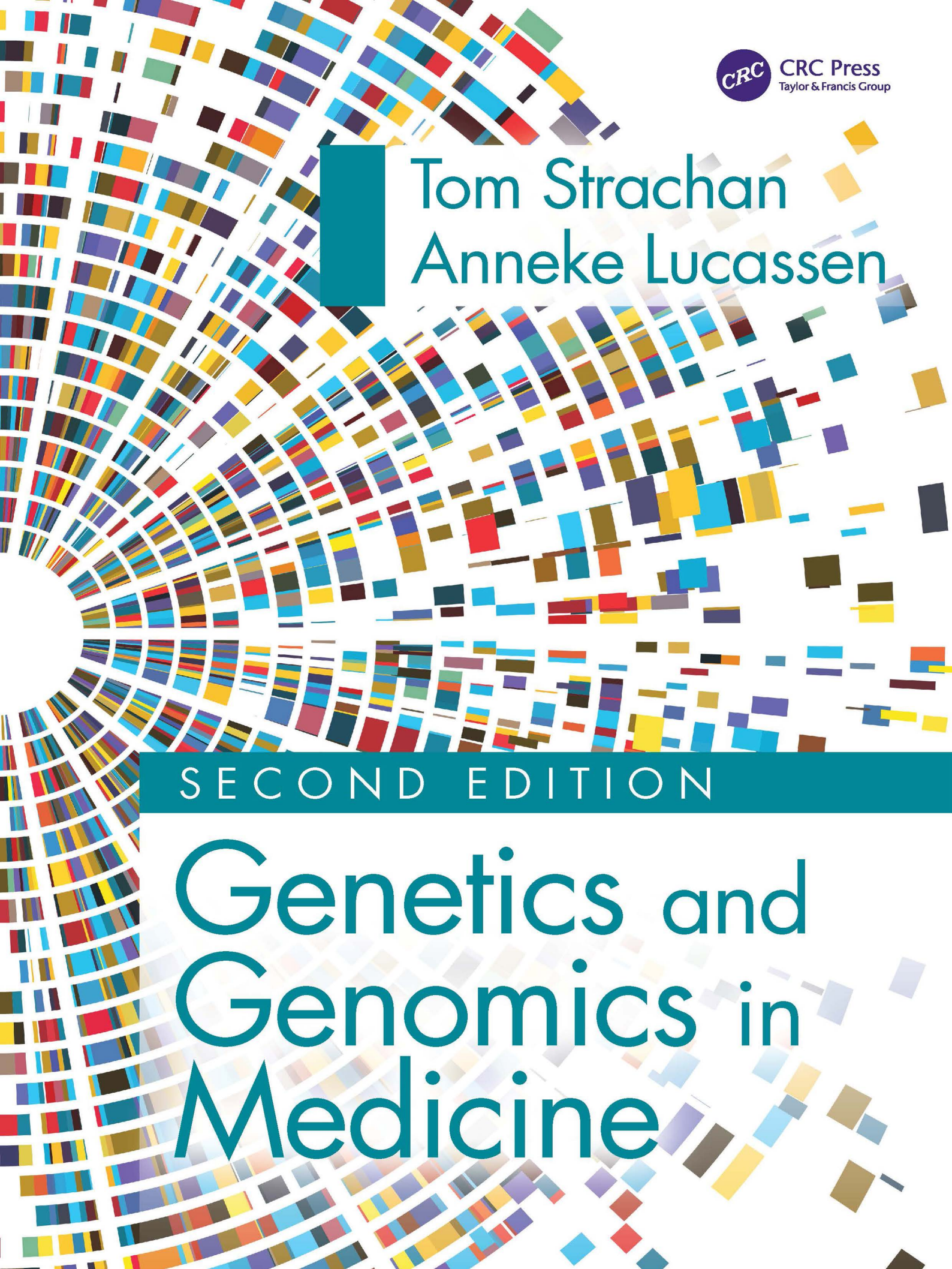
CRC Press
Taylor & Francis Group
Tom Strachan
Anneke Lucassen
SECOND EDITION
Genetics and Genomics in Medicine

GENETICS AND GENOMICS IN MEDICINE

SECOND EDITION

GENETICS AND GENOMICS IN MEDICINE

TOM STRACHAN AND
ANNEKE LUCASSEN

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Second edition published 2023
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-0-367-49082-9 (hbk)
ISBN: 978-0-367-49081-2 (pbk)
ISBN: 978-1-003-04440-6 (ebk)

DOI: 10.1201/b22853

Typeset in Utopia
by Apex CoVantage, LLC

Access the Support Material at: https://www.routledge.com/9780367490812

Contents

Preface

A rationale for establishing the first edition of *Genetics and Genomics in Medicine* was the suspicion that genomewide analyses might transform medicine. Using Sanger dideoxy sequencing the international Human Genome Project took about 13 years to deliver an almost complete genome sequence in 2003. Subsequent technological developments—first, genome-wide microarray technologies and then massively parallel DNA sequencing—have certainly transformed genome analysis, permitting genome data in hours, not years.

The preface to the first edition of this book also included this question: might we soon live in societies where genome sequencing of citizens becomes the norm? Well, that day seems much closer now as millions of people have their genome sequenced, and debate has begun on whether population neonatal genome sequencing should be considered. The genome sequencing revolution found early major applications in medical genetics, then hematology and oncology, but is now being increasingly applied across multiple other medical disciplines. Various national genomic medicine initiatives have recently been established and, in 2020, NHS England became the first national health service to offer whole genome sequencing to patients as part of routine care.

In this book we try to summarize pertinent knowledge, and to structure it in the form of principles, rather than seek to compartmentalize information into chapters on topics such as epigenetics, evolutionary genetics, immunogenetics, pharmacogenetics, and so on. To help readers find broad topics that might be dealt with in two or more chapters, we provide a road map on the inside front cover that charts how some broad themes are distributed between different chapters.

We start with three introductory chapters that provide basic background details. Chapters 1 and 2 cover the fundamentals of DNA, chromosomes, the cell cycle, human genome organization and gene expression. Chapter 3 introduces the basics of three core molecular genetic approaches used to manipulate DNA: DNA amplification (by DNA cloning or PCR), nucleic acid hybridization, and DNA sequencing, but we delay bringing in applications of these fundamental methods until later chapters, setting them against appropriate contexts that directly explain their relevance.

The next three chapters provide some background principles at a higher level. In Chapter 4, we take a broad look at general principles of genetic variation, including DNA repair mechanisms and some detail on functional variation (but we consider how genetic variation contributes to disease in later chapters, notably chapters 7, 8 and 10). Chapter 5 takes a look at how genes are transmitted in families and at allele frequencies in populations. Chapter 6 moves from the basic principles of gene expression covered in chapter 2 to explaining how genes are regulated by a wide range of protein and noncoding RNA regulators, and

the central role of regulatory sequences in both DNA and RNA. In this chapter, too, we outline the principles of chromatin modification and epigenetic regulation and explain how aberrant chromatin structure underlies many single gene disorders.

The remainder of the book is largely devoted to clinical applications. We explain in chapter 7 how chromosome abnormalities arise and their consequences, and how mutations and large-scale DNA changes can directly cause disease. In chapter 8, we look at how genes underlying single gene disorders are identified, and also how genetic variants conferring susceptibility to complex diseases are identified. Then we consider the ways in which genetic variants, epigenetic dysregulation and environmental factors all make important contributions to complex diseases. Chapter 9 briefly covers the wide the range of approaches for treating genetic disorders, before examining in detail how genetic approaches are used directly and indirectly in treating disease. In this chapter, too, we examine how genetic variation affects how we respond to drug treatment. Chapter 10 deals with cancer genetics and genomics and explains how cancers arise from a combination of abnormal genetic variants and epigenetic dysregulation. Finally, Chapter 11 takes a broad look at diagnostic applications (and the exciting applications offered by new genome-wide technologies), plus ethical considerations in diagnosis and in some novel therapies.

Important recent advances have been made in applying genetic and genomic technologies to understanding pathogenesis, and in developing novel genetic testing methods, (including noninvasive ones), and novel treatments. There has been significant improvement, too, in pharmacogenomic approaches and in prenatal and preconception options to avoid serious genetic disease. Now we are no longer bound by the old approach of starting with a phenotype and then searching for a confirmatory genotype but can invert the process to predict phenotypes over a lifetime from a genotype. But challenges remain. Predicting phenotypes over a lifetime from a genotype, for example, is rarely clear-cut; the more we test without medical indications, the less likely we will predict diseases accurately. And, while acquiring genetic and genomic data is no longer the major rate-limiting step it was, data interpretation has become a huge challenge given the inherent complexities of interpreting the 4–5 million variants in a person's genome and their implications for [ill] health.

Mainstreaming of genomic medicine—placing it at the center of healthcare—may be appealing, but its utility can be expected to be limited in the first instance to rare diseases and some easily studied cancers. Complex genetic disease is another matter. Genomewide association studies have undoubtedly been successful, especially in improving our understanding of the molecular pathways in a wide range of complex genetic diseases, but they have their limitations. Increasingly, attention has been devoted to finding rare variants by genomewide sequencing (with considerable recent success in some diseases, such as schizophrenia), and in investigating copy number variants. To properly appreciate the complexity of common genetic disease will require more information, too, from other approaches, investigating modifier genes, environmental factors and so on, and reliance on phenotyping data from large population biobanks will be important.

The familial nature of much genetic information also poses challenges to many modern healthcare services for which there are no clear off-the-shelf solutions. Confidentiality in medicine remains important, yet shared familial inheritances may need disclosing at times, just as we attempted to trace contacts exposed to COVID-19. Sustainability aspects of long-term mass data storage are yet to be examined in any depth, and the lack of population diversity in most of the world's genomic repositories, and thus our understanding of genomic variation, needs urgent attention.

We have tried to convey the excitement of fast-moving research in genetics and genomics and their clinical applications, while explaining how the progress has been achieved. By weaving the ethical, legal and social aspects inherent in

these developments throughout the text we hope to provide the reader with a realistic lens through which to view the promising developments in genetics and genomics. There is a long way to go, notably in understanding complex disease and in developing effective treatments for many disorders. But some impressive recent therapeutic advances, and new technological developments such as the prime editing and base editing refinements to CRISPR-Cas genome editing, have engendered an undeniable sense of excitement and optimism. How far will we move from the commonplace one-size-fits-all approach to disease treatment toward an era of personalized or precision medicine? At the very least, we might expect an era of stratified medicine where, according to the genetic variants exhibited by patients with a specific disease, different medical actions are taken.

We would like to thank the staff at CRC Press and Naughton Project Management Ltd: Jo Koster, Jordan Wearing and Nora Naughton, who have undertaken the job of converting our drafts into the finished product. We are also grateful to our family members: Meryl, Alex, James, Tim, Emily and Isobel for their steadfast support.

LITERATURE ACCESS

We live in a digital age and, accordingly, we have sought to provide electronic access to information. To help readers find references cited under Further Reading we provide the relevant PubMed identification (PMID) numbers for the individual articles—see also the PMID glossary item. We would like to take this opportunity to thank the US National Center for Biotechnology Information (NCBI) for their invaluable PubMed database that is freely available at: http://www.ncbi.nlm.nih.gov/pubmed/. Readers who are interested in new research articles that have emerged since publication of this book, or who might want to study certain areas in depth, may wish to take advantage of literature citation databases such as the freely available Google Scholar database (scholar.google.com).

For background information on single gene disorders, we often provide reference numbers to access OMIM, the Online Mendelian Inheritance in Man database (http://www.omim.org). For the more well-studied of these disorders, individual chapters in the University of Washington's GeneReviews series are highly recommended. They are electronically available at the NCBI's Bookshelf within its PubMed database. For convenience, we have given the PubMed Identifier (PMID) for individual articles that we refer to from the GeneReviews series. Note that all GeneReviews articles can be accessed through PubMed at PMID 20301295, where there is an alphabetic listing of all disorders covered by GeneReviews.

Tom Strachan and Anneke Lucassen

Acknowledgements

In writing this book, we have benefited greatly from the advice of many geneticists, biologists and clinicians. We are also grateful to various colleagues who contributed clinical profiles and/or laboratory data for case studies, or who advised on the contents of chapters and/or commented on some aspects of the text, notably the following: Chiara Bettolo, David Bourn, Gareth Breese, Heather Cordell, Jordi Diaz-Manera, Shaun Haigh, Rachel Horton, Majlinda Lako, Richard Martin, Ciaron McAnulty, Robert McFarland, Sabine Specht, Miranda Splitt, and Volker Straub.

1 Fundamentals of DNA, chromosomes, and cells

CONTENTS

Three structures are the essence of life: cells, chromosomes, and nucleic acids. Cells receive basic sets of instructions from DNA molecules that must also be transmitted to successive generations. And DNA molecules work in the context of larger structures: chromosomes.

Many organisms consist of single cells that can multiply quickly. They are genetically relatively stable, but through changes in their DNA they can adapt rapidly to changes in environmental conditions. Others, including ourselves, animals, plants, and some types of fungi, are multicellular.

Multicellularity offers specialization and complexity: individual cells can be assigned different functions, becoming muscle cells, neurons, or lymphocytes, for example. All the different cells in an individual arise originally from a single cell, and so all nucleated cells carry the same DNA sequences. During development, however, the DNA structure within chromosomes is changed to allow specific changes in gene expression that determine a cell's identity, whether it be a muscle cell or a neuron, for example.

Growth during development and tissue maintenance requires cell division. When a cell divides to produce daughter cells, our chromosomes and the underlying DNA sequences must undergo coordinated duplication and then be carefully segregated to the daughter cells.

Some of our cells can carry our DNA to the next generation. When that happens, chromosomes swap segments and DNA molecules undergo significant changes that make us different from our parents and from other individuals.

DOI: 10.1201/b22853-1

1.1 THE STRUCTURE AND FUNCTION OF NUCLEIC ACIDS

General concepts: the genetic material, genomes, and genes

Nucleic acids provide the *genetic material* of cells and viruses. They carry the instructions that enable cells to function in the way that they do and to divide, allowing the growth and reproduction of living organisms. Nucleic acids also control how viruses function and replicate. As we describe later, viruses can be highly efficient at inserting genes into human cells, and modified viruses are widely used in gene therapy.

Nucleic acids are susceptible to small changes in their structure (**mutations**). Occasionally, that can change the instructions that a nucleic acid gives out. The resulting genetic variation, plus mechanisms for shuffling the genetic material from one generation to the next, explains why individual organisms of the same species are nevertheless different from each other. And genetic variation is the substrate that evolutionary forces work on to produce different species. (But note that the different types of cell in a single multicellular organism cannot be explained by genetic variation—the cells each contain the same DNA and the differences in cell types must arise instead by **epigenetic** mechanisms.)

In all cells the genetic material consists of double-stranded DNA in the form of a double helix. (Viruses are different. Depending on the type of virus, the genetic material may be double-stranded DNA, single-stranded DNA, double-stranded RNA, or single-stranded RNA.) As we describe below, DNA and RNA are highly related nucleic acids. RNA is functionally more versatile than DNA (it is capable of self-replication and individual RNA sequences can also serve as templates to make a protein, or act as regulators of gene expression). RNA is widely believed to have developed at a very early stage in evolution. Subsequently, DNA evolved; being chemically much more stable than RNA, it was more suited to being the store of genetic information in cells.

Genome is the collective term for all the *different* DNA molecules within a cell or organism. In prokaryotes—simple unicellular organisms, such as bacteria, that lack organelles—the genome usually consists of just one type of circular double-stranded DNA molecule that can be quite large and has a small amount of protein attached to it. A very large DNA-protein complex such as this is traditionally described as a **chromosome**.

Eukaryotic cells are more complex and more compartmentalized (containing multiple organelles that serve different functions), and they have multiple different DNA molecules. As we will see below, for example, the cells of a man have 25 different DNA molecules but a woman's cells have a genome made up of 24 types of DNA molecule.

In our cells—and in those of all animals and fungi—the genome is partitioned between the nucleus and the mitochondria. Most of the DNA is found in the nucleus, existing as extremely long linear DNA molecules complexed with a variety of different proteins and some types of RNA to form highly organized chromosomes. However, in mitochondria there is just one type of small circular DNA molecule that is largely devoid of protein. (In plant cells, chloroplasts also have their own type of small circular DNA molecule.)

Genes are the DNA segments that carry the genetic information to make proteins or functional noncoding RNA molecules within cells. The great bulk of the genes in a eukaryotic cell are found in the chromosomes of the nucleus; just a few genes are found in the small mitochondrial or chloroplast DNA molecules.

The underlying chemistry of nucleic acids

Each nucleic acid strand is a polymer, a long chain containing many sequential copies of a simple repeating unit, a **nucleotide**. Each nucleotide in turn consists of a sugar molecule, to which is attached a nitrogenous base and a phosphate group.

In DNA the sugar is deoxyribose, which has five carbon atoms that are labeled 1′ (one *prime)* to 5′. It is very closely related to ribose, the sugar molecule found in RNA—the only difference is that a hydroxyl (-OH) group at carbon 2′ of ribose is replaced by a hydrogen atom in deoxyribose (**Figure 1.1**).

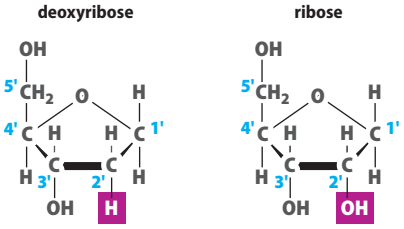

Figure 1.1 Structure of deoxyribose (left) and ribose (right). The f ve carbon atoms are numbered 1′ (one *prime)* to 5′ (f ve *prime)*. The magenta shading is meant to signify the only structural dif erence between deoxyribose (the sugar found in DNA) and ribose (the sugar found in RNA): ribose has a hydroxyl (-OH) group in place of the highlighted hydrogen atom attached to carbon 2′ of deoyribose. The more precise name for deoxyribose is therefore 2′-deoxyribose.

Individual nucleotides are joined to their neighbors by a negatively charged phosphate group that links the sugar components of the neighboring nucleotides. As a result, nucleic acids are polyanions, and have a *sugar-phosphate backbone* with bases bonded to the sugars. As explained in **Box 1.1**, the sugar-phosphate backbone of each nucleic acid strand is asymmetric and the ends of each strand are asymmetric, giving direction to each strand.

BOX 1.1 5′ AND 3′ ENDS, AND STRAND ASYMMETRY OF NUCLEIC ACIDS

In a nucleic acid strand each phosphate group links carbon atom 3′ from the sugar on one nucleotide to a carbon 5′ on the sugar of a neighboring nucleotide. Internal nucleotides will therefore be linked through both carbon 5′ and carbon 3′ of the sugar to the neighboring nucleotides on opposing sides. However, the nucleotides at the extreme ends of a DNA or RNA strand will have different functional groups. At one end, the **5′ end**, the nucleotide has a terminal sugar with a carbon 5′ that is not linked to another nucleotide and is capped by a phosphate group; at the other end, the **3′ end**, the terminal nucleotide has a sugar with a carbon 3′ that is capped by a hydroxyl group (**Figure 1**).

The resulting asymmetry between the two ends of a nucleic acid give it a direction. That is important in packing a nucleic acid because when two single nucleic acid strands pair up to make a stable duplex, they must be *anti-parallel*: the 5′ → 3′ direction of one strand must be opposite to that of its partner strand. And direction is important for synthesis of a nucleic acid: a growing nucleic acid strand always extends in a 5′ → 3′ direction.

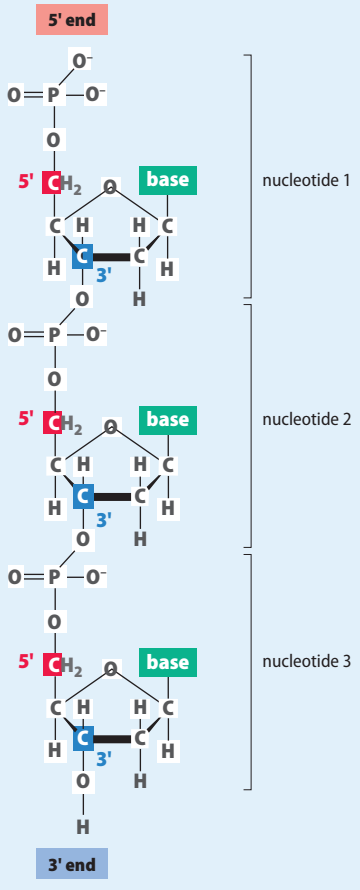

Figure 1 Repeating structure and asymmetric 5′ and 3′ ends in nucleic acids.

Unlike the sugar molecules, the nitrogenous bases come in four different types, and it is the sequence of different bases that identifies the nucleic acid and its function. Two of the bases have a single ring based on carbon and nitrogen atoms (a **pyrimidine**) and two have a double ring structure (a **purine**). In DNA the two purines are adenine (A) and guanine (G), and the two pyrimidines are cytosine (C) and thymine (T). The bases of RNA are very similar; the only difference is that in place of thymine there is a very closely related base, uracil (U) (**Figure 1.2**).

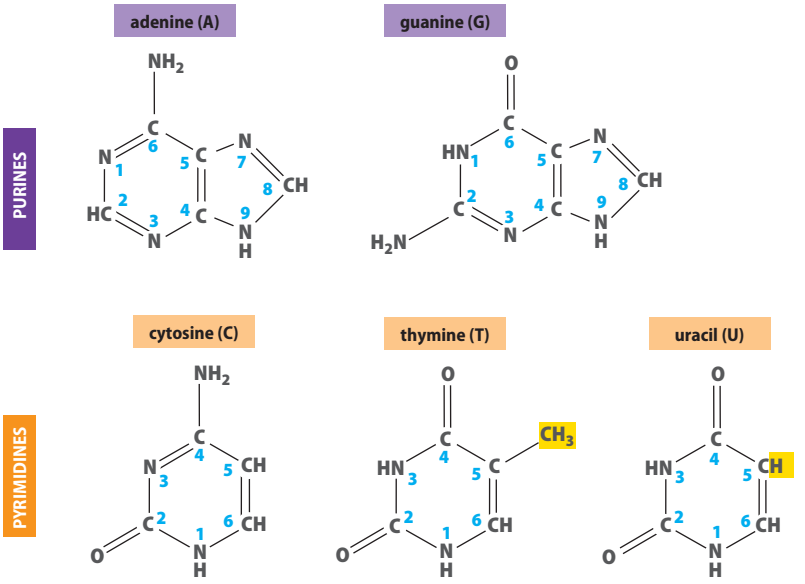

Figure 1.2 Structure of the bases found in nucleic acids. Adenine and guanine are purines with two interlocking rings based on nitrogen and carbon atoms (numbered 1 to 9 as shown). Cytosine and thymine are pyrimidines with a single ring. Adenine, cytosine, and guanine are found in both DNA and RNA, but the fourth base is thymine in DNA and uracil in RNA (they are closely related bases—carbon atom 5 in thymine has an attached methyl group, but in uracil the methyl group is replaced by a hydrogen atom).

Base pairing and the double helix

Cellular DNA exists in a double-stranded (or *duplex*) form, in which the two very long single DNA strands are wrapped round each other. In the resulting double helix each base on one DNA strand is noncovalently linked (by hydrogen bonding) to an opposing base on the opposite DNA strand, forming a **base pair**. However, 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. (Only two types of base pairs are normally tolerated in double-stranded DNA: A–T and G–C base pairs.) G–C base pairs, which are held together by three hydrogen bonds, are stronger than A–T base pairs, which are held together by two base pairs; see **Figure 1.3**.

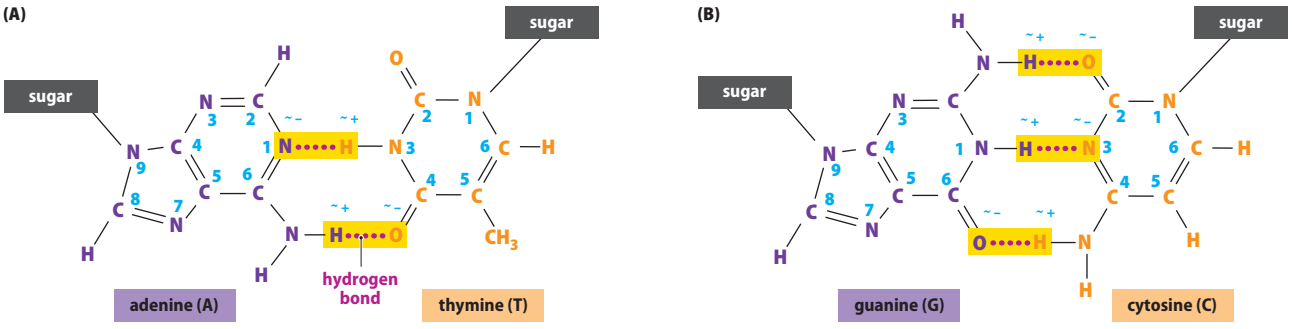

Figure 1.3 Structure of base pairs. In the A–T base pair shown in (A), the adenine is connected to the thymine by two hydrogen bonds. In the G–C base pair shown in (B), three hydrogen bonds link the guanine to the cytosine; a G–C base pair is therefore stronger than an A–T base pair. δ^+ and δ^- indicate fractional positive charges and fractional negative charges.

There is one additional restriction on how two single-stranded nucleic acids form a double-stranded nucleic acid. In addition to a sufficient degree of base pairing, for a duplex to form, the two single strands must be anti-parallel; that is, the 5′ → 3′ direction of one strand is the opposite of the 5′ → 3′ direction of the other strand.

Two single nucleic acid strands that can form a double helix with perfect base matching (according to the base pairing rules given above) are said to have **complementary sequences**. As a result of base pairing rules, the sequence of one DNA strand in a double helix can immediately be used to predict the base sequence of the complementary strand (**Box 1.2**). Note that base pairing can also occur in RNA; when an RNA strand participates in base pairing, the base pairing rules are more relaxed (see Box 1.2).

DNA replication and DNA polymerases

Base pairing rules also explain the mechanism of DNA replication. In preparation for new DNA synthesis before cell division, each DNA double helix must be unwound using a helicase. During the unwinding process the two individual single DNA strands become available as templates for making complementary DNA strands that are synthesized in the 5′ → 3′ direction (**Figure 1.4**).

BOX 1.2 BASE PAIRING PREVALENCE, SEQUENCE COMPLEMENTARITY, AND SEQUENCE NOTATION FOR NUCLEIC ACIDS

THE PREVALENCE OF BASE PAIRING

The DNA of cells—and of viruses that have a double-stranded DNA genome—occurs naturally as double helices in which base pairing is restricted to A–T and C–G base pairs.

Double-stranded RNA also occurs naturally in the genomes of some kinds of RNA viruses. Although cellular RNA is often single-stranded, it can also participate in base pairing in different ways. Many single-stranded RNAs have sequences that allow intramolecular base pairing—the RNA bends back upon itself to form local double-stranded regions for structural stability and/or for functional reasons. Different RNA molecules can also transiently base pair with each other over short to moderately long regions, allowing functionally important interactions (such as base pairing between messenger RNA and transfer RNA during translation, for example; see Section 2.1). G–U base pairs are allowed in RNA–RNA base pairing, in addition to the standard A–U and C–G base pairs.

RNA–DNA hybrids also form transiently in different circumstances. They occur when a DNA strand is transcribed to give an RNA copy, for example, and when an RNA is reverse transcribed to give a DNA copy.

SEQUENCE COMPLEMENTARITY

Double-helical DNA within cells shows perfect base matching over extremely long distances, and the two DNA strands within a double helix are said to exhibit **base complementarity** and to have *complementary sequences*. Because of the strict base pairing rules, knowing the base sequence of just one DNA strand is sufficient to immediately predict the sequence of the complementary strand, as illustrated below.

SEQUENCE NOTATION

Because the base sequence of a nucleic acid governs its biological properties it is customary to define a nucleic acid by its base sequence, which is always written in the 5′ → 3′ direction. While a single-stranded oligonucleotide sequence 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, it is simpler to write it just as CGACCAT.

For a double-stranded DNA the sequence of just one of the two strands is needed (the sequence of the complementary strand can immediately be predicted by the base pairing rules given above). 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'

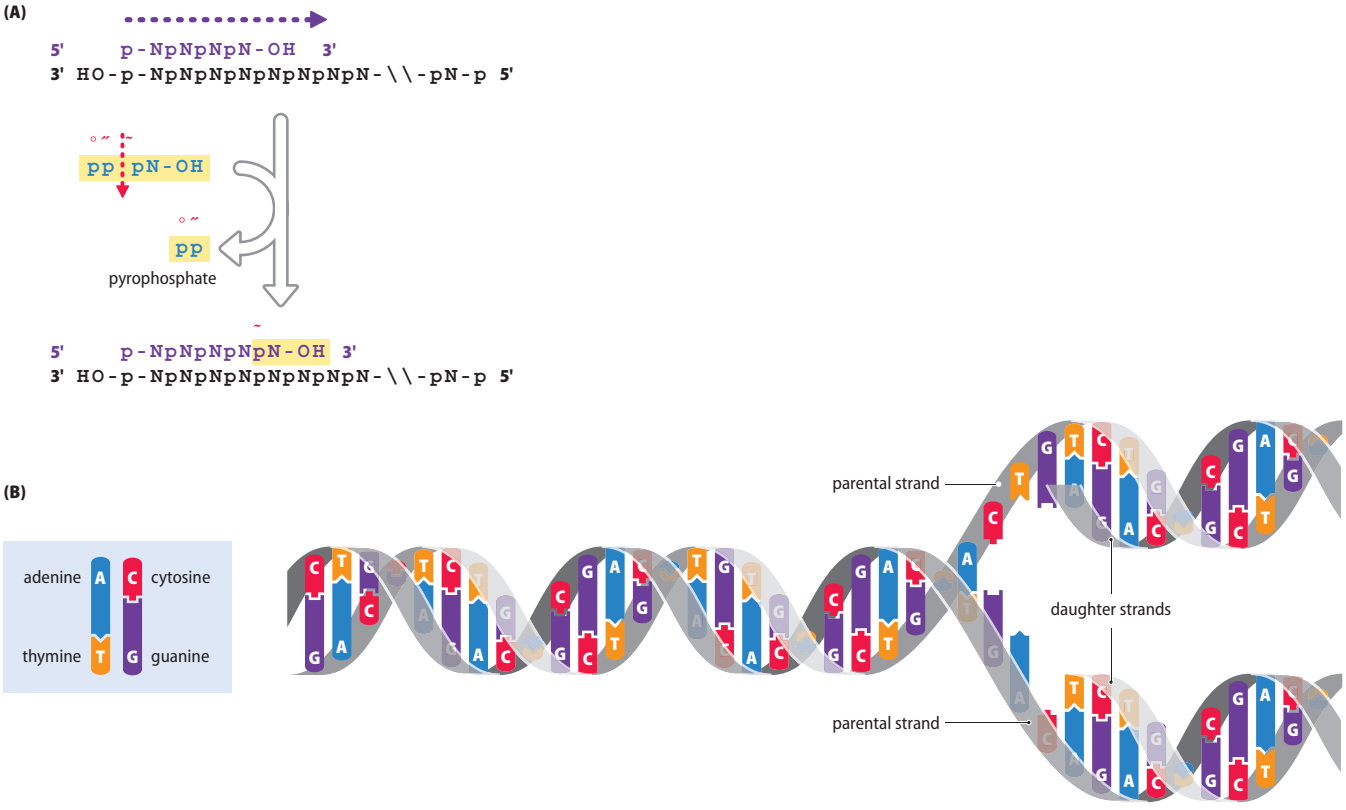

Figure 1.4 DNA synthesis and replication. (A) DNA synthesis. Using a pre-existing DNA strand (black) as a template, a new DNA strand (purple) is synthesized in a 5′ → 3′ direction (dashed arrow) using a DNA polymerase to insert successive dNMPs obtained by cleaving the two external phosphates (α and β), to give pyrophosphate residue (which is discarded). (B) DNA replication. The parental DNA duplex consists of two complementary DNA strands that unwind to serve as templates for the synthesis of new complementary DNA strands (daughter 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.

DNA replication therefore uses one double helix to make two double helices, each containing one strand from the parental double helix and one newly synthesized strand (semi-conservative DNA replication). Because DNA synthesis occurs only in the 5′ → 3′ direction, one new strand (the *leading strand*) can be synthesized continuously; the other strand (the *lagging strand*) needs to be synthesized in pieces, known as Okazaki fragments (**Figure 1.5**).

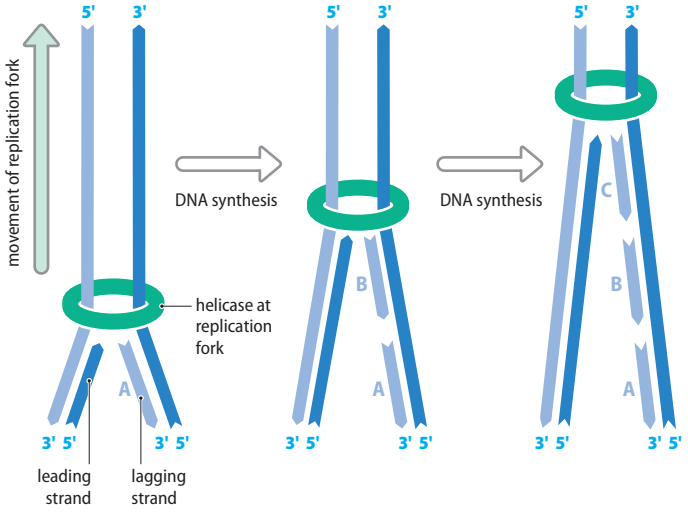

Figure 1.5 Semi-discontinuous DNA replication. The enzyme DNA helicase opens up a **replication fork**, where synthesis of new daughter DNA strands can begin. The overall direction of movement of the replication fork matches that of the continuous 5′ → 3′ synthesis of one daughter DNA strand, the *leading 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 will later be stitched together by a DNA ligase.

Mammalian cells have very many kinds of DNA-dependent DNA polymerases that serve a variety of roles, including DNA replication initiation, synthesis of the leading and lagging strands, and also, as described in Section 4.2, multiple roles in DNA repair. Our cells also contain specialized DNA polymerases that use RNA as a template to synthesize a complementary DNA; see **Table 1.1**.

TABLE 1.1 CLASSICAL DNA-DEPENDENT AND RNA-DEPENDENT DNA POLYMERASES OF MAMMALIAN CELLS

DNA polymerases	Roles
Classical DNA-dependent DNA polymerases	**Standard DNA replication and/or DNA repair**
α (alpha)	initiates DNA synthesis (at replication origins, and also when priming the synthesis of Okazaki fragments on the lagging strand)
δ (delta) and ε (epsilon) β (beta)	major nuclear DNA polymerases and multiple roles in DNA repair
γ (gamma)	base excision repair (repair of deleted bases and simply modifed bases) dedica ted to mitochondrial DNA synthesis and mitochondrial DNA repair
RNA-dependent DNA polymerases	**Genome evolution and telomere function**
Retrosposon reverse transcriptase	occasionally converts mRNA and other RNA into complementary DNA, which can integrate elsewhere into the genome; can occasionally give rise to new genes and new exons, and so on.
TERT (telomerase reverse transcriptase)	replicates DNA at ends of linear chromosomes, using an RNA template

Note: The classical DNA-dependent DNA polymerases are high-fdelit y polymerases—they insert the correct base with high accuracy; however, our cells also have many non-classical DNA-dependent DNA polymerases that exhibit comparatively low fdelit y of DNA replication. We will consider the non-classical DNA polymerases in Chapter 4, because of their roles in certain types of DNA repair and in maximizing the variability of immunoglobulins and T-cell receptors.

Genes, transcription, and the central dogma of molecular biology

As a repository of genetic information, DNA must be stably *transmitted* from mother cell to daughter cells, and from individuals to their progeny; DNA replication provides the necessary mechanism. But within the context of individual cells, the genetic information must also be *interpreted* to dictate how cells work. **Genes** are discrete segments of the DNA whose sequences are selected for this purpose, and gene expression is the mechanism whereby genes are used to direct the synthesis of two kinds of product: RNA and proteins.

The first step of gene expression is to use one of the two DNA strands as a template for synthesizing an RNA copy whose sequence is complementary to the selected template DNA strand. This process is called *transcription*, and the initial RNA copy is known as the primary transcript (**Figure 1.6**). Subsequently, the primary transcript undergoes different processing steps, eventually giving a mature RNA that belongs to one of two broad RNA classes:

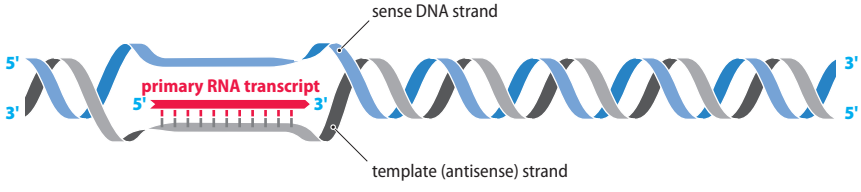

Figure 1.6 Transcription. Transcription results in the synthesis of an RNA transcript in the 5′ → 3′ direction. The nucleotide sequence of the primary RNA transcript is complementary to that of the *template strand* and so is identical to that of the *sense strand,* except that U replaces T. Note: for simplif cation, the diagram does not show the coiling of the RNA transcript around the template DNA strand to form a double helix.

- *Coding RNA*. RNAs in this class contain sequences that direct the synthesis of polypeptides (the major component of proteins) in a process called translation. This type of RNA has traditionally been called a messenger RNA (mRNA) because it carries genetic instructions to be decoded by the protein synthesis machinery.
- *Noncoding RNA*. All other mature functional RNAs fall into this class, and here the RNAs, not proteins, are the functional endpoint of gene expression. Noncoding RNAs have a variety of roles in cells, as described in later chapters.

In all forms of life, genetic information is interpreted in what initially seemed to be one direction only: DNA → RNA → protein, a principle that became known as the central dogma of molecular biology. However, certain DNA polymerases, known as reverse transcriptases, found initially in certain types of viruses, can reverse the flow of genetic information by making a DNA copy of an RNA molecule. The cells of complex organisms also have their reverse transcriptases, as described below. In addition, RNA can sometimes also be used as a template to make a complementary RNA copy. So, although genetic information in cells mostly flows from DNA to RNA to protein, the central dogma is no longer strictly valid.

We will explore gene expression (including protein synthesis) in greater detail in Chapter 2. And in Chapter 6 we will focus on both genetic and epigenetic regulation of gene expression.

1.2 THE STRUCTURE AND FUNCTION OF CHROMOSOMES

In this section we consider general aspects of the structure and function of our chromosomes that are largely shared by the chromosomes of other complex multicellular organisms. We will touch on human chromosomes when we consider aspects of the human genome in Chapter 2, when we first introduce the banded pattern of human chromosomes. In Chapter 7 we consider how disease-causing chromosome abnormalities arise. We describe the methodology and the terminology of human chromosome banding in Box 7.2, and diagnostic chromosome analyses in Chapter 11.

Why we need highly structured chromosomes, and how they are organized

Before replication, each chromosome in the cells of complex multicellular organisms normally contains a single, immensely long DNA double helix. For example, an average-sized human chromosome contains a single DNA double helix that is about 4.8 cm long with 140 million nucleotides on each strand; that is, 140 million base pairs (140 megabases (Mb)) of DNA.

To appreciate the difficulty in dealing with molecules this long in a cell only about 10 μm across, imagine a model of a human cell 1 meter across (a 10^5-fold increase in diameter). Now imagine the problem of fitting into this 1-meter-wide cell 46 DNA double helices that when scaled up by the same factor would each be just 0.2 mm thick but on average 4.8 km (about three miles) long. Then there is the challenge of replicating each of the DNA molecules and arranging for the cell to divide in such a way that the replicated DNA molecules are segregated equally into the two daughter cells. All this must be done in a way that avoids any tangling of the long DNA molecules.

To manage nuclear DNA molecules efficiently and avoid any tangling, they are complexed with various proteins and sometimes noncoding structural RNAs to form **chromatin** that undergoes different levels of coiling and compaction to

form chromosomes. In interphase—the stages of the cell cycle other than mitosis (see Section 1.3)—the nuclear DNA molecules are still in a very highly extended form and normally the very long slender interphase chromosomes remain invisible under the light microscope. But even in interphase cells, the 2 nm-thick double helix is subject to at least two levels of coiling. First, the double helix is periodically wound round a specialized complex of positively charged histone proteins to form a 10 nm nucleosome filament. The nucleosome filament is then coiled into a 30 nm chromatin fiber that undergoes looping and is supported by a scaffold of nonhistone proteins (**Figure 1.7**).

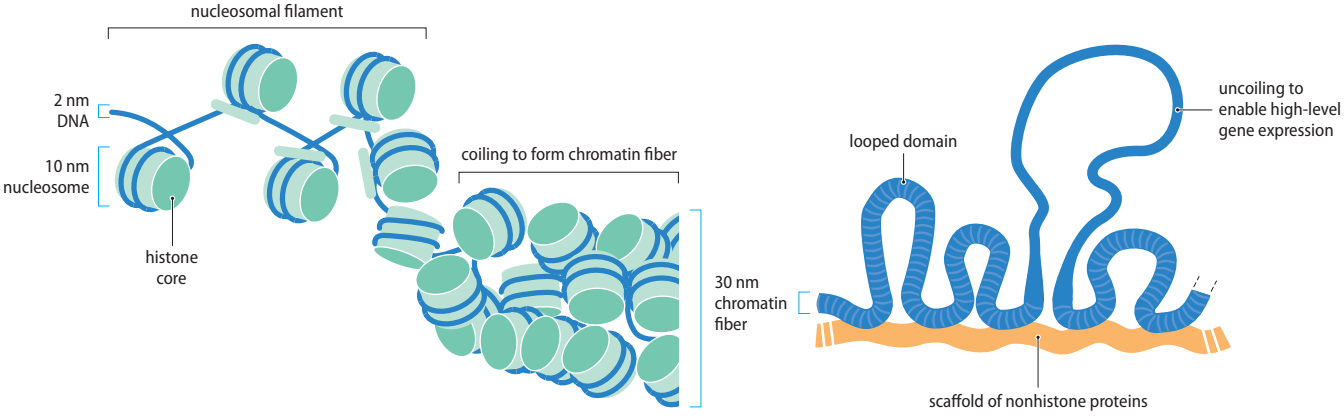

Figure 1.7 From DNA double helix to interphase chromatin. Binding of basic histone proteins causes the 2 nm DNA double helix to undergo coiling, forming f rst a 10 nm f lament studded with nucleosomes that is further coiled to give a 30 nm chromatin f ber. In interphase, the chromatin f ber is organized in looped domains, each containing about 50–200 kilobases of DNA, that are attached to a central scaf old of nonhistone proteins. High levels of gene expression require local uncoiling of the chromatin f ber to give the 10 nm nucleosomal f laments. The diagram does not show structural RNAs that can be important in chromatin. (Adapted from Grunstein M [1992] *SciAm* 267:68–74; PMID 1411455. With permission from Macmillan Publishers Ltd; and Alberts B, Johnson A, Lewis J et al. [2008] Molecular Biology of the Cell, 5th ed. Garland Science.)

During interphase most chromatin exists in an extended state (**euchromatin**) that is dispersed through the nucleus. Euchromatin is not uniform, however—some euchromatic regions are more condensed than others, and genes may or may not be expressed, depending on the cell type and its functional requirements. Some chromatin, however, remains highly condensed throughout the cell cycle and is generally genetically inactive (**heterochromatin**).

As cells prepare to divide, the chromosomes need to be compacted much further to maximize the chances of correct pairing and segregation of chromosomes into daughter cells. Packaging of DNA into **nucleosomes** and then the 30 nm chromatin fiber results in a linear condensation of about 50-fold. During the M (mitosis) phase, higher-order coiling occurs (see Figure 1.7), so that DNA in a human metaphase chromosome is compacted to about 1/10 000 of its stretched-out length. As a result, the short, stubby metaphase chromosomes are readily visible under light microscopes.

Chromosome function: replication origins, centromeres, and telomeres

The DNA within a chromosome contains genes that are expressed according to the needs of a cell. But it also contains specialized sequences that are needed for chromosome function. Three major classes are described below.

Centromeres

When a cell divides the chromosomes must be correctly segregated to the two daughter cells. This requires a **centromere**, a region to which a pair of large protein complexes called kinetochores will bind just before the preparation for cell division (**Figure 1.8**). Centromeres can be seen at metaphase as the primary constriction that separates the short and long chromosome arms. Microtubules attached to each kinetochore are responsible for positioning the chromosomes correctly at metaphase and then pulling the separated chromosomes to opposite poles of the mitotic spindle.

The DNA sequences at centromeres are very different in different organisms. In a mammalian chromosome, the centromeric DNA is a heterochromatic region dominated by highly repetitive DNA sequences that often extend over megabases of DNA.

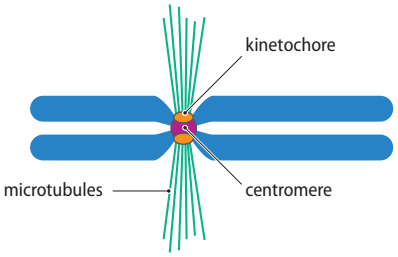

Figure 1.8 Centromere function relies on the assembly of kinetochores and attached microtubules.

Replication origins

For a chromosome to be replicated, it needs one or more replication origins—DNA sequence components to which protein factors bind in preparation for initiating DNA replication. The chromosomes of budding yeast can be replicated using a single very short highly defined DNA sequence, but in the cells of complex organisms, such as mammals, DNA is replicated at multiple initiation sites along each chromosome; the replication origins are quite long and do not have a common base sequence.

Telomeres

Telomeres are specialized structures at the ends of chromosomes that are necessary for the maintenance of chromosome integrity (if a telomere is lost after 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).

Unlike centromeric DNA, telomeric DNA has been well conserved during evolution. In vertebrates, the DNA of telomeres consists of many tandem (sequential) copies of the sequence TTAGGG to which certain telomeric proteins bind. Most of the telomere DNA is double-stranded with one strand containing TTAGGG repeats (the G-rich strand) and the complementary strand containing CCCTAA repeats (the C-rich strand). However, at its 3′ end, the G-rich strand has an overhang (with about 30 TTAGGG repeats) that folds back and base pairs with the C-rich strand. The resulting T-loop is thought to protect the telomere DNA from natural cellular exonucleases that repair double-strand DNA breaks (**Figure 1.9**).

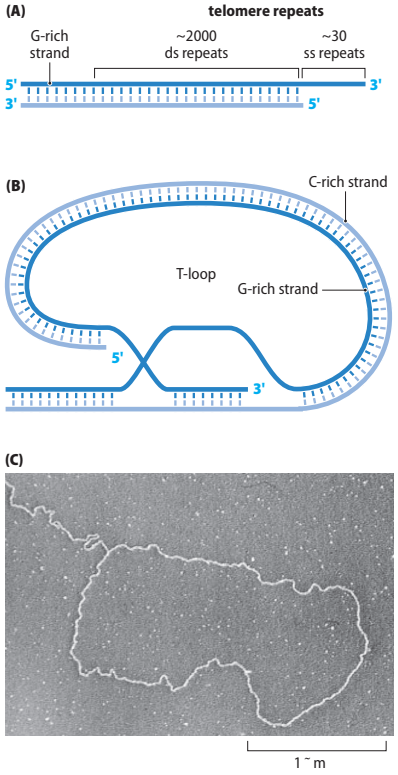

Figure 1.9 Telomere structure and T-loop formation. (A) Human telomere structure. A tandem array of roughly 2000 copies of the double-stranded hexanucleotide (TTAGGG/CCCTAA) repeat followed by a protrusion of about 30 single-stranded TTAGGG repeats. Abbreviations: ss, single-strand; ds, double-strand. (B) T-loop formation. The single-stranded terminus can loop back and invade the double-stranded region by base pairing with the complementary C-rich strand. (C) Electron micrograph showing formation of a roughly 15-kilobase T-loop at the end of an interphase human chromosome. (From Grif th JD et al. [1999] *Cell* 97:503–514; PMID 10338214. With permission from Elsevier.)

1.3 DNA AND CHROMOSOMES IN CELL DIVISION AND THE CELL CYCLE

Differences in DNA copy number between cells

Like other multicellular organisms, we have cells that are structurally and functionally diverse. In each individual the different cell types have the same genetic information, but only a subset of genes is expressed in each cell. What determines the identity of a cell—whether a cell is a B lymphocyte or a hepatocyte, for example—is the pattern of expression of the different genes across the genome.

As well as differences in gene expression, different cells can vary in the number of copies of each DNA molecule. The term *ploidy* describes the number of copies (n) of the basic chromosome set (the collective term for the different chromosomes in a cell) and also describes the copy number of each of the different nuclear DNA molecules.

The DNA content of a single chromosome set is represented as C. Human cells—and the cells of other mammals—are mostly **diploid** (2C), with nuclei containing two copies of each type of chromosome, one paternally inherited

and one maternally inherited. Sperm and egg cells are **haploid** cells that contain only one of each kind of chromosome (1C). Human sperm and eggs each have 23 different types of chromosomes and so $n = 23$ in humans.

Some specialized human cells are nulliploid (0C) because they lack a nucleus—examples include erythrocytes, platelets, and terminally differentiated keratinocytes. Others are naturally polyploid (more than 2C). Polyploidy can occur by two mechanisms. The DNA might undergo multiple rounds of replication without cell division, as when the large megakaryocytes in blood are formed (they have from 16 to 64 copies of each chromosome, and the nucleus is large and multilobed). Alternatively, polyploid cells originate by cell fusion to give cells with multiple nuclei, as in the case of muscle fiber cells.

Mitochondrial DNA copy number

The great majority of our cells are diploid and contain two copies of each nuclear DNA molecule. In stark contrast, the number of copies of the mitochondrial DNA (mtDNA) can vary from hundreds to many thousands according to the cell type, and can even vary over time in some cells. The two types of haploid cells show very large differences in mtDNA copy number: a human sperm typically has about 100 mtDNA copies, but a human egg cell usually has about 250 000 mtDNA molecules.

The cell cycle and segregation of replicated chromosomes and DNA molecules

Cells also differ according to whether they actively participate in the cell cycle and undergo successive rounds of cell division. Each time a cell divides, it gives rise to two daughter cells. To keep the number of chromosomes constant there needs to be a tight regulation of chromosome replication and chromosome segregation. Each chromosome needs to be replicated just once to give rise to two daughter chromosomes, which must then segregate equally so that one passes to each daughter cell.

During normal periods of growth there is a need to expand cell number. In the fully grown adult, the majority of cells are terminally differentiated and do not divide, but stem cells and progenitor cells continue to divide to replace cells that have a high turnover, notably blood, skin, sperm, and intestinal epithelial cells.

Each round of the cell cycle involves a phase in which the DNA replicates—S phase (synthesis of DNA)—and a phase where the cell divides—M phase. Note that M phase involves both nuclear division (mitosis) and cell division (cytokinesis). In the intervals between these two phases are two gap phases—G_1 phase (gap between M phase and S phase) and G_2 phase (gap between S phase and M phase)—see **Figure 1.10**.

Cell division takes up only a brief part of the cell cycle. For actively dividing human cells, a single turn of the cell cycle might take about 24 hours; M phase often occupies about 1 hour. During the short M phase, the chromosomes become extremely highly condensed in preparation for nuclear and cell division. After M phase, cells enter a long growth period called **interphase** ($= G_1 + S + G_2$ phases), during which chromosomes are enormously extended, allowing genes to be expressed.

G_1 is the long-term end state of terminally differentiated nondividing cells. For dividing cells, the cells will enter S phase only if they are committed to mitosis; if not, they are induced to leave the cell cycle to enter a resting phase, the G_0 phase (a modified G_1 stage). When conditions become suitable later on, cells may subsequently move from G_0 to re-enter the cell cycle.

Changes in cell chromosome number and DNA content

During the cell cycle, the amount of DNA in a cell and the number of chromosomes change. In the box panels in Figure 1.10 we follow the fate of a single chromosome through M phase and then through S phase. If we were to

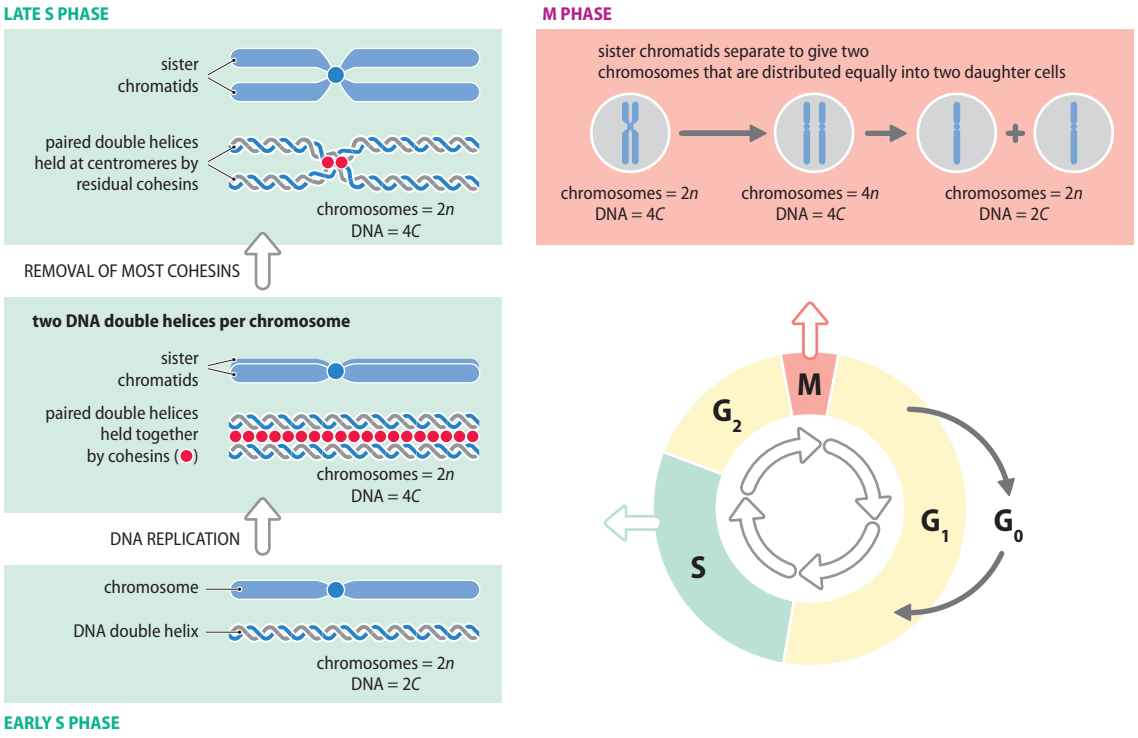

Figure 1.10 Changes in chromosomes and DNA content during the cell cycle. The cell cycle consists of four major phases as shown at the bottom right (in the additional G_0 phase a cell exits from the cell cycle and remains suspended in a stationary phase that resembles G_1 but can subsequently rejoin the cell cycle under certain conditions). In the expanded panels for M and S phases we show for convenience just a single chromosome, and we illustrate in each of the three boxes at left, representing S phase at dif erent stages, how a single chromosome (top) relates to its DNA molecule (bottom). Chromosomes contain one DNA double helix from the end of M phase right through until just before the DNA duplicates in S phase. After duplication, the two double helices are held tightly together along their lengths by binding proteins called cohesins (red circles), and the chromosome now consists of two sister chromatids each having a DNA double helix. The sister chromatids becomes more obvious in late S phase when most of the cohesins are removed except for some at the centromere, which continue to hold the two sister chromatids together. The sister chromatids f nally separate in M phase to form two independent chromosomes that are then segregated into the daughter cells. Note that the S-phase chromosomes in the boxes at left are shown, purely for convenience, in a compact form; in reality they are enormously extended.

consider a diploid human cell this would be one chromosome out of the 46 ($2n$) chromosomes present after daughter cells are first formed. We also show in this figure how a single chromosome (top) relates to its DNA double helix content at different stages in S phase. The progressive changes in the number of chromosomes and the DNA content of cells at different stages of the cell cycle are listed below.

- From the end of the M phase right through until DNA duplication in S phase, each chromosome of a diploid ($2n$) cell contains a single DNA double helix; the total DNA content is therefore 2C.
- After DNA duplication, the total DNA content per cell is 4C, but specialized binding proteins called cohesins hold the duplicated double helices together as **sister chromatids** *within* a single chromosome. The chromosome number remains the same ($2n$), but each chromosome now has double the DNA content of a chromosome in early S phase. In late S phase, most of the cohesins are removed but cohesins at the centromere are retained to keep the sister chromatids together.
- During M phase, the residual cohesins are removed and the duplicated double helices finally separate. That allows sister chromatids to separate to form two daughter chromosomes, giving $4n$ chromosomes. The duplicated chromosomes segregate equally to the two daughter cells so that each will have $2n$ chromosomes and a DNA content of 2C.

Figure 1.10 can give the misleading impression that all the interesting action happens in S and M phases. That is quite wrong—a cell spends most of its life in the G_0 or G_1 phases, and that is where the genome does most of its work, issuing the required instructions to make the diverse protein and RNA products needed for cells to function.

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 1.11**).

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 1.11), the segregation of the mitochondria into daughter cells is thought to be stochastic.

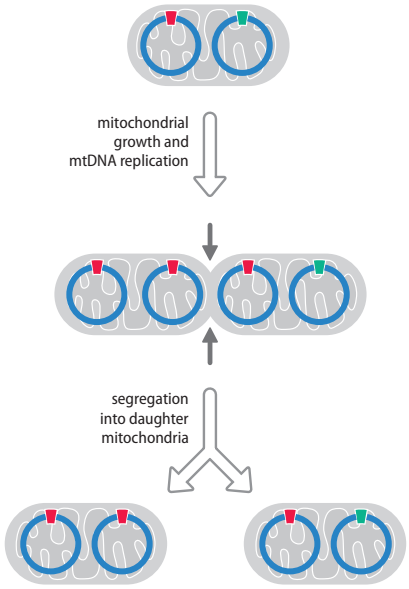

Figure 1.11 Unequal replication of individual mitochondrial DNAs. Unlike in the nucleus, where replication of a chromosomal DNA molecule is tightly controlled and normally produces two copies, mitochondrial DNA (mtDNA) replication is stochastic. When a mitochondrion increases in mass in preparation for 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 5.

Mitosis: the usual form of cell division

Most cells divide by a process known as mitosis. In the human life cycle, mitosis is used to generate extra cells that are needed for periods of growth and to replace various types of short-lived cells. 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 both nuclear division (mitosis, which is divided into the stages of prophase, prometaphase, metaphase, anaphase, and telophase), and also cell division (cytokinesis), which overlaps the final stages of mitosis (**Figure 1.12**). 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; however, after DNA replication, two identical DNA double helices are produced and held together by cohesins. 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, as early as prometaphase (when the chromosomes are now visible under the light microscope), individual chromosomes can be seen to comprise two **sister chromatids** that remain attached at the centromere (bound by some residual cohesins).

Later, at the start of anaphase, the remaining cohesins are removed and 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 1.12).

Meiosis: a specialized reductive cell division giving rise to sperm and egg cells

The **germ line** is the collective term for cells that can pass genetic material to the next generation. It includes haploid sperm and egg cells (the **gametes**) and all the diploid precursor cells from which they arise by cell division, going all the way back to the zygote. The nongermline cells are known as **somatic cells**.

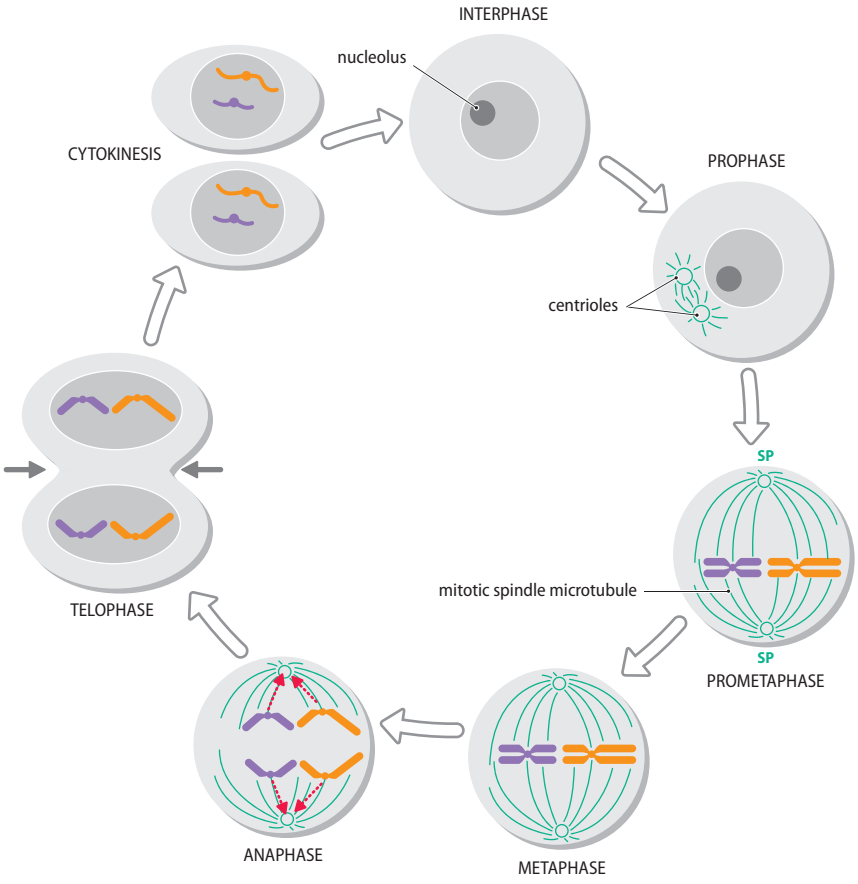

Figure 1.12 Mitosis (nuclear division) and cytokinesis (cell division). Early in prophase, centrioles (short cylindrical structures made up of microtubules and associated proteins) begin to separate and migrate to opposite poles of the cell. They give rise to the spindle poles (SP) from which microtubules will extend to the center of the cell to form the mitotic spindle. In prometaphase, the nuclear envelope breaks down, and the now highly condensed chromosomes become attached at their centromeres to the array of mitotic spindle microtubules. At metaphase, the chromosomes all lie along the middle of the mitotic spindle, still with the sister chromatids bound together (because of residual cohesins at the centromere that hold the duplicated DNA helices together). Removal of the residual cohesins allows the onset of anaphase: the sister chromatids separate and begin to migrate toward opposite poles of the cell (shown by dashed red arrows). The nuclear envelope forms again around the daughter nuclei during telophase, and the chromosomes decondense, completing mitosis. Before the f nal stages of mitosis, and most obviously at telophase, cytokinesis begins. The cell becomes progressively constricted at its middle (shown at telophase by converging horizontal arrows), eventually resulting in full cytokinesis to produce two daughter 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 descendant cells have the chromosome constitution 46,XX (female) or 46,XY (male).

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. The diploid spermatocytes and oocytes can then undergo **meiosis**, the cell division process that produces haploid gametes.

Meiosis is a *reductive* division because it involves two successive cell divisions (meiosis I and meiosis II) but only one round of DNA replication (**Figures 1.13** and **1.14**). As a result, it gives rise to four haploid cells. In males, the two meiotic cell divisions are each symmetric, producing four functionally equivalent spermatozoa. Huge numbers of sperm are produced, and spermatogenesis is a continuous process from puberty onward.

Female meiosis is different: cell division is asymmetric, resulting in unequal division of the cytoplasm. The products of female meiosis I (the first meiotic cell division) are a large secondary oocyte and a small cell, the *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).

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 those primary oocytes that are not used in ovulation until late in life.

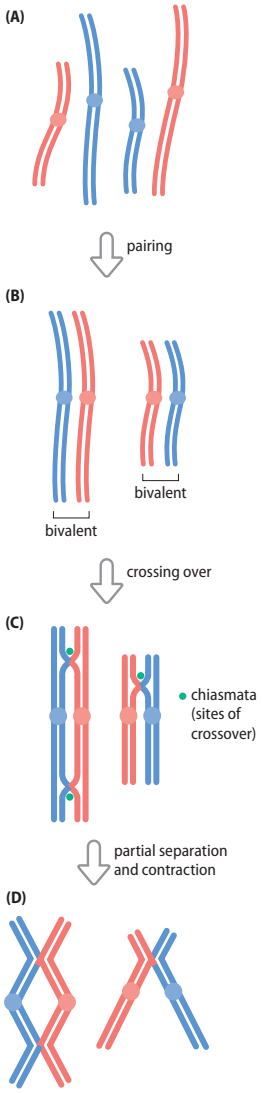

Figure 1.13 Prophase stages in meiosis I. (A) In leptotene, the duplicated chromosomes (each with a pair of sister chromatids) begin to condense but remain unpaired. (B) In zygotene, pairing of maternal and paternal homologous chromosomes (**homologs**) occurs, to form bivalents with four chromatids. (C) In pachytene, recombination (crossing over) occurs through the physical breakage and subsequent rejoining of maternal and paternal chromosome fragments. There are two chiasmata (crossovers) in the bivalent on the left, and one in the bivalent on the right. For simplicity, both chiasmata on the left involve the same two chromatids. In reality, more chiasmata 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. A further stage, diakinesis, is marked by contraction of the bivalents and is the transition to metaphase I. In this f gure, only 2 of 23 possible pairs of homologs are illustrated (with the maternal homolog colored pink, and the paternal homolog blue).

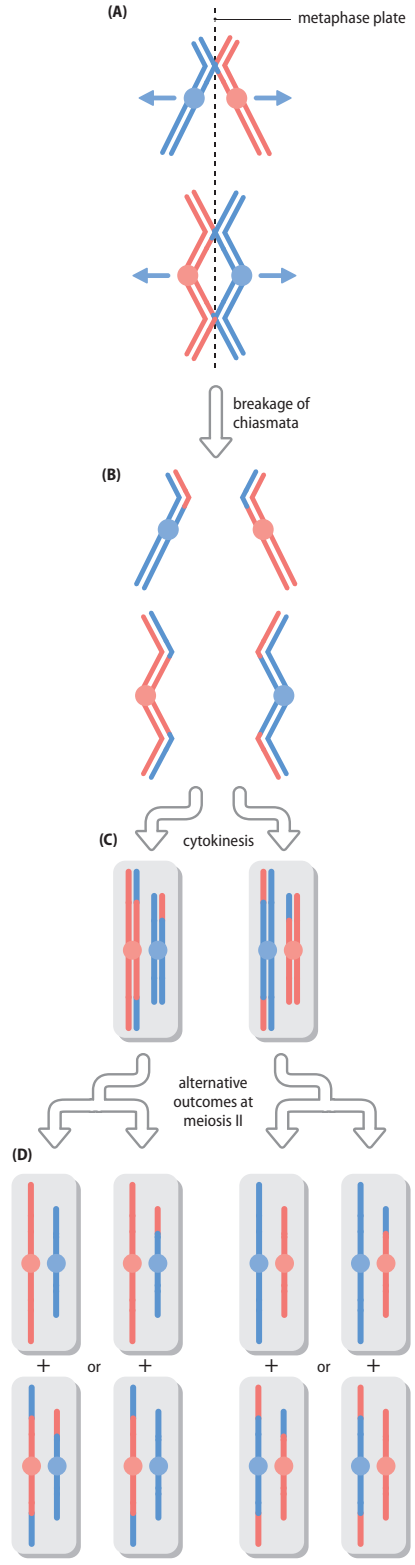

Figure 1.14 Metaphase I to production of gametes. (A) At metaphase I, the bivalents align on the metaphase plate, at the center of the spindle apparatus. Contraction of spindle f bers draws the chromosomes in the direction of the spindle poles (arrows). (B) The transition to anaphase I occurs at the consequent rupture of the chiasmata. (C) Cytokinesis segregates the two chromosome sets, each to a dif erent primary spermatocyte. Note that, after recombination during prophase I (see Figure 1.13C), 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 dif erent human chromosomes are depicted, for clarity, so only 2^2 (that is, 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, the processes by which genetic diversity arises are the same as in spermatogenesis.

Pairing of paternal and maternal homologs (synapsis)

Each of our diploid cells contains two copies (**homologs**) of each type of chromosome, one maternal copy and one paternal copy. So, for example, paternal chromosome 1 and maternal chromosome 1 are homologs. The exception, of course, are the X and Y chromosomes in males.

A special feature of meiosis I—that distinguishes it from mitosis and meiosis II—is the pairing *(synapsis)* of paternal and maternal homologs. Then, the maternal and paternal homologs, each with sister chromatids following DNA replication, align along their lengths and become bound together. The resulting *bivalent* has four strands: two paternally inherited sister chromatids and two maternally inherited sister chromatids (see Figures 1.13B–D and 1.14).

The pairing of homologs is required for recombination to occur (as described in the next subsection). It must ultimately be dictated by high levels of DNA sequence identity between the homologs. The high sequence matching between homologs required for pairing does not need to be complete, however: when there is some kind of chromosome abnormality so that the homologs do not completely match, the matching segments usually manage to pair up.

Pairing of maternal and paternal sex chromosomes is straightforward in female meiosis, but in male meiosis there is the challenge of pairing a maternally inherited X chromosome with a paternally inherited Y. The human X chromosome is very much larger than the Y, and their DNA sequences are very different. However, they do have some sequences in common, notably a major *pseudoautosomal region* located close to the short-arm telomeres. The X and Y chromosomes cannot pair up along their lengths, but because they have some sequences in common, they can always pair up along these regions. We will explore this in greater detail in Chapter 5 when we consider pseudoautosomal inheritance.

Recombination

The prophase of meiosis I begins during fetal life and, in human females, can last for decades. During this extended process, parternal and maternal chromatids within each bivalent normally exchange segments of DNA at randomly positioned but matching locations. This process—called **recombination** (or crossover)—involves physical breakage of the DNA in one paternal and one maternal chromatid, and the subsequent joining of maternal and paternal fragments.

Recombined homologs seem to be physically connected at specific points. Each such connection marks the point of crossover and is known as a chiasma (plural chiasmata—see Figure 1.13C). The distribution of chiasmata across chromosomes is nonrandom. The number of chiasmata per meiosis shows significant sex differences, and there are very significant differences between individuals of the same sex (and even between individual meioses from a single individual). In a large recent study of human meiosis, an average of 38 recombinations were detected per female meiosis, while 24 meioses occurred on average in male meiosis but with very significant variation (shown in Figure 8.3 on page 244). In addition to their role in recombination, chiasmata are thought to be essential for correct chromosome segregation during meiosis I.

There are hotspot regions where recombination is more likely to occur. For example, recombination is more common in subtelomeric regions. In the case of X–Y crossover there is an obligate crossover within a short 2.6 Mb pseudoautosomal region located at the tips of the short arms of the X and Y. This region is so called because it is regularly swapped between the X and Y chromosomes and so the inheritance pattern for any DNA variant here is not X-linked or Y-linked but instead resembles autosomal inheritance.

Why each of our gametes is unique

The sole purpose of sex in biology is to produce novel combinations of gene variants, and the instrument for achieving that aim is meiosis. The whole point of

meiosis is to produce *genetically unique* gametes by selecting different combinations of DNA sequences on maternal and paternal homologs.

Although a single ejaculate may contain hundreds of millions of sperm, meiosis ensures that no two sperm will be genetically identical. Equally, no two eggs are genetically identical. Each zygote must also be unique because at fertilization a unique sperm combines with a unique egg. However, a unique fertilization event can occasionally give rise to two genetically identical (**monozygotic**) twins if the embryo divides into two at a very early stage in development (monozygotic twins are nevertheless unique individuals—genetics is not everything in life!).

The second division of meiosis is identical in form to mitosis; meiosis I is where the genetic diversity originates, and that involves two mechanisms. First, there is independent assortment of paternal and maternal homologs. After 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 will enter 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 1.15**).

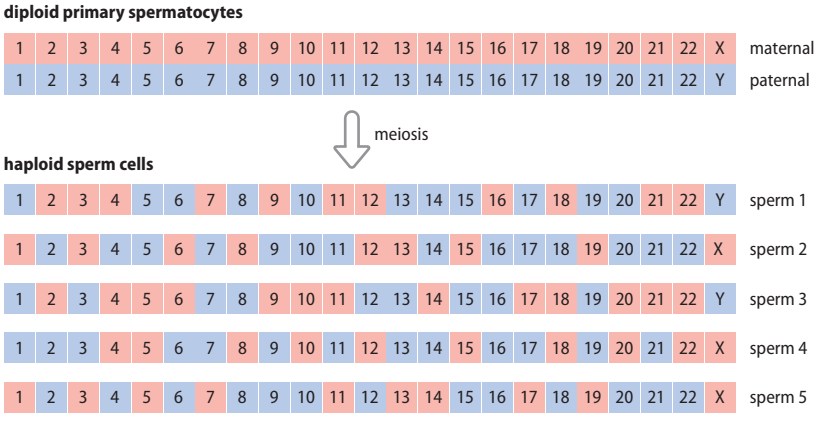

Figure 1.15 Independent assortment of maternal and paternal homologs during meiosis. The f gure shows a random selection of just 5 of the 8 388 608 (2^{23}) theoretically possible combinations of homologs that might occur in haploid human spermatozoa after meiosis in a diploid primary spermatocyte. Maternally derived homologs are represented by pink boxes, and paternally derived homologs by blue boxes. For simplicity, the diagram ignores recombination—but see Figure 1.16.

The second mechanism that contributes to genetic diversity is recombination. Whereas sister chromatids within a bivalent are genetically identical, the paternal and maternal chromatids are not. On average their DNA will differ at roughly 1 in every 1000 nucleotides. Swapping maternal and paternal sequences by recombination will therefore produce an extra level of genetic diversity (**Figure 1.16**). It raises the number of permutations from the 8.4 million that are possible just from the independent assortment of maternal and paternal homologs alone, to a virtually infinite number.

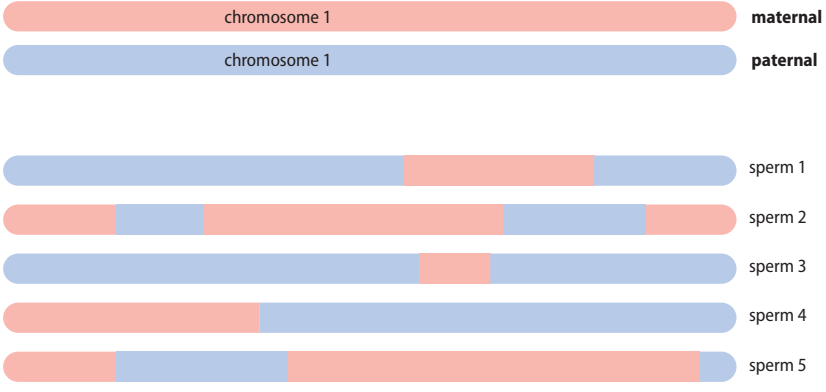

Figure 1.16 Recombination superimposes additional genetic variation at meiosis I. Figure 1.15 illustrates the contribution to genetic variation at meiosis I made by independent assortment of homologs, but for simplicity it ignores the contribution made by recombination. In reality each transmitted chromosome is a mosaic of paternal and maternal DNA sequences, as shown here. See Figure 8.1 on page 243 for a real-life example.

SUMMARY

- Nucleic acids are negatively charged long polymers composed of sequential nucleotides that each consist of a sugar, a nitrogenous base, and a phosphate group. They have a sugar-phosphate backbone with bases projecting from the sugars.
- Nucleic acids have four types of bases: adenine (A), cytosine (C), guanine (G), and either thymine (T) in DNA or uracil (U) in RNA. The sequence of bases determines the identity of a nucleic acid and its function.
- RNA normally consists of a single nucleic acid chain, but in cells a DNA molecule has two chains (strands) that form a stable duplex (in the form of a double helix). Duplex formation requires hydrogen bonding between matched bases (base pairs) on the two strands.
- In DNA two types of base pairing exist: A pairs with T, and C pairs with G. According to these rules, the two strands of a DNA double helix are said to have complementary base sequences.
- Base pairing also occurs in RNA and includes G–U base pairs, as well as G–C and A–U base pairs. Two dif erent RNA molecules with partly complementary sequences can associate by forming hydrogen bonds. Intramolecular hydrogen bonding also allows a single RNA chain to form a complex three-dimensional structure.
- DNA carries primary instructions that determine how cells work and how an individual is formed. Def ned segments of DNA called genes are used to make a single-stranded RNA copy that is complementary in sequence to one of the DNA strands (transcription).
- DNA is propagated from one cell to daughter cells by replicating itself. The two strands of the double helix are unwound, and each strand is used to make a new complementary DNA copy. The two new nuclear DNA double helices (each with one parental DNA strand and one new DNA strand) are segregated so that each daughter cell receives one DNA double helix.
- RNA molecules function in cells either as a mature noncoding RNA, or as a messenger RNA with a coding sequence used to make the polypeptide chain of a protein (translation).
- Each nuclear DNA molecule is complexed with different proteins and some noncoding RNAs to form a chromosome that condenses the DNA and protects it.
- Packaging DNA into chromosomes stops the long DNA chains from getting entangled within cells, and by greatly condensing the DNA in preparation for cell division it allows the DNA to be segregated correctly to daughter cells and to of spring.
- Our sperm and egg cells are haploid cells with a set of 23 dif erent chromosomes (each with a single distinctive DNA molecule). There is one sex chromosome (an X chromosome in eggs; either an X or Y in sperm) and 22 dif erent autosomes (nonsex chromosomes).
- Most of our cells are diploid with two copies of the haploid chromosome set, one set inherited from the mother and one from the father. Maternal and paternal copies of the same chromosome are known as homologs.
- There is one type of mitochondrial DNA (mtDNA); it is present in many copies with wide variation in copy number between dif erent cell types. Both the replication of mtDNA and its segregation to daughter cells occur stochastically.
- Cells need to divide as we grow. In fully formed adults, most of our cells are specialized, nondividing cells, but some cells are required to keep on dividing to replace short-lived cells, such as blood, skin, and intestinal epithelial cells.
- Mitosis is the normal form of cell division. Each chromosome (and chromosomal DNA) replicates once and the duplicated chromosomes are segregated equally into the two daughter cells.
- Meiosis is a specialized form of cell division required to produce haploid sperm and egg cells. The chromosomes in a diploid spermatogonium or oogonium replicate once, but there are two successive cell divisions to reduce the number of chromosomes in each cell.
- Each sperm cell produced by a man is unique, as is each egg cell that a woman produces. During the f rst cell division in meiosis, maternal and paternal homologs associate and exchange sequences by recombination. Largely random recombination results in unpredictable new DNA sequence combinations in each sperm and in each egg.

QUESTIONS

Questions can be downloaded by visiting the following link, under Support Materials: www.routledge.com/9780367490812.

FURTHER READING

More detailed treatment of the subject matter in this chapter can be found in more comprehensive genetics and cell biology textbooks such as:

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K & Walter P (2015) *Molecular Biology of the Cell*, 6th ed. Garland Science.

Strachan T & Read AP (2019) *Human Molecular Genetics*, 5th ed. CRC Press, Taylor & Francis.

Fundamentals of gene structure, gene expression, and human genome organization

2

CONTENTS

Our genome is complex, comprising about 3.2 Gb (3.2×10^9 base pairs; Gb = gigabase) of DNA. One of its main tasks is to produce a huge variety of different proteins that dictate how our cells work. Surprisingly, however, **coding DNA**—DNA sequences that specify the polypeptides of our proteins—account for just about 1.5 % of our DNA.

The remainder of our genome is noncoding DNA that does not make any protein. A significant fraction of the noncoding DNA is functionally important, including many different classes of DNA regulatory sequences that control how our genes work (such as promoters and enhancers), and DNA sequences that specify short regulatory sequence elements that work at the RNA level.

Additionally, we have many thousands of genes that do not make polypeptides; instead they make different classes of functional noncoding RNA. Some of these **RNA genes**—such as genes encoding ribosomal RNA and transfer RNA needed for protein synthesis—have been known for decades, but one of the big surprises in recent years has been the sheer number and variety of noncoding RNAs in our cells. In addition to the RNA genes, our protein-coding genes frequently make noncoding RNA transcripts as well as messenger RNAs (mRNAs).

Like other complex genomes, our genome has a large proportion of moderately to highly repetitive DNA sequences. Some of these are important in centromere and telomere function; others are important in genome evolution.

DOI: 10.1201/b22853-2

By 2003, the Human Genome Project (HGP) provided the first comprehensive insights into our genome, delivering an essentially complete nucleotide sequence of the gene-rich euchromatic component of the genome. Follow-up studies have compared our genome with other genomes, helping us to understand how our genome evolved. The comparative genomics studies, together with genome-wide functional and bioinformatic analyses, are providing major insights into how our genome works.

2.1 PROTEIN-CODING GENES: STRUCTURE AND EXPRESSION

Proteins are the main functional endpoints of gene expression and perform a huge diversity of roles that govern how cells work (acting as structural components, enzymes, carrier proteins, ion channels, signaling molecules, gene regulators, and so on). They each consist of one or more polypeptides, long sequences of amino acids that are encoded by a coding DNA. In many cases a protein also contains carbohydrate or lipid components (which are not genetically determined).

Protein-coding genes come in a startling variety of organizations, as described below, and synthesize one or more polypeptides. Polypeptide synthesis is not the endpoint, however. A newly synthesized polypeptide must undergo multiple different maturation steps, usually involving chemical modification and cleavage events, and often then associates with other polypeptides to form a working protein.

Gene organization: exons and introns

The protein-coding genes of bacteria are small (on average about 1000 bp long) and simple. The gene is transcribed to give an mRNA with a continuous coding sequence that is then translated to give a linear sequence of about 300 amino acids on average. Unexpectedly, the genes of eukaryotes turned out to be much bigger and much more complex than anticipated. And, as we will see, our protein-coding genes often contain a rather small amount of coding DNA.

For most eukaryotic protein-coding genes, the coding DNA is split into segments (**exons)** separated by noncoding DNA sequences (**introns)**. The number of exons and introns in a gene varies considerably (there seems little logic about precisely where introns insert within genes).

Excluding single-exon genes (some genes lack introns), average exon lengths show moderate variation from gene to gene, but introns can show extraordinary size differences. Our genes are therefore often large, sometimes extending over more than a megabase of DNA (**Table 2.1**).

TABLE 2.1 EXAMPLES OF DIFFERENTIAL GENE ORGANIZATION FOR HUMAN PROTEIN-CODING GENES

Human gene	Size in genome (kb)	No. of exons	Average size of exon (bp)*	Average size of intron (bp)**
SRY (male sex-determinant)	0.9	1	850	–
HBB (β-globin)	1.6	3	150	490
TP53 (p53)	39	10	236	3076
F8 (factor VIII)	186	26	375	7100
CFTR (cystic fbr osis transmembrane regulator)	250	27	227	9100
DMD (dystrophin)	2400	79	180	30 770

Items in brackets show the protein name. kb, kilobases (= 1000 bp).

* Note that the shortest human exon is just two nucleotides long, and fnal e xons can quite often be long, the record being 27 303 bp.

** The shortest human intron is 26 bp, and the longest is 1 160 411 bp—see PMID 31164174

RNA splicing: stitching together the genetic information in exons

Like all genes, genes that are split into exons are initially transcribed by an RNA polymerase to give a long RNA transcript. This primary transcript is identical in base sequence to the transcribed region of the sense DNA strand, except that U replaces T (the transcribed region of DNA is called a **transcription unit**). Thereafter, the primary RNA transcript undergoes a form of processing called **RNA splicing**.

RNA splicing involves first cleaving the RNA transcript at the junctions between transcribed exons and introns. The individual transcribed intron sequences are often degraded, but the transcribed exon sequences are then covalently linked (spliced) in turn to make a mature RNA (**Figure 2.1**). RNA splicing is performed within the nucleus by spliceosomes, complex assemblies of protein factors and small nuclear RNA (snRNA) molecules.

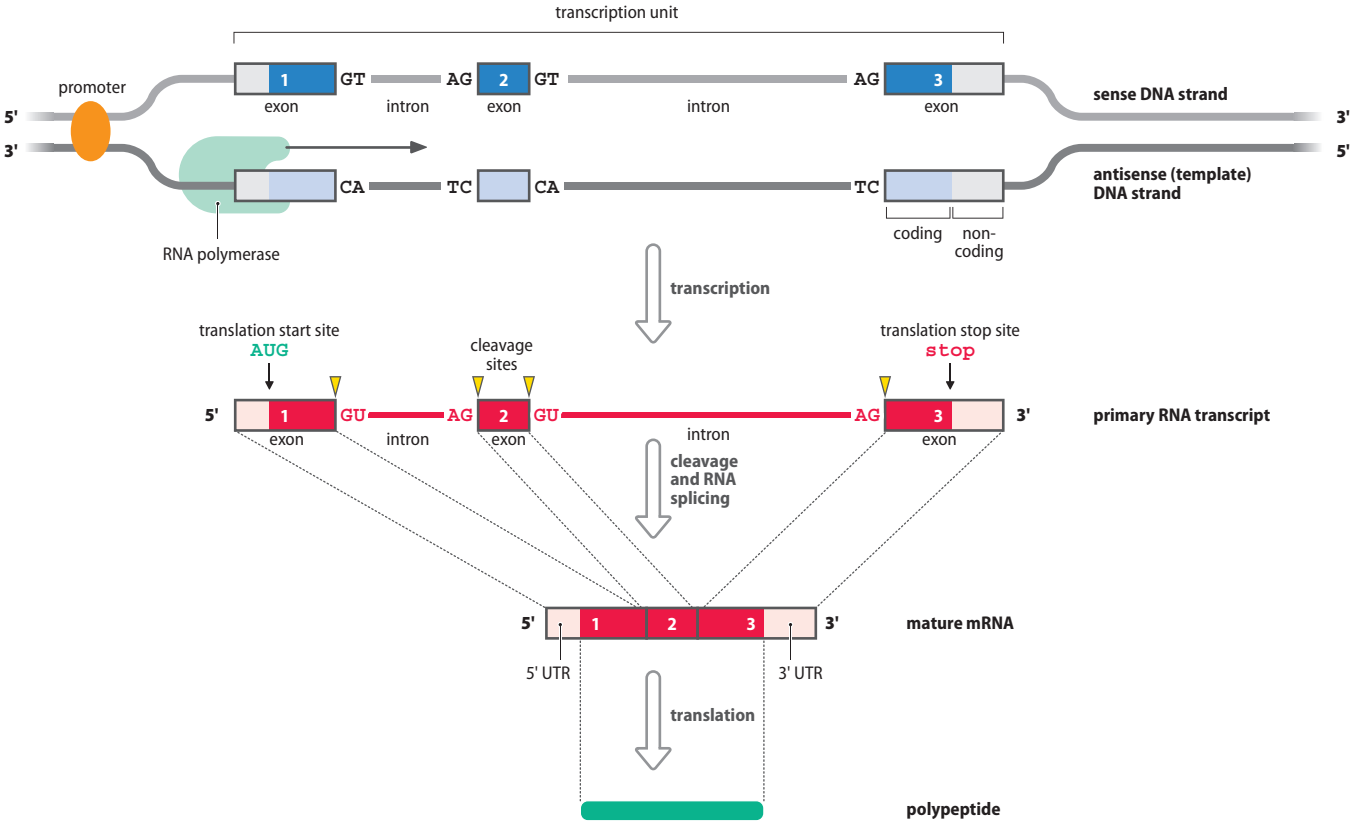

Figure 2.1 RNA splicing brings transcribed exon sequences together. Most of our protein-coding genes (and many RNA genes) undergo RNA splicing. In this generalized example a protein-coding gene is illustrated with an upstream promoter and three exons separated by two introns that each begin with the dinucleotide GT and end in the dinucleotide AG. The central exon (exon 2) is composed entirely of coding DNA (deep red), but exons 1 and 3 have both coding DNA and also noncoding DNA sequences (shown in pink; they 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 def ned 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.

We do not fully understand how the spliceosome is able to recognize and cut the primary transcript at precise positions marking the start and end of introns. However, we do know that certain sequences are important in signaling the

splice sites that define exon-intron boundaries. For example, very nearly all introns begin with a GT dinucleotide on the sense DNA strand and end with an AG so that the transcribed intron sequence begins with a GU (that marks the **splice donor site**) and ends in an AG (marking the **splice acceptor site**). The GT (GU) and AG end sequences need to be embedded in broader splice site consensus sequences that we will describe in Section 6.1 when we consider how gene expression is regulated. As we will see in Chapter 7, mutations at splice sites are important causes of disease.

Figure 2.1 might give the erroneous impression that all protein-coding genes undergo a specific, single type of RNA splicing. However, close to 10% of our protein-coding genes have a single, uninterrupted exon and do not undergo RNA splicing at all—notable examples include histone genes. And most of the genes that go through RNA splicing undergo alternative RNA splicing patterns; a single gene can therefore produce different gene products that may be functionally different. We consider the concept of alternative splicing in greater detail in Chapter 6, in the context of gene regulation.

The evolutionary value of RNA splicing

As we will see in Section 2.2, many RNA genes also undergo RNA splicing. At this stage, one might reasonably wonder why RNA splicing is so important in eukaryotic cells, and so especially prevalent in complex multicellular organisms. Why do we need to split the genetic information in genes into sometimes so many different little exons? The answer is to help stimulate the formation of novel genes and novel gene products that can permit greater functional complexity during evolution.

The huge complexity of humans and other multicellular organisms has been driven by genome evolution. In addition to periodic gene duplication, various genetic mechanisms allow individual exons to be duplicated or swapped from one gene to another on an evolutionary timescale. That allows different ways of combining exons to produce novel hybrid genes. An additional source of complexity comes from using different combinations of exons to make alternative transcripts from the same gene (alternative splicing).

Translation: decoding messenger RNA to make a polypeptide

Messenger RNA (mRNA) molecules produced by RNA splicing in the nucleus are exported to the cytoplasm. Here they are bound by ribosomes, very large complexes consisting of four types of ribosomal RNA (rRNA) and many different proteins.

Although an mRNA is formed from exons only, it has sequences at its 5′ and 3′ ends that are noncoding. Having bound to mRNA, the job of the ribosomes is to scan the mRNA sequence to find and interpret a central coding sequence that will be translated to make a polypeptide. The noncoding sequences at the ends are known as **untranslated regions** (UTRs; as shown in Figure 2.1). and contain sequences that are important in regulating gene expression.

A polypeptide is a polymer made up of a linear sequence of **amino acids** (**Figure 2.2A**). Amino acids have the general formula NH2-CH(R)-COOH, where R is a variable side chain that defines the chemical identity of the amino acid and is connected to the central (alpha) carbon of the NH-CH-CO framework sequence. There are 20 common amino acids (Figure 2.2C). Polypeptides are made by a condensation reaction between the carboxyl (COOH) group of one amino acid and the amino (NH2) group of another amino acid, forming a peptide bond (see Figure 2.2B).

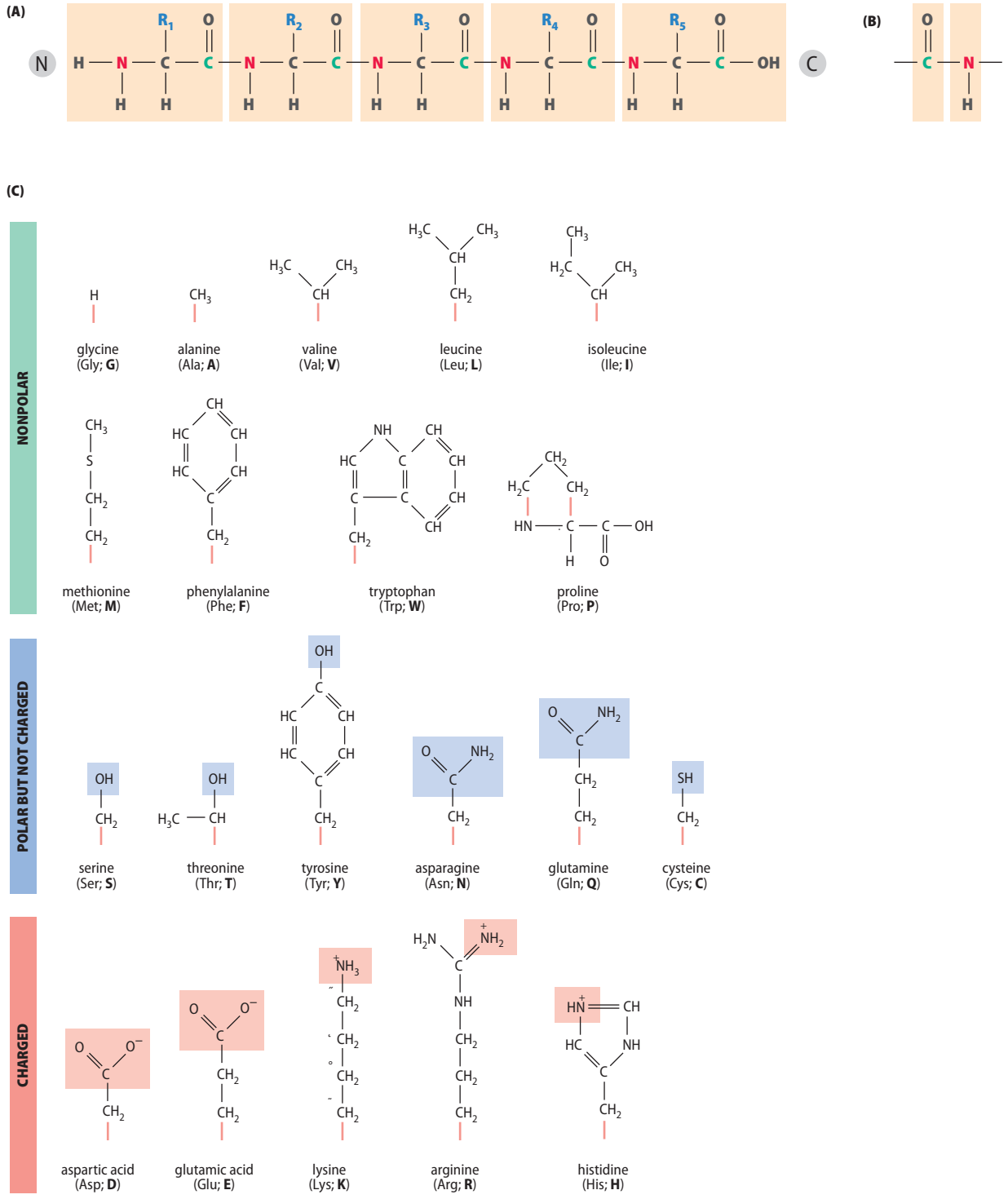

Figure 2.2 Polypeptide and amino acid structure. (A) Polypeptide primary structure. A pentapeptide is shown with its f ve amino acids highlighted. Here, the left end is called the N-terminal end because the amino acid has a free amino (NH_2) group; the right end is the C-terminal end because the last amino acid has a free carboxyl (COOH) group. The side chains (R_1 to R_5) are variable and determine the identity of the amino acid. They are joined to the central carbon atom of the repeating framework sequence: –NH–CH–CO–. Note that at physiological pH the free amino and carboxyl groups will be charged: NH_3^+ and COO^- respectively. (B) Neighboring amino acids are joined by a peptide bond. A peptide bond is formed by a condensation reaction between the end carboxyl group of one amino acid and the end amino group of another: $-COOH + NH_2- \rightarrow -CONH- + H_2O$. (C) Side chains of the 20 principal amino acids. Red lines represent the covalent bond attaching the side chain to the framework protein structure. Note that the structure of proline is unusual and its full structure is given here because its side chain connects to the nitrogen atom of the framework amino group as well as to the alpha carbon, thereby forming a fv e-membered ring.

To make a polypeptide, the coding sequence within an mRNA is translated in groups of three nucleotides at a time, called **codons**. There are 64 possible codons (four possible bases at each of three nucleotide positions makes $4 \times 4 \times 4$ permutations). Of these, 61 are used to specify an amino acid; three others signal an end to protein synthesis. The universal **genetic code**, the set of rules that dictate how codons are interpreted, therefore has some redundancy built into it. For example, the amino acid serine can be specified by any of six codons (UCA, UCC, UCG, UCU, AGU, and AGG), and, on average, an amino acid is specified by any of three codons. As a result, nucleotide substitutions within coding DNA quite often do not cause a change of amino acid. We discuss the genetic code in some detail in Section 7.2 when we consider the effects of single nucleotide substitutions.

The process of translation

Translation begins when ribosomes bind to the 5′ end of an mRNA and then move along the RNA to find a translational start site, the initiation codon—an AUG trinucleotide embedded within the broader, less well defined Kozak consensus sequence (GCC**Pu**CC**AUGG**; the most conserved bases are shown in bold, and Pu represents purine).

The initiation codon is the start of an **open reading frame** of codons that specify successive amino acids in the polypeptide chain (see Box 2.1 for the concept of translational reading frames). As described below, a family of transfer RNAs (tRNAs) is responsible for transporting the correct amino acids to be inserted in the required position of the growing polypeptide chain. Individual types of tRNA carry a specific amino acid; they can recognize and bind to a specific codon, and when they do so they unload their amino acid cargo.

BOX 2.1 TRANSLATIONAL READING FRAMES AND SPLITTING OF CODING SEQUENCES BY INTRONS

TRANSLATIONAL READING FRAMES

In the examples of different translational **reading frames** below, we use sequences of words containing three letters to represent the triplet nature of the genetic code. We designate the reading frames (RF) as 1, 2, or 3 depending on whether the reading frame starts before the first, second, or third nucleotide in the sequence.

Reading frame 1 (RF1) in **Figure 1** makes sense, but a shift to another reading frame produces nonsense. The same principle generally applies to coding sequences. So, for example, if one or two nucleotides are deleted from a coding sequence or there is an insertion of one or two nucleotides, the effect is to produce a **frame-shift** (a change of reading frame) that will result in nonsense.

SPLITTING OF CODING SEQUENCES BY INTRONS

At the DNA level, introns may interrupt a coding sequence at one of three types of position: at a point precisely between two codons (a *phase 0 intron*), after the first nucleotide of a codon (a *phase 1 intron*), or after the second nucleotide of a codon (a *phase 2 intron*).

An internal exon may be flanked by introns of the same phase; in an exon like this the number of nucleotides is always exactly divisible by three. Where an exon is flanked by two introns of a different phase, the exon will have a number of nucleotides that is not exactly divisible by three. That can have important consequences when deletions occur within genes (see **Figure 2**).

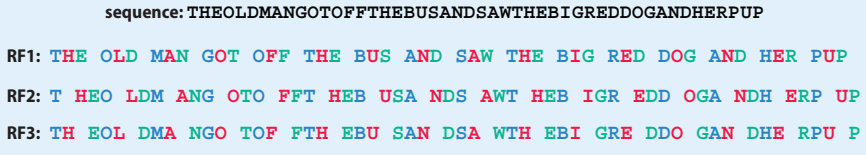

Figure 1 The importance of using the correct translational reading frame. The sequence of letters at the top can be grouped into sets of three (codons) that make sense in reading frame 1 (RF1) but make no sense when using reading frame 2 (RF2) or reading frame 3 (RF3).

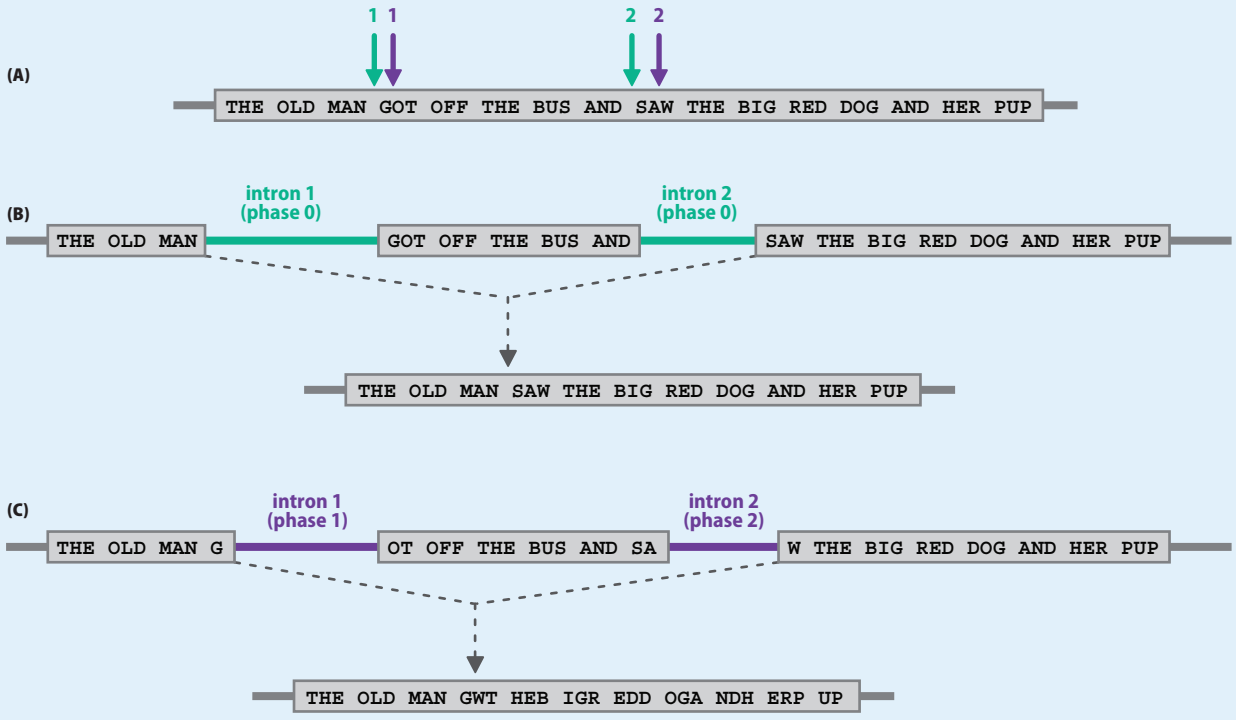

Figure 2 Efe cts on the translational reading frame caused by the deletion of coding exons. (A) Here we show a coding sequence split at the DNA level by a pair of introns, 1 and 2. Now imagine two alternative possibilities, shown by green and purple arrows. Green arrows indicate two f anking introns of the same phase: in this case both are phase 0 introns, having inserted at analogous positions *between* codons. Purple arrows indicate an alternative where the introns are of dif erent phases, respectively phase 1 for intron 1 and phase 2 for intron 2. B) The green introns result in the central exon having a number of nucleotides that is exactly divisible by three; it can be deleted without an ef ect on the downstream reading frame. If the exon does not encode a critical component of the protein, the functional consequences may not be too grave. (C) If instead introns 1 and 2 are located as shown in purple, the central exon has a number of nucleotides that cannot be divided exactly by three. If it were to be deleted, the downstream reading frame would be scrambled with a high chance of a premature termination codon, frequently resulting in lack of function.

As each new amino acid is unloaded it is bonded to the previous amino acid so that a polypeptide chain is formed (**Figure 2.3**). The first amino acid has a free NH2 (amino) group and marks the N-terminal end (N) of the polypeptide. The polypeptide chain terminates after the ribosome encounters a **stop codon** (which signifies that the ribosome should disengage from the mRNA, releasing the polypeptide; for translation on cytoplasmic ribosomes, there are three choices of stop codon: UAA, UAG, or UGA). The last amino acid that was incorporated in the polypeptide chain has a free COOH (carboxyl group) and marks the C-terminal end (C) of the polypeptide.

Transfer RNA as an adaptor RNA

Transfer RNAs have a classic cloverleaf structure resulting from intramolecular hydrogen bonding (**Figure 2.4A**). They serve as adaptor RNAs because their job is to base pair with mRNAs and help decode the coding sequence messages carried by mRNAs. The base pairing is confined to a three-nucleotide sequence in the tRNA called an *anticodon*, which is complementary in sequence to a codon. According to the identity of their anticodons, different tRNAs carry different amino acids covalently linked to their 3′ ends. Through base pairing between codon and anticodon, individual amino acids can be sequentially ordered according to the sequence of codons in an mRNA, and sequentially linked together to form a polypeptide chain (see Figure 2.4B).

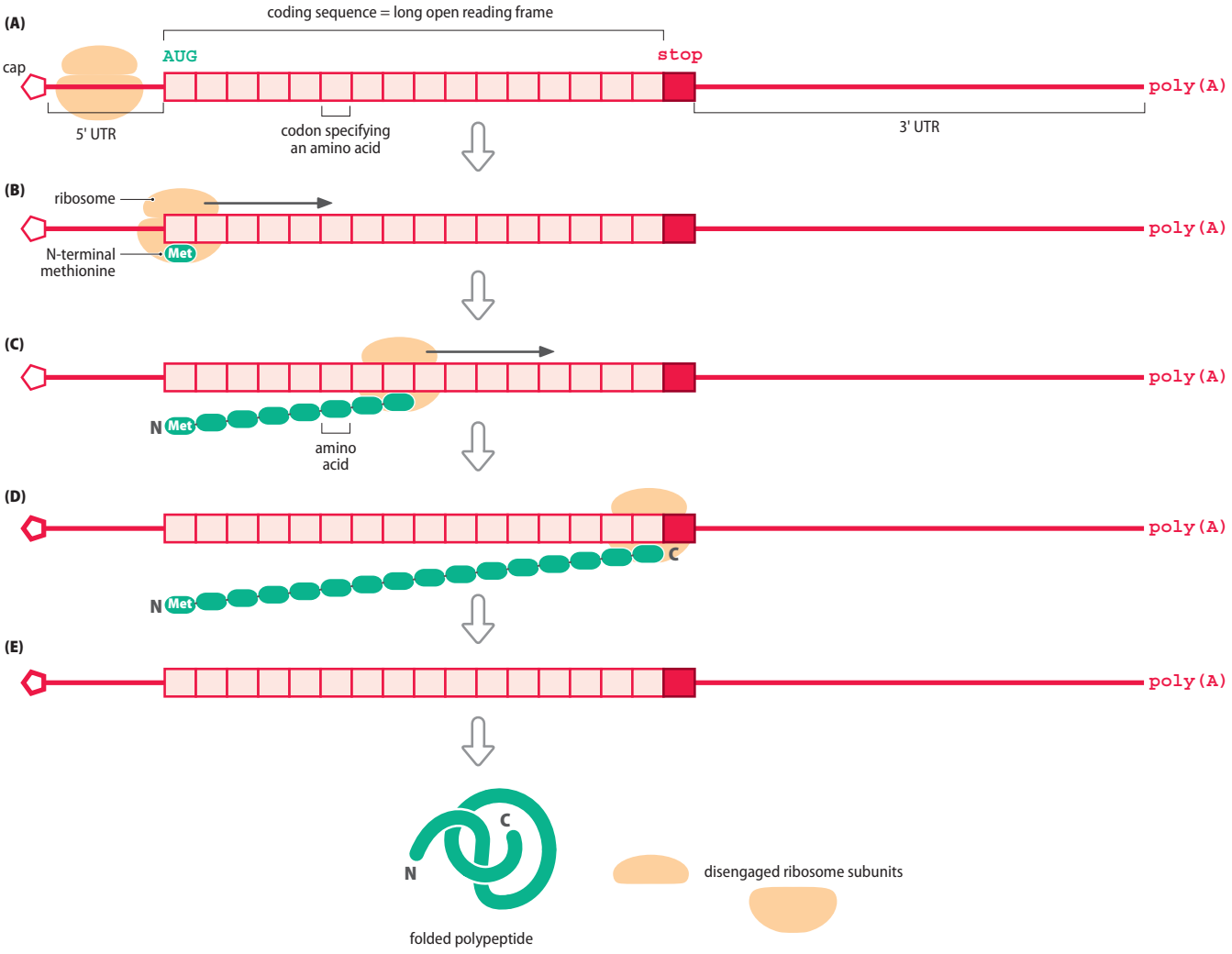

Figure 2.3 The basics of translation. (A) A ribosome attaches to the 5′ untranslated region (5′ UTR) of the mRNA and then slides along until (B) it encounters the initiation codon AUG, at which point a methionine-bearing transfer RNA (not shown) engages with the AUG codon and deposits its methionine cargo (green bar, labeled MET). (C) The ribosome continues to move along the mRNA and as it encounters each codon in turn a specif c amino-acid-bearing tRNA is recruited to recognize the codon and to deposit its amino acid, according to the *genetic code*. The ribosome catalyses the formation of a *peptide bond* (Figure 2.2B) between each new amino acid and the last amino acid, forming a polypeptide chain (shown here, for convenience, as a series of joined green ovals). (D) Finally, the ribosome encounters a stop codon, at which point (E) the ribosome falls of t he mRNA and dissociates into its two subunits, releasing the completed polypeptide. The polypeptide undergoes *post-translational modification* as described in the text, which may sometimes involve cleavage at the N-terminal end so that methionine may not be the N-terminal amino acid in the mature polypeptide.

Untranslated regions and 5′ cap and 3′ poly(A) termini

As illustrated in Figure 2.3, each mature mRNA has a large central coding DNA sequence flanked by two **untranslated regions**, a short 5′ untranslated region (5′ UTR) and a rather longer 3′ untranslated region (3′ UTR). The untranslated regions regulate mRNA stability and contain regulatory sequences that are important in determining how genes are expressed.

As well as sequences copied from the gene sequence, mRNA molecules usually also have end sequences added post-transcriptionally to the pre-mRNA. At the 5′ end a specialized cap is added: 7-methylguanosine linked to the first nucleotide by a distinctive 5′-5′ phosphodiester bond (instead of a normal 5′-3′ phosphodiester bond). The cap protects the transcripts against 5′ → 3′ exonuclease attack and facilitates transport to the cytoplasm and ribosome attachment. At the 3′ end a dedicated poly(A) polymerase sequentially adds adenylate (AMP)

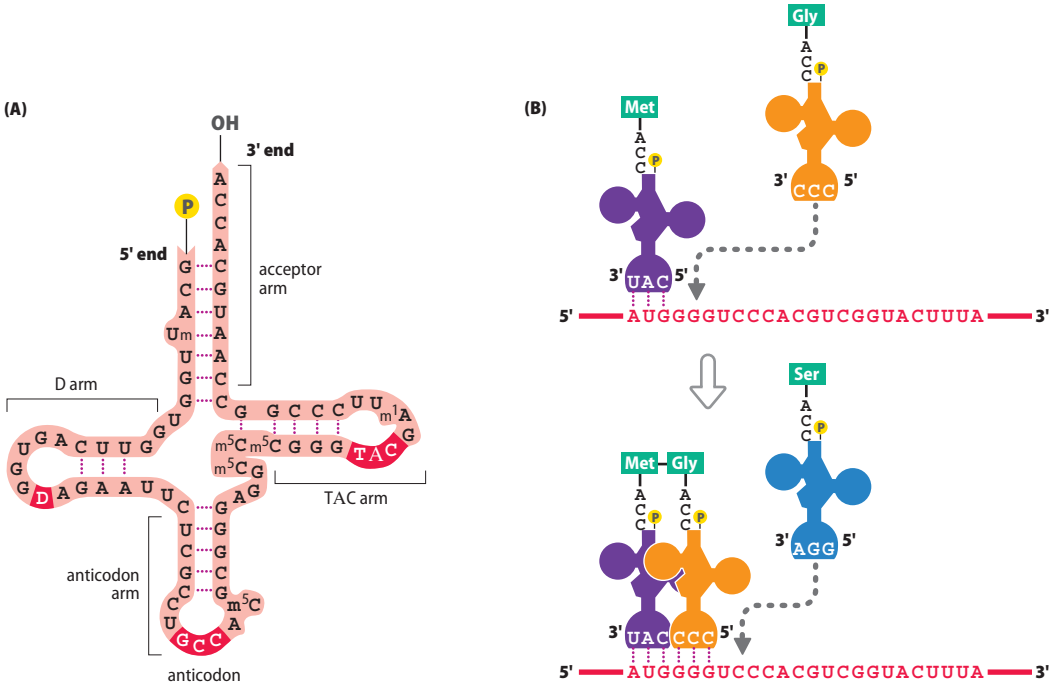

Figure 2.4 Transfer RNA structure, and its role as an adaptor RNA in translation. (A) Transfer RNA structure. The tRNA^{Gly} shown here illustrates the classical cloverleaf tRNA structure. Intramolecular base pairing produces three arms terminating in a loop plus an acceptor arm formed by pairing of 5′ and 3′ end sequences. The latter ends in a -CAA trinucleotide and covalently binds a specific amino acid. The three nucleotides at the centre of the middle loop form the *anticodon,* which identifies the tRNA according to the amino acid it will bear. Minor nucleotides are: D, 5,6-dihydrouridine; Ψ, pseudouridine (5-ribosyluracil); m5C, 5-methylcytidine; m1A, 1-methyladenosine; Um, 2′-*O*-methyluridine. (B) Role of adaptor RNA. Different tRNAs carry different amino acids, according to the type of anticodon they bear. As a ribosome traverses an mRNA it identifies the AUG initation codon. A methionine-bearing tRNA with the complementary CAU anticodon sequences then engages with the ribosome so that the CAU anticodon base pairs with the AUG codon. Note: for ease of illustration we show the tRNAs in the opposite orientation to the standard form shown in (A), with the acceptor arm on the left, not the right. Thereafter, a tRNA bearing glycine engages the second codon, GGG, by base pairing with its CCC anticodon. The ribosome's peptidyltransferase then forms a peptide bond between the N-terminal methionine and glycine. The ribosome moves along by one codon and the tRNA^{Met} is cleaved so that it can be reused and the process continues with an incoming tRNA carrying a serine and an anticodon GGA to bind to the third codon UCC, after which the incoming serine will be covalently bonded to the glycine by the ribosome's peptidyltransferase.

residues to give a poly(A) tail, about 150–200 nucleotides long. The poly(A) helps in transporting mRNA to the cytoplasm, facilitates binding to ribosomes, and is also important in stabilizing mRNAs.

From newly synthesized polypeptide to mature protein

The journey from newly synthesized polypeptide released from the ribosome to fully mature protein requires several steps. The polypeptide typically undergoes post-translational cleavage and chemical modification. Polypeptides also need to fold properly, and they often bind to other polypeptides as part of a multisubunit protein. And then there is a need to be transported to the correct intracellular or extracellular location.

Chemical modification

We describe below one type of chemical modification that involves cross-linking between two cysteine residues within the same polypeptide or on different polypeptides. Often, however, chemical modification involves the simple covalent addition of chemical groups to polypeptides or proteins. Sometimes small chemical groups are attached to the side chains of specific amino acids (**Table 2.2**). These groups can sometimes be particularly important in the structure of a protein (as in the case of collagens, which have high levels of hydroxyproline and hydroxylysine).

TABLE 2.2 COMMON TYPES OF CHEMICAL MODIFICATION OF PROTEINS BY COVALENT ADDITION OF CHEMICAL GROUPS TO A SIDE CHAIN

Type of chemical modification	Target amino acids	Comments
ADDITION OF SMALL CHEMICAL GROUP		
Hydroxylation	Pro; Lys; Asp	can play important structural roles
Carboxylation	Glu	especially in some blood clotting factors
Methylation	Lys	specialized enzymes can add or remove the methyl, acetyl, or phosphate group, causing the protein to switch between two states, with functional consequences
Acetylation	Lys	
Phosphorylation	Tyr; Ser; Thr	
ADDITION OF COMPLEX CARBOHYDRATE OF LIPID GROUP		
N-glycosylation	Asn	added to the amino group of Asn in endoplasmic reticulum and Golgi apparatus
O-glycosylation	Ser; Thr; Hydroxylysine	added to the side-chain hydroxyl group; takes place in Golgi apparatus
N-lipidation	Gly	added to the amino group of an N-terminal glycine; promotes protein-membrane interactions
S-lipidation	Cys	a palmitoyl or prenyl group is added to the thiol of the cysteine. Often helps anchor proteins in a membrane

In other cases, dedicated enzymes add or remove small chemical groups to act as switches that convert a protein from one functional state to another. Thus, specific kinases can add a phosphate group that can be subsequently removed by a dedicated phosphatase. The change between phosphorylated and dephosphorylated states can result in a major conformational change that affects how the protein functions. Similarly, methyltransferases and acetyltransferases add methyl or acetyl groups that can be removed by the respective demethylases and deacetylases. As we will see in Chapter 6, they are particularly important in modifying histone proteins to change the conformation of chromatin and thereby alter gene expression.

In yet other cases, proteins can be modified by covalently attaching complex carbohydrates or lipids to a polypeptide backbone. Thus, for example, secreted proteins and proteins destined to be part of the excretory process of cells routinely have oligosaccharides attached to the side chains of specific amino acids. Different types of lipids are also often added to membrane proteins (see Table 2.2).

Folding

The amino acid sequence, the primary structure, dictates the pattern of folding, but certain regions of polypeptides adopt types of secondary structure important in protein folding (**Box 2.2** gives an outline of protein structure). Until correct folding has been achieved, a protein is unstable; different chaperone molecules help with the folding process (careful supervision is needed because partly folded or misfolded proteins can be toxic to cells).

BOX 2.2 A BRIEF OUTLINE OF PROTEIN STRUCTURE

Four different levels of structure are recognized:

- primary structure—the linear sequence of amino acids in constituent polypeptides
- secondary structure—the path that a polypeptide backbone follows within local regions of the primary structure
- tertiary structure—the overall three-dimensional structure of a polypeptide
- quaternary structure—the aggregate structure of a multimeric protein (composed of two or more polypeptide subunits that may be of more than one type).

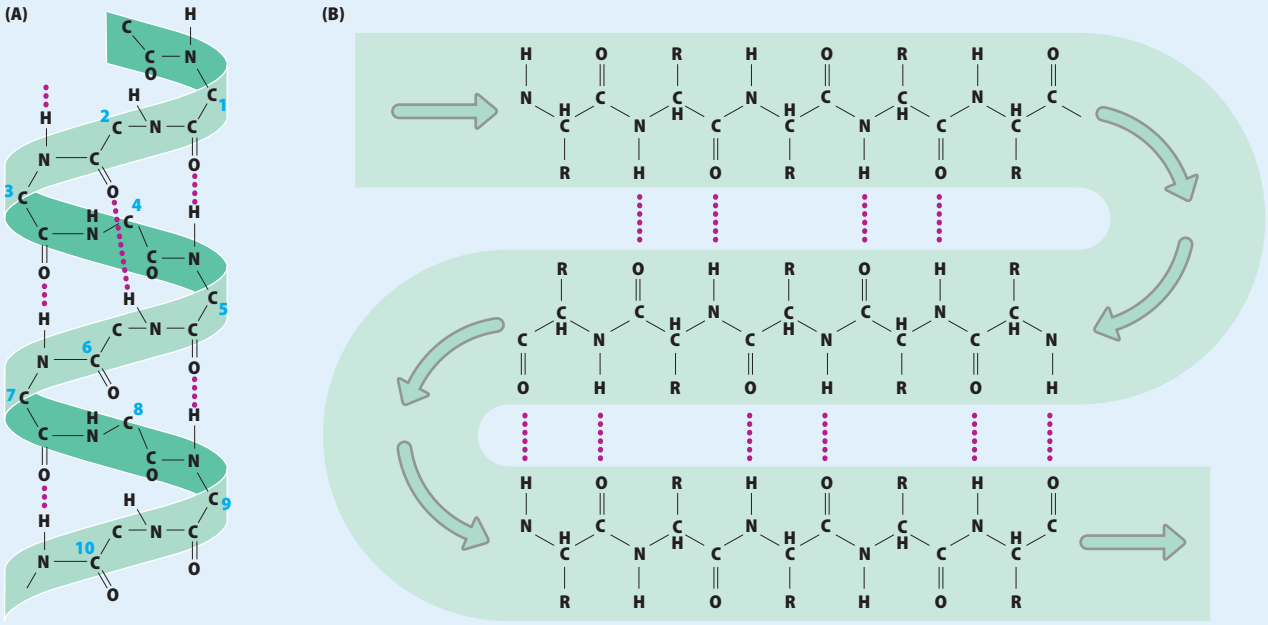

Figure 1 Elements of protein secondary structure. (A) An α-helix. The solid rod structure is stabilized by hydrogen bonding between the oxygen of the carbonyl group (C=O) of each peptide bond and the hydrogen on the peptide bond amide group (NH) of the fourth amino acid away, yielding 3.6 amino acids per turn of the helix. 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) A β-sheet (also called a β-pleated sheet). Here, 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) ....., hydrogen bond.

ELEMENTS OF SECONDARY STRUCTURE

Secondary structure is notably shaped by intramolecular hydrogen bonding. The *α-helix*, for example, is a rigid cylinder 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 1A**). α-Helices often occur in transcription factors and other proteins that perform key cellular functions.

In the *β-sheet* (also called the *β-pleated sheet*), the hydrogen bonds occur between opposed peptide bonds in parallel or antiparallel segments of the same polypeptide chain (**Figure 1B**). β-Sheets occur—often together with α-helices—at the core of most globular proteins.

The *β-turn* involves hydrogen bonding between the peptide-bond carbonyl (C=O) of one amino acid and the peptide-bond NH group of an amino acid located only three places farther along. The resulting hairpin turn allows an abrupt change in the direction of a polypeptide, enabling compact globular shapes to be achieved. β-Turns can connect neighboring segments in a β-sheet, when the polypeptide strand has to undergo a sharp turn.

When placed in an aqueous environment, proteins are stabilized by having amino acids with hydrophobic side chains located in the interior of the protein, whereas hydrophilic amino acids tend to be located toward the surface. For many proteins, notably globular proteins, the folding pattern is also stabilized by a form of covalent cross-linking that can occur between certain distantly located cysteine residues—the sulfhydryl groups of the cysteine side chains interact to form a disulfide bond (alternatively called a disulfide bridge—see **Figure 2.5**).

Cleavage and transport

The initial polypeptide normally undergoes some type of N-terminal cleavage. Sometimes just the N-terminal methionine is removed. But for proteins secreted from cells, the polypeptide precursor carries an N-terminal leader sequence (signal peptide) that is required to assist the protein to cross the plasma membrane, after which the signal peptide is cleaved at the membrane, releasing the

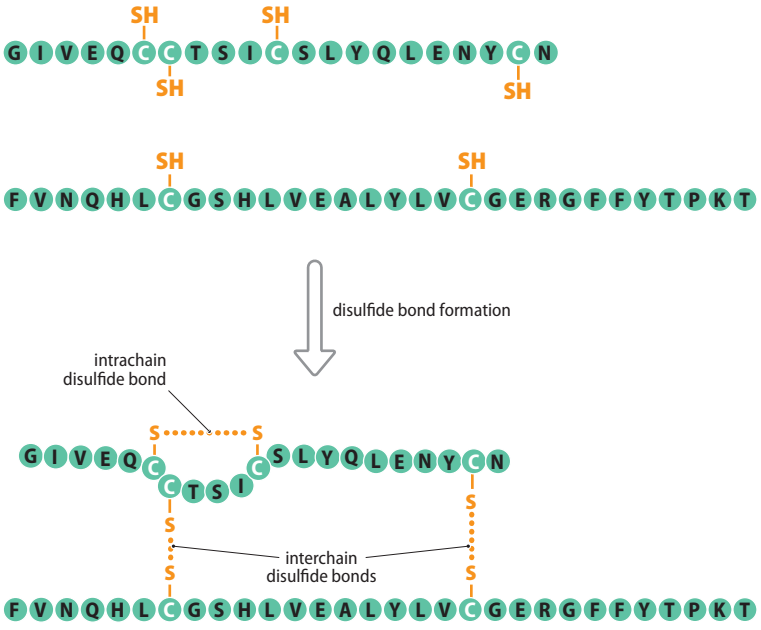

Figure 2.5 Intrachain and interchain disulf de bridges in human insulin. Human insulin is composed of two peptide chains, an A chain with 21 amino acids, and a B chain with 30 amino acids. Disulf de bridges (–S–S–) form by a condensation reaction between the sulfhydryl (–SH) groups on the side chains of cysteine residues. They form between the side chains of cysteines at positions 6 and 11 within the insulin A chain, and also between cysteine side chains on the insulin A and B chains. Note that here all the cysteines participate in disulf de bonding, which is unusual. When disulf de bonding occurs in large proteins, only certain cysteine residues are involved.

mature protein. (The signal peptide, often 10–30 amino acids in length, carries multiple hydrophobic amino acids.)

Other short internal peptide sequences can act simply as address labels for transporting proteins to the nucleus, mitochondria, plasma membrane, and so on. They are retained in the mature protein.

Binding of multiple polypeptide chains

Proteins are often made of two or more polypeptide subunits. Occasionally, constituent polypeptides are covalently linked with disulfide bridges (as in the case of joining the different chains of immunoglobulins; see Figure 4.10 on page 99). Often, however, constituent polypeptides are held together mainly by noncovalent bonds, including nonpolar interactions and hydrogen bonds. For example, hemoglobins are tetramers, composed of two copies each of two different globin chains that associate in this way. Collagens provide a good example of very intimate structural association between polypeptides, consisting of three chains (two of one type; one of another) wrapped round each other to form a triple helix.

2.2 RNA GENES AND NONCODING RNA

The majority of our genes are **RNA genes**, genes devoted to making functional **noncoding RNA** (**ncRNA**) as their end product. (The latest GENCODE data – RELEASE 40, April 2022 – revealed a total of 26 372 human RNA genes and 19 988 protein-coding genes). The vast majority of the RNA genes regulate gene expression in some way, or directly assist in the expression of protein-coding genes, and proteins remain the main functional endpoint in cells.

Like proteins (and mRNA), noncoding RNAs are made as precursors that often undergo enzymatic cleavage to become mature gene expression products. They are also subject to chemical modification: minority bases such as dihydrouridine or pseudouridine and various methylated bases are quite common—see Figure 2.4A for some examples in a tRNA.

Until quite recently, ncRNAs were largely viewed as having important but rather dull functions. For the most part, they seemed to act as ubiquitous accessory molecules that worked directly or indirectly in protein production. After ribosomal and transfer RNAs, we came to know about various other ubiquitous ncRNAs that mostly work in RNA maturation: spliceosomal small nuclear RNAs (snRNAs); small nucleolar RNAs (snoRNAs) that chemically modify specific

bases in rRNA; small Cajal-body RNAs (scaRNAs) that chemically modify spliceosomal snRNA; and certain RNA enzymes (ribozymes) that cleave tRNA and rRNA precursors. All of these types of RNA can be viewed as accessory molecules needed, like rRNA and tRNA, to support protein synthesis in general. In stark contrast to RNA, proteins were viewed as the functionally important endpoints of genetic information, the exciting pacesetters that performed myriad roles in cells.

The view that noncoding RNAs (ncRNAs) are mostly ubiquitous accessory molecules that assist general protein synthesis is no longer tenable. Over the past two decades we have become progressively more aware of the functional diversity of ncRNA and of the many thousands of ncRNA genes in our genome. Multiple new classes of regulatory RNAs have very recently been discovered to be expressed in certain cell types only, or at certain stages of development. Working out what they do has become an exciting area of research.

With hindsight, perhaps we should not be so surprised at the functional diversity of RNA. DNA is simply a self-replicating repository of genetic information, but RNA can serve this function (in the case of RNA viruses) and can also have catalytic functions. In the "RNA world" hypothesis RNA is viewed as the original genetic material and as also being capable of executive functions before DNA and proteins developed. That is possible because, unlike naked double-stranded DNA (which has a comparatively rigid structure), single-stranded RNA has a very flexible structure and can form complex shapes by intramolcular hydrogen bonding, as described below. As will be described in later chapters, the relatively recent understanding of just what RNA does in cells and how it can be manipulated is driving some important advances in medicine. Mutations in certain RNA genes are now known to underlie some genetic disorders and cancers, and RNA therapeutics offers important new approaches to treating disease.

The extraordinary secondary structure and versatility of RNA

The primary structure of nucleic acids and proteins is the sequence of nucleotides or amino acids that defines their identity; however, higher levels of structure determine how they work in cells. Single-stranded RNA molecules are much more flexible than naked double-stranded DNA, and like proteins they have a very high degree of secondary structure where intramolecular hydrogen bonding causes local alterations in structure.

The secondary structure of single-stranded RNA depends on base pairing between complementary sequences on the same RNA strand. Intervening sequences that do not engage in base pairing will loop out, producing stem-loop structures (called hairpins when the loop is short)—see **Figure 2.6**. Higher-level structures can form when, for example, a sequence within the stem of one loop base pairs with another sequence, and extraordinarily intricate structures can develop. Note that base pairing in RNA includes G–U base pairs as well as more stable A–U and G–C base pairs.

Stem-loop structures in RNA have different functions. As described in Chapter 6, they can serve as recognition elements for binding regulatory proteins, and they are crucially important in determining the overall structure of an RNA that can be important for function.

In general, because of the flexible structure of single-stranded RNA, different RNAs can adopt different shapes according to the base sequence; this enables them to do different jobs, such as working as enzymes. Many different classes of RNA enzyme (ribozyme) are known in nature, and some originated very early in evolution. For example, the catalytic activity of the ribosome (the peptidyltransferase responsible for adding amino acids to the growing polypeptide chain) is due solely to the large RNA (28S rRNA) present in the large subunit. In recent years RNAs have been found to work in a large variety of roles (**Figure 2.7**).

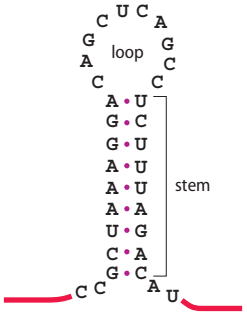

Figure 2.6 A stem-loop structure. This is formed when the RNA folds back on itself so that two short regions can base pair to form the stem while a small intervening sequence loops out. Note that G–U base pairs form in RNA, in addition to G–C and A–U base pairs. Related structures but with shorter stems are important in tRNA structure as shown in Figure 2.4A.

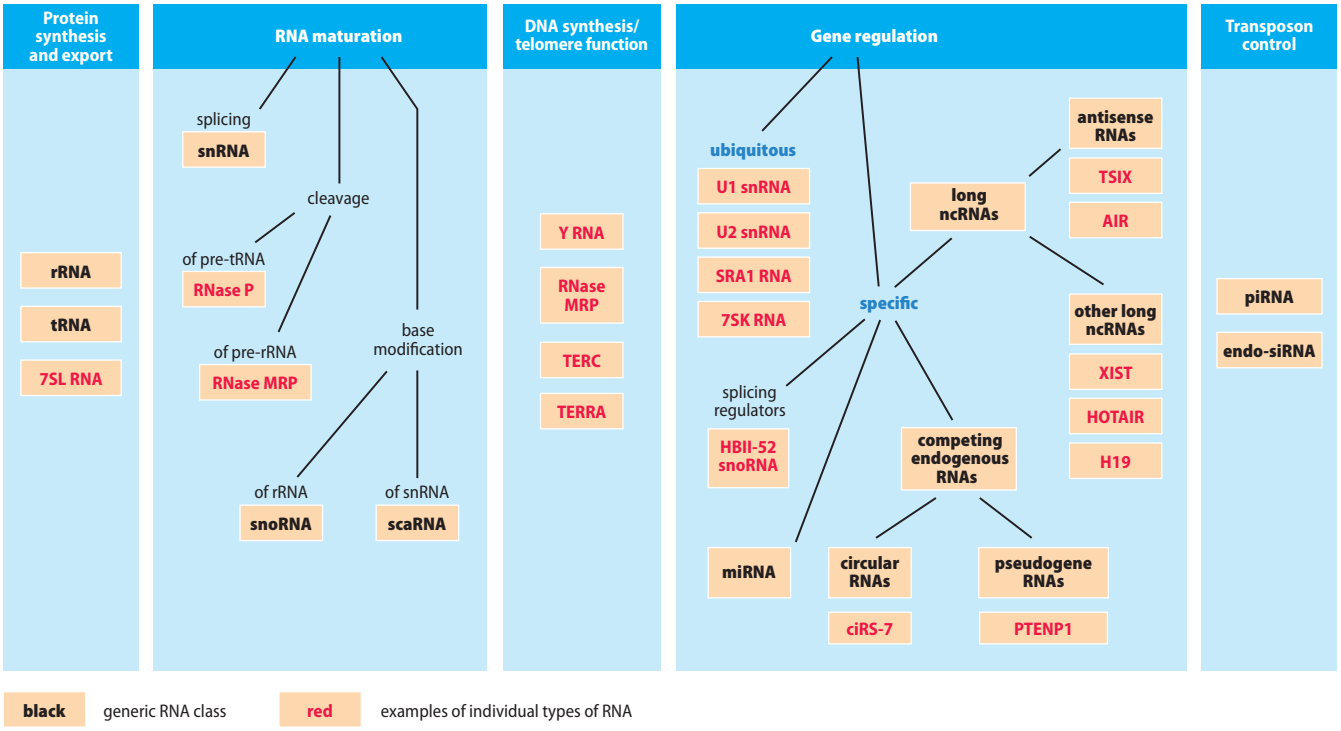

Figure 2.7 The versatility of noncoding RNA. The two panels on the left show ubiquitously expressed RNAs that are important in generally assisting protein production and export, including RNA families that supervise the maturation of other RNAs, notably: small nuclear RNA (snRNA); small nucleolar RNA (snoRNA); and small Cajal-body RNA (scaRNA). The central panel includes RNAs involved in DNA replication (the ribozyme RNase MRP has a crucial role in initiating mtDNA replication, as well as in cleaving pre-rRNA), and developmentally regulated telomere regulators (TERC is the RNA component of telomerase; TERRA is telomere RNA). Diverse classes of noncoding RNA regulate gene expression. In addition to the listed ubiquitous RNAs that have general roles in transcription, many classes of RNA regulate *specific* target genes and are typically restricted in expression. They work at dif erent levels: transcription (such as antisense RNAs), splicing, and translation (notably miRNAs, which bind to certain regulatory sequences in the untranslated regions of target mRNAs). Some RNAs, notably the highly prevalent class of circular RNAs, regulate the interaction between miRNAs and their targets. piRNAs, and to a smaller extent endogenous short interfering RNAs (endo-siRNA), are responsible for silencing transposable elements in germline cells. We describe how RNAs regulate gene expression in detail in Chapter 6.

RNAs that act as specific regulators: from quirky exceptions to the mainstream

The first examples of more specific regulatory RNAs were discovered more than 20 years ago. For a long time they were considered interesting but exceptional cases. They included RNAs working in epigenetic regulation to produce mono-allelic gene expression. For most genes both the maternal and paternal allele are normally expressed, but for a few genes it is *normal* that only one of the two parental alleles is expressed. Some of our genes are *imprinted* so that, according to the specific gene, either the maternal allele or the paternal is consistently expressed, while the other allele is silenced. And in women (and female mammals), genes on one of the two X chromosomes, either the maternal or the paternal X chromosome chosen at random, are *normally* silenced (X-chromosome inactivation). We describe the underlying mechanisms in Chapter 6.

We now know that there are many thousands of different RNA genes in our genome. Many of these genes make regulatory ncRNAs that are expressed in certain cell types only, including some large families of long noncoding RNAs and tiny noncoding RNAs.

Long noncoding RNAs

In addition to the very few ribosomal RNAs, there are a very large number of long noncoding RNAs, the great majority of which are associated with chromatin and act as regulators of gene expression. They come in two broad classes.

Antisense RNAs are transcribed using the *sense* strand of a gene as a template and are not subject to cleavage and RNA splicing. As a result they can be quite large, often many thousands of nucleotides long. They work by binding to the complementary sense RNA produced from the gene, downregulating gene expression.

A second class of long regulatory RNAs are formed from primary transcripts that are typically processed like the primary transcripts of protein-coding genes (and so normally undergo RNA splicing). Many of these RNAs regulate neighboring genes, but some control the expression of genes on other chromosomes. We consider the details of how they work in Chapter 6.

Tiny noncoding RNAs

Thousands of tiny noncoding RNAs (less than 35 nucleotides long) also work in human cells. They include many microRNAs (miRNAs) that are usually 20–22 nucleotides long and are expressed in defined cell types or at specific stages of early development. As described in Chapter 6, a miRNA works by recognizing and binding to defined target regulatory sequences present in specific mRNAs in order to downregulate their expression. MicroRNAs are important in a wide variety of different cellular processes.

Human germ cells also make many thousands of different 26–32-nucleotide Piwi protein-interacting RNAs (piRNAs). The piRNAs work in germ cells to damp down excess activity of **transposons** (mobile DNA elements). Active mobile elements in the human genome can make a copy that migrates to a new location in our genome and can be harmful (by disrupting genes or inappropriately activating some types of cancer gene).

2.3 WORKING OUT THE DETAILS OF OUR GENOME AND WHAT THEY MEAN

The human genome consists of 25 different DNA molecules partitioned between two physically separate genomes, one in the nucleus and one in the mitochondria. In the nucleus there are either 23 or 24 different types of linear DNA molecule (one each for the *different* types of chromosome: 23 in female cells or 24 in male cells). The chromosomal DNA molecules are immensely long (ranging in size from 48 Mb to 249 Mb). In the mitochondria there is just one type of DNA molecule: a comparatively tiny circular DNA just 16.6 kilobases (kb) long, roughly 1/10 000 of the size of an average nuclear DNA molecule. Unlike the chromosomal DNA molecules (each present in only two copies in diploid cells), there are many mitochondrial DNA copies in a cell, and the copy number can vary very significantly according to the type of cell.

In what was a heroic effort at the time, the mitochondrial DNA (often called the mitochondrial genome) was sequenced by a single research team in Cambridge, UK, as far back as 1981. Despite its small size, it is packed with genes. The complexity of the nuclear genome—roughly 200 000 times the size of the mitochondrial genome—posed a much more difficult challenge. That would require an international collaboration between many research teams, as described below.

Working out the nucleotide sequence was only the first step. The next challenge, which is still continuing and may take decades, is to work out the details of how our genome functions and what all the component sequences do.

The Human Genome Project: working out the details of the nuclear genome

For decades, the only available map of the nuclear genome was a low-resolution physical map based on chromosome banding. Chromosomes can be stained with certain dyes, such as Giemsa, to reveal an alternating pattern of dark and light bands for each chromosome, as represented by the image shown in **Figure 2.8**.

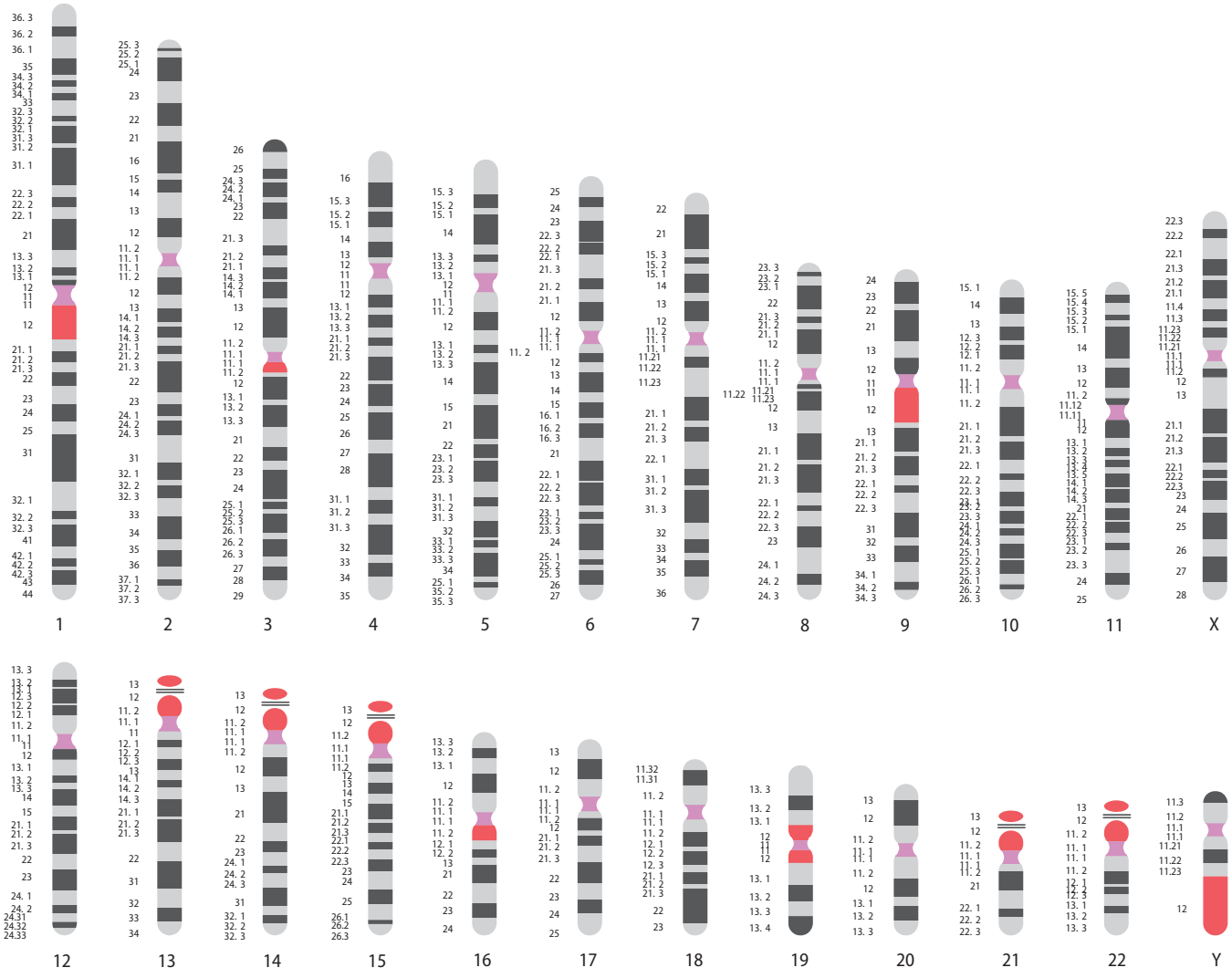

Figure 2.8 Ideogram showing a 550-band Giemsa banding pattern and constitutive heterochromatin within human metaphase chromosomes. Dark bands represent DNA regions where there is a low density of G–C base pairs and a generally low density of exons and genes. Pale bands represent DNA regions where there is a high density of G–C base pairs and a generally high density of exons and genes. Centromeric heterochromatin is illustrated by mauve blocks; non-centromeric and non-telomeric constitutive heterochromatin is shown as bright red blocks. Note the large amounts of non-centromeric heterochromatin on the Y chromosome, the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22), and on chromosomes 1, 9, and 16. Numbers to the left are the numbers of individual chromosome bands; for the nomenclature of chromosome banding, see Box 7.2 on pages 204–5.

We describe the methodology and terminology of human chromosome banding in Box 7.2. For now, there are two salient points to note. First, the alternating pattern of bands reflects different staining intensities. That in turn reflects differences in chromatin organization along chromosomes (as a result of differences in base composition), and differences in gene and exon density (see the legend to Figure 2.8). Secondly, the resolution of the map is low—in even a high-resolution chromosome map the average size of a band is several megabases of DNA. What was needed was a map with a 1 bp resolution, a DNA sequence map.

The principal objective of the international Human Genome Project (HGP) was to obtain a *reference sequence* for our nuclear genome, that is, the aggregated DNA sequences of each of the 24 different human chromosomes. It is necessarily a reference sequence: if we assume 8 billion people on the planet, and without even considering somatic variation, there are at least 16 billion different human genomes (each of us has inherited two different genomes: a maternal