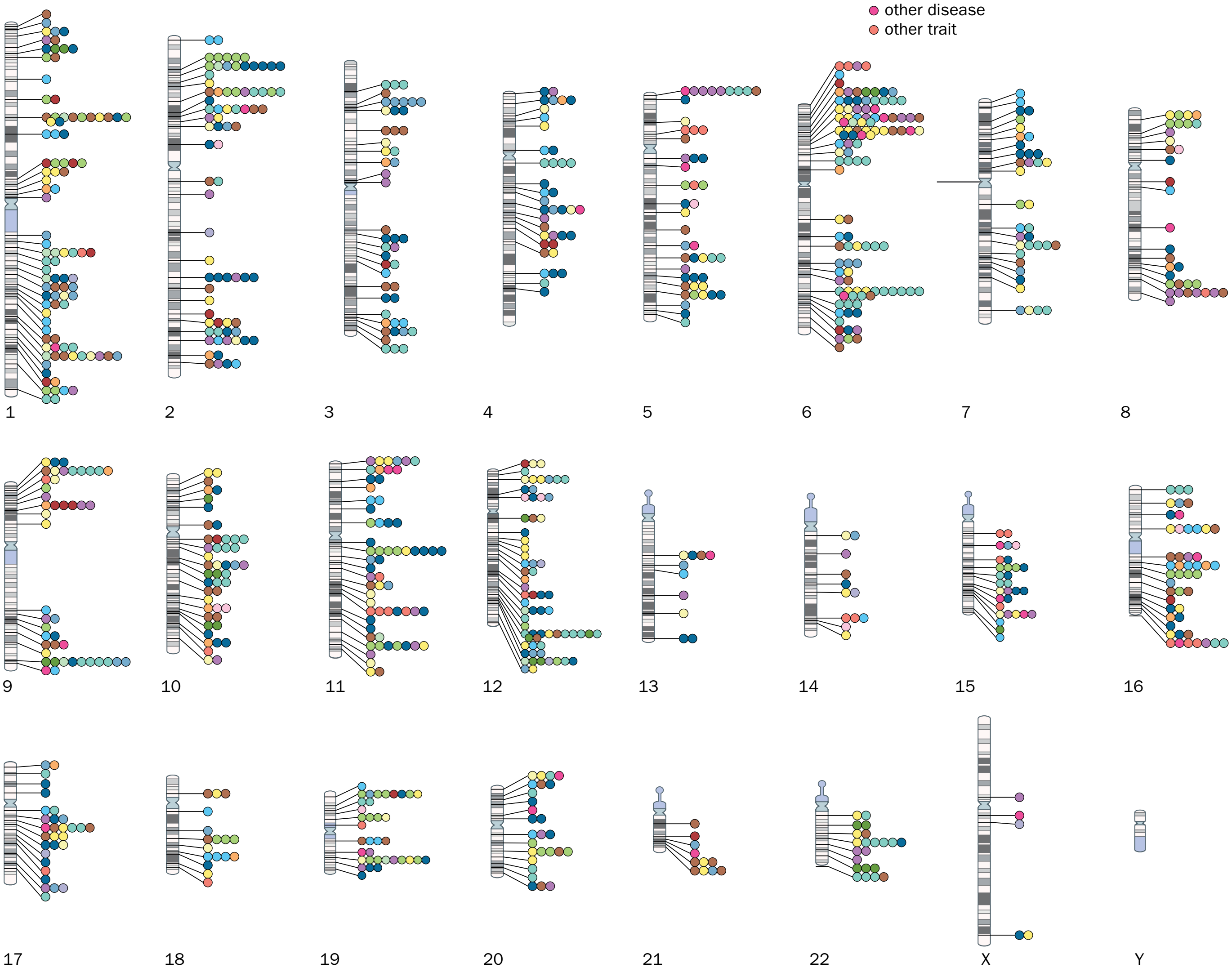
Figure 18.4 Using tag SNPs to characterize haplotype blocks. (**A**) A short segment of four individual copies of the same chromosome shows three biallelic SNPs. (**B**) Extended haplotypes of the same segment of these four chromosomes containing 20 SNPs, showing which allele of each SNP the chromosome carries. The unvarying sequence between SNPs has been omitted. Although there are 2^{20} possible combinations of 20 biallelic SNPs, a population survey shows that most copies of this chromosome have one of these four haplotypes. (**C**) Genotyping just three of the 20 SNPs serves to identify each of these four haplotypes. (Reprinted from The International HapMap Consortium [2003] *Nature* **426**:789–796; PMID 14685227. With permission from Springer Nature. Copyright © 2003.)

In the WTCCC, large consortia of British researchers assembled 2000 well-phenotyped cases of each of seven diseases: bipolar disorder (manic–depressive psychosis), coronary artery disease, Crohn disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes. In addition, 3000 presumed healthy controls were collected. All samples were typed for more than 500,000 SNPs. The estimated power of the study to detect risk factors was 43% for a factor conferring a relative risk of 1.3, and 80% for a factor conferring a relative risk of 1.5, assuming a minor allele frequency of at least 0.05. The first (2007) paper from the WTCCC (PMID 17554300, see Further Reading) is well worth a careful reading because it discusses matters that tend to be taken for granted in more recent studies.

Rigorous checks on data integrity were an essential part of the study (for full details see the Methods section of the online version of the 2007 paper). Samples were excluded from all analyses if data on more than 3% of SNPs was missing (indicating poor DNA quality) or if, over all SNPs, the level of heterozygosity was unusually high (indicating sample contamination). To ensure that only independent samples from individuals of Caucasian ancestry were used, genotypes were checked for duplicated samples, for indications of apparent relatedness between samples, and for indications of non-Caucasian ancestry. Individual SNPs were excluded from all analyses if more than 5% of samples registered missing data for that SNP, or if the genotypes did not fit a Hardy–Weinberg distribution (see Section 12.1). This very elaborate quality control was an important requirement for ensuring that reported associations were genuine.

Across the seven patient groups, 25 independent disease-association signals were identified with p values beyond the threshold of significance, calculated in this study as $p < 5 \times 10^{-7}$ rather than the 5×10^{-8} that is used in most more recent studies (**Figure 18.5**). As a test-bed for a new generation of large-scale GWAS, a key question was how many of the 25 could be replicated? The answer was very reassuring, both in terms of subsequent corroboration and conformity with past studies. Previous studies in the seven diseases were considered to have identified 15 variants with strong, replicated evidence of association with one or other of the seven diseases. Thirteen of the 15 were unambiguously identified in the WTCCC study, although only seven passed the threshold of significance. The great value of the WTCCC study was as an example of what is possible, rather than for the specific candidate genes that were identified. By showing how to perform valid and replicable association studies, the WTCCC set the template for the avalanche of GWAS that followed.

Figure 18.5 Summary data from 779 genome-wide association studies published by 2010. Each circle shows the chromosomal location of a significantly-associated variant, color-coded as shown in the key. A more up-to-date version with far more data is available at www.ebi.ac.uk/gwas/diagram but for present purposes it is uncomfortably overloaded with data.



Linkage disequilibrium, haplotype blocks, and tag-SNPs

Linkage disequilibrium was explained in Section 12.2. It is seen when genotypes at separate loci are correlated, so that knowing a person's genotype at one locus enables one to make a better than random guess about their genotype at another locus. It is a feature of chromosome segments that are shared between relatives. In Section 12.2 we saw how our genomes are structured in a series of haplotype blocks identified through the HapMap and other studies. The blocks are genomic segments where there has been little or no recombination across many generations; boundaries between blocks are recombination hotspots. People who share a haplotype block do so because they have all inherited it from a remote common ancestor, even though they would regard themselves as completely unrelated (we all have common ancestors, otherwise we would not all be the same species). Blocks are defined as regions of linkage disequilibrium: within a block genotypes of

SNPs are correlated. A typical block would be 5–15 kb long and include at least 20 SNPs (**Table 12.2**). If each SNP has two alleles, there are 2^{20} possible combinations of SNP genotypes across such a block, but in fact for most blocks most genomes have one of only 4–5 combinations. The block structure exists because the blocks are ancestral chromosome segments, shared by many apparently unrelated people in the present-day population.

Because of linkage disequilibrium, when scanning a genome for variants associated with susceptibility to a disease, it is not necessary to test every variant in the genome. Suppose a mutation on one block created a variant that increased susceptibility to a common complex disease. Although our hypothetical variant was neither necessary nor sufficient to cause the disease, and indeed might only slightly increase the risk, people with that haplotype block would be more likely to develop the condition than people who had one of the alternative blocks at that location. That would be true even if only a proportion of examples of that haplotype block carried the susceptibility variant. Thus, the basis of GWAS is to search the genome for haplotype blocks that are more frequent in cases—people with the condition under investigation—than in matched controls. In contrast to earlier hypothesis-driven studies that sought variants that directly impacted on disease susceptibility, GWAS simply look for any variants that show association with disease, whether because they directly affect susceptibility or (more likely) because they are in linkage disequilibrium with some unknown causative variant.

A haplotype block may contain 20–30 SNPs, but blocks can be identified by typing a much smaller number of carefully chosen **tag SNPs**. The principle was shown in **Figure 18.4**.

Reporting the results: relative risks and odds ratios

GWAS report the effect sizes in terms of **odds ratios** (**Box 18.2**). It would seem more intuitive to report relative risks—the risk of developing the condition under study for somebody with the relevant variant compared to the risk for somebody without that variant. Unfortunately, the relative risk cannot be calculated from typical GWAS data. All we have is a set of cases who have already developed the condition, and a set of controls who have not. To calculate the relative risk we would need to genotype an unbiased sample of the population, and afterwards check to see who was affected and who was unaffected. Unlike relative risks, odds ratios can be calculated directly from the data.

Odds ratios must be interpreted with caution. As the calculations in **Box 18.2** show, an odds ratio of 1.7 does not usually mean that having the variant increases somebody's risk by 70%. The deviation from the intuitive value is greatest for common variants, but as we have seen, GWAS data are all about common variants because most studies lack the power to detect effects of uncommon variants. A positive feature of odds ratios is that they do not depend on the numbers of cases and controls, only on the proportion of each

BOX 18.2 MEASURES OF RISK

Suppose variant V is associated with disease D.

The relative risk is $\dfrac{\text{risk somebody with V develops D}}{\text{risk somebody without V develops D}}$

This cannot be directly calculated from the results of a case–control study.

The odds ratio is $\dfrac{\text{odds of being a case for people with V}}{\text{odds of being a case for people without V}}$

Tabulating the results of a case–control study, where a, b, c, d are actual numbers of people in each category, we would have:

	Cases	Controls	Odds of being a case	Odds ratio
V present	a	b	a:b or a/b:1	(a/b)/(c/d) = ad/bc
V absent	c	d	c:d or c/d:1	

The odds ratio has some counter-intuitive properties. Suppose we type 1,000 cases and 1,000 controls for two variants. Variant 1 is present in 800 cases and 700 controls. Variant 2 is present in 80 cases and 70 controls. Intuitively we would think both variants have the same effect on risk, just variant 1 is more frequent in the population than variant 2. Now let us calculate the odds ratios:

	Cases	Controls	Odds of being a case	Odds ratio
V1 present	800	700	800/700	(800 × 300)/ (700 × 200) = 1.71
V1 absent	200	300	200/300	
V2 present	80	70	80/70	(80 × 930)/ (70 × 920) = 1.15
V2 absent	920	930	920/930	

Just as with relative risk, an odds ratio of 1 means there is no effect on risk. But when the ratio is significantly different from 1 (either higher or lower) its value depends on the frequency of the variant. Only for rare variants does it approach the intuitive value (8/7 or 1.14 in this example).

that have the risk variant. Odds ratios are normally quoted per allele. Unless there is evidence of dominance or recessiveness it is usually assumed that ratios are multiplicative, being 1 : r : r^2 for –/–, +/– and +/+ genotypes, respectively.

Figure 18.5 shows just how rapidly genome-wide association studies were applied to all manner of common conditions once the necessary technology became available. Since 2010 many other studies have reported. A very crowded updated version of this figure can be seen at www.ebi.ac.uk/gwas/diagram, while a database of studies is accessible at http://www.ebi.ac.uk/gwas/. In August 2018 the database contained 3,541 publications and 69,969 unique SNP-trait associations.

Results of the Wellcome Trust Case–Control Consortium study: Manhattan plots and effect sizes

The large amount of data produced by the WTCCC could be usefully summarized in a so-called Manhattan plot (**Figure 18.6**), while **Table 18.4** details the 25 significant associations corresponding to the red triangles in **Figure 18.6**. There are two main points to notice. First, the minor allele frequencies are all quite high. The power of a study to detect an association involves a trade-off between allele frequency and effect size. Even with thousands of cases and controls, the study would not have had power to detect an association with low frequency variants (minor allele frequency <0.05), unless the effect of the variant were very strong. The second point to notice is the generally small effect sizes. Apart from the well-known association of HLA alleles with the risk of Type 1 diabetes, almost all the odds ratios are below 2 and the majority are below 1.5.

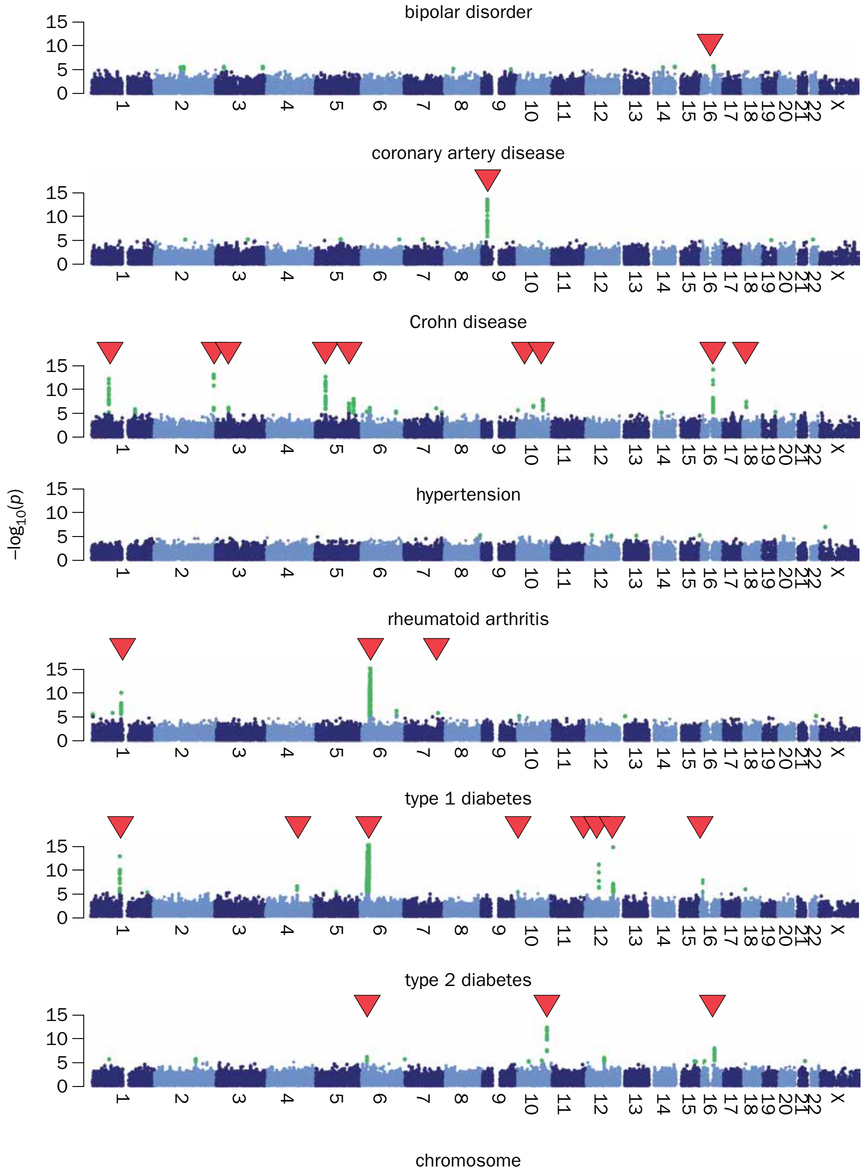


Figure 18.6 Results of the Wellcome Trust Case–Control Consortium (WTCCC) genome-wide association study. For each of the seven diseases studied, the distribution of p values (as $-\log_{10}p$) for the association of each SNP with the disease is shown in this 'Manhattan plot' at the appropriate chromosomal position. Most of the 469,557 SNPs that passed all the quality control checks showed weak or no association with the respective disease (blue dots, merged together). Those showing stronger evidence of association ($p < 10^{-5}$) are marked with green dots. The 25 most strongly associated SNPs or clusters of SNPs ($p < 5 \times 10^{-7}$, the threshold of significance in this study) are marked with red triangles. (Adapted from The Wellcome Trust Case–Control Consortium (2007) *Nature* **447**:661–676; PMID 17554300. With permission from Springer Nature. Copyright © 2007.)

TABLE 18.4 DETAILS OF THE 25 VARIANTS SHOWING SIGNIFICANT ASSOCIATIONS IN THE WTCCC STUDY

Disease	Chromosomal location	SNP	MAF in controls	MAF in cases	Odds ratio (for heterozygotes)
BPD	16p12	rs420259	0.282	0.248	2.08 (1.60–2.71)
CAD	9p21	rs1333049	0.474	0.554	1.47 (1.27–1.70)
CD	1p31*	rs11805303	0.317	0.391	1.39 (1.22–1.58)
CD	2q37	rs10210302	0.481	0.402	1.19 (1.01–1.41)
CD	3p21	rs9858542	0.282	0.331	1.09 (0.96–1.24)
CD	5p13	rs17234657	0.125	0.181	1.54 (1.34–1.76)
CD	5q33	rs1000113	0.067	0.098	1.54 (1.31–1.82)
CD	10q21	rs10761659	0.461	0.406	1.23 (1.05–1.45)
CD	10q24	rs10883365	0.477	0.537	1.2 (1.03–1.39)
CD	16q12*	rs17221417	0.287	0.356	1.29 (1.13–1.46)
CD	18p11	rs2542151	0.163	0.208	1.3 (1.14–1.48)
RA	1p13*	rs6679677	0.096	0.168	1.98 (1.72–2.27)
RA	6p21*	rs6457617	0.489	0.685	2.36 (1.97–2.84)
RA	7q32	rs11761231	0.375	0.327	1.44 (1.19–1.75)
RA+T1D	10p15	rs2104286	0.286	0.245	1.35 (1.11–1.65)
T1D	1p13*	rs6679677	0.096	0.169	1.82 (1.59–2.09)
T1D	4q27	rs6534347	0.351	0.402	1.30 (1.10–1.55)
T1D	6p21*	rs9272346	0.387	0.150	5.49 (4.83–6.24)
T1D	12p13	rs3764021	0.467	0.426	1.57 (1.38–1.79)
T1D	12q13	rs11171739	0.423	0.493	1.34 (1.17–1.54)
T1D	12q24	rs17696736	0.424	0.506	1.34 (1.16–1.53)
T1D	16p13	rs12708716	0.350	0.297	1.19 (0.97–1.45)
T2D	6p22	rs9465871	0.178	0.218	1.18 (1.04–1.34)
T2D	10q25*	rs4506565	0.324	0.395	1.36 (1.2–1.54)
T2D	16q12	rs9939609	0.398	0.453	1.34 (1.17–1.52)

BPD, bipolar disorder; CAD, coronary artery disease; CD, Crohn disease; RA, rheumatoid arthritis; T1D, Type 1 diabetes; T2D, Type 2 diabetes; MAF, minor allele frequency. * Previously reported robust associations. Details of individual SNPs can be accessed in the dbSNP database https://www.ncbi.nlm.nih.gov/projects/SNP/. (Data from The Wellcome Trust Case–Control Consortium [2007] *Nature* **447**: 661–678; PMID 17554300.)

The modest effect sizes were seen at the time as disappointing, but in fact they are only to be expected. Variants that are common in the population today must be ancient—it takes many generations for a mutation to rise to high frequency. If a variant had a substantial impact on reproductive fitness natural selection would quickly eliminate chromosomes carrying it. There are exceptions, for example the sickle cell variant in malarial countries, but in general any variant that has persisted long enough to become common in a population must be benign or, at worst, confer only a very small increase in risk, or perhaps affect risk for a late-onset condition that had little effect on reproductive fitness.

Current genome-wide association studies use phasing, imputation, and meta-analysis

Phasing

The raw data from a traditional SNP chip-based GWAS consist of genotypes. For many purposes it would be desirable to resolve the genotypes into haplotypes. This is the problem of **phasing**. The model we have developed here, explaining population associations in terms of shared ancestral chromosome segments, implies that risk is defined by haplotypes, not genotypes. Haplotype-disease associations should be closer to reality than the allele-disease associations seen when unphased genotype data are analyzed. In addition, imputation, as described below, can only be done on phased data.

Table 18.5 shows how a person's genotypes at the three tag-SNPs in **Figure 18.4** could be converted into haplotypes, provided we have a table of haplotype frequencies in the relevant population. Real haplotypes are not as tidy as those in the figure—linkage disequilibrium comes in shades of gray, not black and white as here—and real phasing programs use the data in a more sophisticated way, but the figures show the principle. The paper by Browning & Browning (2011, PMID 21921926; see Further Reading) gives a good overview of the problems and solutions of phasing, and should be consulted for more detail. But we seem to face a chicken-and-egg situation: phasing as shown here requires a list of frequencies of previously phased haplotypes. How do we get round this?

TABLE 18.5 STATISTICAL PHASING OF GENOTYPE DATA FOR A PERSON HETEROZYGOUS FOR EACH OF THE THREE TAG SNPS SHOWN IN FIGURE 18.4 (A/G, T/C, AND C/G)

Unphased genotypes	Possible phase A	Possible phase B	Possible phase C	Possible phase D
A/G, T/C, C/G	ATC/GCG	ACG/GTC	ATG/GCC	ACC/GTG
Population haplotype frequencies	47% / 0%	23% / 10%	2% / 1%	17% / 0%
Population frequency of haplotype pair	0%	2 × 0.23 × 0.10 = 4.6%	2 × 0.02 × 0.01 = 0.04%	0%
Posterior probability of haplotype pair	0%	4.6/4.64 = 99%	0.04/4.64 = 1%	0%

There are four possible phases (A–D). We suppose there is a database showing the frequencies of each hypothetical 3-locus haplotype in the population. There are four frequent haplotypes, as shown in **Figure 18.4**, plus two rarer ones not shown in that figure. The calculation shows that this individual has a 99% probability of having the phase ACG/GTC, corresponding to haplotypes 2 and 3 in **Figure 18.4**. (Modified from Browning SR & Browning BL [2011] *Nature Rev Genetics* **12**:703–714; PMID 21921926.)

The initial list of haplotype frequencies can be generated in various ways, some through laboratory work and some through computer analysis.

- If genotype data are available on relatives, haplotypes can be deduced by identifying segments identical by descent. Most simply, if we have data on trios of a person and both parents, the haplotypes can immediately be read off, except for loci where all three have the same heterozygous genotype.
- Sequencing provides phased data for loci on the same read. The short reads produced by most laboratory sequencers are of limited value, but single-molecule sequencers like the PacBio machine (see Section 6.5) can phase genotypes over tens of kilobases; whole chromosomes can be phased by using overlapping reads. Clone-based sequencing, as in the original Human Genome Project, is phased over the whole length of the longest clones used.
- Various methods could be used to isolate single homologs of chromosomes in the laboratory for sequencing. These might include microdissection, using a fluorescence-activated cell sorter (FACS) or generating haploid cells.
- Microfluidic tools have been developed that allow randomly-generated long DNA molecules to be partitioned into droplets, where they can be broken down to short fragments, each tagged with a barcode specific to the droplet. The fragments from

many droplets are pooled for conventional short-read sequencing. The barcodes are then used to reassemble the short reads to recover the sequence of the original long molecule. The number of long molecules per droplet is calculated so as to make it very unlikely that two long molecules from the same genomic region would be present in any one droplet.

All these methods are used on relatively small scales to generate reference panels of haplotypes. The thousands of cases and controls in a typical GWAS are then phased computationally. The accuracy of this depends on the size and quality of the reference panel of haplotypes. Data from the 1000 Genomes project are often used. Suitable reference panels are being established for many other populations worldwide.

Imputation

Imputation is the process of using knowledge of linkage disequilibrium to fill in genotypes at loci that were not part of the original experiment. In **Table 18.5** we phased an individual for the haplotypes shown in **Figure 18.4**. The person was only genotyped for the three tag-SNPs, but once we know he has haplotypes 2 and 3, we can impute phased genotypes at the remaining 17 SNPs in the figure. Imputed genotypes can be tested for association in just the same way as experimentally-determined types. Individual SNP alleles may be shared between different haplotypes, thus weakening their specific association with disease susceptibility that is a feature of only one of the haplotypes. An allele at an imputed SNP might by chance be present only on the disease-associated haplotype, and so give a stronger association. It might even be the causative variant. For example, in **Figure 18.4** each allele of the three tag-SNPs is present on more than one of the four haplotypes, but the A allele at SNPs 6 and 9 is present only on haplotype 1, A at SNPs 11 and 17 uniquely marks haplotype 2, and T at SNP 10 and G at SNP 20 mark haplotype 3.

Correct imputation depends on the accuracy of phasing and is more difficult for rare variants. It is estimated that imputation from the 1000 Genomes data identifies 97% of common variants but only 72% of rare (minor allele frequency (MAF) <0.01) variants in typical SNP genotyping studies. As larger reference panels become available the ability to impute rare variants will improve.

Meta-analyses

A major value of imputation is in **meta-analyses**. In the WTCCC study, 2000 cases and 3000 controls per disease gave only 43% power to detect a variant giving an odds ratio of 1.3. Yet it is clear that most susceptibility factors have even weaker effects than this. Identifying them requires ever larger studies, with many thousands of cases and controls. Larger cohorts can be assembled by combining the results of several independent studies of the same condition in a meta-analysis. A typical meta-analysis would include tens of thousands of subjects. A meta-analysis of GWAS of adult height included 253,288 individuals of European ancestry from 79 separate studies (see Wood *et al*. [2014] PMID 25282103; in Further Reading). Different studies often used different genotyping platforms that genotype different SNPs within a haplotype block. Imputation is necessary to generate a common set of SNPs so that the separate datasets can be combined.

Very large studies allow associations with rarer variants to be tested. The WTCCC study, like most first-generation GWAS, was restricted to variants with MAF of 0.05 or greater, and most of the variants identified had rather weak effect sizes. The argument from natural selection implies that any variants with larger pathogenic effect sizes are likely to be rare. There is thus a strong interest in testing uncommon variants (MAF 0.01–0.05) and rare variants (MAF <0.01). To have a sufficient number of people with the variant, the number of cases and controls needs to be very large.

Several different variants in a region may each contribute independently or in combination to susceptibility. Sorting out the causal relationships is extremely difficult when all the variants are in linkage disequilibrium with each other. The main tool for this is **logistic regression**. A principal variant is selected, and the effect of other variants is studied, conditioned on the effect of the principal variant. If there is still an effect using this procedure, the second variant does indeed make an independent contribution.

Moving from statistics to biology

Since the pioneering WTCCC study, GWAS have become ever larger and more complex. **Figure 18.7** gives a good impression of the increasing scale and complexity of studies of Type 2 diabetes, which is in many ways the archetype of genetic analysis of a common

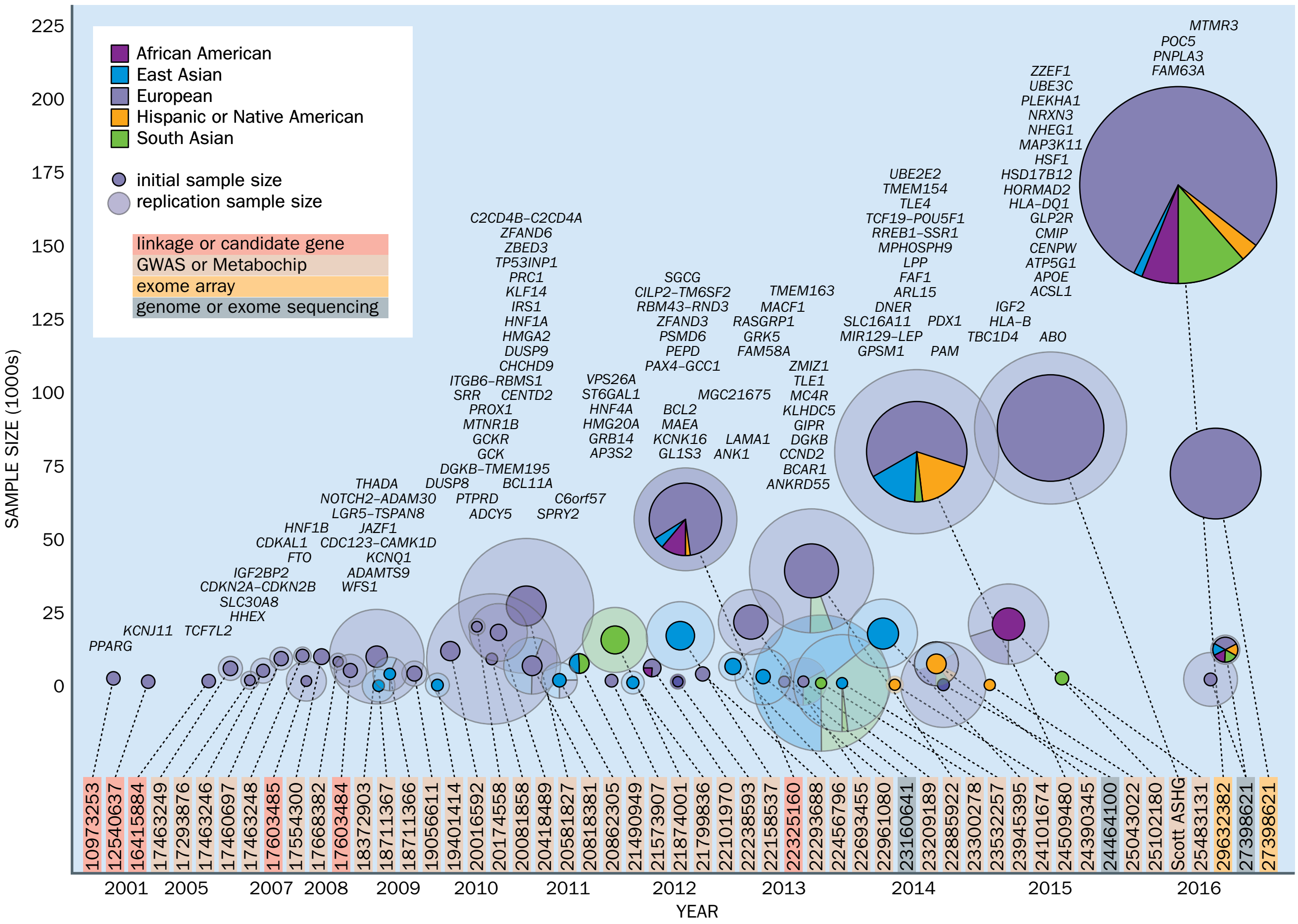
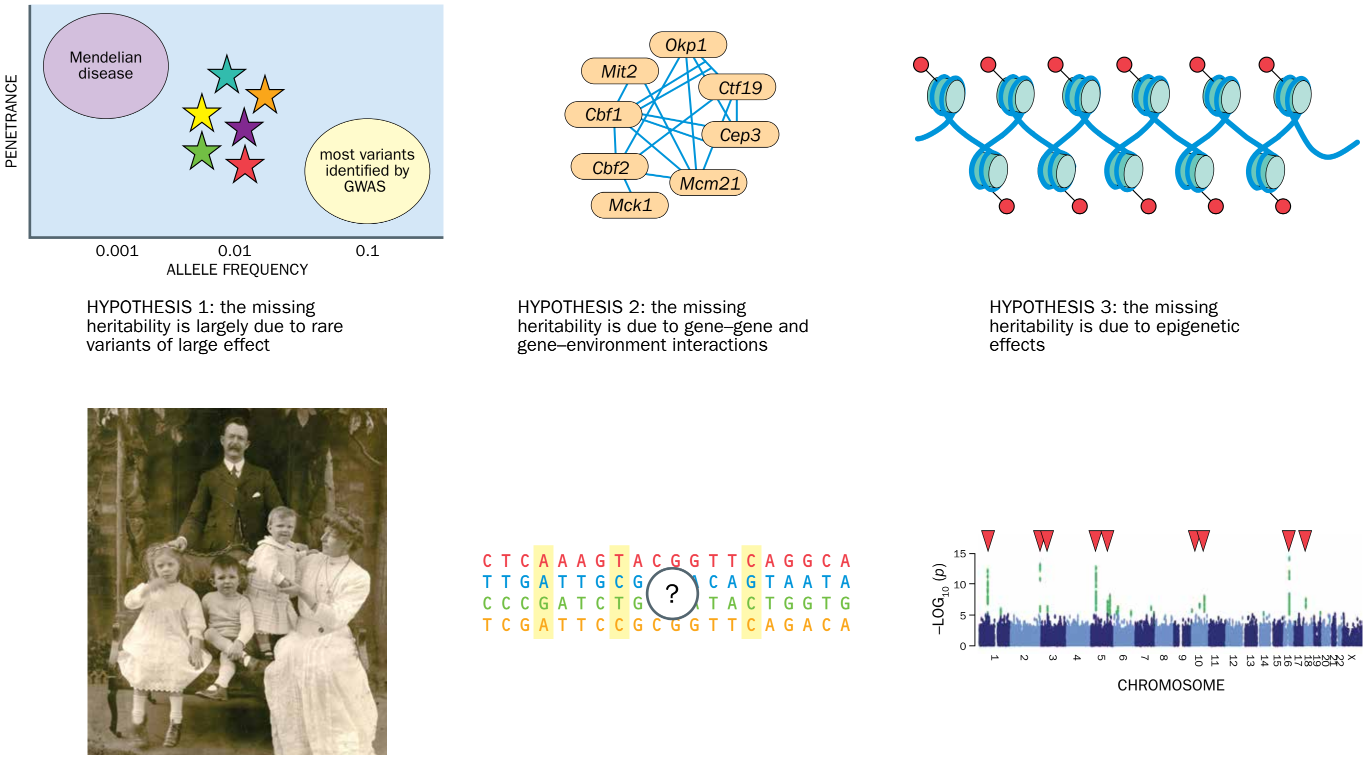


Figure 18.7 A timeline and summary of progress in genetic analysis of Type 2 diabetes. Bars at the bottom show studies, color coded according to methodology and with PMID numbers for the relevant reports. Dotted lines connect these to circles showing the sizes of the initial and replication samples in each study and the ethnic make-up of the study population. Identified susceptibility genes are listed above each circle. (Reprinted from Flannick J & Florez JC [2016] *Nature Rev Genetics* **17**:535–549; PMID 27402621, with permission from Springer Nature. Copyright © 2016.)

complex condition. The review by Flannick & Florez (2016, PMID 27402621; see Further Reading) from which this figure was taken gives a good feel for the many ways in which researchers have tried to move from association to biology. It also introduces discussion of one of the main controversies surrounding GWAS: the opinion of some critics that these studies were of very limited value because they did not explain most of the heritability. This question is considered further in the next section.

18.4 THE LIMITATIONS OF GENOME-WIDE ASSOCIATION STUDIES

The heritability of a character can be estimated in two ways: top-down through family studies (Section 18.1) or bottom-up by identifying individual susceptibility factors (Section 18.3). Almost always the combined effects of all known susceptibility factors identified through bottom-up studies account for less than half the heritability estimated from family studies. For example, a massive meta-analysis of prostate cancer susceptibility by Al Olama and colleagues (2014), covering 87,040 individuals and with 168 authors (PMID 25217961; see Further Reading), brought the number of identified genetic risk factors up to 99—but all together they accounted for only 33% of the familial risk in Europeans. Much effort has been spent on looking for explanations of this '**missing heritability**' (Figure 18.8). The various hypotheses are considered in more detail below. They are not mutually exclusive: different explanations may be relevant for different complex conditions, and more than one of the hypotheses may be true for a single condition.


PENETRANCE
Mendelian disease
most variants identified by GWAS
0.001
0.01
0.1
ALLELE FREQUENCY
Okp1
Mit2
Ctf19
Cbf1
Cep3
Cbf2
Mcm21
Mck1
C T C A A A G T A C G G T T C A G G C A
T T G A T T G C G ? A C A G T A A T A
C C C G A T C T G ? A T A C T G G T G
T C G A T T C C G C G G T T C A G A C A
–LOG10 (p)
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
CHROMOSOME

HYPOTHESIS 1: the missing heritability is largely due to rare variants of large effect

HYPOTHESIS 2: the missing heritability is due to gene–gene and gene–environment interactions

HYPOTHESIS 3: the missing heritability is due to epigenetic effects

HYPOTHESIS 4: there is no missing heritability; family studies overestimate heritability

HYPOTHESIS 5: GWAS underestimate heritability because causative variants are not reliably tagged by SNP chips

HYPOTHESIS 6: much heritability is due to common variants with very small effects

Figure 18.8 Possible explanations of the 'missing heritability'. See text for details.

Hypothesis 1: the missing heritability is largely due to rare variants of large effect

This hypothesis emphasizes the continuity between complex and Mendelian conditions. It is assumed that the large-effect variants are too rare to be assayed by the currently available commercial SNP arrays, and are not well tagged by the SNPs on the arrays. They can only be detected by sequencing. As more and more whole genome sequences have become available many rare variants have been identified in individuals with various conditions, but as a general explanation of the missing heritability this highly plausible hypothesis has not fared well. For example, a 2013 study by Hunt and colleagues (PMID 23698362; see Further Reading) sequenced exons of 25 genes previously associated with autoimmune disease (autoimmune thyroid disease, celiac disease, Crohn disease, psoriasis, multiple sclerosis, and Type 1 diabetes) in 24,892 affected subjects and 17,019 controls. Some rare variants were detected, but it was estimated that collectively they contributed less than 3% of the heritability due to common variants. Similarly, the huge study of Type 2 diabetes by Fuchsberger and colleagues (2016) found little evidence that low-frequency variants have a major role in predisposition to the disease (PMID 27398621; see Further Reading).

Hypothesis 2: the missing heritability is due to gene–gene and gene–environment interactions

Estimates of overall heritability from GWAS assume the contributions of single variants (or maybe single haplotypes) can just be added together to arrive at a final figure. Yet we know that virtually everything that happens in cells is due to combinations of different proteins and/or functional RNA molecules, often in large multimolecular machines. Thus it seems quite reasonable that genotype–genotype interactions could play a significant role in disease susceptibility. Zuk and colleagues (2012, PMID 22223662; see Further Reading) showed theoretically how such interactions could inflate top-down estimates of heritability, so that actually there is less 'missing heritability' needing to be explained. However, evidence for an important role of interactions has been hard to come by. Attempts to identify interactions in GWAS data or twin studies have not produced significant results, though it should be said that the statistical tests used have low power, so negative results do not rule out some role for interaction effects. The top-down heritability that we are trying to explain is normally taken as the narrow heritability—that is, the heritability due to additive genetic effects only. Thus it could be argued that even if interaction effects are biologically important, they are not relevant to the missing heritability problem.

Hypothesis 3: the missing heritability is due to epigenetic effects

Epigenetic modifications such as DNA methylation would not be picked up by SNP chips or by conventional next-generation sequencing, yet we know they play a major role in determining levels of gene expression. However, we should ask, what directs the epigenetic modifications to those particular sites? Conventional genotyping, in conjunction with RNA-seq to measure transcript levels, has identified many eQTLs, quantitative trait loci where sequence variants affect the level of gene expression. No doubt many of these act through modulating binding of transcription factors that in turn can trigger epigenetic modifications. Probably many GWAS variants are in fact eQTLs. Some special reasoning is needed to explain how particular epigenetic modifications that are associated with disease susceptibility could be heritable yet independent of genomic sequence.

Hypothesis 4: there is no missing heritability. Family studies over-estimate heritability

The theoretical study by Zuk and colleagues, mentioned above, suggested one reason why top-down estimates of heritability may be too high. In general, family studies may not take sufficient account of shared family environment or assortative mating as causes of familial clustering.

Hypothesis 5: GWAS underestimate heritability because causative variants are not reliably tagged by SNP chips

Causative variants are likely to have lower MAFs than tag-SNPs. They are unlikely to be selectively neutral and are probably of more recent origin. They probably arose by mutation on an already established ancestral haplotype, so that only some copies of the haplotype, as defined by tag-SNPs, carry the actual causative variant. Associations of disease with the actual causative variants will be stronger than associations with tag-SNPs, so calculations based on tag-SNPs will underestimate the heritability that the genotyping potentially explains.

Hypothesis 6: much heritability is due to common variants with very small effects

Genome-wide association studies necessarily use a very stringent threshold of significance, normally $p = 5 \times 10^{-8}$, to avoid a plethora of false positives. But among the associations that fail to pass the significance threshold, there are probably many true associations. The WTCCC noted 58 signals with single-point p values between 10^{-5} and 5×10^{-7}, a bit short of the significance threshold, but some of them were probably true positives. There are probably many relevant common variants that are well tagged by SNPs on genotyping chips, or that can be imputed with reasonable accuracy from whole-genome sequencing reference panels, but that have effects too small to be detected at the level of genome-wide significance. Even the largest GWAS meta-analyses have low power to detect variants that have weak effect sizes. Thus much of the heritability may not be missing, but rather is hidden below the threshold of significance.

Peter Visscher and Jiang Yang have developed a method that can estimate the overall contribution of all variants to the heritability. In essence their method considers pairs of unrelated individuals; it uses genotyping or sequence data to ask how far their genomes are similar and compares this to information on how far their phenotypes for the trait of interest are similar (see the 2010 [PMID 20562875] and 2015 [PMID 26323059] papers by Yang and colleagues, in Further Reading, for more detail). It cannot identify individual loci, but it provides a figure for the overall effect of all variants. They have shown that their method, maybe in combination with the effects described in Hypothesis 5, can explain all the heritability of height and body mass. For those traits there is no missing heritability. Height and body mass are naturally polygenic characters where everybody would expect there to be innumerable small effects. The question remains how much of the heritability of less obviously polygenic characters will also be explained in this way. For Type 2 diabetes the answer seems to be, around 50%—in other words, less than 100% but still a considerable advance on the heritability attributable to individually identified variants.

Overall, various plausible effects can account for the missing heritability. It is not a mystery, just a problem, and we do not need to postulate exotic mechanisms to explain it.

But the fact remains that GWAS are not able to identify the loci responsible for, in most cases, at least half the heritability of a condition. GWAS have been extremely successful in identifying the 'low-hanging fruit', the variants with reasonably large effect sizes, and these are the most likely to be biologically interesting. One can question the value of pursuing larger and larger studies at greater and greater expense to chase variants with odds ratios of 1.03. Boyle and colleagues (2017, PMID 28622505; in Further Reading) provide a thought-provoking discussion of this topic. It is surely more important to try to understand the associations we already know—to move from tag-SNPs to the causative variants and identify the biology underlying the associations. As GWAS are seen to have run their course, and as sequencing becomes ever cheaper, the emphasis is moving away from SNP chips and towards large-scale sequencing projects as the way to gain further insight into complex diseases.

18.5 WHAT HAVE WE LEARNED ABOUT THE GENETICS OF COMPLEX CHARACTERS?

The first important lesson is the need for a balanced view about the role of genetics in non-Mendelian conditions. This was well understood long before the era of GWAS, but public discourse is still plagued by naïve ideas of an opposition between nature and nurture. Almost nothing is nature *or* nurture. Almost everything is nature *and* nurture. It is quite hard to think of any human attribute that does not have at least some degree of genetic causation. Being knocked down by a truck when crossing the street might seem a good candidate—but there is probably some element of an impulsive personality and maybe poor vision or hearing, all of which are genetically influenced characters. But equally, geneticists have a responsibility not to exaggerate the roles of genetic factors in human variation. We should try to avoid talking about 'the gene for…'. Sometimes it is unavoidable, but at least we should make the effort and be on our guard against verbal bad habits.

Above all, geneticists should avoid exaggerating the utility of genetic tests. Genetic tests test for genotypes, not for phenotypes. They are of course immensely valuable—but only for genotypes that are usefully predictive of phenotypes. Just because a condition has substantial heritability, it does not necessarily follow that genetic tests can usefully predict an individual's risk. The 'lifestyle' genetic testing that companies offer over the Internet is fine as entertainment but should not be confused with clinical diagnosis. We will return to this point in Chapter 20 when we consider the roles and potential of genetic testing. Still less does it mean that if a condition has high heritability, then environmental interventions are irrelevant for preventing or ameliorating it. Genetic determination does not excuse fatalism about a condition. A high heritability means that no *common* variable in the *current* environment has any large influence. It says nothing about the potential of novel environmental interventions. The example of phenylketonuria should suffice to make the point: in this wholly genetic, Mendelian, condition the characteristic intellectual disability can be prevented by a special diet.

The fact that the 'missing heritability' problem is seen with virtually every character that has been studied by GWAS suggests that most non-Mendelian characters have a substantial degree of polygenic determination. A frequently reproduced diagram (**Figure 18.9A**) encapsulated the hopes of the GWAS pioneers: that a significant part of the genetic determination of non-Mendelian characters would be due to variants with frequencies and effect sizes intermediate between Mendelian and polygenic factors. Experience suggests such variants are few and far between. An updated diagram would be more like **Figure 18.9B**.

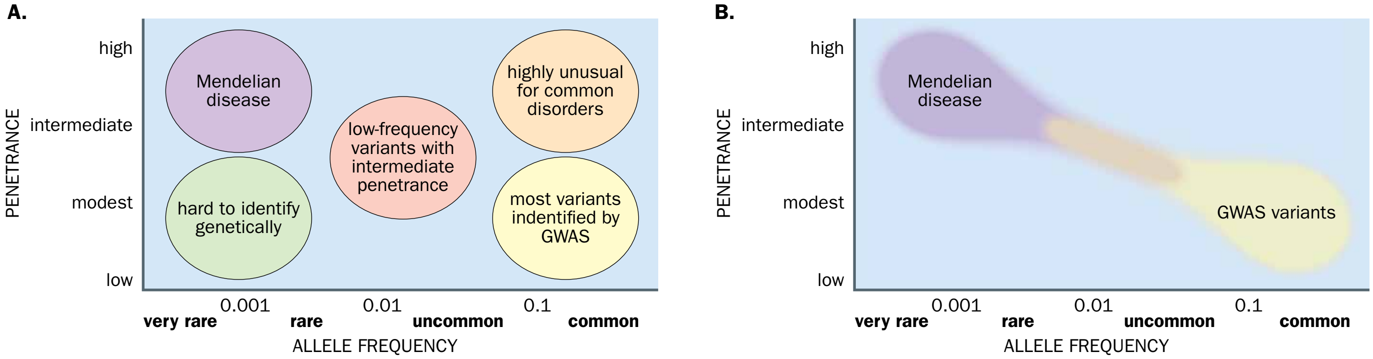


Figure 18.9 The role of genetic factors in determining phenotypes. (**A**). The hope in the early days of GWAS. (**B**) A view after 10 years of GWAS. (A, reprinted from McCarthy MI *et al.* [2008] *Nature Revs Genetics* **9**:356–369; PMID 18398418. With permission from Springer Nature. Copyright © 2008.)

Related to this is the question, how far are complex conditions aggregates of Mendelian or sub-Mendelian conditions? The record on this is mixed, as illustrated by a few examples.

- Valuable progress has been made in cancer, splitting common cancers into subtypes that are driven by different malfunctioning pathways that respond to different treatments—a topic developed in Chapter 19.
- Congenital heart disease can often be ascribed to specific gene mutations, and this information is useful for family testing and counseling about recurrence risks, but it has little influence on treatment, which is largely surgery to correct the specific abnormalities in an infant. For adult-onset heart disease, lifestyle factors are more important.
- On the other hand, Type 2 diabetes, one of the main targets of GWAS efforts, does not seem to be readily resolvable into clinically different diseases, except for the few percent of MODY (maturity-onset diabetes in the young) cases. These latter are Mendelian conditions due to mutations in one or other of about 7 genes. Different drugs are effective with different mutated genes and MODY diagnosis and genotyping is clinically valuable. But for the 95% or so of Type 2 diabetes that is not MODY there are only weak hints of genetic subtypes that might benefit from differential management.
- Psychiatric disease seems inherently complex. Moving diagnostic labels from broad behavioral to specific genetic categories would be a huge advance and would surely lead to much improved treatment. There has been progress in understanding the genetics of schizophrenia and autism, but it has mainly come from identifying tiny subsets caused by copy number changes or *de novo* point mutations. The core of each condition remains defiantly complex. Interestingly, some of the validated genetic susceptibility factors are shared by several conditions, for example schizophrenia, autism, and intellectual disability, as shown in **Table 15.6**. This suggests that there is some sort of genetically-influenced general neurodevelopmental vulnerability, which then manifests as different conditions depending on some specific trigger.

Identifying the causative variants

A significant difference between Mendelian and complex conditions concerns the location in the genome of the relevant genetic factors. Although there are plenty of individual exceptions, the determinants of Mendelian conditions are usually located in protein-coding sequences. Sequencing exons and splice sites of individual genes or whole exomes will identify the causative mutation in over 80% of cases of most Mendelian conditions. But most of the factors identified by GWAS do not lie in coding sequences. This makes good intuitive sense. For a condition to be Mendelian, a single DNA sequence change must cause the condition, regardless of all the other 4 million or so variants in a person's genome, and regardless of their environment, lifestyle, and history. If any of those other factors had a significant influence the condition would not be Mendelian. Only a major change to a protein would be likely to have such a drastic effect. On the other hand, the variants that contribute to complex conditions must have much more subtle effects, so that they only influence the phenotype in combination with many other variants. They are likely to affect regulatory sequences, marginally increasing or decreasing expression of a gene. Indeed, many of the factors identified by GWAS map on to eQTLs, loci where a variant influences the level of expression of a gene. Any variants that do lie in coding sequences would probably only marginally affect the function of the gene product.

This difference means that identifying the causative variants underlying GWAS associations is difficult. For Mendelian conditions one can generally spot causative mutations in coding sequence. They are likely to be deletions, frameshifts, nonsense mutations, or splice-site mutations. Only mis-sense changes or less obvious splicing changes usually cause uncertainty. GWAS data are much harder to interpret. The data will probably have localized the effect to a certain haplotype block, but as **Table 12.2** shows, there may be 50 or more variants in the block, any one of which might be the true cause of the effect, but all of which are associated with the condition because of linkage disequilibrium. Moreover, since any susceptibility variant is likely to have arisen as a new mutation on a pre-existing common block, only some examples of the block may contain the functional variant. And the effect, if any, of a nucleotide change in a sequence that is probably, but not definitely, an enhancer is hard to assess. One can check how a change affects transcription factor binding profiles, but we saw in Chapter 10 that real functional binding sites do not usually have the optimal sequence as assessed by *in vitro* studies—weaker binding allows more flexible control. Within a block one can impute all variants and check the data to see which variants have the strongest effect, but to some extent this may depend on random effects of the way alleles are shared between the various haplotypes at that location. In any case, imputation may not work well for a variant that is present on only a subset of examples of a given block, as is likely to be the case for susceptibility variants. The new gene editing technologies open the possibility of creating each variant, one by one, on a fixed background to isolate its effect. All of these are possible avenues of enquiry, but they are laborious compared to inspecting coding sequence for a major change, and there are many hundreds of candidate variants to check. Thus, progress in moving from associated variants to causative variants has been slow.

Clinical application

Apart from helping our understanding, what practical benefits does the identification of susceptibility factors bring to the prevention or management of complex conditions? At the start of this chapter we wrote "Hopefully, knowledge of such factors will shed light on the pathogenesis of these conditions and suggest approaches for prevention or treatment. Genotyping a person for a set of variants might possibly allow a prediction of individual risk". Regarding prediction of individual risks, progress has been disappointing. We will look at some specific examples in Chapter 20 when we consider genetic testing; suffice it to say here that the hoped-for transition of medicine from a 'diagnose and treat' to a 'predict and prevent' model has been remarkably slow in arriving. Might this be because the 'missing heritability' problem is limiting our current ability to make useful predictions; once we have identified the missing heritability, might this limitation go away?

A theoretical paper by Roberts and colleagues (2012, PMID 22472521; see Further Reading) suggests otherwise. These authors developed a method of leveraging MZ twin concordance data to estimate how far total knowledge of the genetics of a complex condition would allow clinically useful predictions. Taking 24 conditions covering a broad spectrum of types and including major current causes of morbidity and mortality in the USA, the authors asked how often could one use a person's whole genome sequence to make a clinically useful prediction, if one knew every susceptibility factor and understood all their interactions? Their arbitrary definition of 'clinically useful' was a prediction of a risk of 10% or double the population risk, whichever is the greater (but their formulae can be adapted for other definitions). **Figure 18.10** summarizes part of their findings. The optimistic conclusion is that in the best-case scenario 90% of individuals

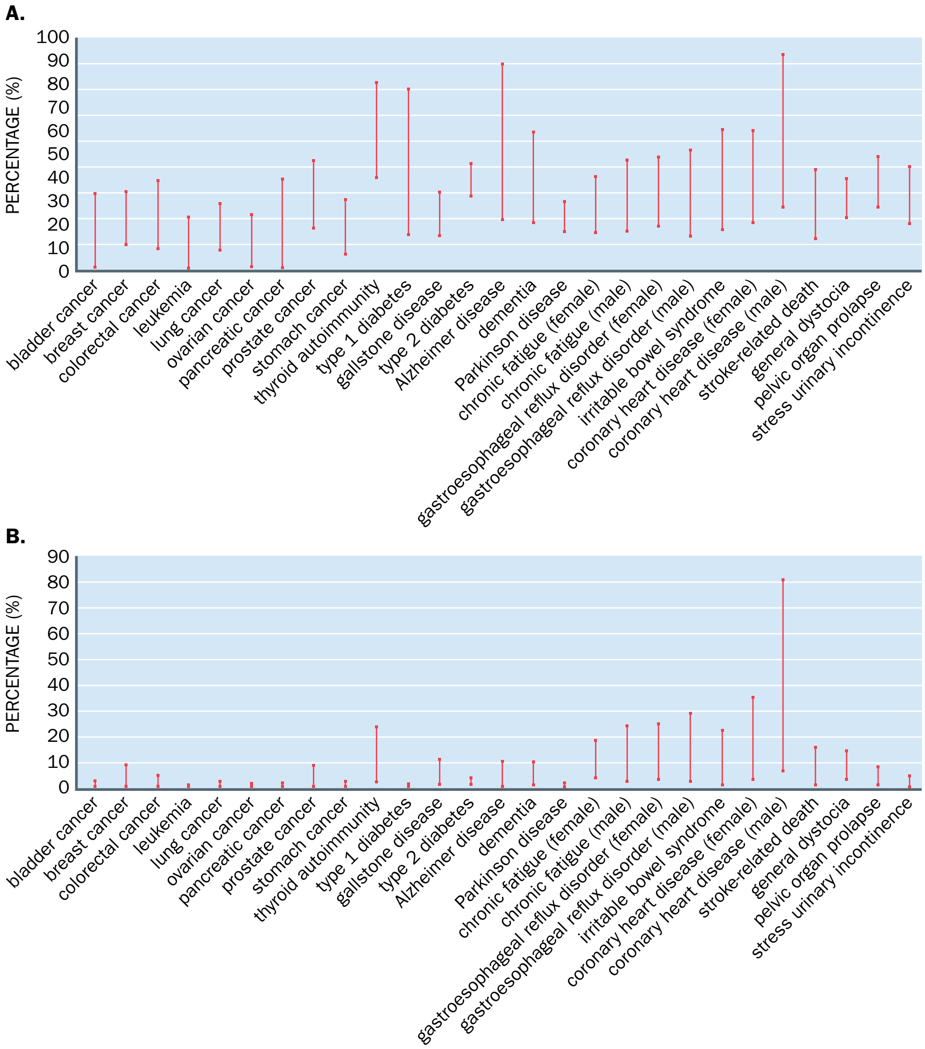


Figure 18.10 How often could we make a clinically useful risk prediction for an individual if we knew his whole genome sequence and had a complete understanding of the genetics of each condition? (**A**) The percentage of affected people and (**B**) the percentage of unselected individuals who would receive a clinically useful prediction of increased risk (the fraction in part B is generally small, as expected, because the incidence of most of the diseases is relatively low). See text for the definition of 'clinically useful'. The error bars do not reflect uncertainties in the predictions. They reflect limits in our *current* knowledge about the genetic architecture of the conditions. In the imagined future when we have total knowledge, those limits will have vanished. The data for each condition will be a single point, though we cannot currently say where that will fall on the bar. (From Roberts NJ *et al.* [2012] *Science Translational Medicine* **4**:133ra58; PMID 22472521. Reprinted with permission from the AAAS.)

would get a clinically useful prediction of increased risk for at least one of the 24 conditions. For most of the conditions most people would get a negative prediction (no significantly increased risk) and these negative predictions would seldom be clinically useful because in most cases any decrease in risk would be small.

Roberts *et al.* conclude that for most patients genetic testing will not displace routine check-ups and risk management based on the history, physical status, and life-style as the main strategy for preventative medicine. Such pessimism may seem unreasonable. Surely if we knew everything we should be able to accurately predict the genetic risk of each and every individual? The problem is that most common complex conditions have many independent weak genetic susceptibility factors. In only a very few individuals will all or most of the risk factors point in one direction—either to increased or to decreased susceptibility. For most people there will be a mix of factors conferring increased and decreased risk, which overall leaves the person at more or less the population risk. Thus, it is indeed true that perfect knowledge would allow perfect individual risk predictions—but only seldom would the risk be sufficiently different from the general population risk for the prediction to be clinically useful. Note that these conclusions apply to the 24 specified common complex conditions and not to Mendelian conditions, where sequence information is often crucial.

Moving forward

If the future does not lie with individualized risk predictions, what other practical benefits can we expect from our increased understanding of complex disease? One main hope is for more efficient drug development. In 2016 the average direct cost of moving a potential drug from initial lead to market approval was estimated at US$ 1.4 billion (see DiMasi *et al.*, PMID 26928437; Further Reading) and the process could take well over 10 years. The great majority of all leads fail in clinical trials, because of either insufficient efficacy or adverse effects. Some rare but serious adverse effects become apparent only during post-marketing surveillance. Withdrawing a drug at this stage has major financial consequences for the company and may lead to years of litigation and possible reputational damage. The example of rofecoxib (Vioxx™) shows just how serious the consequences can be for a company (see Krumholz *et al.* [2007], PMID 17235089; in Further Reading).

Identifying disease susceptibility factors can suggest novel leads for drug development, but the major benefit of genomic knowledge may be in selecting the most promising patients for clinical trials, and in predicting those at risk of adverse effects. A testimony to this interest is the proposal by the Astra-Zeneca pharmaceutical company to sequence two million genomes in collaboration with Human Longevity Inc., the Sanger Institute, Helsinki University, Montreal Heart Institute, and Genomics England (https://www.astrazeneca.com/Harnessing-the-code-of-life-to-develop-new-treatments.html).

While the future of medicine seems unlikely to lie with individualized risk predictions, **stratified medicine** or **precision medicine**—using genomic data to inform the management and treatment of a patient—is a major growth area. Oncology has been the main current beneficiary of this approach, as we will describe in Chapter 19, but the vision is to extend stratification to all areas of medicine. In the USA Barack Obama announced the $215 million Precision Medicine initiative in 2015 (see https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative), while in the UK the £300 million 100,000 Genomes Project was launched in 2014 by the then British prime minister, David Cameron (https://www.genomicsengland.co.uk/the-100000-genomes-project/). Other countries are developing their own ambitious plans—for example in March 2016 China announced a multibillion dollar 15-year initiative in precision medicine. The concepts and processes involved are further considered in Chapter 20.

SUMMARY

- Common diseases are usually complex, having many different possible causes.
- Evidence for the role of genetic factors in many common complex conditions comes from studies of families, twins, and adopted people. Such studies need careful interpretation to disentangle genetic effects from the effects of a shared family environment.
- Genome-wide association studies (GWAS) have been the main means of investigating susceptibility factors for complex conditions. These use large case–control studies to identify common marker alleles that are associated with the condition. The marker alleles identified in this way are not generally supposed to directly cause susceptibility themselves but are thought to be in linkage disequilibrium with the true causal factors.

- Over the years GWAS have expanded from studies of a few thousand cases and controls to meta-analyses of data from tens or hundreds of thousands of subjects.
- Meta-analyses depend on phasing raw genotype data to identify haplotypes, and on imputing genotypes at loci that were not part of the original experimental dataset.
- Although GWAS have identified thousands of susceptibility factors for many different conditions, most such factors make only a small contribution to susceptibility, and for most conditions all known factors combined account for less than half the overall genetic susceptibility deduced from family studies. This has prompted much speculation about the 'missing heritability'.
- The 'missing heritability; is probably accounted in part by the effect of many variants whose effects are too small to pass the rigorous tests used to assess statistical significance; in part because most identified variants are not the actual causal variant but are merely in linkage disequilibrium with it; and in part because the overall heritability may be overestimated by family studies. Interaction effects may also be important.
- Associations detected by GWAS are valid at the population level, but for most conditions they have little clinical utility for predicting risks for individuals. Their main clinical value may be in suggesting pathogenic mechanisms.
- Large-scale genome sequencing projects offer an alternative route to investigating the genetics of complex diseases.

FURTHER READING

Genetic epidemiology

Kety SS (1983) Mental illness in the biological and adoptive relatives of schizophrenic adoptees: findings relevant to genetic and environmental factors in etiology. *Am J Psychiatry* **140**:720–727; PMID; 6342426.

Kety SS, Rosenthal D *et al.* (1976) Mental illness in the biological and adoptive families of adopted individuals who have become schizophrenic. *Behav Genet* **6**: 219–225; PMID: 973827.

Kety SS *et al.* (1994) Mental illness in the biological and adoptive relatives of schizophrenic adoptees. Replication of the Copenhagen Study in the rest of Denmark. *Arch Gen Psychiatry* **51**:442–455; PMID 8192547.

Konigsberg LW, Kammerer CM *et al.* (1989) Segregation analysis of quantitative traits in nuclear families: comparison of three program packages. *Genet Epidemiol* **6**:713–726; PMID: 2606342.

Lalouel JM, Rao DC, Morton NE, Elston RC (1983) A unified model for complex segregation analysis. *Am J Hum Genet* **35**:816–826; PMID 6614001.

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet Epidemiol* **5**:471–472; PMID 3061869.

Polderman TJ, Benyamin B, de Leeuw CA *et al.* (2015) Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**:702–709; PMID 25985137.

Rao DC, Province MA (2000) The future of path analysis, segregation analysis, and combined models for genetic dissection of complex traits. *Hum Hered* **50**:34–42; PMID 10545756.

Rao DC (2008) An overview of the genetic dissection of complex traits. *Adv Genet* **60**:3–34; PMID 18358314.

Model-free linkage analysis

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* **62**:1198–1211; PMID 9545414.

Pericak-Vance MA, Bebout TJL, Gaskell PC Jr *et al.* (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am J Hum Genet* **48**:1034–1050; PMID 2035524.

Risch N (1990) Linkage strategies for genetically complex traits. 1. Multilocus models. 2. The power of affected relative pairs. 3. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* **46**:222–228; 229–241; 242–253; PMID 2301392–4. (Three key papers establishing the statistical basis of familial clustering and shared segment analysis.)

Principles of association studies

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**:703–714; PMID 21921926.

Lohmueller KE, Pearce CL, Pike M *et al.* (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**:177–182; PMID 12524541.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273**:1516–1517; PMID 8801636. (The power calculations in this paper helped trigger the move from linkage to association studies; see also *Science* **275**:1327–1330 [1997] for discussion.)

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**:506–516; PMID 8447318. (Describes the statistical basis for the transmission disequilibrium test, a family-based test of association, and shows an example of its power.)

Wellcome Trust Case–Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**:661–678; PMID 17554300. (An excellent overview of how to perform GWA studies and what they might show.)

Results of genome-wide association studies

Al Olama AA, Kote-Jarai Z, Berndt SI *et al.* (2014) A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer *Nat Genet* **46**: 1103–1109; PMID 25217961.

Flannick J, Florez JC (2016) Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* **17**:535–549; PMID 27402621.

Fuchsberger C, Flannick J, Teslovich TM *et al*. (2016) The genetic architecture of type 2 diabetes. *Nature* **536**:41–47; PMID 27398621.

Welter D, MacArthur J, Morales J *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42** (Database issue):D1001–D1006; PMID 24316577.

Wood AR, Esko T, Yang J *et al*. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**: 1173–1186; PMID 25282103.

Implications and limitations of GWAS

Boyle EA, Li Y, Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**:1177–1186; PMID 28622505.

DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ* **47**:20–33; PMID 26928437.

Hunt KA, Mistry V, Bockett NA *et al*. (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* **498**:232–235; PMID 23698362.

Krumholz HM, Ross JS, Presler AH, Egilman DS (2007) What have we learned from Vioxx? *Br Med J* **334**:120–123; PMID 17235089.

Roberts NJ, Vogelstein JT, Parmigiani G *et al*. (2012) The predictive capacity of personal genome sequencing. *Sci Transl Med* **4**:133ra58; PMID 22472521.

Yang J, Benyamin B, McEvoy BP *et al*. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**:565–569; PMID 20562875.

Yang J, Bakshi A, Zhu Z *et al*. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**:1114–1120; PMID 26323059.

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* **109**:1193–1198; PMID 22223662.

Cancer genetics and genomics

19

INTRODUCTION

Cancer is not a single disease. It does not have a single cause, still less is there a single "cure for cancer" waiting to be discovered. Just like all life on Earth, cancer is the result of natural selection acting on a mutable population to produce changes over the generations. We all understand how the vast array of unique species of animals and plants in the world has been produced by natural selection acting on random mutations over many generations. In just the same way, every tumor is the product of many episodes of mutation and selection. The only difference between the evolution of a tumor and the evolution of a new variety of animal is that in the case of a tumor the natural selection is operating on somatic mutations among the population of cells within a single multicellular organism, rather than on germ-line mutations in different whole organisms. The selection in tumors operates over many generations of cells, but all within the lifetime of a single host individual.

The ability of a population of cells to proliferate depends on the relative rates of cell division and cell death. Both are under genetic control. If a mutation conferred a proliferative advantage on one cell and its progeny, the progeny would outgrow the surrounding cells unless something prevented that happening. Thus, multicellular organisms have a natural tendency to develop tumors. But over their one billion years of evolution, multicellular organisms have developed multilayered and sophisticated controls on cell proliferation to prevent this happening. Cell division is tightly controlled in line with the current requirements of each particular tissue. Any cell that escapes the controls is triggered to kill itself by apoptosis.

A successful cancer cell needs to acquire a specific set of capabilities that were outlined by Hanahan and Weinberg in two important papers in 2000 and 2011 (**Figure 19.1**). These detailed papers (PMID 10647931 and 21376230; see Further Reading) with their logical overview are highly recommended reading for getting a good understanding of tumorigenesis (although they fail to discuss one significant feature of cancer cells, which is that they dedifferentiate). Acquiring each capability is likely to involve inactivating or bypassing a specific control mechanism. To convert a normal cell into a malignant cancer cell usually requires multiple mutations. Tumors develop through stages showing increasing proliferation and decreasing cell differentiation. Mutations occur at random, but to induce cancer they must fall onto fertile ground. Pathologists have long known that tumors most frequently arise in rapidly-dividing tissues. The cells that are most likely to found a tumor are cells that already possess some elements of the necessary capabilities. Primarily these are stem cells that have a high proliferative capacity, either intrinsically or induced by tissue damage or inflammation.

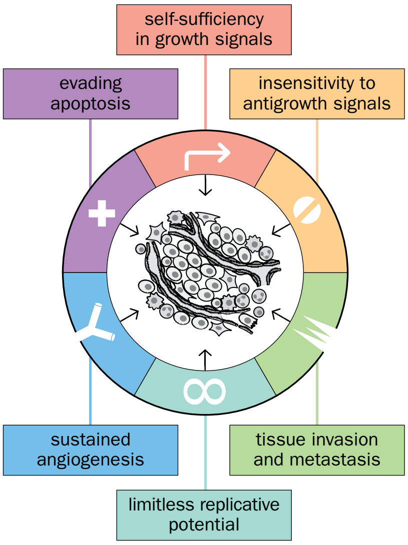

self-sufficiency in growth signals
evading apoptosis
insensitivity to antigrowth signals
sustained angiogenesis
tissue invasion and metastasis
limitless replicative potential

Figure 19.1 Six essential "hallmark" capabilities of cancer cells as proposed by Hanahan & Weinberg in 2000. In a follow-up paper in 2011 the authors suggest two further "emerging" hallmarks might be re-programming of energy metabolism to support continuing cell proliferation and the ability to evade immune surveillance. Two "enabling characteristics" crucial to the acquisition of the hallmark capabilities are genomic instability and inflammation. See Hanahan & Weinberg (2000 and 2011) (PMID 10647931 and 21376230) in Further Reading. (Reproduced from Hanahan D & Weinberg RA [2000] *Cell* **100**:57–70; PMID 10647931. With permission from Elsevier.)

Accumulating random mutations takes time, hence cancer is predominantly a disease of older people. Even so, the normal rate of random mutation is far too low to make it likely that any single cell could acquire all the necessary mutations to become a fully-fledged cancer cell. It might therefore appear that the defenses are impregnable. Since one in three of us will develop cancer at some time in life, this is clearly not the case. Two mechanisms allow cells to overcome the defenses:

- A mutation may give the cell a growth advantage, without it necessarily acquiring the full set of capabilities. If a mutant cell can generate 1000 mutant daughter cells, the chance that one mutant cell will acquire the next necessary mutation has increased 1000-fold;
- A mutation may increase the general mutation rate by destabilizing the genome. Genomic instability is a near-universal feature of tumor cells. The instability can operate at all levels, including disordered segregation of chromosomes at cell division, high frequencies of structural variants and point mutations, and disturbed epigenetic regulation. Most tumor cells have bizarre karyotypes with grossly abnormal numbers of chromosomes and many structural rearrangements (**Figure 19.2**).

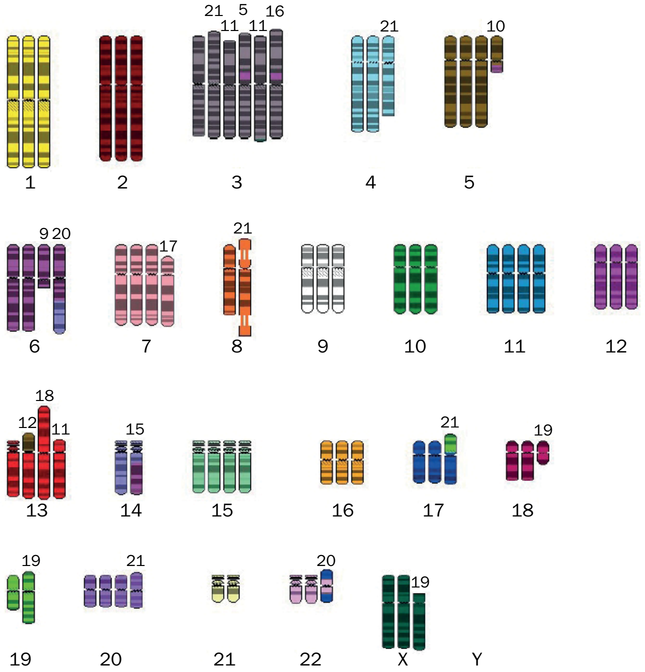


Figure 19.2 Chromosomal abnormalities in tumor cells. A "skygram" showing the results of analysis of 21 cells of the HT29 colon carcinoma cell line by multicolor fluorescence *in situ* hybridization (FISH). The chromosome number per cell varied between 67 and 71. Numbers above each abnormal chromosome show the number of cells in which that chromosome was present. (Reproduced from National Cancer Institute [NCI] and National Center for Biotechnology Information [NCBI] SKY/M-FISH and CGH Database [2001] www.ncbi.nlm.nih.gov/sky/skyweb.cgi.)

Because cancers normally depend on these two mechanisms, they develop in stages, starting with tissue hyperplasia or benign growths. Within a stage, successive random mutations generate an increasingly diverse cell population until eventually, by chance, one cell acquires the capabilities necessary to found a fully-fledged tumor. Exceptions to this gradualist model of tumor evolution can happen when a single event generates a large number of mutations simultaneously (see Section 19.4). Although a tumor may have its origin in a single mutant stem cell, the cells in a developed tumor do not constitute a single identical clone like a bacterial colony. Tumors are organs, containing a variety of cell types (**Figure 19.3**). Hanahan and Weinberg, in their 2011 paper, emphasize the important role of the abnormal but nontumorous stromal cells in the development of tumors. Because this is a textbook of molecular genetics rather than cell biology, we will not discuss these aspects of tumor biology, but that is not to deny their importance.

Even the cancer cells in a tumor are not all identical. Typically, there is a heterogeneous set of rapidly mutating cells, related by branching pathways of mutational history. A major challenge in research and management is to identify the "**driver mutations**" responsible for tumorigenesis against a background of many irrelevant "**passenger mutations**" and extensive cellular heterogeneity. Over the past decade large collaborative projects have described the mutational landscapes of many different types of tumor. Genes harboring driver mutations can be broadly divided into **oncogenes** and **tumor suppressor** (**TS**) **genes**. The natural function of oncogenes is to promote cell proliferation.

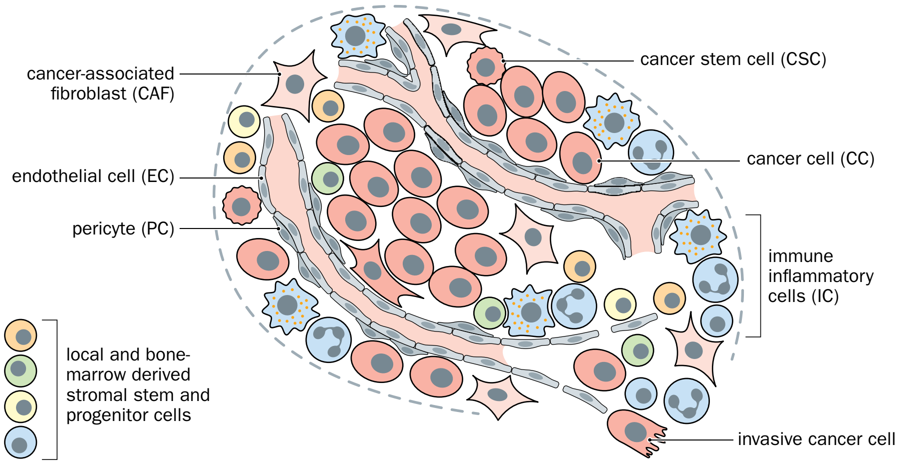

cancer-associated fibroblast (CAF)
endothelial cell (EC)
pericyte (PC)
local and bone-marrow derived stromal stem and progenitor cells
cancer stem cell (CSC)
cancer cell (CC)
immune inflammatory cells (IC)
invasive cancer cell

Figure 19.3 Tumors as organs. A tumor does not consist simply of a mass of cancer cells implanted in normal tissue. An evolving collection of stromal cells plays an integral part in tumor biology. (Reproduced from Hanahan D & Weinberg RA [2011] *Cell* **144**:646–674; PMID 21376230. With permission from Elsevier.)

Normally they act in a highly controlled manner and in response to specific signals; cancer cells carry gain-of-function activating mutations in oncogenes that foster unregulated cell division. TS genes normally act to restrain cell division or to safeguard the integrity of the genome; cancer cells often have loss-of-function mutations that inactivate these controls. The next three sections review these two classes of genes in more detail, then in Section 19.4 we zoom out to the genome-wide view of cancer that is emerging from current research. Finally in section 19.5 we consider how a better understanding of tumor biology and evolution can lead to more effective treatment.

Despite the bewildering heterogeneity, both between and within tumors, cancer is a more tractable research target than, say, schizophrenia or congenital heart malformations. The phenotype—uncontrolled cell proliferation—is relatively simple and amenable to analysis. Comparing the genome, transcriptome, and proteome of a tumor to their normal, constitutional counterparts in a patient gives a full picture of all the mutational events. Point mutations, structural variants, altered levels of gene expression, and epigenetic changes can all be catalogued and analyzed.

19.1 ONCOGENES

The oncogene story began in the 1970s with the discovery of **acute transforming retroviruses**. Some animal cancers (especially leukemias and lymphomas) are caused by viruses. The same is true of a small number of human cancers—examples include cervical cancer caused by human papillomavirus, Kaposi sarcoma caused by human herpesvirus 8, and adult T-cell leukemia caused by human T-cell lymphotropic virus. Some of the viruses, such as SV40 virus and papillomaviruses, have relatively complicated DNA genomes, but others are retroviruses that have very simple 7–10 kb RNA genomes. Acute transforming retroviruses recovered from animal tumors were able to transform cells in culture—that is, change their growth pattern to one resembling that of tumor cells.

The genome of a simple retrovirus has just three transcription units: *gag*, encoding internal proteins; *pol*, encoding a reverse transcriptase and other proteins; and *env*, encoding envelope proteins (see **Figure 8.8**). Acute transforming retroviruses incorporate an extra gene, the "oncogene". For a short time, some enthusiasts hoped that the whole of cancer might be explained by infection with viruses carrying oncogenes, and that once these had been identified, anticancer vaccines could be developed. However, researchers soon discovered that the viral oncogenes were in fact copies of normal cellular genes that had become accidentally incorporated into the retroviral particles through a random processing error. The viruses are noninfectious, being replication-defective.

Driver mutations often activate oncogenes

Functional understanding of oncogenes began with the discovery in 1983 that the viral oncogene v-*sis* (the v- prefix denotes a viral oncogene) was derived from the normal cellular platelet-derived growth factor B (*PDGFB*) gene. In the virus the gene was somehow activated, enabling the viruses to transform infected cells. The forms of oncogenes in normal cells are properly termed **proto-oncogenes**, but it is now common to ignore these distinctions and simply use the term oncogenes for the normal genes. The abnormal versions can be described as activated oncogenes. Many are named after the animal tumor in which they were first identified (**Table 19.1**).

TABLE 19.1 VIRAL AND CELLULAR ONCOGENES

Function	Cellular gene	Location	Viral oncogene	Animal source
SECRETED GROWTH FACTORS				
Platelet-derived growth factor B subunit	*PDGFB*	22q13.1	v-*sis*	Simian sarcoma
CELL SURFACE RECEPTORS				
Epidermal growth factor receptor	*EGFR*	7p11.2	v-*erbb*	Chicken erythroleukemia
Macrophage colony-stimulating factor receptor	*CSF1R*	5q32	v-*fms*	McDonough feline sarcoma
SIGNAL TRANSDUCTION COMPONENTS				
Cytoplasmic tyrosine kinase	*ABL1*	9q34.1	v-*abl*	Abelson mouse leukemia
Small GTPase	*HRAS*	11p15.5	v-*ras*	Harvey rat sarcoma
Small GTPase	*KRAS*	12p12	v-*ras*	Kirsten mouse sarcoma
TRANSCRIPTION FACTORS				
AP-1	*JUN*	1p32.1	v-*jun*	Avian sarcoma 17
MYC	*MYC*	8q24.21	v-*myc*	Avian myelocytomatosis
MYB	*MYB*	6q22	v-*myb*	Avian myeloblastosis
FOS	*FOS*	14q24.3	v-*fos*	Mouse osteosarcoma

The activated viral and normal cellular versions of a gene can if necessary be distinguished by the v- and c- prefixes; for example, v-*myc* and c-*myc*. The normal (nonactivated) cellular genes should strictly be termed proto-oncogenes, but are commonly described simply as oncogenes.

Most cancers, in humans and other animals, are not caused by viruses, but they do depend on abnormal activation of various growth-promoting genes that are therefore classified as oncogenes. Various nonviral mechanisms can activate a cellular (proto-)oncogene. **Table 19.2** shows the four most commonly encountered ways. Epigenetic changes might provide a fifth way, but while tumor suppressor genes are often subject to epigenetic silencing in cancer cells, it is not clear that epigenetic activation of oncogenes is a frequent primary event in tumorigenesis (but see the discussion of isocitrate dehydrogenase mutations in gliomas [Section 19.4] for possible exceptions).

TABLE 19.2 FOUR WAYS OF ACTIVATING (PROTO)ONCOGENES

Activation mechanism	Oncogene	Tumor
Amplification	*ERBB2 (HER2)*	Breast, ovarian, gastric, non-small-cell lung, and colon cancer
	MYCN	Neuroblastoma
Point mutation or small intragenic deletion	*HRAS*	Bladder, lung, and colon cancer; melanoma
	KIT	Gastrointestinal stromal tumors, mastocytosis
	EGFR	Non-small-cell lung cancer
Chromosomal rearrangement creating a novel chimeric gene	*BCR–ABL1*	Chronic myelogenous leukemia (see also **Table 19.3**)
Chromosomal rearrangement placing gene under the control of a powerful enhancer	*MYC*	Translocation to immunoglobulin heavy-chain locus by t(8;14) in Burkitt lymphoma

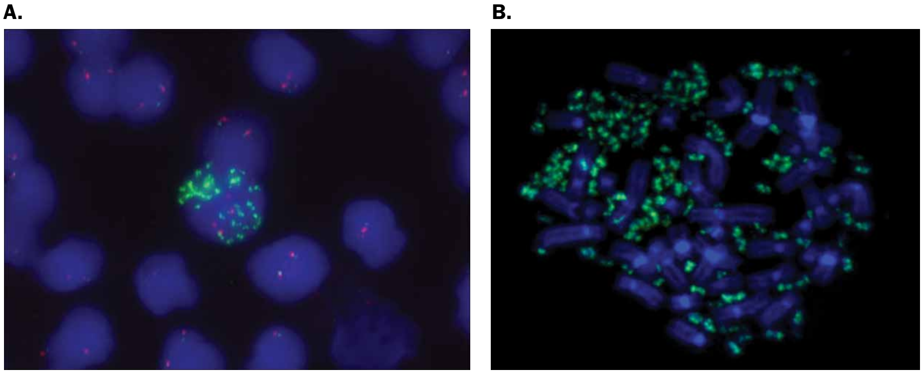


Figure 19.4 Amplification of the *MYCN* oncogene in neuroblastoma tumor cells. (**A**) A single neuroblastoma cell among normal interphase cells from the bone marrow of a patient. Fluorescence *in situ* hybridization was performed with a green-labeled *MYCN* probe and a red-labeled probe for a control sequence, *D2Z* from the centromere of chromosome 2. The normal cells each show two green and two red spots (sometimes superimposed to give a yellow spot). The neuroblastoma cell is chromosomally abnormal, with four copies of chromosome 2 (which carries the normal copy of the *MYCN* gene) but a very large number of extra copies of *MYCN* distributed through the nucleus. (**B**) In a metaphase neuroblastoma cell, the amplified *MYCN* copies can be seen to form small chromatin bodies (double minutes) separate from the normal chromosomes (stained blue). (Courtesy of Peter Ambros, St Anna Kinderkrebsforschung, Vienna.)

Activation by amplification

Many cancer cells contain multiple copies of a structurally normal oncogene. Breast cancers often amplify *ERBB2* (also called *HER2*) and sometimes *MYC*; a related gene, *MYCN*, is usually amplified in late-stage neuroblastomas and rhabdomyosarcomas (**Figure 19.4**). Hundreds of extra copies may be present. They can exist as small, paired chromatin bodies separated from the chromosomes (*double minutes*) or as insertions within the normal chromosomes (*homogeneously-staining regions*). The genetic events producing these may be quite complex because they can contain sequences derived from several different chromosomes. Similar gene amplifications are seen in cells exposed to strong artificial selective regimes—for example, amplified dihydrofolate reductase genes in cells selected for resistance to methotrexate. In all cases, the result is a great increase in the quantity of the gene product.

Amplification of a specific oncogene in tumor cells can be studied by fluorescence *in situ* hybridization (FISH) or by staining with an antibody to the protein product. Amplified sequences across the genome can be identified by array-comparative genomic hybridization or by analyzing the read depth on whole-genome sequencing. The latter two techniques also reveal any loss of material, which may point to TS genes (see below).

Activation by point mutation

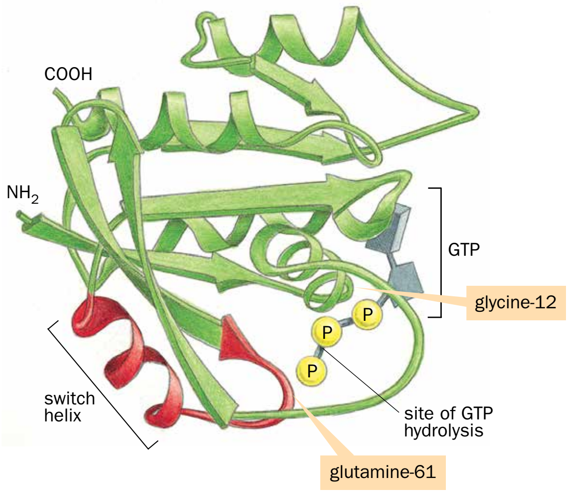
The epidermal growth factor receptor, EGFR or ERBB1, is a prime example of an oncogene that is frequently mutated in cancer. This receptor tyrosine kinase is often mutated in various cancers, especially non-small-cell lung cancer. Common mutations include a point mutation p.L858R or an 18 bp deletion c.2240_2257del18. The mutations all affect an ATP-binding pocket in the cytoplasmic part of the protein (see **Figure 19.28**), and have the effect of enhancing signaling and so producing a gain of function. These very specific mutations are important targets for therapy, as described below (Section 19.5).

The three *RAS* family genes, *HRAS*, *KRAS*, and *NRAS*, encode small intracellular proteins that mediate mitogenic signaling by receptor tyrosine kinases on the cell surface (see **Figure 3.8**). Binding of ligand to the receptor triggers binding of GTP to the Ras protein, and GTP–Ras transmits the signal onward in the cell. Ras proteins have GTPase activity, which rapidly converts GTP–Ras back to the inactive GDP–Ras and switches the signal off. Specific point mutations in *RAS* genes are frequently found in cells from a variety of tumors including colon, lung, breast, and bladder cancers. Almost invariably they encode substitutions of amino acids 12, 13, or 61 (**Figure 19.5**); each such substitution has the effect of decreasing the GTPase activity of the protein so that the GTP–Ras is inactivated more slowly, leading to an excessive cellular response to the signal from the receptor.

The *BRAF* oncogene encodes an intracellular tyrosine kinase that relays the signal from activated Ras proteins to the ERK kinase, eventually stimulating gene transcription (see **Figure 16.10**). Two-thirds of malignant melanomas, and a large number of other tumors, have an amino acid substitution in the kinase domain of BRAF that permanently activates it. A single mutation, p.V600E, accounts for 80% of all *BRAF* mutations in malignant melanoma which, again, presents an important target for therapy. *BRAF* is also sometimes activated by gene fusion, as described below.

In all these examples the activating mutations are very specific, and this suggests that other possible oncogenes may be identified by looking for genes that carry specific mutations in tumors, over and above the general level of random mutation. An application of this approach by Davoli and colleagues (2013 [PMID 24183448]; see Further Reading) identified 250 candidate oncogenes—most of these, however, had far weaker effects than those described here.

Figure 19.5 Structure of the HRAS protein in its GTP-bound active form. Hydrolysis of the GTP triggers a large conformational change in the switch helix (red). Replacement of either glycine-12 or glutamine-61 by almost any other amino acid abolishes the GTPase activity, resulting in oncogenic activation of the RAS signal. (From Alberts B *et al*. [2008] *Molecular Biology of the Cell*, 5th edn. Garland Science. With permission from WW Norton.)



Activation by a translocation that creates a novel chimeric gene

The best-known example of activation by a translocation that creates a novel chimeric gene is the Philadelphia (Ph^1) chromosome. This small acrocentric chromosome is seen in 90% of patients with chronic myelogenous leukemia. It is one product of a balanced reciprocal 9;22 translocation. The breakpoint on chromosome 9 is within an intron of the *ABL1* oncogene. The translocation joins the 3′ part of the *ABL1* genomic sequence onto the 5′ part of the *BCR* (breakpoint cluster region) gene on chromosome 22, creating a novel fusion gene. This chimeric gene is expressed to produce a tyrosine kinase related to the *ABL1* product but with abnormal transforming properties (**Figure 19.6**).

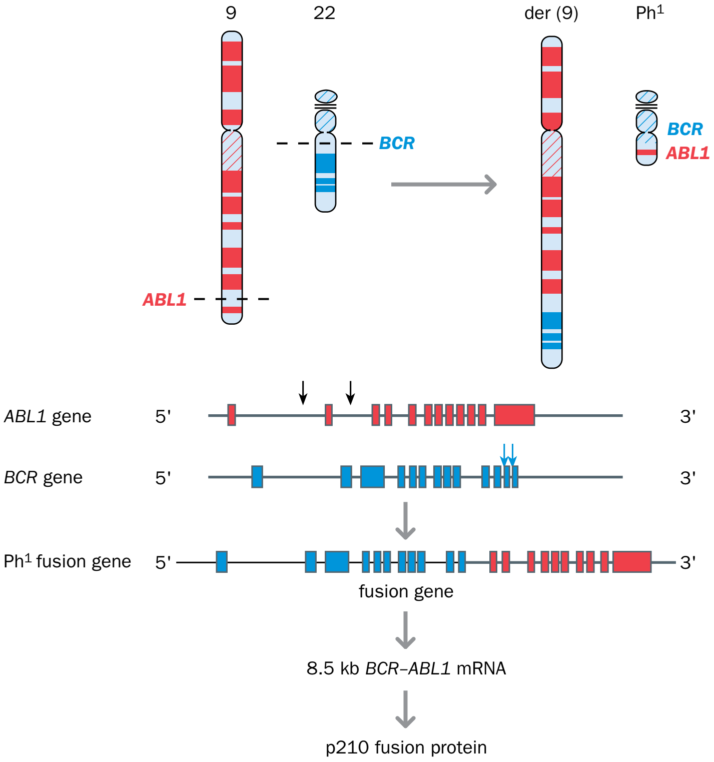


Figure 19.6 A chromosomal rearrangement that activates an oncogene. In chronic myelogenous leukemia, a translocation brings together exons of the *BCR* gene from chromosome 22 and the *ABL1* gene from chromosome 9. Arrows indicate observed breakpoints in different patients. One product of the translocation is the Philadelphia chromosome (Ph^1), containing a chimeric *BCR–ABL1* fusion gene. This encodes a constitutionally-active tyrosine kinase that does not respond to normal controls.

Many other tumor-specific recurrent rearrangements that produce chimeric oncogenes are known (**Table 19.3**). This mechanism is seen in 15–25% of leukemias, lymphomas, and sarcomas; similar rearrangements were less frequently reported in solid tumors, but this is partly because they were harder to identify under the microscope. Genome sequencing has removed this limitation. All known examples, including those in leukemias and lymphomas, are cataloged in the Mitelman database of chromosomal aberrations in cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman). As of May 2018 this listed 21,286 gene fusions. Many genes are promiscuous—that is, they are found in fusions with various different partners. The *MLL* (*KMT2A*) gene at 11q23 has been noted in fusions with more than 40 different partners. Fusions can be produced not only by translocations but also by inversions or, occasionally, by deletions that remove sequences that normally separate two genes. However, probably the great majority of the 20,000 or so fusions in the Mitelman database are passenger events, random rearrangements resulting from the general genomic instability. Only a few hundred are recurrent and likely causative. The review by Mertens and colleagues (2015) summarizes much information (PMID 25998716; see Further Reading).

Clinically, the specific gene fusions present in a patient's tumor cells can have an important bearing on the management and prognosis. They can be identified by sequencing or a targeted FISH assay. For example, to check for the 9;22 translocation in a patient with chronic myelogenous leukemia, differently colored FISH probes for the *BCR* and *ABL1* genes can be hybridized to an interphase cell. If the translocation is present, there will be one BCR signal, one ABL1 signal, and two fusion signals from the reciprocal products of the translocation (see **Figure 15.3**).

TABLE 19.3 EXAMPLES OF CHROMOSOMAL REARRANGEMENTS THAT PRODUCE TUMORIGENIC FUSION GENES

Tumor	Rearrangement	Chimeric gene	Nature of chimeric product
CML	t(9;22)(q34;q11)	*BCR–ABL1*	TK
AML	t(16;21)(p11;q22)	*FUS–ERG*	TF
Acute promyelocytic leukemia	t(15;17)(q22;q12)	*PM–RARA*	TF + RAR
Pre-B-cell ALL	t(1;19)(q23;p13.3)	*E2A–PBX1*	TF
	t(12;21)(p13;q22)	*ETV6–RUNX1*	TF
ALL	t(X;11)(q13;q23)	*MLL–AFX1*	TF
	t(4;11)(q21;q23)	*MLL–AF4*	TF
	t(9;11)(p22;q23)	*MLL–AF9*	TF
	t(11;19)(q23;p13)	*MLL–ENL*	TF
Ewing sarcoma	t(11;22)(q24;q12)	*EWS–FLI1*	TF
Ewing sarcoma (variant)	t(21;22)(q22;q12)	*EWS–ERG*	TF
Malignant melanoma of soft parts	t(12;22)(q13;q12)	*EWS–ATF1*	TF
Desmoplastic small round cell tumor	t(11;22)(p13;q12)	*EWS–WT1*	TF
Liposarcoma	t(12;16)(q13;p11)	*FUS–CHOP*	TF
Alveolar rhabdomyosarcoma	t(2;13)(q35;q14)	*PAX3–FOXO1*	TF
Papillary thyroid carcinoma	inv(1)(q21;q31)	*NTRK1–TPM3 (TRK* oncogene*)*	TK
Papillary thyroid carcinoma	inv(10)(q11.2;q21.2)	*CCDC6–RET*	TK
Non-small-cell lung cancer	inv(10)(p11.2;q11.2)	*KIF5B–RET*	TK
Non-small-cell lung cancer	inv(2)(p21;p23)	*EML4–ALK*	TK
Prostate cancer	del(21q22)	*TMPRSS2–ERG*	TF

Note how the same gene may be involved in several different rearrangements. For a comprehensive list see http://cgap.nci.nih.gov/Chromosomes/Mitelman. ALL, acute lymphoblastoid leukemia; AML, acute myelogenous leukemia; CML, chronic myelogenous leukemia; RAR, retinoic acid receptor; TF, transcription factor; TK, tyrosine kinase.

Activation by enhancer capture

Expression of a gene can be altered by a chromosomal rearrangement that brings it under the influence of a novel enhancer (see **Figure 16.9**). Rearrangements that bring one or another oncogene under the influence of the powerful B-cell-specific enhancers of immunoglobulin genes are a recurrent cause of B-cell malignancies. Burkitt lymphoma is the prime example. This tumor is especially common in malarial regions of Central Africa and Papua New Guinea. Malaria and Epstein–Barr virus (EBV) are part of the causation: malaria causes relative immunosuppression that allows EBV to transform B cells—but over-expression of the *MYC* oncogene is a central event.

A characteristic chromosomal translocation, t(8;14)(q24;q32), is seen in 75–85% of patients (**Figure 19.7**). The remainder have t(2;8)(p12;q24) or t(8;22)(q24;q11). Each of these translocations juxtaposes the *MYC* oncogene (normally located at 8q24) close to an immunoglobulin locus. This may be *IGH* at 14q32, *IGK* at 2p12, or *IGL* at 22q11. Unlike the translocations shown in **Table 19.3**, these translocations do not create novel chimeric genes. In the 8;14 translocation, the *MYC* and *IGH* genes are in opposite transcriptional orientations, head-to-head. Instead, they bring the oncogene under the influence of enhancers that normally ensure high expression of the immunoglobulin genes in antibody-producing B cells. Often, depending on the precise breakpoint, exon 1 of the *MYC* gene (which is noncoding) is not included in the translocated material. Deprived of its normal upstream controls and placed under the influence of a highly active enhancer, *MYC* is expressed at an inappropriately high level. Many T-cell malignancies involve a similar activation by an enhancer at a T-cell receptor locus. The review by Mertens and colleagues (PMID 25998716; see Further Reading) lists other examples, including some in solid tumors.

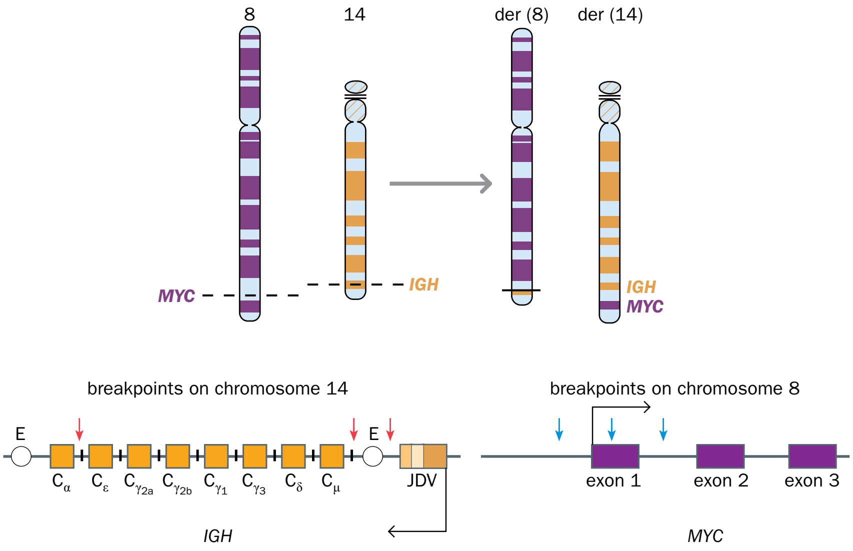


Figure 19.7 Activation of the *MYC* oncogene by capture of B-cell-specific enhancers. In Burkitt lymphoma, an 8;14 translocation places the *MYC* oncogene close to the *IGH* (immunoglobulin heavy chain) locus on chromosome 14. See **Figure 11.16** for details of the *IGH* locus. Unlike the *BCR–ABL1* translocation shown in **Figure 19.6**, this translocation does not create a fusion gene. The two genes are head-to-head, in opposite transcriptional orientations. Instead, the *MYC* gene comes under the influence of B-cell-specific enhancers (marked E) associated with the *IGH* locus. This causes high levels of *MYC* expression in B cells. The translocated *MYC* gene often lacks exon 1 (depending on the exact position of the breakpoint), but this exon is noncoding and its absence does not affect function of the MYC protein.

Enhancers may also be captured by disabling insulator sequences that normally prevent them from controlling expression of genes that are not in the same topologically-associated domain (TAD; see Section 10.2). This may be the result of small deletions, point mutations, or methylation. Some recent studies suggest this may be an under-reported cause of oncogene activation—see the discussion of *IDH1* mutations, below (Section 19.4).

MicroRNAs can act as oncogenes

Given the widespread involvement of miRNAs in control of gene expression, it is not surprising that disturbed function or expression of miRNAs is very common in cancer cells. In fact, aberrant miRNA expression is the rule rather than the exception in cancer. Because a single miRNA can target many mRNA species, and a single mRNA may be regulated by multiple miRNAs, it is difficult to make clean generalizations about the roles of miRNAs in cancer. Most miRNAs down-regulate expression of their targets, and so they might function mainly as TS genes; however, some miRNAs act as oncogenes, being up-regulated in cancer (**Table 19.4**).

TABLE 19.4 EXAMPLES OF microRNAs THAT ACT AS ONCOGENES

miRNA	Targets	Involvement in cancers
miR-17-92 cluster	*TP63*, *E2F1*, *CDKN1A*, *BCL2L11*	Up-regulated in lung and colon cancer, as well as lymphoma, medulloblastoma, and multiple myeloma
miR-21	*PTEN*, *PDCD4*	Over-expressed in multiple solid tumors
miR-106b-93-25 cluster	*CDKN1A*, *BCL2L11*	Over-expressed in multiple solid tumors and multiple myeloma
miR-155	*INPP5D*, *CEPBP*, *SPI1*, *ESPL1*, *PICALM*	Up-regulated in breast, lung, colon, and pancreatic tumors and hematopoietic malignancies
miR-221 and miR-222	*PTEN*, *TIMP3*, *CDKN1B*, *CDKN1C*, *BCL2L11*, *DDIT4*, *FOXO3*	Up-regulated in multiple solid tumors and in chronic lymphocytic leukemia

Data from Malumbres M (2013) *Mol Aspects Med* **34**:863–874; PMID 22771542.

Activation of oncogenes is only oncogenic under certain circumstances

Paradoxically, excessively-elevated signaling by oncoproteins such as RAS, MYC, and RAF usually triggers cell senescence and/or apoptosis rather than proliferation. As one of their many defense mechanisms, cells are able somehow to distinguish between normal and abnormal proliferative signals. One suggestion is that oncogenic stimulation triggers a DNA damage response (see Section 11.2) that shuts down cell cycling; only when the damage response is defective does a cell proliferate in response to the overexpression of an oncogene. This highlights the fact that organisms have multiple layers of defense against misbehaving cells, and that single mutations are not sufficient to cause cancer.

The overall cell biology provides an important context. Even in advanced cancers, the cells retain some characteristics of their tissue of origin, and this may affect the way oncogenic mutations act. For example, in lymphomas and leukemias, specific missense mutations in the *NOTCH1* gene are frequent, establishing *NOTCH1* as an oncogene. However, in squamous cell carcinomas, *NOTCH1* mutations are nonrecurrent and mostly inactivating, suggesting that in those cells it is acting as a TS gene. Notch signaling triggers different functions in different cell types.

19.2 TUMOR SUPPRESSOR GENES

The second major class of genes that are mutated in tumors are the TS genes. These are genes whose function is to keep the behavior of cells under control. This may entail restraining or suppressing inappropriate cell division, maintaining the integrity of the genome, or ensuring that incorrigibly deviant cells are sentenced to death by apoptosis. An early landmark in our understanding of tumor suppressor genes was Knudson's work in 1971 (PMID: 5279523; see Further Reading) on retinoblastoma.

Retinoblastoma provided a paradigm for understanding tumor suppressor genes

Retinoblastoma (OMIM #180200) is an aggressive childhood cancer of the eye. In some families, retinoblastoma is inherited as an autosomal dominant trait with reduced penetrance; other cases occur sporadically. Knudson noticed that children with the hereditary form of retinoblastoma often developed multiple tumors in both eyes. By contrast, people with the "sporadic" form developed a single tumor in only one eye. Also, in cases of hereditary retinoblastoma, the tumors typically occurred before the child was 5 years old; in sporadic cases, they occurred later in development. On the basis of these observations, Knudson reasoned that all retinoblastomas involved two "hits"—probably mutations—but that in the familial cases one hit was inherited (**Figure 19.8**).

Retinoblastoma is one of the so-called embryonal tumors, including neuroblastoma, medulloblastoma, nephroblastoma, and Wilms tumor, that develop in the fetus or very early in life. They arise in populations of cells that are transiently undergoing rapid proliferation before differentiating. Retinoblastomas develop from cone photoreceptor

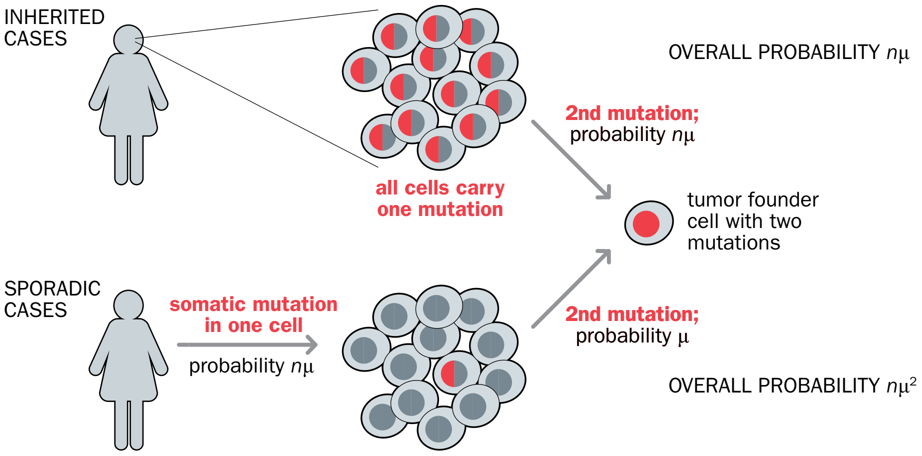

INHERITED CASES
all cells carry one mutation
OVERALL PROBABILITY $n\mu$
2nd mutation; probability $n\mu$
tumor founder cell with two mutations
SPORADIC CASES
somatic mutation in one cell
probability $n\mu$
2nd mutation; probability μ
OVERALL PROBABILITY $n\mu^2$

Figure 19.8 Knudson's two-hit theory for retinoblastoma. There are *n* target cells, and the probability of a cell suffering a hit is µ. In familial cases, one hit is inherited and present in every cell. Given the number of target cells and the likelihood of any one cell suffering a hit, it is highly likely that at least one cell in the target tissue will end up with a second hit, and so develop into a tumor. Thus familial retinoblastoma is quite highly penetrant, frequently involves multiple independent tumors, and has an early age of occurrence. In a person who did not inherit a hit, retinoblastoma will only develop if one cell happens to suffer two hits, a rare event. (Adapted from Read A & Donnai D [2015] *New Clinical Genetics*, 3rd edn. With permission from Scion Publishing.)

progenitor cells. Knudson's hits must be the rate-determining steps in tumorigenesis, but evidently the whole progression happens much more easily than in adult cells. Compared to adult cancers, embryonal tumors show strikingly little evidence of the usual extensive genetic instability. They do, however, show evidence of epigenetic dysregulation. The founder cells of embryonal tumors must be particularly vulnerable to loss of control of proliferation and/or differentiation.

Confirming Knudson's model

A seminal study of retinoblastoma by Cavenee and colleagues in 1983 (PMID 6633649; see Further Reading) both proved Knudson's two-hit hypothesis and established the paradigm for laboratory investigations of TS genes over the next 20 years. Linkage studies in familial retinoblastoma had suggested that the susceptibility locus mapped to chromosome 13q14. Cavenee and colleagues used a combination of cytogenetics and genetic marker (enzyme and restriction fragment length polymorphism) studies to compare constitutional and tumor cells from eight patients with sporadic retinoblastoma. These comparisons revealed several cases in which the constitutional (blood) DNA was heterozygous for one or more markers from 13q but the tumor cells were apparently homozygous.

- In some cases the loss of heterozygosity affected markers all along the chromosome. In those cases the tumor cells had lost one copy of chromosome 13, presumably through mitotic nondisjunction.
- In one of those cases the karyotype showed multiple copies of the chromosome: evidently, following the loss, the remaining chromosome had been re-duplicated.
- In one patient the loss affected all markers distal from 13q14 but not those proximal (**Figure 19.9**). This suggested a mitotic recombination event (the first such described in humans), with subsequent segregation of both mutation-carrying chromatids into one daughter cell.
- In other patients there was no loss of heterozygosity at marker loci. Possible mechanisms for attaining homozygosity for the hypothesised mutant allele included point mutation or a small interstitial deletion on the wild-type chromosome.

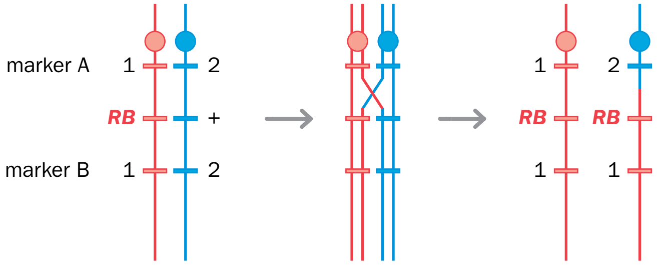


Figure 19.9 Mitotic recombination can reduce markers distal to the crossover point to homozygosity. One of the two copies of chromosome 13 carries a retinoblastoma "hit" (*RB*); the other has the wild-type allele (+). Genotypes for two markers are shown. When the cell replicates after the recombination event, one possible way the chromatids can segregate leads to a daughter cell homozygous for *RB* and all distal markers. Proximal markers remain heterozygous. This cell founds the tumor.

Loss of heterozygosity is only relevant in a cell that already has one mutant TS allele, so the researchers were seeing the second of Knudson's two "hits," which in these cases occurred through cytogenetic mechanisms. Later studies confirmed this interpretation by showing that in inherited cases it was always the wild-type allele that was lost in this way. Eventually, in 1986, the causative gene, *RB1*, was identified, and both first and second hits could then be identified by sequencing or looking for lack of gene expression. Retinoblastoma provides a good reminder that dominance and recessiveness are properties of phenotypes, not genes or alleles. The familial liability is dominant, due to a single mutant allele, but at the level of cell biology retinoblastoma is recessive, requiring loss of both functional alleles.

Implications of the retinoblastoma model

Apart from confirming Knudson's model, the work described above suggested two ways of finding tumor suppressor genes:

- Scanning tumors for losses of specific chromosomal material;
- Mapping and positional cloning of the genes responsible for familial cancers.

The first of these approaches had only limited success in the early days. Sometimes a loss of heterozygosity could be demonstrated by screening paired blood and tumor samples with a panel of polymorphic DNA markers (**Figure 19.10**), but often the results of genotyping were not helpful. Advanced tumors usually have many chromosomal aberrations (see **Figure 19.2**) and may show loss of heterozygosity at up to one-quarter of all loci. In addition, the samples used usually include nontumor stromal material, so that any loss observed was partial and quantitative rather than complete, making interpretation less certain. Now that it is routine to sequence the genomes of tumors and the corresponding blood samples, it is finally possible to document all losses of gene function, whether by chromosomal events, small deletions, or loss-of-function mutations.

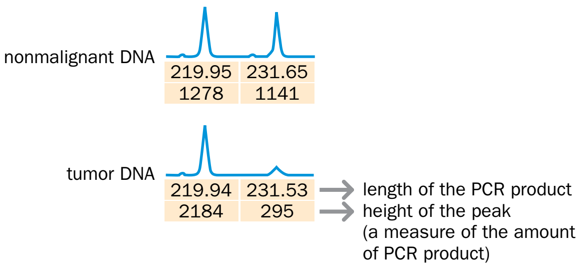


Figure 19.10 Loss of heterozygosity in a tumor. Electropherograms of a PCR-amplified marker. Upper trace: blood DNA from the patient is heterozygous for the marker (two peaks of equal size). Lower trace: tumor DNA shows almost total loss of the 231 bp allele. The small remaining amount is probably amplified from contaminating stromal (normal) tissue. (Courtesy of Lise Hansen, University of Aarhus.)

The second approach, mapping and positional cloning of genes involved in familial cancer syndromes, has been immensely successful in identifying important TS genes (**Table 19.5**).

TABLE 19.5 EXAMPLES OF RARE FAMILIAL CANCERS THAT ENABLED IDENTIFICATION OF TUMOR SUPPRESSOR GENES

Disease	OMIM #	Map location	TS gene
Familial adenomatous polyposis coli	175100	5q21	*APC*
Lynch syndrome I	120435	2p21	*MSH2*
	120436	3p21.3	*MLH1*
Breast–ovarian cancer	113705	17q21	*BRCA1*
Breast cancer (early onset)	600185	13q13.1	*BRCA2*
Li–Fraumeni syndrome	151623	17p13	*TP53*
Gorlin's basal cell nevus syndrome	109400	9q22.3	*PTCH1*
Ataxia telangiectasia	208900	11q22.3	*ATM*
Retinoblastoma	180200	13q14	*RB1*
Neurofibromatosis 1 (von Recklinghausen disease)	162200	17q11.2	*NF1*
Neurofibromatosis 2 (vestibular schwannomas)	101000	22q12.2	*NF2*
Familial melanoma	600160	9p21	*CDKN2A*
Von Hippel–Lindau syndrome	193300	3p25.3	*VHL*

Many tumor suppressor genes show more complicated patterns

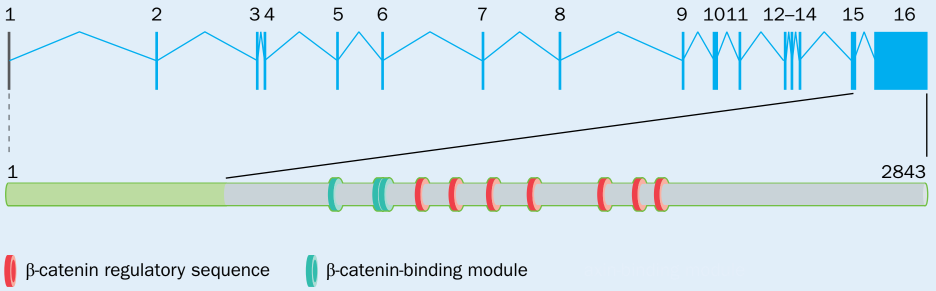
In the case of retinoblastoma all tumors, familial and sporadic, carry similar biallelic mutations in the same gene. Not all TS genes show such a clear-cut pattern.

- Even retinoblastoma is not quite as simple as our model in **Figure 19.8**. There, we assumed that the first and second hits occurred with equal probability μ, but in fact a number of the mechanisms (chromosome loss, mitotic recombination) are only applicable to producing the second hit in a cell that already carries one mutation.
- Some tumor suppressor genes show haploinsufficiency. Monoallelic loss-of-function mutations of the *PTEN* tumor suppressor gene are common in early tumors, without loss or mutation of the second allele. Biallelic loss is uncommon except in advanced cancers.
- *BRCA1* mutations are a frequent cause of familial breast cancer, but are seldom observed in sporadic cancers. A two-hit *BRCA1* mechanism still applies to a few sporadic breast tumors, particularly of the basal-like subtype, but the molecular mechanism of gene inactivation is different. In sporadic tumors, loss of BRCA1 function mainly occurs through methylation of the promoter (see below), often in combination with loss of heterozygosity. A similar difference is seen between *CDKN2A* and *MLH1* mutations in familial and sporadic forms of melanoma and colorectal tumors, respectively.
- Both alleles of the *APC* gene are commonly mutated in sporadic colorectal cancers, and one mutation is inherited in familial adenomatous polyposis coli, as in the classic TS model. However, the first and second hits are not independent: the type of second mutation depends on the type of the first (**Box 19.1**). Moreover, some patients have an attenuated form of the familial disease, with far fewer intestinal polyps and a later average age of onset of cancer. Cancer in these individuals follows a three-hit mechanism: they inherit a weak *APC* mutation, and tumorigenesis requires both inactivation of the remaining wild-type *APC* allele and a second mutation to convert the weak inherited allele into a stronger version.
- Schwannomas—benign tumors of the sheaths of nerves—illustrate further complexities. Vestibular schwannomas or acoustic neuromas (OMIM #101000) affect the acoustic nerve. They can be familial or sporadic and follow a classic two-hit mechanism, with either one inherited and one somatic *NF2* mutation (in familial cases) or two somatic *NF2* mutations in sporadic cases. In a related condition, schwannomatosis (OMIM #162091, #615670), the acoustic nerve is spared. Schwannomatosis can equally be familial or sporadic. Those tumors also lose function of both copies of the *NF2* gene, but even in familial cases the *NF2* mutations are always somatic, never inherited. The inherited mutation in familial cases is usually in either of two genes, *SMARCB1* or *LZTR1,* that lie close to the *NF2* gene on chromosome 22. Familial schwannomatosis follows a three-hit mechanism. After the inherited first hit, the second hit is a chromosomal event that removes a segment of chromosome carrying the wild-type *SMARCB1* or *LZTR1* allele, together with one of the (wild-type) *NF2* alleles. Then in a third hit, the remaining *NF2* allele is mutated. Separately, *SMARCB1* behaves as a classical TS gene in familial and sporadic rhabdoid tumors, while biallelic *LZTR1* driver mutations have been identified in some cases of the brain tumor glioblastoma multiforme. The COSMIC database of genes mutated in cancer (http://cancer.sanger.ac.uk/census) lists somatic *LZTR1* mutations in other cancers, but some loss-of-function mutations in *LZTR1* have also been found in control populations, suggesting reduced penetrance.

All known genes mutated in cancer are listed in the COSMIC database, mentioned above. Vogelstein and colleagues (2013, PMID 23539594; see Further Reading) identified only 74 high-penetrance TS genes. However, Davoli and colleagues identified 320 genes that could be described as TS genes on the basis that they had an excess of loss-of-function mutations in tumors, compared to the expectation of random passenger mutations (PMID 24183448; see Further Reading). Most of those have much weaker effects than the classical TS genes discussed above. Most of the potential TS genes on this extended list do not give rise to inherited familial cancers on the retinoblastoma model. This may be because of haploinsufficiency. If loss of a single copy already affects cell function, natural selection would ensure that constitutionally heterozygous people were rare or nonexistent; but haploinsufficiency would also mean that a single somatic mutation could have a role in tumorigenesis. Somatic loss of the second copy might well have a much more pronounced effect, like loss of both copies of haplo*sufficient* genes such as *RB1*.

BOX 19.1 REGULATION OF β-CATENIN LEVELS IN COLORECTAL TUMORS

The APC protein acts as a negative regulator of Wnt signaling by binding and down-regulating β-catenin. The Wnt signal releases β-catenin from an APC-containing destructive complex, allowing it to move to the cell nucleus where it stimulates transcription of growth-promoting genes including the genes for cyclin D1 and *MYC*. The large APC protein has three β-catenin-binding modules and seven 20-amino acid modules that down-regulate β-catenin levels (**Figure 1**). The gene has an unusual structure. The first 15 of the 16 exons encode only 653 of the 2843 amino acids; all the rest are encoded by the large 3′ exon 16. Unusually, APC genes with a premature stop signal in any codon downstream of codon 640 produce a truncated protein, because nonsense-mediated RNA decay does not occur for stop codons in or close to the last exon of a gene (see **Figure 16.7**). Thus, many protein-truncating mutations produce APC proteins that retain a limited ability to down-regulate β-catenin levels. Fodde and colleagues (2001, PMID 11900252; see Further Reading) showed that if the first *APC* mutation in a tumor produced a protein lacking all β-catenin regulatory modules, the second mutation would always leave the wild-type allele with one or two down-regulation modules. However, if the first hit resulted in a truncated protein that still contained one or two down-regulation modules, the second hit would remove all the modules from the wild-type allele. Thus, in all cases a double-mutant cell would retain a certain modest capacity to down-regulate β-catenin. Maybe complete loss of regulation would trigger cell death.



Box 19.1 Figure 1 Structure and function of the *APC* gene and protein. The function of the protein most relevant to colorectal cancer is its role in regulating levels of β-catenin. Exons 1–15 of the gene encode only the first 653 amino acids of the 2843-amino acid protein. Truncating mutations affecting the parts of the protein colored gray do not trigger nonsense-mediated RNA decay but allow production of a truncated protein. Each truncated protein typically retains 0–2 copies of the β-catenin regulatory sequence. Overall, colorectal cancer cells usually retain 1–2 copies of the β-catenin regulatory sequence, totaled between the first-hit and second-hit mutant proteins. (From Read A & Donnai D [2015] *New Clinical Genetics*, 3rd edn. With permission from Scion Publishing.)

MicroRNAs often act as tumor suppressor genes

As described above, most cancers show disturbed expression of microRNAs. MicroRNAs whose expression is down-regulated in tumor cells are behaving like TS genes. **Table 19.6** shows some examples.

TABLE 19.6 EXAMPLES OF microRNAs THAT ACT AS TUMOR SUPPRESSOR GENES

miRNA	Targets	Involvement in cancers
Let-7 family	*RAS*, *MYC*, *HMGA2*	Down-regulated in multiple solid tumors and hematopoietic malignancies
miR-15-16 cluster	CCND1, *WNT3A*	Translocated and down-regulated in hematopoietic malignancies; down-regulated in pituitary, prostate, and pancreatic tumors
miR-34 family	*CCNE2*, *MET*, *BCL2*, *MYCN*, *NOTCH1/2*, *CDK4/6*	Down-regulated in pancreatic cancer and Burkitt lymphoma
miR-203	*ABL*, *TP63*	Down-regulated in multiple solid tumors and hematopoietic malignancies

Data from Malumbres M (2013) *Mol Aspects Med* **34**:863–874; PMID 22771542.

Tumor suppressor genes are often silenced epigenetically by methylation

Tumor suppressor genes may be silenced by deletion (reflected in loss of heterozygosity) or by point mutations, but a very common third mechanism is methylation of the promoter. As described in Chapter 9, most DNA methylation affects cytosine nucleotides that lie adjacent to a guanine (CpG dinucleotides). CpG dinucleotides are relatively depleted in human DNA except in short “CpG islands” (see **Box 9.1**). In general, the DNA of tumors is hypomethylated compared to the corresponding normal cells; however, we frequently see methylation of normally unmethylated CpG islands in the promoters of certain genes. In normal cells these promoters are unmethylated, regardless of whether the gene is expressed or silent. Methylation in tumor cells prevents expression of the gene.

Specific genes differ in their susceptibility to this type of silencing (**Figure 19.11**). Some TS genes (for example, *MSH2*) are often silenced by mutation but never by methylation. For others (for example, *MLH1*), methylation occurs as a common alternative to point mutation. As noted above for *BRCA1*, methylation can be a mechanism of somatic but not inherited gene silencing since epigenetic changes like DNA methylation are not normally inherited across the generations. For some TS genes (for example, *RASSF1A* at 3p21, and *HIC1* at 17p13.3), methylation is the only known mechanism causing tumor-specific loss of function. MicroRNA genes also often have promoters containing CpG islands, and there are many examples of miRNAs that are specifically silenced in cancer cells by promoter methylation.

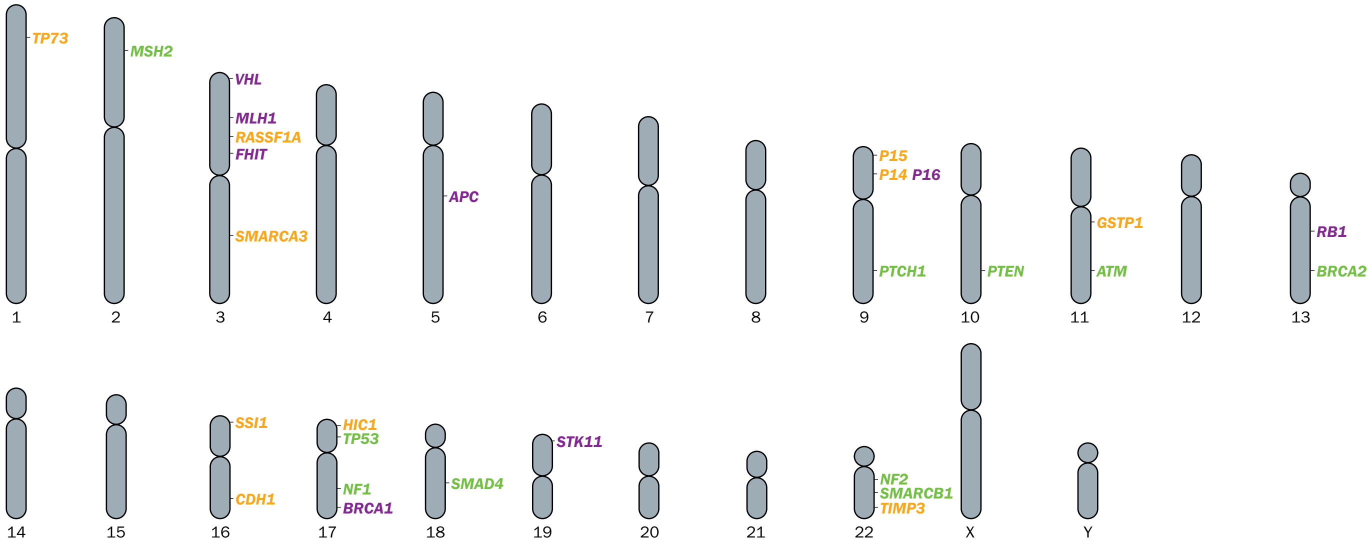


Figure 19.11 Different ways of silencing tumor suppressor genes. Genes shown in green are silenced only by mutation, and those in orange only by methylation, whereas those in purple may be silenced by either mechanism. (Reprinted from Jones PA & Baylin SB [2002] *Nat Rev Genet* **3**:415–428; PMID 12042769. With permission from Springer Nature. Copyright © 2002.)

19.3 KEY ONCOGENES AND TUMOR SUPPRESSOR GENES WORK MAINLY TO REGULATE CELL CYCLE CHECKPOINTS AND GENOME MAINTENANCE

Pathways involved in control of the cell cycle are among those most frequently affected in cancer because, one way or another, they must be derailed in every cancer cell. Control of cell cycling is critical for a properly functioning multicellular organism. Cells must only be allowed to proliferate when there is a need, either for growth or for tissue repair. As outlined in **Figure 3.7**, controls act through a number of so-called checkpoints; unless certain conditions are met, progress is halted at the relevant checkpoint. The G$_1$/S checkpoint is particularly crucial in cancer. Indeed, when the genes controlling the G$_1$/S checkpoint were identified, they turned out to involve a classic cast of the most frequently-mutated oncogenes and TS genes (**Figure 19.12**). Three of them deserve particular mention.

pRb: a key regulator of progression through G$_1$ phase

The *RB1* gene was identified through its role in retinoblastoma, as described above, but it is widely expressed and helps control cycling of all cells. The gene product, pRb, is a 110 kDa nuclear protein. Some cells contain two related proteins, p107 and p130, giving

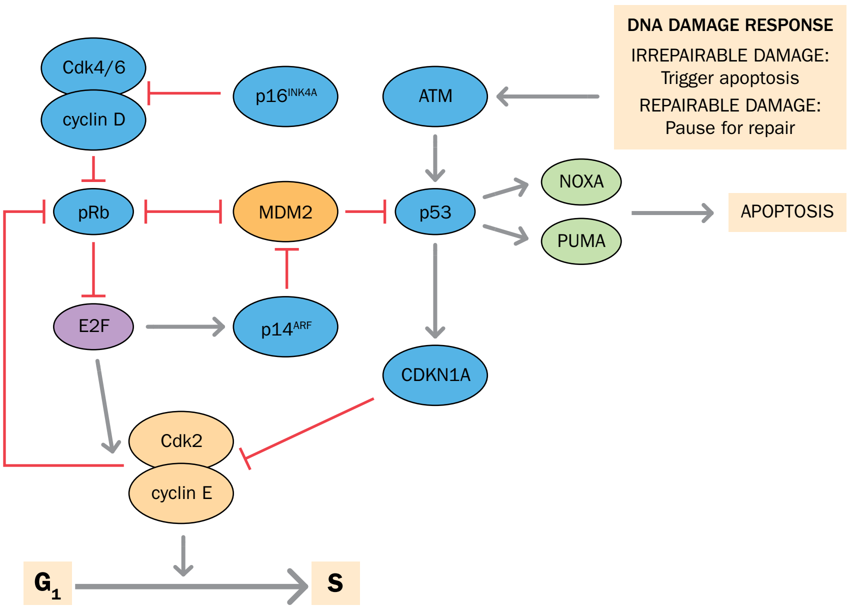


Figure 19.12 The G_1/S cell cycle checkpoint. Stimulatory actions are shown with gray arrows, inhibitory with red. Note that the mouse homolog of p14^{ARF} is called p19^{ARF}.

some redundancy in the pRb pathway. Lack of this redundancy in certain cells may help explain why a loss of *RB1* function results in very specific types of tumor. pRb binds and inactivates the cellular transcription factor E2F, function of which is required for cell cycle progression (see **Figure 19.12**). Two to four hours before a cell enters S phase, complexes of D cyclins and Cdk4 or Cdk6 phosphorylate pRb. This inactivates it, allowing E2F to become free. Once free, E2F stimulates the transcription of a variety of genes whose products are necessary for progression into S phase, including particularly cyclin E. In cells with loss-of-function mutations in *RB1*, E2F is inappropriately activated. Several viral oncoproteins (adenovirus E1A, SV40 T antigen, and human papillomavirus E7 protein) achieve the same result by binding and sequestering or degrading pRb, thus favoring cell cycle progression.

CDKN2A: one gene that encodes two key regulatory proteins

The remarkable *CDKN2A* gene at 9p21 uses alternative promoters and first exons to encode two structurally unrelated proteins (**Figure 19.13**). Exons 1α, 2, and 3 encode the p16^{INK4A} protein. This is an inhibitor of Cdk4/6 and hence serves to keep pRb in its active, dephosphorylated state. This in turn prevents E2F from stimulating the progression of the cell through G_1 toward the G_1/S boundary (see **Figure 19.12**). Thus, p16 is a tumor suppressor protein, whose loss allows inappropriate cell cycling.

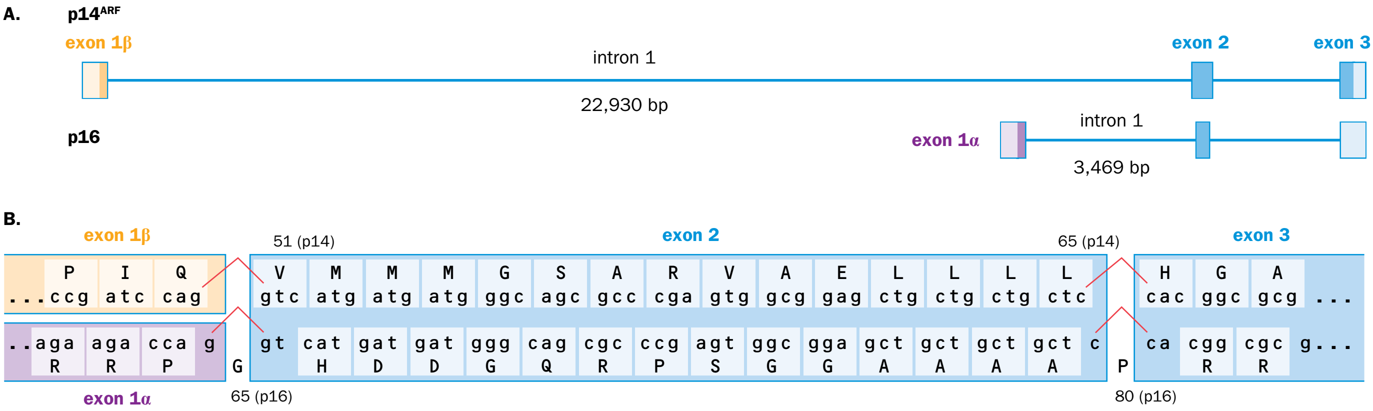


Figure 19.13 The two products of the *CDKN2A* gene. This gene (also known as *MTS* and *INK4A*) encodes two completely unrelated proteins. (**A**) p16^{INK4A} is translated from exons 1α, 2, and 3, and p14^{ARF} from exons 1β, 2, and 3. (**B**) Exons 2 and 3 are read in different reading frames in the two forms, so that the p16^{INK4A} and p14^{ARF} proteins have completely different amino acid sequences. The two gene products are both tumor suppressor proteins, active in the pRb and p53 arms of cell cycle control, respectively, as shown in **Figure 19.12**.

A second promoter starts transcription further upstream, at exon 1β. A second promoter starts transcription further upstream, at exon 1β. Like exon 1α, exon 1β includes the AUG translation initiation site. Exon 1β is spliced onto exons 2 and 3, but the reading frame is shifted so that an entirely unrelated protein, p14$^{\text{ARF}}$ (ARF for alternative reading frame; the mouse homolog is p19$^{\text{ARF}}$), is encoded. p14$^{\text{ARF}}$ mediates G$_1$ arrest by destabilizing MDM2, the oncoprotein responsible for keeping p53 levels low. Loss of p14$^{\text{ARF}}$ function leads to excessive levels of MDM2, excessive destruction of p53, and hence loss of cell cycle control.

Germ-line *CDKN2A* mutations, usually affecting just p16$^{\text{INK4A}}$, are seen in about 20% of families with multiple melanoma. Some of those families also have members with pancreatic cancer. Somatic inactivation of *CDKN2A* is very much more frequent. Deletions at 9p21, the location of the *CDKN2A* gene, are very frequent in a wide range of cancers. Tumor cells probably need to inactivate both the pRb and p53 arms of the G$_1$/S checkpoint for the cell to bypass the usual checks on cycling and to avoid triggering apoptosis. Homozygous deletion of the *CDKN2A* gene is an efficient way of achieving this and is a very common event in tumorigenesis. Some tumors have mutations that affect p16$^{\text{INK4A}}$ but not p14$^{\text{ARF}}$ (for example, specific inactivation of the exon 1α promoter by methylation). Those tumors tend also to have *TP53* mutations, showing the importance of inactivating both arms of the control system shown in **Figure 19.12**.

p53: the guardian of the genome

The p53 transcription factor, encoded by the *TP53* gene, has been called the “guardian of the genome” because of its central role in protecting the integrity of the genome. Normally, p53 levels in a cell are low. The MDM2 E3 ubiquitin ligase targets both pRb and p53 for degradation. *MDM2* is itself a transcriptional target of p53, so there is a negative-feedback loop that keeps p53 concentrations low. Signals from a whole range of cellular stress sensors lead to phosphorylation of p53. In particular, DNA double-strand breaks activate the ATM protein, which then phosphorylates p53 and a number of other proteins important in the DNA damage response. Phosphorylated p53 is no longer a substrate for MDM2, and hence the level of p53 in the cell rises. This triggers various protective mechanisms, in particular increased p53-dependent transcription of a variety of genes:

- CDKN1A (also called p21, WAF1, and CIP1) is an inhibitor of Cdk2 and hence of cell cycling (see **Figure 19.12**). Thus, stabilization of p53 leads to cell cycle arrest, giving the cell time to try to repair the DNA damage;
- *BBC3* (PUMA) and *PMAIP1* (NOXA) encode pro-apoptotic proteins. These proteins act through their BH3 protein–protein interaction domains to sequester antiapoptotic proteins of the Bcl-2 family, hence stimulating apoptosis. This is the fate of the cell if the DNA damage cannot be repaired.

p53 function can be lost by mutation or deletion of its gene *TP53*, or by overactivity of the MDM2 protein. Thus, *MDM2* functions as an oncogene; it is amplified in many sarcomas and de-repressed in cells with loss of function of pRb or p14$^{\text{ARF}}$ (see **Figure 19.12**). Tumor cells with absent or nonfunctional p53 may continue to replicate damaged DNA, leading to genomic instability, and may avoid apoptosis. Loss or mutation of *TP53* is probably the most frequent single genetic change in cancer. *TP53* maps to 17p13, and this is one of the most frequent regions of loss of heterozygosity in a wide range of tumors. Tumors that have not lost *TP53* very often have mutated versions of it. To complete the picture of *TP53* as a tumor suppressor gene, constitutional mutations in *TP53* are found in families with the dominantly inherited Li–Fraumeni syndrome (OMIM #151623). Affected family members suffer multiple primary tumors, typically including soft tissue sarcomas, osteosarcomas, tumors of the breast, brain, and adrenal cortex, and leukemia (**Figure 19.14**).

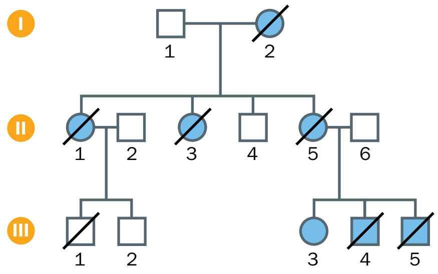


Figure 19.14 A typical pedigree of Li–Fraumeni syndrome. Malignancies typical of Li–Fraumeni syndrome include bilateral breast cancer diagnosed at age 40 in the grandmother (I$_2$). Her daughters were diagnosed with a brain tumor at age 35 (II$_1$), soft tissue sarcoma at age 19 and breast cancer at age 33 (II$_3$), and breast cancer at age 32 (II$_5$). The children of daughter II$_5$ suffered osteosarcoma at age 8 (III$_3$), leukemia at age 2 (III$_4$), and soft tissue sarcoma at age 3 (III$_5$). The grandfather (I$_1$) was diagnosed with colon cancer at age 59; this is assumed to have been coincidental and unrelated to the Li–Fraumeni syndrome. (From Malkin D [1994] *Annu Rev Genet* **28**:443–465; PMID 7893135. With permission from Annual Reviews. Permission conveyed through Copyright Clearance Center, Inc.)

Defects in DNA repair are potent causes of genomic instability

One consequence of the failure of cell cycle checkpoints is that cells with damaged DNA continue to progress through the cycle, rather than pausing while attempts are made to repair the damage. The resulting genomic instability is yet more significant when the DNA repair mechanisms are compromised. In Section 11.2 we described the variety of mechanisms that cells use to repair different types of DNA damage. Some of the proteins involved in one or another of these processes are frequent targets of loss-of-function mutations in cancer cells.

One of these mechanisms, the mismatch repair (MMR) system, is primarily concerned with correcting replication slippage. When a run of short tandem repeats such as a microsatellite or homopolymer run is replicated, if the polymerase temporarily dissociates from the template before re-associating and resuming synthesis, it can easily end up incorporating one or two more or fewer repeat units in the newly synthesized DNA compared to the template strand (see **Figure 11.1**). The MMR system corrects such errors. It uses the MSH2, MSH3, MSH6, MLH1, and PMS2 proteins (see **Figure 11.5**).

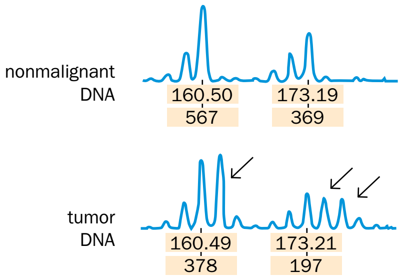


Figure 19.15 Microsatellite instability. Electropherograms of a PCR-amplified microsatellite marker. Upper trace: blood DNA; lower trace: tumor DNA. Both samples give secondary "stutter" peaks, an artifact of the PCR process, but the tumor sample has additional extra peaks (arrowed). (Courtesy of Lise Hansen, University of Aarhus.)

Tumors with defective MMR show instability of microsatellites and/or short homopolymer runs (**Figure 19.15**). The extra microsatellite alleles are probably harmless passenger mutations, being located in noncoding DNA, but they serve to identify tumors with defective MMR. The driver mutations are in other sequences such as the transforming growth factor β receptor. TGFβ is a strong inhibitor of cell proliferation, especially in the colorectum. It acts through a cell surface receptor, one subunit of which is the TGFBR2 protein. Exon 3 of the *TGFBR2* gene has a run of 10 consecutive A nucleotides (**Figure 19.16**). A survey of 111 colorectal tumors with microsatellite instability found somatic mutations in this homopolymer run in 100 of the tumors. These created frameshifts and loss of function.

743	TGC	ATT	ATG	AAG	GAA	AAA	AAA	AAG	CCT	GGT	GAG	ACT	TTC
120	Cys	Ile	Met	Lys	Glu	Lys	Lys	Lys	Pro	Gly	Glu	Thr	Phe

Figure 19.16 A homopolymer run in exon 3 of the *TGFRB2* gene. The run of 10 consecutive A nucleotides (red box) is the target of frequent loss-of-function frameshifting mutations in tumors with defective mismatch repair.

MMR defects are especially found in patients with early-onset colorectal cancer. Often, as with familial adenomatous polyposis coli (OMIM #175100), the cancers are familial, but they are not associated with intestinal polyps and were hence known as hereditary nonpolyposis colon cancer (HNPCC). The preferred name now is Lynch syndrome (OMIM #120435). Other associated cancers include endometrial and pancreatic tumors. Following the classic tumor suppressor model, family members inherit one loss-of-function mutation, and the tumors have a second, somatic mutation causing complete loss of function. As mentioned above, somatic but not germ-line mutations often involve methylation of the promoter. Rare individuals constitutionally homozygous for MMR mutations are at risk of a variety of cancers including colorectal and brain tumors (mismatch repair cancer syndrome or Turcot syndrome, OMIM #276300).

The breast cancer genes *BRCA1* and *BRCA2* also have important roles in genomic stability. As described in Section 11.2, double-strand DNA breaks are the most dangerous type of DNA damage—a single unrepaired double-strand break can result in cell death—and the most difficult to repair. In the G_1 stage of the cell cycle, the cell has to rely on the error-prone process of end-joining to survive double-strand breaks. Double-strand breaks arising in G2 phase can be repaired by the error-free homologous recombination process, using the undamaged sister chromatid as a template. BRCA1, BRCA2, and RAD51 proteins are essential for the homologous recombination pathway (**Figure 19.17**).

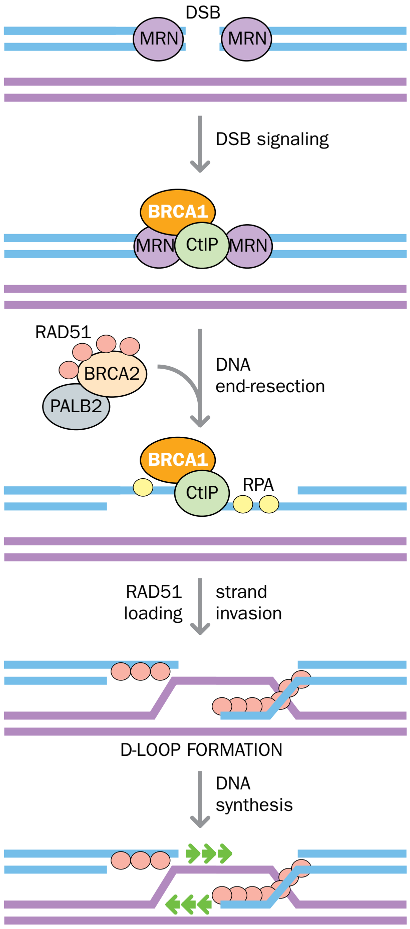


Figure 19.17 Roles of BRCA1 and BRCA2 proteins in repair of DNA double-strand breaks by homologous recombination in S/G_2 phase of the cell cycle. Double-strand breaks (DSB) are first detected by the MRE11–RAD50–NBS1 complex (MRN). This triggers a cascade of phosphorylation and ubiquitylation events (not shown) that promote the recruitment of BRCA1 and CtIP (RBBP8) to the break. RPA (replication protein A) binds single-stranded DNA in the vicinity, while CtIP promotes extensive DNA end-resection. Next, RAD51 is loaded onto the 3′ resected end by the concerted action of BRCA1, PALB2 (**p**artner **a**nd **l**ocalizer of **B**RCA**2**), and BRCA2. This initiates homology searching and invasion of the homologous template by the RAD51-coated DNA strand, forming a D-loop. Later stages of the process, involving re-synthesis and ligation, are shown in **Figure 11.6**. (Adapted from Fradet-Turcotte A *et al.* [2016] *Endocrine-related Cancer* **23**:T1–T17; PMID 27530658. Copyright: © 2016 Society for Endocrinology 2016. Permission conveyed through Copyright Clearance Center, Inc.)

During S phase, replication forks can be stalled by strand breaks, bulky adducts, or interstrand cross-links. Stalled forks can be rescued in various ways. Special DNA polymerases are capable of translesion synthesis, bypassing some obstacles (see **Table 11.2**), but these are low-fidelity, error-prone enzymes. BRCA1 and BRCA2 proteins are involved in alternative processes. Their combined action can stabilize forks and protect them from degradation by loading single-stranded DNA with RAD51. Interstrand cross-links present special problems, which are resolved by the Fanconi anemia complex of proteins that includes BRCA2. (See OMIM #227650 for an overview of Fanconi anemia, a clinically and genetically heterogeneous disorder characterized by a specific inability to repair interstrand cross-links.) Overall, a variety of different pathways are involved in maintaining genomic stability by repairing specific types of DNA damage, and many of the genes involved feature as TS genes with loss-of-function mutations in tumors and cancer-prone syndromes.

19.4 A GENOME-WIDE VIEW OF CANCER

The advent of next-generation sequencing allowed researchers to move from looking at changes in individual oncogenes and TS genes to genome-wide analyses of mutations. A number of large international collaborative projects are using the tools of genomics to try to understand how cancer cells acquire all the capabilities outlined at the start of this chapter (see **Figure 19.1**). For example, the International Cancer Genome Consortium (https://icgc.org/) aims “to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe”. Data release 27 (30th April 2018) comprised data from more than 20,000 cases involving 22 tumor sites collected through 84 projects in 17 countries (**Table 19.7**).

TABLE 19.7 RESEARCH IN CANCER GENOMICS PROCEEDS THROUGH LARGE INTERNATIONAL COLLABORATIVE PROJECTS

Location of research group	Types of cancer investigated
Australia	Melanoma, ovary, pancreas
Brazil	Melanoma
Canada	Pancreas, pediatric medulloblastoma, prostate
China	Bladder, breast, colorectum, esophagus, glioblastoma, kidney, liver, lung, nasopharynx, ovary, pancreas, prostate, stomach, thyroid
European Union	Breast, kidney
France	Bone, breast, eye, leiomyosarcoma, liver, lymphoproliferative syndrome, pancreas, uterus
Germany	Malignant lymphoma, medulloblastoma, prostate
India	Oral
Italy	Pancreas
Japan	Biliary tract, liver, stomach
Mexico	Breast, head and neck, non-Hodgkin lymphoma
Saudi Arabia	Thyroid
Singapore	Biliary tract, lymphoma
South Korea	Acute myeloid leukemia, breast, lung
Spain	Chronic lymphocytic leukemia
UK	Bone, breast, chronic myeloid disorders, esophagus, prostate
USA	Bladder, bone, breast, cervix, colon, endometrium, gastric, head and neck, kidney, leukemia, liposarcoma, liver, lung, lymphoma, melanoma, ovary, pancreas, pediatric brain and solid tumors, prostate, rectum

The table lists current projects by the International Cancer Genome Consortium, https://icgc.org/.

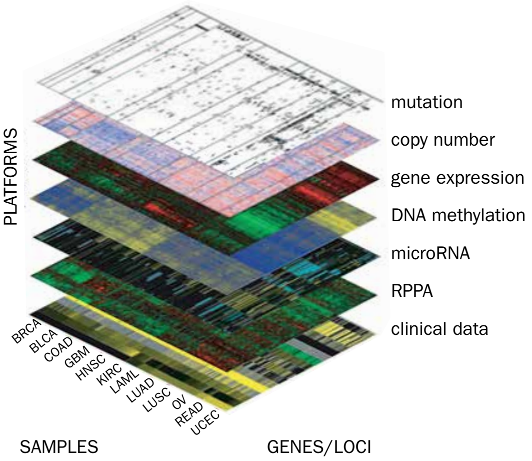


Figure 19.18 A multiplatform, cross-tumor analysis of cancer. Multiple genomic analyses were used on 5074 tumor samples covering 12 types of cancer (BRCA, breast carcinoma; BLCA, bladder carcinoma; COAD, colon adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous carcinoma; KIRC, kidney renal clear-cell carcinoma; LAML, lymphoblastic acute myeloid leukemia; LUAD, lung adenocarcinoma; LUSC, lung squamous carcinoma; OV, ovarian carcinoma; READ, rectal adenocarcinoma; UCEC, uterine cervical and endometrial carcinoma). RPPA is a reverse-phase protein array of antibodies to 131 proteins. Results of analysis of an initial 3527 tumors are shown in **Table 19.8**. (Reprinted from Cancer Genome Atlas Research Network [2013] *Nat Genet* **45**:1113–1120; PMID 24071849. With permission from Springer Nature. Copyright © 2013.)

Multiplatform analyses describe changes at the chromosomal, DNA sequence, epigenetic, RNA, and protein levels

Cells of a tumor can be compared with the normal constitutional cells of the patient on many levels to provide an integrated picture of carcinogenesis. Exome sequencing would reveal all coding-sequence changes, while whole-genome sequencing could document structural variants and copy number changes, in addition to small changes in noncoding sequence. To characterize the transcriptome, early studies used expression arrays, but RNA-Seq is now preferred because it allows all transcripts to be identified, not just those featured on an expression array, and also has a greater dynamic range. DNA methylation can be documented by whole-genome bisulfite sequencing (see **Box 10.3**). Changes in the proteome could be followed by mass spectrometry or antibody-based methods. **Figure 19.18** shows the multiplatform approach used in one large collaborative study, The Cancer Genome Atlas (TCGA). Even this battery of analyses cannot fully characterize a tumor. To fully follow the mutational trajectory, it will be necessary to follow the development of a single tumor over time, to use single-cell technologies to document the differences between different subclones of cancer cells, and to identify the various roles of stromal cells.

Circos plots and heat maps are attempts to present the masses of genomic data in visually digestible forms

Genomic analysis of even a single tumor produces a huge volume and variety of data, which is not only a challenge to process and store, but also a challenge to present to a human reader. Much ingenuity has been devoted to thinking up ways of presenting large volumes of data graphically so as to give scope for the natural pattern-detecting ability of the human eye and brain. One useful tool is the Circos plot (**Figure 19.19A**). This is a way of giving an overall visual impression of the DNA-level changes in a single tumor or class of tumors. Circos plots are useful for highlighting the relative importance of different types of variant in different tumors. **Figure 19.19B** shows a pioneering example of a **heat map**. Heat maps can be used to display various types of quantitative data but are

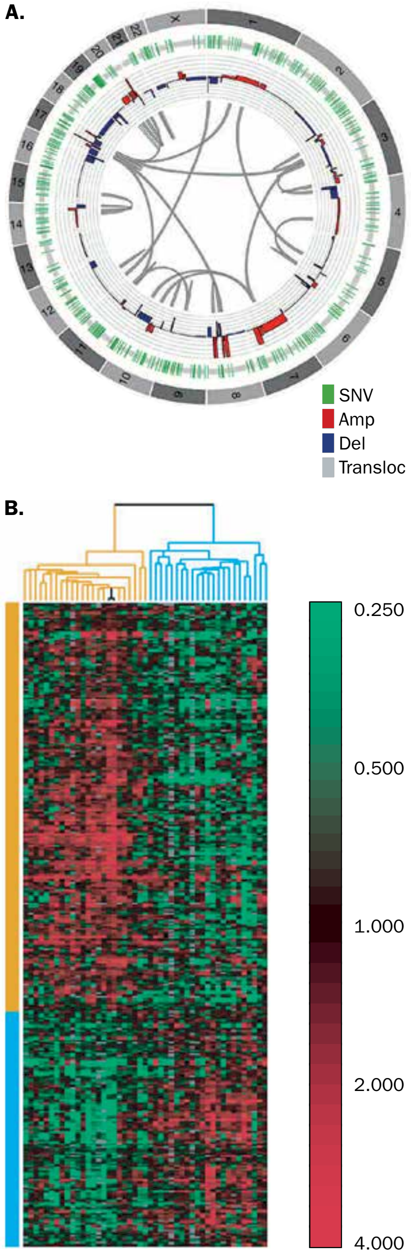


Figure 19.19 Representing complex data. (**A**) A Circos plot showing variants in the SK-BR3 breast cancer cell line. The outermost ring shows the chromosomes. Inner rings show the positions and numbers of single nucleotide variants (SNV), amplifications (Amp), and deletions (Del). Connecting lines across the interior show chromosomal translocations (Transloc). (**B**) A heat map showing hierarchical clustering of gene expression profiles in diffuse large B-cell lymphomas. Each row shows aggregate data for one cDNA, and each column shows total data from mRNA from one tumor. The scale at right shows the color-coding of hybridization intensity relative to a reference mRNA. The tumors can be clearly seen to form two groups, shown by the orange and blue bars, and designated germinal center and activated B-like types, respectively. The distinction is biologically significant because 5-year survival in the two groups was 76% and 16%, respectively. (A, reprinted from Wang Y *et al*. [2014] *Nature* **512**:155–160; PMID 25079324. With permission from Springer Nature. Copyright © 2014; B, reprinted from Alizadeh AA *et al*. [2000] *Nature* **403**:503–511; PMID 10676951. With permission from Springer Nature. Copyright © 2000.)

particularly useful for comparing patterns of gene expression between different tumors. Typically, data on expression levels of a hundred or so genes are shown in each of a large set of tumors, compared to the corresponding normal tissue. Hierarchical clustering methods (see **Box 7.7**) are used to display data from tumors with similar patterns side-by-side. This allows our natural pattern-seeking ability to pick out sets of tumors that have similar overall patterns of disordered gene expression.

Mutational signatures suggest the main mechanisms generating somatic mutations in a tumor

Different mutational processes produce characteristic signatures. For example, there are six possible single nucleotide substitutions (C>A, C>G, C>T, T>A, T>C, T>G, together with their complements on the opposite strand). With four choices for each immediately upstream and downstream nucleotide, there are 96 possible single nucleotide changes affecting a triplet. Analysis of all point mutations (passengers as well as drivers) across panels of tumors can identify the signatures typical of a particular cancer type (**Figure 19.20**). In some cases the mutagenic agent can be identified. Hopefully, identifying the agents responsible for mutations in a tumor will aid in understanding the overall evolutionary process producing the tumor.

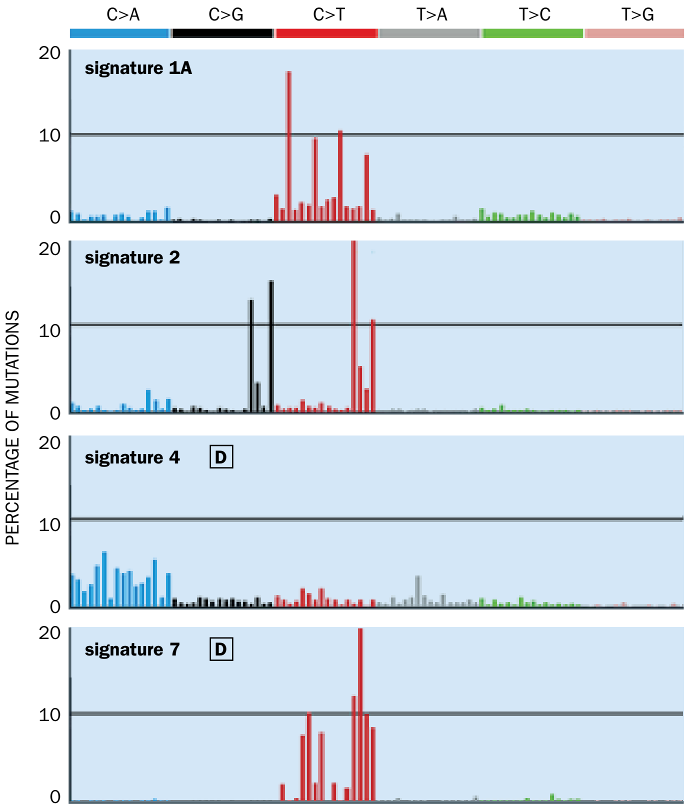


Figure 19.20 Mutational signatures. For each of the changes shown across the top, there are 16 possible contexts, considering just the immediately upstream and immediately downstream nucleotides. Signature 1A seems to reflect random lifelong deamination of cytosine, and is seen in a wide variety of tumor types. Signature 2, also seen in a variety of tumors, is due to uncontrolled activity of the APOBEC family of cytidine deaminases. Signature 4 is smoking-related and seen in head and neck, liver, and lung cancer; signature 7 is restricted to melanomas and shows the mutagenic action of ultraviolet light. D indicates dinucleotide mutations are present. (Adapted from Alexandrov LB *et al.* [2013] *Nature* **500**:415–421; PMID 23945592. With permission from Springer Nature. Copyright © 2013.)

Genomic data allow a new classification of tumors

Tumors are traditionally classified by their tissue of origin, histological appearance, and sometimes other features such as hormone dependence. The B-cell lymphoma example shown in **Figure 19.19B** was an early example of the use of gene expression signatures to subclassify tumors. Two more-recent examples, chosen from among many, are shown below.

- Breast cancers can be classified in a number of ways. A very common scheme is based on the presence or absence of estrogen (ER) and progesterone (PR) receptors and ERBB2 (HER2) amplification. Incorporating gene expression profiles shows that most tumors can be grouped into luminal A, luminal B, and ERBB2-amplified subtypes, but a set of "triple negative" (ER–, PR–, ERBB2–) tumors form a "basal" group with completely different biology. Curtis and colleagues (2012, PMID 22522925; see Further Reading) proposed 12 subgroups based on analysis of 2000 tumors. The different categories have different prognoses, and commercial kits (MammaPrint®, Oncotype DX®, and so on) are available to identify high-risk and low-risk expression signatures.
- Pancreatic cancers have been divided by expression analysis into four subtypes: squamous, pancreatic progenitor, immunogenic, and aberrantly-differentiated endocrine exocrine (ADEX). The subtypes correlate with histopathological characteristics, and tumors of the squamous subtype show significantly worse prognosis than the others (see Bailey and colleagues, 2016 [PMID 26909576], in Further Reading).

The multiplatform pan-cancer analysis illustrated in **Figure 19.18** identified 13 clusters (**Table 19.8**), of which 11 had prognostic value (two contained too few examples to be useful). Five of the clusters showed simple, near one-to-one relationships with the tissue of origin (the cancers labeled GBM, KIRC, LAML, OV, and UCEC in **Figure 19.18**), but other cancer types split between several clusters, and some clusters united cancers of different tissue origins (**Table 19.8**). For example, almost all breast cancers (BRCA) fell into either Cluster 3 or Cluster 4, and each of those clusters encompassed only breast cancers. All READ tumors fell into Cluster 7, but so did all COAD tumors. On the other hand, bladder cancers (BLCA) split between Clusters 1, 2, and 8, of which only Cluster 8 was specific to bladder cancer. Based on this study, one in ten cancer patients would be classified differently by this new molecular taxonomy versus the traditional tissue-of-origin tumor classification system.

TABLE 19.8 USING THE MOLECULAR DATA OF FIGURE 19.18 TO CLUSTER CANCERS

Handle	C1-LUAD-enriched	C2-squamous-like	C3-BRCA/luminal	C4-BRCA/basal	C5-KIRC	C6-UCEC	C7-COAD/READ	C8-BLCA	C9-OV	C10-GBM	C11-small-various	C12-small-various	C13-LAML	Total
BLCA	10	31	0	0	1	0	0	74	0	1	1	2	0	120
BRCA	2	1	688	135	5	0	0	2	0	0	0	0	1	834
COAD	0	0	0	0	0	0	182	0	0	0	0	0	0	182
GBM	3	0	0	0	2	0	0	0	0	190	0	0	0	195
HNSC	1	302	0	0	0	0	0	1	0	1	0	0	0	305
KIRC	1	0	0	0	470	0	0	0	0	2	0	2	0	475
LAML	0	0	0	0	0	0	0	0	0	0	0	0	161	161
LUAD	258	6	0	1	0	1	0	1	0	1	0	2	0	270
LUSC	28	206	0	1	0	0	0	1	0	2	0	0	0	238
OV	1	0	0	0	1	0	0	0	327	0	0	0	0	329
READ	0	0	0	0	0	0	73	0	0	0	0	0	0	73
UCEC	2	0	0	0	0	340	1	0	0	0	2	0	0	345
Totals	306	546	688	137	479	341	256	79	327	197	3	6	162	3527

Molecular features of the 12 types of cancer (see caption to **Figure 19.18** for a key to the "handles") could be analyzed into 13 clusters. Clusters 11 and 12 contained too few samples for further analysis; the remaining 11 clusters had prognostic value independent of the tissue of origin, and in some cases suggested possible treatments. (Reproduced from Hoadley KA *et al*. [2014] *Cell* **158**:929–944; PMID 25109877. With permission from Elsevier.)

Thinking in terms of pathways rather than individual mutations or genes reduces the complexity, and suggests links to Hanahan & Weinberg's list of capabilities

Figure 19.21 shows genes whose mutation frequency differs significantly between the clusters defined by Hoadley and colleagues (PMID 25109877) in **Table 19.8**. The genes are from a list of 291 high-confidence cancer drivers. The overwhelming impression is of great heterogeneity. In addition, not shown in the figure, there were many copy number variants that differed between and within clusters, and within each cluster there was also extensive heterogeneity of mutations.

This analysis has already simplified the picture by considering mutated genes rather than individual mutations. A further major simplification can be achieved by thinking in terms of altered pathways rather than altered genes. For example, **Figure 19.22** shows some of the considerable heterogeneity at the gene level of mutations in glioblastoma multiforme tumors, all of which fall in Cluster 10 of **Table 19.8**. Regardless which of the 14 genes in the figure is mutated, the effect is on just two pathways: the p53-driven pathway producing cell senescence and apoptosis, and the RAS/PI(3)K pathway by which receptor tyrosine kinases affect cell proliferation and survival. Thinking in terms of compromised pathways rather than individual mutations or genes connects fruitfully with the overview of cancer in terms of six capabilities and two "enabling characteristics" outlined at the start of this chapter.

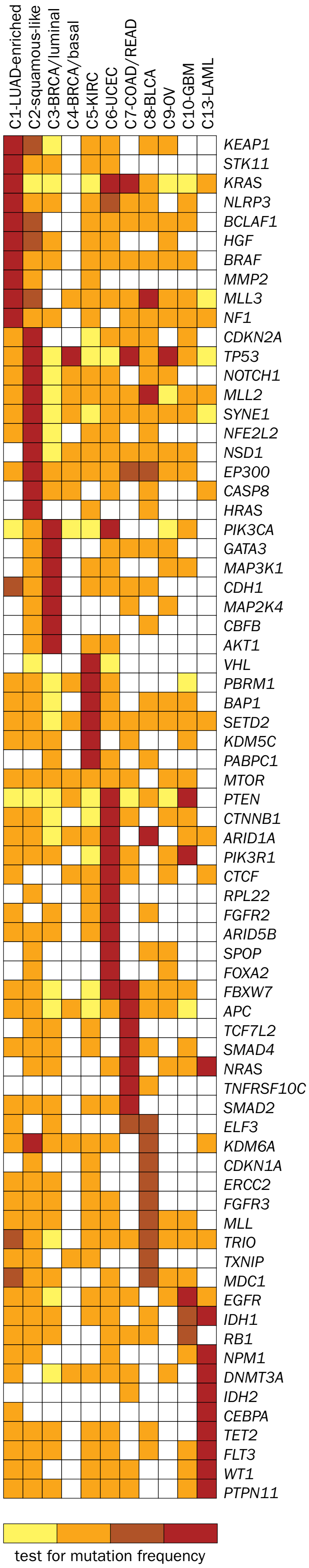


Figure 19.21 Genes frequently mutated in each of the 11 main clusters shown in Table 19.8. Orange, brown, and red squares indicate genes mutated to increasing extents above the frequency averaged across all clusters, as shown in the key. Yellow squares mark genes mutated at background frequency, and white squares mark genes not recorded as mutated in the sample. (Reproduced from Hoadley KA *et al*. [2014] *Cell* **158**:929–944; PMID 25109877. With permission from Elsevier.)

Gene ontology enrichment analysis allows a systematic ascertainment of pathways affected in tumors

The Gene Ontology (GO) database (www.geneontology.org/) classifies all genes across a wide range of organisms in three ways: by molecular function, biological process, and cellular component. The large datasets of mutations produced by whole-genome sequencing of tumors can be checked for enrichment of specific GO terms (http://geneontology.org/page/go-enrichment-analysis). Given a set of genes that are upregulated in a certain tumor type, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set. This provides an unbiased way of identifying target pathways in tumorigenesis. Similar analyses can be performed using the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/), which focuses on molecular interaction networks.

Vogelstein and colleagues, in an overview of molecular events in cancer (2013, PMID 23539594; see Further Reading), identified 64 high-penetrance oncogenes and 74 high-penetrance TS genes, but suggest they all act through one or more of just 12 pathways (**Figure 19.23**).

Genome-wide approaches can identify unsuspected pathways

An unexpected finding from sequencing tumor genomes was the discovery of frequent mutations in *IDH1*. This gene encodes isocitrate dehydrogenase 1, an enzyme of the tricarboxylic acid cycle. IDH1 catalyzes the oxidative decarboxylation of isocitrate to produce α-ketoglutarate, and has no obvious connection to cancer. Yet more than 70% of grade II and III astrocytomas and oligodendrogliomas, and the glioblastomas that develop from these lower-grade lesions, have *IDH1* missense mutations, and evidence suggests these are early events in tumorigenesis. The mutations are very specific: they always affect arginine 132, replacing it with histidine or sometimes serine. Tumors without mutations in *IDH1* often have mutations affecting the homologous amino acid (R172) of the *IDH2* gene. Mutations in R172 of IDH2 are also frequent in acute myelogenous leukemia. The high specificity of the mutations, always changing the same amino acid, and the fact that the effect is present in heterozygotes, show that this must be a gain-of-function effect. But what is the new function?

The first clue emerged when it was shown that the mutant enzyme produced a novel metabolite, 2-hydroxyglutarate (**Figure 19.24**). The likely pathogenic action of 2-hydroxyglutarate is interference with levels of DNA methylation.

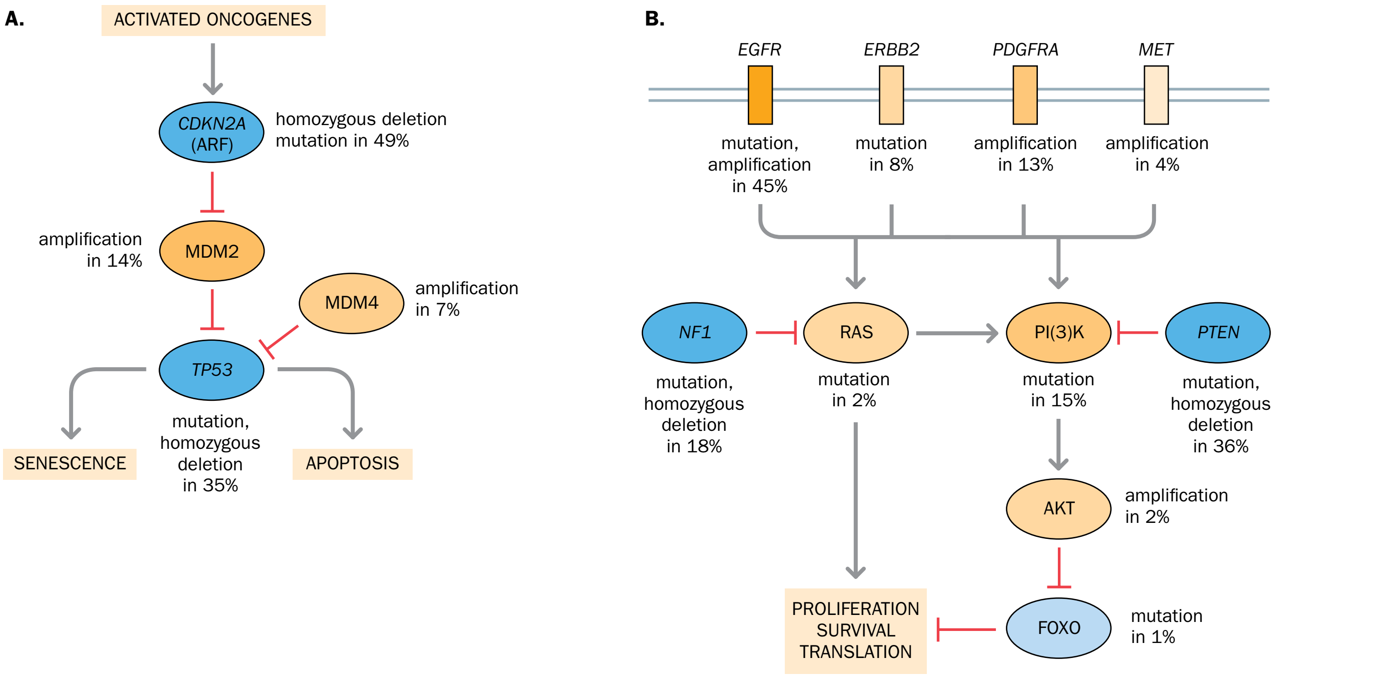


Figure 19.22 Two pathways targeted by frequent mutations in glioblastoma multiforme. In different individual tumors the two pathways are compromised by mutations or copy number changes in a variety of different genes. Some (shown in yellow) act as oncogenes, with activating changes; others, shown in blue, act as tumor suppressor genes with loss-of-function changes; but all effects converge on the two pathways. Frequently altered genes are shown by deeper shades of color. (**A**) 87% of tumors have variants that affect the pathways through which p53 triggers senescence and apoptosis. (**B**) 88% of tumors have genetic changes that affect the pathways by which receptor tyrosine kinases (EGFR, ERBB2, PDGFRA, and MET) acting though RAS and phosphoinositide 3-kinase (PI(3)K) control cell proliferation and survival. (Data from Cancer Genome Atlas Research Network [2008] *Nature* **455**:1061–1068; PMID 18772890.)

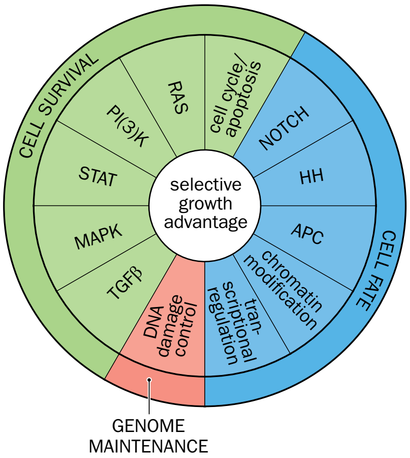


Figure 19.23 Twelve cell signaling pathways that are targets for oncogenes and tumor suppressor genes. (From Vogelstein B *et al.* [2013] *Science* **339**:1546–1558; PMID 23539594. Reprinted with permission from the AAAS.)

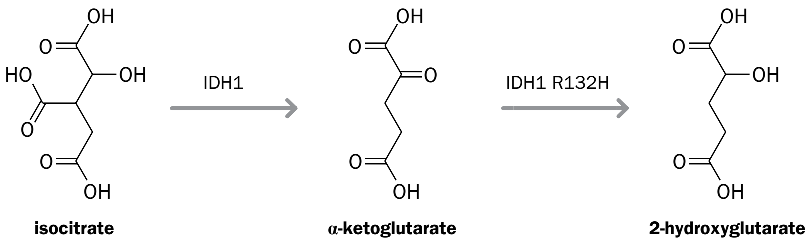


Figure 19.24 Mutant isocitrate dehydrogenase produces the toxic metabolite 2-hydroxyglutarate.

We saw in Chapter 10 how methylation of cytosine in CpG dinucleotides is reversed indirectly through the action of TET dioxygenase enzymes that convert 5-methylcytosine to 5-hydroxymethylcytosine (see **Figure 10.7**). TET enzymes use α-ketoglutarate as a co-factor and are inhibited by 2-hydroxyglutarate. Thus, in cells with the mutant IDH enzyme, DNA is hypermethylated. Hypermethylation of specific sequences can affect chromatin structure and gene expression. Chromosomes are partitioned into megabase-sized topologically-associated domains (TADs). Enhancers may control expression of genes located in the same TAD but have no influence on genes outside their TAD. As discussed in Section 10.2, boundaries of TADs are defined by insulator sequences that bind the CTCF protein. **Figure 10.17** showed how insulator action can be affected by methylation of the DNA. Methylation prevents binding of CTCF, abolishing insulator function and allowing enhancers to affect expression of genes beyond their own TAD. It seems likely that 2-hydroxyglutarate affects patterns of gene expression by inhibiting demethylation of insulator DNA. Critical insulator sequences remain methylated and hence ineffective. Flavahan and colleagues (2016, PMID 26700815; see Further Reading) presented evidence that one effect of 2-hydroxyglutarate-dependent DNA hypermethylation is to bring *PDGFRA* (platelet-derived growth factor receptor A, an oncogene frequently amplified in gliomas) under the influence of a powerful neural-specific enhancer in a neighboring TAD. This epigenetic mechanism suggests an extra way of activating an oncogene, in addition to the four listed in **Table 19.2**. A report by Hnisz and colleagues (2016, PMID 26940867; see Further Reading) suggests that microdeletions or point mutations affecting TAD boundaries may activate oncogenes in other cancers.

Some tumors show evidence of large-scale co-ordinated genetic changes

Although most tumors evolve gradually, slowly acquiring mutations and developing through successive histological stages, sometimes events move more quickly. Occasionally a single event can generate large numbers of mutations. A number of different mechanisms can do this.

- **Breakage–fusion–bridge cycles** are a classic consequence of rearrangements that produce dicentric chromosomes. They were originally described in the 1930s by Barbara McClintock. Broken chromosome ends, or ends lacking telomeres, are joined by nonhomologous end-joining (**Table 11.1**). When ends from two different chromosomes are joined, producing a translocation, the result may be two monocentric chromosomes, or it may be a dicentric and an acentric (see **Figure 15.10**). Acentric chromosomes fail to segregate at mitosis and are lost (or incorporated into micronuclei, see below). Dicentric chromosomes may be pulled in opposite directions by the two centromeres, forming a bridge. Eventually the bridge breaks. The resulting monocentric fragments can segregate to the daughter nuclei, but they have broken ends. Repairing the ends produces a new fusion chromosome that may once again be dicentric. Thus, the original translocation can generate repeated cycles of breakage and fusion, producing multiple chromosome rearrangements.
- In **chromoplexy** a tumor cell has multiple "chained" chromosomal rearrangements. For example, if a translocation forms between chromosomes A and B, rather than the remaining broken ends being joined together to form a conventional reciprocal translocation, they may form fresh translocations with chromosomes C and D. This generates another pair of broken ends that may go on to form further rearrangements, and so on. Baca and colleagues and Shen show examples and discuss possible mechanisms (2013, PMID 23622249, 23680143; see Further Reading).
- **Chromothripsis** is seen when a single chromosome shows tens to hundreds of rearrangements (Stephens *et al*. [2011] PMID 21215367, Kloosterman [2015] PMID 26068832; see Further Reading). It involves fewer chromosomes than chromoplexy, but more rearrangements. Often there is a complete mix of deletions, duplications, and inversions. Sometimes the whole chromosome is affected, in other cases the changes are limited to a small region. Chromothripsis is seen in 2–3% of most cancers. Bone cancers show a particularly high frequency of chromothripsis, and in those tumors several different small chromosomal regions may be involved in complex intrachromosomal and interchromosomal changes.

  One possible mechanism involves anaphase lag. If a chromatid fails to attach correctly to the spindle at metaphase of mitosis it will not get incorporated into the daughter-cell nucleus. Normally such a chromatid would be degraded and lost, but alternatively it may get incorporated into a micronucleus. DNA replication in micronuclei is inefficient, so that when mitotic cyclin-dependent kinases compact the nuclear chromosomes, the micronucleus chromosome may still be only partially replicated. This premature chromosome compaction leads to pulverization of the chromosome. DNA repair systems reassemble the fragments in random order. During a subsequent mitosis the micronucleus chromosome may rejoin the normal chromosomes and be incorporated into a daughter-cell nucleus.

- **Kataegis** produces large numbers of mutations clustered in kilobase- to megabase-sized regions in a single event. The likely cause is the action of APOBEC cytidine deaminases on single-stranded DNA exposed as a result of double-strand breaks or collapsed replication forks (see **Figure 19.17**). Setlur and Lee (2012) provide a commentary (PMID 22632962; see Further Reading).

Each of these processes produces multiple random changes that most usually result in cell death. Cells with a compromised p53 pathway may be better able to survive. If, by chance, one of the events activates an oncogene or inactivates a tumor suppressor gene, the cell may obtain a growth advantage.

The evolution of a tumor can be inferred from comparative analyses or followed directly by single-cell studies

A resected colon from a patient with familial adenomatous polyposis will typically contain many polyps showing, between them, all stages of development from normal but slightly hypertrophic epithelium through to full-blown carcinoma. Many years ago Kinzler and Vogelstein (1996, PMID 8861899; see Further Reading) took advantage of this to explore the sequence of events underlying the tumorigenesis. Within a framework of overall heterogeneity, they were able to identify certain changes that were typical of the early stages and others that normally appeared later (**Figure 19.25**). In 2015 Drost and colleagues (PMID 25924068; see Further Reading) were able to model this progression in an *in vitro* system. Using three-dimensional cultures of intestinal crypt stem cells (organoids) they showed that mutation of just four genes, *APC*, *KRAS*, *SMAD4*, and *TP53*, enabled the organoids to grow independently of all stem cell niche factors, and to produce invasive carcinomas when transplanted into mice.

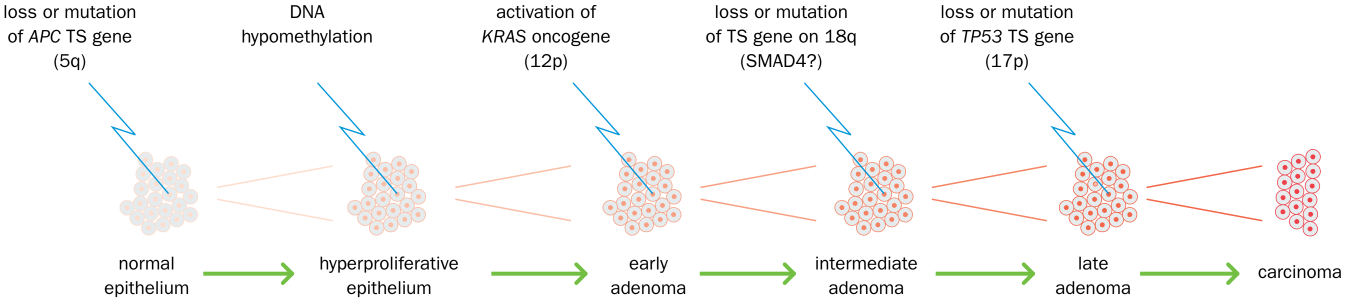

loss or mutation of *APC* TS gene (5q)
DNA hypomethylation
activation of *KRAS* oncogene (12p)
loss or mutation of TS gene on 18q (SMAD4?)
loss or mutation of *TP53* TS gene (17p)
normal epithelium
hyperproliferative epithelium
early adenoma
intermediate adenoma
late adenoma
carcinoma

Figure 19.25 A model for the multistep development of colon cancer. In many tumors loss, activation, or mutation of certain genes is seen at particular histological stages. This is primarily a tool for thinking about how tumors develop, rather than a firm description. Every colorectal cancer is likely to have developed through the same histological stages, but the underlying genetic changes are more varied.

Comparing early-stage tumors with pre-cancerous lesions on the one hand and late-stage tumors on the other, as in **Figure 19.25**, can shed light on the general evolution of a class of tumors, but a deeper understanding requires the ability to characterize diverse cells within a single tumor. It would be a great mistake to think of a tumor as a clonal colony of cells like a bacterial colony on a plate. **Figure 19.3** emphasized the cellular heterogeneity of tumors. Genomic instability and the consequent high mutation rate are characteristic of the great majority of cancers, and so a single tumor will contain heterogeneous populations of cells related by branching mutational trajectories. Identifying these requires the ability to characterize single cells within a tumor.

In Section 7.4 we outlined the technical advances that have made single-cell genomics and transcriptomics possible, and described some of the applications. Cancer research, as briefly mentioned there, is a major application. **Figure 19.26** illustrates the many ways in which single-cell studies are informing cancer research.

Beyond simply documenting the heterogeneity of cells within a tumor, single-cell sequencing can be used to map the clonal evolution of cell populations within a tumor. The heterogeneous cells can be assembled into phylogenetic lineages that identify the nature and sequence of driver mutations. In leukemia, where the mutational landscape is much simpler than in most solid cancers, it is possible to reconstruct the evolution by studying the clonality of mutations in a single blood sample. If driver mutation A is found in all leukemic cells of a patient, but driver mutation B is present only in a subset of cells, it must follow that mutation A appeared first. **Figure 19.27** shows the result of such an analysis. Similar studies in solid tumors confirmed that they evolved from a single somatic cell in the normal tissue of a patient.

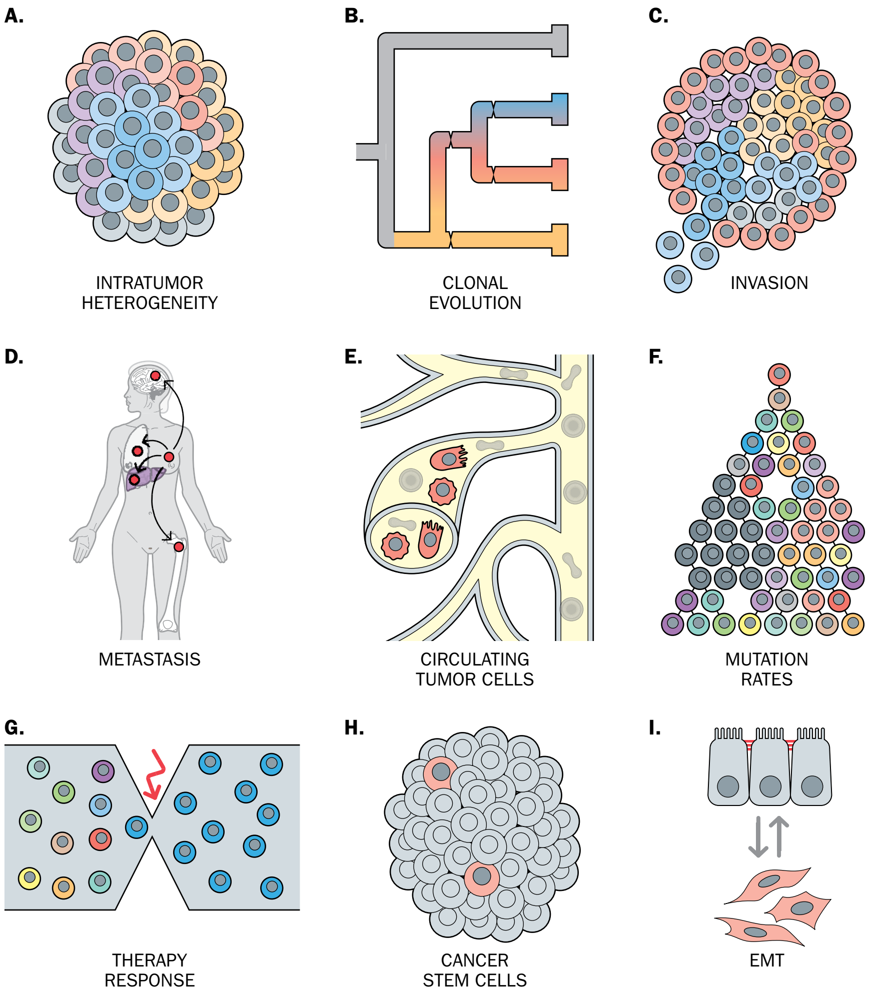


Figure 19.26 Applications of single-cell genomics and transcriptomics in cancer research. EMT, epithelial-mesenchymal transition. (Reproduced with permission from Navin NE [2015] *Genome Res* **25**:1499–1507; PMID 26430160.)

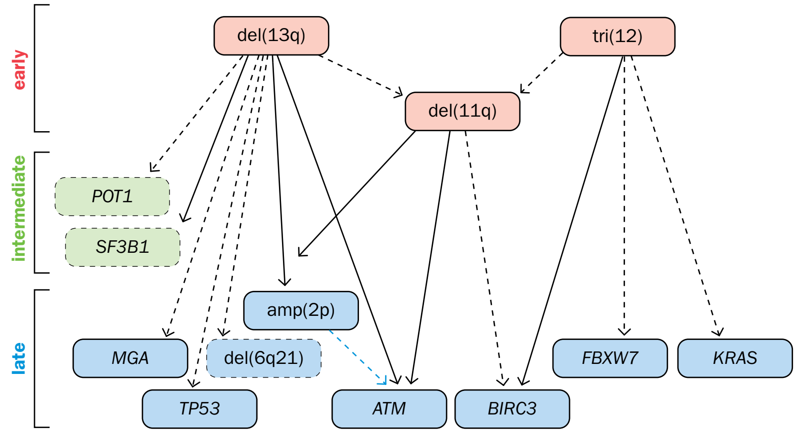


Figure 19.27 Mutational history of chronic lymphocytic leukemia, as inferred from the clonality of mutations. (Reprinted from Landau DA *et al.* [2015] *Nature* **526**:525–530; PMID 26466571. With permission from Springer Nature. Copyright © 2015.)

Metastasis, the formation of disseminated secondary tumors, is the process that kills cancer patients, but the biology of metastasis is poorly understood. There do not seem to be specific mutations that act as general drivers of metastasis, which is unfortunate because they would be excellent targets for antimetastatic drugs. As described in the following section, rare circulating tumor cells can be isolated from the blood of cancer patients, and some subset of these must be the agents of metastasis. Characterizing these single cells and comparing them with both primary and secondary tumors will hopefully lead to a better understanding of metastasis.

Cancer stem cells are well characterized in leukemias but not in solid tumors. It is controversial how far solid tumors depend on small numbers of stem cells. As indicated in **Figure 19.26H**, single-cell analysis of tumors may settle the question. It is thought that cancer stem cells form a very small subpopulation (<1%) of cells in solid tumors, but that they are resistant to chemotherapy and so can persist through treatment and may be the ultimate cause of relapses. It would be important to characterize these cells (if they exist) and identify their vulnerabilities as the basis for long-lasting cancer therapy.

The review by Navin (2015, PMID 26430160; see Further Reading), from which **Figure 19.26** was taken, discusses many applications and achievements of single-cell technologies. As continuing technical development improves the quality and quantity of information that single-cell studies can provide, this is now the cutting edge of cancer research.

19.5 USING OUR NEW UNDERSTANDING OF CANCER

After surgery to remove as much of a tumor as possible, traditional cancer treatment uses radiotherapy and/or chemotherapy. In both cases the treatment is aimed at killing rapidly-dividing cells. Unfortunately, these include normal cells in the gastrointestinal tract, immune system, hair follicles, and so on, hence the long list of often severe side-effects including constipation and diarrhea, nausea, hair loss, tiredness, weakness, and immune suppression.

Targeted anticancer therapies

Our new understanding of driver mutations in cancer has stimulated an intense effort to develop agents to target the specific molecules or pathways that drive tumor development. There are three main approaches:

- Identify small molecules that inhibit a tumor-specific enzyme or signaling molecule or a pathway that is overactive in tumor cells. For example, gefitinib and erlotinib were developed as competitive inhibitors of the epidermal growth factor receptor, which is overactive in many lung and other tumors (**Figure 19.28**). Many of the small-molecule drugs target specific mutations. Patients need to be genotyped to see whether or not their tumor carries the relevant mutation. The combination of a therapeutic agent and companion diagnostic may be the future of personalized medicine;
- Develop monoclonal antibodies (MABs) against tumor-specific cell surface proteins. The effect may be to block the activity of a cell surface receptor or to trigger attack by the immune system. Sometimes the MAB is conjugated to a toxin or radioactive compound to kill targeted cells;
- Engineer T lymphocytes to attack tumor-specific cell-surface antigens.

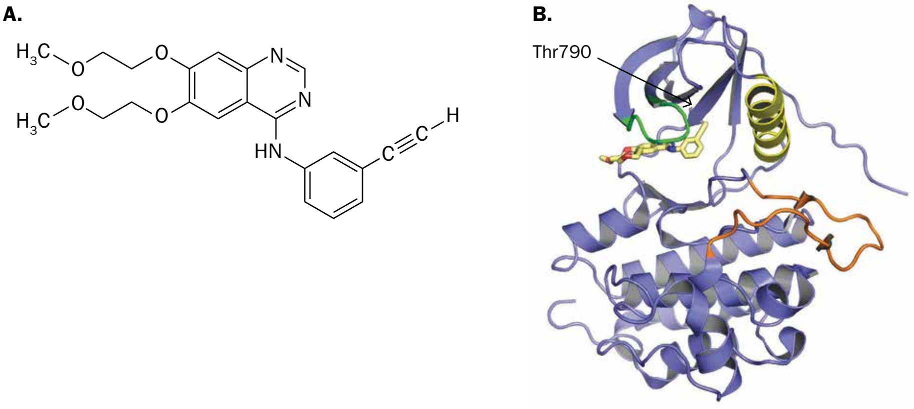


Figure 19.28 A small-molecule tyrosine kinase inhibitor. (**A**) Formula of erlotinib. Erlotinib is a competitive inhibitor of the epidermal growth factor receptor, EGFR. (**B**) The tyrosine kinase domain of a molecule of EGFR with a molecule of erlotinib (yellow) occupying the ATP-binding pocket. Tumors eventually develop resistance to erlotinib, mainly by acquiring the p.T790M mutation. Another small-molecule drug, osimertinib, has been specifically developed as a noncompetitive inhibitor of T790M mutant EGFR. (B, from Yasuda H *et al*. [2012] *Lancet Oncol* **13**:e23–31; PMID 21764376. With permission from Elsevier.)

The prototype targeted small-molecule drug was imatinib (Glivec®/Gleevec®). Imatinib inhibits the tyrosine kinases encoded by the *ABL1*, *KIT*, and *PDGFRA* genes. It has a particular affinity for the chimeric BCR–ABL1 tyrosine kinase encoded by the 9;22 translocation Philadelphia chromosome in chronic myelogenous leukemia (CML). Introduction of imatinib produced a step change in the prognosis of CML. It is also used for patients with gastrointestinal stromal tumors that have mutant *KIT* genes.

A considerable number of other targeted drugs are marketed or in clinical trials. **Table 19.9** shows a small selection. The small molecules can often enter cells; MABs

TABLE 19.9 EXAMPLES OF TARGETED CANCER THERAPEUTICS

Tissue/cancer	Brand name (generic name)	Protein target	Mode of action
SMALL-MOLECULE DRUGS			
Breast	Many brands (tamoxifen)	Estrogen receptor (ER) in ER-positive breast cancers	Blocks ER, preventing growth signals
Leukocytes/leukemia	Glivec® (imatinib)	BCR–ABL1 fusion protein	Inhibits abnormal signaling by fusion protein tyrosine kinase
Skin/melanoma	Zelboraf® (vemurafenib)	BRAF V600E mutant protein	Specifically inhibits V600E mutant BRAF, triggers apoptosis
Non-small-cell lung cancer	Xalkori® (crizotinib)	EML4–ALK fusion protein	Inhibits abnormal signaling by fusion protein tyrosine kinase
Ovarian (advanced)	Lynparza® (olaparib)	PARP1 enzyme	Blocks repair of DNA breaks in *BRCA1*-mutant cancers
Lung/various	Iressa® (gefitinib)	Epidermal growth factor receptor (EGFR) mutants	Binds cytoplasmic part of EGFR, blocks signaling
Lung/various	Tarceva® (erlotinib)	EGFR mutants	Binds cytoplasmic part of EGFR, blocks signaling
Various advanced cancers	Tagrisso® (osimertinib)	EGFR T790M mutant	Binds cytoplasmic part of T790M mutant EGFR, blocks signaling
MONOCLONAL ANTIBODIES			
Breast	Herceptin® (trastuzumab)	EGFR on HER2-positive cells	Attaches to receptor, identifies the cell as a target for the immune system
Skin/melanoma	Yervoy® (ipilimumab)	CTLA4 T-cell inhibitor	Absence of CTLA4 stimulates T cells to attack cancer cells
Skin/melanoma	Opdivo® (nivolumab)	PD-1 T-cell inhibitor	Absence of PD-1 stimulates T cells to attack cancer cells
Leukocytes/leukemia	Mabthera® (rituximab)	CD20 B-cell surface protein	Binds to CD20, identifies cells as targets for natural killer (NK) cells
Colon, lung, head, and neck	Erbitux® (cetuximab)	EGFR	Binds EGFR on outside of cells, blocks signaling; only for tumors with no *KRAS* mutation
Various advanced cancers	Avastin® (bevacizumab)	Vascular endothelial growth factor (VEGF) on cell surface	Inhibits angiogenesis
Lymphatic system/ lymphoma	Zevalin®* (ibritumomab)	CD20 B-cell surface protein	Binds to CD20, radioisotope kills cells
Prostate (advanced)	Provenge®** (sipuleucel-T)	Prostatic acid phosphatase (PAP)	Stimulates T-cell response against PAP

* Zevalin is a monoclonal antibody carrying a radioactive payload that kills target cells.
** Provenge uses a proprietary protein consisting of PAP fused to granulocyte-macrocyte colony-stimulating factor (GM-CSF) to stimulate the patient's own leukocytes *ex vivo*.

can only react with antigens exposed on the cell surface (although so-called intrabodies, single-chain variable fragment [scFv] antibodies, can function within cells; see Section 21.2). MABs also require to be humanized: they are originally produced in mouse hybridoma cells but need mouse-specific amino acids to be replaced by human-specific ones to avoid triggering an immune response in patients. Zevalin® (see Table 19.9) is an example of using the specificity of an antibody to deliver a toxic payload, in this case a radioisotope. Provenge® primes the patient's own immune system to attack cells expressing prostatic acid phosphatase by incubating patient-derived antigen-presenting cells with the enzyme and a growth factor, granulocyte-macrocyte colony-stimulating factor (GM-CSF). The expanded population of primed cells is re-infused into the patient.

Olaparib, the PARP1 inhibitor, demonstrates the potential of **synthetic lethality**. This describes the way a combination of two nonlethal deficiencies can lead to a lethal effect.

Poly(ADP-ribose) polymerase is the key signaling molecule for activating the repair pathway of single-strand DNA breaks. In the absence of PARP, single-strand gaps persist, replication forks collapse in S phase, and the damage has to be repaired by BRCA1/2-mediated homologous recombination. Cells with BRCA1/2 mutations are unable to do this and so are very vulnerable to inhibition of PARP. In the absence of both PARP and BRCA1/2, the cell has to attempt to rescue the damage by nonhomologous end-joining. This introduces many errors that are likely to lead to cell death. Thus, PARP inhibitors are very effective against tumors with BRCA1/2 mutations, but ineffective against other tumors.

Engineered T cells are the subject of intense technical development and hundreds of clinical trials. They have already proven effective against some leukemias and lymphomas. T-cell receptors recognize antigens presented by self-MHC (major histocompatibility complex) molecules, see Section 3.4. One way tumors escape immunosurveillance is by down-regulating expression of MHC molecules. Chimeric antigen receptor T cells (CAR T cells) remove the requirement for MHC recognition. The chimeric receptor includes a scFv single-chain antibody directed against the target antigen (see the review by June *et al*., PMID 29567707 in Further Reading). Initially CAR T cells were made for each individual patient using the patient's own T cells; further genetic engineering raises the prospect of universal cells made by eliminating endogenous MHC or T-cell receptor genes from T cells of healthy donors. Such a development would considerably reduce the labor and cost of generating CAR T cells for a patient. Their main drawback is their propensity to trigger a massive release of cytokines, which at best causes an unpleasant influenza-like reaction and at worst has resulted in several fatalities in clinical trials. Hopefully this can be solved by fine-tuning the protein engineering.

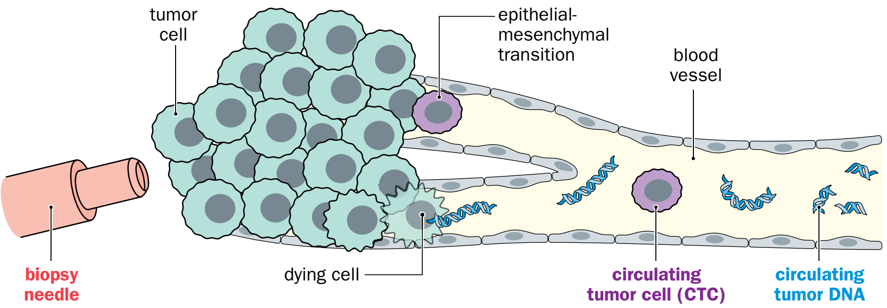
The initial response may be dramatic, but tumors soon develop resistance

The initial results of treatment can be very positive, and although they are not free of side-effects, these drugs are generally better tolerated than conventional chemotherapy. However, it is fair to note that there is a good deal of "hype" surrounding this new wave of targeted drugs. The concept is very beguiling, but the reality is usually more prosaic. Results in leukemia were very encouraging, but leukemias have far less genomic instability than the common epithelial cancers; they can also often be caught at an earlier stage in their development. Almost every patient with chronic myelogenous leukemia has the *BCR–ABL1* translocation and so is eligible for treatment with Glivec®. For most epithelial cancers, only a minority of patients have the genetic changes necessary to obtain benefit, and for those that do benefit, the result is not a cure but a temporary remission.

In a rapidly-evolving and genetically unstable tumor, sooner or later a cell line resistant to the chosen drug will inevitably emerge and the disease will progress. For example, most EGFR-positive tumors treated with erlotinib or gefitinib will eventually develop resistance; of those, two-thirds will carry a particular mutation, p.T790M, that blocks the insertion of the drug molecule into the ATP-binding pocket of EGFR, rendering it ineffective (see **Figure 19.28**). A new molecule, osimertinib (Tagrisso®) was specifically engineered to overcome this resistance. After early clinical trials, the European Union gave approval in February 2016 for use of this drug in metastatic EGFR T790M-positive non-small-cell lung cancer that had progressed from the earlier therapy. Osimertinib was the first new medicine to be approved under the European Commission's expedited process: it took under 3 years from the start of clinical trials to approval, as compared to 12–15 years in a traditional drug-development program. These tumors in turn will eventually become resistant to the drug, and a third-line treatment will be needed.

"Liquid biopsies" allow the emergence of resistance to be followed

It would be highly desirable to be able to monitor a tumor so as to detect any emerging resistant clone at the earliest possible stage, when hopefully some change of treatment could prevent it developing further. With leukemias, serial blood tests allow this possibility. Solid tumors are much more difficult. Patients, perhaps frail and still recovering from surgery, can hardly be subjected to endless repeat biopsies. The development of so-called "liquid biopsies" offers a possible solution. Both circulating tumor cells (CTC) and cell-free circulating tumor DNA (ctDNA) are present in the peripheral circulation of patients with metastatic cancer (**Figure 19.29**). If they could be routinely and reliably characterized, they would offer several advantages over surgical biopsies. Not only would they be more acceptable to patients, but they could also better capture the heterogeneity of tumor cells in an individual patient, both in different regions of the primary tumor and in metastases, and they could be repeated as often as desired to follow



	biopsy	CTC	ctDNA
invasive	+	–	–
all patients eligible	–	+	+
instrumentation required	+	+	–
WGA required	–	+	+/–
RNA profiling	+	+	–
research applicability	+++	++	+
biomarker applicability	–	++	+++

Figure 19.29 Comparison of liquid versus needle biopsies for investigating cancer. CTC, circulating tumor cells; ctDNA, circulating (cell-free) tumor DNA; WGA, whole-genome amplification. (Reproduced from Wyatt AW & Gleave ME [2015] *EMBO Mol Med* **7**:878–894; PMID 25896606. © 2015 The Authors. Published under the terms of the CC BY 4.0 license.)

progression of the disease. Liquid biopsies are particularly promising for guiding treatment by monitoring the emergence of resistant clones.

The average amount of circulating tumor DNA in one study was 17 ng/ml of plasma, but individual levels vary very widely and the proportion of total cell-free DNA that is tumor-derived can be anything from 0.01% to 93%. The DNA can be checked for specific mutations—mutations associated with the emergence of resistance, for example, or mutations known to be present in the patient's primary tumor. Alternatively, exome sequencing or array-comparative genomic hybridization could be used for a less targeted test. Circulating tumor cells are very rare. Typically, there might be one CTC among 10^7 white blood cells in 1 ml of peripheral blood. CTC are often recovered using magnetic beads coated with an epithelial cell adhesion molecule. Other methods use physical properties of the cells to separate them from other cells present in a blood sample. Hopefully the cells could then be characterized by single-cell sequencing or single-cell transcriptomics. All these approaches push the available technology to its limits. The whole field of liquid biopsies holds great promise and is the subject of intensive efforts in technical development and translational research. It is not yet part of routine clinical oncology. The reviews by Crowley *et al*. (2013) and Krebs *et al*. (2014) (PMID 23836314 and 24445517, respectively; see Further Reading) give more detail.

The future may lie with combination treatments

The new drugs give patients a few precious months of extra life, but they do not cure the cancer. Sooner or later resistant clones emerge. They are also exceedingly expensive, especially the monoclonal antibodies. Small-molecule drugs such as Tarceva® or Xalkori® typically cost US$15,000 per month. The cost of treating a single melanoma patient with a combination of two MABs (Yervoy® and Opdivo®, mentioned below) has been estimated at over $250,000. Rational allocation of limited health-care resources is difficult in these circumstances. The amount of money that can give one cancer patient 6 more months of life might confer far more benefit to more people if used, say, in mental health. But it is difficult to resist emotive publicity detailing individual patients dying of cancer who, it may be suggested, are being denied life-saving treatments by cold-hearted bureaucrats.

Despite these caveats it is clear that we are in the early stages of major progress in cancer treatment. The way forward is probably to take a leaf out of human immunodeficiency virus (HIV) treatment. As with cancer, physicians treating HIV patients face a highly mutable adversary that can quickly mutate to resistance against any individual antiviral drug. However, combination treatments (highly-active antiretroviral therapy, HAART) are much more successful because it is very unlikely that one virus molecule could simultaneously mutate to become resistant to two or three drugs that attack different vulnerabilities in the virus. In the same way, combinations of targeted drugs can have a much greater therapeutic effect than single drugs. For example, in one clinical trial involving 945 patients with metastatic or unresectable melanoma, treatment with the combination of Yervoy® and Opdivo® (see **Table 19.9**) resulted in 9% of the tumors disappearing completely and a further 41% shrinking. In this example, the actions of both drugs converge on enhancing immune attack on the cancer cells. Future precision approaches could include multiple agents and drugs that target different vulnerabilities in cancer cells, in order to kill the cells before resistance has a chance to develop. It seems possible that such developments could lead to actual cures for at least some cancer patients, rather than simply temporary remissions.

SUMMARY

- Cancer is the result of Darwinian evolution by natural selection acting on spontaneously arising mutations in somatic cells.
- Tumor cells need to acquire genetic changes that confer on them six general features: (1) independence of external growth signals; (2) insensitivity to external antigrowth signals; (3) the ability to avoid apoptosis; (4) the ability to replicate indefinitely; (5) the ability of a mass of such cells to trigger angiogenesis and vascularization; and (6) the ability to invade tissues and establish secondary tumors. Additionally, metabolism is re-programmed to support cell growth, and tumors must also be able to avoid immune surveillance.
- Genomic instability is a normal feature of tumor cells. Instability can be seen at all levels: chromosomal losses, gains, and rearrangements, structural variants, point mutations, and epigenetic dysregulation.
- Because of this instability, tumors contain large populations of cells carrying a great variety of mutations, on which natural selection can act. Driver mutations contribute to tumor development and are subject to positive selection; passenger mutations are chance by-products of the genomic instability of most tumor cells.
- Highly evolved and sophisticated defense mechanisms protect the body against the proliferation of such mutant cells. The mutations in successful tumor cells disable these defenses.
- Tumors most likely originate from cells that already have a high proliferative capacity, such as stem cells or rapidly multiplying and poorly differentiated embryonal tissues.
- The evolution of a tumor usually occurs gradually, with successive stages showing increasing cell proliferation and decreasing differentiation. Tumors where a catastrophic event has simultaneously generated multiple genomic changes may be partial exceptions to this gradualist model.
- Oncogenes normally act to promote cell division, but a complex regulatory network limits their activity. In tumor cells, one copy of an oncogene is often abnormally activated, through point mutations, copy number amplification, or chromosomal rearrangements, so that it escapes regulation.
- Chromosomal rearrangements in tumor cells often create novel chimeric oncogenes but alternatively may up-regulate the expression of an oncogene by placing it under the influence of a powerful enhancer.
- The normal role of tumor suppressor genes is to limit cell division or to help maintain the integrity of the genome. In tumor cells, this effect is lost through loss-of-function mutations. For genes that show haploinsufficiency, mutation of a single copy can already confer a growth advantage. Often both copies are inactivated by deletions, point mutations, or methylation of the promoter.
- Many tumor suppressor genes have been identified through investigation of familial cancer predisposition syndromes. In these syndromes, individuals are constitutionally heterozygous for an inherited mutation that inactivates one allele of a tumor suppressor gene. A somatic second mutation leads to complete loss of function.
- Oncogenes and tumor suppressor genes normally function in the cell signaling that controls the cell cycle, or in the response to DNA damage. Understanding these processes is central to understanding what goes wrong in cancer.
- Genomic technologies allow the totality of acquired genetic changes in tumor cells to be cataloged, including structural variants, small sequence variants, changes in the transcriptome, and epigenetic changes. The results of these studies emphasize the individuality and large number of such changes in most tumor cells.
- Molecular profiling allows a new classification of tumors, in terms of their gene expression profiles instead of their tissue of origin. Tumors of the same tissue may be divided into different molecular types, with different prognoses and maybe different optimal management strategies. Tumors of different tissues may turn out to have similar molecular profiles and so perhaps benefit from the same drugs.
- Although different tumors have very many different mutations affecting many genes, they all affect a modest number of metabolic or regulatory pathways. Thinking in terms of pathways rather than individual genes helps make sense of the massive heterogeneity of tumor mutations.
- Identifying critical gene mutations or pathways in tumors has allowed the development of a new generation of anticancer drugs that target specific genes or pathways, rather than just targeting any rapidly dividing cells. It is hoped that further development of these drugs will lead to a new era of personalized medicine, in which treatment depends on the patient's tumor genotype. The drugs need to be marketed together with a companion diagnostic.
- Current targeted drugs mostly provide only a temporary remission in the inevitable progression of cancer. It is hoped that simultaneously targeting different vulnerabilities in a tumor by means of combinations of drugs may lead to much more pronounced effects.

FURTHER READING

General overviews of cancer

Davoli T *et al*. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**:948–962; PMID 24183448.

Hanahan D & Weinberg RA (2000) The hallmarks of cancer. *Cell* **100**:57–70; PMID 10647931.

Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**:646–674; PMID 21376230.

Jansson MD & Lund AH (2012) MicroRNA and cancer. *Mol Oncol* **6**:590–610; PMID 23102669.

Jones PA & Baylin SB (2007) The epigenomics of cancer. *Cell* **128**:683–692; PMID 17320506.

Kinzler KW & Vogelstein B (1996) Lessons from hereditary colorectal cancer. *Cell* **87**:159–170; PMID 8861899.

Martincorena I & Campbell PJ (2015) Somatic mutation in cancer and normal cells. *Science* **349**:1483–1489; PMID 26404825.

Melo SA & Esteller M (2011) Dysregulation of microRNAs in cancer: playing with fire. *FEBS Lett* **585**:2087–2099; PMID 20708002.

Vogelstein B *et al.* (2013) Cancer genome landscapes. *Science* **339**:1546–1558; PMID 23539594.

Zhu L *et al.* (2016) Multi-organ mapping of cancer risk. *Cell* **166**:1132–1146; PMID 27565343.

Oncogenes

Flavahan WA *et al.* (2016) Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **529**:110–114; PMID 26700815.

Hnisz D *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**:1454–1458; PMID 26940867.

Mertens F *et al.* (2015) The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* **15**:371–381; PMID 25998716.

Mitelman F, Johansson B, Mertens F (eds) (2016) Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. http://cgap.nci.nih.gov/chromosomes/mitelman

Tumor suppressor genes

Cavenee WK *et al.* (1983) Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**:779–84; PMID 6633649.

Esteller M *et al.* (2000) Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst* **92**:564–569; PMID 10749912.

Herman JG & Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **349**:2042–2054; PMID 14627790.

Knudson AG Jr (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* **68**:820–823; PMID: 5279523.

Piotrowski A *et al.* (2014) Germline loss-of-function mutations in *LZTR1* predispose to an inherited disorder of multiple schwannomas. *Nat Genet* **46**:182–187; PMID 24362817.

Fodde R *et al.* (2001) APC, signal transduction and genetic instability in colorectal cancer. *Nat Rev Genet* **1**:55–67; PMID 11900252.

Genomic analyses of tumors

Alexandrov LB *et al.* (2013) Signatures of mutational processes in human cancer. *Nature* **500**:415–421; PMID 23945592.

Baca SC *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell* **153**:666–677; PMID 23622249.

Bailey P *et al.* (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**:47–52; PMID 26909576.

Curtis C *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**:346–352; PMID 22522925.

Drost J *et al.* (2015) Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521**:43–47; PMID 25924068.

Hoadley KA *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**:929–944; PMID 25109877.

Kloosterman WP (2015) Making heads or tails of shattered chromosomes. *Science* **348**:1205–1206; PMID 26068832.

Landau DA *et al.* (2015) Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**:525–530; PMID 26466571.

Navin NE (2015) The first five years of single-cell cancer genomics and beyond. *Genome Res* **25**:1499–1507; PMID 26430160.

Setlur SR & Lee C (2012) Tumor archaeology reveals that mutations love company. *Cell* **149**:959–961; PMID 22632962.

Shen MM (2013) Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell* **23**: 567–569; PMID 23680143.

Stephens PJ *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**:27–30; PMID 21215367.

Cancer treatment and monitoring strategies

Cheung-Ong K *et al.* (2013) DNA-damaging agents in cancer chemotherapy: serendipity and chemical biology. *Chem Biol* **20**:648–659; PMID 23706631.

Crowley E *et al.* (2013) Liquid biopsy: monitoring cancer-genetics in the blood. *Nat Rev Clin Oncol* **10**:472–484; PMID 23836314.

June CH *et al.* (2018) CAR T cell immunotherapy for human cancer. *Science* **359**:1361–1365; PMID 29567707.

Krebs MG *et al.* (2014) Molecular analysis of circulating tumour cells: biology and biomarkers. *Nat Rev Clin Oncol* **11**: 129–144; PMID 24445517.

Puigvert JC *et al.* (2016) Targeting DNA repair, DNA metabolism and replication stress as anti-cancer strategies. *FEBS J* **283**:232–245; PMID 26507796.

APPLIED HUMAN MOLECULAR GENETICS

PART FIVE

20

Genetic testing in healthcare and the law

Genetic testing is a routine tool for almost all biomedical scientists. PCR-based tests are the standard way of identifying pathogens and are the everyday working tools of microbiologists and virologists. Emerging diseases are identified by sequencing the pathogen's genome and are tracked by genetic tests. Animal and plant breeders use genetic tests to guide their work. Our modern understanding of evolution and development is based on genetic tests. However, this being a book on human molecular genetics, we will confine ourselves here to tests focused on the human genome.

Even here the field is very extensive. Previous chapters have covered many of the principles underlying genetic testing and applications of genetic tests in research, for example to explore the natural variability of human genomes and to understand the roles of genetic variants in disease. Here we consider the methodology of testing in human diagnostics, screening, and forensics. As always in this book, we concentrate on the principles and not the practical details. Best practice guidelines for laboratory diagnosis of the commoner Mendelian diseases are available on the websites of the American College of Medical Genetics and in the UK the Association for Clinical Genomic Science, among others. The reader interested in specific procedures or conditions should consult these.

Genetic tests are unusual among clinical tests because (except when following the evolution of a tumor) a genetic test is normally performed just once, and the result forms a permanent part of a person's health record. Thus, it is especially important to avoid mistakes in diagnostic testing. Any laboratory offering clinical (or forensic) testing must have adequate quality assurance measures in place. These are beyond the scope of this chapter—any interested reader should consult the principles and guidelines published by the Organisation for Economic Co-operation and Development (OECD; see Further Reading).

When a clinician brings a sample from a patient to a laboratory for diagnostic testing of the DNA, there are three possible questions that the laboratory might be asked to answer:

- Does the patient have *a specific variant*, for example, a three-base deletion of the codon for phenylalanine 508 in his or her *CFTR* gene, or a specific chromosomal microdeletion? The circumstances in which this type of very specific question can be asked, and ways of answering it, are considered in Section 20.2;
- Does the patient have *any variant in this particular gene* (or in any one of the genes in this panel of candidate genes) that might cause the disease? Standard ways of answering this question are considered in Section 20.3;
- Does the patient have *any variant anywhere in his genome* that would explain his condition? Although diagnostic laboratories have long dealt with this question in the context of chromosomal abnormalities, for smaller sequence variants this was an impossible question to answer until the widespread adoption of next-generation sequencing by diagnostic laboratories. Nowadays whole exome or whole genome sequencing can provide a list of all candidate genetic variants. The challenge has moved from generating the data to interpreting it.

People sometimes suppose that once a diagnostic test is available for a condition, the logical next step is to move to screening the whole population. Section 20.4 shows that it is not so simple. This section examines some of the particular issues—clinical, organizational, financial, and ethical—involved in schemes for population screening.

The much-promised development of personalized or precision medicine would see genetic information used not just for diagnosis but for management of a patient's condition. Physicians have always known that different patients respond differently to many drugs; increasing genetic knowledge provides some explanations, and these are covered in Section 20.5. Genotypes are only one among many factors causing differential responses to a drug, and it would be unwise to expect an era in which all prescribing is based on genotype. However, specific instances are accumulating where this knowledge is important, particularly for identifying people at risk of severe adverse reactions. Genotyping (in this case of the tumor) is especially relevant in cancer; a number of drugs now target specific mutations in tumors and are supplied together with a companion diagnostic.

Genetic questions might be asked not by a clinician about somebody's diagnosis or prognosis but by a policeman or lawyer about somebody's identity or the relationship between two people. Such questions are addressed by DNA profiling, which is described in Section 20.6. DNA profiling is also occasionally needed to answer clinical questions—for example checking whether a bone marrow transplant has successfully re-populated a patient's immune system.

20.1 WHAT TO TEST AND WHY

DNA for genetic testing can be obtained from any specimen containing nucleated cells, but clinical considerations may dictate the choice of sample. Sometimes it is more appropriate to test RNA or to perform a functional test, for example of enzyme activity; in those cases, the sample must come from a tissue in which the gene is expressed and/or the gene product is normally functional.

Many different types of specimen can be used for genetic testing. Almost always the first step is amplification of the DNA or RNA (as cDNA) by PCR, applying the methods described in Section 6.2. Southern blotting (see **Figure 6.15**) is still used for a few applications, such as testing for fragile X full mutations. The sensitivity of PCR makes it possible to use a wide range of tissue samples (**Table 20.1**). Blood is the most reliable general source of genomic DNA, but mouthwashes or buccal scrapes are used when sampling must be noninvasive. For invasive prenatal diagnosis, chorionic villi or amniotic fluid are normally used. However, the biopsy procedure carries a roughly 1% risk of causing a miscarriage. The development of noninvasive prenatal testing, using fetal DNA in the

TABLE 20.1 SOURCES OF DNA OR RNA FOR GENETIC TESTING

Source	Comments
Peripheral blood	The best general source of DNA
Mouthwash or buccal scrape	Noninvasive, but quality and quantity can be variable
Skin, muscle, etc.	For RNA studies; needs to be a tissue where the gene of interest is expressed; some mitochondrial DNA studies are best performed on a muscle biopsy
Tumor biopsy	Essential tool for management of cancer; 'liquid biopsy' of cell-free tumor DNA in peripheral blood may replace some invasive procedures
Chorionic villi	For prenatal diagnosis at 9–14 weeks
Amniotic fluid	For prenatal testing at 15–20 weeks of pregnancy; a relatively poor source of fetal DNA compared with chorionic villi
Fetal DNA in maternal blood	For noninvasive prenatal screening or diagnostic testing
Single cell from a blastocyst	For pre-implantation diagnosis; technically very demanding
Hair roots, semen, cigarette butts, etc.	Scene-of-crime samples for forensic analysis
Pathological specimens	Vital resource for genotyping dead people; also tumor, etc., biopsies; formalin-fixed paraffin-embedded tissue requires special procedures for DNA purification
Guthrie card	The cards used for neonatal screening have one or more spots of the baby's blood, not all of which are normally used for the screening; if cards are archived they are a possible source of DNA from a deceased baby

maternal bloodstream, has reduced the need for these invasive procedures. Archived pathological specimens are very important sources of DNA from deceased persons, although consent laws in some countries make it difficult to use them. Usually they will have been fixed in formalin and embedded in paraffin wax, which necessitates special DNA extraction procedures. Only short sequences, typically 200 base pairs or less, can usually be recovered.

RNA or DNA?

Only RNA analysis can reliably detect aberrant splicing. It is often hard to predict whether a DNA sequence change will affect splicing; moreover, aberrant splicing may result from activation of a cryptic splice site deep within an intron, which would not be detected by the usual exon-based protocols for sequencing genomic DNA (see Section 16.1). RT-PCR also allows a direct test for nonexpression of a gene, which may be caused by changes outside the coding sequence that are hard to pick up on DNA testing. In addition, a laboratory might prefer to sequence one cDNA rather than 50 or 70 exons of a large gene.

However, analyzing RNA has disadvantages. RNA is much less convenient to obtain and work with than DNA. Samples must be handled and processed with great care to avoid degrading mRNA. Importantly, the gene of interest may not be expressed in any readily accessible tissue. A very low level of ‘illegitimate transcripts’ may sometimes be present in tissues in which the gene is not normally expressed, but reliable analysis of very low-level mRNAs can be difficult. In addition, truncating mutations usually result in unstable mRNA because of nonsense-mediated decay (see **Figure 16.7**), so that the RT-PCR product from a heterozygous person may show only the normal allele. Treating cultured cells or whole blood with the translation inhibitor puromycin has been shown to inhibit nonsense-mediated decay, which may allow cDNA sequencing to detect transcripts that include premature termination codons.

Functional assays

When considering mendelian pedigree patterns (see Section 5.2), we simply distinguished two classes of allele, normal and abnormal, or *A* for the dominant character and *a* for the recessive. Assaying the function of a gene might allow a similar distinction to be made in the laboratory—which is, after all, the essential question in most diagnoses. Function might be tested at the DNA level, for example by a cell transfection assay. Alternatively, the protein product of a gene could be tested biochemically if its function can be adequately assayed in the laboratory. The problem with functional assays is that they are specific to a particular gene or protein; by contrast, DNA technology is generic. This has obvious advantages for the diagnostic laboratory, but in addition it encourages technical development, because any new technique could be used for a wide variety of problems.

20.2 TESTING FOR A SPECIFIC GENETIC VARIANT

Often a genetic test involves checking for the presence of one or more specific pre-defined variants. Typical cases include:

- Diagnosis within a family. Sequencing may be needed to define the family mutation, but once it is characterized, other family members normally need be tested only for that particular mutation;
- Diagnosis or screening of diseases with limited allelic heterogeneity—these are typically diseases caused by a gain of function (see **Figure 16.16**) or with strong founder effects in a certain population (see Section 12.3);
- Checking for a specific microdeletion or microduplication that is suggested by the patient’s phenotype;
- Checking a tumor biopsy for mutations that would govern its response to a targeted drug (see Section 19.4);
- Testing control samples to see if a change seen in a patient is actually a low-frequency population polymorphism;
- Amplifying specific microsatellites to check the repeat number;
- SNP genotyping.

Sequencing can always be used for these purposes, but when the target variants are known, various cheaper and simpler methods are available to genotype them. Some of the main methods are summarized in **Table 20.2**. Many variants of these and other methods have been developed as kits by biotechnology companies.

TABLE 20.2 SOME METHODS OF TESTING FOR A SPECIFIED SEQUENCE VARIANT

Method	Comments
Check size of PCR product	For genotyping microsatellites, e.g. for legal or forensic purposes (see Section 20.6); checking size of repeat expansions in dynamic repeat diseases (see **Table 16.7**, but large expansions, for example in myotonic dystrophy or fragile X, may require Southern blots or specialized methods)
Check size of PCR product after restriction digestion	If the mutation creates or abolishes a natural restriction site or one engineered by the use of special PCR primers (see **Figure 20.1**); tests for one single mutation
PCR using allele-specific primers (ARMS test)	General method for specified point mutations (see **Figure 20.2**); a few dozen tests can be multiplexed
Oligonucleotide ligation assay (OLA)	General method for specified point mutations (see **Figure 20.3**); a few dozen tests can be multiplexed
Quantitative real-time PCR	Checking for specific copy-number variants
Single-nucleotide primer extension	General method for specified point mutations; formatted for readout on a DNA sequencer
Mass spectrometry	Good for repetitive analysis of a fixed panel of up to several hundred SNPs or mutations; quantitative results
Pyrosequencing	Sequencing a few nucleotides at a specified position; quantitative result (see **Box 6.4**)
Hybridize PCR-amplified DNA to allele-specific oligonucleotides (ASO) on a microarray	For genotyping a large panel of single nucleotide variants; the mainstay of the genome-wide association studies described in Chapter 18
Fluorescence *in situ* hybridization	Checking for a specific microdeletion or chromosome rearrangement (see **Figure 15.4**)
PCR with primers located either side of a chromosomal breakpoint	Successful amplification shows the presence of a suspected specific deletion or rearrangement

PCR, polymerase chain reaction; SNP, single nucleotide polymorphism .

Testing for the presence or absence of a restriction site

When a base substitution creates or abolishes the recognition site of a restriction enzyme, this allows a simple direct PCR test for the variant. A suitable length sequence containing the potential variant is PCR amplified. The PCR product is digested with the relevant restriction enzyme, and the products of digestion are separated by electrophoresis to see whether or not a cut has occurred. It is important to check that digestion is complete, otherwise somebody who is homozygous for presence of the site may be misclassified as heterozygous. Many point mutations will not happen to affect a restriction site—although hundreds of restriction enzymes are known, they almost all recognize symmetrical palindromic sites. If a variant does not change a suitable site, sometimes a variable restriction site can be introduced by a form of PCR mutagenesis using carefully designed primers. **Figure 20.1** shows an example.

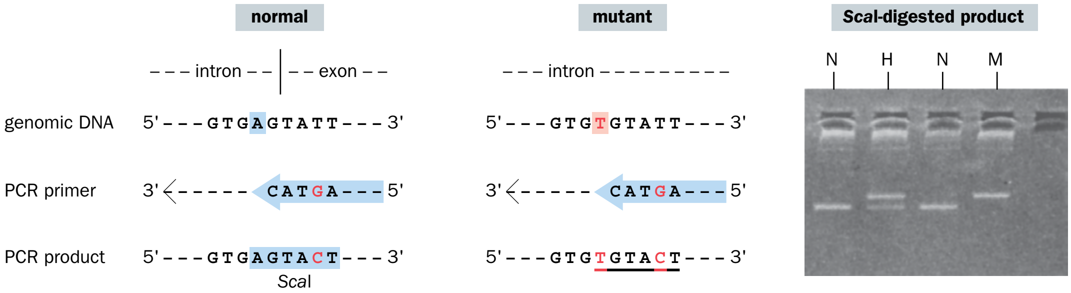

normal
mutant
*Sca*I-digested product
– – – intron – – | – – exon – –
– – – intron – – – – – – – – –
genomic DNA
5' – – – G T G A G T A T T – – – 3'
5' – – – G T G T G T A T T – – – 3'
PCR primer
3' <– – – – – C A T G A – – – 5'
3' <– – – – – C A T G A – – – 5'
PCR product
5' – – – G T G A G T A C T – – – 3'
5' – – – G T G T G T A C T – – – 3'
*Sca*I
N
H
N
M

Figure 20.1 Introducing an artificial diagnostic restriction site. An A→T change in the intron 4 splice site of the *FACC* gene does not create or abolish a restriction site. To detect it a PCR primer is used that stops short of this altered base but has a single base mismatch (red G) in a noncritical position. This does not prevent it from hybridizing to and amplifying both the normal and variant sequences, but the mismatch in the primer introduces an AGTACT restriction site for *Sca*I into the PCR product from the normal sequence but not the variant sequence. The *Sca*I-digested product from homozygous normal (N), heterozygous (H), and homozygous mutant (M) patients is shown. (Courtesy of Rachel Gibson, Guy's Hospital, London.)

Allele-specific PCR amplification

PCR primers do not necessarily have to match their target sequence at every nucleotide position, but the reaction will fail if the 3′ end of the primer where the polymerase must first add nucleotides is mispaired. The ARMS (amplification-refractory mutation system) technique genotypes single nucleotide variants by using two alternative versions of one of the PCR primers, each matching one allele of the SNV at its 3′ end. **Figure 20.2** shows how this could be used to test for the A→T change in the β-globin gene that causes sickle cell disease. The pairs of mutation-specific primers can be made to give distinguishable products so that they can be used together in the same reaction. For example, they can be given different fluorescent labels, or 5′ extensions of different sizes. A deliberate single nucleotide mismatch a few nucleotides upstream of the 3′ end can increase the selectivity. A few dozen ARMS reactions can be multiplexed, with the conserved primers chosen to give different-sized products from each reaction so that they can be separated by gel electrophoresis. Multiplexed allele-specific PCR is well suited to screening fairly large numbers of samples for a given panel of single nucleotide variants.

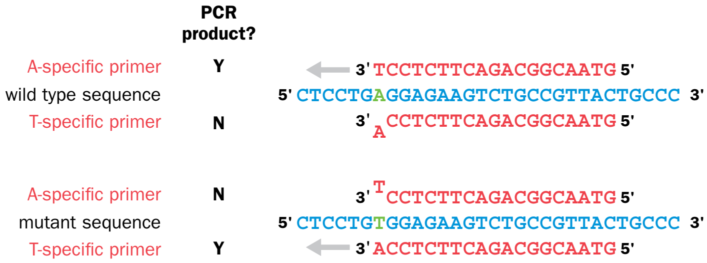

PCR product?
A-specific primer
Y
3' TCCTCTTCAGACGGCAATG 5'
wild type sequence
5' CTCCTGAGGAGAAGTCTGCCGTTACTGCCC 3'
T-specific primer
N
3' A CCTCTTCAGACGGCAATG 5'
A-specific primer
N
3' T CCTCTTCAGACGGCAATG 5'
mutant sequence
5' CTCCTGTGGAGAAGTCTGCCGTTACTGCCC 3'
T-specific primer
Y
3' ACCTCTTCAGACGGCAATG 5'

Figure 20.2 Allele-specific PCR (the ARMS reaction). Here it is used to distinguish wild-type and mutant (sickle cell) β-globin sequences. Primers for the two alleles may be differentially labeled by size or fluorescence to allow the PCR products to be distinguished. An invariant forward primer is off the figure to the left.

The oligonucleotide ligation assay (OLA)

This technique relies on the fact that DNA ligase will only seal a nick in DNA if both nucleotides flanking the nick are correctly base-paired to a complementary strand. To genotype a variable nucleotide in a test DNA, two oligonucleotides are hybridized to the test DNA so that their ends abut at the variable position (**Figure 20.3**). Only perfectly base-paired oligonucleotides can be joined by DNA ligase. The two oligonucleotides

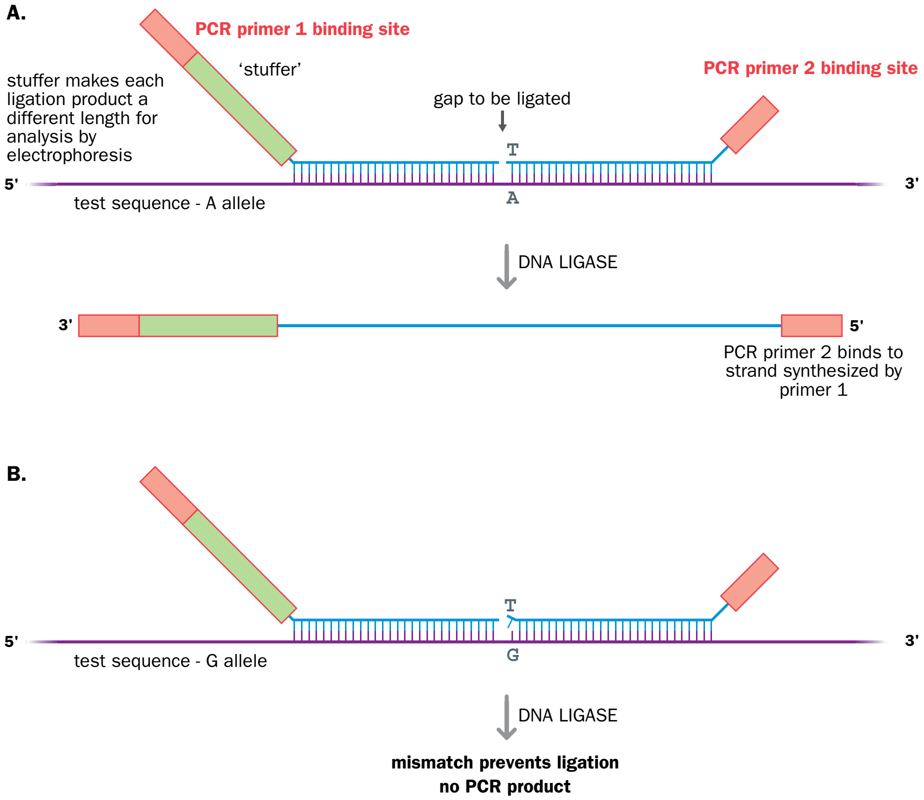

A.
PCR primer 1 binding site
'stuffer'
stuffer makes each ligation product a different length for analysis by electrophoresis
PCR primer 2 binding site
gap to be ligated
T
5'
3'
A
test sequence - A allele
DNA LIGASE
3'
5'
PCR primer 2 binds to strand synthesized by primer 1
B.
T
5'
3'
G
test sequence - G allele
DNA LIGASE
mismatch prevents ligation
no PCR product

Figure 20.3 The oligonucleotide ligation assay (OLA). The two oligonucleotides can only be ligated together if their ends are correctly base-paired. (**A**) The perfectly matched ends can be ligated to form a single PCR-amplifiable molecule. (**B**) The mismatch prevents ligation.

have binding sites for PCR primers, and when ligated form a single PCR-amplifiable molecule. As with the ARMS test, oligonucleotides to match each allele of the test variant can be used, with fluorescent labels or 'stuffer' sequences to give different sized products and allow multiplexing.

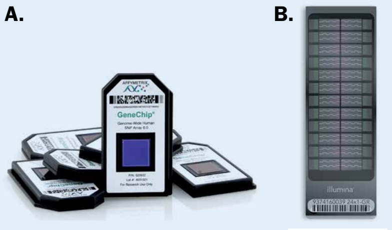
Single base primer extension

SNPs are genotyped by hybridizing a short 'extend' primer to PCR amplicons containing the SNP of interest. The 3' end of the extend primer is immediately adjacent to the variable nucleotide to be identified. As in Sanger sequencing, DNA polymerase and a mix of four fluorescently-labeled dideoxy nucleotides is added. Unlike in Sanger sequencing, no normal deoxy nucleotides are present. Thus, a single dideoxy nucleotide is added to the extend primer, its color depending on the nucleotide present at the SNP. Products are analyzed on a DNA sequencer and can be multiplexed to a moderate degree. The same principle is used on a massively-parallel scale in Illumina SNP chips (see **Box 20.1**).

BOX 20.1 SNP CHIPS

Two widely used systems are the Affymetrix SNP 6 (**Figure 1A**) and Illumina Omni Express-24 (**Figure 1B**).

Both companies offer a range of chips with varying numbers of fixed and user-selectable probes. The SNP 6 chip has probes for 906,600 SNPs and also 946,000 copy number variants; the Omni Express has 710,000 probes covering SNPs, germ-line mutations, structural variants, and copy number variants (CNVs), with the facility to add another 30,000 user-defined probes.



Box 20.1 Figure 1 (**A**) Affymetrix SNP 6; (**B**) Illumina OmniExpress-24.

To use the Affymetrix system, 250 ng DNA samples are digested with *Sty*1 and/or *Nsp*1 restriction enzymes, and linker oligonucleotides are ligated to allow bulk PCR amplification. The cleaned-up PCR product is fragmented and hybridized to the probes on the chip, which are 25 nucleotide-long allele-specific oligonucleotides. For each SNP locus there are two probes, one specific for each allele. After hybridization and washing, bound fragments are end-labeled with a biotin-streptavidin reporting system.

To use the Illumina system, DNA samples (200 ng) are amplified by whole-genome isothermal amplification (see Section 6.2), fragmented and hybridized to the probes. These are 50-mers on microbeads. Each barcoded bead carries many molecules of a single probe that, as with the 'extend' probes used in single-base primer extension or the mass spectrometry systems described, has its 3' end adjacent to the variable SNP nucleotide. Hybridized probes are extended by a single labeled base so as to become differentially labeled depending on the nature of the variable SNP nucleotide. After the hybridization and extension steps all the test DNA (whether or not hybridized) is washed away and the labels on each bead read.

Mass-spectrometric typing of variants

Proprietary systems, for example from the Sequenom or Agena Biosciences companies, can be used to genotype several hundred SNPs or specific variants. The relevant sequences are amplified by multiplex PCR; as for single-base primer extension, after clean-up specific short 'extend' primers are annealed with their 3' ends adjacent to the variant that is to be genotyped. A single nucleotide extension is performed across the variable site but using mass-labeled rather than fluorescently-labeled dideoxynucleotides. The mixed products are then analyzed by MALDI–TOF (matrix-assisted laser-desorption/ionization–time of flight) mass spectrometry. See **Box 7.7** for the principle of MALDI–TOF mass spectrometry. The technique is well-suited to routinely checking many DNA samples for the same set of SNPs or mutations.

SNP chips allow massively-parallel genotyping of SNPs spaced across the genome

The genome-wide association studies described in Section 18.3 were made possible by the development in the late 1990s of microarrays that could genotype a sample for hundreds of thousands of SNPs in a single operation. SNP chips can also be used as an alternative to array-comparative genomic hybridization to define chromosome abnormalities (see Section 15.1). Although sequencing is tending to supersede microarrays for both these applications, SNP arrays remain an important part of the technology of genetic testing. **Box 20.1** describes two leading commercial systems.

20.3 CLINICAL DIAGNOSTIC TESTING

Some diagnostic tests look for one or more specific pre-defined variants, and the methods used for these were covered in the previous section. But very often diagnostic testing involves a more speculative search through all exons of a gene, or through a panel of genes, a whole exome, or a whole genome. Next-generation sequencing has made these tasks far easier, at least for the laboratory if not for bioinformaticians struggling to interpret the raw data. In the past, techniques such as single strand conformation analysis, heteroduplex analysis, and the protein truncation test were used to try to minimize the requirement for sequencing. These methods are largely obsolete now and will not be described here.

Sequencing is the method of choice for checking a sample for small-scale variants

The technologies of Sanger and next-generation sequencing were described in Chapter 6 (Sections 6.4 and 6.5, respectively). In diagnostic work, and depending on work-flows and organization within the laboratory, Sanger sequencing would probably be used if just a single gene was to be checked; where several genes or a whole exome were to be searched, next-generation sequencing would be used. There is a continuing debate about whether variants identified by next-generation sequencing need to be confirmed by Sanger sequencing. Undoubtedly the trend is to use Sanger sequencing less and less, and increasingly to rely on next-generation sequencing. Section 17.3 summarized the processes involved in translating raw next-generation sequencing data into a list of variants.

Deciding what to test: single genes, gene panels, exomes or genomes

A major question for diagnostic laboratories is how wide to cast the search net. There are several options. On might sequence:

- **A single gene**—the clinical features of a patient may point unambiguously to a single malfunctioning gene. For example, although over 1,000 different mutations have been reported in cystic fibrosis (CF) patients, they are all in the *CFTR* gene;
- **A panel of genes**—next-generation sequencing has been particularly useful for phenotypes that can be the result of malfunction of any one of a large but limited number of genes. **Table 20.3** shows some examples. Laboratories or commercial suppliers compile a list of every gene where variants are reported to cause the relevant phenotype and prepare capture probes (see Section 17.3) or PCR primers to capture or amplify each exon of every gene on the list. Alternatively, a laboratory may use virtual gene panels, sequencing the whole exome but restricting the analysis to a specific panel of genes;

TABLE 20.3 EXAMPLES OF GENE PANELS USED IN 2018 BY THE UK 100,000 GENOMES PROJECT

Panel	Total no. of genes
Kabuki syndrome	4
Familial hematuria	8
Congenital hypothyroidism	27
Anophthalmia or microphthalmia	56
Primary ciliary disorders	138
Epileptic encephalopathy	182
Congenital hearing impairment	356
Primary immunodeficiency syndromes	388
Intellectual disability	1997

A total of 224 disease-specific gene panels were developed by crowdsourcing among international experts. They are subject to constant revision—see https://panelapp.genomicsengland.co.uk/panels/. In all cases these are virtual gene panels that specify selective analysis of whole genome sequence data. Other laboratories may use physical gene panels, especially where the number of genes in a panel is modest.

- **The 'clinical exome'**—this is a halfway house between gene panels for specific disorders and whole exomes. A capture kit or set of gene panels cover, between them, all of the 3,000 or so genes that have been reported as having variants that cause Mendelian conditions;
- **The whole exome**—capture kits from various companies isolate 40–60 Mb of DNA (1.5–2% of the genome). Kits typically include around 180,000 protein-coding exons plus variable amounts of nontranslated gene sequence, regulatory sequences, miRNA genes, and so on.
- **The whole genome**.

Whole genome sequencing will likely supplant exome sequencing once prices come down sufficiently. The main advantage is not the ability to identify variants in the 98% of the genome that is noncoding, but the avoidance of exon capture. Most variants in noncoding sequences that are not already covered by some exon capture kits are uninterpretable in the present state of knowledge, and few are expected to be important as high-penetrance causes of disease. But despite many technical improvements, capture remains an inefficient process (see **Figure 17.12**). There are many examples of coding sequence changes that were missed on exome sequencing but detected when the same samples were subjected to whole genome sequencing. An additional benefit of whole genome sequencing is the ability to detect and characterize structural variants, which is not possible from exome data. For most Mendelian conditions this is a somewhat marginal gain, but for analysis of tumors it is crucial.

For most conditions and in most laboratories, the main present choice is between using gene panels or exomes. Gene panels cost less to sequence, the analysis is much less laborious, and the risk of unwanted incidental findings (see below) is much reduced. A carefully designed physical gene panel may also give better coverage of the relevant exons than an exome (**Figure 20.4**). On the other hand, causative mutations in novel genes will be missed, and each panel requires constant revision and updating as new causative genes are identified. For the laboratory it may be inconvenient to have to maintain a whole series of panels with their associated work-flows, rather than just put everything through an exome sequencing pipeline. One option is to use virtual gene panels, sequencing exomes but only analyzing data from a list of genes thought relevant to the patient's condition.

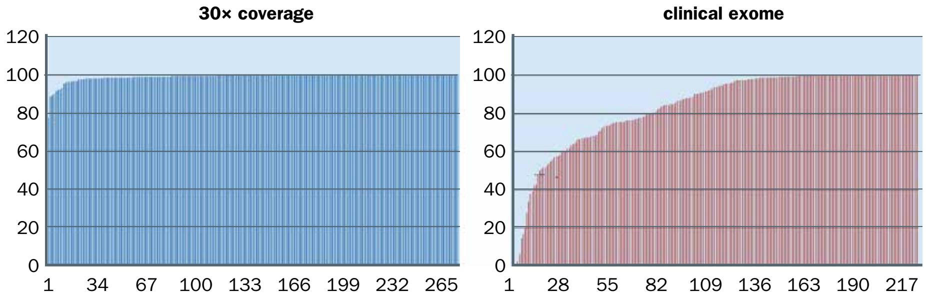


Figure 20.4 A carefully designed physical gene panel (left) may offer superior coverage. The vertical lines show the relative extent of coverage of individual exons achieved in one study in the Manchester laboratory, arranged in order of coverage. (Courtesy of Dr Sid Banka, St Mary's Hospital, Manchester.)

Filtering the list of variants

Once a list of validated variants—typically 20,000 for an exome, but considerably fewer for a gene panel—has been compiled, it must be filtered to identify any that might explain a patient's condition. The basic steps were outlined in Section 17.4. The emphasis there was on detecting novel genes, but in this area diagnostic and research procedures are very similar. As mentioned there, it is critical to distinguish the likely effect of a variant on a gene product from its likely effect on a patient. *In silico* tools like Polyphen and Sift, together with laboratory functional studies, help assess the effect on a gene and its product. But even complete loss of function of a gene is not necessarily pathogenic: the average subject in the ExAC database (see below) has 85 heterozygous and 35 homozygous protein-truncating variants.

In Section 17.5 we described the various ways one would attempt to show that a variant was pathogenic. That was in the context of identifying novel disease genes, but the same considerations apply to deciding whether a variant in a known disease gene explains a patient's condition. The prior probability may be higher than with a variant in a novel gene, but in a clinical context the final conclusion must be more certain. It is more difficult and painful for a patient to be told that a report was mistaken than for a researcher to have to admit to a small error in interpretation.

The three pillars of interpretation

As described in Section 17.5 the three main points to consider when interpreting a variant are precedent, conservation, and rarity. Regarding precedent it is important to realize that many variants labeled as pathogenic in the early heady days of disease gene identification are actually benign. A researcher would identify a candidate disease gene by linkage analysis and seek confirmation by checking a panel of unrelated affected people for mutations. If a reasonable number of people in the panel had variants in that gene the identification and full list of variants would be published, and each variant would be entered into databases as pathogenic. Provided overall enough of the variants were truly pathogenic the gene identification would be correct, but in fact, some of the missense variants might well be chance nonpathogenic variants. Through this process the mutation databases became badly contaminated with nonpathogenic missense variants.

- Precedent can be assessed through the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/). As described in Section 17.5, this contains every reportedly pathogenic variant in every gene, together with details of the phenotype. Where two laboratories have reported different interpretations of a certain variant the two are listed side by side, together with the evidence.
- Conservation of an altered amino acid in a missense variant is assessed using multiple sequence alignments. These can be generated by tools such as POLYPHEN and SIFT, described in **Box 16.2**. Changes to highly conserved amino acids are predicted to be damaging; changes to nonconserved residues are expected to be benign. These assessments are typically around 80% accurate. One cause of false negatives is compensated pathogenic deviation (see **Figure 17.14**). Different tools can give different answers, so it pays to use several and look for a consensus. Remember that they do not consider whether a sequence change might affect splicing (see **Figure 16.4**), nor whether a damaging effect on the protein will actually cause any problems for the whole person.
- Evidence that a change is pathogenic comes primarily from comparing its frequency in patients and controls. Ideally it would have been seen in many patients but no controls. If it is nevertheless present in some healthy people, then the frequency is important. If the condition is dominant, the frequency of the variant among controls must be lower than the frequency of the condition (thus allowing for a degree of reduced penetrance); for a recessive condition the frequency of a candidate variant among controls must be no higher than the estimated frequency of carriers. The ExAC database (https://exac.broadinstitute.org/), expanded in late 2016 to include 126,216 exomes and 15,136 genomes (http://gnomAD.broadinstitute.org/), is an invaluable and very powerful tool for showing that a variant is not pathogenic because it is too frequent among healthy controls. Many variants previously reported as pathogenic fail this test. The average ExAC subject has 54 variants labeled as pathogenic in major disease databases; most of these are present at implausibly high frequencies in one or another population and hence are probably benign.

Rules for clinical reporting

There is general consensus that laboratories should group variants into five categories:

1. Pathogenic;
2. Likely pathogenic;
3. Uncertain significance;
4. Likely benign;
5. Benign.

A working group of the American College of Medical Genetics has provided extensive guidance (see https://www.acmg.net/docs/standards_guidelines_for_the_interpretation_of_sequence_variants.pdf). They consider 16 lines of evidence suggesting that a variant might be pathogenic, and 12 lines suggesting it might be benign; they group these into very strong, strong, moderate, and supporting, and provide rules for making the overall judgement. They suggest that 'likely' in the 5-point classification should mean an estimated 90% certainty (some feel this percentage is too demanding).

The report goes into considerable detail on caveats and problems with each possible line of evidence and should be read carefully by anybody involved in diagnostic work. All these judgements would be applied with much greater caution if the variant in question was in a gene that had not previously been implicated in the patient's condition. Other professional bodies have come up with broadly similar recommendations. Laboratories would normally report variants in the pathogenic or likely pathogenic categories, but not those in the benign or likely benign group. The report by Tarailo-Graovac and colleagues (2016, PMID 27276562; see Further Reading) gives a good flavor of the current contribution of exome sequencing to clinical service.

What to do with the variants of uncertain significance (VUS) is a much-debated problem with no easy solutions. Reporting them presents the patient and the referring physician with unanswerable questions, can cause great anxiety, and may trigger uninformed Internet searches that leave the patient convinced that he will develop a serious disease. On the law of averages, most VUS will turn out to be benign. But not reporting them means there is no way to re-visit the variants in the light of increased knowledge, and it might turn out that some are both pathogenic and actionable (that is, something can be done to avoid or reduce the pathogenic effect). This question is different from the question of *incidental findings*, which are variants known to be pathogenic, but for conditions unrelated to the one for which the patient's DNA was sequenced—for example, a known cancer-pre-disposing variant found in a patient being tested to identify the cause of deteriorating vision. The problem of incidental findings is discussed below, but a common element in minimizing the problems of both VUS and incidental findings is to ensure that the patient understands the possibility of both, and before testing specifically consents to what classes of results will and will not be reported to him.

Nonsequencing methods are used to answer specialized questions

The steady advance of next-generation sequencing has rendered many previous approaches obsolete; however, not every question that a diagnostic laboratory might ask about a patient's DNA is best answered by current sequencing approaches.

Array-comparative genomic hybridization (array-CGH) is a popular method of checking for structural variants, copy number changes, etc.

Most laboratories no longer use karyotyping under the microscope to check for chromosome abnormalities. They use array-CGH for checking across the whole genome for variants including numerical and gross chromosome abnormalities and smaller structural variants, such as microdeletions or duplications and copy number changes. The principle of array-CGH was described in Section 15.1 (see **Figures 15.6** and **15.7**). The resolution (the smallest variant reliably detectable) depends on the number of probes on the array but would normally be in the range 20–60 kb. Array-CGH measures the dosage of sequences in a test DNA compared to a control sample; it cannot detect balanced abnormalities such as balanced translocations, which would be seen on standard (microscope-based) karyotyping. Whole genome sequencing could replace array-CGH for all these purposes if costs fall sufficiently.

Multiplex ligation-dependent probe amplification (MLPA) is widely used to check for deletions or duplications of whole exons

As noted in Section 16.1, most exons are small compared to most introns or the stretches of DNA between genes, and therefore most random breakpoints lie in intergenic DNA or in introns. Thus, deletions or duplications often involve complete exons. Homozygous deletions would be apparent in exome sequencing data; duplications or heterozygous deletions might be noted through the change in read depth of the sequences involved, but many laboratories would not wish to rely on this as the sole criterion on which to base a report. Diagnostic laboratories usually use extra tests to check for deletions or duplications of one or more whole exons. Real-time quantitative PCR (see Section 6.2 and **Box 7.5**) would be one possible method to use, but many laboratories favor MLPA for its reliability, simplicity, and suitability for multiplexing.

The MLPA technique was developed by scientists at the Dutch Medical Research Council, and that organization continues to develop and market MLPA probe sets (Schouten *et al.*, 2002, PMID 12060695; see Further Reading). It is an adaptation of the oligonucleotide ligation assay (see **Figure 20.3**, above). As with the OLA, each

MLPA probe consists of two oligonucleotides that hybridize to a target sequence, leaving a gap that can be sealed by DNA ligase. Unlike in the OLA, they are designed to hybridize to hopefully invariant sequences in a target exon. Whenever the target exon is present the probes can be ligated and then PCR-amplified. Probes for different exons have different-sized 'stuffer' sequences so that the fluorescently-labeled PCR products are of different sizes and can all be distinguished on the output from a gene analyzer.

MLPA can be multiplexed to check 40 target exons in a single reaction, using 50–100 ng of DNA. Probe sets are designed as far as possible to give equal amounts of product from each exon in a multiplex, but that can never be fully achieved. Deletions and duplications are diagnosed by comparing the amount of product from the target exon to the amount from other exons and to the value from a wild-type sample (**Figure 20.5**).

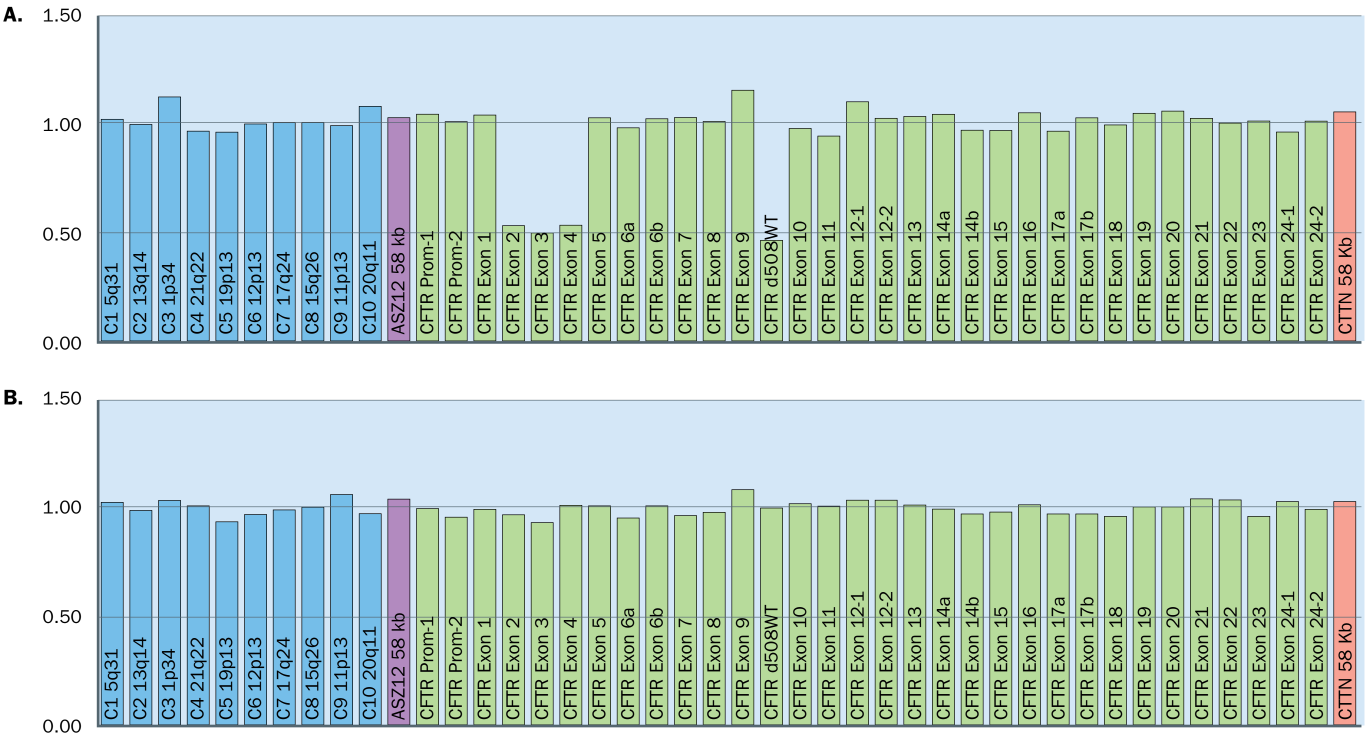


Figure 20.5 In a patient with cystic fibrosis, MLPA reveals compound heterozygosity for the p.F508del mutation and deletion of exons 2–4 of the *CFTR* gene. (**A**) Trace from patient; (**B**) control normal trace. This MLPA kit includes a specific test for the frequent p.F508del mutation as well as for dosage of the *CFTR* promoter and exons 1–24. Blue bars are control sequences from elsewhere in the genome. *ASZ1* and *CTTNBP2* genes (purple and pink bars) flank *CFTR* at 7q31. Note that before reporting that the causative changes had been identified it would be necessary to check parental samples to make sure that the deletion and mutation are in *trans*, that is on different chromosomes, rather than in *cis* where both are inherited from the same parent. (Courtesy of Dr Simon Ramsden, St Mary's Hospital, Manchester.)

Each individual MLPA ligation tests just the 50 nucleotides or so to which the two oligonucleotides hybridize, although it is normally assumed that if that sequence is deleted or duplicated, so is the rest of that exon. Exome sequencing would have revealed any partial deletion or duplication. MLPA is not suitable for checking for deletions or duplications of a whole gene or more.

RNA analysis is used to check for aberrant gene expression or splicing effects

RNA is normally analyzed by making and sequencing cDNA (see Section 6.2). Specific RNA species can be checked by qualitative or quantitative RT-PCR (see **Box 7.5**). For investigating the overall transcriptome (see Section 6.3), expression microarrays were formerly used but sequencing of bulk cDNA can handle a much wider range of expression levels and is not restricted to the particular species that the microarray was designed to detect. RNA is more widely studied in research than for clinical diagnosis; applications in diagnosis might include characterizing tumors and checking a mRNA for splice variants.

Methylation analysis may be used for characterizing tumors and imprinted genes

DNA methylation is a factor controlling gene expression (see Section 10.3). Sometimes a diagnostic laboratory needs to check the overall level of methylation or the methylation status of particular sequences in a tumor biopsy. Other applications include checking expanded fragile X genes, or sequences that control imprinting (see Section 10.4). The techniques to study DNA methylation were described in Chapter 10. In brief, methylation-sensitive restriction enzymes such as *Msp*1 can be used to check methylation of specific cytosines, but the general method is to use sodium bisulfite to convert cytosines, but not 5-methyl cytosines, to uracil, and then PCR-amplify and/or sequence the product (see **Box 10.3**).

Tests must always be rigorously evaluated before being introduced into clinical services

Any proposed test (not just any genetic test) performed for clinical purposes should be evaluated, not only for laboratory aspects, but for the overall effect of introducing the test. The ACCE framework considers four aspects:

- **A**nalytical validity: how well does the test measure what it claims to measure?
- **C**linical validity: how well does the test predict the health outcome that it claims to? This primarily depends on the strength of the genotype–phenotype correlation—a matter discussed in Section 16.5;
- **C**linical utility: what clinical use is the result? Will it lead to any change in management or treatment?
- **E**thical, legal, and social aspects: does the test conform to legal and ethical standards, and is it socially acceptable?

Sometimes a distinction is made between an *assay* and a *test*. An assay is the actual laboratory result—a genotype or measurement of an analyte—whereas a test, in this context, is the overall result of deriving and assembling the information on which to base a clinical decision. For example, the assay could cover reporting of individual genotypes, while the test might include the overall assessment contained in a laboratory report. The distinction can be helpful in avoiding overemphasis on the assay in determining the usefulness of a test.

In the end, the value of a test must be judged by the changes it brings about. Whatever the performance of a test according to the criteria outlined above, in the end what matters is what it achieves for patients. Does it change what people do? For most conditions there will be some threshold for action—for taking X-rays, doing a biopsy, or prescribing a drug. The result of any single test will be considered alongside all other relevant information—the age and sex of the patient, their general health and social situation, and the results of any other tests. A useful test is one that makes a substantial contribution to moving people across the action threshold, in either direction. In this connection, absolute risk is much more important than relative risk. A test may alter somebody's risk tenfold, but if the effect is only to change the risk from 1 in 10,000 to 1 in 1000 it will probably not make any practical difference. On the other hand, a positive test for a *BRCA1* mutation has a relative risk of only about 7 (80% for a carrier versus 12% general population risk), but the test result is important because of the high absolute risk.

20.4 POPULATION SCREENING

The word *screening* is often used loosely as a synonym for testing, but we will use it here in a more restricted sense, meaning large-scale testing performed as part of a program intended to improve public health. Screening, as defined here, is a top-down process, offered by some public authority to whole populations (or maybe specific large subsets of the population), without reference to the health of the individuals involved. This contrasts with conventional genetic testing, which is a bottom-up process initiated by individuals or their physicians to answer specific questions triggered by their symptoms or family history.

Screening tests are not diagnostic tests. Although the distinction can sometimes get blurred, in general the aim of a screening program is to define a high-risk group, who would then be offered a definitive diagnostic test (**Figure 20.6**). Screening programs, as defined here, raise particular issues that are not raised by testing of individuals.

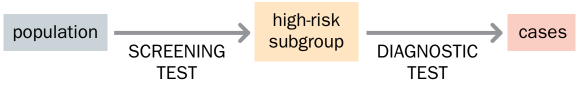


Figure 20.6 The aim of a screening test is usually to identify high-risk individuals who can then be offered a definitive diagnostic test.

Because these are large, centralized programs, the bodies running them have to assess carefully whether or not to spend the money to implement any proposal for screening. Screening programs are matters of formal policy, not individual whims. Individuals may be quixotic in their decisions whether or not to take a genetic test, but central authorities—whether government agencies or insurance companies—have to make formal assessments of proposals for screening.

Quantifying the performance of a test

The basic parameters describing the performance of a test are set out in **Box 20.2**. For any test that is less than perfect it is necessary to set a cut-off—a value above which people are treated as positive and some further action is triggered, but below which no further action is taken. Inevitably there is a trade-off between sensitivity and specificity. A very low cut-off that ensures that all affected people are detected will necessarily mean there will also be many false positives, given the imperfection of the test. Those can be minimized by setting a higher (more stringent) cut-off, but only at the expense of missing some affected people. **Table 20.4** shows that a test that performs well in the laboratory may perform very badly in large-scale population screening. The quaintly named receiver operating characteristic (ROC) curve quantifies the trade-off (see **Box 20.3**) and provides a criterion for judging the value of a test.

BOX 20.2 DEFINING THE PERFORMANCE OF A TEST

If a test aims to detect people with a certain condition, people may score positive or negative on the test, and they may or may not actually have the condition in question. In the Table, a, b, c, and d are actual numbers of people in each category.

	Condition present	Condition absent
Positive on test	a	b
Negative on test	c	d

Sensitivity = a/(a+c). The sensitivity of the test is the proportion of people who actually have the condition that the test picks up.

Specificity = d/(b+d). The specificity of the test is the proportion of people who do not have the condition who are correctly identified by the test.

Positive predictive value = a/(a+b). The positive predictive value is the proportion of people testing positive who actually have the condition.

False-positive rate = b/(a+b+c+d). The false-positive rate is the proportion of all people tested who wrongly test positive for the condition.

False-negative rate = c/(a+b+c+d). The false-negative rate is the proportion of all people tested who wrongly test negative for the condition.

Odds ratio = ad/bc. The odds ratio gives the relative odds of a person having the condition depending whether their test result is positive or negative. See **Box 18.2** for a discussion of the somewhat counter-intuitive properties of this measure.

TABLE 20.4 A TEST THAT PERFORMS WELL IN THE LABORATORY MAY BE USELESS FOR POPULATION SCREENING

Prevalence of condition	True positives in 1 million people screened	True positives detected by screening	True negatives in one million people screened	False positives detected by screening	Positive predictive value
1/100,000	10	10	999,990	10,000	0.001
1/10,000	100	99	999,900	9999	0.0098
1/1000	1000	990	999,000	9990	0.09

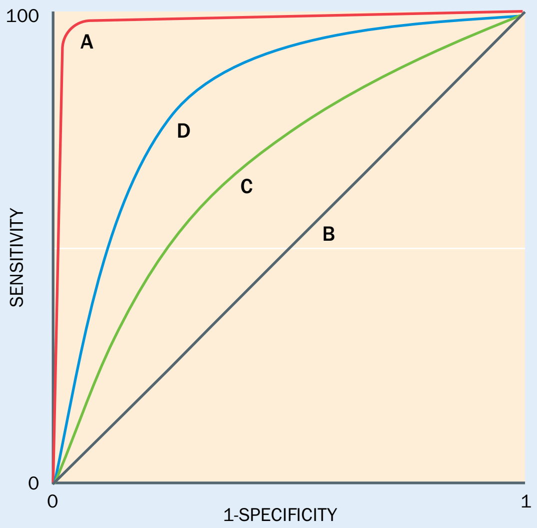
In a laboratory trial of this hypothetical test, 99/100 cases and only 1/100 controls scored positive. The table shows the predicted result of screening 1 million people, for different prevalences of the condition tested. The rarer the target condition, the more crucial is the level of false positives. These include not just samples truly giving a positive result, but also false results due to sample mix-ups, cross-contamination, etc.

BOX 20.3 THE RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

For a given test, the ROC curve (**Figure 1**) shows how varying the cut-off level above which a result is counted as positive affects the number of false-positive results. The sensitivity (proportion of true cases that are above the cut-off) is graphed against the proportion of true negatives that fall above the cut-off and are wrongly labelled as positive.

- Curve A (red) shows an ideal test. A cut-off can be chosen that detects almost 100% of cases while giving hardly any false positives.
- Curve B (black) shows a totally useless test. It completely fails to distinguish between cases and non-cases. If the chosen cut-off identifies x% of cases, it also wrongly labels x% of non cases as positive.
- Curve C (green) shows a marginally useful test. For example, a cut-off that detects 50% of cases would label only 22% of negatives as false positives.
- Curve D (blue) shows a rather better test. Only 10% of negative people are labelled as positive with a cut-off that detects 50% of true cases.

The result can be summarized by the area under the curve (AUC, sometimes called the C-statistic). For curves A, B, C, and D the AUCs are almost 1, 0.5, 0.61, and 0.82, respectively.



Box 20.3 Figure 1 The ROC curve summarizes the predictive power of a test.

Before a test is implemented in a population screening program a broader set of questions must be considered. The criteria used are generally based on proposals made for the World Health Organisation by Wilson and Jungner in 1968. For example in the UK, the National Screening Committee has a list of 20 criteria that should be fulfilled by any screening program conducted through the National Health Service. The full list can be seen at https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes. **Table 20.5** summarizes some general requirements.

TABLE 20.5 REQUIREMENTS FOR A POPULATION SCREENING PROGRAM

Requirement	Examples and comments
A positive result must enable some useful action	Preventive treatment, e.g. special diet for PKU; review and choice of reproductive options in cystic fibrosis carrier screening
The test must have high sensitivity and specificity	Tests with many false negatives undermine confidence in the program; tests with many false positives, even if these are subsequently filtered out by a definitive diagnostic test, can create unacceptably high levels of anxiety
The whole program must be socially and ethically acceptable	Subjects must opt in with informed consent; there must be no pressure to terminate affected pregnancies; screening must not be seen as discriminatory
The benefits of the program must outweigh its costs	It is unethical to use limited health care budgets in an inefficient way

Screening for genetic conditions can be carried out at different times of life

Genetic screening might involve any of the following:

- Prenatal screening;
- Newborn screening;
- Pre-conception screening for carrier status;
- Screening adults for susceptibility to late-onset diseases.

Clinical genetics textbooks or websites should be consulted for lists of conditions and details of programs; here we will use selected examples to bring out some of the technical, personal, and ethical issues involved. Always it is important to distinguish population-based screening programs from individual tests performed because of a person's family history or symptoms.

Prenatal screening for Down syndrome exemplifies some of the real-world issues

Pregnant women are naturally anxious that their baby should be healthy, and a number of prenatal tests for fetal abnormality are available. For chromosomal abnormalities, the traditional diagnostic test requires a sample of fetal cells to be obtained. This is done by chorionic villus biopsy at 10–14 weeks of gestation or by amniocentesis at 16–20 weeks. Both of these procedures are invasive, unpleasant for the mother, are expensive, and carry an approximately 1% risk of triggering a miscarriage. For women at high risk of having a baby with a chromosomal abnormality, these disadvantages may be worth disregarding, but how are high-risk women to be identified?

Some women are at high risk because they or their partner carry a balanced chromosomal translocation (see **Figure 15.14**), but for most women, age is the main risk factor. The probability of having a baby with Down syndrome or another numerical chromosome abnormality rises sharply with the age of the mother (**Table 20.6**). Until fairly recently, it was standard policy in most countries to offer amniocentesis or chorionic villus biopsy to all women over some certain age (usually in the range 35–38 years). But this is a very inefficient method of detecting fetuses with Down syndrome. Although the individual risk is much higher for older mothers, it is not negligible for younger women and, because most babies are born to younger women, so too are most babies with Down syndrome.

TABLE 20.6 RISK OF DOWN SYNDROME DEPENDS ON THE AGE OF THE MOTHER

Mother's age	20	30	34	36	38	40	42	45
Risk: 1 in...	1500	900	500	300	200	100	60	30

Over the years, several biomarkers have been identified in maternal serum that show different, although strongly overlapping, distributions in pregnancies in which the fetus does or does not have Down syndrome. These include levels of alpha-fetoprotein, unconjugated estriol, human chorionic gonadotropin, inhibin-A, and PAPP-A (pregnancy-associated plasma protein A). A noninvasive ultrasound test, the nuchal translucency test, that looks at the neck of the fetus also gives an indicative risk. None of these would individually be useful as a screening test but combining age and measures of several biomarkers can give a valuable composite risk. A policy decision must then be made about what threshold value of the composite risk should trigger intervention. There is a trade-off between sensitivity and specificity. Sensitivity can be maximized only at the cost of plunging more couples into considerable stress and anxiety, as well as triggering risky and invasive diagnostic testing. In contrast, minimizing the number of false positives means that more women who have been given a negative test result will nevertheless go on to have a baby with Down syndrome. The FASTER (First and Second Trimester Evaluation of Risk; Malone *et al*, 2005, PMID 16282175; see Further Reading) study by the USA National Institutes of Health compared several possible protocols. In many programs, the threshold for offering an invasive diagnostic test would be a composite risk of Down syndrome of 1 in 300 or greater. Various options in the FASTER study gave detection rates of 90–95% for a false-positive rate of 5%.

Noninvasive prenatal testing (NIPT)

The development of noninvasive prenatal testing (NIPT) has provided fresh options. The blood of a pregnant woman contains cell-free DNA, around 5–10% of which usually derives from the placenta, a fetal tissue. It also contains a few intact cells of fetal origin but attempts to use those for prenatal diagnosis were put aside because of the difficulty of reliably isolating them in sufficient numbers. When the DNA is sequenced to sufficient depth, if the fetus has an extra chromosome, sequences from that chromosome will be present in slightly greater relative amount in the bulk cell-free DNA, compared to sequences from chromosomes present in normal diploid numbers (**Figure 20.7**).

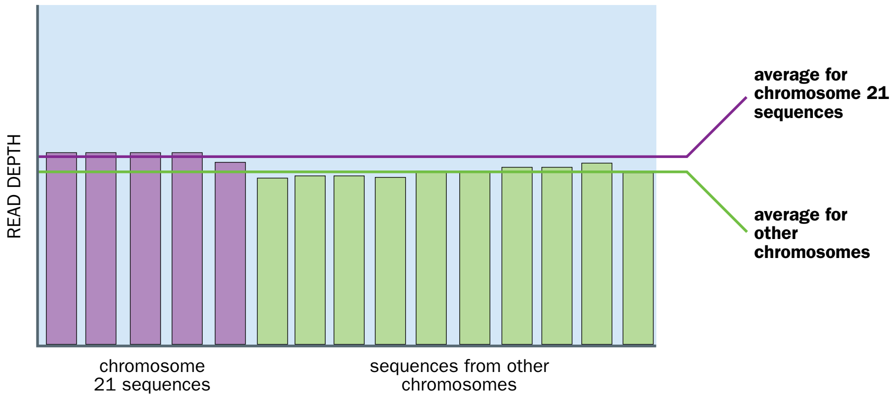


Figure 20.7 Noninvasive prenatal testing for Down syndrome. When the fetus has trisomy 21, sequencing cell-free DNA in the maternal blood (a small percentage of which is of fetal origin) shows a small but significant excess of reads from chromosome 21.

Provided the cell-free DNA includes at least 4–5% of fetal DNA, NIPT has very high sensitivity from around 9 weeks of gestation onwards. It can also in principle reveal much more information about the fetus. It allows reliable fetal sexing, is widely used for Rhesus typing to anticipate problems with hemolytic disease of the newborn and could be used to check for specific mutations or screen for any number of variants (but not variants present in the mother). In private healthcare systems market-driven companies offer a range of analyses. In the UK the National Health Service is much more cautious. Currently (2018) general pregnancy screening is still based on biomarkers plus ultrasound measurement of nuchal translucency, but women whose composite risk from these analyses comes out as 1 in 150 or greater are then offered NIPT. Restricting screening to this high-risk group improves the predictive value of a positive result (as well as saving money and easing the introduction of a novel system). This is still currently regarded as screening, not a definitive diagnostic test. Women who test positive are referred for invasive testing by chorion villus biopsy or amniocentesis, but the number of women requiring those procedures is greatly reduced compared to the old system.

Ethical questions are central to discussions of prenatal screening

Ethical issues loom very large in discussions of prenatal screening. The general point of detecting a fetal anomaly is to enable the couple concerned to do something about it. Some couples will simply want to know, in order better to prepare themselves for coping with a baby with problems, but in general, given the limited opportunities for prenatal treatment, the decision is whether or not to terminate the pregnancy. Clinical geneticists object strongly to being portrayed as carrying out a search-and-destroy operation against abnormal fetuses. Their aim is to allow couples as far as possible to make their own fully informed decisions consistent with their own values and with the laws of their country. This means that couples must opt in to screening with informed consent and awareness of the possible outcomes, and if the result shows an abnormality, must not be pressured in any way to terminate the pregnancy. In reality it can be difficult to ensure couples in a busy antenatal clinic can give fully informed consent to a technical procedure with complex implications, but that is no excuse for not trying.

Some disability rights campaigners and patient groups argue that offering prenatal diagnosis for a condition, with the option of terminating the pregnancy if the fetus is affected, is equivalent to saying that affected people are worthless and should not be alive. On the other side, many people maintain that wanting a healthy baby is not incompatible with loving a child that is born with a disability. This argument has special force for couples who already have one affected child. They often request prenatal testing, saying that however much they love that child, they could not cope with having a second affected child. Whatever one's position on this argument, all civilized people must surely agree that society has an obligation to look after people born with disabilities and do whatever is possible to allow them a full life.

Newborn screening programs are aimed at detecting treatable conditions

Newborn screening is much less contentious than prenatal screening. Every country has a list of conditions (not all genetic) for which every newborn baby should be checked. Conditions are chosen because treatments are available and early diagnosis has been shown to lead to improved outcomes. In addition, early information can also lead to knowledge of the condition for the family, thus avoiding a potential diagnostic odyssey or inappropriate therapies. To quote the American College of Medical Genetics (ACMG) "Newborn screening is more than testing. It is a co-ordinated and comprehensive system consisting of education, screening, follow-up, diagnosis, treatment and management, and program evaluation". The ACMG recommends screening for 29 conditions, mostly inborn errors of metabolism (see http://www.acmg.net/resources/policies/NBS/). The UK National Health Service more cautiously screens for 9 genetic conditions (sickle cell disease, CF, congenital hypothyroidism, and 6 inborn errors of metabolism).

Phenylketonuria (PKU) is on the list in every advanced country. The screening test measures the level of phenylalanine in a blood spot taken by pricking the baby's heel. Note that, as with almost all screening tests, the PKU screen is not DNA-based because many different loss of function mutations can cause PKU. Testing cannot be done immediately after birth because while the baby is *in utero* excess phenylalanine crosses the placenta and is cleared by the mother (who would normally be a heterozygous carrier). Ideally the sample would be taken 5 days after birth to give time for the level to build up, but provided testing is delayed for at least 24 hours there are few false negatives.

A positive screening test triggers diagnostic testing. Babies testing positive may have benign hyperphenylalaninemia, requiring no further action, or may have a variant form due to tetrahydrobiopterin deficiency. The latter is a rare but very severe condition requiring specific dietary supplements. Treatment for classic PKU involves restricting dietary phenylalanine by a special low-protein diet, with supplements of the other amino acids. Careful monitoring is necessary—phenylalanine is an essential amino acid and the baby needs enough to support normal growth, but not enough to accumulate and damage its developing brain. Untreated PKU usually results in severe intellectual disability, often also with behavioral problems and epilepsy. Clinical opinions differ as to whether the diet should be maintained throughout life or whether it might be discontinued once brain growth is complete. Regardless, a woman with PKU needs to go back on the diet during pregnancy, otherwise her excessive phenylalanine will cross the placenta and cause severe brain damage in her fetus, even though the fetus itself would most likely be just heterozygous.

People contemplating reproduction might be screened for carrier status for recessive conditions

We are all carriers of one or more severe recessive disorders. People might wish to know which disorders, either in order to avoid partnering somebody who carries the same disorder, or to allow prenatal diagnosis if both partners are carriers. Carriers are usually phenotypically normal, and often do not show clear biochemical differences from non-carriers. Thus, carrier testing usually has to be done by genotyping.

Recessive conditions are almost always due to loss-of-function mutations, and so often show extensive allelic heterogeneity (see Section 16.5). Where a recessive condition is common in a particular population there may be one or a few founder mutations (see Section 12.3) that can be checked by a specific DNA test. However, such tests for specific variants will miss some individuals who carry a different variant. For example, a CF carrier screening program based on detecting only the common p.F508del *CFTR* mutation would pick up only 70–80% of Northern European carriers of CF. Thus, comprehensive DNA-based carrier testing only became possible with the availability of exome sequencing.

Tay–Sachs screening illustrates the arguments about biochemical versus DNA screening. This severe autosomal recessive condition (MIM 272800) is lethal in early childhood. It is generally rare, but among Ashkenazi Jews the carrier frequency is about 1 in 30 due to a founder effect. In many countries Jewish communities have organized carrier screening programs based on an assay of hexosaminidase-A enzyme activity. Among people of pure Ashkenazi descent, three specific hexosaminidase A mutations account for 92–98% of carriers. This allows efficient DNA-based screening using buccal smears. However, as the Ashkenazi community has opened up more over the past few generations, the proportion of self-identifying Ashkenazi Tay–Sachs carriers who have those specific mutations has decreased. Thus, it is generally recommended to stay with the enzyme-based carrier test. Some individuals show an intermediate enzyme level, and some have a variant enzyme that is functional *in vivo* but fails to work on the artificial substrate used in the test (a pseudodeficiency allele), and so confirmation by DNA analysis is recommended. In any case, any couple wanting prenatal diagnosis needs to have their mutations defined.

Carrier testing, for this or any other recessive disease, could be carried out on individuals of any age. Parents of a child with a recessive condition quite often want to know whether their healthy children are carriers. The general feeling among clinical geneticists is that such requests should be resisted. There is no benefit to a child in knowing its carrier status, and it should retain the freedom to decide for itself whether or not to have carrier testing at an appropriate age. Carrier screening of newborns is still less desirable. However, a hospital may use exome sequencing to arrive at a diagnosis on a sick infant, and the result will inevitably identify any conditions for which it is a carrier. In those circumstances the clinician will probably feel bound to report those results to the parents.

A second ethical concern is about the risk of stigmatizing people who turn out to be carriers. Geneticists understand that this has no sinister implications, but other people may not ('Joe Brown is a mutant…!'). This is particularly a problem for children and is another reason not to test children. Among Orthodox Jewish communities where marriages are semi-arranged, the result of carrier testing for Tay–Sachs and a number of other 'Jewish' recessive diseases is not communicated to the young person involved, but is given to a match-maker who will know the results for both potential partners and indicate if a proposed marriage would be inadvisable. Most young people in more open communities would probably not accept such guidance but might still appreciate advance notice that they might wish to have prenatal diagnosis.

Screening for susceptibility to multifactorial conditions

In Section 18.3 we saw how numerous genetic factors have been identified that contribute to susceptibility to common late-onset diseases. For many of these conditions lifestyle factors are also relevant, and public health doctors spend much time trying to persuade us to adopt healthier lifestyles to reduce our risk (and costs to the public purse). If genetic screening could identify individuals at particularly high risk, the lifestyle advice could be much better targeted and hopefully more effective. Section 18.5 casts doubt on our ability to generate useful predictions of individual risk, but what if screening programs could identify high-risk subsets of the population?

A number of prospective studies have addressed this question. For example, several studies have looked at the ability to define a set of people at high risk of type 2 diabetes. A cohort of currently healthy individuals is recruited. Baseline measurements are used to predict their risk. These might include age, sex, fasting glucose level, and body mass index. Subjects are then followed for many years to see who does actually develop the condition. They are also genotyped for a range of susceptibility factors (the genotyping could not be done at the beginning because the markers had not been defined at that time, but this doesn't matter because a person's genotypes do not change over time) and the question is asked, how much better would the prediction have been if it had included genotype data? **Table 20.7** shows results of four such studies. The ability to predict is measured by the AUC, the area under the ROC curve (see **Box 20.3**). The question to ask is how much in each individual study is the AUC improved by adding in the genotype data?

TABLE 20.7 PREDICTING WHO WILL DEVELOP TYPE 2 DIABETES

Clinical indicators	AUC from clinical indicators	Susceptibility loci genotyped	AUC from clinical + genetic indicators	Reference
Age, sex, BMI	0.78	18 loci	0.80	Lango *et al.* (2008) PMID 18591388
Age, sex, BMI	0.66	18 loci	0.68	Van Hoek *et al.* (2008) PMID 18694974
Age, sex, BMI, family history, liver enzyme levels, smoking, measures of insulin secretion and action	0.74	16 loci	0.75	Lyssenko *et al.* (2008) PMID 19020324
Age, sex, BMI, family history, blood pressure, blood glucose, HDL cholesterol, triglycerides	0.903	62 loci	0.906	Vassy *et al.* (2014) PMID 24520119

Once sufficient baseline clinical indicators have been taken into account, genotyping for known susceptibility factors adds little to the ability to predict, as measured by the change in the AUC statistic. (See Hivert *et al.* [2014], PMID 24535206 in Further Reading for further discussion.)

The clear conclusion from the data in **Table 20.7** is that there is little mileage in population screening for genetic susceptibility factors for type 2 diabetes—a disappointing conclusion, given the immense public health problem posed by the rapidly increasing incidence of this disease. A modeling study of breast cancer screening gave a somewhat more positive message (**Figure 20.8**).

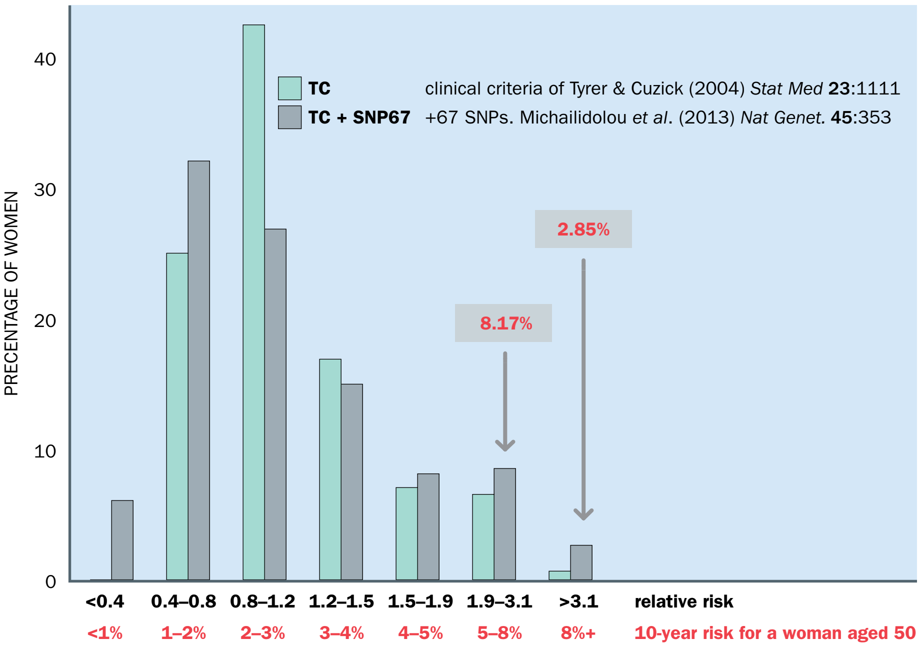


Figure 20.8 Genotyping for susceptibility-associated SNPs can modify a woman's estimated risk of breast cancer. Using SNP genotypes in additional to clinical criteria can identify subsets of healthy women who are at significantly increased or decreased risk of developing breast cancer. This information could be used to modify policies for routine mammographic screening. (Adapted by permission from Springer Nature on behalf of Cancer Research UK: Brentnall *et al.* [2014] *Br J Cancer* **110**:827; PMID 24448363. Copyright © 2014.)

Testing for risk of Alzheimer disease has been particularly controversial. A few percent of cases have onset before age 60. In these cases the condition is often inherited as an autosomal dominant Mendelian condition caused by mutations in the *PSEN1*, *PSEN2*, or *APP* genes. For these cases genetic testing is standard, following protocols similar to those developed for predictive testing for Huntington disease. Most Alzheimer disease is of late onset and is not Mendelian. However, a strong susceptibility factor is known. The E4 allele of *APOE* has a frequency of 0.07–0.15 in many populations and is a significant risk factor. Heterozygotes and homozygotes for E4 have about a twofold and a tenfold risk, respectively, of developing the disease. This is a much more significant effect than with most of the susceptibility factors identified for other common diseases. Older people naturally worry about their risk of becoming demented, but there are currently no drugs that significantly either delay the onset or reduce the severity of Alzheimer disease. Several professional bodies have advised against using APOE as a predictive test because of its poor sensitivity and specificity. There is a substantial risk of giving either false reassurance or a wrong bad prognosis. Some people think this attitude is over-paternalistic: people should be trusted to make their own decisions (see Green *et al.*, [2009] PMID 19605829, in Further Reading, for example). Attitudes would be very different if there were an effective treatment to delay or prevent onset of the condition.

In conclusion, it seems unlikely that measures of genetic susceptibility will play a major role in population screening programs, but as we move into an era where most people have already had their genomes sequenced, there will probably be specific programs where it proves useful to use the pre-existing genetic data.

Lifestyle genetic testing

Numerous companies, operating over the Internet, offer lifestyle genetic testing. You spit into a tube, send it away with your payment, and some time later you receive a report. It may cover purely recreational matters, like predicting your eye color and hair color; it may suggest your ethnic origin or offer to put you in touch with people with genotypes similar to yours; it may offer advice as to whether you would be more suited to sprinting or endurance running (probably without knowing that you have only one leg); and it may suggest your risk of a variety of common late-onset diseases. These predictions are usually based on SNP genotyping, although increasingly genome sequencing may be used.

They are usually made without any knowledge of your current health, your medical history, or your family history. Kalf and colleagues (2014, PMID 23807614, see Further Reading) report a thought-provoking study on the ways different companies calculate disease risks.

Applying the ACCE framework (see above), hopefully the company is able to deliver good analytical validity—that is, your genotypes are what they say they are—although it is worth noting that, unlike clinical laboratories, they are unlikely to be enrolled in any external quality assurance scheme. Clinical validity is another thing altogether. Even if the SNPs they use have been well validated as susceptibility factors (and in your ethnic group), we have seen that such factors have extremely weak predictive power for individuals. Evidence for the clinical validity of the great majority of tests offered in this way is nonexistent, but it is a fair assumption that it would be extremely low.

If a relative risk is given, it is worth asking two questions:

- How large is the relative risk? Is it 20, 10, 5 or 1.1? In most cases it will be much nearer 1.1 (a 10% increase) than 20. Do you care about a 10% increased risk of developing obesity?
- Does the report give your absolute risk as well as your relative risk? If the absolute risk of an outcome is very low, then even quite a large relative risk may not matter. And remember that risks calculated without knowledge of your medical and family history and your lifestyle mean very little.

Nobody would dispute a person's right to find out about their own genome, provided they do it at their own expense, but it should be regarded as a strictly recreational activity, like tracing family history, and not as any sort of medical investigation. Despite all these negative conclusions, there is some scientific value in all these activities. The largest companies have genotype data on millions of people. People who use their services are also more likely than average to be interested in genetic prediction and willing to take part in surveys. Thus, if the company puts out a call for people to report whether or not they or their family have some particular characteristic, they can very rapidly perform huge genome-wide association studies at negligible cost. With all the caveats about using unconfirmed self-reported phenotypic data, this can at very least produce hypotheses that can be tested in more controlled studies.

Incidental findings represent a sort of opportunistic screening

With the rapid growth of diagnostic exome and genome sequencing, the question what to do about incidental findings becomes acute. These are findings that may be clinically relevant but are not related to the diagnostic question originally asked. For example, a young person whose exome was sequenced to identify the cause of a retinal degeneration might be noted to have a mutation in the *BRCA2* gene. Should they be told? Shkedi-Rafid and colleagues in 2014 gave a wide-ranging discussion of the problem and referred to many reports and sets of guidelines addressing it (PMID 25228303, see Further Reading).

The first imperative is to prepare the ground in advance. When a patient consents to having their DNA analyzed they must be made aware of the possible results. They can then give informed consent. They certainly have the right to be given all the information generated, but they should be able to choose how much to take. They might consent to knowing everything, or to knowing all likely significant findings, or to knowing just all actionable findings (those where something can be done to avoid or reduce the risk), or to being told just results that are relevant to the diagnostic question being asked.

The American College of Medical Genetics caused great controversy when in 2013 they recommended that laboratories performing clinical sequencing should actively seek and report mutations in a list of 56 genes. This was to be done for all clinical germline (constitutional) exome and genome sequencing, including the "normal" of tumor-normal subtractive analyses in all subjects, regardless of the original indication and irrespective of age, but excluding fetal samples. There was to be no opt-out. For the details see Green *et al.* (2013, PMID 23788249) in Further Reading. Not every variant in the 56 genes was to be reported, only variants that had been previously reported and were a recognized cause of the relevant disorder, or previously unreported variants that were of the type expected to cause the disorder. Reporting was to the referring clinician, who presumably could then decide whether or not to communicate the findings to the patient. The list was drawn up after extensive consultation focusing on the clinical validity and clinical utility of the variants. Despite the care with which the recommendations were developed, the absence of any opt-out proved too controversial to uphold.

Later, despite the objections of some members of the ACMG working group, the recommendations were made optional.

One way of minimizing the problem of incidental findings is to use gene panels rather than the whole exome or whole genome. Only genes relevant to the diagnostic question would be analyzed. This might be done by selecting and sequencing just those genes, or by applying a filter to whole exome or whole genome data. The European Society of Human Genetics recommended this approach where possible, but also said that 'if the detection of an unsolicited genetic variant is indicative of serious health problems (either in the person tested or his or her close relatives) that allow for treatment or prevention, in principle a health-care professional should report such genetic variants' (see Van El *et al.*, [2013] PMID 23676617, in Further Reading).

20.5 PHARMACOGENETICS AND PERSONALIZED MEDICINE

All too often a generally useful drug will fail to work in certain people or will cause an adverse reaction. Even if a drug works, different individuals often require different doses to achieve the same therapeutic effect. Some of these differential effects are due to environmental causes: a person's ability to absorb or metabolize a drug may be changed by their illness or lifestyle (drinking, smoking, exercise, etc.). Sick people are often taking multiple drugs and some combinations may interact. But many differences are due to genetic variation between people. Pharmacogenetics and pharmacogenomics explore these effects. **Figure 20.9** summarizes the processes that can be affected by genetic variation.

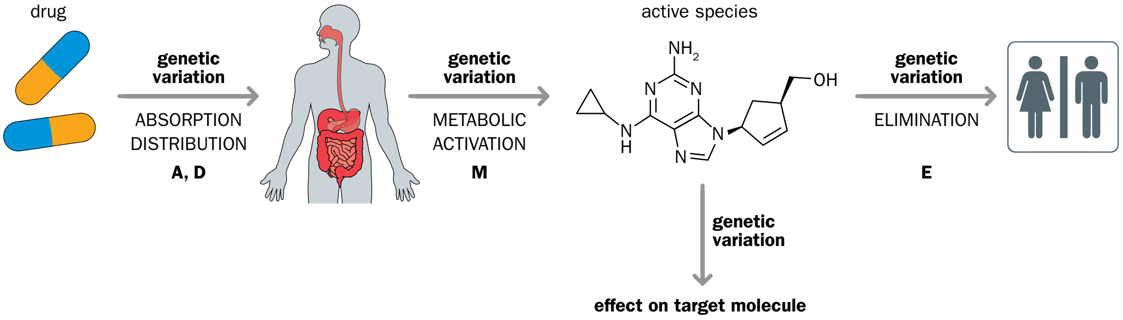


Figure 20.9 Pharmacogenetic variation can affect all aspects of drug action.

Pharmacogenetic effects can be divided into two categories:

- **Pharmacokinetics** covers genetic variations in the way a drug is absorbed, distributed, metabolized, and eliminated—in other words, what the person does to the drug;
- **Pharmacodynamics** covers genetic variation in the way a drug target responds to a given drug—in other words, what the drug does to the person.

The problem of adverse drug reactions

Adverse drug reactions (ADRs) are a serious problem. It has been estimated that in the USA they are responsible for about 100,000 deaths a year. In the UK, a study by Pirmohamed and colleagues (2004; PMID 15231615, see Further Reading) of 18,820 consecutive admissions to two large general hospitals in 2001–2 concluded that 6.5% of admissions were related to ADRs, and 2% of those patients died. The projected annual cost to the NHS was up to £466 million. **Table 20.8** shows examples. ADRs can be divided into two types:

- **Type A ADRs** (the great majority) are an exaggerated response to a standard dose of a drug, because of variants that make an individual particularly sensitive to that drug;
- **Type B ADRs** are unrelated to the normal action of the drug and are the consequence of some quite unexpected interaction. Type B reactions are rare, can be very serious and are hard to predict or understand.

TABLE 20.8 EXAMPLES OF ADVERSE DRUG REACTIONS

Drug	Effect
Abacavir	Serious and sometimes fatal hypersensitivity reactions in patients with HLA-B*5701 genotype
Azathioprine	Life-threatening bone marrow suppression from normal dose in people with low-activity thiopurine methyltransferase
Carbamazepine	Life-threatening Stevens–Johnson syndrome in East Asians with HLA-B*1502 and Europeans with HLA-A*3101 genotypes
Fluorouracil	Potentially fatal toxicity in people with deficiency of dihydropyrimidine dehydrogenase
Irinotecan	Severe neutropenia and diarrhea in people homozygous for a low-activity variant of the *UGT1A1* gene
Succinylcholine	Prolonged apnea in people with butyrylcholinesterase deficiency
Warfarin	Excessive bleeding in people with low-activity CYP2C9 or VKORC1
All are Type A reactions, dependent on the sorts of genetic variation described below, except for the adverse reactions with abacavir and carbamazepine, which are idiosyncratic Type B reactions.	

Many genetic differences affect the metabolism of drugs

The reactions of drug metabolism are traditionally divided into two phases. Phase 1 reactions (oxidation, hydroxylation, and hydrolysis) often produce the biologically active molecule, although sometimes the phase 1 product might be an intermediate in the inactivation and degradation of the drug. Phase 2 reactions (conjugation reactions such as acetylation, glucuronidation, or sulfation) produce a water-soluble compound that is more easily excreted. Not every drug is processed through both phases, but **Figure 20.10** shows a common sequence of events. Enzymes involved in both phases often show polymorphic variations in activity that affect responses to many drugs. The natural function of these enzymes is in handling xenobiotics (foreign substances), especially perhaps plant alkaloids present in the diet. A wide variety of enzymes are involved in these reactions (**Figure 20.11**).

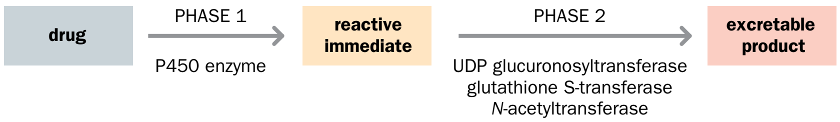


Figure 20.10 Stages in the metabolism of a drug. Phase 1 often involves oxidation, hydroxylation, or hydrolysis to produce a polar compound. This may be the active molecule itself or a degradation product. In Phase 2, the molecule may be conjugated with an acetyl, glucuronosyl, or glutathionyl group to form a water-soluble molecule that can be excreted.

The P450 cytochromes are responsible for much of the phase 1 metabolism of drugs

P450 cytochromes constitute a large family of enzymes that have an iron–sulfur active site and a spectral absorption peak at 450 nm. They act by inserting a single oxygen atom derived from molecular oxygen into a very wide range of organic compounds. The end product is usually a polar, hydroxylated derivative of the substrate. Humans have about 60 P450 genes, encoding enzymes that, between them, are responsible for the phase 1 metabolism of maybe 60% of all prescribed drugs. They are grouped into several families. Genes have names such as *CYP2D6* (cytochrome P450 family 2, subfamily D, polypeptide 6).

At least 10 different P450 enzymes have significant roles in phase 1 drug metabolism. Many of them can metabolize a considerable range of different drugs and have widely variable activity in different people. Individual drugs are often substrates for more than one P450 enzyme. These wide and overlapping specificities mean that there are seldom close correlations between any one P450 activity in a patient and their handling of a specific drug. Some drugs have an additional effect of inducing or inhibiting specific P450 enzymes, which can lead to unforeseen interactions between these drugs and those that are substrates of the enzyme.

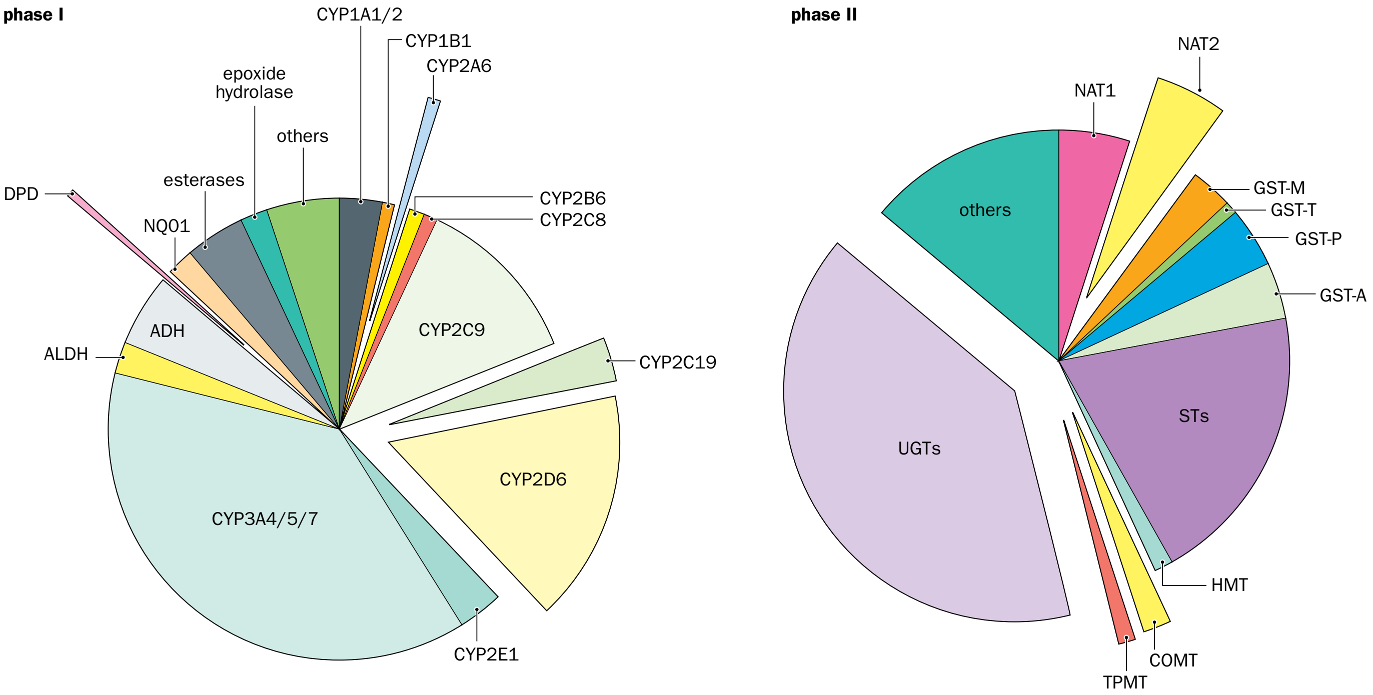


Figure 20.11 Enzymes involved in phase 1 and phase 2 drug metabolism. ADH, alcohol dehydrogenase; ALDH, aldehyde dehydrogenase; COMT, catechol O-methyltransferase; CYP, cytyochrome P450 enzymes; DPD, dihydropyrimidine dehydrogenase; GST, glutathione S-transferase; HMT, histone methyltransferase; NAT, N-acetyltransferase; NQO1, NAD(P)H quinone dehydrogenase 1; ST, sulfate transferase; TPMT, thiopurine methyltransferase; UGT, UDP-glucuronosyltransferase. (From Evans WE & Relling MV [1999] *Science* **286**:487–491; PMID 10521338. Reprinted with permission from the AAAS.)

The role of P450 enzymes first surfaced in the 1970s when it was noted that some individuals showed markedly enhanced sensitivity to the antihypertensive drug debrisoquine, and to the antiarrhythmic drug sparteine. On investigation, these individuals showed high plasma levels of the drugs but low urinary levels of the catabolism products, implying that they were failing to metabolize and excrete the drugs. The cause was eventually identified as low activity of the CYP2D6 enzyme. Individuals can be classified into poor, intermediate, extensive, and ultra-rapid metabolizers. Poor metabolizers have loss-of-function mutations in the *CYP2D6* gene, while ultra-rapid metabolizers have increased copy numbers of the gene (up to 13 copies). **Figure 20.12** shows how CYP2D6 activity governs the effective dose of the antidepressant drug nortriptyline, which is metabolized by this enzyme.

phenotype	ultra-rapid metabolizers	extensive metabolizers	intermediate metabolizers	poor metabolizers
frequency (Caucasians)	5–10%	65–80%	10–15%	5–10%

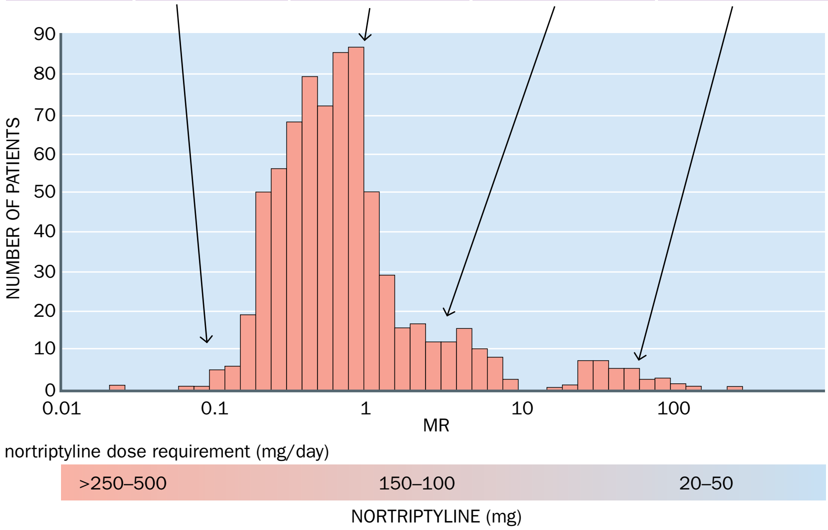


Figure 20.12 Pharmacogenetics of debrisoquine and nortriptyline. The activity of the CYP2D6 enzyme is measured by the metabolic ratio (MR), the ratio of amounts of a substrate drug and its metabolic product in urine after a standard dose of the drug. The graph shows observed ratios for a range of patients. Note the logarithmic scale of MR. High ratios show poor conversion due to low enzyme activity. The same enzyme is largely responsible for phase 1 catabolism of the antidepressant drug, nortriptyline. Depending on the CYP2D6 phenotype as measured by the MR, patients require different doses of nortriptyline (bottom bar). (Adapted from Meyer UA [2004] *Nature Rev Genetics* **5**:669–676; PMID 15372089. With permission from Springer Nature. Copyright © 2004.)

CYP2D6 is involved in the metabolism of perhaps 25% of all drugs. Variation in its activity has significant effects on the response to some beta-blockers used to treat hypertension and heart disease, and to several psychiatric drugs including tricyclic antidepressants. Poor metabolizers are at risk of overdose effects of these drugs. In contrast, CYP2D6 is also required to convert codeine into its active form, morphine. Codeine is ineffective in pain relief for poor metabolizers, whereas ultra-rapid metabolizers risk adverse effects such as sedation and impaired breathing. Other P450 enzymes also show variable activity, so that people can be classified into poor, intermediate, and extensive metabolizers (**Figure 20.13**), although they do not show the copy-number variation seen in CYP2D6 ultra-rapid metabolizers.

CYP2C9 hydroxylates drugs including nonsteroidal anti-inflammatory drugs, sulfonylureas, inhibitors of angiotensin-converting enzyme, and oral hypoglycemics. For example, rare poor metabolizers have an exaggerated response to tolbutamide, a hypoglycemic agent that is used to treat type 2 diabetes. The role of CYP2C9 variants in sensitivity to warfarin is discussed below.

CYP2C19 metabolizes drugs including anticonvulsants such as mephenytoin, proton pump inhibitors such as omeprazole (used to treat stomach ulcers), proguanil (an antimalarial), and certain antidepressants. Intermediate metabolizers are compound heterozygotes for an active and an inactive allele; poor metabolizers have two inactive alleles. The frequency of poor metabolizers is 3–5% among Caucasians and African-Americans, but higher in Orientals and very high in Polynesians. A particular risk for poor metabolizers is unacceptably prolonged sedation with a standard dose of diazepam as a result of slow demethylation by CYP2C19. In contrast, proguanil is a prodrug that requires activation by CYP2C19 to form the active molecule, cycloguanil; poor metabolizers therefore show a decreased effect of that drug.

CYP3A4 is the most abundant P450 cytochrome in liver and is involved in the metabolism of maybe 40% of all drugs. Its activity varies up to 30-fold between individuals. The enzyme is highly inducible by many different substances, and the variability is mainly due to variation in the upstream regulatory sequences governing inducibility, although there are also some coding sequence variants.

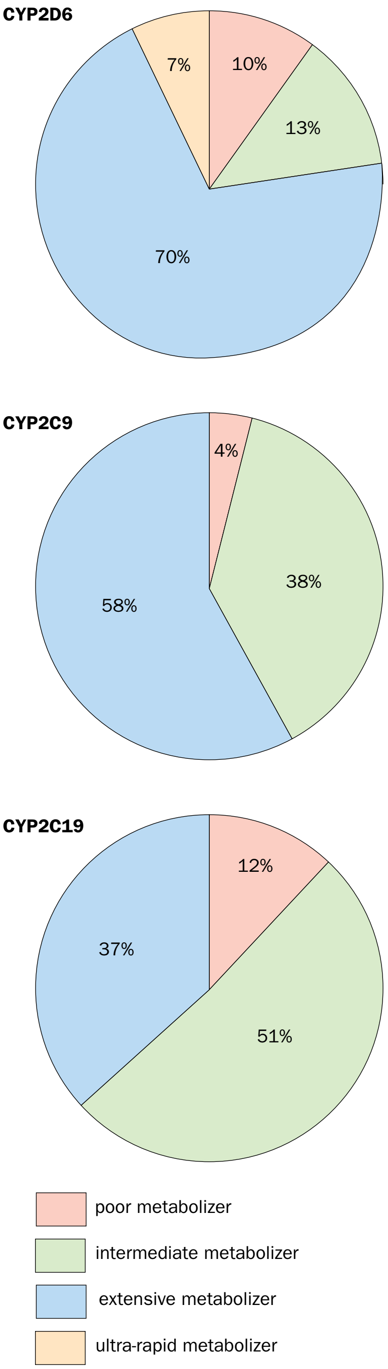


Figure 20.13 Population frequencies of P450 activity variants. The distributions differ in different populations. (Data for CYP2D6 adapted from Meyer UA [2004] *Nat Rev Genet* **5**:669–676, PMID: 15372089; data for CYP2C9 for northern Europeans from Service RF [2005] *Science* **308**:1858–1860, PMID: 15976283; data for CYP2C19 for Taiwanese from Liou YH, Lin CT, Wu YJ, Wu LS [2006] *J Hum Genet* **51**:857–863, PMID: 16924387.)

Another phase 1 enzyme variant causes a problem in surgery

Suxamethonium (succinyl choline) is a muscle relaxant used in surgery. About 1 in 3500 Europeans suffers prolonged apnea (failure to breathe spontaneously) after a standard dose. Spontaneous breathing resumes when the drug is inactivated by the enzyme butyrylcholinesterase (also known as pseudocholinesterase). The prolonged effect is seen in people who are homozygous for low-activity variants of the enzyme and who therefore inactivate suxamethonium abnormally slowly.

Phase 2 conjugation reactions produce excretable water-soluble derivatives of a drug

Phase 2 reactions can involve acetylation, glucuronidation, sulfation or methylation of the drug. Genetic variations in N-acetyltransferases, glutathione-S-transferases and UDP-glucuronosyltransferases underlie many variable phase 2 activities. People with deficient phase 2 activity inactivate and excrete the relevant drugs abnormally slowly.

Humans have two aryl-*N*-acetyltransferase enzymes, each involved in phase 2 metabolism but for different spectra of drugs. They are encoded by the highly homologous *NAT1* and *NAT2* genes that lie close together on chromosome 8p22. The NAT1 enzyme is relatively invariant, but all human populations show frequent polymorphism for NAT2 variants with different enzymatic activity. Rapid acetylation, the wild type, is dominant over slow acetylation. Slow acetylators eliminate drugs and other xenobiotics more slowly, and so show enhanced sensitivity to their effects. Variable acetylation of the antitubercular drug isoniazid was one of the earliest observations in pharmacogenetics. Many years ago, it was noticed that individuals had greatly variable plasma concentrations of the drug after receiving a standard dose, and this had important clinical consequences. Slow acetylators are at increased risk of developing peripheral neuropathy, a known adverse effect of the drug. Other drugs for which variations in acetylation rate can be clinically important include procainamide (an antiarrhythmic), hydralazine (an antihypertensive), dapsone (antileprosy), and several sulfa drugs.

Glutathione S-transferases (GSTs) are a large family of enzymes that are involved in the detoxification of a variety of xenobiotics and carcinogens (**Figure 20.14**). There are six main classes—alpha (a), kappa (k), mu (m), pi (p), sigma (s), and theta (t), each encoded by several closely related genes. Different enzymes differ in their tissue and substrate specificities. The genes encoding the GSTM1 and GSTT1 enzymes have been the most extensively investigated. At each of these loci, gene deletions are common (about 50% and 15%, respectively, in white northern Europeans), probably as a result of unequal crossover in tandemly repeated gene clusters (see **Figure 15.17**).

People can be classified into nonconjugators, low conjugators, and high conjugators with respect to any particular GST activity. Numerous studies have reported associations of low conjugation with susceptibility to genotoxic effects. People with low GST activity may be unable to cope with high doses of drugs whose phase 2 detoxification involves conjugation with glutathione. In some cases, the effect has been in the opposite direction—for example, GSTT1-dependent conjugation of the industrial chemical trichloroethylene produces compounds with increased toxicity.

Many drugs are excreted in the form of glucuronide conjugates. These are formed by the action of UDP glucuronosyltransferases. The *UGT1A* locus on chromosome 2q37 has a remarkable structure, with 13 alternate first exons that are spliced onto invariable exons 2–5. Exon 1 determines the substrate specificity of the enzyme, while exons 2–5 encode the active site. One of the variants, UGT1A1, is responsible for catabolism of both bilirubin, the normal breakdown product of heme from red blood cells, and the anticancer drug irinotecan. Many variants of UGT1A1 with decreased enzymatic activity have been described, primarily in connection with hyperbilirubinemias, but patients with low UGT1A1 activity also suffer severe side-effects when treated with irinotecan.

Thiopurine S-methyltransferase (TPMT) transfers a methyl group from *S*-adenosylmethionine onto the immunosuppressant drugs azathioprine and 6-mercaptopurine, leading to their inactivation. About 10% of Europeans are heterozygous, and 0.3% homozygous, for low-activity variants of TPMT. These individuals require a lower dose of the drugs. Homozygotes can suffer life-threatening bone marrow toxicity when given a standard dose of either drug. Three relatively common variants account for about 90% of the low-activity alleles.

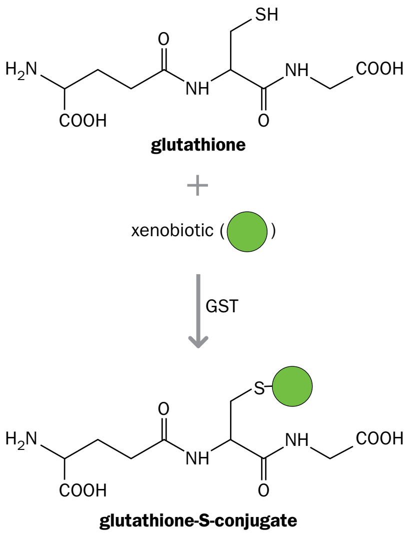


Figure 20.14 Glutathione conjugation. Phase 2 metabolism of many drugs and other xenobiotics involves conjugation with glutathione. The reaction is catalyzed by glutathione S-transferase (GST).

Genetic variation in its target can influence the pharmacodynamics of a drug

The effects considered above affect the pharmacokinetics of a drug: how fast it is activated, inactivated, and excreted. Another way in which genetic differences can affect drug responses is through pharmacodynamics—that is, the specific response of a drug target to a given drug. Drug targets include receptors, enzymes, and signal transduction systems. Genetic variants in the target can affect the efficacy of a drug.

Many of the most important applications of pharmacodynamics are in the anticancer drugs that are designed to be effective against specific mutant versions of a cell-surface receptor or other critical molecule. These were described in Section 19.4 and will not be further covered here. A few naturally-occurring variants in other drug targets are associated with significantly different clinical responses.

The angiotensin-converting enzyme (ACE) is a peptidase that converts angiotensin I into angiotensin II. The latter is an important regulator of blood pressure and has many other physiological functions. A polymorphic insertion/deletion of an Alu sequence in an intron of the *ACE* gene is associated with variations in enzyme activity (**Figure 20.15**). DD (deletion) homozygotes have about twice the level of circulating ACE as II (insertion) homozygotes. ACE inhibitors such as enalapril and captopril are widely used drugs for treating heart failure. Several reports suggest that ACE inhibitors are more effective in DD than II patients. The insertion–deletion polymorphism has also been extensively

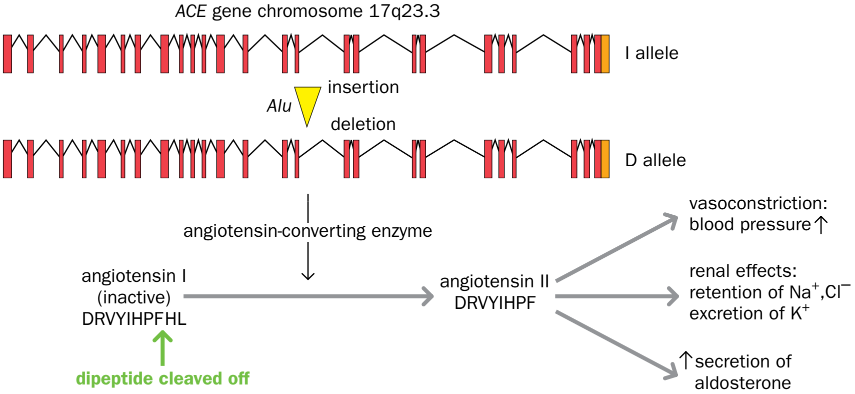


Figure 20.15 An insertion–deletion polymorphism in the *ACE* gene that encodes the angiotensin-converting enzyme. Insertion (I) alleles have an insertion of an *Alu* element into an intron of the gene. Homozygotes for the deletion allele (D) have higher levels of circulating enzyme than those with the I allele. Peptide sequences are shown with the single-letter code.

studied for association with diseases. DD homozygotes are at increased risk of myocardial infarction and coronary artery disease, and possibly also for complications of type 2 diabetes, but at slightly decreased risk of late-onset Alzheimer disease.

The β_2 adrenergic receptor has two common variants, p.Arg16Gly and p.Gln27Glu, that have frequencies of 0.4–0.6 in many populations and are in strong linkage disequilibrium with each other. β_2 agonists are the most widely used drugs for treating asthma. Individuals homozygous and heterozygous for the Arg16 variant are 5.3-fold and 2.3-fold, respectively, more likely to respond to a single dose of the antiasthmatic drug albuterol than those who are homozygous for Gly16; however, they are also at increased risk of long-term deterioration with chronic drug use.

The *ADRB1* gene located on chromosome 10q24–26 encodes the β_1 receptor. This has only limited homology to the β_2 receptor and is targeted by different drugs. A common polymorphism, p.Arg389Gly, has been tested for pharmacodynamic effects. The Gly389 allele is associated with a decreased cardiovascular response to beta-blocker drugs. For example, Arg389 homozygotes had a much better response to the beta-blocker bucindolol than those who are heterozygous or homozygous for the Gly389 allele.

Warfarin is a test-bed for using pharmacogenetic data in personalized medicine

Warfarin is a powerful anticoagulant, used for patients at risk of embolism or thrombosis (and also to poison rats and mice by triggering internal bleeding). It works by decreasing the availability of vitamin K. This is an essential cofactor for the enzyme γ-glutamyl carboxylase that activates blood clotting factors II (prothrombin), VII, IX, and X. During these reactions, vitamin K is converted to the inactive vitamin K epoxide, and this is recycled by the action of vitamin K epoxide reductase (VKOR). Warfarin inhibits VKOR (**Figure 20.16**).

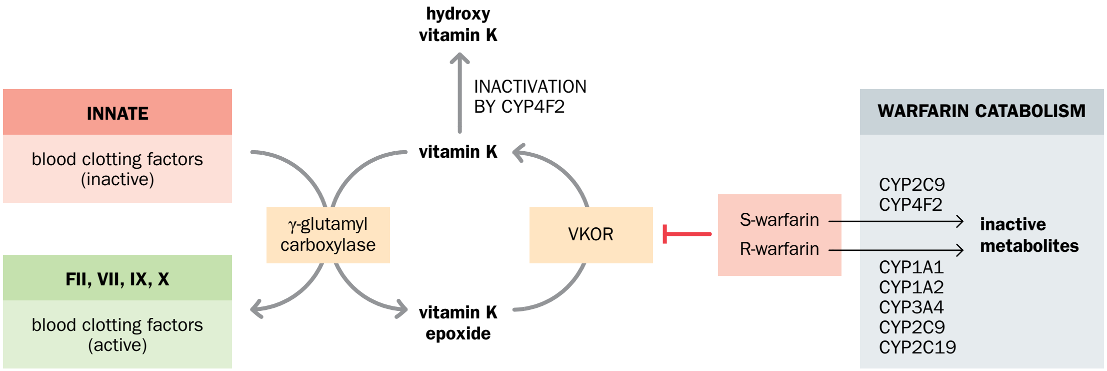


Figure 20.16 Action and metabolism of vitamin K and warfarin. See text for details.

Achieving the correct level of anticoagulation is clinically very important. If the level is too low, the patient remains at risk of thrombosis or embolism; if it is too high, there is a risk of hemorrhage, which can be life-threatening. The therapeutic window (the range of doses that are efficacious and not harmful) is narrow, but the effective warfarin dose varies up to 20-fold between individuals. After insulin, warfarin is the most common prescription drug responsible for emergency hospital admissions. The average cost of a bleeding episode has been estimated as USD $16,000. This has made warfarin something of a test case for the general utility of pharmacogenetically-informed prescribing.

Warfarin as normally used is a mixture of two stereoisomers, *R*-warfarin and *S*-warfarin. Both isomers are active, but the *S*-form is 3–5 times more potent than the *R*-form. CYP2C9 is the principal enzyme that catalyzes the conversion of *S*-warfarin to inactive 6-hydroxy and 7-hydroxy metabolites, whereas the oxidative metabolism of *R*-warfarin is catalyzed mainly by CYP1A2 and CYP3A4. However, as shown in **Figure 20.16**, several other P450 enzymes have a role in removing warfarin or otherwise inactivating vitamin K. Variants in *CYP2C9* and *VKORC1* (the gene encoding subunit 1 of VKOR) can explain about 30–40% of the variation of response to warfarin. In 2007, the USA Food and Drug Administration (FDA) put wording on the label of warfarin that recommended (but did not mandate) genotyping these two loci. Trials of prescribing following dosage algorithms based on *CYP2C9* and *VKORC1* genotypes, maybe also including *CYP4F2*, have demonstrated advantages over the traditional trial and error procedure, especially for patients who need particularly large or small doses. Many clinicians, however, remain skeptical about the value of genotyping their patients because the algorithms do not fully predict the response. A major goal in pharmacogenetics is to identify the remaining factors and develop a simple test that will accurately predict the correct dose of warfarin for a patient.

Pharmacogenetics and pharmacogenomics have tended to be of more academic than practical interest, but this may be changing

Some of the effects described above have been known for decades, yet they have had little influence on prescribing. They explained only some of the many factors affecting how well a drug would deal with a patient's problem, and by and large physicians proceeded by trial and error. There was also a major logistic issue. Both the patient and the physician want a consultation to end with the patient clutching a prescription. They do not want it to involve taking a blood sample and coming back for the prescription a week later when the laboratory has reported the relevant genotype.

The targeted cancer drugs described in Section 19.4 may be marketed together with a companion diagnostic test to identify appropriate patients. In the cancer case this is driven by the extremely high cost of the drugs. With less expensive drugs there is less incentive to test where the risk is just that the drug will be ineffective, but there is an incentive where the need is to detect people with genotypes that put them at risk of a severe adverse reaction. Thus, genotyping before prescribing is widely accepted for drugs such as abacavir (see **Table 20.8**), which is a treatment of choice for most HIV patients, but where people with certain specific genotypes may suffer severe adverse effects. It is less widely accepted where the aim is just to arrive at the most effective dose for the individual patient, and where genotypes are only one factor in achieving this, as with warfarin.

In developing novel drugs, companies now try to avoid compounds that are metabolized by highly-variable enzymes like the P450 cytochromes. The drugs that are subject to these variable effects are often long-established and are now off-patent. Companies have little financial incentive to develop genotyping assays for drugs where they do not hold a patent. In the USA the FDA maintains a list of approved drugs with pharmacogenomic information in their labelling (see http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm). In August 2018 this had some 280 entries, but most of these were advisory rather than mandatory. The pharmgkb database (https://www.pharmgkb.org) is an invaluable source of specific information.

Probably the main current limitation for pharmacogenetics in mainstream medicine is the logistic problem alluded to above. What is needed is a handheld device like a smartphone that can be used at the bedside to take a drop of blood or saliva and report the relevant genotype within a few minutes. Some such devices are already on the market for limited applications, and companies are working hard to develop cheap and versatile successors. Alternatively, we may arrive at the point where most people's full genome sequence is part of their standard medical record, and in that case the relevant information will be available right at the start of a consultation. Whether we then move forward into the long-promised era of personalized medicine where genotypes determine prescribing depends on how predictive genotypes are for a given clinical situation. It is unlikely that there will be a one-size-fits-all answer to that question.

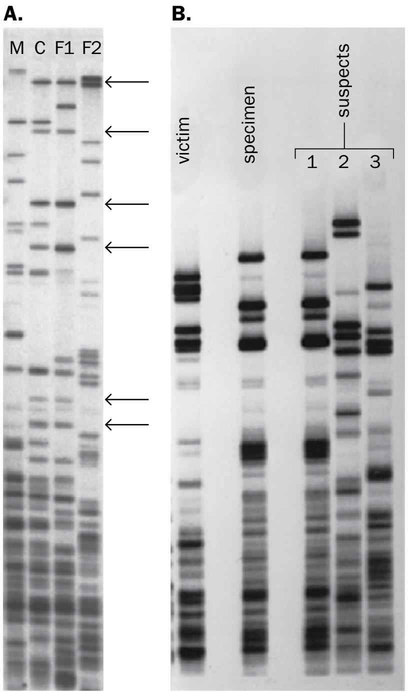


Figure 20.17 Using DNA fingerprinting. (**A**) Resolving disputed paternity. DNA fingerprints were obtained from a mother (M), her child (C), and two possible fathers (F1 and F2). Arrows show bands present in the child but not in the mother. They could have come from F1 but not F2, thus ruling out F2 as the father. (**B**) A criminal investigation. A DNA fingerprint from a vaginal swab taken from a rape victim is shown, together with DNA fingerprints from three suspects. Because the swab may contain DNA of the victim (even after a procedure to selectively isolate sperm heads), her own fingerprint is also shown. The fingerprint of suspect 1 matches the specimen. (Images courtesy of Cellmark Diagnostics, Abingdon, Oxfordshire, UK.)

20.6 DNA FORENSICS: IDENTIFYING INDIVIDUALS AND RELATIONSHIPS

The use of DNA variants for identifying individuals started with the pioneering work of the British scientist Alec Jeffreys. In 1984 Jeffreys had identified a 33 bp tandem repeated sequence from intron 1 of the human myoglobin gene. When he used this sequence as a probe on Southern blots of restriction-digested human DNA, several different hybridizing bands were detected. Based on this he developed a series of probes containing related but shorter tandem repeats. These probes hybridized to a number of different hypervariable minisatellites spread across the genome, producing individual-specific 'fingerprints' on Southern blots of *Hinf*1-digested human DNA. Individual bands in the fingerprints were reproducible and behaved as single alleles in family studies. Two of the new probes, 33–6 and 33–15, formed the basis of forensic DNA fingerprinting (**Figure 20.17**). The chance that two random unrelated Europeans would have identical fingerprints was calculated as 3×10^{-11} when just probe 33–15 was used, or 5×10^{-19} with both probes (see Jeffreys *et al*. [1985a] PMID 2989708, see Further Reading).

Limitations of the original DNA fingerprinting technique

DNA fingerprinting revolutionized forensic science and rapidly showed its value, both in criminal investigations and family disputes. Nevertheless, it had some major limitations.

- Being based on Southern blotting, it required significant amounts of undegraded starting material: 0.5–5 μg of DNA, 60 μl of blood, or 5 μl of semen. The procedure is also laborious, time-consuming and requires considerable skill in the laboratory.
- Importantly, the fingerprint could not be analyzed genetically. Bands could not be assigned to specific loci or paired up as alleles; the fingerprint just had to be studied as an image. Thus, when comparing two DNA fingerprints, the investigator matched each band individually by position and intensity. The continuously variable distance along the gel had to be divided into a number of 'bins'. Bands with similar intensity and falling within the same bin were deemed to match. Then if, say, 10 out of 10 bands matched, the odds that the suspect, rather than a random person from the population, was the source of the sample are 1 in p^{10}, where p is the chance that a band in a random person would match a given band (we have simplified by assuming p to be the same for every bin). Even for $p = 0.2$, p^{10} is as low as 10^{-7}. It is imperative that the same binning criteria be used for judging matches between two profiles as for calculating p. The criteria can be arbitrary within certain limits, but they must be consistent.
- The same limitation meant that there was no obvious way of constructing large-scale searchable databases of fingerprints—fingerprints just had to be compared side by side. This limitation turned out to be important in the first major application of DNA fingerprinting to a criminal case. In 1983 and 1986 two 15-year-old women had been raped and murdered near the same village in Eastern England. The circumstances of the cases led police to conclude that the same man was responsible for both crimes. The village in question was near to Leicester, where Dr Jeffreys worked. When all other lines of enquiry had proved fruitless, the police turned to his then novel DNA technique. All adult men in three local villages were asked to volunteer blood samples so that they could be eliminated from the enquiry. After over 4000 men had been tested, no match had been found to semen recovered from the victims (a local man with learning difficulties had previously confessed to the crimes, but the DNA fingerprints proved that this was a false confession). Then a woman reported having heard a man boasting in a local pub that he had provided a blood sample while falsely claiming to be his friend Colin Pitchfork, for which Pitchfork had paid him £200. When DNA was taken from the real Colin Pitchfork it matched the crime scene fingerprint. Pitchfork was found guilty of the crimes and sentenced to life imprisonment. A modern searchable database would no doubt have flagged up that the man in the pub had provided two profiles under different names but noticing two identical fingerprints among 4000 would not have been possible.

Scene-of-crime samples often contain DNA from more than one individual, and this again makes crime scene DNA fingerprints hard to match to the fingerprint of a suspect. For all these reasons, some forensic laboratories switched to using single-locus probes to detect restriction fragment length polymorphisms. Several probes would be used sequentially, with the Southern blot filter being stripped and rehybridized, or if the amount of DNA allowed, fresh blots prepared. All these procedures were quickly abandoned when it became possible to type PCR-amplified DNA for short tandem repeat markers (STRs or microsatellites).

DNA profiling uses PCR-amplified short tandem repeats

Alleles of STR markers can be defined unambiguously by the precise repeat number, which avoids the binning problem and allows easy reporting and recording of a DNA profile. If the frequency of each allele in the relevant population is known, an exact calculation can be made of the odds that a suspect, rather than an unrelated member of that population, is the source of a DNA sample. Many different microsatellites might be used, but the development of national DNA databases makes it necessary to fix on a standard set. Apart from being highly polymorphic, suitable markers should be spaced out across the genome and amplify reliably—as far as possible they should not risk failing because of SNPs under primer binding sites. Primers are chosen to give products of nonoverlapping sizes from different markers so that they can be multiplexed. If necessary, stuffer sequences can be incorporated into the 5′ ends of primers to give suitable-sized products. Tetranucleotide repeats are preferred to dinucleotides, because they give fewer stutter bands and the different alleles are readily separated. It is usual also to include amelogenin as a sex marker. The X and Y chromosomes each have a copy of the amelogenin gene (*AMELX*, *AMELY*), but the copies differ so that each gives a different-sized PCR product.

Table 20.9 shows the markers used in the main USA and UK forensic databases. The UK DNA17 set is the result of European-level agreements and is designed to allow easy data exchange with the databases of other European countries. **Figure 20.18** shows a typical DNA17 profile.

Over the years national DNA databases have grown very large. In the USA in June 2018 the National DNA Index (NDIS) contained 13,413,029 offender profiles, 3,174,013 arrestee

TABLE 20.9 EXAMPLES OF STR MARKER SETS USED FOR FORENSIC DNA PROFILING

Marker	C'me	CODIS (USA)	UK first panel	SGM (UK)	SGM+ (UK)	DNA17 UK)
D1S1656	1					+
D2S441	2					+
D2S1338	2				+	+
TPOX	2	+				
D3S1358	3	+			+	+
FGA	4	+		+	+	+
CSF1PO	5	+				
D5S818	5	+				
F13A1	6		+			
SE33	6					+
D7S820	7	+				
D8S1179	8	+		+	+	+
D10S1248	10					+
THO1	11	+	+	+	+	+
D12S391	12					+
VWA	12	+	+	+	+	+
D13S317	13	+				
FES/FPS	15		+			
D16S539	16	+			+	+
D18S51	18	+		+	+	+
D19S433	19				+	+
D21S11	21	+		+	+	+
D22S1045	22					+
Amelogenin	X, Y	+		+	+	+
Average match probability		1:10^{13}	1:10,000	1:5 × 10^7	1:10^9	1:10^9

The current (2018) sets in the USA and UK are CODIS and DNA17 respectively; the other sets are shown to illustrate the development of the UK marker panel. Note the way UK panels after SGM have been designed to allow backward compatibility. By a political decision the match probability for the UK SGM+ and DNA17 sets is quoted in court as 1:10^9, although in reality it is much lower. In special cases an exact probability can be calculated. Details of all markers can be found in the STR factsheets at http://www.cstl.nist.gov/biotech/strbase/str_fact.htm.

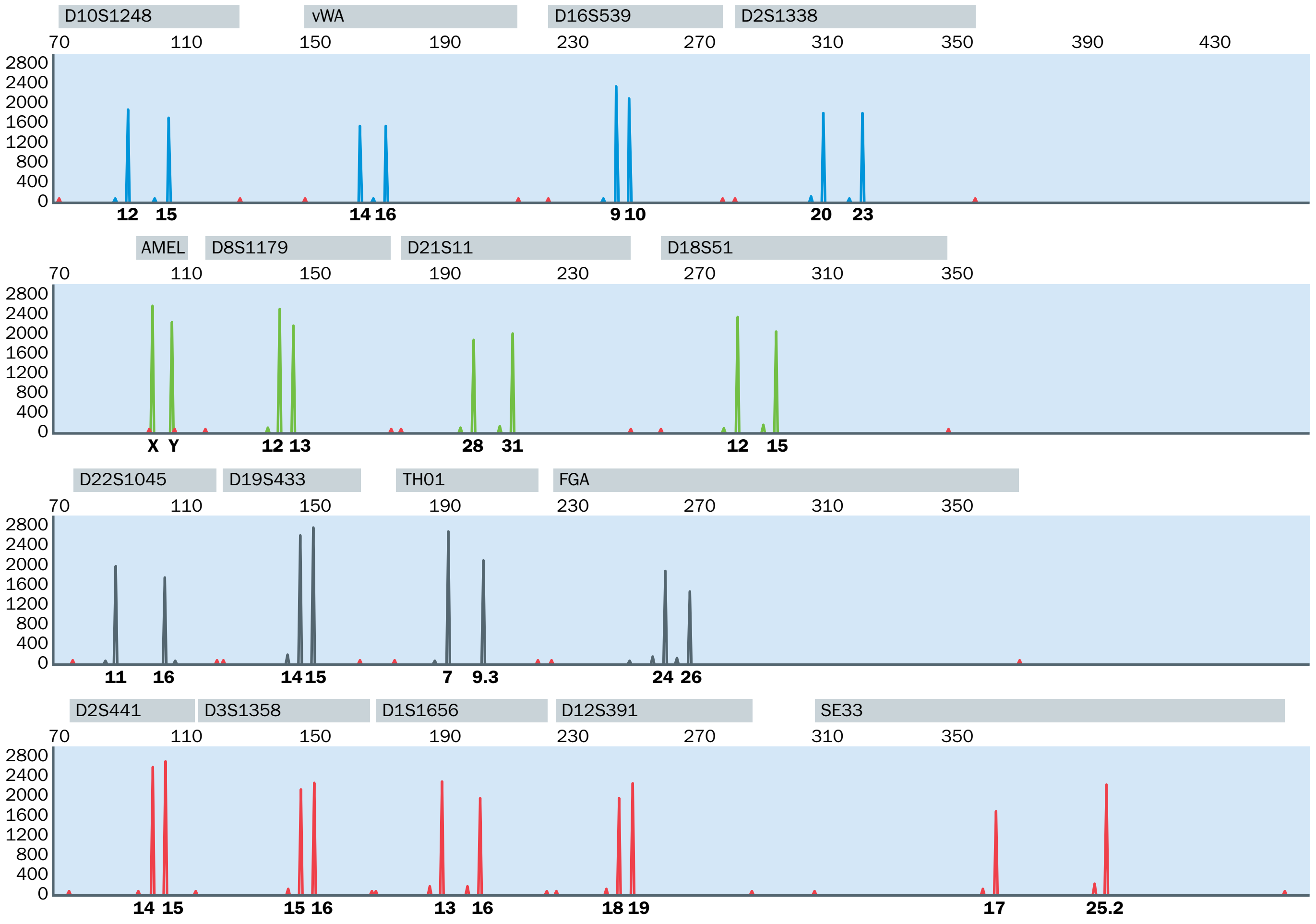


Figure 20.18 **A DNA profile obtained with the DNA17 marker set.** The green bars show the range of expected allele sizes for each marker. Note the small stutter bands just ahead of most peaks. Profile obtained using the Applied Biosystems AmpFℓSTR ® NGM SElect™ kit.

profiles, and 864,128 forensic profiles. At 30 June 2018 the UK National DNA database NDNAD held 5,405,095 DNA profiles from individuals and 599,834 from crime scenes. 62% of crime scene samples booked into the NDNAD in 2014–5 matched an offender profile. Note, however, that 15,108 out of the 34,201 crime scene profiles loaded in that year related to burglary, compared to only 720 for rape and 543 for homicide—in other words the high hit rate must have owed something to a limited number of prolific burglars.

Using Y-chromosome markers

Y chromosome haplotypes can be defined using a mix of STR and SNP markers (see Table 14.3). Because, barring rare mutations, a man's Y-haplotype is shared with all his male-line relatives including very distant ones, a Y-haplotype match has little power to incriminate a suspect, but a mismatch is powerful exclusionary evidence. Y-STR results may be obtained from samples containing an overwhelming preponderance of female DNA, for example after a sexual assault where the standard sperm separation protocol failed. After a gang rape Y-typing may identify individual perpetrators among a pool of suspects when autosomal markers give a hopelessly confused pattern. Outside criminal investigations, Y markers can identify distant male-line relationships. An interesting example concerns the story that USA president Thomas Jefferson may have fathered one or more children by his slave Sally Hemings (Figure 20.19A). Distant male-line descendants of both Jefferson and a son of Hemings were tracked down and demonstrated to carry identical Y chromosome haplotypes. This proved beyond reasonable doubt that a Jefferson male fathered children with Hemings; it did not prove that the male in question was President Thomas Jefferson. Records showing which Jefferson males were present at the Monticello estate at the times the children were conceived do make it highly likely that it was indeed President Thomas Jefferson, but do not prove it beyond all reasonable doubt (see https://www.monticello.org/site/plantation-and-slavery/jefferson-hemings-resources).

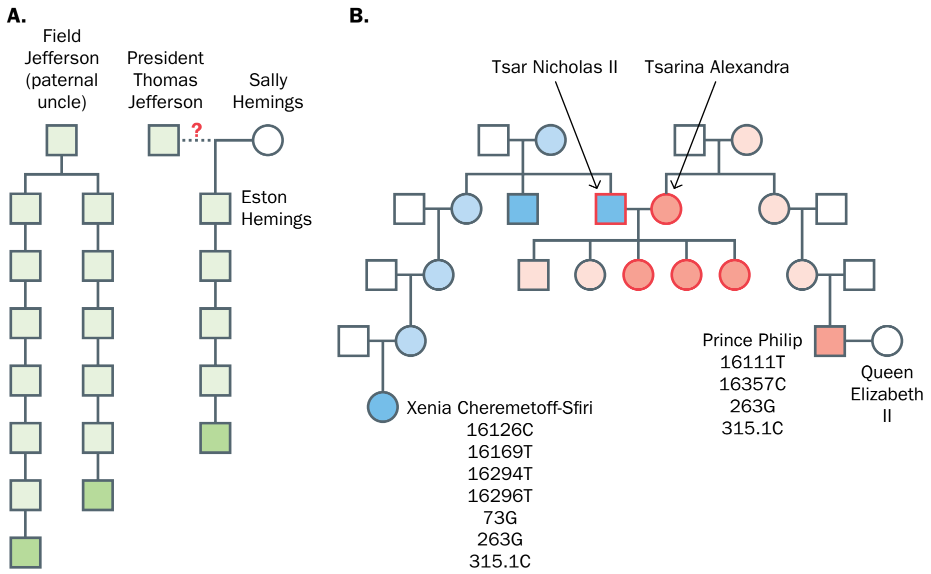


Figure 20.19 The use of Y-chromosome and mitochondrial DNA variants to follow distant family relationships. (**A**) Following USA President Thomas Jefferson's Y chromosome in Eston, born in 1808 to his slave Sally Hemings. The Y haplotype was defined by 11 STRs, 7 SNPs, and one microsatellite and was found in all three samples marked in dark green. (**B**) Using mitochondrial DNA variants to identify the remains of the family of Russian Tsar Nicholas II, murdered by the Bolsheviks in 1917. Symbols outlined in red show the skeletal remains unearthed in 1991. In each part strong colors show experimentally observed types, weak colors show inferred types. See text for details.

Using mitochondrial DNA variants

A mother's mt-DNA is transmitted intact without recombination to all her children and so, like the Y-haplotypes, can be used to follow distant relationships. Common variation is mainly found in the HV1 and HV2 hypervariable regions of the D-loop (see **Figure 9.1**). Variants are usually typed by PCR-amplifying and then sequencing the HV1 and HV2 regions (typically as 342 bp and 268 bp amplicons, respectively). Heteroplasmy (see Section 16.4) is not uncommon and can complicate the interpretation. Since cells contain many mitochondria, mt-DNA sequences are present at much higher copy number than nuclear sequences. Thus mt-DNA typing can be useful for samples that are too small or too degraded to allow reliable typing of nuclear markers. A classic application of mitochondrial DNA analysis was the identification of the remains of the family of Russian Tsar Nicholas II, murdered by the Bolsheviks in 1917 (**Figure 20.19B**).

Nine skeletons unearthed near Ekaterinburg in 1991 were thought to include the remains of the Tsar, his wife, and three of their children. Five autosomal STRs (the four of the UK First Panel plus SE33) were used to show that five of the nine skeletons were from a family group; the other four were probably three servants and the family doctor. Living persons distantly related to the Tsar and to the Tsarina entirely through the female line were identified and provided reference mt-DNA samples (**Figure 20.19B**). Prince Philip, the husband of the British Queen, was the great-grandson through the female line of the Tsarina's mother; his mt-DNA precisely matched the mt-DNA of the presumed Tsarina and the children. On the Tsar's side a living great-granddaughter of the Tsar's sister through the female line had mt-DNA that matched the presumed Tsar except at position 16169, where the presumed Tsar's mt-DNA was heteroplasmic T/C while the living relative's DNA was homoplasmic T. This remaining uncertainty was resolved by exhuming the body of the Tsar's brother and demonstrating that his mt-DNA was identical to the presumed Tsar's mt-DNA, including the same heteroplasmic variant.

A DNA photofit?

Since monozygotic twins are usually extremely similar in appearance, it must follow that our facial features are largely encoded in our DNA. If the encoding were fully understood, it might be possible by analyzing a crime scene sample to produce a photofit portrait of the person who left it. We are currently a very long way from being able to do this, but limited deductions about hair and eye color can be made. This is an area of active research by companies that can see a large potential market.

Difficult samples

Unlike most samples referred to genetic testing laboratories, crime scene samples are often degraded, mixed, or very small. Each of these features poses problems for the laboratory.

- **Degraded DNA** is often of low molecular weight. Large amplicons like SE33 (310–450 bp in the system shown in **Figure 20.18**) will often fail to amplify from samples where smaller amplicons still work. So-called mini-STRs give a higher success rate with degraded DNA. Thus, makers of the standard kits have repositioned primers so as to reduce the size of amplicons as far as possible. Samples may also contain inhibitors of the PCR reaction—for example heme or the indigo dye used to dye jeans.

- **Mixed samples** give confusing profiles. If the different individuals represented in the sample contributed unequal amounts of DNA it may be possible to use the peak heights, or peak areas, to separate out the individual profiles. Various software solutions are available to help with this, but it is a difficult area and much may depend on expert judgement (and therefore be challengeable in court).
- Forensic laboratories are under constant pressure to produce results from **smaller and smaller samples**. Samples containing 100–200 pg of DNA (the DNA of 16–30 cells) can usually give a full profile, but below this, random (stochastic) events become significant. Profiles show **allele dropout**, where alleles present in the donor fail to amplify. Allele drop-in is also observed, as contamination is an increasingly serious problem with very small samples. Again, interpreting the results from very small samples calls on the expert judgement of the scientists, and so is less certain as evidence in court.

Contamination is an ever-present problem. Elimination databases contain the profiles of all the people—police and laboratory personnel—whose DNA might accidentally contaminate a scene-of-crime sample. Sometimes the source of contamination is unexpected. It is said that Austrian police were hunting a master criminal whose DNA had been found at a number of different crime scenes—until it turned out that the DNA in question came from an individual who worked in the firm that supplied the laboratory with its plastic reaction tubes.

Courtroom issues

The days are long past when lawyers would contest the science behind DNA profiling. But when entering a courtroom, scientists enter a very alien environment. Scientists accept experimental results if they judge the procedure reliable and the controls adequate. In a courtroom they may face insistent allegations that samples have been deliberately interfered with or that expert witnesses are manipulating the facts. Never mind the likelihood, the aim is to sow doubt in the minds of the jury. In the notorious 1994 OJ Simpson trial, for example, the defense lawyers did not dispute that various bloodstains contained Simpson's DNA, but they variously alleged that the police had deliberately conspired to plant Simpson's DNA or that the police or the laboratory had handled samples carelessly and contaminated them with their reference sample of Simpson's blood, collected the day after the murders. See Weir (1995, PMID 7493014) in Further Reading for a commentary by a leading forensic scientist.

Even if all parties accept that a scene-of-crime sample does indeed contain the suspect's DNA, that does not automatically make the suspect guilty. There can be disputes about how the DNA got there—was it deliberately planted by a malicious person in order to incriminate an innocent suspect? Might it be the result of passive transfer—from a person or object that the suspect touched, without him ever having been present at the crime scene? Even if the suspect had indeed been present, the mere presence of his DNA does not prove when he was there or what he had been doing (with the exception, maybe, of DNA in a vaginal swab from a rape victim). A DNA match does not in itself prove guilt; it is just one piece of evidence—albeit a very powerful one—in the overall case.

In the so-called 'prosecutor's fallacy' an unscrupulous advocate tries to mislead the jury into asking the wrong question about the significance of a DNA match. Suppose the suspect's DNA does match the crime scene DNA. The jury needs to consider the likelihood that he is innocent despite the match. In the prosecutor's fallacy they are instead led to ask what is the probability of a match, given that he is innocent. To see the difference, suppose that the chance a random person would have a matching profile is 1 in a million. If there is no other evidence, the suspect might be any one of 10 million people in a city. Then there would actually be 10 individuals in the city whose profiles matched, and our suspect is only one of those ten. Thus, while the probability of a match, given that he is innocent, is only one in a million, the probability he is innocent, given the match, is 9 out of 10. Fortunately, a full CODIS or DNA17 match gives likelihoods so extreme that the prosecutor's fallacy could not overturn them—but the issue might become more relevant if the crime scene profile is very partial.

Ethical issues

The many ethical issues surrounding DNA forensics focus on three areas: who can be required to provide a sample, how long samples or profiles should be retained, and what questions can be asked about somebody's DNA. As with all law enforcement matters, there are also concerns that national DNA databases contain disproportionate numbers of profiles of individuals from ethnic minorities.

Countries differ in the circumstances where compulsory DNA testing is allowed. In some jurisdictions, only somebody accused of a serious crime can be compelled to provide a sample. In the UK, on the other hand, samples can be taken without consent from anybody arrested in connection with a 'recordable' offence—a category that includes all but the most minor offences. This is regardless of whether or not the person is eventually charged with any offence, still less whether or not they are convicted. At the extreme end, in 2016 the government of Kuwait proposed to make DNA profiling compulsory for all residents and all visitors, even tourists or businessmen visiting the country for just a few days. Refusal was to be punished with a heavy fine or imprisonment. The stated purpose was to help combat terrorism. There were concerns that the real purpose might be something different (for example, adultery is illegal, and citizenship, with its attendant access to the generous welfare system, depends partly on tribal origins), and in any case, whatever the intentions of the present government, a future government or hackers could use the resulting database for sinister purposes. In the light of arguments raised by a local law firm and by the European Society of Human Genetics, the government rowed back on these proposals, and suggested limiting sampling to accused or convicted criminals.

A further set of issues concerns the retention of samples or profiles. If DNA samples are taken when a person is arrested, what happens if they are never actually charged with any offence (as often happens), or charged but found not guilty at a subsequent trial? Will their profile remain on the database, available for all future speculative searches against crime scene samples? Will their DNA sample remain in a freezer, available for genotyping in currently unforeseen ways? Police forces have been extremely reluctant to destroy samples or delete profiles from the database. In the UK, following a 2008 judgement by the European Court of Human Rights, the 2012 Protection of Freedoms Act set detailed rules for the retention of samples and profiles. The actual DNA sample must normally be destroyed once a profile has been obtained, while profiles can only be retained indefinitely from people convicted of an offence. There are special provisions for profiles of young persons. These limitations apply to reference samples taken from suspects, not to scene-of-crime samples or profiles.

A final set of questions concerns the types of searches that may be made on a profile or other DNA data. For example, are police allowed to try to infer the ethnicity of the person who left DNA at a crime scene? Particular controversy surrounds familial searching. Criminality often runs in families. If no match to a crime scene sample is obtained in the database, might it contain the profile of the father or brother of the person who left the sample? Since most criminals are male, Y chromosome markers are particularly helpful. Familial searching inevitably produces a long list of possible fits, which must then be whittled down by considering age, location, and normal police questions. This is a lot of work, so familial searching is only used in the most serious cases. Nevertheless, people worry that it may improperly reveal family secrets. In the USA, familial searching is allowed by some states but prohibited by others. **Figure 20.20** looks at these concerns.

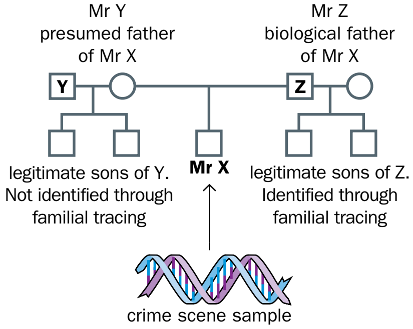


Figure 20.20 Familial searching in a family with secrets. Mr. X has left his DNA at a crime scene. The profile from this DNA does not match any in the database, but the database does contain profiles of Mr. X's presumed father, Y and his actual biological father Z. What family secrets might be revealed by familial searching? Y's DNA is unrelated to X's and therefore familial searching would not identify Y. It could possibly identify Y's sons if their profiles were on the database, although as half-brothers the match would be weaker with many more false positives. Any such match would not involve any family secrets. Z's DNA gives a possible partial match and the police therefore request samples from Z's sons. If Z and X acknowledge they are father and son, no family secrets are involved. If Z either does not mention X, or maybe is unaware that X is his son, the police will not be given X's name. If Z acknowledges X as his son, but X is unaware of this, then the police will need to invent a story to explain how they have located X, otherwise the secret would be revealed.

Concerns about the potential for DNA submitted for one purpose—often recreational—to be used for a quite different purpose were intensified by the use in 2018 of familial searching to identify the suspected "Golden State Killer" (see https://www.bbc.co.uk/news/world-us-canada-43916830). This man was thought to have committed multiple rapes, murders, and burglaries across California in the 1970s and 1980s, but DNA samples linking the crimes had no match in the CODIS database. The DNA evidence was uploaded to the GEDmatch database, which supports recreational DNA and genealogical analysis by amateur and professional researchers and genealogists. This identified several possible relatives of the killer. After narrowing the list by orthodox police enquiries, the search eventually led to a former police officer Joseph James DeAngelo, who is currently (2018) awaiting trial.

Paternity and relationship testing

DNA profiling provides a far superior method of establishing paternity or other relationships, compared to previous methods based on blood groups and so on. The well-validated forensic marker kits could be used, or any other suitable DNA markers. Y-chromosome and mitochondrial markers have many applications. Paternity testing can readily exclude an alleged father but can never absolutely prove that a man is the father of a child. The aim is to establish a paternity index, the relative likelihood that the suspect rather than a random man from the same population is the father. This will be based on the frequencies of the paternal alleles in the relevant population (**Figure 20.21**). For more distant relationships the calculations can become quite elaborate. The first application of Jeffreys's fingerprinting technique was in a difficult immigration case, and while the particular technique is now obsolete, the calculations are still interesting (see Jeffreys *et al.* [1985b] PMID 4058586).

Unlike in forensic matching, in paternity testing the possibility of mutation must be taken into account. STR markers have much higher mutation rates than SNPs, and with very highly polymorphic markers such as *SE33* or *D18S51*, mutations are quite frequently observed. Specific mutation rates for each allele of each marker are documented and can be incorporated into calculations (longer alleles have higher mutation rates than shorter alleles). The general rule is that a single mismatch, particularly if the mismatch is by only one repeat unit, does not exclude paternity.

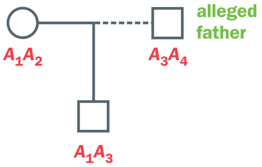


Figure 20.21 Using single-locus markers for a paternity test. The odds that the alleged father, rather than a random member of the population, is the true father are ½ : q_3, where q_3 is the allele frequency of A_3. A series of *n* unlinked markers would be used and, if paternity were not excluded, the odds would be (½)^n: q_a x q_b x q_c x ... x q_n.

SUMMARY

- Genetic testing procedures have evolved rapidly in the past few years, with the widespread adoption of next-generation sequencing by diagnostic laboratories.
- The choice of testing method depends on the precise question being asked.
- Often the question is whether a test sample contains one or more specific sequence variants. Various cheap and rapid methods are available to answer that question; SNP chips allow testing for a large panel of specified single-nucleotide variants.
- At other times the question is whether the patient has a pathogenic variant anywhere in one gene or a set of candidate genes. This is addressed by sequencing all relevant exons. Next-generation sequencing of large panels of candidate genes has made many highly heterogeneous conditions amenable to genetic testing. A number of methods formerly used to scan a gene to prioritize exons for sequencing have lost popularity as sequencing capacity has risen and costs have fallen.
- When there is no candidate gene, whole exome sequencing is used to seek a pathogenic change that might be in any gene. The main problem with this approach is the bioinformatic effort needed to identify the sought-after pathogenic variant among the very large number of variants detected.
- Whole genome sequencing avoids problems associated with exon capture but is currently too expensive to be the default method except for testing tumors, where it is crucial to identify structural variants.
- Deletions and duplications of one or a few exons may be identified by MLPA (multiplex ligand-dependent probe amplification); larger ones are usually defined by array-CGH (array-comparative genomic hybridization). Structural variants could also be identified by whole genome sequencing.
- For chromosomal abnormalities, traditional karyotyping has been largely supplanted by array-based or sequencing-based approaches but is still valuable for balanced abnormalities.
- Effects on splicing can often be predicted computationally, but definitive confirmation requires mRNA to be sequenced (as cDNA).
- Variants that affect the metabolism (pharmacokinetics) or action (pharmacodynamics) of drugs are important causes of adverse drug reactions and the limited efficacy of some drugs. Some drugs are prescribed with a companion diagnostic. Fully personalized medicine remains a project for the future.
- In forensic testing, panels of microsatellites are used to link DNA recovered from a crime scene to a suspect, to identify individuals after disasters, and to verify family relationships. Y-chromosome and mitochondrial haplotypes are valuable for tracing distant family relationships.
- Forensic testing involves some extra issues compared to other testing, including problems with small, mixed, or degraded samples; the need to have a documented chain of custody of scene of crime samples and to consider the possibility of deliberate misconduct; and a number of special ethical issues.

FURTHER READING

Diagnostic testing

American College of Medical Genetics https://www.acmg.net

British Society for Genetic Medicine http://www.bsgm.org.uk/

Lek M, Karczewski KJ, Minikel EV *et al*. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**:285–291; PMID 27535533.

OECD guidelines on quality assurance in genetic testing http://www.oecd.org/sti/biotech/oecdguidelinesforqualityassuranceingenetictesting.htm

Richards S, Aziz N, Bale S *et al*. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**:405–424; PMID 25741868. https://www.acmg.net/docs/standards_guidelines_for_the_interpretation_of_sequence_variants.pdf

Schouten JP, McElgunn CJ, Waaijer R *et al*. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification *Nucl Acids Res* **30**:e57; PMID 12060695.

Tarailo-Graovac M, Shyr C, Ross CJ *et al*. (2016) Exome sequencing and the management of neurometabolic disorders. *New Engl J Med* **374**:2246–2255; PMID 27276562.

van El CG, Cornel MC, Borry P *et al*. (2013) Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics. *Eur J Hum Genet* **21**:580–584; PMID 23676617.

Population screening

Green RC, Roberts JS, Cupples LA *et al*. (2009) Disclosure of APOE genotype for risk of Alzheimer's disease. *N Engl J Med* **361**:245–254; PMID 19605829.

Green RC, Berg JS, Grody WW *et al*. (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* **15**:565–574; PMID 23788249.

Hivert M-F, Vassy JL, Meigs JB (2014) Susceptibility to type 2 diabetes mellitus—from genes to prevention. *Nat Rev Endocrinol* **10**:198–205; PMID 24535206.

Kalf RR, Mihaescu R, Kundu S *et al*. (2014) Variations in predicted risks in personal genome testing for common complex diseases. *Genet Med* **16**:85–91; PMID 23807614.

Malone FD, Canick JA, Ball RH *et al*. (2005) First-trimester or second-trimester screening, or both, for Down's syndrome. *New Engl J Med* **353**:2001–2011; PMID 16282175.

Shkedi-Rafid S, Dheensa S, Crawford G *et al*. (2014) Defining and managing incidental findings in genetic and genomic practice. *J Med Genet* **51**:715–723; PMID 25228303.

Pharmacogenetics

Pirmohamed M, James S, Meakin S, *et al*. (2004) Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Br Med J* **329**:15–19; PMID 15231615.

FDA Table of Pharmacogenomic Biomarkers in Drug Labels www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

Pharmgkb database https://www.pharmgkb.org

Wang L, McLeod HL, Weinshilboum RM (2011) Genomics and drug response. *New Engl J Med* **364**:1144–1153; PMID 21428770.

DNA forensics

Annual report of the UK National DNA Database strategy board 2014/5 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/484937/52921_NPCC_National_DNA_Database_web_pdf.pdf

Jeffreys AJ, Wilson V, Thein SL (1985a) Individual-specific 'fingerprints' of human DNA. *Nature* **316**:76–79; PMID 2989708.

Jeffreys AJ, Brookfield JFY, Semeonoff R (1985b) Positive identification of an immigration test case using DNA fingerprints. *Nature* **317**:818–819; PMID 4058586.

Nuffield Council on Bioethics report on forensic bioinformation http://nuffieldbioethics.org/wp-content/uploads/The-forensic-use-of-bioinformation-ethical-issues.pdf (A 2007 UK report that gives a good overview of the ethical issues.)

US CODIS/NDIS Frequently Asked Questions https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet

US CODIS/NDIS statistics https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics

Weir BS (1995) DNA statistics in the Simpson matter. *Nature Genetics* **11**:365–368; PMID 7493014.

Model organisms and modeling disease

21

We begin in Section 21.1 by giving an overview of the range of unicellular and multicellular model organisms and their utility in both scientific and medical research. In the remainder of the chapter we look at different ways of modeling disease, explaining how different types of artificial disease model are produced, what they are useful for, and their limitations. In Section 21.2 we cover cellular disease models and how new technologies using stem cells are driving progress in this area. Pluripotent stem cells are greatly extending the range of cellular disease models. And new three-dimensional cell cultures, often based on self-organizing stem cells, allow different types of cell to assemble *in vitro* into structures that resemble tissues and organs (organoids).

Thereafter we consider modeling disease at the level of the whole organism. Animal disease models are widely used because they allow intensive exploration of the whole organism. They permit broad studies of the pathogenesis, allowing a whole range of invasive analyses to produce data that illuminate our understanding of equivalent diseases in humans. And they can serve as *pre-clinical models*, frontline systems for testing the efficacy and safety of conventional drugs and novel therapeutic strategies before proceeding to clinical trials on humans.

Some animal disease models can be induced by nongenetic factors, but in Section 21.3 we describe how animal disease models are produced as a result of genetic changes, mostly artificially-induced germline mutation or transgenesis. Finally, in Section 21.4 we consider the utility of models for genetic disorders and the difficulties in reproducing human disease phenotypes.

21.1 AN OVERVIEW OF MODEL ORGANISMS

Our planet teems with countless organisms, but only a very few have been studied in the laboratory. Certain species that are amenable to experimental investigation have been well investigated as models, and in this section we are mostly concerned with the scientific utility of model organisms and the interest in microbial pathogens; applications in medical research that rely on model organisms as disease models are covered in Sections 21.2–21.4.

A major advantage of studying model organisms is that they help us understand how human genes function. Because gene function in animal cells has been generally strongly conserved during evolution, we can gain insights from a large range of model organisms—from primates to microbes. For model organisms evolutionarily distant from us, however, the amount that we can infer about human biological processes is limited to only the most highly conserved aspects of cell function, and to fundamental cellular processes. For access to databases on model organisms widely used in genetic studies, see Further Reading for the Model Organism Databases portal supported by the US National Human Genome Research Institute. In addition to the more widely studied model organisms are various nontraditional and emerging model organisms that are especially of interest from the scientific point of view. Many are not described here but interested readers can find examples in PMID 28662661 and 27639630.

Unicellular model organisms aid understanding of basic cell biology and microbial pathogens

Various microbes are particularly suited to genetic and biochemical analyses and offer important advantages such as extremely rapid generation times and convenient large-scale culture. Species studied include representatives of the two prokaryotic kingdoms, bacteria and archaea, plus yeasts (unicellular fungi), protozoa (unicellular animals), and unicellular algae.

A variety of normally nonpathogenic bacteria have been long-standing and popular model organisms, notably *Escherichia coli*. Through intensive studies over decades we have built up more knowledge of *E. coli* than of any other type of cell, and most of our understanding of the fundamental mechanisms of life, including DNA replication, transcription, and protein synthesis, has come from studies of this organism. Other bacteria are also studied because of their economic importance (many are used in preparing fermented foods, in waste processing, in biological pest control, and in the manufacture of antibiotics and other chemicals), or because they are pathogens.

Archaea are not known to be associated with disease; the major interest in studying them is to know more about how they evolved to be so different. Initially found in unusual, often extreme environments, such as at very high temperatures in ocean vents and hot springs, they are now known to inhabit more familiar environments, such as in soils and lakes, and inside the digestive tracts of cows, termites, and marine life where they produce methane. The metabolic and energy conversion systems of archaea resemble those of bacteria, but the systems used to handle and process genetic information (DNA replication, transcription, translation) are more closely aligned to those of eukaryotes.

Yeasts are the unicellular organisms most widely used to model eukaryotic cell functions. They can be easy to culture, are genetically very amenable, and can provide insights into more complex eukaryotes because certain key proteins and fundamentally important cellular functions are known to have been conserved from yeasts to mammals. Most yeasts replicate asexually by budding: the cytoplasm and dividing nucleus from the parent cell is initially a continuum with the *bud*, or daughter yeast, before a new cell wall is deposited to separate the two. Some yeasts, however, replicate like our cells do, by binary fission. Usually, yeasts are not associated with disease, but some *Candida* species can cause common health problems, including candidiasis (thrush).

Two genetically amenable yeasts, the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe*, have been particularly well-studied. *S. cerevisiae* has been used as a model to dissect various aspects of cell biology, notably cell cycle control, protein trafficking, and transcriptional regulation. *S. pombe* closely resembles higher eukaryotes in some aspects of chromosome structure and RNA processing and has been used primarily as a model of cell cycle control. Genetic screens have provided a host of valuable yeast mutants that have been particularly useful for understanding aspects of the cell cycle and DNA repair that are very relevant to our understanding of human cells and cancer. The incremental gain in knowledge of yeast (and some other microbial models) might soon reach a plateau, however: a 2005 perspective by Stanley Fields and Mark Johnston argued that we will know essentially all that is worth knowing about this yeast *S. cerevisiae* somewhere around the years 2025–2035.

Various protozoan models are also studied, including amoebae, flagellates, and ciliates that move using, respectively, pseudopodia, flagella, or cilia. They are of interest to biomedical researchers as models of various facets of cell and developmental biology. Many are also parasites associated with disease, acting as hosts for pathogenic bacteria that cause diseases such as Legionnaire's disease, salmonellosis, and tuberculosis, or they may cause disease directly as in the case of some trypanosomes and malaria-causing *Plasmodium* species. As models of cell and developmental biology, most interest has been focused on the amoeba *Dictyostelium discoideum* and the ciliate *Tetrahymena thermophila*.

Now that unicellular genome sequencing has become routine, a second type of human genome project, the Human Microbiome Project, aims to sequence the diverse microbial genomes within the human **microbiome**, the collective term for the microorganisms that live and interact inside and on humans. Residing mostly on the skin and in the digestive tract, our resident microorganisms outnumber our cells by a factor of 10 or so.

Synthetic unicellular model organisms—made by removing the genome from a microbial cell and replacing it with a synthetic genome—are also expected to be important in the future. The recently constructed Syn 3.0 cell, which has a synthetic minimal bacterial genome containing 530 kb of DNA with just 473 genes (PMID 27013737), is an important step forward. Together with the ability to mutate genes, this type of approach is the beginning of a new era of **synthetic biology**. Synthetic unicellular organisms hold great promise for biomedical research and the possibility of developing novel solutions to environmental problems, such as the production of green biofuels and helping break down toxic waste. Even with in-built protection systems, there are, however, important safety issues and the threat of bioterrorism on a large scale.

Some invertebrate models offer high-throughput genetic screening and insights into gene function

To understand complex cell–cell interactions we need to study multicellular organisms. Metazoan (multicellular animal) models include various invertebrates and vertebrates. Invertebrate models are often easy and inexpensive to maintain and can offer very large numbers of offspring and rapid generation times. These characteristics make them ideally suited to high-throughput genetic screening.

The roundworm *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* are the two most widely studied invertebrates. They are particularly amenable to experimentation and high throughput genetic screening, and because they have homologs of many human genes they have provided valuable insights into how human genes function.

Like other nematodes, *C. elegans* possesses very simple digestive, nervous, excretory, and reproductive systems, but lacks a discrete circulatory or respiratory system. Its small size and short life-cycle make it easy to culture in the lab, and its transparent body facilitates study of cell lineage and its nervous system. These and other features make it a very useful model organism (**Box 21.1**). *C. elegans* is an example of a free-living roundworm but many roundworms are symbiotic or parasitic. They can cause great damage by devouring crops, and by infecting a billion humans to spread diseases including river blindness and elephantiasis.

Although considerably more complex than a nematode, the fruit fly *D. melanogaster* also has a short life cycle and is particularly amenable to sophisticated genetic

BOX 21.1 CHARACTERISTICS OF THE TWO PRINCIPAL INVERTEBRATE ANIMAL MODELS

Caenorhabditis elegans (**Figure 1**, left), a 1 mm-long roundworm, can be cultured easily in the lab, on agar plates (feeding on bacteria), or in liquid culture, and is very amenable to genetic analyses. The predominant sex is hermaphrodite (XX), a modified female that has 959 somatic cells in the adult. By producing both sperm and eggs it can self-fertilize, resulting in increased homozygosity. Rarely, a male (X0) form (with 1031 cells in adults) develops through occasional loss of one of the two hermaphrodite X chromosomes, and hermaphrodites will mate preferentially with males when available. Several areas of study take advantage of the unique features of *C. elegans*, as listed below.

- **Lineage studies.** *C. elegans* is transparent throughout its life cycle; every cell can be seen and followed during development. As a result, it is the only multicellular organism where we know all the cells.
- **Nervous system.** Because *C. elegans* is transparent, the connections between its neurons are readily observed. By the 1980s, a complete wiring diagram of the nervous system was established, identifying all 302 neurons and all the interneuron connections. There are significant similarities to vertebrate nervous systems: the neurotransmitters are mostly the same, and most of the vertebrate genes that make molecular components of the brain have homologs in *C. elegans*.
- **Apoptosis.** Apoptosis forms a key part of *C. elegans* development: 1090 cells develop initially in the hermaphrodite but 131 cells are programmed to die. Genes involved in apoptosis are often highly conserved.
- **Aging.** Because *C. elegans* develops from one cell into the fully-grown form within 3 days, and survives for only a fortnight, the aging process is readily studied.
- **Gene expression and functional studies.** The transparency of *C. elegans* allows protein expression patterns to be followed by linking the green fluorescent protein gene to any *C. elegans* gene or cDNA. Transient inactivation of expression can be achieved for specific genes by *RNA interference* (*RNAi*) and various large-scale RNAi screens have been carried out to understand gene function.

The fruit fly ***Drosophila melanogaster*** (**Figure 1**, middle) has been studied extensively for decades, and a vast amount of information has been amassed on the function of its genes. Some of the major features and approaches that have assisted genetic mapping and functional analyses are listed below.

- **Polytene chromosomes** (**Figure 1**, right). Present in the salivary gland cells in the larval stages, these interphase chromosomes arise through repeated DNA replication without separation into daughter nuclei. Eventually, 1024 copies of a normally single DNA duplex are arranged side-by-side, like drinking straws in a box. As a result, these exceptional interphase chromosomes are visible under the light microscope

Box 21.1 Figure 1 The two most studied invertebrate models. Images are as follows: left, *Caenorhabditis elegans*; middle, *Drosophila melanogaster*; right, polytene chromosomes from *Drosophila* salivary gland cells.

and so chromosome breakpoints and hybridized DNA clones can be readily mapped on chromosomes at high resolution.

- **Spatially and temporally restricted expression of transgenes** is possible using the GAL4-UAS system of conditional gene expression. Large-scale mutagenesis screens are possible. In many cases, loss of function does not result in a mutant phenotype, but transgene misexpression often gives clues to gene function by producing dominant/dominant negative phenotypes. One-generation screens for suppressors/enhancers of dominant mutant phenotypes can identify interacting genes. The yeast flp-frt recombinase system can be used to induce mitotic clones and so form homozygous patches, permitting observation of phenotypes of lethal recessive mutations at late stages of development. Mitotic recombination can also be used in a one-generation screen to score mutant phenotypes in clones and recover lethal mutations that affect late development.
- **The P-element.** This *Drosophila* transposable element permits several types of experimental manipulation, including mutagenesis and transgenesis. Unequal recombination between adjacent P-element inserts can also produce precise deletions.

analyses (**Box 21.1**). Studies over many decades have generated and systematically analyzed a large number of mutants, including many developmental mutants. As a result, we know more about how the fruit fly develops than for any other multicellular organism. It is also a principal model for studying behavior and neuroscience.

A variety of other invertebrate animal models are also being studied to understand aspects of evolution and invertebrate development, or to understand disease, either because they are human parasites or because they transmit pathogenic protozoan parasites or viruses (see **Table 21.1**).

TABLE 21.1 PRINCIPAL REASONS FOR STUDYING MAJOR INVERTEBRATE MODEL ORGANISMS

Organism class	Principal models	Major reasons why studied
Cnidarians and Planaria	*Hydra* *Lineus ruber*	Models of regeneration in a diploblast (*Hydra*) or in a triploblast (*Lineus ruber*)
Molluscs	*Aplysia californica* (sea-slug)	Model of cell behavior and development
Nonparasitic nematodes	*Caenorhabditis elegans*	Model of development, behavior, and aging
Parasitic nematodes	*Wuchereria bancrofti* *Oncocherca volvulus*	To understand pathogenesis (*W. bancrofti* causes elephantiasis; *O. volvulus* causes river blindness), and for economic reasons (nematodes that attack crops)
Nonpathogenic arthropods	*Drosophila melanogaster* (fruit fly)	Model of development, behavior, and neuroscience. Understanding gene function and modeling some human diseases at the cellular level. Rapid drug screening
	Danaus plexippus (monarch butterfly)	To study the cellular and molecular mechanisms underlying a sophisticated circadian clock
	Apis mellifera (honey bee)	Model of learning, memory, and social behavior
Pathogenic arthropods	*Anopheles gambiae* *Aedes aegypti*	*A. gambiae* transmits malaria-causing *P. falciparum. A. aegypti* transmits the Zika virus and viruses causing yellow fever and dengue fever
Echinoderms	Sea urchin (*S. purpuratus*)	Model of development and evolutionary interest in a basic deuterostome
Ascidians	*Ciona intestinalis* (sea squirt)	Model of development and evolutionary interest in a basic chordate
Cephalochordates	*Amphioxus*	Model of development and because evolutionarily closely related to vertebrates

Various fish, frog and bird models offer accessible routes to study vertebrate development

Mammalian development is not easy to analyze. First, egg cells are very small and difficult to manipulate. Secondly, developing embryos implant into the uterine epithelium; development then proceeds inside the mother making it difficult to access material for study. Fish, frogs, and birds produce eggs that can be large in size and easy to manipulate, and the subsequent development proceeds external to the mother. For these reasons, various nonmammalian vertebrate models are routinely studied to illuminate our understanding of vertebrate development.

Two types of small freshwater fish have been popular developmental models: the zebrafish and the medaka (Japanese killifish—see **Box 21.2**). Pufferfish have been studied for a different reason: their genomes are remarkably free of repetitive DNA and so have been useful in comparative genomics studies to aid the discovery of genes and regulatory elements in other vertebrate genomes.

BOX 21.2 CHARACTERISTICS OF SOME POPULAR NONPRIMATE MODEL VERTEBRATES

The **zebrafish** (*Danio rerio*) (**Figure 1A**) has a short generation time and breeds prolifically. Because fertilization is external, all aspects of development are accessible, and the embryo is transparent, facilitating identification of mutants. Mutagenesis screens have produced a large number of valuable developmental mutants. Gene silencing methods are widely used to inactivate specific genes.

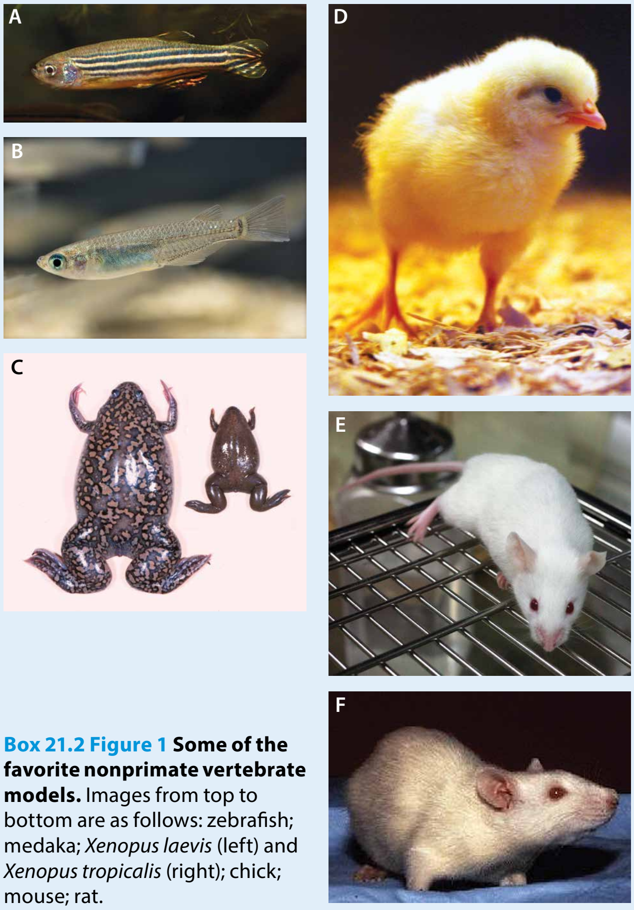
The **medaka** (*Oryzias latipes*) (**Figure 1B**), a distant relative of the zebrafish, is well suited to genetical and embryological analyses, with a short generation time (2–3 months), inbred strains, genetic maps, transgenesis, enhancer trapping, and availability of stem cells.

Two species of ***Xenopus*** (**Figure 1C**), a genus of African clawed frogs, have been widely used in studying development: *X. laevis* (left) and the much smaller *X. tropicalis* (right). All developmental stages are accessible, and the comparatively large size of both eggs and embryos facilitates micromanipulation (microinjections, cell grafting, and labeling). Because each cell of an amphibian embryo contains its own yolk supply, the cells cope well with being transplanted, and continue to differentiate in an explant, or as an individual cell in simple salt solutions. *X. laevis* has a 1–2-year generation time and can be induced to produce 300–1000 large (1–1.3 mm) eggs at a time, but *X. tropicalis* has a <5 months generation time, and can be induced to produce 1000–3000 medium-sized (0.6–0.7 mm) eggs at a time. *X. laevis* is not suited to genetic analyses (it is disadvantaged by having a *pseudotetraploid* genome because of a recent genome duplication), unlike *X. tropicalis*, which has a diploid genome. Transgenesis has long been established.

The **chick** (*Gallus gallus*) (**Figure 1D**) has a very large, and relatively translucent embryo, making delicate microsurgical manipulations easy, and allowing molecular studies to be combined with classical embryology. Popular experimental manipulations of chick embryos include: surgical manipulations and tissue grafting; retrovirus-mediated gene transfer; electroporation of developing embryos; and embryo culture. Rapid advances are being made in chicken transgenics, embryonic stem cell (ESC) technology and cryopreservation of sperm, blastodisc cells, primordial germ cells, and ESCs.

The laboratory **mouse** (*Mus musculus*) (**Figure 1E**) is particularly well suited to genetic studies and is an extensively used model of mammalian development. Its small size and short generation time have allowed large-scale mutagenesis programs and extensive genetic crosses, and various features aid in mapping genes and phenotypes. The ability to construct mice with pre-determined genetic modifications to the germline (by transgenic technology and gene targeting in ESCs) has been a powerful tool in studying gene function.

The laboratory **rat** (*Rattus norvegicus*) (**Figure 1F**), being 10× larger than the mouse, has for many years been the mammal of choice for physiological, neurologic, pharmacological, and biochemical analysis. Genetic analysis in laboratory rats, however, is much less advanced than in mice, partly because of the relatively high cost of rat breeding programs and because until recently, it has been much more difficult to modify the rat germline by gene targeting.



Box 21.2 Figure 1 Some of the favorite nonprimate vertebrate models. Images from top to bottom are as follows: zebrafish; medaka; *Xenopus laevis* (left) and *Xenopus tropicalis* (right); chick; mouse; rat.

Frogs of the genus *Xenopus* (African clawed frog) have been particularly important models for investigating both embryonic development and cell biology. The name *Xenopus* ("strange foot") originates from the sharp claws on the toes of the large, strong, webbed hind feet. *Xenopus* has been an important model for establishing the mechanisms for early fate decisions, patterning of the basic body plan, and organogenesis. There has also been seminal work on chromosome replication, chromatin and nuclear assembly, cell cycle components, cytoskeletal elements, and signaling pathways. All developmental stages are accessible and the comparatively large size of *Xenopus* eggs and embryos facilitates micromanipulation, including microinjections (of mRNA, antibodies, and antisense oligonucleotides), cell grafting, and labeling experiments. Importantly, as for all amphibians, each cell of the embryo contains its own yolk supply. As a result, the cells are well able to cope with being transplanted and to continue to differentiate in an explant (where they can be incubated with selected protein factors) or as an individual cell in simple salt solutions.

Birds have amniotic membranes and their development closely resembles that of mammals. However, while a mammalian embryo depends on the mother for its nutrition (with exchange occurring across the placenta), avian embryos do not have a placenta and so are self-developing systems. Because a bird embryo develops outside the body, it is accessible at all stages in development. The chick is a good model largely because its embryo is easily obtained (see **Box 21.2**).

Mammals are the most relevant animal models but are disadvantaged by practical limitations and ethical concerns

Mammals are closely related to us at the biochemical, developmental, and physiological levels, and almost all human genes have an easily identifiable counterpart in other mammals.

Mammalian models are, however, disadvantaged by various practical considerations. Accommodating and breeding mammals is often very expensive. Mammalian generation times are generally long and the number of offspring is often low, making genetic studies difficult. There are also ethical concerns. Some people are opposed to all animal experimentation. Those who feel that it is justified are often reluctant to accept the need for experimentation on our closest evolutionary cousins. Livestock animals, such as sheep and pigs, have some useful applications, but rodents are by a long way the most studied mammalian models, partly because of their greater amenability to genetic studies, and partly because of financial and ethical considerations. The mouse is the premier model for a variety of reasons, although an increasing number of studies involve larger models, such as rat (see **Box 21.2**), pigs and so on.

Nonhuman primates resemble humans in physiology, cognitive capabilities, detailed brain organization, social complexity, reproduction, and development, and have been particularly important in neuroscience research (rodents, by comparison, are not good models for understanding human brain function). Great apes, being large animals that live in complex social groups, are expensive to maintain. They also have quite long generation times, and their use in animal experimentation has been contentious. Partly as a result, most primate studies have been conducted on smaller primates with shorter generation times, notably macaques and marmosets (see **Table 21.2**). We must bear in mind, however that even primate models are at best imperfect models.

TABLE 21.2 USES OF MAJOR MAMMALIAN MODEL ORGANISMS IN SCIENTIFIC RESEARCH AND IN STUDIES OF PATHOGEN INTERACTIONS

Model	Uses/characteristics
NONPRIMATES	
Cat	Host gene–infectious pathogen interactions. Reproductive physiology, endocrinology, and behavior
Pig/sheep/cow	Mostly as models of human endocrinology/physiology. Host–pathogen interactions in relation to food safety
Dog	An important model for the genetics of behavior
Mouse	The principal model for understanding human gene function and embryonic development. Sophisticated genetic analyses have long been available. Large numbers of mutants available
Rat	Relatively large size makes it more suited to physiological, neurological, pharmacological, and biochemical analyses. Some sophisticated genetic analyses have recently become available
PRIMATES	
Chimpanzee	Evolutionary studies, and to identify the basis of its resistance to developing various infectious diseases
Marmoset	Studies in psychology and physiology, notably to aid understanding of brain function
Rhesus macaque	Most widely used primate model, and an important model in neuroanatomy and neurophysiology

A series of "white paper" discussion documents on individual model organisms and proposed genome projects have been archived at https://www.genome.gov/10002154/approved-sequencing-targets-archive/.

21.2 CELLULAR DISEASE MODELS

Disease modeling can be conducted at different levels. In this section we focus on *in vitro* models that depend on growing cells in culture, that is, cellular disease models. In Section 21.3 we describe *in vivo* models, that is, animal disease models.

One might reasonably ask why we should need cellular disease models at all; surely we need to be looking at the complete picture, the whole organism? Two major reasons for the interest in cellular disease models are: they offer quick and inexpensive analyses, and they allow *human* cells to be analyzed. Animal disease models are disadvantaged because of important species differences (often causing the phenotype to be different from the human disease phenotype). To minimize differences in phenotype we might contemplate using just chimpanzees, our closest evolutionary cousins, as disease models. That, however, is simply not practical because of huge costs, very long timescales required for analyses, and for ethical reasons (chimpanzees are so evolutionarily close to us that they should be in the same genus as us—being anthropocentric, we decided to award ourselves a separate genus, *Homo*). As a result, cellular disease models and animal disease models have complementary advantages and disadvantages; both types of disease model are needed—see **Figure 21.1** for an overview.

	CELLULAR DISEASE MODELS	ANIMAL DISEASE MODELS
MAJOR ADVANTAGES	Rapid analyses of some (mostly molecular) aspects of pathogenesis Human models (mostly) Low cost Usually no ethical concerns	Comprehensive analyses of all aspects of phenotype from embryo to late adult
MAJOR DISADVANTAGES	Limited range of analyses	Not human! High costs Time-consuming analyses Some ethical concerns
MAJOR CLASSES BY ORIGIN	Traditional 2D monolayer cultures Pluripotent stem cell-derived 2D monolayers Organoid and other 3D cultures	Spontaneous models Models derived by expressing transgenes Models derived using targeted mutagenesis Models derived using random mutagenesis

Figure 21.1 Advantages, disadvantages and origins of cellular and animal disease models. A few cellular disease models have been produced from human embryonic stem cells whose construction has involved the destruction of human embryos, which has raised ethical concerns. In the case of animal disease models, many people are opposed to experimentation on primates, and some people are opposed to all animal experimentation. 2D, two-dimensional; 3D, three-dimensional.

Disease modeling using traditional cell cultures

Disease modeling has long been possible using cell lines established from diseased human cells and tissues, and occasional cell lines originating from various animal disease models. Cell lines representing all three major classes of genetic disorder—monogenic disorders, chromosomal disorders, and complex diseases—have been established.

Early cellular disease models were derived from natural cell sources from affected tissues. However, primary cells are often difficult to culture and to maintain, and the cell populations are heterogeneous. To obtain long-lasting homogeneous cultures that offered large numbers of cells for study, immortalized cells were needed. Artificial methods of immortalizing cells were often employed therefore, as detailed in Section 8.1, or natural tumors were sampled (immortalization occurred because the cellular growth control mechanisms had been subverted in tumorigenesis). In each case the cells must exhibit some disease-associated phenotype; but immortalization procedures and continued passaging of cells can mean that the behavior of the cells is different to that of primary cells.

Cellular disease models can offer many advantages. They allow multiple simultaneous tests to be carried out over short time periods, and at very much lower costs than is possible for animal disease models. Genetic and pharmacological intervention, imaging, and biochemical analyses are all much easier than for animal models; ethical and regulatory issues are much less of a concern.

Another major advantage is that *human* cells can be analyzed. As described below, the phenotype of animal disease models can vary significantly from that of the human disease they are intended to mirror, and some fail completely to replicate the human phenotype. In general, cellular disease models offer the opportunity to carry out various types of experimental analyses with two principal motivations, as listed below.

- *Pathogenesis studies.* By culturing cells that are abnormal and comparing them with the equivalent cells from normal individuals, there is the opportunity to explore the relevant disease pathways and to understand the ways in which underlying genes behave abnormally at different levels. That can involve passive molecular analyses, such as whole-genome transcriptome profiling, interactive molecular analyses, such as by externally introducing specific drugs or other reagents to the cell system, and even physiological analyses in the case of certain cells such as muscle and neuron cultures.
- *Drug screening.* Drug screening is most effectively carried out using robotic manipulations. Thousands of cell samples can be deposited in individual miniaturized reaction chambers and exposed to different individual drugs. The hope is that some small-molecule drugs are found to interfere with the disease pathway in a way that lessens the phenotype (a small-molecule drug typically works by binding to a specific protein target with a cleft on its surface; the cleft provides a stereochemical fit for a drug with the right conformation, and if the protein target is key to the pathogenesis, binding of the drug can inhibit it or affect it in some other way). Preliminary indications of possible drug toxicity can also be obtained by exposing cell samples to different concentrations of promising drug candidates.

Cell lines that originate from naturally-occurring disease cells have two main disadvantages. First, the cell lines are inevitably skewed towards easily accessible cells; as a result, disorders where the pathogenesis is in tissues from which biopsies are not generally accessible, cannot readily be modeled. Secondly, traditional cell cultures consist of 2D (two-dimensional) monolayers of a single cell type, and so are not very representative of *in vivo* conditions (where tissues and organs are affected that normally have an ordered three-dimensional [3D] microenvironment where different types of cell interact). Recent developments using stem cells, however, notably the use of artificially-derived human pluripotent stem cells and organoid cultures formed by self-organizing stem cells, go a long way to overcoming the deficiencies of traditional cell cultures.

Extending the range of human cellular disease models using pluripotent stem cell cultures

Recall from Section 4.2 that immortal mammalian pluripotent stem cells are an artificial construction; they do not exist *in vivo*. Certain cells in early development are totipotent (the zygote and its immediate descendants) or pluripotent (such as cells in the inner cell mass of the blastocyst), but they are *transient* cells. They rapidly give rise to more differentiated descendant cells during early development but do not self-renew (unlike naturally immortal or very long-lived tissue stem cells that both self-renew and produce differentiated cells).

Two classes of artificial human pluripotent stem cell lines—notably, induced pluripotent stem cells (iPSCs), and, to a lesser extent, embryonic stem cells (ESCs)—have enormously extended the range of cellular disease models. In each case, pluripotent stem cells are constructed that have disease-predisposing DNA variants and are then manipulated to give suitable differentiated cells that can act as disease models. Recall that pluripotent stem cells can give rise to a wide range of differentiated cells normally formed during development (including cells originating from each of the fundamental germ layers—mesoderm, ectoderm, and endoderm). As a result, it is possible to differentiate pluripotent stem cells derived from the cells of a patient to give rise to disease models of interest, including for disorders where access to cell models has been very limited because of difficulties in accessing biopsy samples.

ESCs as platforms for disease modeling

ESCs were the first human pluripotent stem cells to be constructed. They have usually been made by culturing cells from the inner cell mass of blastocysts from surplus human embryos obtained during in vitro fertilization (IVF) procedures. Culturing occurs under special conditions to obtain cells that can function as immortal stem cells (able to both self-renew and give rise to more differentiated descendant cells). In some cases, disease models have been made by using genome-editing techniques to genetically modify ESC cultures obtained from normal human embryos (**Figure 21.2A**).

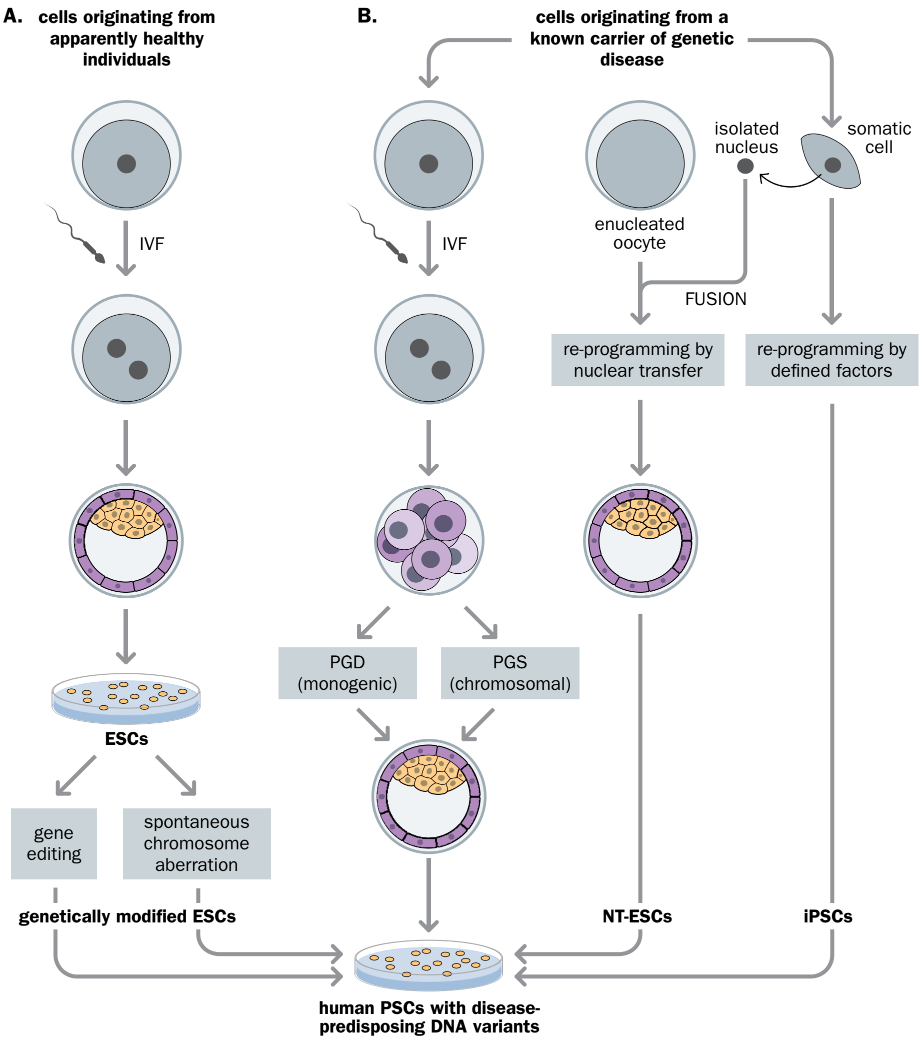

A. cells originating from apparently healthy individuals
B.
cells originating from a known carrier of genetic disease
IVF
IVF
enucleated oocyte
isolated nucleus
somatic cell
FUSION
re-programming by nuclear transfer
re-programming by defined factors
ESCs
PGD (monogenic)
PGS (chromosomal)
gene editing
spontaneous chromosome aberration
genetically modified ESCs
NT-ESCs
iPSCs
human PSCs with disease-predisposing DNA variants

Figure 21.2 Strategies for generating disease models using human pluripotent stem cells (PSCs). Human PSCs carrying a genetic abnormality can be generated by utilizing apparently healthy cells or cells carrying known disease-predisposing mutations. (**A**) Isolated embryonic stem cells (ESCs) generated from apparently normal surplus embryos made available after in vitro fertilization (IVF) can be genetically edited at a specific locus, to generate *de novo* disease-predisposing mutations. As ESCs can acquire spontaneous chromosomal aberrations in culture, they can sometimes be used to model chromosomal disorders such as Turner syndrome. (**B**) Utilizing cells from a carrier of a genetic disorder provides other alternatives. If pre-implantation genetic diagnosis (PGD) or pre-implantation genetic screening (PGS) has been carried out following IVF, embryos carrying mutant genotypes or chromosome abnormalities can be identified and used to produce ESCs carrying the disease-associated genetic change to serve as models for monogenic or chromosomal disorders. Alternatively, somatic cells from patients can be re-programmed into PSCs either by transferring their nucleus into an enucleated oocyte to generate nuclear transfer embryonic stem cells (NT-ESCs), or by the use of defined factors to generate induced pluripotent stem cells (iPSCs). The endpoint for each method, therefore, is human PSCs with disease-associated DNA variants that can be used to model genetic disorders. (Reprinted from Avior Y *et al.* [2016] *Nature Rev Mol Cell Biol* **17**:170–182; PMID 26818440. With permission from Springer Nature. Copyright © 2016.)

ESC lines can also be prepared from cells contributed by carriers of a genetic disease. During IVF, for example, an IVF embryo judged to be homozygous for a recessive disorder during pre-implantation genetic diagnosis, or one that is found to have a chromosomal abnormality during embryo screening, can be used as a source of cells from the inner cell mass to make disease-associated ESC lines. The alternative is to use somatic cell nuclear transfer to re-program an enucleated oocyte (from an oocyte donor) by a nucleus from the somatic cell from a patient, but this is a very difficult and laborious method (see Figure 21.2B).

Disease modeling using iPSC-derived cells

Differentiated human cells can be re-programmed to pluripotency after exposing them to certain factors, such as a cocktail of four transcription factors known to be important in embryonic development (detailed in Section 4.2). The re-programmed cells, known as induced pluripotent stem cells (iPSCs), can then be induced to differentiate into cell types that show associated pathology in patients with the disease (Figure 21.2B).

Producing iPSC lines is much simpler than producing ESC lines, and the method is not disadvantaged by the type of ethical concerns associated with human ESCs (the normal procedure to obtain human ESCs involves destruction of human embryos). Added to that is a huge advantage: easily accessible cells, such as skin fibroblasts, can be re-programmed to pluripotency, so that *patient-specific* cell lines can be produced. As a result, the iPSC route has become the predominantly-used method for producing new cellular disease models, and large banks of disease-specific iPSCs and patient-specific iPSCs, are being produced as platforms for disease modeling and drug discovery. In the past, cellular models of some disorders were not available. Models of cardiac and neurologic disorders were lacking, for example, because heart and brain tissues are particularly inaccessible via patient biopsies; now iPSCs can be conveniently made from patients and then differentiated to make cardiomyocytes or neurons.

For monogenic disorders, iPSC-based disease models are preferentially suited to modeling highly penetrant disorders in which the differentiated cells of affected tissues display **cell-autonomous** defects, that is, defects exhibited by genotypically mutant cells only. For modeling complex disease, the iPSC route provides some advantages over animal disease models (which have generally been very disappointing). As well as being disadvantaged by human–animal differences, there is the major problem that the full genetic contribution to complex diseases is not appreciated: animal models of complex disease can be designed to have the animal equivalent of just the *known* disease-predisposing DNA variants only. By contrast, the iPSC-based route can offer **isogenic** disease models: iPSCs from a patient with a complex disease, and differentiated cells derived from them, will have the same, full genetic contribution that predisposed to the disease in the first place.

Modeling disease and development *in vitro* using organoid cultures based on stem cells

Although very useful for many purposes, traditional 2D monolayer cultures have clear limitations because the cells are unable to achieve the structural organization and connectivity found *in vivo*. As a result, the properties of the cells in 2D cultures—morphology, proliferation, differentiation, gene expression, and so on—are inevitably affected. However, **organoids**, *in vitro* 3D clusters of cells deriving exclusively from primary tissue or stem cells, are capable of self-renewal and self-organization. In the latter case, for example, cell sorting occurs whereby cells preferentially adhere to other cells of the same type; cells of the same type have the same adhesive properties, and so a mixture of two cell types can spontaneously undergo cell sorting to form homogeneous regions formed by a single cell type (**Figure 21.3A**). Cell differentiation can also be regulated by spatial positioning of cells (see **Figure 21.3B**).

The self-organizing capacity of mammalian cells has long been appreciated, as has the role of the extracellular matrix in providing a 3D structural support for cells. Attempts to develop 3D cell cultures began in the 1980s and over time different 3D cell culture methods were developed. Some use scaffolds, often hydrogels such as Matrigel, a

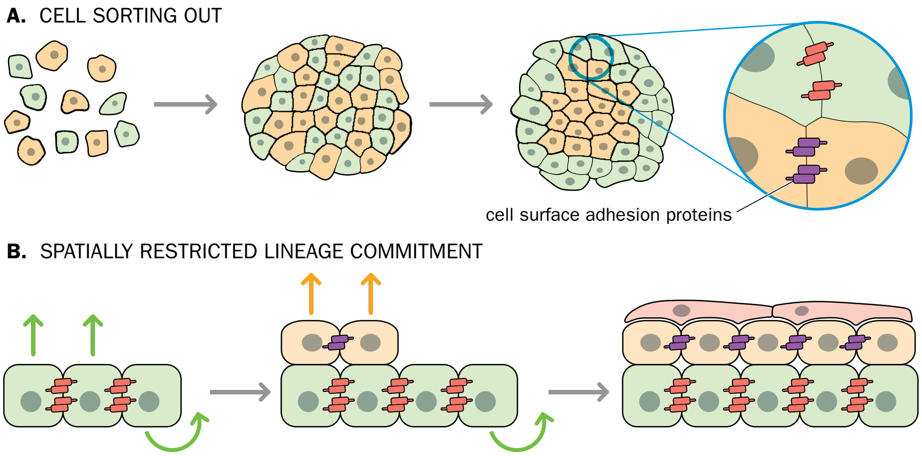

A. CELL SORTING OUT
cell surface adhesion proteins
B. SPATIALLY RESTRICTED LINEAGE COMMITMENT

Figure 21.3 Principles of natural cell self-organization that also occur in organoid formation. (**A**) Cell sorting out describes the movement of cells into different domains. Different cell types (beige or green) sort themselves because of different adhesive properties conferred by their differential expression of distinct cell adhesion molecules (shown as purple or orange bars). (**B**) Spatially restricted cell-fate decisions also contribute to self-organization *in vivo* and in organoids. Progenitors (green) give rise to more differentiated progeny (beige), which, because of spatial constraints of the tissue and/or division orientation, are forced into a more superficial position that promotes their differentiation. These cells can sometimes further divide to give rise to more differentiated progeny (pink), which are further displaced. (From Lancaster MA & Knoblich J [2014] *Science* **345**:1247125; PMID 25035496. Reprinted with permission from the AAAS.)

gelatinous protein mix that serves as a substitute for an extracellular matrix. But the field really began to develop by using pluripotent stem cells as a way of directing the formation of different types of organoid and by applying knowledge of endogenous stem cells and their microenvironments (*stem cell niches*; detailed in Section 4.2). Two key advances were made. In 2008 Yoshiki Sasai and colleagues reported making 3D cerebral cortex tissue from pluripotent stem cells. In 2009 Hans Clevers and co-workers described making gut organoids from defined adult intestinal stem cells after 3D culture in Matrigel plus epidermal growth factor and R-spondin, a Wnt agonist. Subsequently, a wide range of different organoids have been produced by defined differentiation steps beginning from pluripotent stem cells—see **Figure 21.4** for some examples. Unlike in 2D monolayer cultures, the structure of organoids is much closer to the *in vivo* situation (see **Figure 21.5A** for the example of a cerebral organoid).

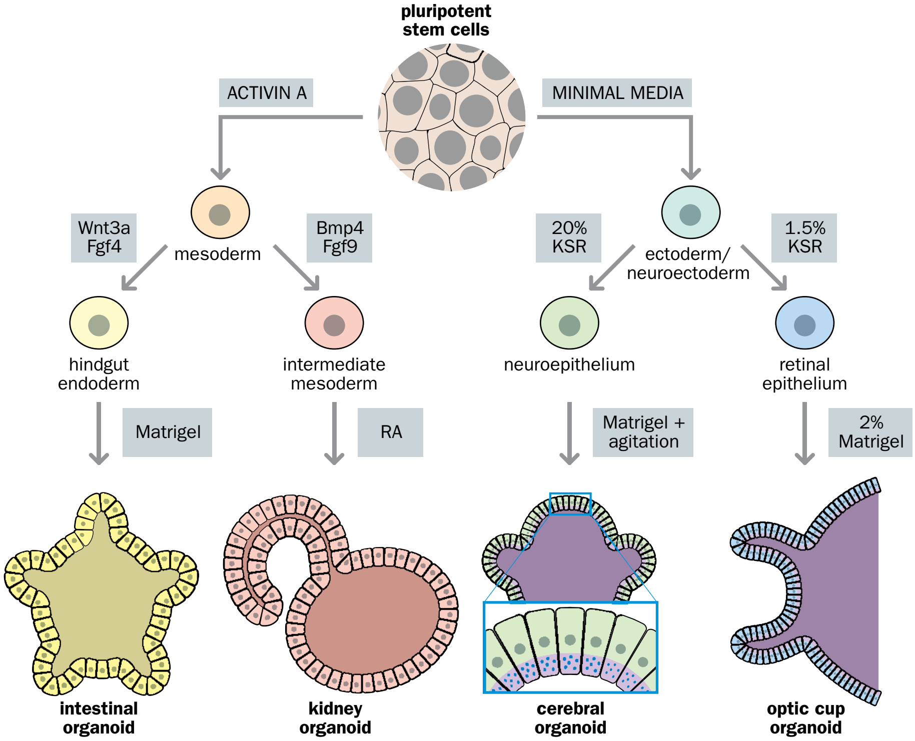


Figure 21.4 Examples of strategies for organoid differentiation from pluripotent stem cells (PSCs). Conditions and growth factors are indicated for the derivation of progenitor identities. For neuroectoderm, minimal medium without serum is used. KSR is knockout serum replacement, a serum-free growth-promoting alternative. Low concentrations of KSR, along with a low concentration of Matrigel dissolved in the medium, promotes retinal neuroepithelium, whereas higher KSR concentrations and embedding in pure Matrigel promotes the formation of various brain regions. Renal organoids have been generated in several ways, but growth factors in common are shown. RA, retinoic acid. (From Lancaster MA & Knoblich J [2014] *Science* **345**:1247125; PMID 25035496. Reprinted with permission from the AAAS.)

By exhibiting functional properties similar to the tissue of origin, organoids constitute a bridge between traditional 2D cell cultures and the way in which cells are organized *in vivo*. As a result, they might be expected to lead to more representative cellular disease models. Because iPSCs can be directed to differentiate to give specific types of organoid, disease-specific and patient-specific organoids can be conveniently made: a skin biopsy is taken from the patient and skin fibroblasts are re-programmed to make iPSCs, which in turn are induced to differentiate to give a desired type of organoid. That has permitted analyses of the pathogenesis in human tissues that have not previously been accessible, such as cerebral organoids in the case of various neurologic and neurodevelopmental disorders—see **Figure 21.5B** for an example.

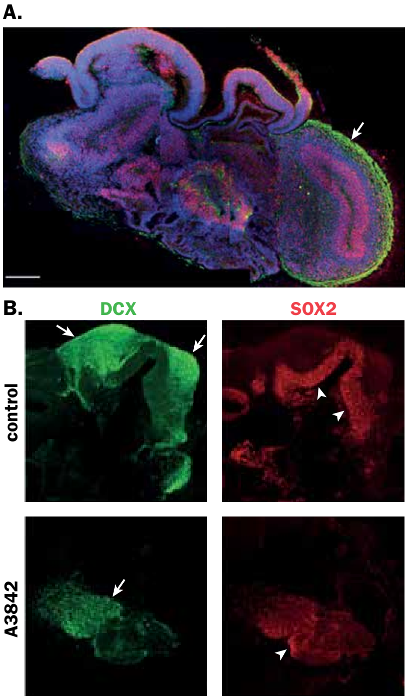


Figure 21.5 Complexity of cerebral organoid structure and altered structure in inherited microcephaly. (**A**) Sectioning and immunohistochemistry reveal complex morphology and cellular heterogeneity in a cerebral organoid prepared by differentiation from normal pluripotent stem cells. Heterogeneous regions are revealed here by staining with a SOX2-specific antibody (revealing neural progenitors in red) and a TUJ1-specific antibody that recognizes neuron-specific class III beta-tubulin (revealing neurons in green). (**B**) Immunohistochemistry in control cerebral organoid tissues and tissues derived from a cerebral organoid model of microcephaly (A3842). The latter was prepared by differentiation of induced pluripotent stem cells from a patient who has primary microcephaly as a result of inactivating mutations in both alleles of the *CDK5RAP2* gene. The figure shows smaller neural tissues in the microcephaly model at day 30 of differentiation, with fewer neurons (revealed using an antibody specific for doublecortin, DCX, arrows) and smaller progenitor zones (SOX2, arrowheads). (Reprinted from Lancaster MA *et al*. [2013] *Nature* **501**:373–379; PMID 23995685. With permission from Springer Nature. Copyright © 2013.)

Although organoids show considerable promise as disease models, they have their limitations. For example, cerebral organoids are highly variable, and although they share some similarities in organization to the developing brain there are differences, too.

21.3 ORIGINS OF ANIMAL MODELS OF GENETIC DISORDERS

The great majority of animal disease models are artificially constructed. Many are produced by genetic methods, but some animal disease models have been constructed by alternative approaches. For example, harmful chemicals or neurotoxins can be injected into animals to destroy certain cells (pancreatic beta cells in the case of diabetes) or certain neurons (in the case of Parkinson disease), or specific antibodies can be injected to cause disease (such as collagen-specific antibodies to induce rheumatoid arthritis, for example). Infectious diseases have been modeled by infecting animals with pathogens.

Here we focus on animal disease models that originate through *genetic* methods, that is, through modification of germline DNA. That can happen as a result of spontaneous mutations or chromosomal abnormalities in germline DNA, which have been identified in a wide range of experimental organisms, pets, livestock, and so on. As described in Section 21.4 the current range of animal models for Duchenne muscular dystrophy contains many spontaneous disease models.

The great majority of genetic animal disease models are constructed artificially, either by mutagenesis (mutation of endogenous genes) or transgenesis (introducing and expressing additional genes). Different genetic protocols can be used, but they essentially fall into three major classes as listed below (and as described more fully in the subsections that follow).

- *Transgene/transchromosome expression.* A transgene designed to produce a positively harmful gene product, or over-express a gene product, is inserted into germ-line DNA, or an artificial chromosome is introduced into the germ line of an animal to produce a transchromosomic animal. This approach is taken when one wants to construct a model for a specific phenotype arising from DNA changes that cause some harmful gain of function.
- *Targeted mutagenesis.* Genome editing is used to make specific changes to one or more pre-determined sequences in the genome of cells contributing to the germ line. Chromosome engineering can be carried out using site-specific recombination with Cre-*loxP* (detailed in Section 8.3) and can be used to make models of defined aneuploidies and chromosome microdeletions/microduplications. However, in most cases targeted mutagenesis is employed to produce desired models of monogenic disorders caused by loss-of-function phenotypes.
- *Random mutagenesis.* Unlike the two classes above in which the intention is to make specific disease models, disease models can also be produced in an unpredictable way by random mutagenesis, using a chemical mutagen or ionizing radiation. This type of mutagenesis is phenotype-driven: we do not know which genes will be affected, but we know that abnormal phenotypes will be produced. By carefully studying the phenotypes, models of different types of genetic disorders can be identified (by comparing them with known disease phenotypes).

Modeling disorders due to DNA changes that cause a harmful gain of function

Some pathogenic mutations and large-scale DNA changes cause a gain of function. The problem is not a deficiency of a gene product: instead, a gene product (or products) becomes actively harmful to cells. It may be a mutant protein or toxic RNA, or it may be that too much gene product is made (by overexpression of some dosage-sensitive genes, or inappropriate expression of an oncogene). In all these cases the result is aberrant cell behavior or cell death causing disease, and the damage is caused even in the presence of the wild-type allele. Somatic DNA changes causing gain of function are common in tumors, and as detailed in Section 16.2 many dominantly-inherited disorders result from inherited gain-of-function mutations.

Modeling disease due to harmful mutant proteins

Modeling the effect of positively harmful mutant proteins involves **transgenesis**, a procedure that we previously introduced in Section 8.6. First, a transgene is prepared with a coding DNA that can be used to express the mutant protein (sometimes a mutant human cDNA is used; or the equivalent animal cDNA is genetically engineered to replicate the pathogenic human mutation). Then the transgene is inserted into germ-line DNA, often by injecting it into the pronucleus of a fertilized oocyte. For an illustration, see **Figure 8.21** that illustrates transgenesis by pronuclear microinjection in the case of the mouse, but similar transgenesis procedures have long been carried out in a wide variety of animal species.

Huntington disease arises by gain-of-function mutations resulting in unstable expansion of CAG repeats in exon 1 of the large *HTT* (huntingtin) gene. Although the precise molecular mechanism is poorly understood, the pathogenesis is believed to result from a harmful mutant protein that has a polyglutamine tract with >36 glutamine residues. Some disease models have simply been designed to express a mutant *HTT* exon 1 (containing an expanded number of CAG repeats) under the control of the human *HTT* promoter. Other models express the full length huntingtin protein, either by a *knock-in* strategy (a mutant human *HTT* exon 1 is inserted into the endogenous mouse *Htt* gene), or by introducing a mutant human transgene that can make the full length mutant human protein, often by expressing a full length mutant *HTT* gene in a yeast or bacterial artificial chromosome under the control of the human *HTT* promoter. Whereas, the models expressing mutant exon 1 can show very pronounced somatic (and germ line) repeat instability, full length mutant *HTT* in yeast or bacterial artificial chromosomes is comparatively stable. We consider progress in modeling Huntington disease in Section 21.4.

Modeling pathogenic overexpression

Pathogenesis due to overexpression of a gene can arise in different ways. Sometimes it arises through mutations in regulatory sequences, but mutations like this are generally not easy to replicate in animals: short regulatory sequences evolve rapidly (and can show substantial differences between different mammals), and our understanding of individual regulatory sequences is often poor.

Pathogenic overexpression of a gene often arises through aberrant gene duplication. Having three gene copies instead of two in diploid cells means that more gene product is made; but for some genes where the amount of product needs to be very tightly controlled, that causes major problems. Some common pathogenic duplications extend over megabase regions of DNA containing multiple genes (*chromosome microduplications*; detailed in Section 15.3), and even whole chromosomes, as in trisomy 21. The neuropathy type 1A Charcot–Marie–Tooth disease, for example, usually results from a 1.4 Mb duplication that spans multiple genes, but the disease arises because of dosage-sensitivity in one gene, *PMP22* (peripheral myelin protein 22). As a result, the disease can be modeled by making transgenic animals with extra copies of a *PMP22* transgene. Transgenic mice and rats with additional copies of a *PMP22* transgene have proved to be reasonable models of the disease.

Trisomies constitute special cases of the harmful effects of gene overexpression: pathogenesis results from the combined effects of overexpressing multiple dosage-sensitive genes. For human autosomes with moderate to high gene density, many dosage-sensitive genes are affected in this way and embryonic lethality results. But trisomies are viable for some chromosomes with low gene content—chromosomes 13, 18, and 21 (trisomy X is also viable, but only because two of the three X chromosome copies are subject to X-inactivation). Human trisomy 21, the major cause of Down syndrome, is especially common and three different approaches have been taken to make animal models of Down syndrome (see **Box 21.3**).

BOX 21.3 THREE DIFFERENT WAYS OF MAKING ANIMAL MODELS OF DOWN SYNDROME

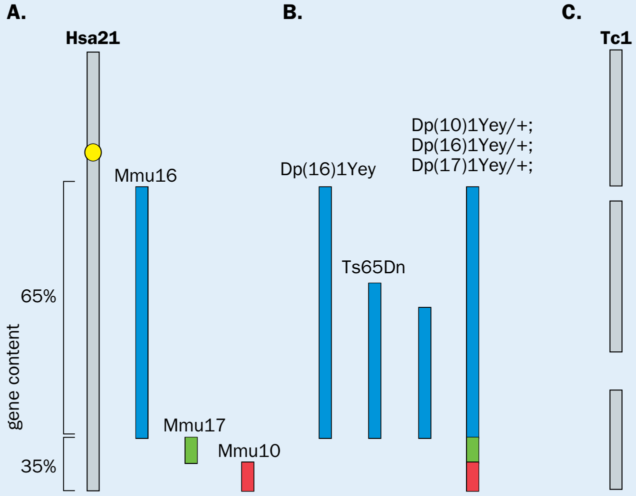
Down syndrome, the most common genetic cause of intellectual disability, is also associated with significant heart abnormalities, psychiatric problems, and risk of developing early-onset Alzheimer disease. Animal disease models offer the opportunity of invasive experimental analyses to better understand the pathogenesis, and various mouse models have been constructed, but none are perfect. The most common cause is trisomy 21, but modeling trisomy 21 is difficult because of human–mouse differences in gene order. Genes on human chromosome 21 have orthologs on multiple mouse chromosomes, but because 21p is essentially made up of heterochromatin and of rRNA genes that have multiple, nearly identical copies on four other chromosomes, the problem lies with overexpression of certain genes on 21q.

Most of the human 21q genes have counterparts on mouse chromosome 16, but those located in the distal 4.5 Mb (~35% of them) have orthologs on mouse chromosome 17 or 10 (see **Figure 1A**). Mouse trisomy 16 mice are not useful models (they die *in utero* and do not resemble human trisomy 21; importantly, the majority of genes on mouse chromosome 16 have orthologs on chromosomes other than human chromosome 21). To obtain more accurate models, all three different ways of producing animal models—random mutagenesis, targeted mutagenesis, and transgenesis/transchromosomics—have been employed.

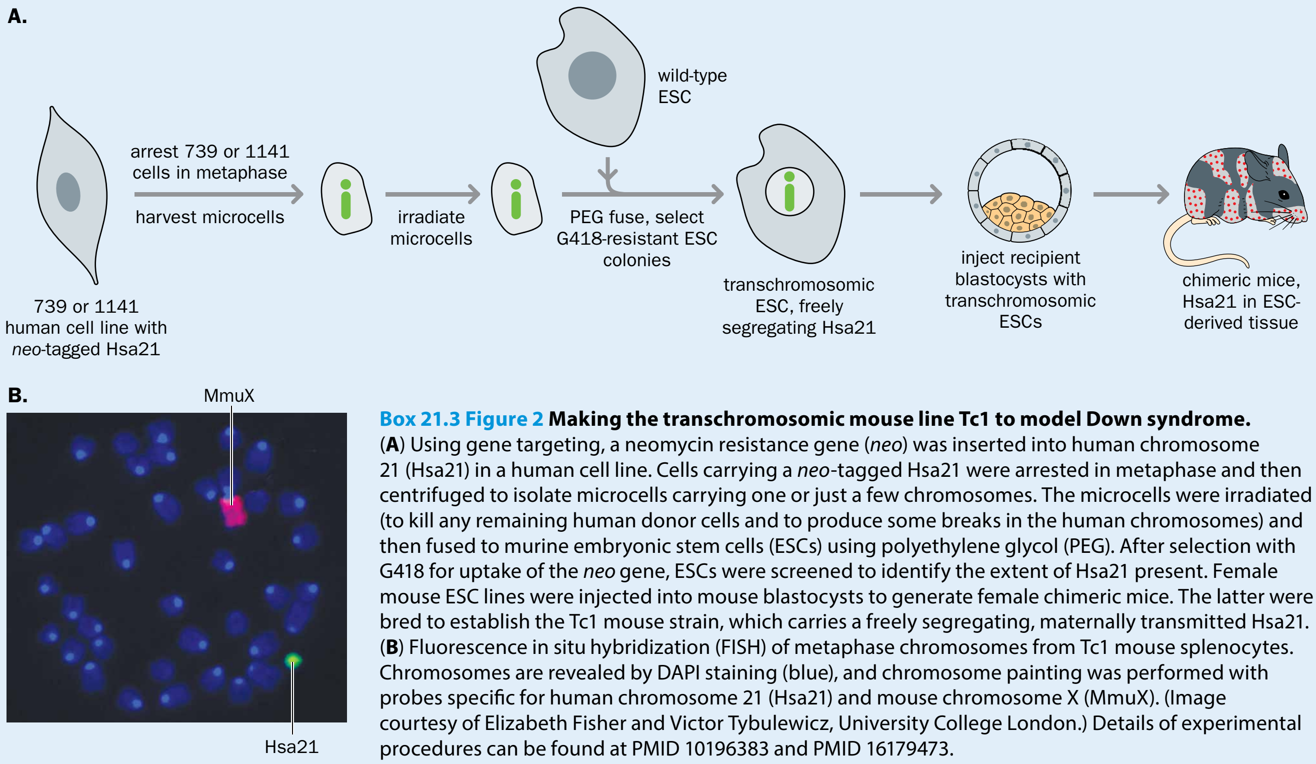
The first Down syndrome models were made using random mutagenesis: after having their testes irradiated, male mice were bred from and their offspring were screened for chromosomal rearrangements involving mouse chromosome 16, including partial trisomy 16 arising through translocations. The most widely-used model produced in this way is the Ts65Dn mouse, because of its learning and behavior deficits. It was found to be trisomic for <60% of the human chromosome 21 syntenic region on mouse chromosome 16 (see **Figure 1B**) and the trisomic region was found to be present as a freely segregating extra chromosome. However, it is also trisomic for many genes on a region of chromosome 17 that has homologs on human chromosome 16.

Targeted mutagenesis has also been applied to make models, involving chromosome engineering using the Cre-*loxP* system (for the background, see Section 8.3 and **Figure 8.14**). In the first case *loxP* sequences were stitched into mouse chromosome 16 at positions marking the ends of the 22.9 Mb region that is syntenic with human chromosome 21 (Hsa21). Cre-driven recombination between *loxP* sequences on two mouse chromosome 16 homologs sometimes results in duplication of the full 22.9 Mb region, giving rise to the Dp(16)1Yey mouse model (**Figure 1B**). Similar engineering on mouse chromosomes 10 and 17 produced mice with duplications of the regions syntenic with distal human 21q, the Dp(10)1Yey and Dp(17)1Yey mice, and crossbreeding these with Dp(16)1Yey mice has produced a promising model that is trisomic for the entire Hsa21 syntenic regions on Mmu10, Mmu16, and Mmu17 (**Figure 1B**) and exhibits multiple Down syndrome phenotypes.

A final mouse model, a transchromosomic mouse with a freely segregating single human chromosome 21, was made by transferring chromosomes from human fibroblast cells into mouse embryonic stem cells (ESCs) using microcell-mediated chromosome transfer (**Figure 2**). Using a marker gene, ESCs that had picked up an almost complete human chromosome 21 were selected and used to make the Tc1 transchromosomic mouse line. The artificial human chromosome 21 lacks about 8 Mb of the original DNA (see **Figure 1C**), but the main difficulty with this model is random loss of the human chromosome during development, causing variable levels of mosaicism in different tissues and thereby complicating the analyses.



Box 21.3 Figure 1 Segmental trisomy for chromosomal regions corresponding to human chromosome 21q in some mouse models of Down syndrome. (**A**) Location of mouse orthologs of genes on human 21q. Thick vertical bars at left indicate that genes in the most proximal 23 Mb of human 21q have orthologs on mouse chromosome 16. Although quite small, the most distal 4.5 Mb of human 21q has 35% of the genes on 21q, with orthologs on mouse chromosomes 17 (Mmu17) or 10 (Mmu10). (**B**) Vertical bars represent the extent of trisomy for mouse orthologs of genes on human 21q in indicated mouse models. The three-colored vertical bar at right represents a promising model that is trisomic for the entire HSA21 syntenic regions on Mmu10, Mmu16, and Mmu17. (**C**) Extent of human chromosome 21 sequences in the Tc1 transchromosomic mouse line that carries an artificial human chromosome 21. Gaps indicate missing DNA sequence. (Adapted from Rueda N *et al*. [2012] *Neural Plasticity* Article ID 584071; PMID 22685678.) Hsa, *Homo sapiens*; Mmu, *Mus musculus* (mouse).


A.
wild-type ESC
arrest 739 or 1141 cells in metaphase
harvest microcells
irradiate microcells
PEG fuse, select G418-resistant ESC colonies
transchromosomic ESC, freely segregating Hsa21
inject recipient blastocysts with transchromosomic ESCs
chimeric mice, Hsa21 in ESC-derived tissue
739 or 1141 human cell line with *neo*-tagged Hsa21
B.
MmuX
Hsa21

Box 21.3 Figure 2 Making the transchromosomic mouse line Tc1 to model Down syndrome. (**A**) Using gene targeting, a neomycin resistance gene (*neo*) was inserted into human chromosome 21 (Hsa21) in a human cell line. Cells carrying a *neo*-tagged Hsa21 were arrested in metaphase and then centrifuged to isolate microcells carrying one or just a few chromosomes. The microcells were irradiated (to kill any remaining human donor cells and to produce some breaks in the human chromosomes) and then fused to murine embryonic stem cells (ESCs) using polyethylene glycol (PEG). After selection with G418 for uptake of the *neo* gene, ESCs were screened to identify the extent of Hsa21 present. Female mouse ESC lines were injected into mouse blastocysts to generate female chimeric mice. The latter were bred to establish the Tc1 mouse strain, which carries a freely segregating, maternally transmitted Hsa21. (**B**) Fluorescence in situ hybridization (FISH) of metaphase chromosomes from Tc1 mouse splenocytes. Chromosomes are revealed by DAPI staining (blue), and chromosome painting was performed with probes specific for human chromosome 21 (Hsa21) and mouse chromosome X (MmuX). (Image courtesy of Elizabeth Fisher and Victor Tybulewicz, University College London.) Details of experimental procedures can be found at PMID 10196383 and PMID 16179473.

Modeling disorders caused by loss-of-function mutations

Recessive disorders and dominantly inherited disorders due to haploinsufficiency can be modeled using targeted inactivation of an orthologous animal gene to produce a **gene knockout** (where the gene is completely inactivated). Figure 21.6 gives an overview of three different approaches that have been taken, and we consider these in turn immediately below.

Since the 1980s until quite recently, the models that were made in this way were almost always mouse models (unlike modeling disease due to gain-of-function mutations where transgenesis was used and disease models could be made easily in other mammals). In addition to other favorable characteristics, the mouse was the number one choice for modeling disease caused by loss of function because embryonic stem cell (ESC) lines derived from the mouse 129 strain provided an especially efficient way of introducing mutant alleles into germline DNA through homologous recombination-based genome editing; ESC lines from other strains of mice and from other mammals were much less effective.

To replicate the human phenotype in mice, the orthologous mouse gene was traditionally inactivated in cultured ESCs by some type of insertional inactivation. Often, for example, homologous recombination in ESCs would be used to replace one or more early exons in the endogenous gene by a short reporter sequence, effectively making a deletion at the start of the gene that was typically intended to cause a frameshift early in the translational reading frame. Figure 21.6A gives a very general overview of the ESC route to making a gene knockout. We covered the detail of this approach in Figure 8.22, and Figure 8.24 shows an example of a specific targeting strategy that was used to make a mouse model of Ellis van Creveld syndrome. For autosomal recessive disorders, disease models are produced by first identifying carriers of the desired mutation, and then mating carrier males and females. Now, as a result of the International Knockout Mouse Consortium (detailed below), efforts are being made to make mouse knockouts for nearly all of the protein-coding genes in the mouse. The resulting knockout mouse strains will be made widely available.

Making gene knockouts in other mammals

For decades, suitably efficient ESC lines, comparable to mouse strain 129 ESCs, were not available in other mammalian species. Instead, different routes into the germ line were sometimes taken, notably by the inefficient and rather laborious process of somatic cell nuclear transfer (the same methodology used when making the world's first cloned mammal, Dolly the sheep). It involves fusing a somatic cell nucleus with an enucleated

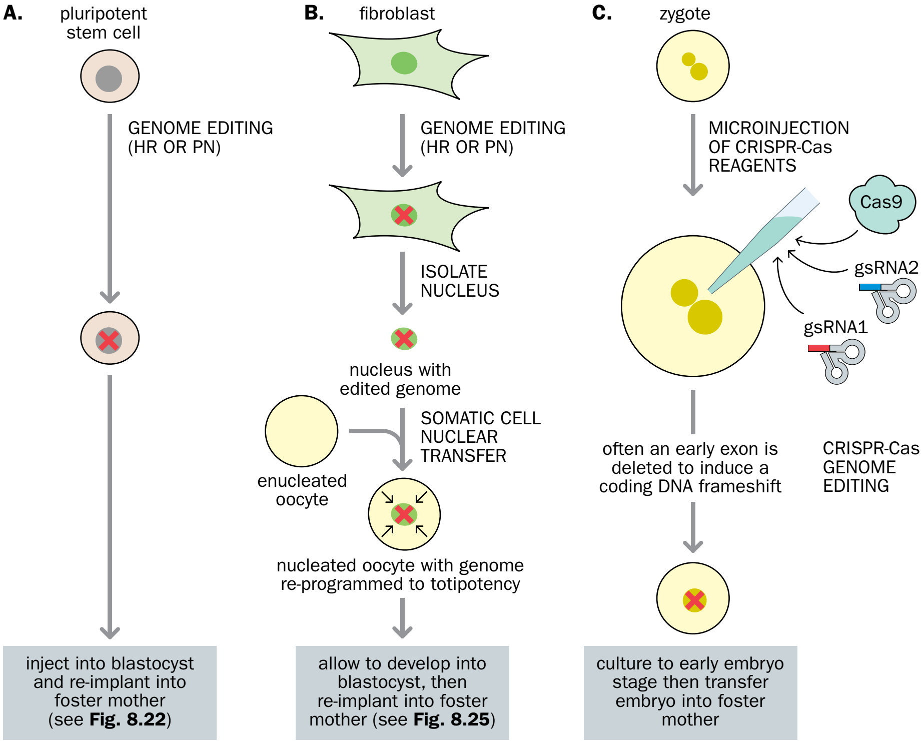

A.
pluripotent stem cell
GENOME EDITING (HR OR PN)
inject into blastocyst and re-implant into foster mother (see Fig. 8.22)
B.
fibroblast
GENOME EDITING (HR OR PN)
ISOLATE NUCLEUS
nucleus with edited genome
SOMATIC CELL NUCLEAR TRANSFER
enucleated oocyte
nucleated oocyte with genome re-programmed to totipotency
allow to develop into blastocyst, then re-implant into foster mother (see Fig. 8.25)
C.
zygote
MICROINJECTION OF CRISPR-Cas REAGENTS
Cas9
gsRNA2
gsRNA1
often an early exon is deleted to induce a coding DNA frameshift
CRISPR-Cas GENOME EDITING
culture to early embryo stage then transfer embryo into foster mother

Figure 21.6 Three routes towards making specific mammalian models of disease due to loss-of-function mutations. The general object is to inactivate a gene or genes by genome editing, often by deleting a short early exon that causes a frameshift in the coding DNA. Genome editing can be carried out in cultured cells (which are used as vehicles to transfer edited genomes into the germ line), or directly in germ line cells, most commonly by microinjection of the zygote. (**A**) Until quite recently, genome editing to inactivate genes was carried out almost exclusively in pluripotent stem cells, notably mouse strain 129 embryonic stem cells, and editing was achieved by homologous recombination (HR). More recently, pluripotent stem cells from other species have been used, and genome editing using programmable nucleases (PN) has also been carried out. (**B**). An alternative approach involves carrying out genome editing in somatic cells, notably fibroblasts, and microinjecting a nucleus with an edited genome into an enucleated oocyte. Cytoplasmic factors in the oocyte then re-program the nucleus (indicated by black arrows) to make a totipotent cell that can give rise to an animal with the desired mutant allele. Breeding of heterozygote animals can then be used to produce homozygotes. (**C**) Genome editing by microinjection of CRISPR-Cas9 reagents into isolated zygotes. Often an indel is induced in coding DNA to inactivate a gene. As shown, a Cas9 nuclease plus two guide RNA sequences (gRNA) can be injected to make one double-stranded DNA break (which is incorrectly repaired by the nonhomologous end joining DNA repair pathway to produce the indel). Or two double-strand breaks can be induced to ensure a large deletion. Interested readers can find the CRISPR-Cas9 technical details in PMIDs 25271304 and 25058643.

oocyte, after which oocyte cytoplasmic factors re-program the somatic cell nucleus to totipotency. An early application was to make the first large-animal models of cystic fibrosis, a pig model and a ferret model that were first reported in 2008. In these cases, gene editing using homologous recombination was used to knock out the *Cftr* (cystic fibrosis transmembrane regulator) gene, or produce the delF508 mutant, in cultured fibroblasts. Then the mutant allele was introduced into the germ line by somatic cell nuclear transfer (see **Figure 21.6B**).

Recent major technological advances have removed the impediments to making gene knockouts in mammals other than mice. Successive pluripotent stem cell lines permitting efficient transfer of mutant alleles into the germ line have been constructed for many other mammals (detailed in Section 4.2) and simple, but efficient, CRISPR-Cas genome editing has become routine. The latter method can even be carried out by microinjection of reagents into the zygote (see **Figure 21.6C**).

Making animal models by gene knockdown

Gene silencing means selectively inhibiting the expression of a desired target gene by targeting the RNA transcripts. Unlike gene knockouts (where the object is complete gene inactivation), gene silencing often results in at least some residual gene function. As a result, the effect is described as a *gene knockdown*, but gene silencing can nevertheless produce interesting phenotypes. The methods are simple and fast and can be highly efficient in some model systems, being often accomplished in *C. elegans* using RNA interference, and by using antisense morpholino oligonucleotides in zebrafish (Section 8.5 gives the details of the technologies used).

Many animal models of human genetic disorders have been identified by phenotype screening in random genome-wide mutagenesis programs

In the sections above we have been concerned either with transgenesis, or with targeted mutagenesis (specific mutations are targeted to occur in pre-determined genes; mutant phenotypes are produced after confirming a desired change in the targeted gene). However, many animal disease models have been created by a quite different route: *random mutagenesis*, where genes are mutated essentially at *random*, often by using chemical mutagens or highly active transposons, as described below. Because the mutation is random, the animal models are not made to order (it is not possible to predict which genes will be affected). However, thousands of mutants can quickly be produced with a range of interesting phenotypes that can be analyzed. As well as being important for understanding gene function, the mutant phenotypes can provide disease models.

Chemical mutagenesis

Ethylnitrosurea (ENU) and ethylmethanesulfonate (EMS) are alkylating agents that are widely used in animal mutagenesis programs (see Figure 21.7A). They induce mutations at random across the genome, some of which adversely affect important functional DNA sequences, causing abnormal gene expression and an altered phenotype. Because the vast majority of the induced mutations are point mutations, there used to be no easy way of identifying the mutations associated with abnormal phenotypes. However, careful examination of mutants can reveal aspects of phenotype reminiscent of a human disease phenotype (in which case the murine homologs of the suggested human disease loci are screened by sequencing).

Extensive mutagenesis programs have been carried out in the two invertebrate models amenable to highly sophisticated genetic analyses, *D. melanogaster* and *C. elegans*. As a result, a wide variety of mutant phenotypes have been uncovered in known homologs of human disease genes, thereby providing disease models—we give some examples below. Although these animals, especially *C. elegans*, are quite simple organisms, ~60% of human disease genes are estimated to have *C. elegans* homologs, and close to 75% of them have homologs in *Drosphila*.

Genome-wide vertebrate mutagenesis began in 1996 with ENU mutagenesis of male zebrafish. Mutation in spermatogonial stem cells led to identification of thousands of developmental phenotypes in the descendants of the mutagenized fish. In the late 1990s, large-scale ENU mutagenesis programs were established as a systematic approach to generating new mouse models of human disease (see Figure 21.7B and see also the section below).

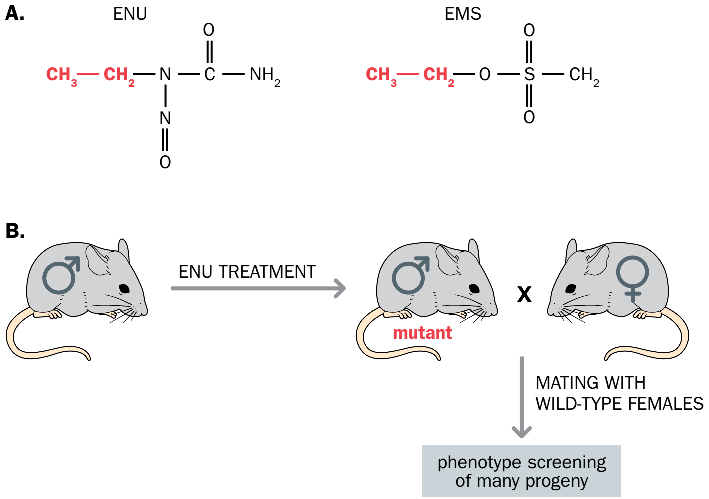


Figure 21.7 Common chemical mutagens used in mutagenesis projects and random genome-wide ENU mutagenesis in the mouse. (**A**) Structures of *N*-ethyl-*N*-nitrosourea (ENU) and ethyl methane sulfonate (EMS). These chemical mutagens act as alkylating agents, causing point mutations by transferring their ethyl groups (in red font) to bases in DNA. In the case of ENU, A-T base pairs are preferentially mutated. (**B**) Genome-wide ENU mutagenesis in the mouse. Male mice are treated with a controlled dose of ENU to produce a moderate rate of random point mutations in spermatogenesis (often the aim will be that at each gene locus, roughly 1 in 1000 sperm carry a new point mutation). Subsequently, mutagenized males are bred with wild-type females over one to two generations to generate mutants.

Transposon mutagenesis

Transposon mutagenesis is a type of random insertional mutagenesis: the method relies on transposons to jump into genes, often causing insertional inactivation. Unlike chemical mutagenesis, therefore, affected genes are tagged by a large insert (a transposon copy) whose sequence is known, greatly aiding identification of genes associated with the observed phenotypes.

DNA transposons have been widely used in germ-line mutagenesis in *D. melanogaster* and *C. elegans*. The 2.9 kb P element of *D. melanogaster*, which moves by a *cut-and-paste* mechanism, is a celebrated example. As well as revolutionizing the study of gene function in the fruit fly, it allowed identification of some phenotypes reminiscent of certain human diseases, thereby providing disease models.

For decades, transposons simply were not available to mutagenize vertebrate genomes in a meaningful way (vertebrate transposons are frequently inactive or have low transposition frequencies). Recently, however, two transposon systems have been developed and used widely for mouse mutagenesis: piggyback, a DNA transposon originally derived from the cabbage looper moth, and *Sleeping Beauty*, a defective fish transposon that was resuscitated by molecular engineering to become an active transposon. Because in each case the transposon is a foreign sequence in the mouse genome, the location of the transposon insertion can be rapidly identified.

Delivering new mouse mutant phenotypes and disease models through large-scale mutagenesis projects and high-throughput phenomics

Various large-scale mouse mutagenesis projects are now delivering many new mutant phenotypes and a wealth of disease models. They include both targeted mutagenesis projects (where individual pre-determined genes are selected to be mutated), and random mutagenesis screens (in which case the screens are driven by the phenotype rather than the genotype).

Genome-wide targeted mutagenesis of mouse genes began with the International Mouse Knockout Consortium (IMKC—see http://www.mousephenotype.org/about-ikmc). In this ambitious project specific mutations were generated in panels of mouse embryonic stem cells (ESCs) for 18,500 mouse genes (representing more than 90% of the protein-coding genes). The major aim was to inactivate individual genes simply to understand their function (after examining any associated mutant phenotype). The mutations generated were therefore selected to be null mutations that would subsequently be introduced into the germ line. Of course, any resulting loss-of-function phenotypes were expected to be of clinical interest because of their potential for generating models of human disease.

Most of the ESC mutations generated by the IMKC were *conditional mutations*, allowing them to be induced in a way that was advantageous (helping to avoid early embryonic lethality for universally important genes, for example). The resulting mutant ESCs then provided the starting material for the International Mouse Phenotyping Consortium (IMPC) to introduce the mutations into the germ line, and thereby generate knockout mouse strains (see http://www.mousephenotype.org). By early 2018, over 7000 mutant lines had been created, and the initial returns have been impressive. About 90% of the gene-phenotype annotations had not previously been reported. Of the first 3,238 genes analyzed, 889 known human disease genes were found to have an orthologous IMPC strain. And of the IMPC knockout lines showing phenotypes overlapping with diseases at those 889 human disease loci, 78% provided the first mouse models for these diseases.

The alternative approach has been to use large-scale random mutagenesis and phenotype-driven screens. For example, a recent large-scale ENU mutagenesis program sought to identify models of age-related disease. Using a high-throughput automated phenotype-detection strategy, pedigrees of the mutagenized mice were screened for mutant phenotypes at various time points as the mice aged. 105 distinct mutant lines were identified from 157 pedigrees analyzed, out of which 27 were late-onset phenotypes. Whole genome sequencing was used to identify genes underlying the phenotypes of 44 mutant strains, including 12 late-onset phenotypes and a novel mouse model of age-related deafness was produced (for more details, see Potter *et al.*, [2016] PMID 27534441, in Further Reading).

Standardized phenotyping and high-throughput mouse phenomics

Following large-scale mouse mutagenesis screens, comprehensive studies of the phenotype are carried out, in which data from a very wide range of physiological systems and processes in the mutagenized mice are systematically recorded (**phenomics**). That is in marked contrast with previous small-scale mouse phenotyping analyses where quite limited aspects of mutant phenotypes would typically be investigated (simply because the researchers would primarily be interested in specific areas, such as cardiac or kidney abnormalities, for example). Often, however, a gene has *pleiotropic* effects, that is it can have multiple functions with differing effects in different organ systems or time points during development. By carrying out a battery of extensive and standardized tests at different time points, phenomics is able to offer the prospect of hypothesis-free phenotyping and the ability to capture data on the pleiotropic effects of individual mutations.

When large-scale mouse mutagenesis projects were devised, standardization of mouse phenotyping tests became a priority, and the IMPC is following standardized procedures known as IMPReSS (**I**nternational **M**ouse **P**henotyping **Re**source of **S**tandardised Screens—see https://www.mousephenotype.org/impress). Dedicated mouse clinics have been established at prominent large research centers in many countries, and standardization of phenotyping procedures facilitates collaboration with other researchers, data reproducibility, and aggregation of data from collaborating centers. In the future, statistical machine-learning methods can be expected to integrate the multiple mouse phenotype data streams, and there will be increasing integration of massive mouse genotype–phenotype data sets with equivalent data sets from large human studies.

21.4 HOW USEFUL ARE ANIMAL MODELS OF GENETIC DISORDERS?

A very wide range of animal species are being used as disease models. Take Duchenne muscular dystrophy (DMD) as an example. In addition to various *Drosophila, C. elegans*, and zebrafish models, a total of 90 mammalian models are listed in a 2015 review in *Disease Models and Mechanisms*: 61 mouse models, 22 dog models, 3 pig models, 2 rat models, and 2 cat models. A rhesus monkey model has subsequently been reported. The majority of these models were artificially created, but the *mdx* mouse, the most widely studied DMD model, is a spontaneous mutant, as are both cat models and essentially all the dog models (which comprise a wide variety of breeds).

No animal model is perfect and different species offer particular advantages and disadvantages. In this section we take a look at how useful different animal species have been in modeling genetic disorders, beginning with nonmammalian models, then moving on to consider rodent models, and then larger nonprimate mammals. Finally, we consider primate disease models, and conclude by examining the increasing interest in human disease models. As we progress through different groups of species we examine the utility of the disease models with reference to the three major uses of animal disease models, as listed below.

- ***Pre-clinical models for testing novel therapies***. The primary requirement is that the chosen animal model should replicate the clinical phenotype as faithfully as possible so that it can be used for testing novel drugs or therapeutic strategies in advance of clinical trials. Ideally, it should have the same genetic basis as the human disorder, mirror the hallmarks and progression of the human pathology, and have a robust and reproducible phenotype. Animals phylogenetically closely related to humans can be expected to be the best pre-clinical models.
- ***Models of molecular/cellular pathogenesis***. If the disease process involves a very highly conserved pathway, the range of animal models useful for this purpose can be very broad, extending to genetically tractable animals phylogenetically distant from humans.
- ***Models for drug screening and toxicity testing***. The object is to assess the likely efficacy of drugs and possible toxicity effects. Depending on whether the major pathways affected by disease have been highly conserved in evolution, models phylogenetically distant from humans may be used as models for drug screening, but mammalian models are preferred for toxicity testing.

The advantages of genetically tractable nonmammalian models of genetic disorders

Outside of mammalian models, three species have been widely used in modeling genetic disorders: the zebrafish, and perhaps more surprisingly, two invertebrate models—the fruit fly *D. melanogaster* and the nematode *C. elegans*. All three species are amenable to highly sophisticated genetic analyses, and extensive mutagenesis screens have been undertaken. They are too phylogenetically distant from humans to serve as good pre-clinical models but have been used to explore highly-conserved molecular pathways in disease and for drug screening.

Zebrafish as disease models

71% of human genes (and 82% of human disease genes listed in OMIM) have a homolog in zebrafish. Zebrafish models of genetic disorders have often been obtained as a product of mutagenesis screens (large-scale genetic screens are less technically challenging than in the case of mice, and a pair of fish can produce hundreds of embryos each week that develop rapidly outside of the mother and can be easily visualized and manipulated experimentally). Identifying developmental mutants is facilitated by the optical transparency of the embryos, leading to identification of mutants with various types of abnormalities. Many additional disease models have been generated by RNA silencing, typically using morpholino antisense oligonucleotides (as detailed in Section 8.5).

In addition to supplying various models of cardiovascular disease, zebrafish have become popular models for studying kidney disease (assisted by the anatomical simplicity of zebrafish kidneys) and eye disorders (the morphology, physiology, and function of the zebrafish eye is similar to that of the human eye). A variety of models of neurodegenerative disorders have also been obtained, including models of Parkinson, Alzheimer and Huntington disease that may help elucidate molecular pathways involved in the pathogenesis of these disorders. (Although Parkinson and Alzheimer disease are complex diseases, both have Mendelian subsets where a single identified gene has been shown to have a major effect). Zebrafish models have also been used for functional analysis of disease-associated variants, as in the case of ciliopathies (see **Box 17.3**). Different types of tumor have also been modeled in zebrafish, such as a model of melanoma generated by overexpressing a mutant human *BRAF* allele in zebrafish embryos. And drug screening is facilitated by being able to expose mutant larvae to individual drugs deposited in wells of microtiter dishes.

Invertebrate disease models

Invertebrates are evolutionarily distant to humans. They lack some organs and vertebrate adaptations, such as a central nervous system, an adaptive immune system, and skeletal muscle. Two invertebrate animals—*D. melanogaster* and *C. elegans*—have nevertheless been widely used as disease models because they offer sophisticated genetic analyses. Despite the roughly 800 million years of evolution that have elapsed since the lineage leading to modern vertebrates separated from those leading to fruit flies and nematodes, nearly 75% and ~60% of human disease genes have homologs in *Drosophila* and *C. elegans*, respectively. With only 959 or 1031 cells, depending on the sex, *C. elegans* is a very simple organism, but *Drosophila* is much more complex with a brain that has more than 100,000 neurons in adults.

Various high-throughput genetic screens can be carried out rapidly and inexpensively in *Drosophila* and *C. elegans*, and the sophisticated genetics of these organisms allows rapid analyses of disease pathways that have highly-conserved components (see **Table 21.3** for examples). The analyses can lead to the identification of novel modifier genes that regulate molecular components of the disease pathway and can influence the phenotype (see legend to **Table 21.3**).

TABLE 21.3 SOME DISEASE MODELING APPLICATIONS FOR THE TWO MOST WIDELY USED INVERTEBRATE MODELS, *C. ELEGANS* AND *D. MELANOGASTER*

Modeling molecular pathways in	Examples/comments
Neurodegenerative disorders	Alzheimer, Parkinson and polyglutamine diseases (notably Huntington disease)
Metabolic diseases	*Drosophila* models of types I and II diabetes, metabolic syndrome and others; *C. elegans* was the first animal where viable and fertile mutants were obtained with knockout alleles of genes encoding OGT and OGA, highly-conserved enzymes working in *O*-GlcNAc signaling pathways important in metabolic diseases
Muscular dystrophies	Notably Duchenne muscular dystrophy
Aging and cell death	An important area of study in *C. elegans*
Cancers	Various

Note that both models are also important in high-throughput screening for drugs that interact with components of the disease pathway so as to reduce the severity of the phenotype. Molecular investigations can also identify natural *modifier genes* that can influence the severity of the phenotype; as an example, identification of a *C. elegans* modifier gene that regulates aggregation of amyloid-beta in Alzheimer disease and alpha-synuclein in Parkinson disease, led to identification of human homologs with a similar function (see van Ham *et al*. [2010] *Cell* **142**:601–612; PMID 20723760).

Rodents provide the most widely used mammalian models of genetic disorders

Two types of rodent have been especially used as genetic models of disease: the mouse (the premier disease model) and, to a much lesser extent, the rat. In addition to various practical advantages—comparatively short generation times, large numbers of offspring and relatively inexpensive maintenance costs (for a mammalian model)—the mouse offers the opportunity of carrying out sophisticated genetic analyses. Until the arrival of genome editing using programmable nucleases, the mouse offered an unrivaled opportunity of making gene knockouts (using homologous recombination in the favorable 129 strain of embryonic stem cells, as described above). Rats have the advantage of being 10 times larger than mice, and are better suited to physiological analyses than mice.

Mice and rats also the advantage that the genetic background can be controlled (because highly inbred strains are used, enabling consistent phenotypes), and breeding strategies can allow mutant alleles to be placed on different genetic backgrounds. Just as in humans, the phenotypes of single gene disorders can be quite strongly influenced by the genetic background: certain *modifier genes* make products that interact with disease pathway components in ways that can modify the effect of the mutant allele. Placing a mutant allele on a different genetic background may produce better animal models, and after identifying modifier genes by experimental analyses in mice, human orthologs of the modifier genes can be sought, with the prospect of obtaining additional therapeutic targets. **Box 21.4** gives an overview of mouse genetic backgrounds and modifier genes.

BOX 21.4 MOUSE GENETIC BACKGROUNDS AND MODIFIER GENES

The **genetic background** of a mouse describes the genetic constitution (all alleles at all loci) except for the mutated gene of interest and a very small amount of other genetic material (generally from one or two other mouse strains). The genetic background is important because it can influence the phenotype of a mutant allele in different ways.

Inbred strains are produced from brother–sister matings for at least 20 successive generations. Because such strains are essentially homozygous at all loci, some may not breed so well and show some abnormalities (as shown by the 129 strain that has traditionally been used to make embryonic stem cells); others, such as the C57BL/6 strain are quite robust. A mutant allele is generally maintained on a background that displays a strong phenotype and breeds well, and knockout lines produced using ESCs from the 129 strain have often been back-crossed with a robust strain, such as C57BL/6, to increase fertility and assist phenotype analysis.

Congenic strains can be produced by back-crossing to a parental inbred strain for at least ten generations, while selecting for a specific marker from the donor strain. The resulting mouse contains a small genetic region (ideally containing a single gene of interest) from the donor strain but is otherwise identical to the original inbred strain.

MODIFIER GENES

When identical mutations are placed on different genetic backgrounds, large differences may be seen in the mutant phenotype, because the different strains can have different alleles at one or more number of modifier loci. A useful example comes from the *Min* (multiple intestinal neoplasia) mouse that has a mutant *Apc* allele. Mutations in the human ortholog, *APC*, cause adenomatous polyposis coli and related colon cancers (OMIM 175100), and the *Min* mouse has been regarded as a good model for such disorders.

The *Min* phenotype, however, is strongly influenced by the genetic background. For example, the number of colonic polyps in mice carrying the *Apc*^{Min} mutation is strikingly dependent on the mouse strain: B6 mice heterozygous for the *Apc*^{Min} mutation are very susceptible to developing intestinal polyps, but offspring of these mice mated with AKR/J, MA/MyJ, or CAST strains are significantly less susceptible. Various modifier genes affect the *Min* phenotype in mice, notably *Mom1* (modifier of Min1) which was subsequently identified as the mouse *Pla2g2a* (phospholipase A2, group IIA) gene. *Pla2g2a* alleles making the normal protein product inhibit the development of polyps, but in the B6 mouse strain *Pla2g2a* has a natural inactivating mutation. (Similar phenotypic variability is found in human adenomatous polyposis coli families [different members of a family with identical *APC* mutations may have strikingly different tumor phenotypes], but although the human counterpart of *Pla2g2a* is overexpressed in the human colorectal adenomas, it does not seem to influence the number of polyps.)

A plethora of mouse models

Recall from the start of Section 21.4 that a 2015 survey of animal models of Duchenne muscular dystrophy (DMD) identified 61 mouse models and 2 rat models. Because genome editing (via homologous recombination) used to be so much easier in mouse than in any other mammal, one would expect that there would be more mouse models than rat models, but one might reasonably ask why so many different mouse DMD models have been produced. Part of the explanation is that different strategies were used to generate the models: random mutagenesis (using chemical mutagens or insertion vectors); targeted mutagenesis (introducing specific inactivating mutations to create dystrophin gene knockouts); double-knockouts (knocking out both the dystrophin gene, *Dmd*, and an interacting gene); transgenesis (overexpression of dystrophin or interacting transgenes); placing mutant alleles on different genetic backgrounds (by breeding of mutants with other genetic strains); and construction of immune-deficient dystrophin-deficient mice. See **Table 21.4** for a summary.

Another partial explanation for the number of mouse models is simply the desire to obtain better models. The well-studied *mdx* muscular dystrophy mouse has a quite mild phenotype, and because of the importance of genetic background (**Box 21.4**), the *mdx* mutation was placed on different genetic backgrounds in an attempt to obtain a more severe phenotype, as observed in the case of the DBA2 background. Similarly, the incentive for producing some double-knockouts was to produce more severe phenotypes. For example, the mouse *Utrn* gene makes utrophin, a protein related to dystrophin that can partially compensate for the absence of dystrophin. Knocking out the genes for both dystrophin and utrophin might be expected to produce a severe phenotype.

The number of mouse models in general is now set to escalate as a result of the International Knockout Mouse Consortium and the International Mouse Phenotype

TABLE 21.4 THE DIVERSITY OF MOUSE DUCHENNE MUSCULAR DYSTROPHY MODELS

Category of mouse model by origin	Number	Examples/comments
Spontaneous mutant	1	The *mdx* mouse, with a point mutation in exon 23 of the *Dmd* gene, originally on a C57BL/10 genetic background
Mutant allele placed on different genetic background	6	*Dmd*^{mdx} allele transferred to Albino, BALB/c, C57BL/10, C3H, DBA2, or FVB background
Random mutagenesis (ENU or insertional mutation)	7	Alleles resulting from random mutations causing dystrophin gene inactivation
Targeted inactivation of the mouse dystrophin gene)	3	Mutants have whole gene deletion (using Cre-*loxP*-based genome editing), or an inactivating exon deletion/duplication
Double-knockout (dystrophin gene plus another relevant gene)	27	13 with severe dystrophic phenotype; 3 with milder phenotype, similar to the mdx mouse; 11 with reduced disease phenotype
Transgenics	13	Overexpression of full length dystrophin or shorter isoforms, or of interacting proteins in *mdx* or other backgrounds
Immune-deficient dystrophin-deficient mice	4	The mice were intended to be recipients of transplants of allogeneic/xenogeneic satellite cells, multipotent precursors of skeletal muscle cells (in order to test a type of cell therapy)

(Data abstracted from McGreevy JW *et al*. [2015] *Disease Models Mech* **8**:195–213; PMID 25740330.) ENU, ethylnitrosurea.

Consortium (IMPC) (detailed in Section 9.4). By September 2018 the IMPC website at http://www.mousephenotype.org/ had listed 7466 human diseases associated with IMPC mouse models. In addition to knockout models, high-throughput phenotype screens have been carried out for different types of phenotypes, such as age-related disorders. See Dickinson *et al*. (2016, PMID 27626380) and Potter *et al*. (2016, PMID 27534441) in Further Reading for examples of recent outputs.

The limitations of rodent and other mammalian models and why modeling some single gene disorders has been challenging

Although the mouse has been the premier animal disease model, mice and other rodents can never be great models for some diseases. They have less developed brains and limited cognitive abilities, for example, and so are not well qualified to be models for disorders that affect the brain, such as neurologic and neuropsychiatric disorders. And the short life of the mouse, in particular, means that it may be disadvantaged as a model for some age-related diseases. Because of their small size, mice, and rats to a lesser extent, are also not as well suited to physiological analyses as larger mammalian models.

There are many other important human–rodent differences, too, that we describe below, and while mouse phenotypes depend a great deal on genetic background, the very much larger differences in genetic background between humans and mice will have much greater effects. In one project, for example, where knockout mouse mutations were created in 37 known homologs of genes underlying human recessive disorders, 17 of the homozygous mouse knockouts were lethal (possibly the human phenotypes are caused by mutations that are not as disruptive as the mouse knockout alleles—see White *et al*. [2013, PMID 23870131] in Further Reading). And when modeling dominant disorders arising from haploinsufficiency, heterozygous mouse mutants quite often show very little or no evidence of a mutant phenotype while the homozygous mutants are often embryonic lethals. In addition to full knockout mutations, therefore, there may often be a need for other types of inactivating mutation that permit a small amount of residual gene function.

Faced with difficulties in modeling disorders in rodents, other mammalian models have subsequently been made. Pigs, dogs, and sheep are phylogenetically closer to humans and have larger brains. They are quite long-lived (compared to rodents), and because of their larger sizes, they are more amenable to physiological analyses (pigs are especially physiologically close to humans). They also make more appropriate preclinical models, both on the basis of their larger size and because unlike mice, they show the same type of immune responses to gene therapy vectors as humans. Miniature pig breeds ("mini pigs") have the advantage of cheaper maintenance costs and being easier to handle.

Challenges in modeling single gene disorders

Single gene disorders might be expected to be the easiest genetic disorders to model. In the example of DMD described above, certain double-knockout mice and a rat have phenotypes comparable to human DMD, as does a pig model (but with rapid disease progression) and a golden retriever dog model. However, disease modeling for some other single gene disorders has been challenging. Here we give two examples: cystic fibrosis (where rodent models have been disappointing, largely because of a human–mouse difference in chloride ion channels), and Huntington disease (which illustrates the intrinsic difficulty in modeling neurologic disorders in animals).

- ***Cystic fibrosis (CF) models.*** Mouse models of CF with inactivating mutations in *Cftr* (cystic fibrosis transmembrane regulator gene) have not been good models. Although they can display severe intestinal obstruction, similar to that seen in CF patients, none of them develop the spontaneous lung inflammation seen in CF patients, limiting their usefulness in studying progression of lung disease in CF. Additionally, most mouse models display only mild complications in the pancreas, liver, and vas deferens, unlike the associated severe complications seen in CF patients. The species differences are largely due to differences in chloride ion channels: mice can express a second Cl^- ion channel (with an alternative signaling pathway) that may compensate for the defective CFTR protein. This was an unexpected difference but is one of many human–mouse differences (see **Box 21.5** for an overview). As a result, other animal models have been needed. Both pig and ferret models of CF replicate the CF phenotype more accurately than do mouse models, but their general utility is hampered by severity of the phenotype (see **Table 21.5**).
- ***Huntington disease (HD) models.*** Many animal HD models have been made, notably rodent models, as described in Section 21.3. Various neuropathological aspects of the phenotype can be replicated at the molecular and cellular levels, but replicating the clinical phenotype is intrinsically difficult in animals because of major human–animal brain differences. Like HD patients, the animal models do show some learning and memory deficits but reproducing specific features of HD-associated motor dysfunction—such as chorea (the involuntary dance-like movements) and the motor abnormalities leading to dysarthria (speech abnormalities)—has been difficult in tetrapod, speechless animal models. And modeling psychiatric disturbance—anxiety, irritability, impulsivity, aggression, apathy, depressed mood, delusions, hallucinations, obsessions, and compulsions—has been challenging (but some rodent models appear to show depression-like and anxiety-like behavior). Various large animal HD models—pig, sheep, and rhesus monkey—have been constructed, but none of them appear to be useful disease models (there is little evidence that any of them mirrors the human phenotype, and first-generation HD monkeys developed an aggressive disease phenotype causing them to die after just a few months).

BOX 21.5 WHY HUMAN PHENOTYPES CAN BE DIFFICULT TO REPLICATE IN MICE

Although the mouse is the premier disease model, it is disadvantaged by important human–mouse differences, some of which we list below.

- ***Gene catalogs.*** For ~14,000 of human and mouse genes there is a 1:1 orthologous relationship, but about 2,000 human and mouse protein-coding genes resulted from duplication since the human and mouse lineages diverged from a common ancestor ~90 million years ago, resulting in copy number changes and functional divergence. In total, over 4,000 human and mouse protein-coding genes lack evidence of shared ancestry. More than 20% of essential human genes have nonessential mouse orthologs. Some genes sufficiently important in humans to be a disease locus have no counterpart in rodents, such as *SHOX*, a locus for two human genetic disorders (OMIM §127300, §249700), and *KAL*, the locus for Kallman syndrome type I (OMIM §308700).
- ***Genetic background.*** Human populations are mostly outbred; laboratory mice are highly inbred. When making mouse models, the mutant phenotype can vary very significantly according to the genetic background (see **Box 21.4**).
- ***Gene regulation.*** During evolution, regulatory elements diverge more rapidly than coding sequences (Section 13.4). Even if human and mouse orthologs appear highly conserved, divergent regulatory elements may result in divergent expression of orthologs and differential interaction with other gene products.
- ***Biochemical and metabolic pathways.*** The metabolic rate of mice is ~7 times that of humans and there are significant human–mouse differences in various biochemical pathways, such as in drug metabolism, inflammatory responses, and so on. Important human–mouse differences in the expression of glycosylating enzymes can result in divergent protein distribution and function.
- ***Cell physiology and developmental pathways.*** For example, melanocytes are found in the outermost layer of human skin, but in mice, they are confined to hair follicles, making it difficult to model disorders such

as melanoma by simply exposing mice to excess UV irradiation.

- ***Brain organization and capacity.*** The size and complexity of the human brain far exceeds that of the mouse and our unique cognitive capacities mean that aspects of brain function and brain disease can never be accurately modeled in rodents.
- ***Longevity.*** Humans live ~30 times as long as mice. The pathogenesis of some human disorders that manifest in middle to old age could involve temporal components, making it difficult to replicate the phenotype in short-lived mice.
- ***Cell division and telomere regulation.*** About 3000 times larger than mice, humans have correspondingly more cells, and in an average lifetime there are ~10^{16} human mitoses but ~10^{11} mouse mitoses, with much the same incidences of developing cancer. As cells divide, telomeres get progressively shorter. Telomere erosion has been implicated in cell senescence and cancer, but there are major human–mouse differences in telomere length (4–6 × longer in mouse) and telomerase (functionally active in most mouse cells, but scarcely detectable in most somatic cells in adult humans). Possibly, some distinct cancer containment mechanisms are intrinsic to human cells, partly explaining why it is difficult to model human cancers in the mouse.

HUMANIZED MICE

In some areas, the biology of humans and mice are too divergent to accurately reproduce aspects of human disease with standard mouse models. Some human–mouse differences can be overcome, nevertheless, by making *humanized mice*. Thus, modeling susceptibility to viral infectious disease can often be difficult because of human–mouse differences in cell-surface receptors. Most cases of the common cold in humans, for example, are due to rhinoviruses that bind to the ICAM-1 receptor, but mice have a substantially different ICAM-1 receptor and are resistant to infection. Genetic engineering of mice to give them a modified ICAM-1 receptor, more similar to the human one, allowed production of the first mice able to catch the common cold. In addition to studying infectious disease, other applications of humanized mouse models include studying autoimmune disorders and aneuploidies (the transchromosomic mouse model of trisomy 21 described in **Box 21.3** is a type of humanized mouse). We also describe the use of humanized mice in Chapter 22 when we consider animal models in systems for producing therapeutic antibodies and for testing drug treatment (useful because of human–mouse differences in drug metabolism).

TABLE 21.5 PIG AND FERRET DISEASE MODELS MORE ACCURATELY REPLICATE THE CLINICAL FEATURES OF CYSTIC FIBROSIS THAN DO MOUSE MODELS

Species	Spontaneous lung infection	Pancreatic disease	Intestinal disease	Liver and gall bladder disease	Reproduction
Human	Yes	Pancreatic insufficiency	15% of infants have meconium ileus**	Biliary cirrhosis	Severe vas deferens defect
Mouse*	No	No	Intestinal obstruction, often fatal	No	Reduced fertility in females
Pig	Yes	Pancreatic insufficiency	100% of piglets have meconium ileus	Biliary cirrhosis	Severe vas deferens defect
Ferret	Yes	Pancreatic insufficiency	75% have meconium ileus	Liver disease	Severe vas deferens defect

*The phenotypes of different mouse models can show some differences. ** Unlike later feces, the meconium, the earliest stool of a mammalian infant, is normally viscous and sticky, but in meconium ileus the stool becomes thickened and congested in the intestines. The utility of the pig and ferret models is hampered by the presentation of meconium ileus in 100% of piglets (which is fatal without early surgical intervention), and because the ferret models rarely reach adolescence due to the severity of the phenotype. (Reproduced from Lavelle GM *et al*. [2016] PMID 27340661, see Further Reading.)

The promise and difficulties of modeling disease in primates

There are many disadvantages to using nonhuman primates as disease models. They are very expensive to buy and maintain, are long-lived, and produce limited numbers of offspring. The experimental analyses take a long time and are costly, and the data are inevitably limited by having small numbers of animals. Unlike mice and rats, they are not inbred, and so phenotypes are less consistent. And experimenting on primates raises very significant ethical concerns.

Despite the very significant ethical concerns raised by research on nonhuman primates, there has been increasing interest in primate disease models. Being phylogenetically very closely related to humans, nonhuman primates might be expected to offer the most accurate pre-clinical models and the best models for human brain disorders. Small primates—macaques and marmosets—have been preferred because of comparatively short generation times, early sexual maturity (18 months in the case of marmosets), comparatively cheap maintenance costs, and easier handling (but these animals are much more expensive and more complicated to care for than rodents). Primates are not inbred, however, and so there can be significant phenotype variation unlike when using specific inbred strains of mice and rats.

The construction of primate disease models using genetic methods is largely a recent development: although transgenesis has been available for some time, refined CRISPR-Cas germline genome editing has recently made it possible to alter germ-line DNA with high efficiency and high precision at pre-determined sites. Although some models have been produced that replicate aspects of human disease phenotypes not often seen in mouse models, it remains to be seen just how useful nonhuman primates will be as disease models.

Humans—the ultimate model organism?

Humans are set to become much more thoroughly investigated. Human phenotypes are already the most intensively studied on the planet—a continuous and global screen of abnormal function and disease is maintained by individuals, families, and health professionals. As we move further into the age of population genomics, the number of individual human genome sequences obtained will rapidly escalate—some predictions suggest that by the early to mid 2020s personal genome sequencing will have become so routine that the majority of individuals in many regions of the planet will have had their genomes decoded.

As a result of a dramatic advances in high-throughput DNA sequencing, the scale of correlating human genetic variation with human phenotypes will rapidly expand. At the moment, healthcare systems carry out many different types of tests on individual humans throughout the world, but the documentation of the resulting phenotype data is currently haphazard. Increasing use of computerized records and electronic networks will mean easier access to, and integration of, human phenotype analyses. Eventually, we will be able to comprehensively correlate genome data with detailed personal phenotypes and lifestyle information.

Concluding remarks

No animal species will provide perfect models of human disease, and we may come to rely more and more on correlating human DNA variants with clinical phenotypes. Modeling single gene disorders in rodents has often led to imperfect models, but the ease with which CRISPR-Cas genome editing can be applied to other animals may now lead to a strong upsurge in various large-animal models, such as pigs and sheep (where costs of keeping the animals and ethical objections are significantly less than for nonhuman primates).

As for modeling brain disorders and complex disease in animals, that is going to remain very challenging. The use of common mammalian models to model brain disorders has resulted in inadequate models. The modeling of neuropsychiatric disorders is especially challenging given the subjective nature of many of the symptoms, plus the lack of objective diagnostic tests and specific biomarkers. Reasonable models for some complex diseases do exist—the NOD (nonobese diabetes) mouse model, for example, serves as a satisfactory model of type I diabetes (in which an autoimmune mechanism causes pancreatic beta cell destruction). But going ahead to use genetic methods to make models of complex disease is generally going to be limited because only a proportion of the disease-causing genetic variation may be known, and there can be significant environmental components. (At least in the case of diseased cells derived from induced patient-specific pluripotent stem cells, all the genetic variants predisposing to disease should be present.)

SUMMARY

- Model organisms have long been used for understanding facets of basic biology and evolution, and for applied research.
- Microbes and simple invertebrates are important for understanding the most highly conserved functions of cells and animals and for understanding microbial and animal pathogens.
- Genetically tractable invertebrates such as *C. elegans* and *D. melanogaster* are important for understanding fundamental features of development and nervous systems. They can also serve as disease models and in drug screening (by testing to see if individual drugs can reduce the severity of mutant phenotypes).
- Vertebrate and mammalian model organisms are important for providing insights into aspects of cell and developmental biology, for inferring evolutionary relationships, and for modeling disease.
- Disease can be modeled *in vitro* using cell cultures, (which offer rapid, mostly molecular analyses of human disease cells), as well as *in vivo* using animal disease models (which offer extensive analyses of the whole organism).

- Traditional cellular disease models are based on 2D monolayers of cultured cells derived from diseased tissue/tumors. They are disadvantaged in two ways: some types of disease cells are not readily accessible as biopsies; and 2D cultures are not good representations of the 3D environment of cells within organisms.
- Induced pluripotent stem cells prepared from skin fibroblasts cells of patients can be directed to differentiate to give patient-specific diseased cells, including cells that are difficult to obtain routinely from biopsies.
- 3D cell cultures organized by stem cells, notably pluripotent stem cells, can give rise to a wide variety of organoids that can resemble the organization of cells within living tissue.
- Some animal disease models arise through spontaneous mutation, but most are artificially produced, usually using mutagenesis to modify endogenous genes in germline DNA or transgenesis, in which an exogenous gene (or genes) introduced in a transgene or artificial chromosome are expressed.
- Mutagenesis may involve targeting DNA changes to specific sequences of interest (by gene editing) to create a desired animal disease model. Alternatively, random mutation (through chemical mutagens, X-rays) is applied, in which case the phenotypes produced are unpredictable, and large numbers of progeny are screened for mutant phenotypes of interest.
- Disease resulting from DNA changes causing gain of function can be modeled by making transgenic animals with mutant transgenes, or by overexpressing a gene.
- Disease resulting from DNA changes causing loss of function can be modeled by making a knockout. An inactivating mutation is introduced into the gene, such as by deleting a small early exon to cause a frameshift in the translational reading frame, or by creating a large deletion. Heterozygote animals can be bred to produce homozygotes to model recessive disorders. Alternatively, gene silencing (using RNA interference or antisense oligonucleotides) is used to specifically inhibit the expression of alleles at the disease gene locus.
- Animal disease models are used for three purposes: as preclinical models (allowing testing of drugs or other novel therapeutic strategies in advance of clinical trials); as models to understand the molecular basis of disease; and for drug screening.
- For pre-clinical models, faithfully reproducing the clinical phenotype is a priority and animals phylogenetically very close to humans might be expected to be the best models. But for understanding the molecular basis of disease and for drug screening, phylogenetically distant animals are also often used because they can confer some practical advantages.
- None of the many animal species used as disease models is ideal. Despite some important human–mouse differences, the mouse has been the premier disease model because it offers various practical advantages. Some larger animal models offer easier physiological analyses or are more suitable for investigating normal and abnormal brain functions.
- Humans are the ultimate disease models, and as the genomes of more and more patients are sequenced, human disease models are expected to increase in importance as increasingly massive data sets are established linking human genotypes to human phenotypes.

FURTHER READING

Model organisms: general reviews and databases

Beck CW, Slack JMW (2001) An amphibian with ambition: a new role for *Xenopus* in the 21st century. *Genome Biology* **2** Reviews:1029.1–1029.5; PMID 11597339.

Beckingham KM *et al.* (2005) *Drosophila melanogaster*—the model organism of choice for the complex biology of multi-cellular organisms. *Gravit Space Biol Bull* **18**:17–29; PMID 16038090.

Carlsson HE *et al.* (2004) Use of primates in research: a global view. *Am J Primatol* **63**:225–237; PMID 15300710.

Davis RH (2004) The age of model organisms. *Nature Rev Genet* **5**:69–76; PMID 14708017.

Model Organism Databases portal at the National Human Genome Research Institute (NHGRI). At: https://www.genome.gov/10001837/model-organism-databases/. (Offers quick access to databases for principal models.)

Stern CD (2005) The chick: a great model system becomes even greater. *Developmental Cell* **8**:9–17; PMID 15621526.

WormBook. The online review of *C. elegans* biology at http://www.wormbook.org.

Disease modeling using pluripotent stem cells and organoid cultures

Avior Y *et al.* (2016) Pluripotent stem cells in disease modeling and drug discovery. *Nature Rev Mol Cell Biol* **17**:170–182; PMID 26818440.

Bellin M *et al.* (2012) Induced pluripotent stem cells: the new patient? *Nature Rev Mol Cell Biol* **13**:713–726; PMID 23034453.

Clevers H (2016) Modeling development and disease with organoids. *Cell* **165**:1586–1597; PMID 27315476.

Fatehullah A *et al.* (2016) Organoids as an in vitro model of human development and disease. *Nature Cell Biol* **18**: 246–254; PMID 26911908.

Lancaster MA, Knoblich JA (2014) Organogenesis in a dish: modeling development and disease using organoid technologies. *Science* **345**:1247125; PMID 25035496.

Sterneckert JL *et al.* (2014) Investigating human disease using stem cell models. *Nature Rev Genet* **15**:625–639; PMID 25069490.

Animal disease models (general)

Altman TJ *et al.* (2011) The future of model organisms in human disease research. *Nature Rev Genet* **12**:575–582; PMID 21765459.

Spradling A *et al.* (2006) New roles for model genetic organisms in understanding and treating human disease. *Genetics* **172**:2025–2032; PMID 16636111.

Technological aspects in making animal models (see also Chapter 8)

Doyle A *et al.* (2012) The construction of transgenic and gene knockout/knockin animal models of disease. *Transgenic Res* **21**:327–349; PMID 21800101.

Justice MJ *et al.* (2011) Technical approaches for mouse models of human disease. *Dis Model Mech* **4**:305–310; PMID 21558063.

Yu Y, Bradley A (2001) Engineering chromosomal rearrangements in mice. *Nature Rev Genet* **2**:780–790; PMID 11584294.

Nonmammalian disease models

Santoriello C, Zon LI (2012) Hooked! Modeling human disease in zebrafish. *J Clin Invest* **122**:2337–2343; PMID 22751109.

Segalat L (2007) Invertebrate animal models of disease as screening tools in drug discovery. *ACS Chem Biol* **2**:231–236; PMID 17455900.

Silverman GA *et al.* (2009) Modeling molecular and cellular aspects of human disease using the nematode *Caenorhabditis elegans*. *Pediatr Res* **65**:10–18; PMID 18852689.

The Interactive Fly: *Drosophila* as a model for human diseases. Available at: https://www.sdbonline.org/sites/fly/modelsystem/aamodelsystem.htm

General reviews on mouse disease modeling and relevant human–mouse differences

Elsea SH & Lucas RE (2003) The mousetrap: what can we learn when the mouse model does not mimic the human disease. *ILAR J* **43**:66–79; PMID 11917158.

Ernst PB, Carvunis A-R (2018) Of mice, men and immunity: a case for evolutionary systems biology. *Nature Immunol* **19**:421–425; PMID 29670240. (Considers human–mouse differences and their consequences from an evolutionary perspective.)

Guenet J-L (2011) Animal models of human genetic diseases: do they need to be faithful to be useful? *Mol Genet Genomics* **286**:1–20; PMID 21547562.

Liao B-Y & Zhang J (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci USA* **105**:6987–6992; PMID 18458337.

Perlman RL (2016) Mouse models of human disease. An evolutionary perspective. *Evol Med Pub Health* **1**:170–176; PMID 27121451.

Rosental N, Brown S (2007) The mouse ascending: perspectives for human-disease models. *Nature Cell Biol* **9**:993–999; PMID 17762889.

Scheer N *et al.* (2013) Generation and utility of genetically humanized mouse models. *Drug Discovery Today* **18**: 1200–1211; PMID 23872278.

Large-scale mouse mutagenesis screens and high-throughput phenomics

Brown SDM *et al.* (2018) High-throughput mouse phenomics for characterizing mammalian gene function. *Nature Rev Genet* **19**:357–370; PMID 29626206. (Also describes the importance of mouse phenomics in disease modeling.)

Dickinson ME *et al.* (2016) High throughput discovery of novel developmental phenotypes. *Nature* **537**:508–514; PMID 27626380. (An example from large-scale mouse knockout output.)

Fuchs H *et al.* (2018) Understanding gene functions and disease mechanisms: phenotyping pipelines in the German Mouse Clinic. *Behavioural Brain Res* **352**:187–196; PMID 28966146.

Potter PK *et al.* (2016) Novel gene function revealed by mouse mutagenesis screens for models of age-related disease. *Nature Comm* **7**:12444; PMID 27534441. (An example of phenotype-driven screening.)

White JK *et al.* (2013) Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154**:452–464; PMID 23870131.

Large nonprimate mammalian models

Prather RS *et al.* (2013) Genetically engineered pig models for human diseases. *Annu Rev Anim Biosci* **1**:203–219; PMID 25387017.

Whitelaw CBA *et al.* (2016) Engineering large animal models of human disease. *J Pathol* **238**:247–256; PMID 26414877.

Primate disease models

Butte AJ (2008) Medicine. The ultimate model organism. *Science* **320**:325–327; PMID 18420921. (Humans as models of disease.)

Chan AWS (2013) Progress and prospects for genetic modification of nonhuman primate models in biomedical research. *ILAR J* **54**:211–223; PMID 24174443.

Garbarini N (2010) Primates as a model for research. *Dis Models Mech* **3**:15–19; PMID 20075377.

Jennings CG *et al.* (2016) Opportunities and challenges in modeling human brain disorders in transgenic primates. *Nature Neurosci* **19**:1123–1130; PMID 27571191.

Animal models of specific diseases

Conn PM (ed.) (2017) *Animal Models for the Study of Human Disease*, 2nd edn. Academic Press. (With 44 chapters, most of them devoted to models for specific diseases or disease categories.)

Lavelle GM *et al.* (2016) Animal models of cystic fibrosis pathology: phenotypic parallels and divergences. *Bio Med Res Intl* Article ID 5258727, 14 pages; PMID 27340661.

Pouladi MA *et al.* (2013) Choosing an animal model for the study of Huntington disease. *Nature Rev Neurosci* **14**: 708–721; PMID 24052178.

McCammon JM, Sive H (2015) Addressing the genetics of human mental health disorders in model organisms. *Annu Rev Genomics Hum Genet* **16**:173–197; PMID 26002061.

McGreevy JW *et al.* (2015) Animal models of Duchenne muscular dystrophy: from basic mechanisms to gene therapy. *Dis Models Mech* **8**:195–213; PMID 25740330.

Sheppard O *et al.* (2012) Mouse models of aneuploidy. *Sci World J* Article ID 214078, 6 pages; PMID 22262951.

22 Genetic approaches to treating disease

Treatment of genetic disease and genetic treatment of disease are two separate matters. The cause of a disease (whether mostly genetic or mostly environmental) and its treatability are quite unconnected. Standard medical treatments that are intended to alleviate disease symptoms—hearing aids or cochlear implants for treating profound deafness, for example—are just as applicable if the disease is mostly genetic or mostly environmental. In this chapter the primary focus is on how genetic technologies are being applied to treat disease, but we begin by taking a broader look at different treatment strategies for genetic disorders. For the great majority of genetic conditions, even for single gene disorders, existing treatments are lacking or unsatisfactory, but important new inroads have been made recently using genetic technologies.

Causative genes for many single gene disorders have been identified quite recently, and many of them are genes that make previously unstudied proteins. In these cases, it typically takes decades of research to fully understand how the underlying genes work normally and to gain deep insights into the molecular pathogenesis of the monogenic disorder (which usually involves studying artificially constructed animal disease models and cellular models of the disease). Armed with that knowledge, there is the hope that novel treatments may be developed. For some monogenic disorders, there are extremely difficult obstacles to devising effective treatments; for others, novel gene therapies have recently been successfully developed—we provide examples below.

Reasonably satisfactory treatments exist for some complex diseases, such as diabetes; for many others the treatments are less than satisfactory, or ineffective. By definition, complex diseases are complex at the genetic level: until very recently we knew very few of the underlying genetic factors, but important ones have been revealed by genetic studies, notably for disorders with an autoimmune basis, and increasingly for cancers. In many cases, genetic studies will be able to divide individual complex diseases into subtypes, allowing different treatments to be tailored to suit different disease subtypes (**stratified medicine**). The emerging information will place us in a better position to develop novel, more effective treatments.

Environmental factors are clearly very important in complex diseases and have been notably well documented in many cancers. Some environmental factors are also well recognized in some noncancer conditions. Cigarette smoking is a powerful factor in age-related macular degeneration and emphysema, for example, and the importance of a healthy diet and regular exercise is well recognized in conditions like type 2 diabetes. Considerable work needs to be done to extend our knowledge of contributing environmental factors. That will provide opportunities for effective interventions because exposure to an environmental factor can often be modified.

In this chapter we will primarily be concerned with molecular approaches to treating disease, and we deal principally with genetic approaches. The majority of the applications are in treating genetic disease (mostly monogenic disorders but also approaches towards treating certain types of complex disease), but we also take a brief look at approaches towards treating some types of infectious disease that depend on conferring resistance to a pathogen. Cancer therapies are covered separately in Chapter 19.

In Section 22.1 we give an overview. First, we look at how treatments can be classified into different categories. We take a broad view of the different levels at which disease can be treated and explore the different genetic technology inputs that can be applied. In Section 22.2 we cover treating disease with genetically-modified therapeutic proteins. In Section 22.3 we cover the principles and general methodology of different therapeutic methods involving genetic modification of a patient's cells or transplantation of cells that have been genetically modified (gene therapy). We describe in Section 22.4 how standard gene therapy for recessive disorders has been applied in clinical trials and assess the progress. Finally, in Section 22.5 we cover newer developments: RNA therapeutic approaches, therapeutic genome editing using programmable nucleases, and some new genetic approaches to preventing disease.

22.1 AN OVERVIEW OF TREATING GENETIC DISEASE AND OF GENETIC TREATMENT OF DISEASE

In this introductory section we first look at broad categories of treating genetic disease. Then we consider the different levels at which molecular-based disease treatments can be applied.

Three different broad approaches to treating genetic disorders

Two types of treatment can be used, according to whether pathogenesis is due to genetic deficiency or to some positively harmful effect, rather than a deficiency. A third type of treatment seeks to reduce susceptibility to disease, based on understanding the pathway involved (**Figure 22.1**). We expand on these themes in the sections below, taking into account both current practice and experimental therapies.

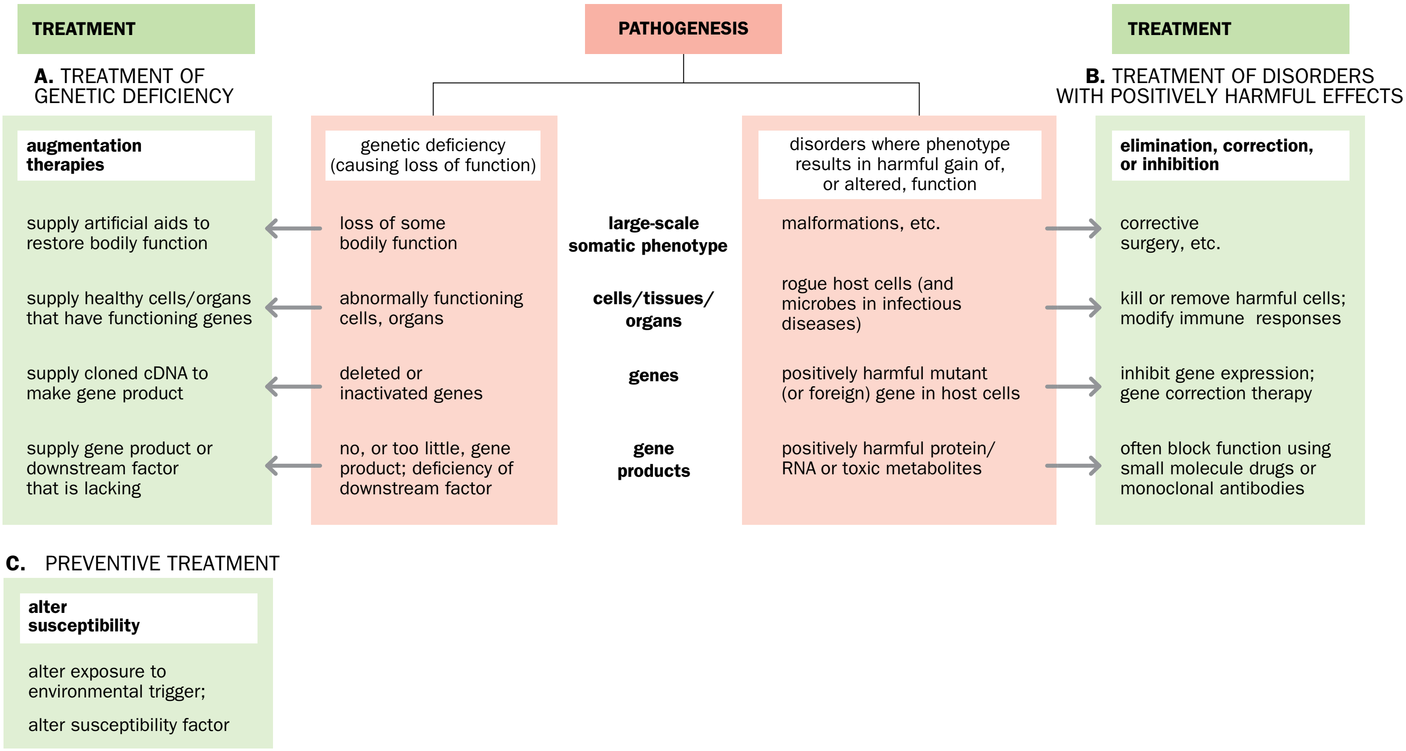

TREATMENT
PATHOGENESIS
TREATMENT
A. TREATMENT OF GENETIC DEFICIENCY
B. TREATMENT OF DISORDERS WITH POSITIVELY HARMFUL EFFECTS
augmentation therapies
genetic deficiency (causing loss of function)
disorders where phenotype results in harmful gain of, or altered, function
elimination, correction, or inhibition
supply artificial aids to restore bodily function
loss of some bodily function
large-scale somatic phenotype
malformations, etc.
corrective surgery, etc.
supply healthy cells/organs that have functioning genes
abnormally functioning cells, organs
cells/tissues/organs
rogue host cells (and microbes in infectious diseases)
kill or remove harmful cells; modify immune responses
supply cloned cDNA to make gene product
deleted or inactivated genes
genes
positively harmful mutant (or foreign) gene in host cells
inhibit gene expression; gene correction therapy
supply gene product or downstream factor that is lacking
no, or too little, gene product; deficiency of downstream factor
gene products
positively harmful protein/RNA or toxic metabolites
often block function using small molecule drugs or monoclonal antibodies
C. PREVENTIVE TREATMENT
alter susceptibility
alter exposure to environmental trigger;
alter susceptibility factor

Figure 22.1 Different major treatment strategies for genetic disorders. Note that some treatment strategies are experimental. (**A**) Augmentation therapies for phenotypes that result from deficiency of some gene product. The idea is to compensate for the genetic deficiency by supplying purified functional gene product directly (protein augmentation) or a purified downstream factor that is required but lacking, or by indirectly supplying cloned DNA or healthy cells (either from a donor or genetically-modified cells from the patient) to make the missing gene product. (**B**) Therapies for phenotypes that result from positively harmful cells or molecules. Brackets indicate applicability of a strategy to infectious disease. Some therapies work at the cell/tissue level to deal with rogue cells that behave abnormally to cause disease (cancer cells, immune system cells that attack host cells in autoimmune and inflammatory diseases). Others work at the gene or gene product level to prevent the harmful effects of a gain-of-function mutation (or a gene from a pathogenic microorganism) or seek to eliminate or reduce production of elevated toxic metabolites in inborn errors of metabolism. (**C**) Disease-prevention strategies include altering exposure to environmental triggers, such as through extreme dietary modifications in some inborn errors of metabolism and the use of drugs, such as statins, to alter disease susceptibility (see text).

Augmentation therapy for genetic deficiencies

In some genetic disorders the problem is loss of some normal function. In principle, these disorders might be treated by **augmentation therapy**: something is provided to the patient that *supplements* a severely depleted, or missing, factor, thereby overcoming the deficiency and restoring function. Different types of supplement can be provided to restore function at different levels. At the level of the somatic phenotype, treatment can be conventional—providing cochlear implants or hearing aids to treat hereditary deafness, for example.

At the molecular level, the phenotype can be restored by providing a purified gene product that is lacking—a missing enzyme, say, in many inborn errors of metabolism. Or when the gene product works in a biological pathway required to synthesize some important downstream factor, such as a lipid hormone, it might be lack of the downstream factor that is treated (by providing purified lipid hormone, in this case). At a higher molecular level, many types of gene augmentation therapy involve transferring a cloned cDNA into the affected tissues of a patient where it can be expressed to make a missing protein.

At the cellular and organ levels, healthy cells and organs can be transplanted into a patient where they make a product that the patient lacks. That can involve transplanting cells from a donor, as in bone marrow transplantation or organ transplantation. More recently, some cellular gene therapies have been used very successfully; here, the cells of the patient are genetically modified so that they can now express the desired gene product.

Applicability of molecular augmentation therapy

Recessive disorders (where both alleles lose their function) are more suited to molecular augmentation therapy than dominant disorders. Affected individuals often cannot make any functional copies of some normal gene product. Even a modest efficiency in delivery (of healthy cells, genes, or proteins) to an affected individual can often allow effective treatment, and there have been recent dramatic breakthroughs. As illustrated below, however, augmentation therapy is currently not practical for some recessive disorders—it can often be difficult to get efficient delivery and production of the desired molecules.

In dominant disorders due to haploinsufficiency, the disease occurs even when one allele is normal and present in all diploid cells. Very efficient delivery and high-level production of the missing gene product would be essential and is currently unavailable for treating single-gene disorders resulting from haploinsufficiency. Augmentation therapy can also be applied to certain complex diseases, however, such as by treating diabetes using purified insulin, or by transplantation of pancreatic islet cells.

Treatment for disorders producing positively harmful effects

A second, different approach to treatment is needed for diseases where the pathogenesis involves a positively harmful effect, rather than a deficiency. Here, augmentation therapy cannot be used: something has gone wrong that cannot be corrected by simply administering some normal gene, normal gene product, or normal cells to the patient. Different methods are needed (**Figure 22.1B**).

The harmful effect might be treatable at the somatic phenotype level, as in the case of some developmental malformations: corrective surgery is highly effective, for example, in treating various complex disorders such as congenital heart defects, cleft lip and palate, and pyloric stenosis.

At the molecular level, treatments can be conducted at different stages. In many inborn errors of metabolism the problem is elevated levels of harmful metabolites that can be tackled in different ways. A more general problem is presented by actively harmful gene products from a mutant gene. Examples include mutant prion proteins and β-amyloid that are liable to form protein aggregates harmful to cells, and also harmful proteins or RNAs formed after unstable expansion of short oligonucleotide repeats. Dangerous mutant gene products may be combatted by using a small-molecule drug or therapeutic monoclonal antibody to selectively bind to the mutant molecule and inhibit its activity.

In some cases, the therapy seeks to selectively inhibit the expression of a harmful gene at the mRNA level by using RNA interference strategies, as described in Section 22.5. (As well as targeting harmful mutant RNAs the same approach can also be targeted to silence the genes of intracellular pathogens in infectious diseases). And, at the gene level, therapeutic gene editing has the potential to reverse a mutation and restore the original sequence, as described below.

At the cellular level, the problem may manifest as harmful cells. Some mutations can induce cells to behave abnormally, proliferating excessively to cause cancers that can be treated by long-standing methods (surgical excision, radiation, and chemotherapy) and by targeted chemical and biological drugs and cancer gene therapies. In some genetic disorders, the problem is excessive immune responses in which certain immune system cells inappropriately attack host cells (in autoimmune disorders, such as rheumatoid arthritis, and in inflammatory diseases, such as Crohn disease). Here, there is the potential to employ therapies that down-regulate immune responses, but in some cancer gene therapies, the exact opposite approach has been taken (upregulating immune responses in an attempt to kill cancer cells).

Treatment by altering disease susceptibility

A third way of treating disease seeks to reduce susceptibility to disease in some way and offers ways of treating certain monogenic disorders and also some complex diseases (**Figure 22.1C**). In some inborn errors of metabolism the blockage at one step in a metabolic pathway can drive alternative pathways that cause build up of toxic metabolites. But that can sometimes be overcome by reducing disease susceptibility, such as by removing an environmental trigger. And in some diseases, key susceptibility factors can be manipulated to reduce the chances that a disease recurs, or to reduce the effects of a progressive disease. We consider some novel genetic approaches aimed at preventing disease in Section 22.5.

Genetic treatment of disease may be conducted at many different levels

Any disease, whether it has a genetic cause or not, is potentially treatable using a range of different procedures that apply genetic manipulations or genetic knowledge in some way (see **Figure 22.2**). Sometimes genetic techniques form part of a treatment regime that also involve conventional small-molecule drugs or vaccines. Pharmacogenetics is concerned with how the actions of drugs and the reactions to them vary according to variation in the patient's genes. Genotyping of individuals might then be used to predict patterns of favorable and adverse responses to specific drug treatments, as described in Section 20.5. Such genotyping may become routine as massively-parallel DNA sequencing permits extensive screening of genes in vast numbers of people.

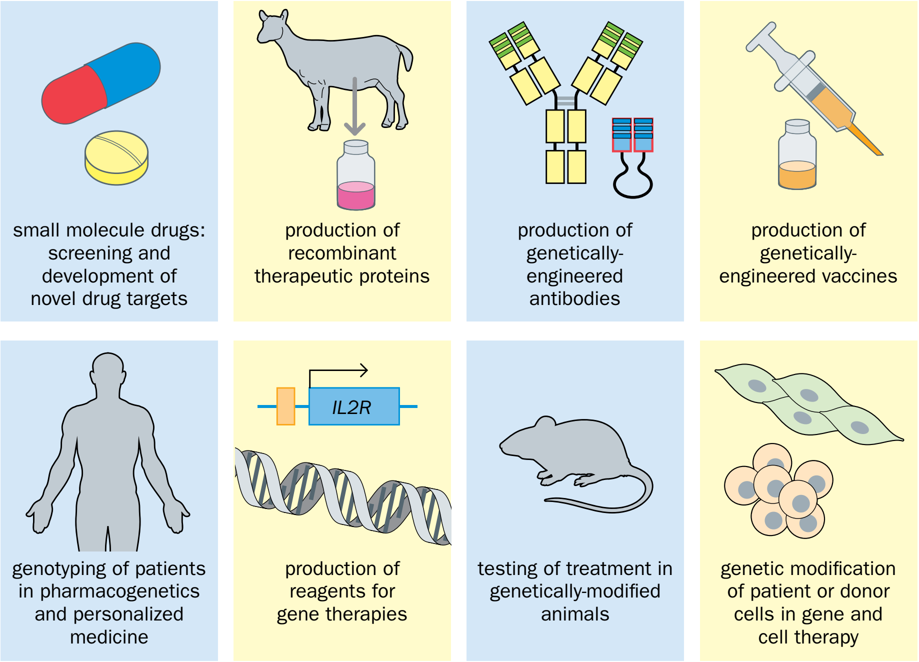

small molecule drugs: screening and development of novel drug targets
production of recombinant therapeutic proteins
production of genetically-engineered antibodies
production of genetically-engineered vaccines
genotyping of patients in pharmacogenetics and personalized medicine
IL2R
production of reagents for gene therapies
testing of treatment in genetically-modified animals
genetic modification of patient or donor cells in gene and cell therapy

Figure 22.2 Some of the many different ways in which genetic technologies are used in the treatment of disease. See text for detail.

New targets for drug development are being identified using knowledge of genetics and cell biology. Genetic techniques can also be used directly in producing drugs and vaccines for treating disease. Another active area concerns treating disease with therapeutic proteins that are produced or modified by genetic engineering. Genes are cloned and expressed in suitable cultured cells or organisms to make large amounts of a specific protein that is then purified (so-called "recombinant" proteins), including hormones, blood factors and enzymes, and especially genetically-engineered antibodies.

Gene therapies are the ultimate genetic application in treating disease and rely on genetically modifying the cells of a patient. That can involve transplanting genetically-modified cells into a patient, or *in situ* delivery of genetic material directly into the cells of the patient. Animal models are particularly important resources for testing new therapies before they are used in clinical trials. As described in Chapter 21, the vast majority of animal models of disease have been generated by genetic manipulation of rodents, notably mice.

22.2 TREATING DISEASE WITH GENETICALLY-ENGINEERED THERAPEUTIC PROTEINS

Chemical treatments for disease are developed by the pharmaceutical industry. Previously, they relied almost exclusively on chemical drugs, hydrocarbon-based small molecules synthesized by standard chemical reactions. More recently, they have been joined by a new class of biological drugs (*biologics*): therapeutic proteins produced using genetic technologies.

Therapeutic recombinant proteins produced by genetic engineering

Certain genetic disorders resulting from deficiency of a specific protein hormone or blood protein can be treated by obtaining and administering an external supply of the missing protein. To ensure greater stability and activity, the proteins are often conjugated with polyethylene glycol (PEG). The increased size of the protein–PEG complex means reduced renal clearance, so that the protein spends more time in the circulation. Adding PEG can also make the protein less immunogenic.

Therapeutic proteins were often previously extracted from animal or human sources, but there have been safety issues. Many hemophiliacs contracted AIDS and/or hepatitis C, after being treated with factor VIII prepared from unscreened donated blood. And some children succumbed to Creutzfeldt–Jakob disease, after having injections of growth hormone extracted from unscreened cadaver pituitary glands.

A safer, but rather expensive, alternative is to use therapeutic **recombinant proteins**. They are produced by cloning human genes and expressing them to make protein, usually within mammalian cells, such as human fibroblasts or the Chinese hamster ovary cell line. (Mammalian cells are often needed because many proteins undergo post-translation modification, such as glycosylsation where the pattern of modification shows differences between species). Recombinant human insulin was first marketed in 1982; **Table 22.1** also gives a number of subsequent examples.

TABLE 22.1 EXAMPLES OF THERAPEUTIC RECOMBINANT PROTEINS

Recombinant protein	For treatment of
Insulin	Diabetes
Growth hormone	Growth hormone deficiency
Blood clotting factor VIII	Hemophilia A
Blood clotting factor IX	Hemophilia B
α-Interferon	Hairy cell leukemia; chronic hepatitis
β-Interferon	Multiple sclerosis
γ-Interferon	Infections in patients with chronic granulomatous disease
Tissue plasminogen activator	Thrombotic disorders
Leptin	Obesity
Erythropoietin	Anemia

For genetically-engineered therapeutic antibodies, see **Table 22.2**.

Some human proteins are required in very high therapeutic doses, beyond the production capabilities of cultured cell lines. Transgenic animals are an alternative source, such as transgenic sheep or goats, where the desired protein is secreted in the animal's milk, aiding purification. In 2009, ATryn became the first therapeutic protein produced by a transgenic animal to be approved by the US Food and Drug Administration. Expressed in the milk of goats, ATryn was designed to be used as an antithrombin in anti-blood clotting therapy.

Genetically-engineered antibodies with improved therapeutic potential

One class of recombinant protein has notably been put to therapeutic use: genetically-engineered antibodies. Each one of us has a huge repertoire of different antibodies that act as a defense system against innumerable foreign antigens. Antibody molecules function as adapters: they have binding sites for foreign antigen at the variable end, and binding sites for effector molecules at the constant end. Binding of an antibody may be sufficient to neutralize some toxins and viruses; more usually, the bound antibody triggers the complement system and cell-mediated killing.

Artificially produced therapeutic antibodies are designed to be monospecific (specific for a single antigen). Traditional monoclonal antibodies (mAbs) are secreted by *hybridomas*, immortalized cells produced by fusing antibody-producing B lymphocytes from an immunized mouse or rat with cells from an immortal mouse B-lymphocyte tumor. Hybridomas are propagated as individual clones, each of which can provide a permanent and stable source of a *single* mAb.

The therapeutic potential of mAbs produced like this is, unfortunately, limited. Rodent mAbs, raised against human pathogens and so on, have a short half-life in human serum (they often cause the recipient to make antirodent antibodies). And only some of the different classes can trigger human effector functions.

Genetically-engineered antibodies

Genetic engineering can modify rodent monoclonal antibodies so that they become more stable in humans: some or all of the rodent protein sequence is replaced by the human equivalent after genetically manipulating coding DNA sequences.

Initially, hybrid human–rodent cDNA sequences were used to generate hybrid antibodies. The first examples were *chimeric* antibodies, having rodent variable chains but human constant regions. Subsequently, *humanized* antibodies were constructed: all the rodent sequence was replaced by human sequence, except for the complementarity-determining regions (CDRs), the hypervariable sequences of the antigen binding site (see **Figure 22.3**). More recently, fully human antibodies have been prepared by different routes (see **Box 22.1**).

From inauspicious beginnings in the 1980s, mAbs have become the most successful biotech drugs ever, and the market for mAbs has been the fastest-growing component of

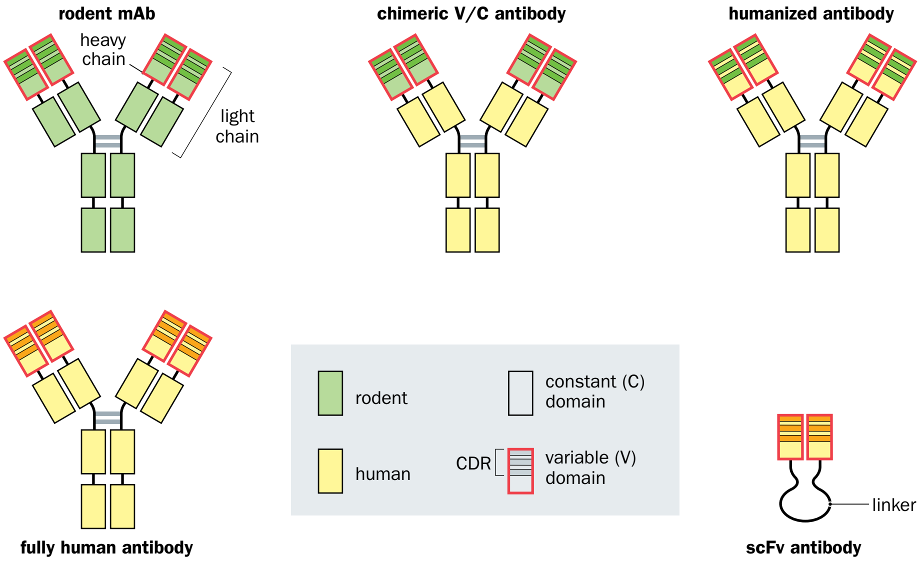


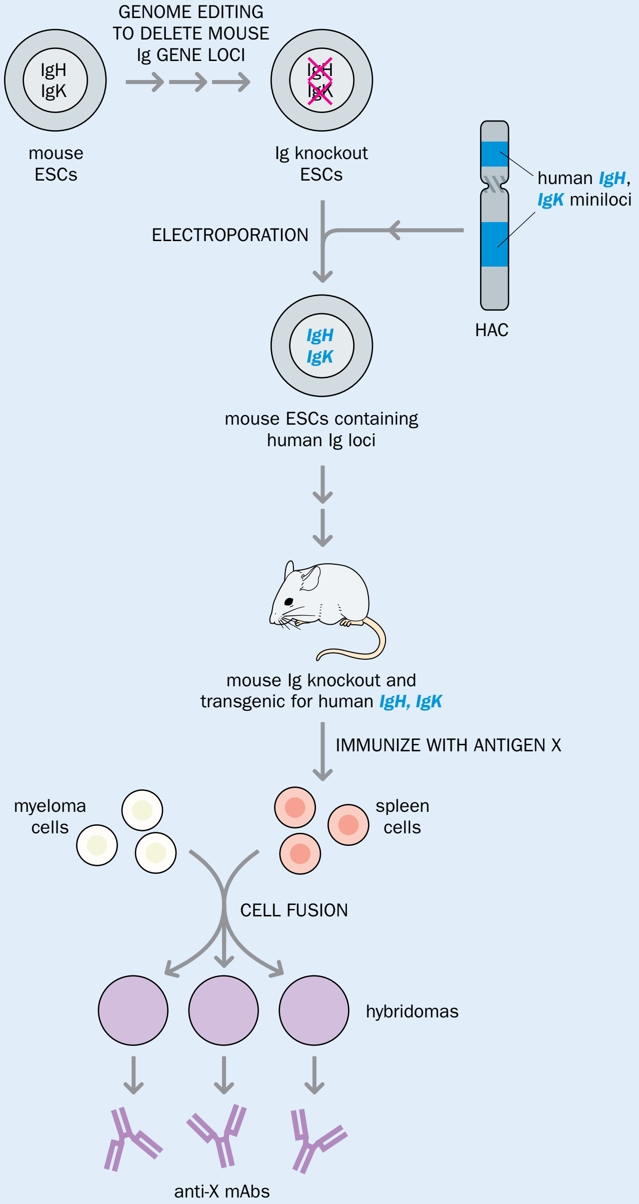
Figure 22.3. Using genetic engineering to make improved therapeutic antibodies. Classical antibodies consist of heavy (H) and light (L) chains with variable (V) and constant (C) domains. Rodent monoclonal antibodies (mAbs) are monospecific antibodies synthesized by hybridomas (see text). Chimeric V/C antibodies are genetically engineered to have human constant domains joined to rodent variable domain sequences (which contain the critically important hypervariable sequence known as the complementarity determining region, CDR). Humanized antibodies can be engineered so that all the sequence is human, except for the hypervariable CDR. More recently, it has been possible to obtain fully human antibodies by different routes (see **Box 22.1**). Genetic engineering has also been used to make single chain antibodies composed of two variable domains only, connected by a linker peptide. These single-chain variable fragment (scFv) antibodies are particularly well suited to working within the reducing environment of cells and can serve as *intrabodies* (intracellular antibodies) by binding to specific antigens within cells. Depending on the length of the linker, they bind their target as monomers, dimers, or trimers. Multimers bind their target more strongly than monomers.

BOX 22.1 MAKING FULLY HUMAN THERAPEUTIC MONOCLONAL ANTIBODIES

Monoclonal antibodies have traditionally been made in rabbits and other rodents by repeated injections with a protein or peptide antigen of interest, then isolating B cells specific for the antigen and fusing them with myeloma cells. The resulting immortal hybridoma cell lines are used to make desired antibodies with diverse research applications (including helping to track human proteins in cells and tissues). Rodent antibodies, however, usually have very limited therapeutic utility: they have limited stability in human serum, and provoke a host immune response against what is a foreign protein. Accordingly, therapeutic antibodies have been engineered to replace some of the rodent amino acid sequences by the equivalent human sequences, beginning with chimeric antibodies and progressing to humanized antibodies, as illustrated in **Figure 22.3**.

The ultimate goal was to make fully human monoclonal antibodies. The key question was how to do that without using humans as the antibody source (which could not be considered for ethical reasons). One successful solution was to bypass hybridoma technology altogether using phage display technology (which has the merit of being fast; for details see the end of Section 6.1). However antibodies discovered using phage libraries show significant disadvantages, including limited diversity, suboptimal biophysical attributes, and poor pharmacokinetics, and can also be immunogenic in patients, with diminishing efficacy over time. Additional, time-consuming *in vitro* engineering methods are typically needed to improve their affinity.

An alternative solution—humanizing transgenic animals—focused on mouse embryonic stem cell lines. Cre–*lox*P-based genome editing in mouse embryonic stem cells (ESCs) can be used to delete the endogenous immunoglobulin loci. Thereafter, series of BAC clones containing segments of human immunoglobulin loci can be stitched into the genome of the ESCs to construct human immunoglobulin loci; details can be found in Lee *et al*. (2014) PMID 24633243, in Further Reading. An alternative approach is to construct artificial chromosomes containing whole human immunoglobulin loci and introduce them into ESCs (see **Figure 1**). In either case, the modified ESCs are used to make transgenic mice in which DNA rearrangements can occur at the human Ig loci in maturing B cells to make human antibodies. The humanized mice are immunized against the antigen of interest, and antigen-specific B cells are isolated and utilized in traditional hybridoma production systems. While not so fast as phage display, the transgenic animal route permits natural immune selection in an intact organism and offers the advantage of high-affinity antibodies without need for further *in vitro* engineering. Different transgenic mice for producing human mAbs are commercially available; rat and bovine systems are also being investigated.



Box 22.1 Figure 1 Humanized transgenic mice can make human monoclonal antibodies. Cre–*lox*P-based genetic engineering in mouse embryonic stem cells (ESCs) can be used to delete the endogenous immunoglobulin loci. BAC clones or yeast artificial chromosomes containing components of, or whole human immunoglobulin loci are then introduced to make fully human monoclonal antibodies (mAbs), endogenous mouse Ig loci can be deleted and replaced by introduced human Ig loci. This figure illustrates how Isao Ishida and colleagues generated a human artificial chromosome (HAC) containing the entire human Ig heavy-chain locus (IGH) plus the entire human Igγ light-chain loci. This HAC was then introduced into mouse ESCs that had previously been subjected to rounds of Cre–*loxP* gene targeting to delete the endogenous mouse Ig loci. The resulting mouse can generate immunoglobulins by rearrangements of the introduced human Ig loci, and so can be used to produce human mAbs of any desired specificity. For full details see US patent 7041870, accessible by searching the Google patent database at https://patents.google.com/. (Reproduced with permission from Thomas Evans Photography.)

the pharmaceutical industry. Of the therapeutic mAbs currently in use, the seven best-sellers together generate an annual income of over US $50 billion. Several hundred additional mAb products are in the pipeline.

The great majority of the approved therapeutic mAbs are aimed at treating diseases where pathogenesis results from positively harmful effects, notably autoimmune/immunologic diseases or cancers (**Table 22.2**). In these cases the antibodies are designed to work by binding specific target proteins on the surface of immune system cells or tumor cells, and thereby inhibit their harmful effects. Some of the latest antibodies developed to treat cancers are being developed as antibody–drug conjugates, so that the antibodies deliver powerful toxins to kill cancer cells.

TABLE 22.2 EXAMPLES OF LICENSED THERAPEUTIC MONOCLONAL ANTIBODIES (mAbs)

Disease category	Target protein	mAb trade name (generic name)	mAb type*	Disease(s) treated
Autoimmune disease/ immunologic	CD11a	Raptiva (efalizumab)	Humanized	Psoriasis
	IgE	Xolair (omalizumab)	Humanized	Asthma
	Integrin α4	Tysabri (natalizumab)	Humanized	Multiple sclerosis
	TNFα	Remicade (infliximab)	Chimeric	Rheumatoid arthritis, ankylosing spondylitis, psoriatic arthritis, Crohn disease, ulcerative colitis
		Humira (adalimumab)	Human	Crohn disease, rheumatoid arthritis and others
Cancer	CD20	Rituxan (rituximab)	Chimeric	Lymphomas, leukemias
	EGFR	Erbitux (cetuximab)	Chimeric	Metastastic colon cancer; head and neck cancer
		Vectibix (panitumumab)	Human	Colorectal cancer
	HER2	Herceptin (trastuzumab)	Humanized	Metastatic breast cancer
	VEGF	Avastin (bevacizumab)	Humanized	Colorectal, breast, renal, NSCL cancer
Other diseases	F protein (RSV)	Synagis (palivizumab)	Humanized	Respiratory syncytial virus prophylaxis
	VEGF	Lucentis (ranibizumab)	Humanized	Age-related macular degeneration

* See **Figure 22.3** for an illustration of mAb types; human, fully human antibody; CD11a, integrin αL, a white blood cell surface antigen; IgE, immunoglobulin E; TNFα, tumor necrosis factor alpha; CD20, a specific B-cell surface protein; EGFR, epidermal growth factor receptor; HER2, human epidermal growth factor receptor 2; NSCL, non-small cell lung cancer; RSV, respiratory syncytial virus; VEGF, vascular endothelial growth factor.

Intrabodies

A more recently developed—and potentially promising—class of therapeutic antibody has been engineered to have a single polypeptide chain. Single chain variable fragment (scFv) antibodies have almost all the binding specificity of a mAb but are restricted to a single nonglycosylated variable chain (**Figure 22.3**). They can be made on a large scale in bacterial, yeast, or even plant cells. They have the advantage of being stable in the reducing environment within cells (in which conventional four-chain antibodies would be unstable). Accordingly, scFV antibodies are well suited to acting as intracellular antibodies (**intrabodies**). Instead of being secreted like normal antibodies, they are designed to bind specific target molecules within cells and can be directed as required to specific subcellular compartments.

Intrabodies can carry effector molecules that perform specific functions when antigen binding occurs. However, in most cases the therapeutic aim envisages that they simply block specific protein–protein associations within cells. As such, they complement conventional drugs. Protein–protein interactions usually occur across large, flat surfaces and are often unsuitable targets for small-molecule drugs (which normally operate by fitting snugly into clefts on the surface of macromolecules). Promising therapeutic target proteins for intrabodies include mutant proteins that tend to misfold in a way that causes neurons to die, as in various neurodegenerative diseases including Alzheimer, Huntington, and prion diseases.

22.3 BASIC PRINCIPLES OF GENE THERAPY AND RNA THERAPEUTICS

Gene therapy involves the direct genetic modification of the cells of a person (or animal disease model) in order to achieve a therapeutic goal. The genetic modification involves the transfer of some artificial genetic construct, often a DNA molecule; but when used in a broad sense, gene therapy may involve transfer of RNA or oligonucleotides. In some cases, the genetic constructs are specifically intended to target RNA transcripts, in which case the term **RNA therapeutics** is often used. According to whether genetically-modified cells can be potentially transmitted to future generations or not, two classes of potential gene therapy can be recognized, as listed below.

- ***Somatic gene therapy.*** The therapy is targeted at somatic cells or tissues of the patient, and any consequences of the genetic modification should be confined to that patient. This type of gene therapy is used when modifying nuclear DNA.
- ***Germ-line gene therapy.*** The aim is to genetically modify the DNA of a gamete, zygote, or early embryo during *in vitro* fertilization. Germ-line gene therapy aimed at modifying nuclear DNA has been widely banned in humans for ethical reasons (the genetic modification can be transmitted to, and affect, descendants). However, germ-line gene therapy aimed at modifying mitochondrial DNA to prevent transmission of severe mtDNA disorders is a rather different matter and has already been legalized in the UK.

Essentially all current human gene therapy trials and protocols involve modifying the genome of somatic cells. However, as gene therapy successes accumulate, and as the technologies become increasingly refined and safe, the idea of extending the technology to modify nuclear DNA in the germ line will come more to the forefront. The possibility of genetic enhancement and "designer babies" inevitably raises ethical concerns.

Gene therapy has had a chequered history. Tremendous initial excitement—and quite a bit of hype—was followed by a fallow period of disappointing results and major safety concerns (with unexpected deaths of patients arising from unforeseen deficiencies in the treatment methods). More recently, there have been very significant successes, and a greater appreciation of safety risks.

In this section and the following sections, we are mostly concerned with gene therapy for inherited disorders, but we also describe some approaches to treating infectious disease. Cancer gene therapy has often been of limited practical use; we describe different approaches to treating cancer in Section 19.5.

Gene therapy for inherited disorders focused initially on recessive disorders. The first real successes were not achieved until the early 2000s and involved very rare cases of severe combined immunodeficiencies. The therapies took advantage of previous experience of bone marrow transplantation, which is effectively a type of stem cell therapy (bone marrow is enriched in hematopoietic stem cells that can both renew themselves and give rise to both blood cells and certain types of tissue cells with an immune system function). Cell therapies based on genetic modification of stem cells are also fundamental in *regenerative medicine* in which the object is to treat disease by replacing cells or tissues that have been lost through disease or injury.

In this section we consider the principles underlying gene therapy. In Section 22.4 we deal with the progress that has been made in standard gene therapy that is designed to transfer normally-functioning genes to overcome a genetic deficiency, and in Section 22.5 we consider progress in RNA therapeutics, therapeutic genome editing, and genetic strategies to prevent disease.

Two broad strategies in somatic gene therapy

In somatic cell gene therapy, the cells that are targeted are often those directly involved in the pathogenic process (but in some types of cancer gene therapy the aim has been to genetically modify normal immune system cells in a patient in an effort to provoke an enhanced immune response against tumor cells).

Using molecular genetic approaches to treat disease might involve many different strategies. But at the level of the diseased cells there are two basic strategies: disease cells are simply modified in some way so as to alleviate disease; or they are selectively killed. Within each of the two main strategies are different substrategies, as described below.

- ***Modifying disease cells*** (**Figure 22.4A**). According to the molecular pathology, different strategies are used. If the problem is loss of function, a simple solution (in theory) is to add functioning copies of the relevant gene. In genetic disorders where the pathogenesis results from a gain-of-function, there is some harmful or toxic gene product within cells. The approach then might be to selectively inhibit the expression of the harmful gene product without affecting the expression of any normal genes. This can often be done by selectively blocking transcription or by targeting transcripts of a specific gene to be destroyed (*gene silencing*). Yet other approaches seek to repair a genetic lesion by genome editing (bottom panel of Figure 22.4A) or find a way of minimizing its effect. We detail the approaches below.

- ***Killing disease cells*** (Figure 22.4B). This approach is particularly appropriate for cancer gene therapy. Traditional cancer treatments have often relied on killing disease cells using "blunt instruments", such as high-energy radiation and harmful chemicals that selectively kill dividing cells. Gene therapy approaches seek to kill harmful cells either directly, or by modifying immune system cells to enhance immune responses that can kill the harmful cells.

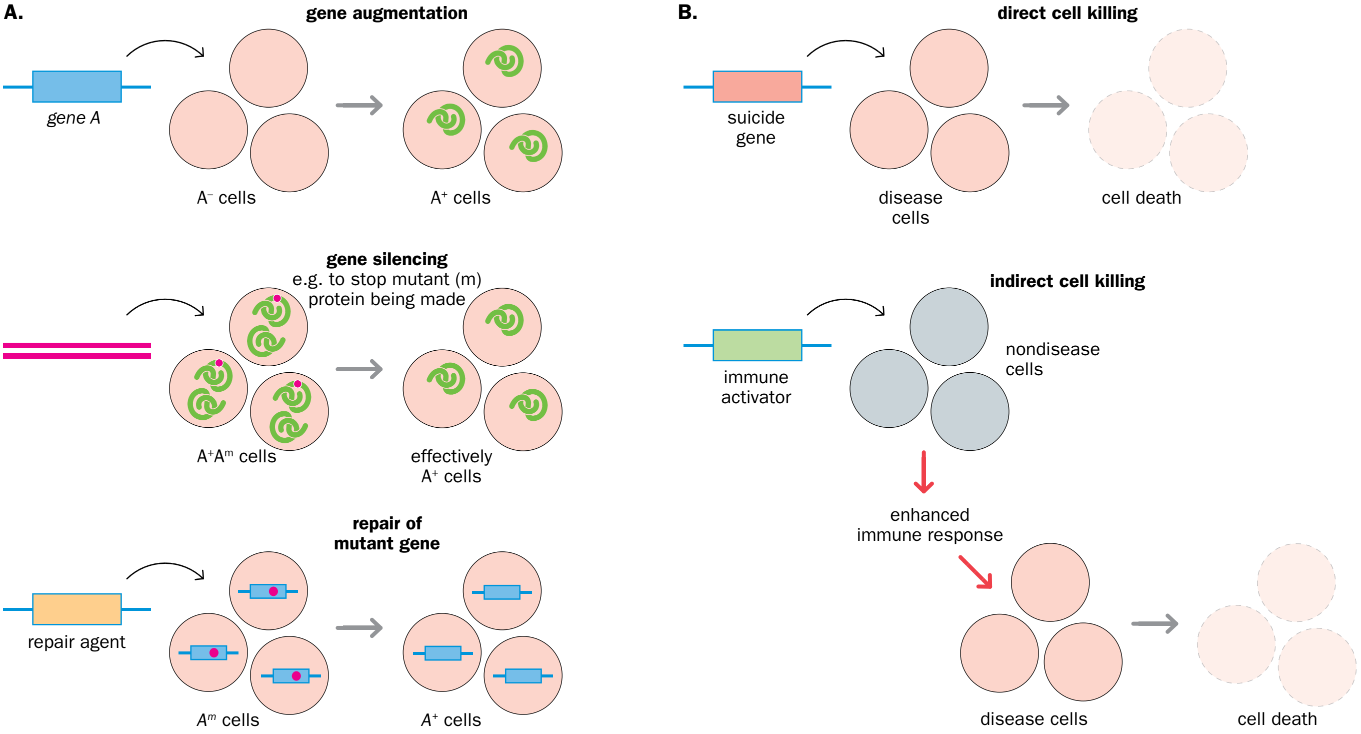


Figure 22.4 Different general types of gene therapy strategy. (**A**) Therapies aimed at modifying disease cells. Gene augmentation therapy can be applied to loss-of-function disorders but currently is limited to treating recessive disorders (where the disease results from lack, or almost complete lack, of some gene product). The object is simply to transfer a cloned working gene copy into the cells of the patient in order to make some gene product that is lacking. Gene silencing therapy is applied to disorders that result from positively harmful gene products. If a gain-of-function mutation produces a harmful mutant gene product, A^m, in addition to the normal gene product A^+, one might try to selectively inhibit the *expression* of the mutant allele by targeting the mutant RNA transcripts (an example of RNA therapeutics). The same approach can be applied to treating autoimmune and infectious diseases. Additional approaches seek to repair the DNA lesion (by using genome editing to convert the mutant allele sequence, A^m, to the normal allele sequence, *A*), or to minimize the effect of a pathogenic mutation (not shown). (**B**) Therapies aimed at killing cells. Cancer gene therapy strategies often rely on killing cancer cells. The aim is to kill cells directly—by inserting and expressing cloned genes that will give rise to some cytotoxic product and cell death—or indirectly, by transferring genes into healthy immune system cells in order to stimulate an enhanced immune response directed at tumors.

The delivery problem: designing optimal and safe strategies for getting genetic constructs into the cells of patients

In gene therapy, a therapeutic *genetic construct* of some type—often a modified cDNA, but sometimes an RNA or oligonucleotides—is transferred into the cells of a patient. (A nucleic acid molecule introduced in this way is often referred to as a **transgene**). According to the disease class, different strategies may be favoured. Depending on the nature of the cells that need to be targeted and their local environment, some disorders are easier to treat in principle than others.

Consider access to the desired target cells. Some cells and tissues—such as blood, skin, muscle, and eyes—are very accessible; others are less so, such as brain cells. Then there is the question of overcoming various barriers that impede transfer and expression of genetic constructs. Strong immune responses constitute important barriers, and as we will see, mechanical barriers can also be important.

Another significant difference occurs between short-lived cells that need to divide periodically to replenish the lost cells (such as blood and skin cells), and long-lived cells, such as terminally-differentiated muscle cells. That is an important distinction because for nondividing cells the key parameters would simply be the efficiency of transfer of the therapeutic construct into the cells of the patient and the degree to which the introduced construct was able to function in the expected way. But for dividing cells we also need to take into account what happens to the descendant cells.

Even if we were to achieve significant success in getting the desired genetic construct into short-lived cells, the cells that have taken up the genetic construct are going to die and will be replaced by new cells. Certain **stem cells** divide to continuously replenish cells lost through aging, illness, or injury. To ensure that copies of the therapeutic construct keep getting into newly dividing cells, therefore, it would be most efficient to target the relevant stem cells, if possible, and get the therapeutic construct integrated into chromosomes (so that it gets replicated, allowing copies to be passed to both daughter cells at cell division).

Efficiency and safety aspects

In any gene-delivery system used in gene therapy two key parameters are fundamental: efficiency and safety. Most gene-therapy methods rely on transferring genes into the cells of a patient and expressing them to make some product. In order for the gene-delivery method to be effective it is important to maximize transfection efficiencies for optimal target cells and to get long-lasting high-level expression of the therapeutic genes.

For disorders where target cells are short-lived, the relevant stem cells should be targeted, but the problem is that the stem cells might occur at very low frequencies. For blood disorders, happily, it is possible to obtain bone marrow cell preparations or peripheral blood lymphocytes from patients, grow the cells in culture, and enrich for hematopoietic stem cells. The purified cells can be genetically modified in culture to overcome a genetic defect and then be returned to the patient, a type of *ex vivo* gene therapy as described in the next section.

As we describe below, viral vectors are commonly used to get therapeutic gene constructs into cells at high efficiency and they often allow high-level expression of the therapeutic transgenes. Some viral vectors are deliberately used because they are adept at getting DNA inserted into chromosomes, which is important when targeting tissues where cells are short-lived. But the features that make the gene therapy process efficient come with significant safety risks.

One important risk concerns the integration of some therapeutic recombinant viruses into chromosomes—there has been little control over where they will insert into the genomic DNA of patient cells. They might insert by accident into an endogenous gene and block its function, but that has consequences just for that cell. The greatest danger from transgene integration is that it activates a neighboring oncogene, causing tumor formation. An additional risk is that the patient might mount a strong immune/inflammatory response to large amounts of what might be viewed to be foreign molecules. Components of viral vectors might pose such risks but even if a perfectly normal therapeutic human gene were inserted and expressed to give a desired protein that the patient completely lacked, (through constitutional homozygous gene deletion, for example) an immune response might occur if the protein had never been produced by the patient. We will revisit these issues in Section 22.4.

Gene therapy is sometimes carried out *in vivo*, but *ex vivo* gene therapy provides significant advantages

In gene therapy a genetic construct is inserted into the cells of a patient (or animal disease model) in an attempt to alleviate disease. The genetic construct is transferred into the cells using either virus vectors (**transduction**), or through nonviral delivery systems (**transfection**). We previously provided technical details of viral and nonviral delivery systems for transporting desired nucleic acids into mammalian cells in Chapter 8; interested readers are referred especially to Section 8.1. In general, viral vector systems are much more efficient than nonviral methods but pose greater safety risks, as described more fully below.

Some types of gene therapy procedure are designed to occur *in vivo*: the transfer of the therapeutic constructs is carried out *in situ* within the patient. Often the therapeutic construct is directly injected into a tissue or organ (such as muscle, eye, brain, and so on). As described below, certain viruses are known to infect human cells of a particular type, and that property has also been exploited to increase the efficiency of delivering therapeutic genes to the desired target cells *in vivo*. As there is no way of selecting and amplifying cells that have both taken up and expressed the genetic construct, the success of *in vivo* gene therapy is crucially dependent on the general efficiency of gene transfer and expression in the correct tissue.

Ex vivo gene therapy

Ex vivo gene therapy means removing cells from a patient, culturing them and genetically modifying them *in vitro*, and then returning suitably modified cells back to the patient (**Figure 22.5**). Because the cells of the patient are genetically modified in the laboratory, there is an enormous advantage: cells can be analyzed at length to identify those cells where the intended genetic modification has been successful. Correctly-modified cells can then be selected, amplified in culture, and transplanted back into the patient. The transplant of genetically-modified self-cells is most effective when the transplanted cells include stem cells, and *ex vivo* gene therapy is especially suited to blood disorders, capitalizing on the long experience of bone marrow transplantation, which effectively is a type of stem cell therapy (see **Box 22.2**).

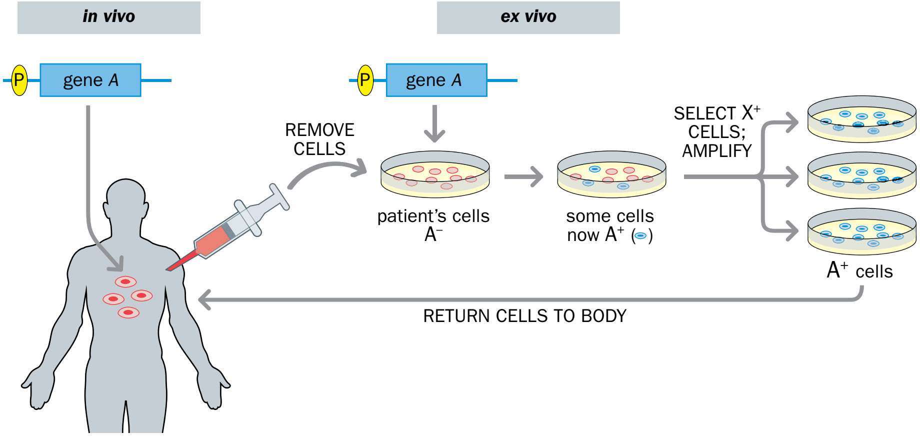


Figure 22.5 ***Ex vivo* and *in vivo* gene therapy.** In *ex vivo* gene therapy, cells are removed from the patient, genetically modified in some way in the laboratory (in this case we illustrate a gene augmentation procedure where a therapeutic transgene *A* is expressed to make a gene product A that is lacking in the cells of the patient). The modified cells are selected, amplified in culture, and returned to the patient. The procedure allows detailed checking of genetically-modified cells to ensure that they have the correct genetic modification before they are returned to the patient. For many tissues this is not possible, and the cells must be modified within the patient's body (*in vivo* gene therapy). P, promoter.

BOX 22.2 FROM BONE MARROW TRANSPLANATION TO *EX VIVO* GENE THERAPY

Bone marrow transplantation has long been used to treat certain cancers of the blood or bone marrow, such as multiple myeloma or leukemia. Effectively, it is a crude type of hematopoietic stem cell transplantation: the bone marrow is enriched in hematopoietic stem cells (which give rise to all of the different blood cells plus tissue dendritic cells and tissue macrophages—see **Figure 3.17**). Prior to transplantation, the patient often has radiation treatment to kill much of the original hematopoietic stem cells (which would otherwise continue to produce abnormal blood cells). The idea is that after the transplant, donor hematopoietic stem cells can proliferate and become the dominant stem cell type, giving rise to healthy blood cells.

Like other types of standard organ and tissue transplantation, bone marrow transplantation normally involves **allogeneic** cell transplantation: the donor and recipient are genetically different (but in very rare cases an identical twin may provide a *syngeneic* transplant). In all allogeneic transplants there is a risk that the graft will be rejected; the major genetic loci determining graft rejection are located in the HLA complex, and finding a good HLA match between donor and recipient is a priority. Bone marrow transplantation, however, is especially dangerous for two reasons. First, because the graft of transplanted donor cells includes immune system cells, life-threatening graft-versus-host disease can occur (immune system cells originating from the donor bone marrow interpret the cells of the patient as being foreign and mount a strong immune response against them). Secondly, if radiation treatment has been chosen, the patient will have a suppressed immune system and be especially vulnerable to infections. Without prior irradiation treatment, bone marrow transplantation has a 10–15% mortality risk that increases to more than 35% with irradiation treatment.

Ex vivo gene therapy for recessive blood disorders combines augmentation gene therapy with hematopoietic stem cell therapy, but in this case transplant of **autologous cells** is involved: the donor cells come from the patient and because they are genetically identical to the recipient cells (except for an inserted transgene), any resulting immune responses are normally insignificant. In this case, the donor cells can be peripheral blood cells that have been enriched in hematopoietic stem cells (by selecting for cells containing the CD34 cell surface marker).

FUTURE *EX VIVO* GENE THERAPY USING CELL REPROGRAMING

The accessibility of blood cells and the long experience of transplanting crude preparations of hematopoietic stem cells have led to *ex vivo* gene therapy being directed at recessive blood disorders (with very considerable success, as detailed in Section 22.4). Extension of *in vivo* therapy to other disorders may be possible by cell reprogramming. A sample of blood or skin cells from the patient might be taken and the cells reprogrammed *in vitro* to make *induced pluripotent stem cells* that are genetically modified, then directed to differentiate to give a required type of stem cell or progenitor cell.

Nonviral systems for delivering therapeutic genetic constructs: safety at the expense of efficiency

Interest in nonviral vector delivery systems has mostly been propelled by safety concerns using viral vectors. The nonviral vector systems are certainly safer—they do not integrate into chromosomes and they are not very immunogenic. Additionally, it is possible to transfer very large DNA molecules using nonviral methods. The big downsides are low transfer efficiencies and often low-level transgene expression.

The therapeutic gene is typically carried in a plasmid vector but transport of plasmid DNAs into the nucleus of nondividing cells is normally very inefficient (the plasmid DNA often cannot enter nuclear membrane pores). Various tricks can be used to help get the plasmids into the nucleus (such as by conjugating specific DNA or protein sequences known to facilitate nuclear entry, or compacting the DNA to a small enough size to pass through the nuclear pores). Because the transfected DNA cannot be stably integrated into the chromosomes of the host cell, nonviral methods of therapeutic gene delivery are more suited to delivery into tissues where the cells rarely divide, such as muscle, and where the injected DNA may continue to be expressed for several months.

Different delivery systems can be used. In some cases, naked DNA has been injected directly with a syringe and needle into a target tissue such as skeletal muscle. More efficient transfer and greater stability of nucleic acids can be achieved using lipid-based transfer systems or by conjugating certain chemicals to the nucleic acid that can cause the DNA to be compacted, as detailed below.

- ***Lipid-based transfer systems.*** Transporting nuclei acids into cells is aided by complexing the nucleic acid with lipids. When certain lipids are mixed in aqueous solution lipid vesicles can form spontaneously (phospholipids, for example, can form bilayered vesicles that mimic the structure of biological membranes, with the hydrophilic phosphate groups on the outside and the hydrophobic lipid tails in the inside). A lipid coating can protect the nucleic acid from nucleases and helps them in binding to the plasma membrane and in being endocytosed. Cationic lipids are often used alongside helper lipids. The cationic lipids help in binding negatively-charged nucleic acids, aid binding to plasma membranes (which have frequent surface negative charges, including from phosphate and sulfate groups found within membrane-bound glycoproteins and phospholipids), and permit increased escape from endosomes. Unlike viral vectors, the nucleic acid/lipid complexes are easy to prepare and there are is no limit to the size of nucleic acid that is transferred.

  Cationic lipid vesicles are the method of choice for introducing short interfering RNA (siRNA) into cells. They have also been used to transfer DNA transgenes (but the efficiency of gene transfer is low with comparatively weak transgene expression, and because the introduced DNA is not designed to integrate into chromosomal DNA, transgene expression may not be long-lasting). In gene therapy the two most commonly used lipid transfer vehicles are cationic liposomes and cationic lipid nanoparticles (see Figure 22.6).

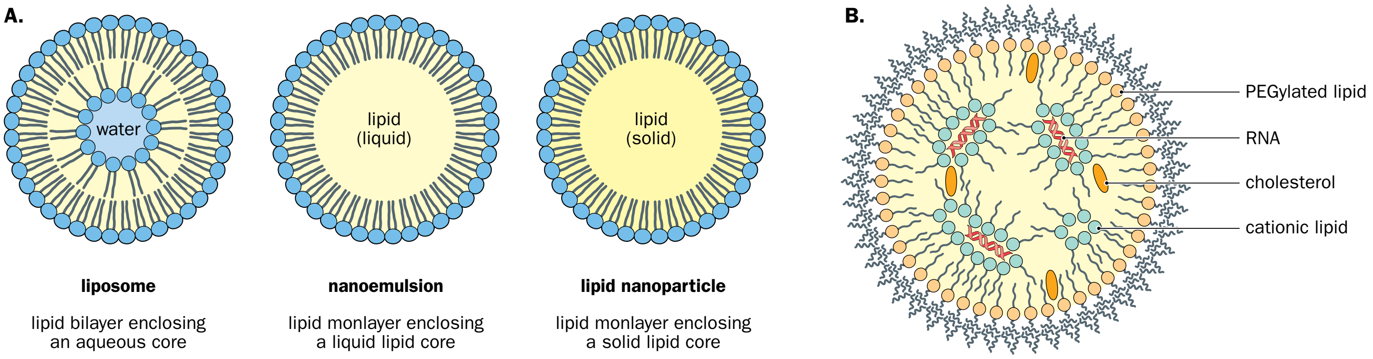


Figure 22.6 Structure of lipid nano-dispersed vehicle systems. (**A**) Three major classes of nano-dispersed vehicle system. (**B**) An example of a cationic lipid nanoparticle transporting a siRNA. PEGylated means complexed with poly(ethylene glycol), which results in greater stability. (A, reproduced from Urechi O *et al*. [2014] *Application of Nanotechnology in Drug Delivery*, ed. Sezer AD INTECH; http://dx.doi.org/10.5772/58672, published under CC BY license 3.0; B, reproduced with permission of Dove Press from Zatsepin TS *et al*. [2016] *Int J Nanomed* **11**:3077–3086; PMID 27462152. Permission conveyed through Copyright Clearance Center, Inc.)

- ***Compacted DNA nanoparticles.*** Polycations bind strongly to negatively-charged DNA, causing the DNA to be significantly compacted (histones, which have a high frequency of positively-charged arginine and lysine side chains play this role *in vivo*). To form DNA nanoparticles, DNA is often complexed with a polyethylene glycol-substituted poly-L-lysine known as PEG-CK30, so-named because it has an N-terminal cysteine (C), to which polyethylene glycol (PEG) is covalently bound, plus 30 lysine (K) residues. Within this complex, the DNA forms a very condensed structure. Because of their greatly reduced size, compacted DNA nanoparticles can pass through nuclear membrane pores, and so are comparatively efficient at transferring genes to both dividing and nondividing cells. They have a plasmid capacity of at least 20 kb.

Viral delivery of therapeutic gene constructs: relatively high efficiency but safety concerns

Viruses have a DNA or RNA genome packaged within an outer protein coat (capsid). They normally attach to suitable host cells by recognizing and binding specific receptor proteins on the host cell surface. Some viruses infect a broad range of human cell types and are said to have a broad **tropism**. Other viruses have a narrow tropism: they bind to receptors present on the surface of just a few cell types. For example, herpes viruses are tropic for central nervous system cells. The natural tropism of viruses may be retained in vectors, or genetically modified in some way so as to target a particular tissue.

Enveloped viruses have the capsid enclosed by a lipid bilayer containing viral proteins. Some of them enter cells by fusing with the host plasma membrane to release their genome and capsid proteins into the cytosol. Other enveloped viruses first bind to cell-surface receptors and trigger receptor-mediated endocytosis, fusion-based transfer, or endocytosis-based transfer.

Some viruses can gain access to the nucleus only after the nuclear envelope dissolves during mitosis. They are limited to infecting dividing cells. Other viruses, such as HIV and other lentiviruses, have devised ways to transfer their genomes efficiently through nuclear membrane pores. As a result, both dividing and nondividing cells can be infected.

To allow easy insertion of a therapeutic gene construct into a viral vector, the vector is in a double-stranded DNA form. Some of the most popular vectors used in gene therapy have been based on **retroviruses**, single-stranded RNA viruses that encode a reverse transcriptase after infecting cell to make a cDNA copy of their RNA genome. The resulting single-stranded DNA is then used to make a double-stranded DNA (replicative form) that gets inserted into a host cell chromosome (details are given in Section 8.1).

Integrating and nonintegrating viral vectors

Integrating vectors allow therapeutic genes to be inserted into chromosomes of cells, and to be passed on to any descendant cells (an important advantage when the target cells have a high cell turnover rate). They are typically based on retroviruses, which are adept at inserting genes into chromosomes. The vector is made by isolating viral replicative forms (consisting of double-stranded DNA), and genetically modifying them in various ways.

Initially, integrating vectors were based on gammaretroviruses (formerly called oncoretroviruses), a class of retroviruses with simple genomes (detailed in Section 8.1). As described below, however, there have been important safety issues with gammaretrovirus vectors. As a result, modern clinical trials use safer integrating vectors, often ones based on a class of more complex retroviruses known as lentiviruses, including HIV (**Figure 22.7A**). Lentiviral vectors also have the ability to target nondividing cells as well as dividing cells.

Nonintegrating vectors are traditionally based on DNA viruses and they can be especially useful when the object is to get high-level expression in nondividing target cells, such as muscle. Using adenovirus virus vectors used to be popular because they can permit very high levels of gene expression, but there have been safety issues (which relate to their immunogenicity, as described below). Safer vectors based on adeno-associated virus (**Figure 22.7B**) subsequently became popular. See **Table 22.3** for properties of viral vectors commonly used in gene therapy.

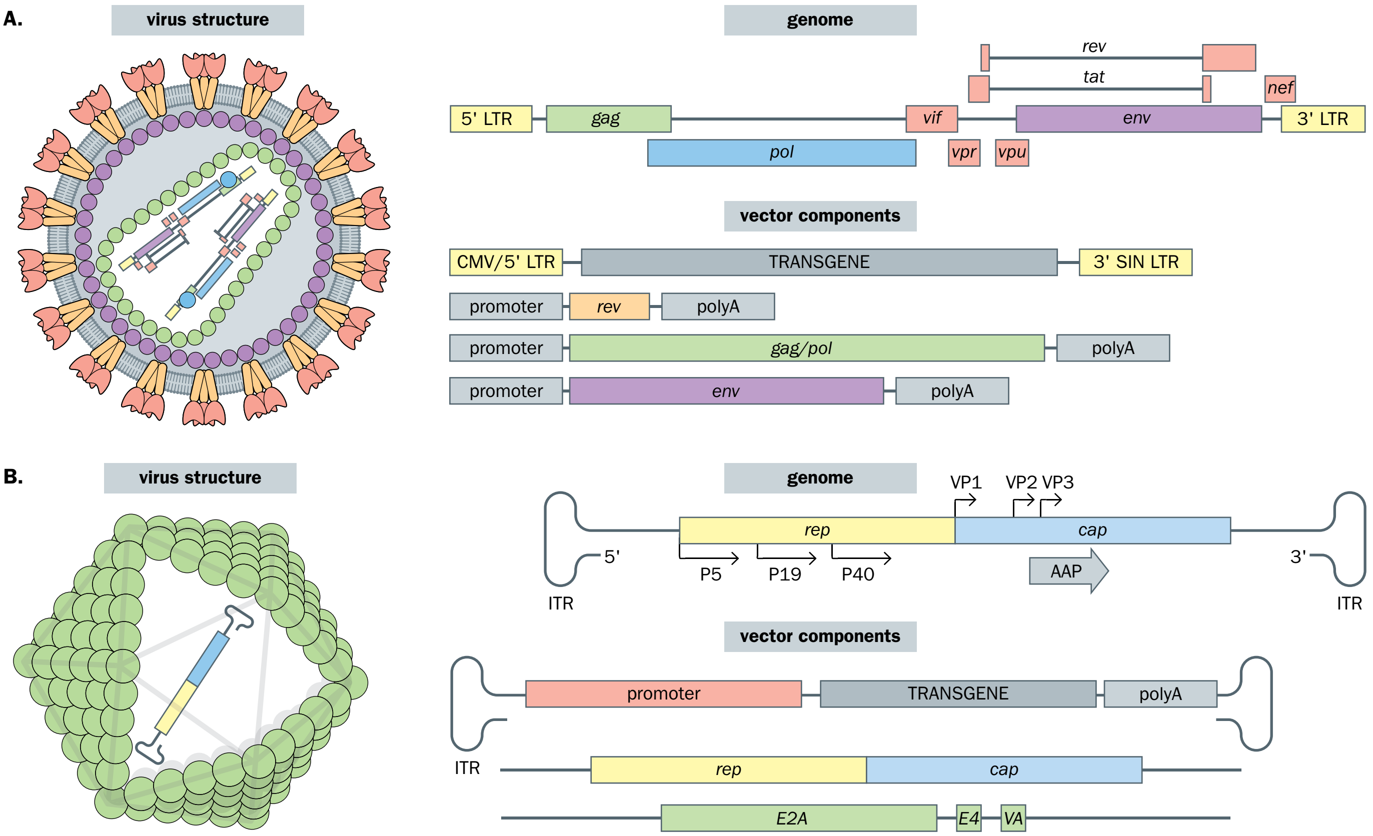


Figure 22.7 Lentivirus and adeno-associated virus: structure and vectors. (**A**) Lentivirus. Like gammaretroviruses, lentiviruses have two copies of positive-strand RNA surrounded by the protein capsid and envelope, and the RNA genome has *gag/pol* and *env* genes flanked by long terminal repeats (LTRs). But lentiviruses have additional protein-coding genes: *tat* and *rev* make regulatory proteins, and *nef*, *vif*, *vpr*, and *vpu*, encode accessory proteins. In lentivirus vectors the therapeutic transgene is inserted between the viral LTRs, which also function as a promoter sequence. Other vector components include *gag/pol*, *env*, and *rev* RNA sequences that are supplied in *trans*. CMV, cytomegalovirus; SIN, self-inactivating. (**B**) Adeno-associated virus (AAV). The virus has a single-stranded DNA genome enclosed within the protein capsid. The 4.7 kb genome is flanked by inverted terminal repeats (ITR) and has three classes of open reading frames: *rep* (replicase/integrase) makes proteins required for the AAV lifecycle; *cap* makes capsid proteins; and *aap* makes the assembly-activating protein, AAP, needed for capsid assembly. In recombinant AAV (rAAV) vectors, the therapeutic transgene, along with associated promoter and polyadenylation sequences, is inserted between the viral ITRs. To produce the vector AAV *rep* and *cap*, and adenovirus E2A, E4, and VA RNA sequences are supplied in *trans*. VA, viral-associated.

TABLE 22.3 FOUR CLASSES OF VIRAL VECTORS THAT HAVE BEEN WIDELY USED IN GENE THERAPY PROTOCOLS

Virus class	Viral genome	Cloning capacity	Integrating?	Target cells	Transgene expression	Vector yield** and other comments
Gamma-retroviruses (oncoretroviruses); see **Figures 8.6–8.9**	ssRNA; ~8–10 kb	7–8 kb	Yes	Dividing cells only	Long-lasting	Moderate vector yield Risk of activating cellular oncogene
Lentiviruses, notably HIV; see **Figure 22.7A**	ssRNA; ~9 kb	Up to 8 kb	Yes	Dividing/nondividing cells Tropism varies	Long-lasting and high level expression	High vector yield Low risk of oncogene activation
Adenoviruses	dsDNA; 38–39 kb	often 7.5 kb but up to 34 kb	No	Dividing and nondividing cells	Transient but high level expression	High vector yield Immunogencity a major problem
Adeno-associated viruses (AAV); see **Figure 22.7B**	ssDNA; ~5 kb	<4.5 kb	No (mostly)*	Dividing/nondividing cells Strains can be selectively tropic	High level expression in medium to long-term (year)	High vector yield Small cloning capacity Immunogenicity is less of a problem than for adenovirus

Note that gamma-retroviruses and adenoviruses used to be the most commonly-used gene therapy vectors until major safety concerns were raised in both cases (**Box 22.3**). In modern gene therapy the preferred virus vectors are based on lentiviruses and adeno-associated viruses (AAVs). * Recombinant AAVs very occasionally integrate but are mostly episomal. ** Moderate and high vector yields signify 10^{10} transducing units/ml and 10^{12} transducing units/ml, respectively. ds, double-stranded; ss, single stranded.

22.4 THE PRACTICE OF GENE AUGMENTATION THERAPY FOR TREATING RECESSIVELY INHERITED DISORDERS

Gene therapy has had a roller-coaster ride over three decades; periods of overoptimism would be followed by bouts of excessive pessimism in response to significant set-backs. The first undoubted successes were reported in the early 2000s and the number of successful reports is beginning to increase significantly.

By November 2017, the Wiley database of gene therapy clinical trials worldwide (available at http:// www.abedia.com/wiley) had listed over 2597 such trials—see **Table 22.4** that lists trials according to disease/indication. The majority (65%) have been aimed at treating cancers. Many cancer gene therapies focus simply on killing cancer cells and so use different types of gene therapy approach; they have been of limited clinical value.

TABLE 22.4 INDICATIONS FOR GENE THERAPY CLINICAL TRIALS

Disease(s) or other indications	Proportion of gene therapy clinical trial (%)
Cancers	65.0
Cardiovascular diseases	6.9
Infectious diseases	7.0
Inflammatory diseases	0.6
Monogenic disorders	11.1
Neurological diseases	1.8
Ocular diseases	1.3
Other diseases	2.2
Other indications*	4.1

Data from the Gene Therapy Clinical Trials Worldwide website at www.abedia.com/wiley).
* Includes trials on healthy volunteers, and gene marking studies.

Of the 36% or so of gene therapy trials not focused on treating cancer, the most popular have been aimed at treating monogenic disorders, but significant numbers have been carried out for infectious diseases and complex diseases. However, the vast majority of the listed trials are phase I and II trials; only 3.8% of the listed trials are phase III trials, where the efficacy of the therapy is tested on a large scale. Here, we focus on gene therapies where the object has been to genetically modify disease cells. They include gene therapies for monogenic disorders (where there have been many definitive successes), and also for some infectious diseases.

Multiple successes for *ex vivo* gene augmentation therapy targeted at hematopoietic stem cells

Successful *ex vivo* gene therapy trials have been carried out for various recessively inherited blood disorders and some storage disorders by targeting bone marrow cells or peripheral blood lymphocytes that had been enriched for hematopoietic stem cells. Our blood cells are short-lived, and need to be replaced by new cells derived from self-renewing hematopoietic stem cells. These cells, found mostly in the bone marrow (and to a lesser extent in peripheral blood), give rise to all of the many different types of blood cells, and also to some tissue cells that have immune functions, including tissue macrophages, such as brain microglia, and tissue dendritic cells (see **Figure 3.17**).

For some of the disorders treated in this way, alternative treatments have sometimes been used. For some blood disorders treatment with purified gene product (such as recombinant proteins) is an option, but it is extremely expensive. Bone marrow transplantation has been occasionally used, but it has very significant risks. In *allogeneic* bone marrow transplantation the donor is often a family relative, such as a sibling, but complete HLA matching of donor and patient is rare (even for siblings there is only a 1 in 4 chance), and sometimes transplantation is attempted using partial HLA matching between donor and recipient. Severe, life-threatening, graft-versus-host disease can result (see **Box 22.2** for details). The advantage of *ex vivo* gene therapy for blood disorders, therefore, is that it is significantly less expensive than using purified proteins, and much less risky than bone marrow transplantation.

Safety issues in gammaretroviral integration

The first successes came in treating severe immunodeficiencies. In severe combined immunodeficiency (SCID) the functions of both B and T lymphocytes are defective. Affected individuals have virtually no functioning immune system and are extremely vulnerable to infectious disease.

The most common form of SCID is X-linked; inactivating mutations in the *IL2RG* gene means a lack of the common gamma (γc) subunit for multiple interleukin receptors, including interleukin receptor 2. (Lymphocytes use interleukins as *cytokines*, chemical messengers that help in intercellular signaling, in this case between different types of lymphocyte and other immune system cells; lack of the γc cytokine receptor subunit has devastating effects on lymphocyte and immune system function). Another common form of SCID is due to adenosine deaminase (ADA) deficiency; the resulting build-up of toxic purine metabolites kills T cells. B-cell function is also impaired because B cells are normally regulated by certain types of regulatory T cells.

The first SCID gene therapy trials involved *ex vivo* gammaretroviral transfer of *IL2RG* or *ADA* coding sequences into autologous patient cells. To aid the chances of success, bone marrow cells from the patient were further enriched for hematopoietic stem cells by selecting for cells expressing the CD34 surface antigen, a marker of hematopoetic stem cells (**Figure 22.8**). By 2008 17/20 X-linked SCID patients and 11/11 ADA-deficient SCID patients had been successfully treated and retained a functional immune system (for >9 years post-treatment in the earliest patients).

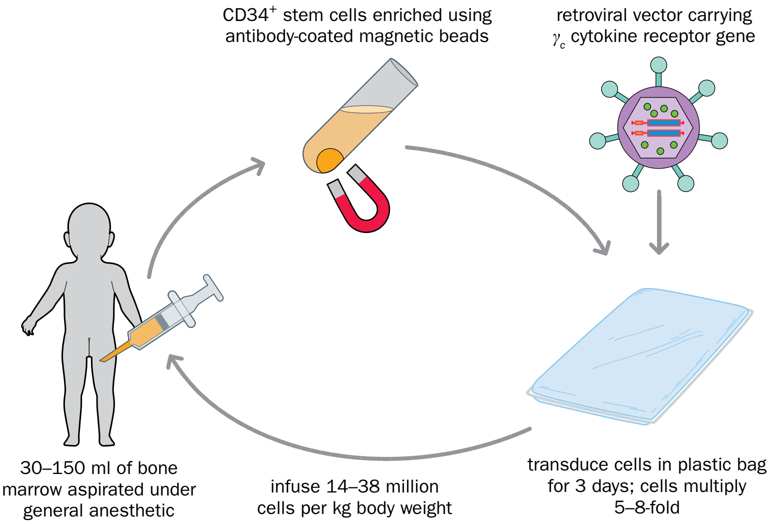

$CD34^+$ stem cells enriched using antibody-coated magnetic beads
retroviral vector carrying γ_c cytokine receptor gene
30–150 ml of bone marrow aspirated under general anesthetic
infuse 14–38 million cells per kg body weight
transduce cells in plastic bag for 3 days; cells multiply 5–8-fold

Figure 22.8 The first successful gene therapy: *ex vivo* gene therapy for X-linked severe combined immunodeficiency disease (X-SCID). Bone marrow cells were removed from the patient and antibody affinity was used to enrich for cells expressing the CD34 antigen, a marker of hematopoietic stem cells. To do this, bone marrow cells were mixed with paramagnetic beads coated with a CD34-specific monoclonal antibody; beads containing bound cells were removed by using a magnet. The transduced stem cells were expanded in culture before being returned to the patient. For details, see articles with PubMed identification (PMID) numbers 10784449 and 11961146.

Although the use of integrating retrovirus vectors was beneficial in terms of efficiency, the chromosomal insertion of the transgenes was unsafe and led to several patients developing leukemia (see **Box 22.3**). The same kind of approach has been successfully applied to some other blood disorders. However, oncogene activation in some other cases also led to leukemia in patients or silencing of the inserted transgenes. In the first major breakthrough in gene therapy for β-thalassemia, when a patient benefited from treatment, the transgene inserted into the *HMGA2* gene (which encodes a protein that works in gene regulation). That resulted in overexpression of a truncated HMGA2 protein; long-term follow-up will be required to confirm safety and efficacy.

Increased safety profiles using lentiviral vectors

More recently, *ex vivo* gene therapy trials have largely used self-inactivating lentivirus vectors. They have the advantage of long-lasting high level expression, and are much safer. Abnormal activation of an endogenous gene is very rarely triggered when lentivirus vectors integrate into chromosomes. That is so for two major reasons. First, lentivirus vectors don't have quite the same tendency to insert close to transcriptional start sites as do gammaretroviruses. Secondly, in self-inactivating (SIN) lentiviral vectors the very strong viral promoter and enhancer sequences in the long terminal repeats have been deleted and replaced by more appropriate mammalian promoter sequences.

The first successful gene therapy using a lentivirus vector was for a lipid storage disorder that primarily affects the brain, but nevertheless depended on *ex vivo* gene transfer into hematopoietic stem cells. X-linked adrenoleukodystrophy (OMIM #300100) is a progressive neurodegenerative disorder that is a accompanied by adrenal insufficiency. The adrenal insufficiency is treatable, but there is no effective treatment for the neurodegeneration. Affected boys have inactivating mutations in the *ABCD1* gene and usually die in adolescence. The *ABCD1* product, the peroxisomal membrane protein ALDP, is important for natural degradation of very-long-chain fatty acids in the peroxisomes; a deficiency of this protein results in harmful build-up of very-long-chain fatty acids. The result is progressive loss of the lipid-rich myelin sheath of nerve cells (and axon degeneration), and an impaired ability to convert cholesterol into steroids, so that the adrenal glands fail to make steroid hormones.

Ex vivo gene therapy was designed to halt the progression of the disease by transducing an autologous cell population enriched in hematopoietic stem cells with a recombinant HIV vector containing an *ABCD1* coding sequence. The transduced hematopoietic stem cells gave rise to myelomonocytic cells (with characteristics of both granulocytes and monocytes) that migrated into the central nervous system to replace diseased microglial cells and relieve the lipid storage problem.

In vivo gene therapy: approaches, barriers, and recent successes

In *in vivo* gene therapy, a genetic construct is transferred (usually directly) into postmitotic disease cells at specific sites in the body (such as muscles, eyes, brain, liver, lung, heart, and joints). Because the intended target cells are nondividing cells, there is no need to insert genes into chromosomes, and the viral vectors are typically based on nonintegrating DNA viruses.

Delivery using adenovirus and AAV vectors

Early *in vivo* gene therapy trials often used adenovirus vectors to transfer therapeutic transgenes. They allow high-level expression and some adenovirus vectors can accept inserts as large as 35 kb (much larger than the vast majority of full-length human cDNA sequences). But harmful immune and inflammatory responses have sometimes resulted (**Box 22.3**). The vectors are nonintegrating and expression of introduced transgenes is often somewhat transient; short-term and repeated administration would be necessary for sustained expression, which could only exacerbate the immune response.

Adeno-associated viruses (AAV) are nonpathogenic and are quite unrelated to adenoviruses (their name comes from their natural reliance on coinfection by a *helper virus,* often an adenovirus). Their most important advantage is that they can permit robust *in vivo* expression of transgenes in various tissues over several years while exhibiting little immunogenicity and little or no toxicity or inflammation. Multiple different serotypes of AAV have been isolated and some have usefully narrow tropism, such as AAV8 (strongly tropic for the liver). A major downside is that a maximum of just 4.5 kb of foreign DNA can be inserted into an rAAV (recombinant AAV) vector (to maximize insert capacity, the *rep* gene that is involved in directing occasional integration of AAV into a specific site on human chromosome 19 is deleted).

Amenability of disorders to *in vivo* gene therapy

Different disorders may be more or less amenable to *in vivo* gene therapy, largely depending on the efficiency of trangene transfer and expression. That, in turn, partly depends on different types of barrier. Immunologic barriers are particularly important when using recombinant virus vectors: as well as posing safety risks, immunologic responses can result in transgene silencing (increased host cell cytokine signalling often attenuates the influence of viral promoters).

In addition to immunologic barriers, mechanical barriers can also be a major obstacle. Take disorders that primarily affect the lungs, such as cystic fibrosis, for example. Gene delivery to the airways might seem a very attractive option, given that lung epithelial cells directly interface with the environment. But a combination of immunologic and mechanical barriers is a formidable obstacle to gene therapy. Lung epithelial cells are locked together by intercellular *tight junctions*, and large numbers of macrophages are on patrol, ready to intercept and destroy viral vectors. And finally, there is a natural,

BOX 22.3 TWO MAJOR SAFETY ISSUES IN GENE THERAPY TRIALS

As described below, the practice of gene therapy has been marred by serious illness and even fatalities in the case of a few participating subjects. (Although tragic, the partial successes provided some hope and benefit to seriously affected patients, and fundamental information for how to improve the gene therapy design.) Two serious safety issues have been highlighted. One applies to integrating vectors—the lack of control over the integration event can lead to altered expression of genes at the integration site, and subsequent tumor formation. The second safety issue is a general one and concerns the immunogenicity of the therapeutic genetic construct.

PROBLEMS WITH INTEGRATING VECTORS

From an efficiency point of view, integrating retroviral vectors can be highly desirable. By allowing therapeutic transgenes to be stably transmitted following cell division, they can offer long-lasting transgene expression in populations of cells that need to undergo cell division, such as white blood cells. But early gene therapy trials for severe combined immunodeficiency (SCID) showed that using conventional gammaretroviral vectors (based on the murine Moloney leukemia virus) was not safe. Five out of the 19 individuals who had been treated for X-linked SCID as shown in **Figure 22.8** went on to develop T-acute lymphoblastoid leukemia.

It subsequently became clear that gammaretroviral vectors have a pronounced tendency to integrate close to transcriptional start sites. Additionally, the long terminal repeats carry very powerful promoter and enhancer sequences that can readily activate expression of neighboring host cell genes. The integration site location might have been expected to vary considerably, but surprisingly in four out of the five patients who developed leukemia, the same gene was inactivated by transgene insertion, the proto-oncogene *LMO2*. (Activation of *LMO2* is now known to promote the self-renewal of thymocytes so that committed T cells accumulate additional genetic mutations required for leukemia transformation.)

Subsequent clinical trials have used safer, self-inactivating retrovirus vectors in which powerful promoter/enhancers in the LTRs are deleted, then replaced by more moderate mammalian control sequences without significant effects on recombinant vector performance. Lentivirus vectors have been preferred because they are less disposed to integrate next to transcriptional start sites. But even the modern vectors are not entirely safe: there is little control over where they integrate. To further reduce safety risks, efforts are being made to direct transgene integration to certain safe sites within the genome (“safe harbors”; see the 2016 review by Papapetrou and Schambach, PMID 26867951, in Further Reading).

PROBLEMS WITH IMMUNOGENICITY

Any introduced therapeutic genetic construct is potentially immunogenic. Nonviral vectors are less of a problem here, and synthetic lipid-based transfection systems can be designed to be weakly immunogenic. By contrast, viral vectors can be a very significant problem. Humans often have antibodies to strains of viruses from which gene therapy vectors have been developed. And in the case of nonintegrating, or mostly nonintegrating, vectors (such as adenoviruses and adeno-associated virus [AAVs]), where the expression of therapeutic transgenes often declines with time, gene therapy protocols usually involve giving repeated doses, which tends to exacerbate immune responses.

That immunogenicity problem was highlighted by the tragic death of Jesse Gelsinger, just a few days after receiving recombinant adenoviral particles by intrahepatic injection in a clinical gene therapy trial for ornithine transcarbamylase deficiency—the procedure provoked a massive inflammatory reaction that resulted in multiple organ failure. Since then, AAV vectors have been preferred to adenovirus vectors because of their greater safety profiles.

protective layer of mucus on the epithelial surface that becomes thicker in individuals with cystic fibrosis.

Some parts of the body are *immunologically privileged sites* in which immune responses to foreign antigen are much weaker than in most others parts of the body (due to blood–tissue barriers, lack of lymphatics etc). They include the brain and much of the eyes. Additional advantages of the eyes are their accessibility, and also their compactness (compare the need for multiple injections at diverse skeletal muscle sites in disorders such as Duchenne muscular dystrophy).

The liver, too, is a quite accessible organ (via direct injection, injection into the hepatic portal vein, or even injection into a peripheral vein), and because it has a primary role in biosynthesis, the liver has become a popular target for gene delivery. A wide range of metabolic disorders are caused by defective synthesis of proteins manufactured in the liver (such as blood clotting factors VIII and IX that are deficient in hemophilia, and many enzymes in inborn errors of metabolism).

Examples of successful *in vivo* gene therapy

Although the spotlight has mostly shone on *ex vivo* gene therapy, there have been some successes too with *in vivo* gene therapy. Two recent examples are given below.

- Hemophilia B (OMIM #306900). This X-linked recessive disorder is caused by deficiency of blood clotting factor IX. The disorder can be treated by protein therapy (using clotting factor concentrates) but at huge cost. Remarkably, a single intravenous injection of a recombinant AAV (rAAV) construct with a factor IX cDNA sequence could successfully treat patients with hemophilia for more than a year, even although factor IX expression levels were about 10% or less of the normal values.

- Type 2 Leber congenital amaurosis (OMIM #204100). The principal clinical feature—profound loss of vision—usually presents at birth. In the type 2 form, the blindness results from inactivating mutations in both copies of the *RPE65* gene, causing severe retinal degeneration (*RPE65* encodes a retinal pigment epithelium enzyme). Different *in vivo* gene therapy trials have involved injecting an rAAV construct containing a transgene with the *RPE65* coding sequence into the subretinal space, allowing transduction of retinal pigment epithelial cells. The trials showed the procedure was both safe, and of considerable clinical benefit. In a large clinical trial all patients demonstrated increased pupillary and increased visual field and a majority of patients demonstrated improved visual acuity.

22.5 RNA THERAPEUTICS, THERAPEUTIC GENOME EDITING PROSPECTS, AND GENETIC APPROACHES TO PREVENTING DISEASE

The classical gene augmentation therapy described in Section 22.4 has dominated the headlines because of the recent major successes reported for treating certain recessive diseases, led by the hematopoietic stem cell gene therapies described above. Here we describe some additional therapeutic approaches that use different genetics-based approaches to treat or prevent disease.

RNA therapeutics: gene silencing by RNA interference and modulation of RNA splicing

Different types of RNA therapeutic strategies can be used. The most common approaches seek to specifically down-regulate the expression of a harmful gene that produces a toxic product or has some other harmful gain of function (*gene silencing*). We also describe here a more specialized RNA therapeutic method designed to reduce the severity of the disease phenotype by forcing the mutant gene to undergo altered splicing.

Gene silencing therapy using RNA interference

Different diseases are potentially amenable to treatment based on gene silencing. In each case the problem is a gene that is doing something positively harmful, such as in inherited disorders where the disease is due to a gain-of-function mutation, or a dominant negative effect. Here, the strategy must be to selectively inhibit the expression of the mutant gene, with minimal effect on any normal allele.

Different technologies can be used to achieve gene silencing. Initial attempts used gene-specific antisense RNAs to bind to transcripts of a mutant gene, and then selectively block expression, or modified ribozymes (RNA enzymes) would be designed to cleave a specific RNA transcript. However, large single-stranded RNA is very prone to degradation. The most popular therapeutic gene silencing method utilized RNA interference (RNAi), an innate defense mechanism that protects cells against invading viruses and also from excess activity by transposable elements (**Box 8.2** details the mechanism).

Therapeutic RNAi involves delivery of small RNA duplexes designed to have a sequence corresponding to part of the normal RNA transcript of the gene of interest. In response, RNA-induced silencing complexes (RISC) initiate a pathway leading to functional inactivation of any RNA transcripts containing the same nucleotide sequence as the introduced RNA. Different types of short RNA can be transferred into cells, or plasmid DNAs are provided that are transcribed and processed to give duplex RNAs (**Figure 22.9**).

Over the years, therapeutic gene silencing has been beset by various difficulties. Initially there were significant concerns about toxicity deriving from off-target effects (RNA transcripts from other genes were affected in addition to the intended transcripts), and sometimes also immunogenic effects of the introduced RNA. Efficient delivery of small RNAs to the desired target cells has also been a challenge, but recent advances have been made in delivery. Cationic lipid transfer systems are the most widely used gene delivery system, but different conjugates have recently been used to target delivery to certain cell types. (For delivery to hepatocytes, for example, the siRNA can be conjugated to *N*-acetyl galactosamine, a sugar that interacts with the asialoglycoprotein receptor present on hepatocytes, facilitating endocytosis).

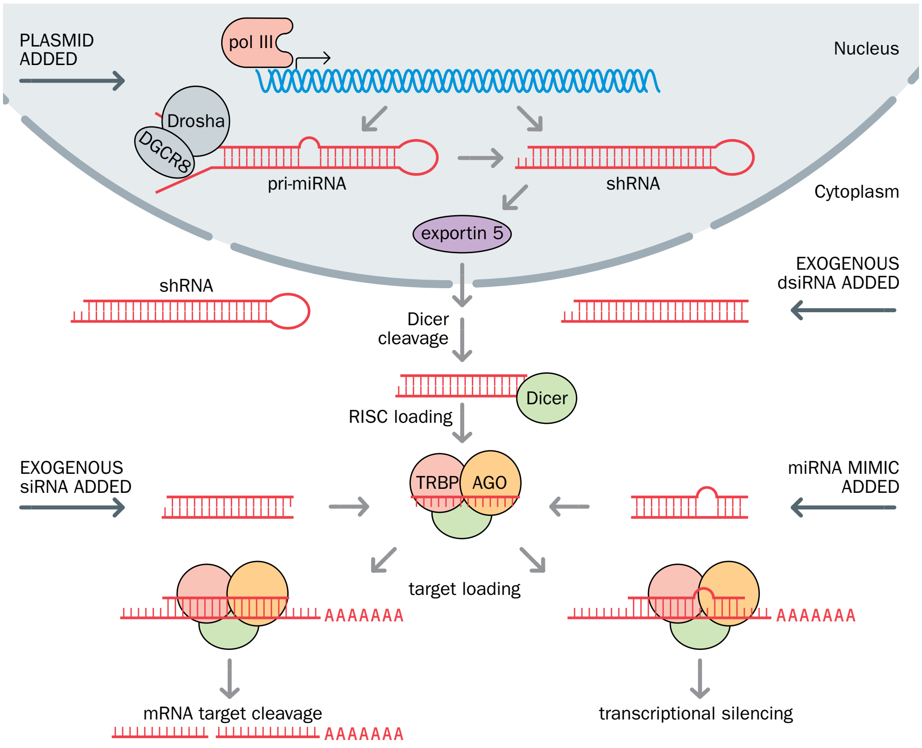


Figure 22.9 RNAi therapeutics enter the RNAi pathway at one of four places. The most upstream step is to add a plasmid DNA that is transcribed by RNA polymerase III (pol III) to give a short hairpin RNA (shRNA) or a primary microRNA (pri-miRNA). This requires nuclear processing followed by export into the cytoplasm. There, shRNA and Dicer substrate RNA (dsiRNA) are processed into canonical short interfering RNA (siRNA) or microRNA (miRNA). The shRNAs bind to target sequences via the RNA-induced silencing complex (RISC) and result in messenger RNA (mRNA) target cleavage and degradation. If miRNA is delivered, the target sequence is bound by RISC, which results in translational repression followed by segregation into p-bodies for degradation. pol III, polymerase III. (Reproduced from Bobbin ML, Rossi JJ [2016] *Annu Rev Pharmacol Toxicol* **56**:103–122; PMID 26738473. With permission from Annual Reviews. Permission conveyed through Copyright Clearance Center, Inc.)

Different tissues have been amenable targets, notably the eye. Macugen is the only approved RNA therapeutic at the time of writing. Designed to suppress expression of vascular endothelial growth factor in the eye, it is used to treat macular degeneration, a degenerative condition that is a leading cause of blindness in adults. Advances in gene delivery have made the liver, too, an amenable target using intravenous or subcutaneous injections. By 2018, more than 40 different clinical trials using RNA therapeutics were ongoing, several at the phase III stage.

Modulation of splicing

One unusual approach to treat disease is to force a disease gene to undergo a specific altered splicing pattern to significantly reduce the normally harmful effect of the pathogenic mutation. To do that antisense oligonucleotides are designed to bind to specific splice-junctions in pre-mRNA transcripts. Blockading the splice junction from interacting with the spliceosomal machinery causes exon skipping. The induced exon skipping might be applicable in cases where the exon contains a harmful mutation and the number of nucleotides in the exon is exactly divisible by three (to maintain the reading frame). Or it might be used to correct a frameshifting deletion. But scenarios like these would be appropriate only in rare cases where loss of an exon does not have a catastrophic effect.

One prominent example relates to Duchenne muscular dystrophy. The very long dystrophin protein joins the contractile machinery of muscle cells to the plasma membrane, using functionally important sequences at the N- and C-terminal end as hooks. A sizeable proportion of the central protein segment, however, is functionally much less important: if it is deleted by mutation, a shorter protein is made that continues to work, but less efficiently. Large non-frameshifting deletions spanning multiple central exons of the X-linked dystrophin gene are associated with a mild phenotype, therefore, whereas nonsense mutations and frameshifting point mutations in the central exons (and frameshifting large deletions) cause a severe phenotype (**Figure 22.10A**). Exon skipping can be induced to restore the reading frame for mutant genes with a frameshifting deletion (**Figure 22.10B**), or a harmful point mutation, and that can result in significant restoration of dystrophin expression.

The prospects of therapeutic genome editing using programmable nucleases

Early successes in gene therapy used homologous recombination-based genome editing, but genome editing using programmable nucleases is increasingly being adopted.

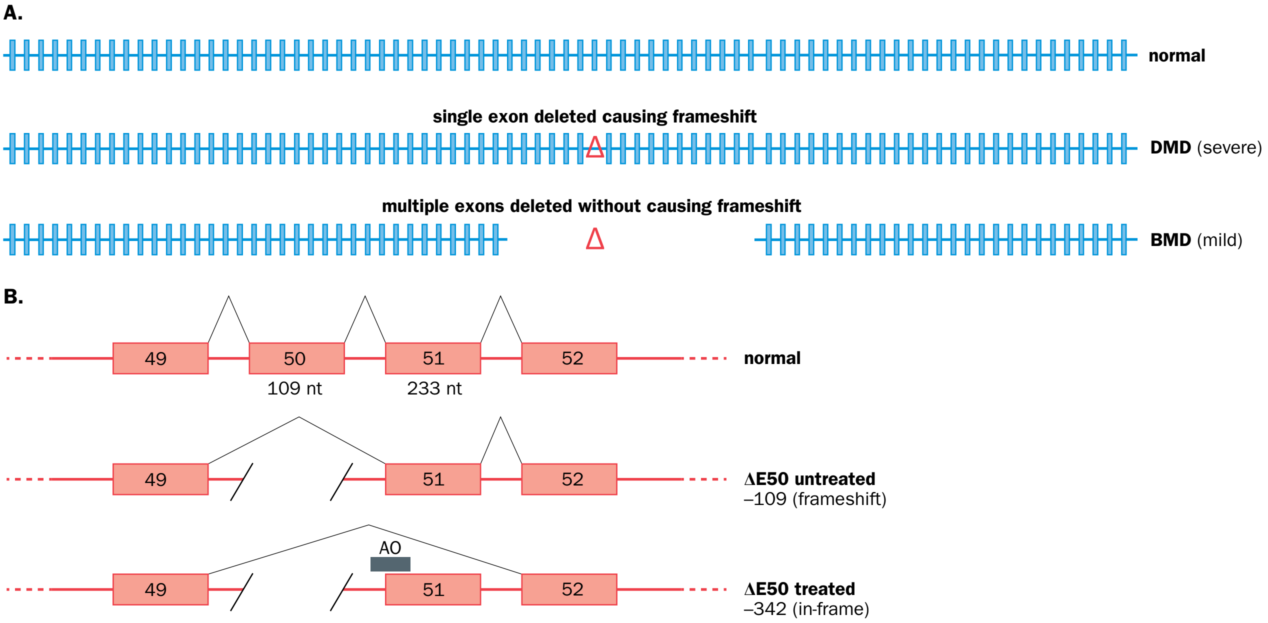


Figure 22.10 Inducing exon skipping to mitigate the effects of pathogenic mutations in the dystrophin gene. (**A**) Nonsense mutations and frameshifting point mutations in a central exon of the 79-exon X-linked dystrophin gene cause severe Duchenne muscular dystrophy (DMD) in boys; large nonframeshifting deletions that remove multiple central exons of the dystrophin gene are associated with the much milder Becker muscular dystrophy (BMD). (**B**) Illustration of how skipping exon 51 could restore the translational reading frame in patients who have a deletion of exon 50 (ΔE50) in the dystrophin gene. Deletion of the 109-nucleotide (nt) exon 50 results in the splicing of exon 49 to exon 51. The loss of the 109 nucleotides (not a multiple of three) produces a frameshift and severe DMD. Therapy that causes skipping of exon 51 in ΔE50 patients will result in the splicing of exon 49 to exon 52. That type of skipping takes out 109 + 233 = 342 nucleotides (a multiple of three), and the translational reading frame is maintained even although the coding capacity is reduced (with a loss of 114 amino acids). The exon skipping occurs because a specific antisense oligonucleotide (AO) is used to bind to, and blockade, the splice junction at the start of exon 51. Administration of an antisense oligonucleotide (by local intramuscular injections) to induce skipping of exon 51 has been reported to restore dystophin production in muscle fibers of patients with the appropriate types of dystrophin exon deletion, and to provide significant clinical benefit. PMID 18160687 gives the experimental details; for a general review, see Spitali P & Aartsma-Rus A (2012) PMID 22424220 in Further Reading.

It involves making precise changes to a predetermined sequence within the genome of intact cells using artificially engineered endonucleases to make a double-strand break at the desired target site (Section 8.4 gives the background).

Site-specific endonucleases, such as zinc finger nucleases and TALE nucleases (TALENs), may be used, and have two key characteristics: a DNA-cleaving domain that cuts the DNA, and a DNA-binding domain. The latter acts as a *protein guide sequence*: it is designed to specifically recognize and bind the target DNA sequence, thereby positioning the DNA-cleaving domain at the target site. In these cases, a transgene is transferred into cells and expressed to produce the site-specific endonuclease.

A recent alternative is to transfer transgenes that make an endonuclease and a related RNA guide sequence that can bind the endonuclease (as notably employed in the CRISPR-Cas system). Here the RNA guide sequence is designed to be complementary in sequence to a desired target sequence. After binding the nuclease, the guide RNA then seeks out the complementary sequence in DNA, bringing the endonuclease to the desired DNA sequence to make the double-strand break there.

In both cases, the aim is to make a double-strand break at a specific desired position within intact cells. After the double-strand break is made, the method relies on natural cellular DNA repair, and different repair pathways can be used. In the nonhomologous end joining pathway, random errors frequently occur during DNA repair so that the sequence of the repaired segment can differ from cell to cell, and cells with a desired sequence change are selected. Alternatively, the homologous recombination pathway of DNA repair is artificially directed: a suitable transgene sequence is copied to make a desired sequence change (see **Figure 8.15**).

Therapeutic genome editing involves changing the sequence of a predetermined gene in cells taken from a patient, selecting cells that have the desired sequence, and re-inserting them in the patient.

Many therapeutic applications can be envisaged, and different approaches can be taken (**Figure 22.11**). The most general type of approach, one that can be applied in

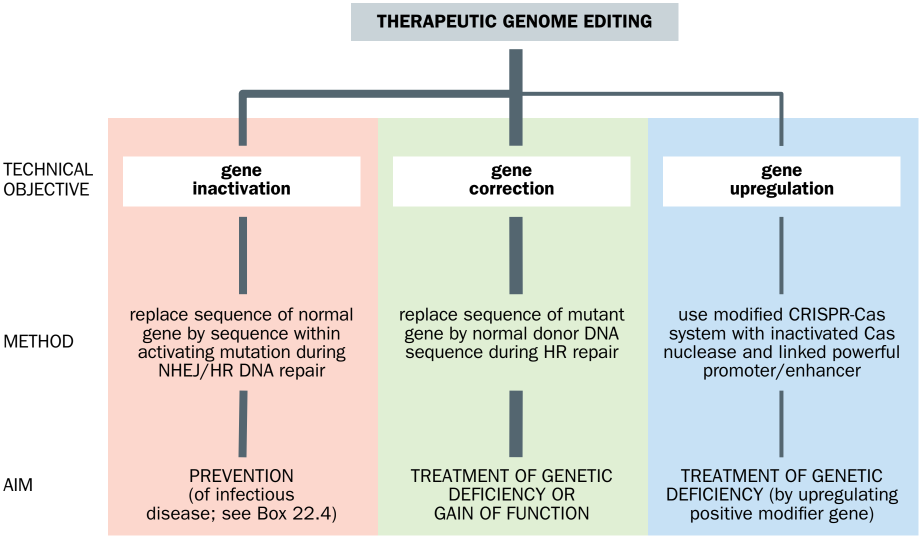

THERAPEUTIC GENOME EDITING
TECHNICAL OBJECTIVE
METHOD
AIM
gene inactivation
gene correction
gene upregulation
replace sequence of normal gene by sequence within activating mutation during NHEJ/HR DNA repair
replace sequence of mutant gene by normal donor DNA sequence during HR repair
use modified CRISPR-Cas system with inactivated Cas nuclease and linked powerful promoter/enhancer
PREVENTION (of infectious disease; see Box 22.4)
TREATMENT OF GENETIC DEFICIENCY OR GAIN OF FUNCTION
TREATMENT OF GENETIC DEFICIENCY (by upregulating positive modifier gene)

Figure 22.11 Different strategies for therapeutic genome editing. A general strategy for treating genetic disease using genome editing is to replace a mutant sequence in a disease-causing gene by the equivalent normal sequence (gene correction); the bulk of therapeutic genome editing might eventually be expected to be in this area. A more specialized application might be in upregulating a gene whose expression can mitigate disease arising from loss of function (such as compensating for loss of dystrophin by up-regulating a gene that makes a functionally-related protein, utrophin). Gene inactivation can also be applied to the specialized case of infectious disease to prevent production of a human cell surface protein that a pathogen needs to recognize in order to infect cells. HR, homologous recombination; NHEJ, nonhomologous end joining.

principle to any genetic condition (including gain-of-function mutations), involves replacing the sequence of a mutant gene by the equivalent normal sequence.

One potential application of therapeutic genome editing, which was first considered at an early stage in the technology, envisaged treating infectious disease by making host cells resistant to viral infection (see **Box 22.4** for the example of HIV therapy). Subsequently, as the technologies using programmable nucleases have become mainstream, this route towards gene therapy has become more generally used. The CRISPR-Cas technology is especially easy to carry out and it is efficient, but there is the drawback that Cas9 cleavage specificity is highly dependent on the CRISPR-RNA used, and off-target effects can occur. That is, the nuclease may sometimes cut at a similar sequence elsewhere in the genome (which is a particular concern for therapeutic applications: a single off-target effect might, for example, activate an oncogene). Ways of increasing Cas9 nuclease specificity are actively being explored (such as by incorporating structurally rigid modified oligoribonucleotides known as bridged nucleic acids into CRISPR RNAs—see PMID 29654299).

BOX 22.4 INVESTIGATING THE POTENTIAL FOR CURATIVE HIV THERAPY USING STEM CELL TRANSPLANTATION AND GENE EDITING

Genetic variation between people causes differences in susceptibility to infectious diseases. In some extreme cases people appear to be highly resistant to infection by a disease-causing virus. That can happen when a person naturally does not make a specific host cell receptor that the virus must recognize and bind to before infecting cells.

Take the human immunodeficiency virus HIV. It begins its attack by infecting CD4$^+$ helper T cells, regulatory immune system cells with a major role in helping to protect us against viruses. To latch onto a helper T cell, HIV requires to bind to the CD4 receptor on the T cell surface, and then interacts with a co-receptor that, for most HIV strains, is the chemokine (C-C motif) receptor 5 (CCR5). By attacking and killing helper T cells, HIV causes the immune system to be compromised; people with AIDS are unable to fight off common infections, and they develop various virus-induced cancers. Although T cells have a limited life span, HIV-AIDS persists because new T cells in an infected person also become infected with HIV.

Unlike CD4, the CCR5 receptor is not important in T-cell function, and some normal people naturally have defective CCR5 receptors: a *CCR5* allele with an inactivating 32 bp deletion (*CCR5-Δ32*) is carried by 5–14% of normal individuals of European descent. Heterozygotes with one *CCR5-Δ32* allele are more resistant to HIV infection than the normal population, and normal *CCR5-Δ32* homozygotes are highly resistant to HIV infection.

THE POTENTIAL FOR CURATIVE HIV THERAPY USING STEM CELL TRANSPLANTATION

HIV infection can be kept in check by a maintenance strategy of daily antiretroviral therapy, using a combination of drugs designed to reduce viral replication but that work at different stages in the HIV lifecycle. However, the famous "Berlin patient" study led to the idea of potentially curative HIV therapy, simply by transplanting blood stem cells (see **Figure 1**).

The Berlin patient study was clearly an exceptional situation (HLA-identical donors who happen to be *CCR5-Δ32* homozygotes are rare: about 1–2% of people in Caucasian populations are *CCR5-Δ32* homozygotes; in other populations the incidence is very much lower). All transplants of

Box 22.4 Figure 1 The "Berlin patient", the first person to have been cured of HIV. Timothy Ray Brown, an American living in Berlin, was diagnosed with HIV in 1995 but later went on to also develop acute myeloid leukemia (AML). His doctor in Berlin decided to try to cure both diseases at once by hematopoietic stem cell transplantation. After checking donor data at the German Bone Marrow Centre, multiple prospective HLA-identical donors were listed, one of whom was subsequently identified to be a *CCR5-Δ32* homozygote. The first transplant of allogeneic CD34$^+$ peripheral blood stem cells from the HLA-identical *CCR5-Δ32* homozygote donor occurred in 2007. After an AML relapse, a second transplant was required a year later, and immediately afterwards antiretroviral therapy was discontinued. Nine years after the study was published in 2009 (PMID 19213682), Mr. Brown appears to be free from HIV (and AML).

allogeneic CD34$^+$ peripheral blood stem cells, however, come with very significant risks of graft-versus-host (GVH) disease (where donor lymphocytes originating from the stem cells attack the recipient's cells). Timothy Brown nearly died after the second transplant (despite the apparent HLA identity between donor and recipient tissue, other mismatched loci outside the HLA system can contribute to GVH disease). In follow-up allogeneic stem cell transplantation in other patients with both HIV and cancer, several patients have died of side-effects related to the allogeneic stem cell transplantation.

GENE EDITING TO RENDER HOST CELLS RESISTANT TO HIV INFECTION

Various studies have sought to extend HIV-resistance to other HIV$^+$ individuals by using genome editing to homozygously inactivate *CCR5* in autologous helper (CD4$^+$) T cells or CD34$^+$ stem cells. Because the stem cell preparations were prepared from the blood of the patients, subjected to *ex vivo* genome editing and then re-infused into the patient, the autologous stem cell transplantation avoids the high risks of allogeneic stem cell transplantation. Zinc finger nucleases were used in genome editing of *CCR5* in the first clinical trials (which investigated the safety of the procedure). Because off-target effects were increasingly found to be common, more recent, ongoing trials are employing TALE nucleases or CRISPR-Cas-based genome editing to inactivate *CCR5.*

Mitochondrial replacement therapy: a specialized case of preventing transmission of genetic disorders by modification of the germ line

As illustrated above, the path towards successful somatic gene therapies has certainly not been a smooth one (and has resulted in several fatalities), but at least the direct consequences are restricted to the person that has been treated. Genetic modification of the human germ line is a different proposition: here, the consequences may also extend to future generations.

Proponents of germ-line gene therapy argue that as genetic modification techniques become ever more refined, germ-line genetic modification could be used to eradicate a variety of severe genetic disorders, such as Huntington disease. The counter argument is that there is no need for germ-line gene therapy. Even if genetic modification of embryos were 100% accurate and reliable, surely it would be simpler to carry out *in vitro* fertilization procedures for prospective parents at risk of passing on a severe pathogenic variant, and use genetic screening to select embryos free of the harmful variant? Currently, gene therapy employing genetic modification of nuclear DNA is widely banned, but modification of human germ-line mtDNA is a different story, as described below.

Prevention of mtDNA disorders by mitochondrial replacement therapy

Mutations in mitochondrial DNA (mtDNA) are a significant cause of human disease. They are transmitted exclusively by mothers, but in many women the inheritance is complicated by the co-existence of wild-type and mutant mtDNA molecules (heteroplasmy). For a given pathogenic mutation, the severity of the mtDNA disorder is determined by the mitochondrial *mutation load*, the proportion of mutant to wild-type mtDNA. For many mtDNA disorders, disease is manifest when the mutation load exceeds a threshold of around 80%.

There is no adequate treatment for mtDNA disorders and some of them are severe, and sometimes fatal, causing miscarriage or early death in childood. As a result, clinical efforts have focused on prevention. However, there is a dramatic reduction in mtDNA copy number during early oogenesis (the germ-line bottleneck), and different oocytes from the same mother can show substantial variation in the mutation load. As a result, accurate prediction of disease risk to future children is impossible. Embryos with low mitochondrial mutation loads can be identified by pre-implantation genetic diagnosis, but this type of diagnosis is not suitable for women whose oocytes contain consistently high mutation loads. Instead, in these cases a different type of genetic intervention has been developed, known initially as nuclear genome replacement but subsequently renamed as mitochondrial replacement therapy (also called mitochondrial donation).

In mitochondrial replacement therapy, enucleated oocytes from healthy women donors are used as a source of normal mitochondria and to form hybrid embryos for IVF, where the nuclear genome is provided by the prospective parents. The nuclear genome can be provided in the form of a *karyoplast* (a combination of male and female pronuclei) obtained after IVF of an affected maternal oocyte by paternal sperm (pronuclear transfer, see Figure 22.12A), or before fertilization (spindle transfer, see Figure 22.12B). Because the zygote is formed using healthy donor mtDNA to replace the abnormal maternal mtDNA, this can be regarded as a type of germ-line gene therapy and is banned in many countries. In 2015, however, a change in the law of the UK was enacted, permitting mitochondrial replacement by pronuclear transfer and in late 2016 the procedure received approval by the UK's regulating authority, the Human Embryo and Fertilization Authority.

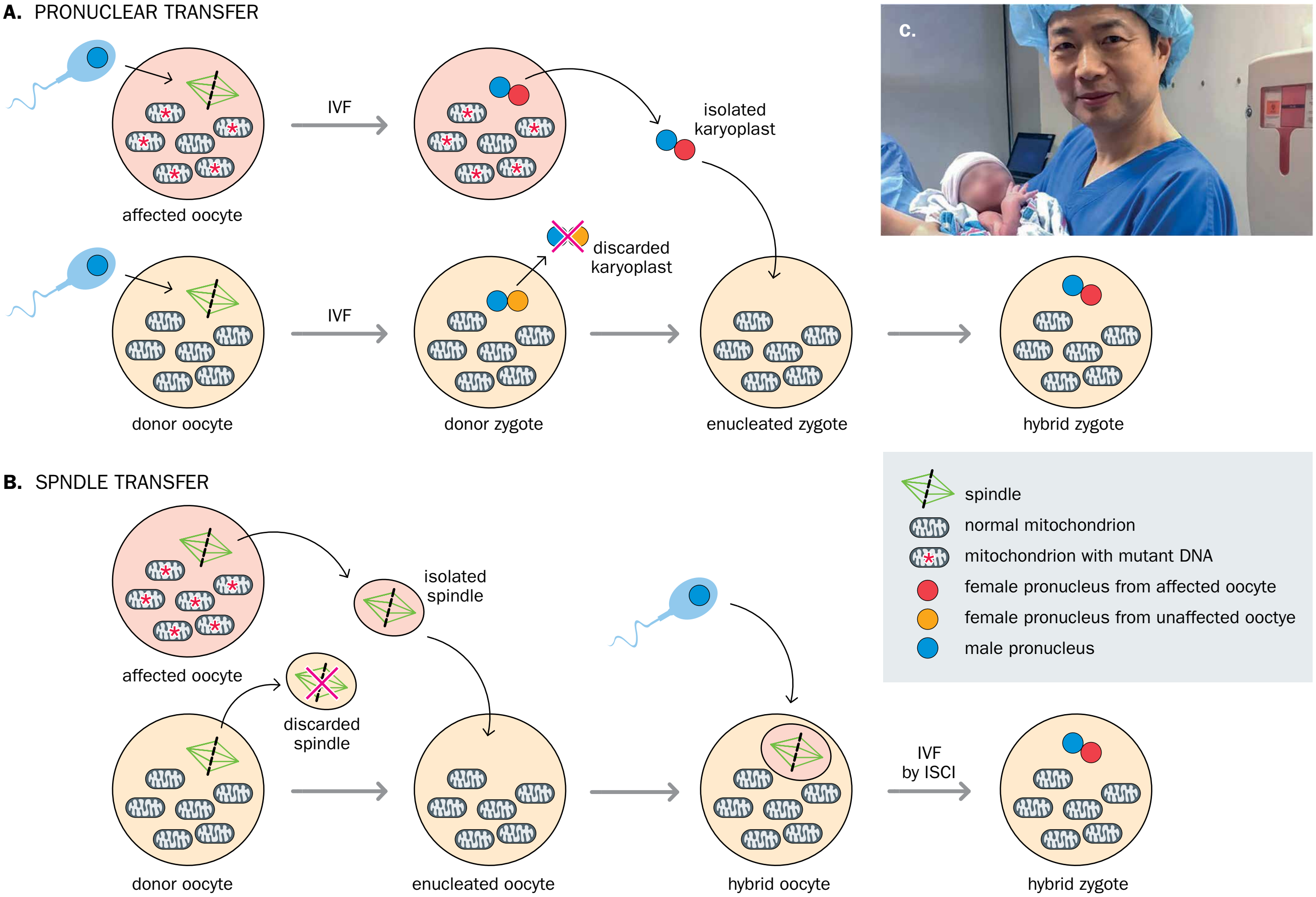


Figure 22.12 Mitochondrial replacement therapy, a type of germ-line gene therapy, may prevent transmission of severe mtDNA disease. The idea is to use an enucleated oocyte with healthy mitochondria from a donor to provide normal mtDNA and for the prospective parents to provide the nuclear genome. The nuclear genome can be provided after or before *in vitro* fertilization (IVF). (**A**) Pronuclear transfer technique. An affected oocyte from the prospective mother (with a high mutation load where many, if not all, of the mtDNAs carry the pathogenic mutation) is fertilized by her partner's sperm and the resulting normal *karyoplast* (combined male and female pronuclei) is isolated, then transferred into an enucleated donor zygote with normal mitochondria. The result is a hybrid zygote with "foreign" but normal mtDNA. (**B**) Metaphase II spindle transfer technique. The metaphase II spindle is transferred from an oocyte with mutant mtDNA into a mitochondrial donor oocyte, producing a hybrid oocyte with a nuclear genome from the prospective mother but mtDNA from the donor. Fertilization by intracytoplasmic sperm injection (ICSI) produces a hybrid zygote. (**C**) In 2016 a hybrid human zygote produced by mitochondrial donation gave rise to a three-parent baby in an attempt to prevent transmission of the severe neurological disorder, Leigh syndrome. The image shows the newborn boy who when tested some months later had a mutation load of just 1%. Holding him is Dr. John Zhang from the New Hope Fertility Center in New York. See text for more details. (A and B adapted from Craven L *et al*. [2011] *Hum Mol Genet* **20**:R168–174; PMID 21852248, with permission from Oxford University Press; C reproduced courtesy of New Hope Fertility Center, New York.)

Mitochondrial replacement has been used in mouse and primate models, with encouraging results, but at the time of writing, mitochondrial donation is still illegal in most countries, including the USA. However, the first use of human mitochondrial replacement therapy was recently carried out by an American doctor in Mexico (where the procedure is not illegal), and the first three-parent baby resulting from mitochondrial donation was born in 2016 (**Figure 22.12C**). The treatment was intended to prevent transmission of Leigh syndrome, a severe neurological disorder that leads to miscarriage or death in early childhood (typically within 2–3 years). The Jordanian parents had had four previous miscarriages and two children that died at an early age from the disease. Encouragingly, the baby had a mitochondrial mutation load of only 1%, but will need to be carefully monitored in case there should be any replicative advantage for the mutant mtDNA, leading to a progressive increase in the mutation load.

SUMMARY

- Different therapeutic strategies are required according to the nature of pathogenesis. If the problem is a deficiency, some form of augmentation therapy is used to provide a missing component. If the problem is some positively harmful factor, the aim is to destroy it or block it in some way, minimizing the damage. In some cases prevention may be the best option.
- Therapeutic proteins are sometimes made by expressing cloned human genes in cells to make a human “recombinant protein” that can be purified and used to treat a genetic deficiency of that protein.
- Therapeutic antibodies are usually designed to bind to positively harmful gene products to block their effects. Rodent monoclonal antibodies are not ideal (with limited lifetimes after injection into patients); genetic engineering allows replacement of rodent sequences by human sequences to make more effective antibodies.
- Genetically-engineered antibodies with a single variable polypeptide chain can work as intracellular antibodies (intrabodies), by binding harmful proteins within cells.
- Gene therapy means inserting a nucleic acid or oligonucleotide into the cells of a patient in order to counteract or alleviate disease.
- According to the delivery method, a transferred nucleic acid (transgene) or oligonucleotide may be transported into cells. Retrovirus vectors are useful for integrating a transgene into a chromosomal location, but there is normally very little control over where the vector integrates into the host cell genome.
- Virus vectors are highly efficient at transporting therapeutic genetic constructs into cells, but nonviral vectors are safer. A virus vector integration event can sometimes cause tumorigenesis, and some virus vectors sometimes cause dangerous immune or inflammation responses.
- Some vectors can allow a transgene to be inserted into the chromosomes of a cell. That is highly desirable when targeting short-lived cells that are replenished by stem cells; if a therapeutic transgene integrates into the stem cell, the stem cell will continue to divide and produce differentiated cells with the desired transgene.
- In gene augmentation therapy diseased cells that are genetically deficient for some product are supplemented by transfecting a transgene with a coding DNA sequence to make the missing product inside the cells.
- *Ex vivo* gene therapy involves removing cells from a patient, genetically modifying them in culture and returning the genetically-modified autologous cells to the patient. It has been focused on treating disorders by genetic modification of impure populations of hematopoietic stem cells (which give rise to all blood cells, and some tissue cells, notably tissue macrophages and dendritic cells).
- Some therapies target RNA (RNA therapeutics). In gene silencing the expression of a positively harmful gene (such as a gene with a gain-of-function mutation or one expressed by a pathogen) is selectively repressed, usually via RNA interference. To counteract disease RNAs can sometimes also be induced to undergo alterative splicing by using antisense oligonucleotides to selectively blockade specific exon–intron boundaries.
- Genome editing depends on making RNA or protein guide sequences and DNA-cleaving protein domains in intact cells. The guide sequences, covalently or noncovalently linked to the DNA-cleaving domain, bind to a predetermined specific target sequence in the genome, and the attached DNA-cleaving domains make a double-strand cut at that DNA site. The double-stranded break is imperfectly repaired, either naturally or with artificial intervention, resulting in a desired sequence change.
- Therapeutic genome editing is expected to be mostly involved in repairing defective genes in autologous cell preparations enriched in stem cells but may also be used to inactivate genes to render host cells resistant to viruses.
- In human gene therapy trials, genetic modification has been limited to somatic cells only (where the direct consequences are limited to the patient).
- Germ-line gene therapy involving genetic modification of the nuclear genome is widely banned, but mitochondrial replacement therapy is a specialized type of germ-line gene therapy that has been legalized in the UK to prevent transmission of severe mtDNA disorders. A woman donor provides oocytes with healthy mitochondria and normal mtDNA to replace the damaged mitochondria that would otherwise be transmitted by a woman carrying pathogenic mtDNA.

FURTHER READING

General overviews of genetic disease treatment

Dietz H (2010) New therapeutic approaches to mendelian disorders. *N Engl J Med* **363**:852–863; PMID 20818846.

Treacy EP *et al*. (2008) Treatment of genetic disease. In: *The Metabolic and Molecular Bases of Inherited Disease*. Scriver CR *et al*. (eds.), 8th edn, pp. 175–191.

Therapeutic antibodies and proteins

Bruggemann M *et al*. (2015) Human antibody production in transgenic animals. *Arch Immunol Ther Exp* **63**:101–108; PMID 25467949.

Cardinale A, Biocca S (2008) The potential of intracellular antibodies for therapeutic targeting of protein-misfolding diseases. *Trends Mol Med* **14**:373–380; PMID 18693139.

Dimitrov DS (2012) Therapeutic proteins. *Meth Mol Biol* **899**:1–26; PMID 22735943.

Lee E-C *et al*. (2014) Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. *Nature Biotechnol* **32**:356–363; PMID 24633243.

Rodgers KR, Chou RC (2016) Therapeutic monoclonal antibodies and derivatives: Historical perspectives and future directions. *Biotechnol Adv* **34**:1149–1158; PMID 27460206.

Gene therapy general

Kumar SR *et al*. (2016) Clinical development of gene therapy: results and lessons from recent successes. *Mol Therap Meth Clin Dev* **3**:16034; PMID 27257611.

Naldini L (2015) Gene therapy returns to centre stage. *Nature* **526**:351–360; PMID 26469046.

Gene therapy: gene transfer technology

Kotterman MA *et al*. (2015) Virus vectors for gene therapy: translational and clinical outlook. *Annu Rev Biomed Eng* **17**:63–89; 26643018.

Papapetrou EP, Schambach A (2016) Gene insertion into genomic safe harbors for human gene therapy. *Mol Therap* **24**:678–684; PMID 26867951.

Ramamoorth M, Narvekar A (2015) Non-viral vectors in gene therapy—an overview. *J Clin Diagn Res* **9**:1–6; PMID 25738007.

Zatsepin TS *et al*. (2016) Lipid nanoparticles for targeted siRNA delivery—going from bench to bedside. *Int J Nanomed* **11**:3077–3086; PMID 27462152.

Augmentation gene therapy for specific diseases

Booth C *et al*. (2016) Treating immunodeficiency through HSC gene therapy. *Trends Mol Med* **22**:317–327; PMID 26993219.

Cavazzano-Calvo M *et al*. (2010) Transfusion-independence and HGMA2 activation after gene therapy of human β-thalassemia. *Nature* **467**:318–322; PMID 20844535.

Maguire AM *et al*. (2008) Safety and efficacy of gene transfer for Leber's congenital amaurosis. *N Eng J Med* **358**:2240–2248; PMID 18441370.

Nathwani AC *et al*. (2011) Adenovirus-associated virus vector-mediated gene transfer in hemophilia B. *N Eng J Med* **365**:2357–2365; PMID 22149959.

Clinical trials databases

ClinicalTrials.gov (comprehensive general clinical trials website at www.clinicaltrials.gov)

Wiley database of world-wide gene therapy clinical trials: http://www.abedia.com/wiley

Therapeutic gene silencing and modulated splicing

Bobbin ML, Rossi JJ (2016) RNA interference (RNAi)-based therapeutics: delivering on the promise? *Annu Rev Pharmacol Toxicol* **56**:103–122; PMID 26738473.

Spitali P, Aartsma-Rus A (2012) Splice modulating therapies for human disease. *Cell* **148**:1085–1088; PMID 22424220.

Yu D *et al*. (2012) Single-stranded RNAs use RNAi to potently and allele-selectively inhibit mutant huntingtin expression. *Cell* **150**:895–908; PMID 22939619.

Therapeutic genome editing and prospects for curative HIV therapy

Allers K, Schneider T (2015) CCR5Δ32 mutation and HIV infection: basis for curative HIV therapy. *Curr Opin Virol* **14**:24–29; PMID 26143158.

Cox DB *et al*. (2015) Therapeutic genome editing: prospects and challenges. *Nature Med* **21**:121–131; PMID 25654603.

Gu W-G (2015) Genome editing-based HIV therapies. *Trends Biotechnol* **33**:172–179; PMID 25600622.

Porteus M (2016) Genome editing: a new approach to human therapeutics. *Annu Rev Pharmacol Toxicol* **56**:163–190; PMID 26566154.

Preventing transmission of mtDNA disorders

Herbert M, Turnbull D (2017) Mitochondrial donation—clearing the final regulatory hurdle in the United Kingdom. *N Eng J Med* **376**:171–173; PMID 28030773.

Wolf DP *et al*. (2015) Mitochondrial replacement therapy in reproductive medicine. *Trends Mol Med* **21**:68–76; PMID 25573721.

Glossary

3′ end: the end of a DNA or RNA strand that is linked to the rest of the chain only by carbon 5 of the sugar, not carbon 3. See **Figure 1.8**.

3′, 5′- phosphodiester bond: the link between adjacent nucleotides in DNA or RNA. See **Figure 1.5**.

5′ RACE: **r**apid **a**mplification of **c**DNA **e**nds—a technique for characterizing the ends of mRNAs, in this case the 5′ ends.

5′ end: the end of a DNA or RNA strand that is linked to the rest of the chain only by carbon 3 of the sugar, not carbon 5. See **Figure 1.8**.

Acentric: of a chromosome, lacking a centromere.

ACCE framework: a framework for evaluating a test. See Section 20.3.

a-CGH: array-comparative genomic hybridization. See **Figure 15.6**.

Acute transforming retrovirus: a retrovirus that has accidentally incorporated and activated a cellular (proto) oncogene, enabling it to transform cells in culture. Replication-defective because of loss of normal viral sequences.

Adaptive immune system: the immune system of vertebrates that creates immunological memory.

Adaptor (or linker) oligonucleotides: short oligonucleotides ligated to the ends of a heterogeneous collection of DNA molecules to allow them all to be PCR-amplified using a single pair of primers. See **Figure 6.10**.

Affected sib pair (ASP) analysis: a form of model-free linkage analysis based on measuring haplotype sharing by sibs who both have the same disease. See **Figure 18.2**.

Affinity tag: in genetic manipulation, a short peptide that is attached to a recombinant protein in order to allow the protein to be isolated by affinity chromatography.

Alleles: alternative forms of the same gene.

Allele dropout: absence in a PCR product of an allele present in the original sample.

Allele frequency: the proportion of all alleles at a locus that are the allele in question.

Allele-specific PCR: using primers designed to amplify just one allele of a SNP. See **Figure 20.2**.

Allelic heterogeneity: the existence of many different mutations, but all within the same gene, in unrelated people with the same phenotype.

Allogeneic: describing tissues or cells that are genetically dissimilar, as is normally the case for donor and recipient in organ transplantation. cf. Isogenic, Autologous.

Amino acid: the building blocks of proteins. See **Figure 1.4** for names and formulae of the 20 amino acids used in natural proteins.

Amplicon: a PCR-amplified DNA sequence.

Amplification: an increase in the copy number of a DNA sequence. May occur naturally by evolution. Also achieved by PCR, isothermal amplification, or DNA cloning.

Anaphase lag: loss of a chromosome because it moves too slowly at anaphase to get incorporated into a daughter nucleus.

Ancestral chromosome segments, shared: chromosomal segments that are shared by apparently unrelated people because they are inherited from an unknown distant common ancestor. See **Figure 12.5**, **Table 12.2**.

Ancient DNA (aDNA): DNA recovered from archeological specimens.

Aneuploid: of a cell, having one or more chromosomes extra or missing from the normal full euploid set.

Annealing: allowing two complementary single-stranded nucleic acids to form a base-paired double strand. The reverse of denaturation.

Antibody: a protein produced by activated B cells in response to a foreign molecule or microorganism. See **Figure 3.24**, **Table 3.6**.

Anticipation: the tendency for the severity of a condition to increase in successive generations. Commonly due to bias of ascertainment (see Section 5.2) but seen for real with dynamic mutations (Section 16.3).

Anticodon: the 3-base sequence in a tRNA molecule that base-pairs with the codon in mRNA.

Antigen: a molecule that can induce an adaptive immune response or that can bind to an antibody or T-cell receptor.

Antiparallel: of the strands in a double-stranded nucleic acid molecule, running in opposite directions so that where one strand has its 5′ end the complementary strand has its 3′ end.

Antisense RNA: a transcript complementary to a normal mRNA. Naturally occurring antisense RNAs, made using the non-template strand of a gene, are important regulators of gene expression. Synthetic antisense RNAs are used for gene silencing.

Antisense strand (template strand): the DNA strand of a gene, which, during transcription, is used as a template by RNA polymerase for synthesis of mRNA. See **Figure 1.15**.

Antisense technology: experimental inhibition of expression of a gene by use of an RNA or modified oligonucleotide complementary to the mRNA of the gene.

Apoptosis: a common type of programmed cell death.

Aptamers: single-stranded oligonucleotides or peptides designed to bind specifically to an antigen, mimicking antibodies.

Archaea: single-celled prokaryotes superficially resembling bacteria, but with molecular features indicative of a third kingdom of life.

ARMS (amplification refractory mutation system): allele-specific PCR. See **Figure 20.2**.

Array CGH (comparative genomic hybridization): competitive hybridization of a test and control sample to a microarray of mapped clones to detect copy number variations. See **Figures 15.6**, **15.7**.

ASOs (allele-specific oligonucleotides): under stringent hybridization conditions, oligonucleotides 15–20 nt long will hybridize only to a perfectly matched target. This provides the basis for various methods of distinguishing alleles that differ by only a single nucleotide. See, for example, **Figure 6.13**.

Asymmetric cell division: a cell division where the two daughter cells have different fates, as when a stem cell produces one daughter stem cell and a more specialized daughter cell. See **Box 2.2**.

Association: a tendency of two characters (diseases, marker alleles, etc.) to occur together at nonrandom frequencies. Association is a simple statistical observation, not a genetic phenomenon, but can sometimes be caused by linkage disequilibrium. See Section 18.3.

Assortative mating: mating where the partner is chosen on the basis of phenotypic or genotypic similarity (e.g. tall people tend to marry tall people, deaf people tend to marry deaf people; some people prefer to marry relatives). Assortative mating can produce a non-Hardy–Weinberg distribution of genotypes in a population. See Section 12.4.

ATAC (assay for transposase accessible chromatin): a method for identifying DNA within the interphase cell nucleus that is accessible by regulatory molecules. See Section 10.1.

Augmentation therapy: therapy that seeks to overcome a genetic deficiency by providing some agent (such as a missing gene product or cells expressing such) that compensates for the deficiency.

Autocatalytic introns: self-splicing introns that do not depend on the spliceosome.

Autoimmunity: an abnormal state in which the distinction between self and non-self fails, so that the body mounts an adaptive immune response against one or more self molecules.

Autologous: describing cells or tissues that were obtained from or pertain to the same individual.

Autophagy: digestion of worn-out organelles by a cell's own lysosomes.

Autoradiography: using photographic film to make a radiolabeled molecule reveal its location on a gel or in a cell.

Autosome: any chromosome other than the sex chromosomes, X and Y.

Autozygosity: in an inbred person, homozygosity for alleles identical by descent. See Section 17.2.

B cells (B lymphocytes): the lymphocytes that secrete antibodies when mature. Immature B lymphoblasts make immunoglobulins that remain on the cell membrane as primary components of B-cell receptors.

Bacterial artificial chromosome (BAC): a cloning vector in which inserts up to 300 kb long can be propagated in bacterial cells. See **Box 7.1**.

Bacteriophage (phage): a virus that infects bacteria. Modified phages are used as vectors for cloning in bacterial cells.

Balanced: of a structural rearrangement, involving no loss or gain of material (also used for Robertsonian translocations, where the loss has no effect).

Balancing (overdominant) selection: selection in favor of heterozygotes, at the expense of both homozygous genotypes. Also called heterozygote advantage.

Banding: in a preparation of chromosomes, treatments to make the chromosomes stain in a reproducible pattern of dark and light bands to aid identification of chromosomes and detection of structural abnormalities. See Section 15.1.

Barr body: the chromatin of an inactive X chromosome, seen as a blob of condensed chromatin at the edge of the nucleus of interphase cells that contain one or more inactive X chromosomes. Also called sex chromatin. See **Figure 10.13**.

Basal lamina: thin mat of extracellular matrix that separates epithelial sheets and many types of cell from the underlying connective tissue.

Basal transcription apparatus: the multiprotein complex that is required for RNA polymerase II to transcribe any gene. Additional, tissue-specific, proteins are usually also needed. See Section 1.3.

Base complementarity: the relationship between bases on opposite strands of a double-stranded nucleic acid that enables stable base pairing. The dominant base pairing rules are: A opposite T (or U in RNA) and G opposite C, but in RNA G:U base pairs can also be found. cf. Complementary sequences.

Base cross-linking: a type of base modification where abnormal covalent bonds form between bases on the same strand or between bases on complementary strands.

Base pair: the unit of length of a double-stranded nucleic acid. Also, more narrowly, a purine base hydrogen-bonded to a pyrimidine base on opposite strands of a double-stranded nucleic acid. See **Figure 1.7**.

Bayesian statistics: a method of combining a number of independent probabilities. It forms the basis of much genetic risk estimation. See **Box 17.2**.

Benign: (1) of a variant, not causing disease; (2) of a tumor, not invasive.

Bias of ascertainment: distorted proportions of phenotypes in a dataset caused by the way cases are collected. See Section 5.2.

Biometrics: the statistical study of quantitative characters.

Biotin–streptavidin system: a tool for isolating labeled molecules. The bacterial protein streptavidin binds the B-vitamin biotin with exceptionally high affinity. Biotinylated molecules can be isolated using streptavidin-coated magnetic beads. See **Box 6.2**.

Bisulfite sequencing: a method for identifying methylated cytosines in a DNA sample. Sodium bisulfite converts unmethylated cytosines, but not methylated cytosines, to uracil. When the product is sequenced, cytosines that were originally methylated are still read as cytosines, but those that were unmethylated are read as thymine. See **Box 10.3**.

Bivalent: in cytogenetics, the four-stranded structure seen in prophase I of meiosis, comprising two synapsed homologous chromosomes. See **Figure 2.14**.

Blastocyst: an embryo at a very early stage of development when it consists of a hollow ball of cells with a fluid-filled internal compartment, the blastocele. See Figure 4.4.

Blastomere: one of many cells formed by cleavage of a fertilized egg, with minimal cell growth before division.

Bonferroni correction: in multiple testing, correcting the threshold of significance to take account of the number of independent tests. For n tests, the threshold is divided by n.

Bootstrapping: a statistical method designed to check the accuracy of an evolutionary tree constructed from comparative sequence analysis. See Section 13.5.

Bottleneck: an event that greatly reduces the numbers contributing to the next generation. Can apply to a population of individuals (see Figure 12.7) or of DNA sequences (e.g. mitochondrial bottleneck, see Figure 16.14).

Branch site: in mRNA processing, a rather poorly-defined sequence (consensus YNCTRAY; R = purine, Y = pyrimidine, N = any nucleotide) located 10–50 bases upstream of the splice acceptor, containing the adenosine at which the lariat splicing intermediate is formed. See Figure 1.19.

Bromodomain: a protein domain that binds to acetylated lysine residues, primarily in histones.

C statistic: a measure of the performance of a test. See Box 20.3.

C-value paradox: the lack of a direct relationship between the amount of DNA in the cells of an organism (the C value) and the complexity of the organism.

CAAT box: a short sequence, GGCCAATCT or a close variant that is found in the promoter of many genes that are transcribed by RNA polymerase II.

CAGE (cap analysis of gene expression): a high-throughput technique for cataloging bulk mRNAs by isolating around 18 nucleotides adjacent to the 5′ cap for sequencing.

Capping: a stage in RNA processing, addition of a special nucleotide, 7-methylguanosine triphosphate, by a 5′–5′ bond to the 5′ end of a primary transcript. Capping is important for the stability of the RNA. See Figure 1.22.

cDNA: complementary DNA—a DNA copy of an RNA, made by reverse transcriptase.

Cell cortex: the network of microfilaments lying beneath the plasma membrane of most cells.

Cell cycle: the reproductive cycle of a cell, comprising mitosis (M phase), the first gap (G1 phase), DNA synthesis (S phase), a second gap (G2 phase), then mitosis of the next cycle. See Figure 2.9.

Cell autonomous: pertaining to a trait in a multicellular organism in which only genotypically mutant cells exhibit the mutant phenotype.

Cell differentiation: the process by which a less specialized cell becomes more specialized.

Cell junctions: specialized structures that control the passage of molecules between cells. They may form totally impermeable barriers, or they may allow specific types or sizes of molecule to pass. See Figure 3.11.

Cell lineage: in development, the ancestry and descendants of a cell, as traced backwards or forwards through successive cell divisions. See Figure 7.17.

Cell polarity: asymmetry of a cell. In embryos, asymmetric cells often divide to form daughter cells that follow different fates in development.

CentiMorgan (cM): the unit of genetic distance. Loci 1 cM apart have a 1% probability of recombination during meiosis.

Central dogma: the unidirectional information flow of DNA → RNA → protein—that is, the DNA specifies the nucleotide sequence of an RNA, which in turn specifies the amino acid sequence of a protein. Not strictly true because of natural reverse transcription in RNA viruses and cells.

Centriole: a cylinder of short microtubules located in the centrosome. See Box 2.3.

Centromere: the primary constriction of a chromosome, separating the short arm from the long arm, and the point at which spindle fibers attach to pull chromatids apart during cell division.

Centrosome: in cell division, the microtubule organizing center that forms a spindle pole. See Box 2.3.

CEPH families: a set of families assembled by the Centre d'Etude du Polymorphisme Humain in Paris to assist the production of marker–marker framework maps.

CGH: see Comparative genomic hybridization.

Character: an observable property, phenotype, or trait of an individual.

Checkpoint: a biochemical quality control check that prevents further progress through the cell cycle unless the genome and the cell are adjudged to be in a suitable state to proceed. See Figure 19.12.

Chiasma (plural chiasmata): the physical manifestation of meiotic recombination, as seen under the microscope.

Chimera: (1) an organism derived from more than one zygote. See Figures 4.18, 5.16; (2) a chimeric gene is a gene created when a chromosomal rearrangement brings together parts of two different genes to create a novel functional gene—a frequent event in tumors. See Figure 19.6.

Chromatin: a general term for the packaged DNA in a cell nucleus. The basic conformation is a 30 nm coiled coil of DNA and histones. See Figure 2.19.

Chromatin flavors: combinations of histone modifications that help define the activity of a region of chromatin. See Figure 10.4.

Chromatin immunoprecipitation (ChIP): a technique for identifying the DNA sequences that bind a specific protein. Protein and DNA are reversibly cross-linked, the chosen protein is precipitated with an antibody, and the associated DNA is sequenced. See Box 9.3.

Chromatin remodeling complexes: protein complexes that can move, dissociate, or reconstitute nucleosomes in chromatin, as part of the systems controlling chromatin conformation. See Figure 10.2.

Chromodomain: a protein domain that stimulates binding to methylated lysine residues, primarily in histones.

Chromoplexy: in a tumor, multiple "chained" chromosomal rearrangements. See Section 19.4.

Chromothripsis: an abnormal event in which a chromosomal segment appears to have been pulverized and the fragments reassembled in random order. Seen mainly in cancer cells.

Chromosome conformation capture: a set of techniques (3C, 4C, 5C, HiC) for identifying DNA sequences that lie close together in interphase nuclei. See Box 10.1 for details.

Chromosome engineering: genetic engineering techniques to produce specific large-scale chromosomal deletions or rearrangements. See Figure 8.14B.

Chromosome painting: fluorescence labeling of a whole chromosome by a FISH procedure in which the probe is a cocktail of many different DNA sequences from that particular chromosome. See Figure 15.5.

Chromosome set: the chromosomes of a haploid genome.

***Cis*-acting (of a regulatory factor):** controlling the activity of a gene only when it is part of the same DNA molecule or chromosome as the regulatory factor. Compare *trans*-acting regulatory factors, which can control their target sequences irrespective of their chromosome location.

Cladogram: a rooted evolutionary tree.

Clinical exome: a list of around 3000 genes that have been reported as having variants that cause Mendelian conditions.

Clonal selection and expansion: the process responsible for immunological memory. Binding of an antigen to a B- or T cell stimulates it to multiply, forming a clone of cells that react to the same antigen. However, clones that respond to self-antigens are eliminated. See **Figure 3.23**.

Clones: identical copies (of a DNA sequence, a cell, an organism). In genetic research, often means cells containing identical recombinant DNA molecules (the cells themselves may or may not be identical).

Clone fingerprinting: identifying independent clones that contain overlapping inserts by comparing the pattern of fragments produced by a series of restriction enzymes.

Cloning: production of many identical copies of a DNA sequence, a cell, or a whole organism.

Co-activators: proteins that enhance transcription of a gene through protein–protein rather than protein–DNA interactions.

Coding RNA: messenger RNA that codes for protein.

Codon: a nucleotide triplet (strictly in mRNA, but by extension, in genomic coding DNA) that specifies an amino acid or a translation stop signal.

Coefficient of inbreeding: the proportion of loci at which a person is homozygous by virtue of the consanguinity of their parents. See Section 12.4.

Coefficient of relationship: of two people, the proportion of loci at which they share alleles identical by descent. See Section 12.4.

Coefficient of selection: the chance of reproductive failure for a certain genotype, relative to the most successful genotype. See Section 12.3.

Cofactor: a small molecule or metal ion that is required for the biological activity of a protein. More generally, any factor that assists the principal factor in a process.

Co-immunoprecipitation (co-IP): using an antibody to precipitate a known molecule complexed with its binding partners, as a way of identifying the binding partners.

Common variant: conventionally, a variant whose frequency is >0.05.

Compaction: in embryogenesis, the tightening of cell–cell contacts that converts the loosely bound products of the initial cleavage divisions of the zygote into a compact morula. See **Figure 4.4**, **Box 4.2**.

Companion diagnostic: a diagnostic test used when a drug is prescribed to check for genetic variants that influence its safety or efficacy.

Comparative genomic hybridization: competitive hybridization of a test and control sample, normally to a microarray of mapped clones (but originally to a spread of normal chromosomes). Employed to detect chromosomal regions in the test sample that are amplified or deleted compared to the control sample.

Comparative genomics: a systematic and comprehensive comparison of the genomes of different organisms.

Compensated pathogenic deviation: occurs when a loss of function due to one variant is rescued by another change in the same gene. A cause of false-negative pathogenicity scores. See **Figure 17.14**.

Complement system: a system of serum proteins activated by antigen–antibody complexes or by microorganisms. See **Figures 3.20**, **3.21**.

Complementarity determining region: a region having hypervariable sequences located near the N-terminal ends of heavy and light immunoglobulin chains that are responsible for antigen recognition and binding. See **Figure 22.3**.

Complementary: of two nucleic acid strands, having the ability to form sufficient base pairs to establish a stable duplex. Also used to describe the sequences of complementary strands.

Complementation: two alleles complement if in combination they restore the wild-type. Normally alleles complement only if they are at different loci, although some cases of interallelic complementation occur.

Complex: of a phenotype, one that can have a variety of different causes and modes of inheritance in different people.

Compound heterozygote: a person with two different mutant alleles at a locus.

Concatemers: molecules joined end-to-end in a chain.

Concordance: of twins, the frequency with which co-twins have the same phenotype.

Conformation: of a complex molecule, the 3-dimensional shape—the result of the combined effects of many weak noncovalent bonds.

Congenic strains: in experimental animals, two strains that have identical genetic backgrounds and differ only in some chosen gene or location.

Congenital: of a character, present at birth.

Consanguinity: where two individuals, normally mating partners, share one or more identifiable recent common ancestors. See **Figure 12.12**.

Consensus sequence: a representation of the main shared elements of a family of functionally related DNA sequences.

Conservative change: in a protein, replacement of one amino acid by a chemically similar one.

Conserved sequence: a sequence (of DNA or sometimes protein) that is identical or recognizably similar across a range of organisms.

Constant region (of an immunoglobulin): the C-terminal portion of an Ig protein that is comparatively conserved in sequence (see **Figure 3.24**). It is largely comprised of globular domains (constant domains), with three such domains in the heavy chain, and one in the light chain (**Figures 3.25**, **22.3**).

Constitutional abnormality: an abnormality that was present in the zygote, and so is present in every cell of a person.

Constitutive heterochromatin: heterochromatin that remains condensed throughout the cell cycle. It is found at centromeres plus some other regions. See Section 9.3.

Contig: a clone contig is an ordered set of overlapping clones. See **Figure 7.3**.

Contiguous gene syndrome: a syndrome that is the result of a deletion that inactivates two or more contiguous genes, each of which contributes to the phenotype. See **Table 15.4** for some examples.

Continuous character: a character such as height, which everybody has, but to differing degree—as compared with a dichotomous character like polydactyly, which some people have and others do not.

Copy number variant (CNV): variation between individuals in the number of copies of a particular DNA sequence in their genomes. Normally used only for relatively large changes (e.g. >50 nucleotides). See **Figure 11.9**.

Co-repressors: proteins that suppress transcription of a gene through protein–protein rather than protein–DNA interactions.

COS cells: artificial cells derived from African Green Monkey kidney cells that permit replication of any DNA containing an SV40 origin of replication.

Cosmid: a vector for cloning in *E. coli*. Cosmids have an insert capacity of up to 44 kb. See **Table 6.1**.

Cousin: A and B are (first) cousins if one of A's parents is the sib of one of B's parents. If both parents are sibs A and B are double first cousins. The children of A and B are second cousins. See **Figure 12.12**.

CpG: in DNA, cytosine followed by guanine (the p represents the phosphate linking them). CpG sequences occur symmetrically on both strands and are targets for DNA methylation.

CpG island: short stretch of DNA, often <1 kb, containing frequent unmethylated CpG dinucleotides. CpG islands tend to mark the 5' prime ends of genes.

Cre-*lox*P recombination: natural site-specific recombination by the phage P1 Cre recombinase that cuts DNA at a specific 34 bp recognition sequence (*lox*P). Artificially employed to engineer desired chromosomal deletions, inversions, or translocations in cells or animal models (see **Figure 8.14**), or to allow tissue- or stage-specific knockout of a gene in an intact animal.

CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR associated): a type of natural prokaryotic adaptive immunity. Adapted as a genome editing technique that uses artificial RNA guide sequences. See **Figure 8.17**.

Crossover: an act of meiotic recombination, or the physical manifestation of that, as seen under the microscope.

Cryptic splice site: a sequence in pre-mRNA with some homology to a splice site, that may be used as a splice site when splicing is disturbed or after a base substitution mutation that increases the resemblance to a normal splice site. See **Figure 16.4**, **Table 16.3**.

Cytogenetics: the study of chromosomes.

Cytokines: extracellular signaling proteins or peptides that act as local mediators in cell–cell communication.

Cytokinesis: the final event of cell division, when the cytoplasm of the cell divides. See **Figure 2.11**.

Cytoskeleton: the internal scaffold of protein filaments in a cell.

Cytosol: the contents of the cytoplasm of a cell, excluding membrane-bound organelles such as mitochondria or lysosomes.

***De novo* mutation:** a mutation that is present in an individual but not in DNA samples obtained from their parents. See **Figures 5.17**, **5.18**, **Box 11.2**.

Degenerate: used in genetics to describe a many-to-one relation between structure and function. The genetic code is degenerate because most amino acids can be incorporated into a polypeptide in response to any of several different codons in the mRNA. An oligonucleotide is degenerate if it is a mixture of several related sequences.

Denaturation: dissociation of double-stranded nucleic acid to give single strands. Also destruction of the 3-dimensional structure of a protein by heat or high pH.

Denisovans: an enigmatic extinct species of human, named after Denisova cave in Siberia where their remains have been found. See **Figure 14.13**.

Dichotomous character: a character such as polydactyly, which some people have and others do not have—as compared to a continuous character like height, which everybody has, but to differing degree.

Dideoxy (Sanger) sequencing: the standard method of DNA sequencing, developed by Fred Sanger and using dideoxynucleotide chain terminators. See Section 6.4.

Diploid: having two copies of each type of chromosome; the normal constitution of most human somatic cells.

Distal (of chromosome): positioned comparatively distant from the centromere.

Distance matrix: for constructing an evolutionary tree, the distance matrix shows the evolutionary distance between each pair of organisms involved. See **Figure 13.30**.

Disulfide (S–S) bridge or bond: in proteins, an intramolecular or intermolecular link between the SH groups of two cysteine residues. Important for maintaining the 3-dimensional folding of proteins. See **Figure 1.35**.

Divergence time: in evolutionary studies, the time when the ancestors of two current species became separate.

Dizygotic: of twins, resulting from the independent fertilization of two eggs. See **Box 4.3**.

DNA chip: a high-density microarray carrying oligonucleotides or longer single-stranded DNA molecules.

DNA cloning: making identical copies of a DNA molecule.

DNA fingerprinting: a now obsolete method of identifying a person for legal or forensic purposes based on probing Southern blots with a hypervariable minisatellite probe. See **Figure 20.17**. cf. DNA profiling.

DNA libraries: a collection of many different DNA molecules, the result of cloning or amplifying random fragments of genomic DNA or cDNA.

DNA methylation: conversion of cytosine in DNA to 5-methylcytosine, a signal that helps regulate gene expression. See Section 10.3.

DNA polymerases: the family of enzymes that can add nucleotides to the 3' end of a DNA molecule.

DNA profiling: using genotypes at a series of polymorphic loci to recognize a person, usually for legal or forensic purposes. See Section 20.6.

DNA repair: correcting lesions in DNA caused by mistakes during replication or by external agents such as radiation or chemicals.

DNA replication: the process of copying a DNA molecule to make two identical daughter molecules.

DNAse I-hypersensitive sites: regions of chromatin that are rapidly digested by DNAse I because the DNA is relatively exposed rather than being tightly packaged in nucleosomes. They are believed to mark important long-range control sequences. See Section 10.1.

Dominant: in human genetics, any trait that is expressed in a heterozygote. See also Semi-dominant.

Dominant negative effect: the situation where a mutant protein interferes with the function of its normal counterpart in a heterozygous person. See **Figure 16.8**.

Double helix: the normal structure of DNA, with two antiparallel DNA strands wrapped round one another. Also describes stable double-stranded RNA genomes in certain RNA viruses, and regions where there is transient DNA–RNA base pairing.

Downstream: in the 3' direction on the sense strand.

Driver mutations: in cancer, mutations that are subject to positive selection during tumorigenesis because they assist development of the tumor (cf. Passenger mutations).

Droplet digital PCR: a method for detecting very low-level mosaicism. See **Figure 5.19**.

Duplex: any double-stranded nucleic acid.

Dynamic mutation: an unstable expanded repeat that changes size between parent and child. See Section 16.3.

Ectoderm: one of the three germ layers of the embryo. Formed during gastrulation from cells of the epiblast, it gives rise to the nervous system and outer epithelia. See Figure 4.9.

Effective population size (N_e): the size of a population that would feature the same level of genetic drift as the actual population under consideration. See Box 14.2.

Electroporation: a method of transferring DNA into cells *in vitro* by use of a brief high-voltage pulse.

Elongation factors: factors that assist progression of RNA polymerase along a DNA sequence once transcription has been initiated.

Embryonic germ cells: pluripotent cells derived from cultured primordial germ cells of embryos.

Embryonic stem cells (ESCs): artificial pluripotent cells derived from cells of the inner cell mass of a blastocyst. See Figure 8.22.

Empiric risks: risks calculated from survey data rather than from genetic theory. Genetic counseling in most non-Mendelian conditions is based on empiric risks.

ENCODE (Encyclopedia of DNA elements) project: a large collaborative effort to identify all functional elements in the human genome. See Section 9.4.

Endocytosis: the process in which a portion of the plasma membrane of a cell invaginates to form a pit and then pinches off to form an intracellular vesicle enclosing some of the extracellular fluid.

Endoderm: one of the three germ layers of the embryo. It is formed during gastrulation from cells migrating out of the epiblast layer. See Figure 4.9.

Endonuclease: an enzyme that cuts DNA or RNA at an internal position in the chain.

Endophenotype: a phenotype of a complex disease that is hopefully closely related to the underlying biology.

Endoplasmic reticulum: a meshwork of membranes in the cytoplasm of cells, forming a compartment where membrane-bound and secreted proteins are made.

Endosymbiosis: one cell engulfs another without destroying it, so that both contribute to the overall function. See Figure 2.5.

Enhancer: a set of short sequence elements that stimulate transcription of a gene and whose function is not critically dependent on their precise position or orientation. See Figure 10.24.

Epiblast: the layer of cells in the pre-gastrulation embryo that will give rise to all three germ layers of the embryo proper, plus the extra-embryonic ectoderm and mesoderm. See Figure 4.4. cf. Hypoblast.

Epigenetic: heritable (from mother cell to daughter cell, or sometimes from parent to child), but not produced by a change in DNA sequence. DNA methylation is the best understood epigenetic mechanism.

Epigenetic marks: molecular tags attached to DNA or histones that act as signals for epigenetic mechanisms. See Sections 10.2, 10.3.

Epigenetic memory: the phenomenon of epigenetic marks being "remembered" from mother cell to daughter cells across mitosis. Sometimes the memory occurs between generations, across meiosis, but this is unusual and controversial.

Episome: any DNA sequence that can exist in an autonomous extra-chromosomal form in the cell. Often used to describe self-replicating and extra-chromosomal forms of DNA.

Epistasis: literally "standing above". Gene A is epistatic to gene B if A functions upstream of B in a common pathway. Loss of function of A will cause all the effects of loss of function of B, and maybe other effects as well.

Epitope: the part of an immunogenic molecule to which an antibody responds.

Epitope tagging: a method for visualizing a specific protein in cells or tissues. A recombinant version of the protein is produced that has a marker peptide attached, for which a fluorescently-labeled antibody is available.

Euchromatin: the fraction of the nuclear genome that contains transcriptionally active DNA and which, unlike heterochromatin, adopts a relatively extended conformation. See Figure 2.19.

Eukaryotes: organisms made of cells with a membrane-bound nucleus and other organelles (see Box 2.1). One of the three kingdoms of life.

Euploid: of a cell, having one or more complete sets of chromosomes. cf. Aneuploid.

Exaptation: an unusual evolutionary process in which sequences derived from a transposable element are used by the host genome for a novel function. See Figure 13.27.

Exome: the totality of exons in a genome.

Exon: a segment of a gene that continues to be represented in RNA transcripts after splicing. Individual exons may contain coding and/or noncoding DNA (untranslated sequences). See Figure 1.18.

Exon junction complex: a set of proteins that are bound to mRNAs during splicing, at the positions where introns have been removed. See Figure 16.7.

Exonization: in evolution, formation of a new exon from noncoding DNA sequence often supplied by a transposable element.

Exonuclease: an enzyme that digests DNA or RNA from one end. May be a 3' prime or a 5' prime exonuclease.

Expressed sequence tag (EST): short partial sequences of cDNAs that can be used to follow gene expression or to isolate a full-length cDNA.

Expression array: a microarray of probes for expressed sequences, used to analyze the pattern of gene expression in a given cell type or tissue by hybridizing to labeled cDNA from the cell or tissue. See Boxes 7.7, 17.21.

Expression cloning: cloning in vectors that are deigned to allow genes in the insert to be expressed. Used to make purified gene product. See Figure 6.5.

Extracellular matrix: a meshwork of polysaccharide and protein molecules found within the extracellular space and in association with the basement membrane of the cell surface. It provides a scaffold to which cells adhere and serves to promote cellular proliferation.

Facultative heterochromatin: heterochromatin that may reversibly decondense to form euchromatin, depending on the requirements of the cell.

FAIRE (formaldehyde-assisted isolation of regulatory elements): a method of identifying DNA in the interphase cell nucleus that is accessible by regulatory proteins. See Section 10.1.

First-degree relatives: parents, children, or sibs.

Fitness (f): in population genetics, a measure of the success in transmitting genotypes to the next generation, relative to the most successful genotype. Also called biological or reproductive fitness, f is unrelated to fitness in the athletic sense. f always lies between 0 and 1.

Fluorescence *in situ* hybridization: *in situ* hybridization using a fluorescently-labeled DNA or RNA probe. A key technique in modern molecular genetics. See Figures 15.3, 15.4.

Fluorophore: a fluorescent chemical group, used for labeling nucleic acids or proteins. See Box 6.2.

Founder effect: high frequency of a particular allele in a population because the population is derived from a small number of founders, one or more of whom carried that allele. See Figure 12.7.

Fragile sites: locations on chromosomes where, under special culture conditions, the chromatin of metaphase chromosomes appears uncondensed. Most are nonpathogenic variants present at varying frequencies in normal healthy individuals, but a few are pathogenic.

Frameshift mutation: a mutation that alters the triplet reading frame of a mRNA (by inserting or deleting a number of nucleotides that is not a multiple of 3). See Figures 16.5, 16.6.

Framework map: a map of the locations of some physical entities—genetic markers, sequence-tagged sites, or clones—across a genome or chromosome. Used as a step towards a full genome sequence.

Functional genomics: analysis of gene function on a large scale, by conducting parallel analyses of gene expression/function for large numbers of genes, even all genes in a genome.

Fusion protein: the product of a natural or engineered fusion gene: a single polypeptide chain containing amino acid sequences that are normally part of two or more separate polypeptides. See Figure 6.6.

G-banding: the standard method of identifying chromosomes under the microscope. See Figure 15.1.

G-value paradox: the lack of a direct relationship between the number of genes in an organism (the G value) and the complexity of the organism.

Gain-of-function mutations: mutations that cause the gene product to do something abnormal, rather than simply to lose function. Usually the gain is a change in the timing or level of expression. See Section 16.2.

Gamete: sperm or egg; a haploid cell formed when a primordial germ cell undergoes meiosis.

Gametologs: homologous functional X–Y gene pairs.

Gastrulation: conversion of the two-layer embryo (consisting of epiblast plus hypoblast) to one that contains the three germ layers: ectoderm, mesoderm, and endoderm. See Figure 4.8.

GC box: a short sequence, GGGCGG or a close variant, that is found in the promoters of many genes that are transcribed by RNA polymerase II.

Gene: (1) a segment of DNA that is transcribed to give a mRNA or a functional noncoding RNA; (2) a factor that controls or affects a phenotype and segregates in pedigrees according to Mendel's laws.

Gene conversion: a naturally occurring nonreciprocal genetic exchange in which a sequence of one DNA strand is altered so as to become identical to the sequence of another DNA strand. See Figures 15.22, 15.23.

Gene expression: production of the gene product (a protein or a functional RNA).

Gene family: a set of related genes with a presumed common ancestry. See Table 9.10 for examples.

Gene frequency: the proportion of all alleles at a locus that are the allele in question. Really we mean allele frequency. See Section 12.1.

Gene knockdown: targeted inhibition of expression of a gene by, for example, using siRNA or a morpholino oligonucleotide to bind to RNA transcripts.

Gene knock-in: a targeted mutation that replaces activity of one gene by that of an introduced gene (usually an allele or a reporter gene). See Figure 8.23.

Gene knockout: the targeted inactivation of a gene within an intact cell.

Gene ontology: a formal controlled vocabulary for describing the functions of genes, as an aid to automated cross-referencing.

Gene pool: all the genes (in the whole genome or at a specified locus) in a particular population.

Gene silencing: preventing expression of a gene, usually by targeting its transcript.

Gene superfamily: a set of multiple genes and gene families that show signs of overall distant structural and functional relationships—for example the immunoglobulin and the G-protein coupled receptor superfamilies.

Gene targeting: targeted modification of a gene in a cell or organism using some type of genome editing. See Figure 8.22 for an example.

Gene therapy: treating disease by genetic modification. May involve adding a functional copy of a gene that has lost its function, inhibiting a gene showing a pathological gain of function, or more generally, replacing a defective gene.

Gene tracking: following a disease gene through a pedigree by use of linked markers rather than a direct test for the pathogenic change.

Gene trap: using random insertions of a reporter construct into genes of embryonic stem cells to generate embryos with random genes inactivated. The affected genes can be identified via the reporter.

General transcription factors: DNA-binding proteins that are always required to allow transcription to take place (as distinct from tissue-specific or stage-specific transcription factors).

Genetic background: the genotypes at all loci other than one under active investigation. Variations in genetic background are a major reason for imperfect genotype–phenotype correlations.

Genetic code: the relationship between a codon and the amino acid it specifies. See Figure 1.29.

Genetic distance: distance on a genetic map, defined by recombination fractions and the mapping function, and measured in centiMorgans. See Section 17.1.

Genetic drift: random changes in gene frequencies over generations because of random fluctuations in the proportions of the alleles in the parental population that are transmitted to offspring. Only significant in small populations. See Figure 12.6.

Genetic map: a map showing the sequence and recombination fractions between genes, based on breeding experiments or observation of human pedigrees.

Genetic marker: any character that can be used to follow the segregation of a particular chromosomal segment through a pedigree or in a population. Normally a DNA sequence polymorphism.

Genetic redundancy: partially or completely overlapping function of genes at more than one locus, so that loss of function mutations at one locus do not cause overall loss of function.

Genome: the total set of different DNA molecules of an organelle, cell, or organism. The human genome consists of 3.1 × 10^9 bp of DNA divided into 25 molecules, the mitochondrial DNA molecule plus the 24 different chromosomal DNA molecules. cf. Transcriptome, Proteome.

Genome assembly: assembling many short sequences into an overall sequence.

Genome browser: a program that provides a graphical interface for interrogating genome databases. See **Table 7.1**.

Genome defense: mechanisms that suppress transposon and retrovirus activity in germ cells, including the use of piRNAs and endogenous siRNAs working through RNA interference pathways and KRAB-ZNF proteins acting through histone modification.

Genome editing: making desired changes to a specific target sequence in the genome of cultured cells or cells that can be introduced into the germ line. Formerly done using homologous recombination between an introduced transgene and the target sequence within the genome. More recently, CRISPR-Cas-based genome editing is often used.

Genotype: a list of the alleles present in a person at one or more loci.

Genotype–phenotype correlation: the extent to which a phenotype can be predicted from a genotype. Typically poor in humans, better in experimental animals, which are inbred and live under standard laboratory conditions. See Section 16.5.

Germ line: the germ cells and those cells that give rise to them; other cells of the body constitute the soma.

Germinal (gonadal or gonosomal) mosaic: an individual who has a subset of germ-line cells carrying a mutation that is not found in other germ-line cells.

Glycolipid: a lipid molecule with a covalently attached sugar or oligosaccharide.

Glycosaminoglycans: long polysaccharide molecules made of pairs of sugar units, one of which is always an amino sugar. A major component of extracellular matrix.

Glycosylation: covalent addition of sugars, usually to a protein or lipid molecule.

Golden path: a unique clone tiling path that can be reduced to one nonredundant haploid representation of the genome.

Golgi apparatus: a membranous organelle in which proteins and lipids are modified and sorted for transport to different destinations. See **Box 2.1**.

Gonadal mosaic: see Germinal mosaic.

Great apes: chimpanzees, bonobos, gorillas, and orangutans.

GT-AG rule: the rule that almost all human introns begin with GT (GU in the RNA) and end in AG. A few follow an AT-AC rule and use a different spliceosome machine.

GWAS (genome-wide association study): the standard approach to identifying factors governing susceptibility to complex disease. See **Figure 18.5**.

Haplogroup: a set of related haplotypes.

Haploid: describing a cell (typically a gamete) that has only a single copy of each chromosome (as in the 23 different chromosomes in human sperm and eggs).

Haploinsufficiency: a locus shows haploinsufficiency if producing a normal phenotype requires more gene product than the amount produced by a single functional allele. See **Figure 16.17**.

Haplotype: a series of alleles found at linked loci on a single chromosome.

Haplotype blocks: blocks of variants that are in linkage disequilibrium with each other, but not with variants in adjacent blocks. Cataloged by the International HapMap project. The consequence of shared remote ancestry. See **Table 12.2**.

Hardy–Weinberg distribution: the simple relationship between gene frequencies and genotype frequencies that is found in a population under certain conditions. See Section 12.1.

HAT medium: a medium that allows cells to grow only if they have a functional thymidine kinase gene. See **Figure 8.11**.

Heat map: a form of data display used particularly for expression array data. A table of cells, with each row representing a gene, each column a sample, and the color of each cell representing the level of expression of that gene in that sample. See **Box 7.7**.

Helicase: a protein that acts to separate the two strands of double-stranded nucleic acid, as part of the machinery for replication, recombination, and repair.

Hematopoietic stem cell: self-renewing bone marrow cell that gives rise to all the various types of blood cell plus certain types of tissue immune system cells. See **Figure 3.17**.

Hemizygous: having only one copy of a gene or DNA sequence in diploid cells. Males are hemizygous for most genes on the sex chromosomes. Deletions occurring on one autosome produce hemizygosity in males and in females.

Heritability: the proportion of the variation in a character that is due to genetic variation. See Section 5.4.

Heterochromatin: chromatin that is highly condensed and shows little or no evidence of active gene expression (cf. Euchromatin). Heterochromatin may be constitutive or facultative (qv.). See **Figure 2.19**.

Heterochrony: describes the situation where different regulatory elements control the activity of the same gene in different developmental stages.

Heteroduplex: double-stranded DNA in which there is some mismatch between the two strands. Important in mutation detection.

Heterogametic: in organisms with a chromosomal sex determination system, the sex with two different sex chromosomes is heterogametic (XY human males, ZW female birds).

Heteroplasmy: mosaicism, usually within a single cell, for mitochondrial DNA variants. See Section 16.4.

Heterotopy: describes the situation where different regulatory elements control the activity of a single gene in different tissues.

Heterozygote: an individual is heterozygous at a particular locus if they have two different alleles.

Heterozygote advantage: the situation when somebody heterozygous for a mutation has a reproductive advantage over both homozygotes. Sometimes called overdominance. Heterozygote advantage is one reason why severe recessive diseases may remain common.

Histones: a family of small basic proteins that complex with DNA to form nucleosomes. See **Figures 2.18**, **10.3**.

Histone code: the idea that the pattern of covalent modification of histones in nucleosomes determines the activity of the DNA in the vicinity. In fact, histone modification is only one of several factors that, between them, determine gene expression. See Section 10.2.

Homeobox: a 180 bp module found in many genes that have functions in development. The products of homeobox genes regulate the expression of target genes through a 60 amino acid DNA-binding homeodomain.

Hominids: humans, great apes, and a number of their extinct ancestors.

Hominins: extinct species (some classified within the genus *Homo* and others classified within other genera, notably *Australopithecus*), that are more closely related to modern-day humans than to great apes. See **Figure 14.2**.

Hominoids: humans, great apes, and gibbons.

Horizontal gene transfer: transfer of genes from one organism to another.

Homoduplex: double-stranded DNA in which the two strands match perfectly. cf. Heteroduplex.

Homologs (chromosomes): the two copies of a chromosome in a diploid cell. Unlike sister chromatids, homologous chromosomes are not copies of each other; one was inherited from the father and the other from the mother.

Homologs (genes): two or more genes whose sequences are significantly related because of a close evolutionary relationship, either between species (orthologs) or within a species (paralogs).

Homologous recombination: (1). the normal recombination that is part of meiosis; (2) a mechanism for repairing DNA damage, see **Figure 11.6**.

Homoplasmy: of a cell or organism, having all copies of the mitochondrial DNA identical. cf. Heteroplasmy. See Section 16.4.

Homozygous: an individual is homozygous at a locus if they have two identical alleles at that locus. For clinical purposes a person is often described as homozygous if they have two normally functioning alleles or two pathogenic alleles at a locus, regardless whether the alleles are in fact completely identical at the DNA sequence level. Homozygosity for alleles identical by descent is called autozygosity.

Housekeeping gene: a gene that provides some basic aspect of cell function, common to most or all cells of an organism.

Human accelerated region (HAR): a sequence that is generally well conserved across species, but that has undergone rapid changes in the human lineage. HARs might be important for making us different from other apes.

Humanized antibodies: monoclonal antibodies made in rodent systems but genetically engineered so that all but the complementarity-determining regions are replaced by human sequence. See **Figure 22.3**.

Humanized mice: mice that have been genetically modified so that some chosen aspect of their genetics or physiology more closely resembles its human equivalent.

Hybridization: of nucleic acids, allowing complementary single strands to base-pair (anneal).

Hybridization stringency: the degree to which the conditions (temperature, salt concentration) during a hybridization assay permit sequences with some mismatches to hybridize. High stringency conditions allow only perfect matches. See **Figure 6.13**.

Hybridoma: a cell line made by fusing an antibody-producing B cell with a cell of a B-lymphocyte tumor. The source of monoclonal antibodies.

Hydrogen bond: a weak chemical bond that forms when a hydrogen atom lies in line between two atoms that may individually be oxygen, nitrogen, or fluorine (**Figure 1.7**). The basis of base-pairing in nucleic acids (see **Figure 1.7**) and crucially important in protein structure (**Figures 1.33**, **1.34**).

Hydrophilic: of a chemical group, having energetically favorable interactions with water and other polar molecules. A property of charged or polar groups.

Hydrophobic: of a chemical group, repelled by water and other polar groups. Hydrophobic groups associate together in the interior of protein molecules, membranes, etc.

Hypoblast: the layer of cells in the pre-gastrulation embryo that gives rise to extra-embryonic endoderm. See **Figure 4.4**.

Identity by descent (IBD): alleles in an individual or in two people that are known to be identical because they have both been inherited from a demonstrable common ancestor. See **Figure 12.10**.

Identity by state (IBS): alleles that appear identical, but may or may not be identical by descent because there is no demonstrable common source. See **Figure 12.10**.

Immunoblotting: using an antibody to identify proteins that have been fractionated by size and charge by electrophoresis and then transferred to a nitrocellulose membrane.

Immunogen: any molecule that elicits an immune response.

Immunological memory: the ability of the adaptive immune system to mount a rapid and strong response to an antigen that it has previously encountered.

Imprinting: in genetics, determination of the expression of a gene by its parental origin. See **Figure 10.16**, **Table 10.4** for some examples.

Imprinting control center: a short sequence within an imprinted gene cluster where differential methylation controls the imprinting status of genes within the cluster. See **Figures 10.17**, **10.18**.

Imputation: in genome-wide association studies, guessing (imputing) genotypes at loci that were not typed, using experimentally-determined genotypes and information about linkage disequilibrium.

***In situ* hybridization:** hybridization of a labeled DNA or RNA probe to an immobilized nucleic acid target. The target may be denatured DNA within a chromosome preparation (chromosome *in situ* hybridization), RNA within the cells of a tissue section on a microscope slide (tissue *in situ* hybridization), or RNA within a whole embryo (whole mount *in situ* hybridization).

Inbreeding: marrying a blood relative. The term is comparative, since ultimately everybody is related. The coefficient of inbreeding is the proportion of a person's genes that are identical by descent. See Section 12.4.

Incidental finding: in diagnostic work, a finding that is clinically relevant but unrelated to the reason for which the patient was tested.

Indel: an insertion/deletion variant. See **Figure 11.8** for discussion of subtleties.

Induced pluripotent stem cells: somatic cells that have been treated with specific genes or gene products to reprogram them to resemble pluripotent stem cells. They can then be induced to differentiate into desired cell types. A great hope for regenerative medicine.

Induction: in development, the process whereby one tissue changes the state or fate of an adjacent tissue.

Initiation codon: the AUG sequence in messenger RNA that signals the start of translation.

Innate immune system: system of nonspecific response to a pathogen using the natural defenses of the body; cf. Adaptive immune system.

Inner cell mass: a group of cells located internally within the blastocyst that will give rise to the embryo proper.

Insertional mutagenesis: mutation (usually causing abolition of function) of a gene by insertion of an unrelated DNA sequence into a gene.

Insulators: DNA elements that act as barriers to the spread of chromatin changes or the influence of *cis*-acting elements. Insulators are usually binding sites for the CTCF protein.

Interference: in meiosis, the tendency of one crossover to inhibit further crossing over within the same region of the chromosomes.

Interphase: all the time in the cell cycle when a cell is not dividing.

Interphase FISH: fluorescence *in situ* hybridization of a probe to interphase cell nuclei. Used to detect aneuploidies or other chromosomal abnormalities without the need to culture cells, or to examine the subnuclear localization of chromosomes in nondividing cells. See **Figure 15.4**.

Intrabodies: engineered nonsecreted intracellular antibodies that can be used to inactivate selected molecules inside a cell.

Intracytoplasmic sperm injection (ICSI): a human infertility treatment in which sperm heads are injected into unfertilized eggs. Sometimes used experimentally to make transgenic animals by first coating the sperm head with the desired transgene DNA.

Intron: traditionally, segments of a transcript that are cut out during splicing, but also used widely to describe the corresponding DNA sequences (see **Figure 1.18**). Some introns are the source of small nucleolar RNAs, microRNAs, and other functional RNAs (see **Figures 9.5B**, **10.34**).

Inversion: a sequence variant in which a stretch of sequence (which may be anything between a few dozen nucleotides and a large segment of a chromosome) is present in the opposite orientation to the normal. See **Figure 15.11** for chromosomal inversions.

Iron response element (IRE): a sequence element in certain mRNA species that changes the activity of the mRNA in response to excess or deficiency of Fe^{++}. See **Figure 10.32**.

ISCN: International System for Cytogenetic Nomenclature. See **Box 15.1**.

Isochromosome: an abnormal symmetrical chromosome, consisting of two identical arms, which are normally either the short arm or the long arm of a normal chromosome.

Isoforms: variants of a protein or noncoding RNA produced at a single locus (often because of alternative splicing).

Isogenic: describing tissues or cells that are genetically identical. cf. Allogeneic.

Isothermal amplification: *in vitro* DNA amplification carried out at constant temperature. See **Figure 7.18**.

K_a/K_s ratio (also called the dN/dS ratio): an indicator of selection affecting a gene sequence that involves calculating the number of nonsynonymous substitutions per nonsynonymous site and dividing by the number of synonymous substitutions per synonymous site in a comparison of two organisms. See **Box 13.1**.

Karyotype: a summary of the chromosome constitution of a cell or person, such as 46,XY. Often used more loosely to mean an image showing the chromosomes of a cell that have been sorted in order and arranged in pairs (strictly, a karyogram). See **Figure 15.1**.

Kataegis: in a tumor, occurrence of large numbers of mutations clustered in a small chromosomal region. See Section 19.4.

Kinetochore: the structure at chromosomal centromeres to which the spindle fibers attach. See **Box 2.3**.

Kozak consensus sequence: the necessary context for a functional AUG initiation codon, GCCRCCAUGG (R = purine).

Kurgan hypothesis: in linguistics, the hypothesis that Indo-European languages originated with the Asian Kurgan culture around 4500 years ago. See **Box 14.1**.

Lagging strand: in DNA replication, the strand that is synthesized as Okazaki fragments. See **Figure 1.12**.

Lateral inhibition: a process during embryogenesis in which cells that differentiate inhibit neighboring cells from doing the same. The result is to develop spaced sets of differentiated cells.

Leader sequence: a sequence of a dozen or so amino acids at the N-terminal end of some proteins that serves as a signal defining the location to which the protein must be transported. Leader sequences are usually cleaved off once the sorting process is completed.

Leading strand: in DNA replication, the strand that is synthesized continuously. See **Figure 1.12**.

Leucine zipper: a structure found in many DNA-binding proteins. See **Box 3.1**.

Ligand: any molecule that binds specifically to a receptor or other molecule. See **Figure 3.1**.

Ligase: DNA ligase is an enzyme that can seal single-strand nicks in double-stranded DNA or covalently join two oligonucleotides that are hybridized at adjacent positions on a DNA strand.

LINE (long interspersed nuclear element): a class of repetitive DNA sequences that make up about 20% of the human genome. Some are active transposable elements. See **Figures 9.12**, **9.13**.

Linkage disequilibrium (LD): a statistical association between particular alleles at separate but linked loci, normally the result of a particular ancestral haplotype being common in the population studied. See **Figures 12.4**, **12.5**, and **Box 12.2**.

Linker (adapter) oligonucleotide: a double-stranded oligonucleotide that can be ligated to a DNA molecule of interest and which has been designed to contain some desirable characteristic, e.g. a favorable restriction site or a binding site for a PCR primer.

Liposome: a synthetic lipid vesicle used to introduce DNA into cells. See **Figure 8.4**.

Liquid biopsy: genetic diagnosis, particularly of cancer, using cell-free DNA or tumor cells in the peripheral circulation.

Locus: a unique chromosomal location defining the position of an individual gene or DNA sequence.

Locus control region (LCR): a stretch of DNA containing regulatory elements that control the expression of genes in a gene cluster that may be located tens of kilobases away.

Locus heterogeneity: determination of the same disease or phenotype by mutations at more than one locus.

Lod score (z): in linkage analysis, the log (base 10) of the odds that the loci are linked (with recombination fraction θ) rather than unlinked. For Mendelian characters a lod score greater than +3 is evidence of linkage; one that is less than –2 is evidence against linkage. See **Boxes 17.1**, **17.2**.

Logistic regression: a statistical method of deciding which of a series of factors affecting an outcome are independent.

Loss-of-function mutations: mutations that cause the gene product to lose its function, partially or totally. See Section 16.1.

Loss of heterozygosity (LoH): homozygosity or hemizygosity in a tumor or other somatic cell when the constitutional genotype is heterozygous. Evidence of a somatic genetic change. See **Figures 19.9**, **19.10**.

Lymphoblastoid cell line: immortalized EBV-transformed cells. See **Box 8.1**.

Lymphocytes: white blood cells involved in the immune response. The two main classes are T cells and B cells.

Lyonization: X-chromosome inactivation, the process by which cells adapt to differing numbers of X chromosomes. See Section 10.4.

Lysosomes: membrane-bound organelles full of digestive enzymes where macromolecules are broken down to their basic subunits. See **Box 2.1**.

Macrophages: the general scavenger cells of the body.

Major groove: in a DNA double helix, the larger of the two spiral grooves that run the length of the molecule. Many DNA-binding proteins recognize sequence-specific features in the major groove.

Maintenance DNA methyltransferase: the DNMT1 DNA methyltransferase, which acts to maintain the parental pattern of DNA methylation in daughter cells. See **Figure 10.9**.

Major histocompatibility complex (MHC) proteins: proteins encoded by the Class I and Class II regions of the MHC that function in antigen recognition by binding fragments of antigens and presenting them on the surface of T cells. See **Box 3.2**.

Major pseudoautosomal region: a 2.6 Mb homologous region at the tips of Xp and Yp. These sequences pair in male meiosis and have an obligatory recombination.

Manhattan plot: a graphical presentation of the results of a genome-wide association study, plotting odds ratio vs. chromosomal location for each SNP. See **Figure 18.6**.

Manifesting heterozygote: a female carrier of an X-linked recessive condition who shows some clinical symptoms, presumably because of skewed X-inactivation.

Map function: a mathematical equation describing the relation between recombination fraction and genetic distance. See Section 17.1.

Marker gene: in genetic manipulation, a gene whose product enables cells containing a construct to be recognized or selected. See, for example, **Figure 8.13**.

Matrilineal inheritance: transmission from just the mother, but to children of either sex; the pattern of mitochondrial inheritance. See **Box 5.1**.

Matrix-assisted laser desorption/ionization (MALDI): a method for analyzing large nonvolatile molecules on a mass spectrometer. The molecules are mixed into a light-absorbing matrix that is vaporized by a laser pulse. See **Box 7.8**. Often combined with time-of-flight analysis (MALDI-TOF).

MeDIP: immunoprecipitation of methylated DNA fragments.

Meiosis: the specialized reductive form of cell division used only to produce gametes. See **Figures 2.12–2.16**.

Melting temperature (Tm): in denaturing double-stranded DNA, the temperature at the mid-point of the transition from double to single strands.

Mendelian: a character whose pattern of inheritance suggests it is caused by variation at a single genetic locus; a monogenic character. See **Box 5.1**.

Mesenchymal stem cells: the stem cells of connective tissues.

Mesoderm: embryonic tissue that is the precursor to muscle, connective tissue, the skeleton, and many internal organs. See **Figure 4.9**.

Messenger RNA: a processed gene transcript that carries protein-coding information to the ribosomes.

Meta-analysis: an analysis of combined data from a number of independent studies of the same topic.

Metabolic interference: a hypothetical situation where two alleles, each in itself normal, conflict in a heterozygote so as to produce an abnormal phenotype, although both homozygotes are normal.

Methyl-Seq: identifying the methylation pattern of a DNA fragment by bisulfite sequencing (qv.).

Methylation-sensitive PCR: PCR amplification of the sample with and without bisulfite treatment, to identify methylated cytosines. See **Box 10.3**.

MHC restriction: the process that restricts recognition of foreign antigens by T cells to fragments that are associated with an MHC molecule on the surface of an antigen-presenting cell.

Microbiome: the totality of microorganisms in a particular environment. The human microbiome includes all the microbes resident on and within the body.

Microcell-mediated chromosome transfer: a technique for introducing a selected single chromosome into a mutant cell, to see if it can correct the mutant phenotype.

Microfilament: actin filaments forming part of the cell cytoskeleton.

Micronucleus: small membrane-bound vesicle containing DNA, formed artificially or as a result of anaphase lag of chromosomes during cell division.

MicroRNAs (miRNAs): short (21–22 nt) RNA molecules encoded within normal genomes that have a role in regulation of gene expression and maybe also of chromatin structure. See **Figures 8.19**, **10.33**.

Microsatellite: small run (usually less than 0.1 kb) of tandem repeats of a very simple DNA sequence, usually 1–4 bp, for example $(CA)_n$. Often polymorphic, providing the primary tool for genetic mapping during the 1990s. Sometimes also described as STR (short tandem repeat) polymorphism. See **Figure 7.4**.

Microtubules: long hollow cylinders constructed from tubulin polymers. Form the spindle fibers that move chromosomes in mitosis and meiosis, and contribute to the cytoskeleton.

Microsatellite instability: tumor-specific generation of extra alleles at microsatellites due to defective mismatch repair. See **Figure 19.15**.

Minigene assay: a test of splicing in which a portion of a gene is cloned into an expression vector to check splicing. See **Figure 16.3**.

Minisatellites: 0.1–20 kb arrays of tandem repeats.

Minor groove: in a DNA double helix, the smaller of the two spiral grooves that run along the length of the molecule.

Missense changes: changes in a coding sequence that cause one amino acid in the gene product to be replaced by a different one.

Missing heritability: describes the gap between heritability as estimated from family studies and heritability estimated from the effects of all known susceptibility factors. See **Figure 18.8** for possible explanations.

Mitogen: a substance that stimulates cells to divide.

Mitosis: the normal process of cell division, which produces daughter cells genetically identical to the parent cell. See **Figure 2.11**.

Molecular barcoding: attaching random short synthetic oligonucleotides to a mix of DNA fragments, so that individual fragments are uniquely labeled. See **Figure 7.15**.

Molecular clock: in evolutionary studies, the idea that DNA sequence changes accumulate at a fixed rate, allowing one to use molecular data to date events.

Monoclonal antibody (mAb): a pure antibody with a single specificity, produced by hybridoma technology, as distinct from polyclonal antibodies that are raised by immunization. See **Box 7.6**.

Monosomy: having just one copy of a particular chromosome e.g. 45,X in Turner syndrome.

Morphogenesis: the formation of structures during embryonic development.

Morphogens: signaling molecules that can impose a pattern on a field of cells in response to a gradient of concentration of the morphogen.

Morpholino: a type of stable chemically-modified RNA analog used to inhibit expression of a gene under study. See **Figure 8.5**.

Morula: an early stage of embryonic development; a loosely packed ball of cells that will give rise to the blastocyst. See **Figure 4.4**.

Mosaic: an individual who has two or more genetically different cell lines derived from a single zygote. The differences may be point mutations, chromosomal changes, etc. See **Figure 5.16**.

Mosaic pleiotropism: a gene that has different functions in different cell types, or at different developmental stages.

Motif: a short sequence or structure (usually in a protein) that forms a recognizable signature of a structure or function.

Missense changes: changes in a coding sequence that cause one amino acid in the gene product to be replaced by a different one.

Monozygotic: of twins, resulting from splitting of a single embryo. See **Box 4.3**.

mtDNA: mitochondrial DNA—DNA of the 16,569 nt mitochondrial genome. See **Figure 9.1**.

Multifactorial: a character that is determined by some unspecified combination of genetic and environmental factors. cf. Polygenic.

Multiplex ligation-dependent probe amplification (MLPA): a technique for detecting whole-exon deletions or duplications. See **Figure 20.5**.

Mutagen: an external agent that can cause mutations.

Mutagenesis: creation of mutations, *in vitro* or *in vivo*. Artificially induced mutagenesis of cells may be designed to occur randomly (such as by using chemical mutagens or transposons), or be targeted to specific loci by some type of genome editing.

Mutation: (1) the process of DNA sequence change; (2) the resultant changed sequence.

Mutational signature: in a tumor, a pattern of base-substitution mutations that allows the mutational mechanism to be inferred. See **Figure 19.20**.

N-glycosidic bond: a bond linking a sugar molecule to a nitrogen atom, typically one in a purine or pyrimidine base. See **Figure 1.2**.

Neanderthals: an extinct species of human, named after the Neander valley (thal in 19th century German) where their remains were first found. See **Figures 14.12**, **14.13**.

Necrosis: cell death as a result of irreparable external damage.

Neocentromere: formation of a novel functional centromere in a hitherto euchromatic part of a chromosome.

Neofunctionalization: one of a pair of duplicated genes evolves to have a novel function. See **Figure 13.11**.

Neogene: in evolution, a novel protein-coding gene originating from noncoding DNA, such as by exaptation of a transposable element.

Neoteny: in evolutionary studies, the idea that species may evolve by extending their period of infantile development and retaining features previously lost during maturation.

Neural crest: a group of migratory cells in an embryo that form along the lateral margin of the neural folds and give rise to many different tissues. Neural crest derivatives include part of the peripheral nervous system, melanocytes, some bone and muscle, the retina, and other structures.

Neural tube: in vertebrate embryos, a tube of ectoderm that will form the brain and spinal cord.

Next-generation (massively-parallel) sequencing: a collection of methods for very high-throughput DNA sequencing by sequencing many molecules in parallel. See Section 6.5.

Nonallelic homologous recombination (NAHR): recombination between misaligned DNA repeats, either on the same chromosome, on sister chromatids or on homologous chromosomes. NAHR generates recurrent deletions, duplications, or inversions. See **Figure 15.17**.

Noncoding RNA (ncRNA): RNA that does not contain genetic code for a protein. Noncoding RNAs have many different functions in cells.

Nondisjunction: failure of chromosomes (sister chromatids in mitosis or meiosis II; paired homologs in meiosis I) to separate (disjoin) at anaphase. The major cause of numerical chromosome abnormalities.

Nonhomologous end-joining (NHEJ): a DNA repair mechanism in which broken DNA ends are joined regardless of homology. See **Figure 15.18**.

Noninvasive prenatal testing (NIPT): specifically, the analysis of cell-free DNA in the maternal circulation to detect fetal anomalies.

Nonparametric (model-free) linkage analysis: a method such as affected sib pair analysis that does not depend on a specific genetic model. See Section 18.2.

Nonpenetrance: the situation when somebody carrying an allele that normally causes a dominant phenotype does not show that phenotype. An effect of other genetic loci or of the environment. A pitfall in genetic counseling. **Figure 5.11** shows an example.

Nonrecombinant: in a pedigree, two loci are nonrecombinant in a gamete that contains the same combination of alleles as the person received from his or her parent. See **Figure 17.3**.

Nonsense-mediated mRNA decay: a cellular mechanism that degrades mRNA molecules that contain a premature termination codon (>50 nt upstream of the last splice junction). See **Figure 16.7**.

Nonsense mutation: a mutation that replaces the codon for an amino acid with a termination codon.

Nonsynonymous substitution (or mutation): one that replaces one codon by another that specifies a different amino acid.

Northern blot: a membrane bearing RNA molecules that have been size-fractionated by gel electrophoresis, used as a target for a hybridization assay. Used to detect the presence and size of transcripts of a gene of interest.

Notochord: a flexible rod-like structure that in mammalian embryos induces formation of the central nervous system.

Nuclear reprogramming: large-scale epigenetic changes to convert the pattern of gene expression in a cell to that typical of another cell type or state.

Nucleic acid: DNA or RNA.

Nucleoid: the loosely organized DNA in a prokaryotic cell or mitochondrion. See **Box 2.1**.

Nucleolar organizer region (NOR): the satellite stalks of human chromosomes 13, 14, 15, 21, and 22. NORs contain arrays of ribosomal RNA genes and can be selectively stained with silver. Each NOR forms a nucleolus in telophase of cell division; the nucleoli fuse in interphase.

Nucleolus: the site within the nucleus where ribosomal RNA is transcribed and assembled into the ribosomal subunits.

Nucleoside: a purine or pyrimidine base linked to a sugar (ribose or deoxyribose). See **Table 1.1**.

Nucleosome: the basic structural unit of chromatin, comprising 186 bp of DNA wound round an octamer of histone molecules. See **Figure 2.18**.

Nucleotide: base + sugar + phosphate. The basic building block of DNA and RNA. See **Table 1.1**.

Odds ratio: in case–control studies, the relative odds of a person with or without a factor under study being a case. See **Box 18.2**.

Okazaki fragments: short strands of DNA, the immediate product of lagging-strand replication. See **Figure 1.12**.

Oligogenic: a character that is determined by a small number of genes acting together.

Oligonucleotide ligation assay (OLA): a test to identify which allele of a SNP is present in a sample. See **Figure 20.3**.

Oligosaccharide: a molecule consisting of a few linked sugar units.

OMIM: the Online Inheritance in Man database at https://www.ncbi.nlm.nih.gov/omim.

Oncogene: a gene involved in control of cell proliferation which, when overactive can help to transform a normal cell into a tumor cell. See Section 19.1. Originally the word was used only for the activated forms of the gene, and the normal cellular gene was called a proto-oncogene, but this distinction is now widely ignored.

One gene-one enzyme hypothesis: the hypothesis advanced by Beadle and Tatum in 1941 that the primary action of each gene was to specify the structure of an enzyme. Historically very important, but now seen to be only part of the range of gene functions.

Open reading frame: a DNA sequence that does not contain a stop codon in a selected reading frame.

Organelle: a discrete body within a eukaryotic cell. See **Box 2.1**.

Organoid: *in vitro* miniature organs derived from multiple cell types, typically supported on a hydrogel composed of extracellular matrix proteins. In organoids generated from stem cells, cell co-operation allows impressive self-assembly and self-organizing abilities.

Origin of replication: a site on DNA where replication can be initiated.

Ortholog: orthologous genes are genes present in different organisms that are related through descent from a common ancestral gene. See **Figure 13.10**.

Orphan gene: a gene without detectable homologues in other lineages, and so may be restricted to a narrow taxon, such as a species.

Out-of-Africa model: the idea that the ancestors of all current non-African populations originated in a migration out of Africa around 50,000 years ago. See **Figure 14.12**.

P450 enzymes: a large family of enzymes involved in the metabolism of drugs and other foreign substances. See Section 20.5.

Paired-end sequencing: comparing the number of nucleotides separating two known sequences in a person's DNA with the number in a reference genome, as a way of identifying structural rearrangements. See **Figure 6.21**.

Paleontology: the study of fossils.

Palindrome: a DNA sequence such as ATCGAT that reads the same when read in the 5' → 3' direction on each strand. DNA–protein recognition, for example by restriction enzymes, often relies on palindromic sequences.

Paralog: one of a set of homologous genes within a single species. See **Figure 13.10**.

Parametric linkage analysis: a method such as standard lod score analysis, that requires a tightly specified genetic model.

Paramutation: an unusual situation where an inherited phenotype is due to transgenerational epigenetic inheritance rather than a DNA sequence change. See **Figure 10.20**.

Passenger mutations: in cancer, mutations that arise incidentally during development of a tumor and do not play any causative role in the process. cf. Driver mutations.

Paternity index: in paternity testing, the relative likelihood that the suspect rather than a random man from the same population is the father. See **Figure 20.21**.

Pathogenic: causing disease.

Penetrance: the frequency with which a genotype manifests itself in a given phenotype.

Peptide bond: the bond linking amino acids in a polypeptide. See **Figure 1.3**.

Peptide mass fingerprinting: a way of identifying proteins in a mixture by digesting with trypsin, analyzing the mixture of peptides on a mass spectrometer, and comparing the resulting peaks against a database showing the patterns produced by known proteins.

Personalized (stratified) medicine: using genomic data to inform the management and treatment of patients.

Phage: a bacteriophage. A virus that replicates in bacterial cells.

Phage display: an expression cloning method in which foreign genes are inserted into a phage vector and are expressed to give polypeptides that are displayed on the surface (protein coat) of the phage. See **Figure 6.7**.

Phagocyte: a cell that specializes in engulfing and killing microbial pathogens. See **Figure 3.19**.

Phagocytosis: digestion in lysosomes of material imported into a cell.

Pharmacogenetics: the study of the influence of genetic factors on the response to drugs. Divided into pharmacokinetics (the absorption, activation, catabolism, and elimination of a drug) and pharmacodynamics (the response of a target organ or cell to a drug).

Phase: (1) in a pedigree, if we know which combination of alleles a person inherited from each individual parent they are phase-known; otherwise they are phase-unknown. See **Figure 17.5**; (2) in genome-wide association studies, phasing means converting genotypes into haplotypes. See **Table 18.5**.

Phenotype: the observable characters (traits) of an individual.

Phenocopy: a person (or organism) who has a phenotype normally caused by a certain genotype, but who does not have that genotype. Phenocopies may be the result of a different genetic variant, or of an environmental factor.

Phenome: the totality of phenotypes, that is, the set of physical and biochemical traits, exhibited by an organism. cf. Genome.

Phenomics: measurement of phenomes as they change in response to mutation and/or environmental influences.

Phenotype: the observable characteristics of a cell or organism, including the result of any test that is not a direct test of the genotype.

Phylogeny: classification of organisms according to perceived evolutionary relatedness. See **Figures 13.28 –13.31**.

Physical distance: the distance between two objects, measured in bp, kb, Mb, etc.

Physical map: a map showing the locations of some physical entities on a chromosome or genome. The entities might be DNA sequences or features such as natural or radiation-induced breakpoints. cf. Genetic map.

Pitch: of a spiral, the distance occupied by a single turn, which is 3.4 nm in the standard B-DNA double helix.

Plasma membrane: the membrane that surrounds a cell.

Plasmid: a small circular DNA molecule that can replicate independently in a cell. Modified plasmids are widely used as cloning vectors.

Pleiotropy: the common situation where variation in one gene affects several different aspects of the phenotype.

Ploidy: the number of complete sets of chromosomes in a cell. Cells can be haploid, diploid, triploid… polyploid.

PMID: PubMed identifier, a unique number, currently eight digits in length, assigned to a biomedical literature article at https://www.ncbi.nlm.nih.gov.

Polar body: in female meiosis, the small product of the asymmetrical division of the cell mass during each division of meiosis. The polar bodies eventually degenerate.

Poly(A) tail: the string of 200 or so A residues that are attached to the 3' end of a mRNA. The poly(A) tail is important for stabilizing mRNA. See **Figure 1.23**.

Polyadenylation: addition of the poly(A) tail to the 3' end of a mRNA. See **Figure 1.23**.

Polyclonal antibodies: natural antibodies produced by the adaptive immune system in response to an antigen. Polyclonal antibodies are typically a mixture of species that respond to different epitopes of the stimulating antigen.

Polygenic: a character determined by the combined action of a number of genetic loci. Mathematical polygenic theory (see Section 5.4) assumes there are very many loci, each with a small effect.

Polylinker: in a cloning vector, a short sequence containing recognition sites for several different restriction enzymes, as an aid to making recombinant molecules. See **Figure 6.4**.

Polymerase chain reaction (PCR): the standard technique used to amplify short DNA sequences. See **Figure 6.8**.

Polymorphism: strictly, the existence of two or more variants (alleles, phenotypes, sequence variants, chromosomal structure variants) in the population at frequencies too high to be maintained by recurrent mutation. Looser usages among molecular geneticists include: (1) a sequence variant present at a frequency >1% in a population; (2) a nonpathogenic sequence variant, regardless of frequency. See Section 11.3.

Polypeptide: a string of amino acids linked by peptide bonds. See **Figure 1.3**. Proteins may consist of one or more polypeptide chains.

Polyploid: having many (>2) copies of the genome.

Population attributable risk (PAR): in epidemiology, the contribution that a particular factor or combination of factors makes to the overall incidence of a condition.

Population mutation parameter: the diversity per nucleotide site that is expected due to the balance between mutation creating new alleles and drift to fixation eliminating them. See Section 12.3.

Position effect: complete or partial silencing of a gene when a chromosomal rearrangement moves it close to heterochromatin.

Positional cloning: identifying a disease gene using knowledge of its chromosomal location. The way the great majority of Mendelian disease genes were identified. See **Figure 17.2**.

Positive selection: selection in favor of a particular genotype.

Potency: of a cell, its potential for dividing into different cell types. Mammalian cells can be totipotent, pluripotent, oligopotent, or committed to one fate.

Pluripotent: when applied to mammalian cells, this term describes the ability to give rise to all of the cell types in the body but not to extra-embryonic/placental cells, such as in the natural case of cells in the inner cell mass or artificial embryonic stem cells.

Premutation alleles: among diseases caused by dynamic mutations (expanding repeats), a repeat expansion that is large enough to be unstable on transmission, but not large enough to cause disease. See Section 16.3.

Primary cilium: a cell surface structure carrying many types of receptor molecules. See **Box 2.1**.

Primary structure: of a polypeptide or nucleic acid, the linear sequence of amino acids or nucleotides in the molecule. See **Table 1.7**.

Primary transcript: the RNA product of transcription of a gene by RNA polymerase, before splicing. The primary transcript of a gene contains all the exons and introns.

Primase: DNA primase is an enzyme that synthesizes a short RNA molecule that serves as a primer for DNA replication.

Primer: a short oligonucleotide, often 15–25 bases long, which base-pairs specifically to a target sequence to allow a polymerase to initiate synthesis of a complementary strand.

Primitive streak: in early embryos, a transient structure that defines the longitudinal axis. See **Figure 4.8**.

Primordial germ cells (PGCs): cells in the embryo and fetus that will ultimately give rise to germ-line cells.

Prion: an infectious pathogenic protein.

Proband or propositus: the person through whom a family was ascertained.

Probe: a known DNA or RNA fragment (or a collection of such fragments) used in a hybridization assay to identify closely related DNA or RNA sequences within a complex, poorly understood test sample nucleic acid population. In standard hybridization assays, the probe is labeled but in reverse hybridization assays the test sample nucleic acid is labeled.

Programmable nuclease: an endonuclease that is engineered to be able to cut DNA strands at a specific desired target sequence within the genome of intact cells. The endonuclease is steered to the target site by being bound to an RNA or protein *guide sequence* designed to bind to a specific sequence at the target site.

Programmed cell death: programmed death of an animal cell, in which a "suicide" program is activated in the cell. See **Table 3.3**, **Figure 3.9**.

Prokaryotes: single-celled microorganisms (bacteria or archaea) that lack a membrane-bound nucleus.

Prometaphase: in mitosis, late prophase, when chromosomes are well separated but not yet maximally contracted; the optimum stage for normal cytogenetic analysis. See **Figure 2.11**.

Promoter: a combination of short sequence elements, normally just upstream of a gene, to which RNA polymerase binds in order to initiate transcription of the gene. See **Figures 1.16**, **10.21**.

Proofreading: an enzymic mechanism by which DNA replication errors are identified and corrected.

Propositus: the proband, the person through whom a family was ascertained.

Prosecutor's fallacy: a misuse of statistics in court cases to exaggerate the likelihood the defendant is guilty. See Section 20.6.

Protein domain: a structural (and often functional) subunit of a protein; a structural module that may be found in several different proteins.

Proteome: the totality of proteins in a cell or organism. Highly variable between different cell types, unlike the genome.

Proteomics: global or large-scale studies of the proteins in a cell or organism. See Section 7.3.

Proto-oncogenes: normal cellular genes whose function is to promote cell proliferation. Activated versions in tumors are called oncogenes, but nowadays the proto-oncogene/oncogene terminology is widely disregarded, and all versions are called oncogenes.

Proximal: of a chromosomal location, comparatively close to the centromere.

Pseudoautosomal regions (PAR): regions at each tip of the X and Y chromosomes containing X–Y homologous genes (see **Figures 13.16**, **13.17**). Because of X–Y recombination, alleles in these regions show an apparently autosomal mode of inheritance.

Pseudogene: a DNA sequence that shows a high degree of sequence homology to a nonallelic functional gene, but which is itself nonfunctional. See **Box 9.2**.

Purifying (negative) selection: selection against unfavorable genotypes.

Purines: nitrogenous bases having a specific double-ring chemical structure. Adenine and guanine are purines. See **Figure 1.2**.

Pyrimidines: nitrogenous bases having a specific single-ring chemical structure. Cytosine, thymine, and uracil are pyrimidines. See **Figure 1.2**.

Pyrosequencing®: a technique for sequencing a few nucleotides from a defined start point. See **Box 6.4**.

Quantitative PCR (qPCR): PCR methods that allow accurate estimation of the amount of template present. Reliable qPCR methods are based on real-time techniques. See **Box 7.5**.

Quantitative character: a character such as height, which everybody has, but to differing degree—as compared with a dichotomous character like polydactyly, which some people have and others do not.

Quantitative trait locus (QTL): a locus that contributes to determining the phenotype of a continuous (quantitative) character.

Quaternary structure: the overall structure of a multimeric protein. See **Table 1.7**.

Race: in anthropology and much popular thought, the idea that humans can be divided into separate races (Caucasoid, Negroid, Mongoloid, etc.). This has been largely discredited by the evidence that within-population variation far exceeds between-population variation. This is not to deny that group differences do exist, but they need to be seen in a more nuanced way.

Radiation hybrid: an artificially produced cell in culture that contains radiation-generated fragments of human chromosomes in a mouse background.

RAN (repeat-associated non-ATG) translation: abnormal translation of a mRNA containing expanded repeats. See **Table 16.6**.

Rare variant: a variant whose frequency is <0.01.

Reactive oxygen species (ROS): superoxide ions (O_2^-), hydroxyl radicals (OH·) and hydrogen peroxide (H_2O_2) formed mainly in mitochondria, that can damage DNA or proteins. See **Figure 11.3**.

Read depth: in next-generation sequencing, the number of times a given stretch of DNA is sequenced in independent fragments (primarily a function of the amount of sample loaded into the machine). See **Figure 17.12**.

Reading frame: during translation, the way the continuous sequence of the mRNA is read as a series of triplet codons. See **Figure 16.5**. The correct reading frame is set by correct recognition of the AUG initiation codon.

Real-time PCR: a PCR process in which the accumulation of product is followed in real time, which allows accurate quantitation of the amount of template present.

Receptor-mediated endocytosis: invagination of the plasma membrane after ligands have bound to receptor proteins on the cell surface. See **Figure 8.3**.

Recessive: a character is recessive if it is manifest only in the homozygote.

Recombinant: in linkage analysis, a gamete that contains a combination of alleles that is different from the combination that the parent inherited from their parent. See **Figure 17.3**.

Recombinant DNA: an artificially constructed hybrid DNA containing covalently linked sequences from two or more different sources. See **Figure 6.2**.

Recombinant proteins: proteins produced in expression cloning systems. Although the vector is recombinant, the protein is not actually recombinant.

Recombination fraction: for a given pair of loci, the proportion of meioses in which they are separated by recombination. Usually signified as θ. θ values vary between 0 and 0.5. See Section 17.1.

Regression to the mean: the phenomenon whereby parents with extreme values of quantitative characters have, on average, children with less extreme values. This is a purely statistical phenomenon, and has no bearing on whether or not a character is genetically determined. See Section 5.4.

Relative risk: in epidemiology, the relative risks of developing a condition in people with and without a susceptibility factor. See **Box 18.2**.

Replication fork: in DNA replication, the point along a DNA strand where the replication machinery is currently at work.

Replication slippage: an error during replication of a tandem repeated DNA sequence, causing the newly synthesized strand to have one or more too few or too many repeat units. See **Figure 11.1**.

Replicon: any nucleic acid that is capable of self-replication. Many cloning vectors use extrachromosomal replicons (as in the case of plasmids). Others use chromosomal replicons, either directly (as in the case of yeast artificial chromosome vectors), or indirectly, by allowing integration into chromosomal DNA.

Reporter gene: a gene used to test the ability of an upstream sequence joined onto it to cause its expression. Putative *cis*-acting regulatory sequences can be coupled to a reporter gene and transfected into suitable cells to study their function. Alternatively, transgenic animals (and other organisms) are often made with a promotorless reporter gene integrated at random into the chromosomes, so that expression of the reporter marks the presence of an efficient promoter.

Reporter molecule: a molecule whose presence is readily detected (for example, a fluorescent molecule) that is attached to a DNA sequence we wish to monitor.

Restriction endonuclease: a bacterial enzyme that cuts double-stranded DNA at, or close to, a short (normally 4, 6, or 8 bp) recognition sequence. See **Box 6.1**.

Restriction fragment length polymorphism (RFLP): a DNA polymorphism that creates or abolishes a recognition sequence for a restriction endonuclease. When DNA is digested with the relevant enzyme, the sizes of the fragments will differ, depending on the presence or absence of the restriction site. See **Figure 7.4**.

Retrogene: a functional gene that appears to be derived from a reverse-transcribed RNA. See **Box 9.2**.

Retrotransposon: a genetic element that (potentially) uses reverse transcription of its RNA to insert DNA copies elsewhere in the genome. Sometimes called a retroposon. See **Figure 9.12**.

Retrovirus: an RNA virus with a reverse transcriptase function, enabling the RNA genome to be copied into cDNA prior to integration into the chromosomes of a host cell. See **Figure 8.7**.

Reverse genetics: inferring phenotypes from knowledge of genes (a reversal of the classical pathway in which genes are identified through the study of phenotypes).

Reverse transcriptase: an enzyme, often of viral origin, that makes a DNA copy of an RNA template; an RNA-dependent DNA polymerase.

Reverse transcriptase PCR: indirect PCR amplification of RNA by first making a cDNA copy using reverse transcriptase.

Revertant mosaicism: in a constitutionally mutant individual, mosaicism for cells that have reverted to the nonmutant type.

Ribonuclease: an enzyme that breaks down RNA.

Ribonuclease protection assay: a method for quantitating one specific RNA transcript in a complex mixture. Uses a labeled antisense probe to protect the transcript of interest from degradation by ribonuclease.

Ribosomal DNA: the DNA from which ribosomal RNA is transcribed using RNA polymerase I. In human cells, located on the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22). See **Figure 1.24**.

Ribosome: the large cytoplasmic protein-RNA complex where polypeptides are assembled using information in a messenger RNA.

Ribozyme: a natural or synthetic catalytic RNA molecule.

Risk ratio: in family studies, the relative risk of disease in a relative of an affected person, compared to a member of the general population. See Section 18.1.

RNA editing: insertion, deletion, or substitution of specific nucleotides in a mRNA after transcription. An unusual event in humans. See **Figure 10.30**.

RNA interference: (1) a natural mechanism used to protect cells against viruses and transposons; (2) a powerful tool for studying gene function using siRNAs to knock down expression of specified genes.

RNA polymerase: an enzyme that can add ribonucleotides to the 3′ end of an RNA chain. Most RNA polymerases use a DNA template, but some use an RNA template, and hence synthesize double-stranded RNA. See **Table 1.3**.

RNA processing: the processes required to convert a primary transcript into a mature messenger RNA—capping, splicing, and polyadenylation.

RNA-Seq: sequencing cDNA as an indirect method of sequencing RNA. Some new technologies in principle allow direct sequencing of RNA.

RNA splicing: see Splicing.

RNA therapeutics: therapies directed at interfering with gene expression at the RNA level, including RNA interference-based gene silencing and other methods.

Robertsonian translocation: a translocation between the short arms of two acrocentric human chromosomes (nos. 13, 14, 15, 21, or 22). See **Figure 15.10**.

ROC (receiver operating characteristic) curve: for a test, a graph of sensitivity versus specificity that gives a measure of the discriminatory power of the test. See **Box 20.3**.

Rooted evolutionary tree: one that infers the existence of a common ancestor. See **Figure 13.29**.

Rough endoplasmic reticulum: endoplasmic reticulum (qv.) that is studded with ribosomes.

RT-PCR: see Reverse transcriptase PCR. NB. *not* real-time PCR).

SAGE (serial analysis of gene expression): a method of expression profiling based on sequencing.

Sanger sequencing: sequencing DNA using dideoxynucleotide chain terminators—the technique invented by Fred Sanger. See Section 6.4.

Satellite (chromosome): stalked knobs variably seen on the ends of the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22).

Satellite (DNA): a DNA fraction that forms minor "satellite" bands on density gradient centrifugation because of its unusual base composition. The DNA is composed of very long arrays of tandemly repeated sequences. See **Table 9.13**.

Screening: sometimes used loosely to mean simply testing, but in this book used to mean a planned population-based program to define a high-risk subgroup.

Second-degree relatives: uncles, aunts, nephews, nieces, grandparents, grandchildren, and half-sibs.

Second messenger: intracellular signaling molecules that relay signals from cell surface receptors to downstream targets. See **Table 3.2**.

Secondary structure: the path of the backbone of a folded polypeptide or single-stranded nucleic acid, determined by weak interactions between residues in different parts of the sequence. See **Table 1.7**.

Segmental aneuploidy syndrome: a syndrome caused by deletion or duplication of a segment of a chromosome. See **Table 15.4** for some examples.

Segmental duplication: the existence of very highly related DNA sequence blocks on different chromosomes or at more than one location within a chromosome.

Segregation: (1) the distribution of allelic sequences between daughter cells at meiosis. Allelic sequences are said to segregate, nonallelic sequences to assort; (2) in pedigree analysis, the probability of a child inheriting a phenotype from a parent.

Segregation analysis: the statistical methodology for inferring modes of inheritance. See Section 18.1.

Selective sweep: reduced heterozygosity in a population around a locus that has recently been subject to strong positive selection. See **Box 11.3**.

Semi-conservative: of DNA replication: each daughter double helix contains one parental and one newly synthesized strand. See **Figure 1.11**.

Semi-discontinuous: of DNA replication, where one newly synthesized strand has to be made in short pieces (Okazaki fragments) because DNA polymerase can only extend a chain in the 5′ → 3′ direction. See **Figure 1.12**.

Semi-dominant: used in animal and plant studies to describe a character that is present in heterozygotes, but in a less marked form than in homozygotes (e.g. pink vs. red flowers). Not much used in human genetics, see Section 5.2.

Sense strand: the DNA strand of a gene that is complementary in sequence to the template (antisense) strand, and identical to the transcribed RNA sequence (except that DNA contains T where RNA has U). Quoted gene sequences are always of the sense strand, in the 5' → 3' direction. See **Figure 1.15**.

Sensitivity of a test: the proportion of all true positives that the test is able to detect.

Sequence similarity: the degree to which two nucleic acid or protein sequences are identical. A looser measure of similarity takes into account synonymous codon replacements in DNA and conservative amino acid replacements in proteins.

Sequence-tagged site (STS): any unique piece of DNA for which a specific PCR assay has been designed, so that any DNA sample can be easily tested for its presence or absence. See **Figure 7.5**.

Serial founder model: the theory that non-African populations originated from several different migrations out of Africa, rather than a single event.

Sex chromatin: the Barr body (qv.).

Short interfering RNA (siRNA): 21–22 nt double-stranded RNA molecules that can dramatically shut down expression of genes (RNA interference). siRNAs are a major tool for studying gene function. See **Box 8.2**.

Short tandem repeat polymorphism (STR): a polymorphic microsatellite.

Shotgun sequencing: sequencing a genome by mass sequencing of random fragments, then using computers to assemble the mass of short reads into an overall sequence. Assembly is difficult for genomes with much repetitive DNA, as with humans. See **Figure 7.2**.

Sib: a brother or sister.

Signal sequence: a short N-terminal sequence of amino acids in a nascent protein that specifies its destination. Signal sequences are normally cleaved off extracellular proteins before export from the cell. See **Figure 1.32**.

Silencer: a combination of short DNA sequence elements that suppress transcription of a gene.

Silent change: a nucleotide substitution in a coding sequence that does not alter the amino acid encoded. Silent mutations may nevertheless cause problems by interfering with splicing. See **Figure 16.4B**.

SINE (short interspersed nuclear element): a class of moderate to highly repetitive DNA sequence families, of which the best known in humans is the Alu repeat family. See **Figures 9.12**, **9.13**.

Single nucleotide polymorphism (SNP): a position in the genome where two or occasionally three alternative nucleotides are common in the population. May be pathogenic or neutral. The dbSNP database lists human SNPs, but includes some rare pathogenic variants and some variants that involve two or more contiguous nucleotides.

Single-stranded binding protein: a protein that binds and stabilizes single-stranded DNA. Important in recombination and DNA repair.

Sister chromatids: the two chromatids of a post-replication chromosome.

Small nucleolar RNA (snoRNA): a large family of small RNA molecules present in the nucleolus that act as guides to modify specific bases in other RNA molecules, especially ribosomal RNAs.

SNP: see Single nucleotide polymorphism.

SNP chip: a microarray carrying allele-specific oligonucleotides for genotyping many SNPs in a single operation. See **Box 20.1**, **Figure 15.8**.

SNV: single nucleotide variant (includes those not frequent enough to qualify as a polymorphism).

Somatic cell: any cell in the body that is not part of the germ line.

Somatic cell hybrid: an artificially produced cell in culture that contains chromosomes from two different species, e.g. human and rodent. See **Figure 7.5**.

Somatic cell nuclear transfer (SCNT): an experimental manipulation in which the nucleus of an unfertilized egg is removed and replaced by the nucleus of a somatic cell from another animal. The technique used to produce Dolly the sheep. See **Figure 8.25**.

Somatic mosaicism: mosaicism that affects only somatic cells, not the germ line.

Somites: in an embryo, paired blocks of segmental mesoderm that will establish the segmental organization of the body by giving rise to most of the axial skeleton (including the vertebral column), the voluntary muscles, and part of the dermis of the skin.

Southern blot: transfer of DNA fragments from an electrophoretic gel to a nylon or nitrocellulose membrane (filter), in preparation for a hybridization assay. See **Figure 6.15**.

Specificity: in testing, a measure of the performance of a test. Specificity = (1 – false-positive rate), see **Box 20.2**.

Spindle: the set of microtubules that move chromosomes during cell division. See **Box 2.3**.

Splice isoforms: variant mature messenger RNAs produced by alternative splicing of the same primary gene transcript. See **Figures 10.27**, **10.28**.

Spliceosome: the large ribonucleoprotein complex that splices primary transcripts to remove introns.

Splicing: cutting out the introns from an RNA primary transcript and joining together the exons. See **Figures 1.18**, **1.20**.

SSCP: single-strand conformation polymorphism. A method for scanning a short piece of DNA (up to 200 bp) for sequence variants compared to a control sample.

Start codon: in mRNA, the AUG codon at which the ribosome initiates protein synthesis.

Stem cell: a cell that can act as a precursor to differentiated cells but which retains the capacity for self-renewal.

Stem cell niches: the special locations of stem cells. See **Figure 4.15**.

Sticky end: a short single-stranded protrusion at one end of a double-stranded nucleic acid. Molecules with complementary sticky ends can associate and then be covalently joined by DNA ligase—a key step in making recombinant DNA.

Stop codon: in mRNA, a UAA, UAG, or UGA triplet. When the ribosome encounters an in-frame stop codon it dissociates from the mRNA and releases the nascent polypeptide.

Stratification: a population is stratified if it consists of several subpopulations that do not interbreed freely. Stratification is a source of error in association studies and risk estimation.

Stratified (personalized) medicine: using genomic data to inform the management and treatment of patients.

Stringency (of hybridization): the choice of conditions that will allow either imperfectly matched sequences or only perfectly matched sequences to hybridize.

Subfunctionalization: when each of a pair of duplicated genes gradually sheds different aspects of the overall function, so that they come to have rather different (but probably overlapping) specialized functions. See Figure 13.11.

Substitution rate: in evolutionary studies, the rate at which nucleotide substitutions accumulate.

Supercoiling: coiling an already coiled strand.

Super-enhancers (stretch enhancers): particularly long enhancers with multiple binding sites for regulatory factors, associated with important developmental genes.

Susceptibility gene: a gene, variation in which influences susceptibility or resistance to a complex disease.

Swadesh list: in linguistic studies, a list of supposedly universal words that can be used to identify deep relationships between languages. See Box 14.1.

Synapsis: a close functional association of two partners, e.g. homologous chromosomes in prophase I of meiosis (see Figure 2.14) or neurons in the nervous system.

Synaptonemal complex: a proteinaceous substance that helps link paired homologous chromosomes during prophase I of meiosis.

Syncytial: of a cell, containing multiple nuclei, as in skeletal muscle fiber cells (see Figure 2.8).

Synonymous (silent) substitution (or mutation): one that replaces a codon by another for the same amino acid.

Synteny: loci are syntenic if they are on the same chromosome. Syntenic loci are not necessarily linked: loci sufficiently far apart on the chromosome assort at random, with 50% recombinants.

Synthetic biology: producing and studying wholly synthetic living microorganisms.

Synthetic lethality: seen when two nonlethal changes are lethal in combination.

Systems biology: the attempt to get a full understanding of how cells and organisms function by quantitative modeling of the network of interactions between genes, pathways, and metabolism that link inputs and outputs.

T cells (T lymphocytes): a heterogeneous set of lymphocytes including T-helper cells and cytotoxic T lymphocytes that, between them, are responsible for adaptive cell-mediated immunity.

Tag-SNPs: single nucleotide polymorphisms selected because the combined genotypes of a small number of such tag-SNPs serve to identify haplotype blocks and make it unnecessary to genotype every SNP in the block. See Figure 18.4.

TALEN (transcription activator-like effector nuclease): a synthetic protein used in gene editing. See Figure 8.16.

TATA box: a short sequence, TATAAA or a close variant, that is part of the promoter of many genes that are transcribed by RNA polymerase II in a tissue-specific or stage-specific way.

Taxon: a phylogenetic group. See Figure 13.28.

Telomere: the structure at the end of a chromosome. See Figure 2.23.

Terminal differentiation: the state of a cell that has ceased dividing and has become irreversibly committed to some specialized function.

Termination codon: the UAG, UAA, or UGA codons in messenger RNA (AGA, AGG, UAA, or UAG in mitochondrial mRNA) that signal the end of the translated sequence.

Tertiary structure: the 3-dimensional structure of a polypeptide. See Table 1.7.

Third-degree relatives: the parents or children of second-degree relatives of a person, most commonly the first cousins.

Tiling path: an ordered series of partially overlapping DNA fragments that define the overall sequence.

Tissue *in situ* hybridization: hybridization of a labeled probe to RNA molecules in a tissue section to show their distribution.

Tm: see Melting temperature

Topoisomerase: an enzyme that can unwind DNA, relax coiling, or even pass one DNA double helix through another by making temporary cuts and then rejoining the ends.

Topologically-associated domains (TADs): relatively compact 500 kb–1 Mb genomic regions in the interphase cell nucleus. Enhancers act only within their own TAD.

Totipotent: in general, a cell that is able to give rise to all the cell types in an organism. For mammalian cells, the meaning extends to the ability to form all cell types in the body plus extra-embryonic and placental cells, such as is displayed by the zygote and early cleavage cells.

Trait: a character or phenotype.

***Trans*-acting:** of a regulatory factor, affecting expression of all copies of the target gene, irrespective of chromosomal location. *Trans*-acting regulatory factors are usually proteins that can diffuse to their target sites.

Transgenic: an animal or cell containing an artificially inserted gene. See Figure 8.20.

Transcription factor: DNA-binding protein that promotes transcription of genes. See Box 3.1, Section 10.2.

Transcriptome: the total set of different RNA transcripts in a cell or tissue. cf. Genome, Proteome.

Transdifferentiation: artificial conversion of a cell to a different type of cell, such as from fibroblast to neuron.

Transduction: (1) relaying a signal from a cell surface receptor to a target within a cell; (2) using recombinant viruses to introduce foreign DNA into a cell.

Transfection: direct introduction of an exogenous DNA molecule into a cell without using a vector.

Transformation (of a cell): (1) uptake by a competent bacterial cell of naked high molecular weight DNA from the environment; (2) alteration of the growth properties of a normal eukaryotic cell as a step towards evolving into a tumor cell.

Transgene: an exogenous gene that has been transfected into cells of an animal or plant. It may be present in some tissues (as in human gene therapies) or in all tissues (as in germ-line engineering, e.g. in mouse). Introduced transgenes may be episomal and be transiently expressed, or can be integrated into host cell chromosomes.

Transgenesis: genetic modification of cells that involves transfer of exogenous genetic material into the cells of an organism.

Transgenic: an animal or cell containing an artificially inserted gene (or other exogenous DNA). See Figure 8.20.

Transit amplifying cells: the immediate progeny by which stem cells give rise to differentiated cells. Transit amplifying cells go through many cycles of division, but eventually differentiate.

Translesion synthesis: DNA synthesis passing over a damaged site. Uses special low-fidelity DNA polymerases. See Table 11.2.

Translocation: transfer of chromosomal regions between nonhomologous chromosomes. See Figure 15.10.

Transposon: a (potentially) mobile genetic element. See Figures 9.12, 9.13.

Triploid: having three copies of the genome (in humans, 69 chromosomes).

Trisomy: having three copies of one particular chromosome, e.g. trisomy 21 in Down syndrome.

Trisomy rescue: in a nonviable trisomic embryo, chance loss of one of the trisomic chromosomes by mitotic nondisjunction, producing a disomic cell that can eventually form a viable embryo. A cause of uniparental disomy.

Trophoblast (or trophectoderm): outer layer of polarized cells in the blastocyst, which will go on to form the chorion, the embryonic component of the placenta.

Tropism (of virus strains): the range of different cells that a virus can infect. Depends on the extent to which different cell types display cell receptors that the virus needs to bind to in order to be able to infect cells.

Tumor-suppressor gene (TSG): a gene whose normal function is to inhibit or control cell division. TSG are typically inactivated in tumors. See Section 19.2.

Ultraconserved sequence: genomic sequences >200 bp long that are 100% conserved between human, rat, and mouse genomes. They are likely to be important regulatory elements.

Unbalanced: of a structural rearrangement, involving gain or loss of material.

Uniparental disomy (UPD): a cell or organism in which both copies of one particular chromosome pair are derived from one parent. Depending on the chromosome involved, this may or may not be pathogenic.

Unrelated: ultimately everybody is related; the word is used in this book to mean people who do not have an identified common ancestor in the last four or so generations.

Unrooted evolutionary tree: one that shows relationships within the tree, but does not attempt to identify a common ancestor. See **Figure 13.29**.

Untranslated region (5′ UTR, 3′ UTR): regions at the 5′ end of mRNA before the AUG translation start codon, or at the 3′ end after the UAG, UAA, or UGA stop codon. See **Figure 1.25**.

Upstream: in the 5′ direction on the sense strand.

Variable expression: variable extent or intensity of phenotypic signs among people with a given genotype. See for example, **Figure 5.10**.

Variable region (of an immunoglobulin): the N-terminal portion of an Ig protein that is comparatively variable in sequence (see **Figure 3.24**), and is largely comprised of a single globular domain on each chain (**Figure 3.25**).

Variant of uncertain significance (VUS): a variant where the present state of knowledge does not allow a judgement whether it is clinically significant. cf. Incidental finding.

Vector: a nucleic acid that is able to replicate and maintain itself within a host cell, and that can be used to confer similar properties on any sequence covalently linked to it. See **Table 6.1**.

Virtual gene panel: when analyzing exome or whole genome data from a patient, deciding to check only a specific list of potentially relevant genes (to ease analysis and lessen the risk of incidental findings).

Watson–Crick rules: describe the normal base-pairing in double-stranded nucleic acid: A with T (or U); G with C.

Western blotting: a procedure analogous to Southern blotting but involving proteins separated by charge and size on electrophoretic gels, blotted onto a membrane and detected using antibodies or stains.

Whole genome amplification: *in vitro* amplification of all the DNA of a genome as a way of making a precious small genomic DNA sample go further. Done using isothermal amplification with phi29 DNA polymerase.

Wobble pairing: a special relaxed base-pairing that occurs between the 3′ nucleotide of a codon and a tRNA anticodon. See **Table 1.5**.

X-inactivation (lyonization): the epigenetic inactivation of all but one of the X chromosomes in the cells of mammals that have more than one X. See **Figures 10.13**, **10.14**.

X-inactivation center: the location on the proximal long arm of the human X chromosome from which the spreading X-inactivation is initiated. See Section 10.4.

XY body: in male meiosis, the X and Y chromosomes associate in a condensed body of heterochromatin.

Yeast artificial chromosome: a vector able to propagate inserts of up to 2 megabases in yeast cells. See **Box 7.1**.

Yeast two-hybrid screening: a technique for identifying protein–protein interactions. Proteins that physically associate are identified by their ability to bring together two separated parts of a transcription factor, and so stimulate transcription of a reporter gene in yeast cells. See **Box 9.4**.

Zinc finger: a structure found in many DNA-binding proteins. See **Box 3.1**.

Zinc finger nucleases: synthetic enzymes that combine an endonuclease module with a sequence-specific targeting module, so as to cleave DNA at a selected sequence. See **Figure 8.16**.

Zona pellucida: a layer of glycoprotein that surrounds an unfertilized egg and acts as a barrier to fertilization. See **Figure 4.3**.

Zoo-blot: a Southern blot containing DNA samples from a range of different species, used to check for conserved sequences.

Zygote: the fertilized egg cell.

Index

Notes: Page entries followed by “f” refer to figures and those followed by “t” and “b” refer to tables and boxed material respectively. A page entry such as 177–8b indicates a box that continues on two successive pages. Page entries followed by “ff.” refer to substantial number of pages beginning on the page that is referenced. Please note that all entries refer to humans unless otherwise stated.