CHAPTER 2

Introduction to the Human Genome

Understanding the organization, variation, and transmission of the **human genome** is central to appreciating the role of genetics in medicine, as well as the emerging principles of genomic and personalized medicine. With the availability of the sequence of the human genome and a growing awareness of the role of genome variation in disease, it is now possible to begin to exploit the impact of that variation on human health on a broad scale. The comparison of individual genomes underscores the first major take-home lesson of this book—*every individual has his or her own unique constitution of gene products, produced in response to the combined inputs of the genome sequence and one's particular set of environmental exposures and experiences*. As pointed out in the previous chapter, this realization reflects what Garrod termed *chemical individuality* over a century ago and provides a conceptual foundation for the practice of genomic and personalized medicine.

Advances in genome technology and the resulting explosion in knowledge and information stemming from the **Human Genome Project** are thus playing an increasingly transformational role in integrating and applying concepts and discoveries in genetics to the practice of medicine.

CHROMOSOME AND GENOME ANALYSIS IN CLINICAL MEDICINE

Chromosome and genome analysis has become an important diagnostic procedure in clinical medicine. As described more fully in subsequent chapters, these applications include the following:

- ***Clinical diagnosis.*** Numerous medical conditions, including some that are common, are associated with changes in chromosome number or structure and require chromosome or genome analysis for diagnosis and genetic counseling (see Chapters 5 and 6).
- ***Gene identification.*** A major goal of medical genetics and genomics today is the identification of specific genes and elucidating their roles in health and disease. This topic is referred to repeatedly but is discussed in detail in Chapter 10.
- ***Cancer genomics.*** Genomic and chromosomal changes in somatic cells are involved in the initiation and progression of many types of cancer (see Chapter 15).
- ***Disease treatment.*** Evaluation of the integrity, composition, and differentiation state of the genome is critical for the development of patient-specific pluripotent stem cells for therapeutic use (see Chapter 13).
- ***Prenatal diagnosis.*** Chromosome and genome analysis is an essential procedure in prenatal diagnosis (see Chapter 17).

THE HUMAN GENOME AND THE CHROMOSOMAL BASIS OF HEREDITY

Appreciation of the importance of genetics to medicine requires an understanding of the nature of the hereditary material, how it is packaged into the human genome, and how it is transmitted from cell to cell during cell division and from generation to generation during reproduction. The human genome consists of large amounts of the chemical deoxyribonucleic acid (**DNA**) that contains within its structure the genetic information needed to specify all aspects of embryogenesis, development, growth, metabolism, and reproduction—essentially all aspects of what makes a human being a functional organism. Every nucleated cell in the body carries its own copy of the human genome, which contains, depending on how one defines the term, approximately 20,000 to 50,000 **genes** (see Box later). Genes, which at this point we consider simply and most broadly as functional units of genetic information, are encoded in the DNA of the genome, organized into a number of rod-shaped organelles called **chromosomes** in the nucleus of each cell. The influence of genes and genetics on states of health and disease is profound, and its roots are found in the information encoded in the DNA that makes up the human genome.

Each species has a characteristic chromosome complement (**karyotype**) in terms of the number, morphology, and content of the chromosomes that make up its genome. The genes are in linear order along the chromosomes, each gene having a precise position or **locus**. A **gene map** is the map of the genomic location of the genes and is characteristic of each species and the individuals within a species.

The study of chromosomes, their structure, and their inheritance is called **cytogenetics**. The science of human cytogenetics dates from 1956, when it was first established that the normal human chromosome number is 46. Since that time, much has been learned about human chromosomes, their normal structure and composition, and the identity of the genes that they contain, as well as their numerous and varied abnormalities.

With the exception of cells that develop into gametes (the **germline**), all cells that contribute to one's body are called **somatic cells** (*soma,* body). The genome contained in the nucleus of human somatic cells consists of 46 chromosomes, made up of 24 different types and arranged in 23 pairs (Fig. 2-1). Of those 23 pairs, 22 are alike in males and females and are called **autosomes,** originally numbered in order of their apparent size from the largest to the smallest. The remaining pair comprises the two different types of **sex chromosomes:** an X and a Y chromosome in males and two X chromosomes in females. Central to the concept of the human genome, each chromosome carries a different subset of genes that are arranged linearly along its DNA. Members of a pair of chromosomes (referred to as **homologous chromosomes** or **homologues**) carry matching genetic information; that is, they typically have the same genes in the same order. At any specific locus, however, the homologues either may be identical or may vary slightly in sequence; these different forms of a gene are called **alleles**. One member of each pair of chromosomes is inherited from the father, the other from the mother. Normally, the members of a pair of autosomes are microscopically indistinguishable from each other. In females, the sex chromosomes, the two **X chromosomes,** are likewise largely indistinguishable. In males, however, the sex chromosomes differ. One is an X, identical to the Xs of the female, inherited by a male from his mother and transmitted to his daughters; the other, the **Y chromosome,** is inherited from his father and transmitted to his sons. In Chapter 6, as we explore the chromosomal and genomic basis of disease, we will look at some exceptions to the simple and almost universal rule that human females are XX and human males are XY.

In addition to the nuclear genome, a small but important part of the human genome resides in mitochondria in the cytoplasm (see Fig. 2-1). The mitochondrial chromosome, to be described later in this chapter, has a number of unusual features that distinguish it from the rest of the human genome.

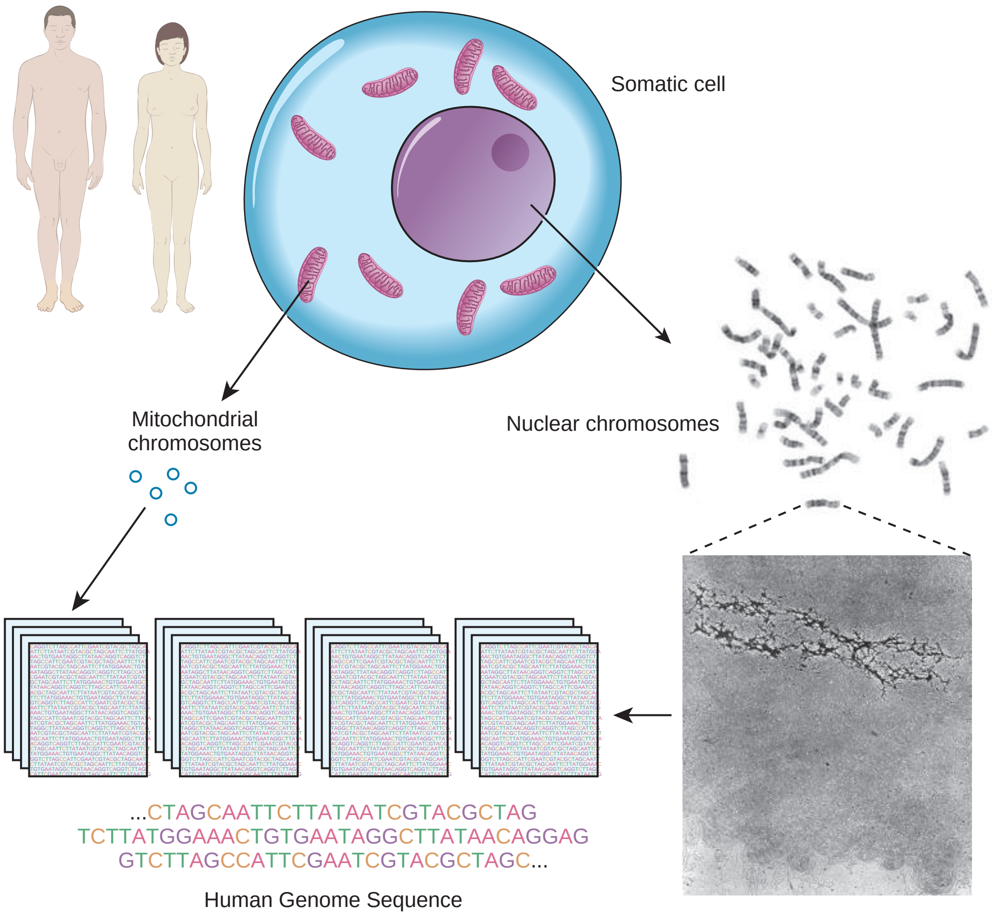
Somatic cell
Mitochondrial chromosomes
Nuclear chromosomes
...CTAGCAATTCTTATAATCGTACGCTAG
TCTTATGGAAACTGTGAATAGGCTTATAACAGGAG
GTCTTAGCCATTCGAATCGTACGCTAGC...
Human Genome Sequence

Figure 2-1 The human genome, encoded on both nuclear and mitochondrial chromosomes. *See Sources & Acknowledgments.*

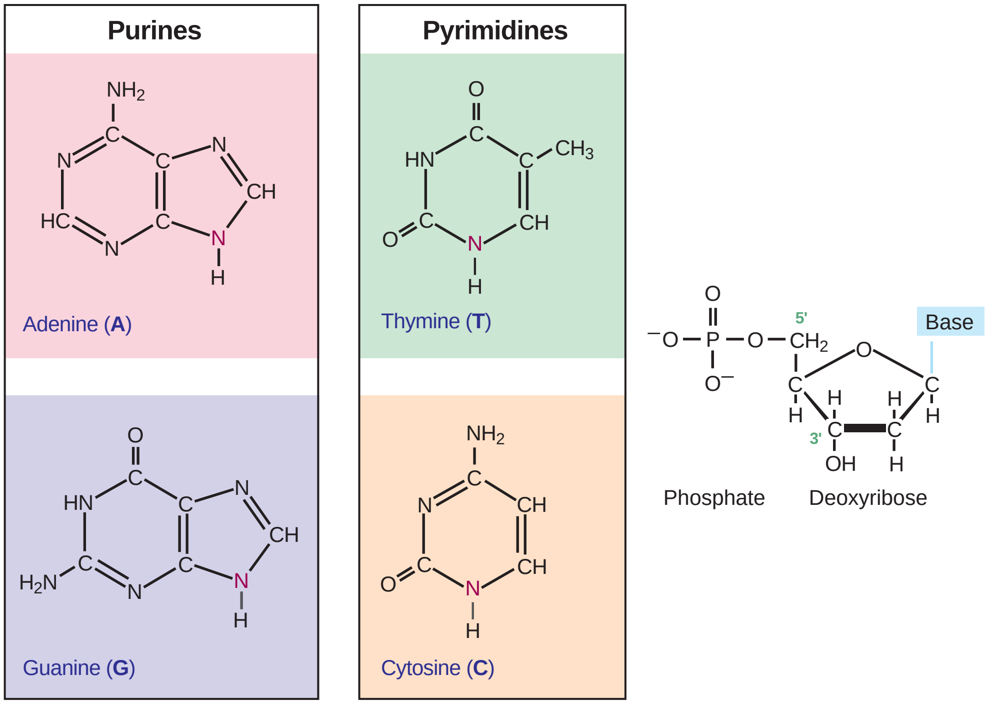

Figure 2-2 **The four bases of DNA and the general structure of a nucleotide in DNA.** Each of the four bases bonds with deoxyribose (through the nitrogen shown in *magenta*) and a phosphate group to form the corresponding nucleotides.

GENES IN THE HUMAN GENOME

What is a gene? And how many genes do we have? These questions are more difficult to answer than it might seem.

The word *gene*, first introduced in 1908, has been used in many different contexts since the essential features of heritable "unit characters" were first outlined by Mendel over 150 years ago. To physicians (and indeed to Mendel and other early geneticists), a gene can be defined by its observable impact on an organism and on its statistically determined transmission from generation to generation. To medical geneticists, a gene is recognized clinically in the context of an observable variant that leads to a characteristic clinical disorder, and today we recognize approximately 5000 such conditions (see Chapter 7).

The Human Genome Project provided a more systematic basis for delineating human genes, relying on DNA sequence analysis rather than clinical acumen and family studies alone; indeed, this was one of the most compelling rationales for initiating the project in the late 1980s. However, even with the finished sequence product in 2003, it was apparent that our ability to recognize features of the sequence that point to the existence or identity of a gene was sorely lacking. Interpreting the human genome sequence and relating its variation to human biology in both health and disease is thus an ongoing challenge for biomedical research.

Although the ultimate catalogue of human genes remains an elusive target, we recognize two general types of gene, those whose product is a protein and those whose product is a functional RNA.

- The number of **protein-coding genes**—recognized by features in the genome that will be discussed in Chapter 3—is estimated to be somewhere between 20,000 and 25,000. In this book, we typically use approximately 20,000 as the number, and the reader should recognize that this is both imprecise and perhaps an underestimate.
- In addition, however, it has been clear for several decades that the ultimate product of some genes is not a protein at all but rather an RNA transcribed from the DNA sequence. There are many different types of such RNA genes (typically called **noncoding genes** to distinguish them from protein-coding genes), and it is currently estimated that there are at least another 20,000 to 25,000 noncoding RNA genes around the human genome.

Thus overall—and depending on what one means by the term—the total number of genes in the human genome is of the order of approximately 20,000 to 50,000. However, the reader will appreciate that this remains a moving target, subject to evolving definitions, increases in technological capabilities and analytical precision, advances in informatics and digital medicine, and more complete genome annotation.

DNA Structure: A Brief Review

Before the organization of the human genome and its chromosomes is considered in detail, it is necessary to review the nature of the DNA that makes up the genome. DNA is a polymeric nucleic acid macromolecule composed of three types of units: a five-carbon sugar, deoxyribose; a nitrogen-containing base; and a phosphate group (Fig. 2-2). The bases are of two types, **purines** and **pyrimidines**. In DNA, there are two purine bases, **adenine** (A) and **guanine** (G), and two pyrimidine

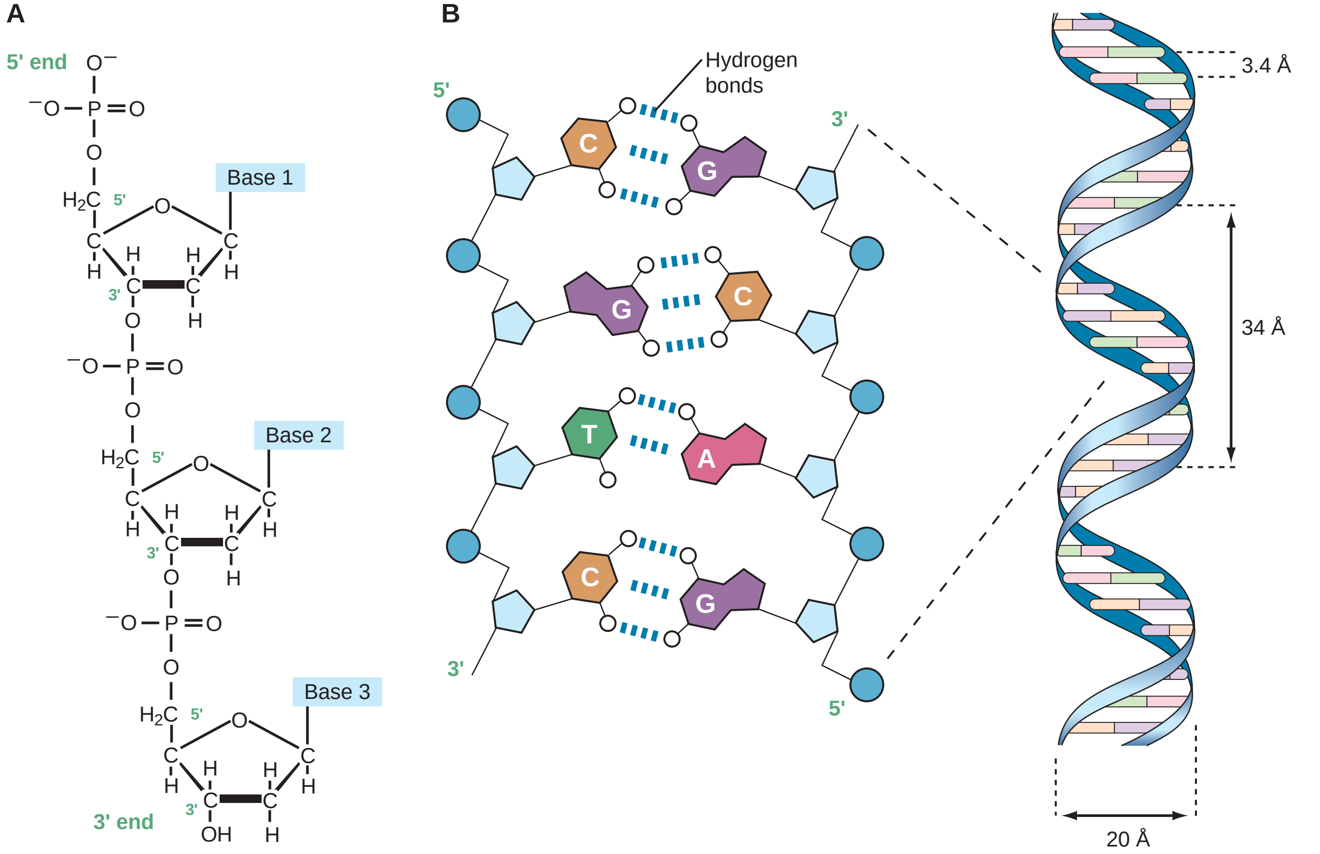

Figure 2-3 **The structure of DNA. A,** A portion of a DNA polynucleotide chain, showing the 3′-5′ phosphodiester bonds that link adjacent nucleotides. **B,** The double-helix model of DNA, as proposed by Watson and Crick. The horizontal "rungs" represent the paired bases. The helix is said to be right-handed because the strand going from lower left to upper right crosses over the opposite strand. The detailed portion of the figure illustrates the two complementary strands of DNA, showing the AT and GC base pairs. Note that the orientation of the two strands is antiparallel. *See Sources & Acknowledgments.*

bases, **thymine** (T) and **cytosine** (C). Nucleotides, each composed of a base, a phosphate, and a sugar moiety, polymerize into long polynucleotide chains held together by 5′-3′ phosphodiester bonds formed between adjacent deoxyribose units (Fig. 2-3A). In the human genome, these polynucleotide chains exist in the form of a double helix (Fig. 2-3B) that can be hundreds of millions of nucleotides long in the case of the largest human chromosomes.

The anatomical structure of DNA carries the chemical information that allows the exact transmission of genetic information from one cell to its daughter cells and from one generation to the next. At the same time, the primary structure of DNA specifies the amino acid sequences of the polypeptide chains of proteins, as described in the next chapter. DNA has elegant features that give it these properties. The native state of DNA, as elucidated by James Watson and Francis Crick in 1953, is a double helix (see Fig. 2-3B). The helical structure resembles a right-handed spiral staircase in which its two polynucleotide chains run in opposite directions, held together by hydrogen bonds between pairs of bases: T of one chain paired with A of the other, and G with C. The specific nature of the genetic information encoded in the human genome lies in the sequence of C's, A's, G's, and T's on the two strands of the double helix along each of the chromosomes, both in the nucleus and in mitochondria (see Fig. 2-1). Because of the complementary nature of the two strands of DNA, knowledge of the sequence of nucleotide bases on one strand automatically allows one to determine the sequence of bases on the other strand. The double-stranded structure of DNA molecules allows them to replicate precisely by separation of the two strands, followed by synthesis of two new complementary strands, in accordance with the sequence of the original template strands (Fig. 2-4). Similarly, when necessary, the base complementarity allows efficient and correct repair of damaged DNA molecules.

Structure of Human Chromosomes

The composition of genes in the human genome, as well as the determinants of their expression, is specified in the DNA of the 46 human chromosomes in the nucleus plus the mitochondrial chromosome. *Each human* *chromosome consists of a single, continuous DNA double helix;* that is, each chromosome is one long, double-stranded DNA molecule, and the nuclear genome consists, therefore, of 46 linear DNA molecules, totaling more than 6 billion nucleotide pairs (see Fig. 2-1).

Chromosomes are not naked DNA double helices, however. Within each cell, the genome is packaged as **chromatin,** in which genomic DNA is complexed with several classes of specialized proteins. Except during cell division, chromatin is distributed throughout the nucleus and is relatively homogeneous in appearance under the microscope. When a cell divides, however, its genome condenses to appear as microscopically visible chromosomes. Chromosomes are thus visible as discrete structures only in dividing cells, although they retain their integrity between cell divisions.

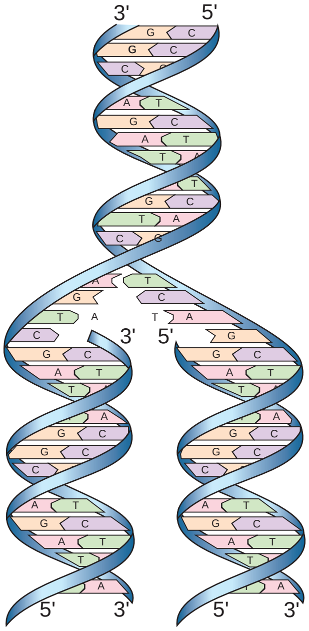

Figure 2-4 Replication of a DNA double helix, resulting in two identical daughter molecules, each composed of one parental strand and one newly synthesized strand.

The DNA molecule of a chromosome exists in chromatin as a complex with a family of basic chromosomal proteins called *histones*. This fundamental unit interacts with a heterogeneous group of nonhistone proteins, which are involved in establishing a proper spatial and functional environment to ensure normal chromosome behavior and appropriate gene expression.

Five major types of histones play a critical role in the proper packaging of chromatin. Two copies each of the four core histones H2A, H2B, H3, and H4 constitute an octamer, around which a segment of DNA double helix winds, like thread around a spool (Fig. 2-5). Approximately 140 base pairs (bp) of DNA are associated with each histone core, making just under two turns around the octamer. After a short (20- to 60-bp) "spacer" segment of DNA, the next core DNA complex forms, and so on, giving chromatin the appearance of beads on a string. Each complex of DNA with core histones is called a **nucleosome** (see Fig. 2-5), which is the basic structural unit of chromatin, and each of the 46 human chromosomes contains several hundred thousand to well over a million nucleosomes. A fifth histone, H1, appears to bind to DNA at the edge of each nucleosome, in the internucleosomal spacer region. The amount of

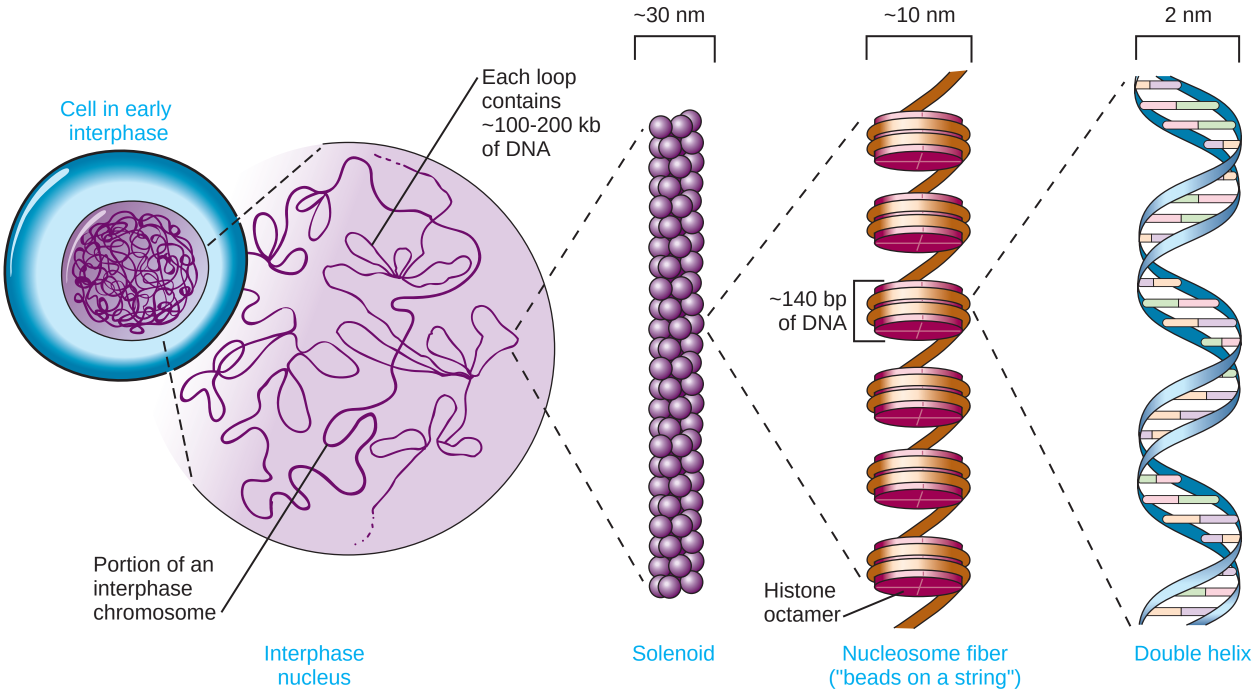
Cell in early interphase
Each loop contains ~100-200 kb of DNA
~30 nm
~10 nm
2 nm
~140 bp of DNA
Portion of an interphase chromosome
Histone octamer
Interphase nucleus
Solenoid
Nucleosome fiber ("beads on a string")
Double helix

Figure 2-5 Hierarchical levels of chromatin packaging in a human chromosome.

DNA associated with a core nucleosome, together with the spacer region, is approximately 200 bp.

In addition to the major histone types, a number of specialized histones can substitute for H3 or H2A and confer specific characteristics on the genomic DNA at that location. Histones can also be modified by chemical changes, and these modifications can change the properties of nucleosomes that contain them. As discussed further in Chapter 3, the pattern of major and specialized histone types and their modifications can vary from cell type to cell type and is thought to specify how DNA is packaged and how accessible it is to regulatory molecules that determine gene expression or other genome functions.

During the cell cycle, as we will see later in this chapter, chromosomes pass through orderly stages of condensation and decondensation. However, even when chromosomes are in their most decondensed state, in a stage of the cell cycle called **interphase,** DNA packaged in chromatin is substantially more condensed than it would be as a native, protein-free, double helix. Further, the long strings of nucleosomes are themselves compacted into a secondary helical structure, a cylindrical "solenoid" fiber (from the Greek *solenoeides*, pipe-shaped) that appears to be the fundamental unit of chromatin organization (see Fig. 2-5). The solenoids are themselves packed into **loops** or domains attached at intervals of approximately 100,000 bp (equivalent to 100 kilobase pairs [kb], because 1 kb = 1000 bp) to a protein **scaffold** within the nucleus. It has been speculated that these loops are the functional units of the genome and that the attachment points of each loop are specified along the chromosomal DNA. As we shall see, one level of control of gene expression depends on how DNA and genes are packaged into chromosomes and on their association with chromatin proteins in the packaging process.

The enormous amount of genomic DNA packaged into a chromosome can be appreciated when chromosomes are treated to release the DNA from the underlying protein scaffold (see Fig. 2-1). When DNA is released in this manner, long loops of DNA can be visualized, and the residual scaffolding can be seen to reproduce the outline of a typical chromosome.

The Mitochondrial Chromosome

As mentioned earlier, a small but important subset of genes encoded in the human genome resides in the cytoplasm in the mitochondria (see Fig. 2-1). Mitochondrial genes exhibit exclusively maternal inheritance (see Chapter 7). Human cells can have hundreds to thousands of mitochondria, each containing a number of copies of a small circular molecule, the mitochondrial chromosome. The mitochondrial DNA molecule is only 16 kb in length (just a tiny fraction of the length of even the smallest nuclear chromosome) and encodes only 37 genes. The products of these genes function in mitochondria, although the vast majority of proteins within the mitochondria are, in fact, the products of nuclear genes. Mutations in mitochondrial genes have been demonstrated in several maternally inherited as well as sporadic disorders (Case 33) (see Chapters 7 and 12).

The Human Genome Sequence

With a general understanding of the structure and clinical importance of chromosomes and the genes they carry, scientists turned attention to the identification of specific genes and their location in the human genome. From this broad effort emerged the **Human Genome Project,** an international consortium of hundreds of laboratories around the world, formed to determine and assemble the sequence of the 3.3 billion base pairs of DNA located among the 24 types of human chromosome.

Over the course of a decade and a half, powered by major developments in DNA-sequencing technology, large sequencing centers collaborated to assemble sequences of each chromosome. The genomes actually being sequenced came from several different individuals, and the consensus sequence that resulted at the conclusion of the Human Genome Project was reported in 2003 as a "reference" sequence assembly, to be used as a basis for later comparison with sequences of individual genomes. This reference sequence is maintained in publicly accessible databases to facilitate scientific discovery and its translation into useful advances for medicine. Genome sequences are typically presented in a 5′ to 3′ direction on just one of the two strands of the double helix, because—owing to the complementary nature of DNA structure described earlier—if one knows the sequence of one strand, one can infer the sequence of the other strand (Fig. 2-6).

Organization of the Human Genome

Chromosomes are not just a random collection of different types of genes and other DNA sequences. Regions of the genome with similar characteristics tend to be clustered together, and the functional organization of the genome reflects its structural organization and sequence. Some chromosome regions, or even whole chromosomes, are high in gene content ("gene rich"), whereas others are low ("gene poor") (Fig. 2-7). The clinical consequences of abnormalities of genome structure reflect the specific nature of the genes and sequences involved. Thus abnormalities of gene-rich chromosomes or chromosomal regions tend to be much more severe clinically than similar-sized defects involving gene-poor parts of the genome.

As a result of knowledge gained from the Human Genome Project, it is apparent that the organization of DNA in the human genome is both more varied and

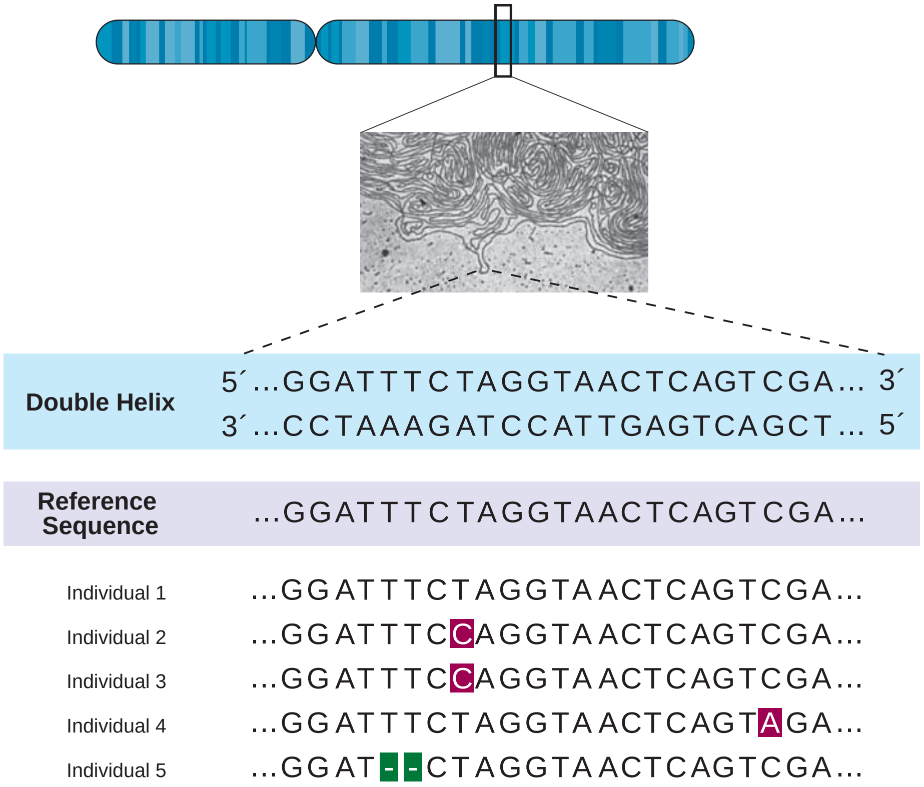
Double Helix
5´...GGATTTCTAGGTAACTCAGTCGA... 3´
3´...CCTAAAGATCCATTGAGTCAGCT... 5´
Reference Sequence
...GGATTTCTAGGTAACTCAGTCGA...
Individual 1 ...GGATTTCTAGGTAACTCAGTCGA...
Individual 2 ...GGATTTCCAGGTAACTCAGTCGA...
Individual 3 ...GGATTTCCAGGTAACTCAGTCGA...
Individual 4 ...GGATTTCTAGGTAACTCAGTAGA...
Individual 5 ...GGAT--CTAGGTAACTCAGTCGA...

Figure 2-6 **A portion of the reference human genome sequence.** By convention, sequences are presented from one strand of DNA only, because the sequence of the complementary strand can be inferred from the double-stranded nature of DNA (shown above the reference sequence). The sequence of DNA from a group of individuals is similar but not identical to the reference, with single nucleotide changes in some individuals and a small deletion of two bases in another.

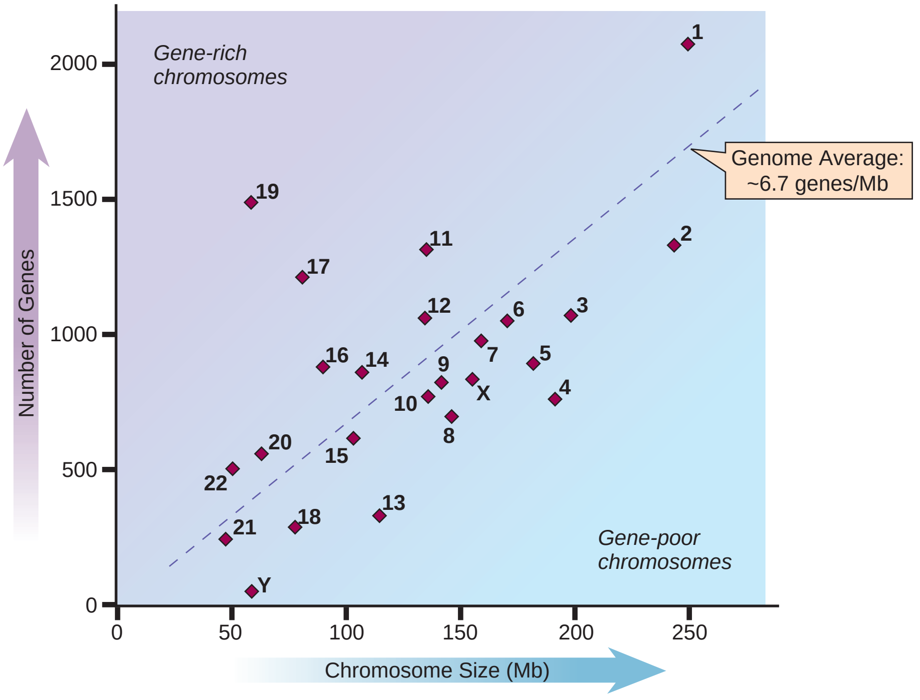

Figure 2-7 **Size and gene content of the 24 human chromosomes.** *Dotted diagonal line* corresponds to the average density of genes in the genome, approximately 6.7 protein-coding genes per megabase (Mb). Chromosomes that are relatively gene rich are above the diagonal and trend to the upper left. Chromosomes that are relatively gene poor are below the diagonal and trend to the lower right. *See Sources & Acknowledgments.*

more complex than was once appreciated. Of the billions of base pairs of DNA in any genome, less than 1.5% actually encodes proteins. Regulatory elements that influence or determine patterns of gene expression during development or in tissues were believed to account for only approximately 5% of additional sequence, although more recent analyses of chromatin characteristics suggest that a much higher proportion of the genome may provide signals that are relevant to genome functions. Only approximately half of the total linear length of the genome consists of so-called **single-copy** or **unique DNA,** that is, DNA whose linear order of specific nucleotides is represented only once (or at most a few times) around the entire genome. This concept may appear surprising to some, given that there are only four different nucleotides in DNA. But, consider even a tiny stretch of the genome that is only 10 bases long; with four types of bases, there are over a million possible sequences. And, although the order of bases in the genome is not entirely random, any particular 16-base sequence would be predicted by chance alone to appear only once in any given genome.

The rest of the genome consists of several classes of **repetitive DNA** and includes DNA whose nucleotide sequence is repeated, either perfectly or with some variation, hundreds to millions of times in the genome. Whereas most (but not all) of the estimated 20,000 protein-coding genes in the genome (see Box earlier in this chapter) are represented in single-copy DNA, sequences in the repetitive DNA fraction contribute to maintaining chromosome structure and are an important source of variation between different individuals; some of this variation can predispose to pathological events in the genome, as we will see in Chapters 5 and 6.

Single-Copy DNA Sequences

Although single-copy DNA makes up at least half of the DNA in the genome, much of its function remains a mystery because, as mentioned, sequences actually encoding proteins (i.e., the coding portion of genes) constitute only a small proportion of all the single-copy DNA. Most single-copy DNA is found in short stretches (several kilobase pairs or less), interspersed with members of various repetitive DNA families. The organization of genes in single-copy DNA is addressed in depth in Chapter 3.

Repetitive DNA Sequences

Several different categories of repetitive DNA are recognized. A useful distinguishing feature is whether the repeated sequences ("repeats") are clustered in one or a few locations or whether they are interspersed with single-copy sequences along the chromosome. Clustered repeated sequences constitute an estimated 10% to 15% of the genome and consist of arrays of various short repeats organized in tandem in a head-to-tail fashion. The different types of such tandem repeats are collectively called **satellite DNAs,** so named because many of the original tandem repeat families could be separated by biochemical methods from the bulk of the genome as distinct ("satellite") fractions of DNA.

Tandem repeat families vary with regard to their location in the genome and the nature of sequences that make up the array. In general, such arrays can stretch several million base pairs or more in length and constitute up to several percent of the DNA content of an individual human chromosome. Some tandem repeat sequences are important as tools that are useful in clinical cytogenetic analysis (see Chapter 5). Long arrays of repeats based on repetitions (with some variation) of a short sequence such as a pentanucleotide are found in large genetically inert regions on chromosomes 1, 9, and 16 and make up more than half of the Y chromosome (see Chapter 6). Other tandem repeat families are based on somewhat longer basic repeats. For example, the α-satellite family of DNA is composed of tandem arrays of an approximately 171-bp unit, found at the **centromere** of each human chromosome, which is critical for attachment of chromosomes to microtubules of the spindle apparatus during cell division.

In addition to tandem repeat DNAs, another major class of repetitive DNA in the genome consists of related sequences that are dispersed throughout the genome rather than clustered in one or a few locations. Although many DNA families meet this general description, two in particular warrant discussion because together they make up a significant proportion of the genome and because they have been implicated in genetic diseases. Among the best-studied dispersed repetitive elements are those belonging to the so-called ***Alu* family**. The members of this family are approximately 300 bp in length and are related to each other although not identical in DNA sequence. In total, there are more than a million *Alu* family members in the genome, making up at least 10% of human DNA. A second major dispersed repetitive DNA family is called the long interspersed nuclear element (**LINE,** sometimes called L1) family. LINEs are up to 6 kb in length and are found in approximately 850,000 copies per genome, accounting for nearly 20% of the genome. Both of these families are plentiful in some regions of the genome but relatively sparse in others—regions rich in GC content tend to be enriched in *Alu* elements but depleted of LINE sequences, whereas the opposite is true of more AT-rich regions of the genome.

Repetitive DNA and Disease. Both *Alu* and LINE sequences have been implicated as the cause of mutations in hereditary disease. At least a few copies of the LINE and *Alu* families generate copies of themselves that can integrate elsewhere in the genome, occasionally causing insertional inactivation of a medically important gene. The frequency of such events causing genetic disease in humans is unknown, but they may account for as many as 1 in 500 mutations. In addition, aberrant recombination events between different LINE repeats or *Alu* repeats can also be a cause of mutation in some genetic diseases (see Chapter 12).

An important additional type of repetitive DNA found in many different locations around the genome includes sequences that are duplicated, often with extraordinarily high sequence conservation. Duplications involving substantial segments of a chromosome, called **segmental duplications,** can span hundreds of kilobase pairs and account for at least 5% of the genome. When the duplicated regions contain genes, genomic rearrangements involving the duplicated sequences can result in the deletion of the region (and the genes) between the copies and thus give rise to disease (see Chapters 5 and 6).

VARIATION IN THE HUMAN GENOME

With completion of the reference human genome sequence, much attention has turned to the discovery and cataloguing of variation in sequence among different individuals (including both healthy individuals and those with various diseases) and among different populations around the globe. As we will explore in much more detail in Chapter 4, there are many tens of millions of common sequence variants that are seen at significant frequency in one or more populations; any given individual carries at least 5 million of these sequence variants. In addition, there are countless very rare variants, many of which probably exist in only a single or a few individuals. In fact, given the number of individuals in our species, *essentially each and every base pair in the human genome is expected to vary in someone somewhere around the globe*. It is for this reason that the original human genome sequence is considered a "reference" sequence for our species, but one that is actually identical to no individual's genome.

Early estimates were that any two randomly selected individuals would have sequences that are 99.9% identical or, put another way, that an individual genome would carry two *different* versions (**alleles**) of the human genome sequence at some 3 to 5 million positions, with different bases (e.g., a T or a G) at the maternally and paternally inherited copies of that particular sequence position (see Fig. 2-6). Although many of these allelic differences involve simply one nucleotide, much of the variation consists of insertions or deletions of (usually) short sequence stretches, variation in the number of copies of repeated elements (including genes), or inversions in the order of sequences at a particular position (**locus**) in the genome (see Chapter 4).

The total amount of the genome involved in such variation is now known to be substantially more than originally estimated and approaches 0.5% between any two randomly selected individuals. As will be addressed in future chapters, any and all of these types of variation can influence biological function and thus must be accounted for in any attempt to understand the contribution of genetics to human health.

TRANSMISSION OF THE GENOME

The chromosomal basis of heredity lies in the copying of the genome and its transmission from a cell to its progeny during typical cell division and from one generation to the next during reproduction, when single copies of the genome from each parent come together in a new embryo.

To achieve these related but distinct forms of genome inheritance, there are two kinds of cell division, mitosis and meiosis. **Mitosis** is ordinary somatic cell division by which the body grows, differentiates, and effects tissue regeneration. Mitotic division normally results in two daughter cells, each with chromosomes and genes identical to those of the parent cell. There may be dozens or even hundreds of successive mitoses in a lineage of somatic cells. In contrast, **meiosis** occurs only in cells of the germline. Meiosis results in the formation of reproductive cells (**gametes**), each of which has only 23 chromosomes—one of each kind of autosome and either an X or a Y. Thus, whereas somatic cells have the **diploid** (*diploos*, double) or the 2n chromosome complement (i.e., 46 chromosomes), gametes have the **haploid** (*haploos*, single) or the n complement (i.e., 23 chromosomes). Abnormalities of chromosome number or structure, which are usually clinically significant, can arise either in somatic cells or in cells of the germline by errors in cell division.

The Cell Cycle

A human being begins life as a fertilized ovum (**zygote**), a diploid cell from which all the cells of the body (estimated to be approximately 100 trillion in number) are derived by a series of dozens or even hundreds of mitoses. Mitosis is obviously crucial for growth and differentiation, but it takes up only a small part of the life cycle of a cell. The period between two successive mitoses is called **interphase,** the state in which most of the life of a cell is spent.

Immediately after mitosis, the cell enters a phase, called G_1, in which there is no DNA synthesis (Fig. 2-8). Some cells pass through this stage in hours; others spend a long time, days or years, in G_1. In fact, some cell types, such as neurons and red blood cells, do not divide at all once they are fully differentiated; rather, they are permanently arrested in a distinct phase known as G_0 ("G zero"). Other cells, such as liver cells, may enter G_0 but, after organ damage, return to G_1 and continue through the cell cycle.

The cell cycle is governed by a series of **checkpoints** that determine the timing of each step in mitosis. In

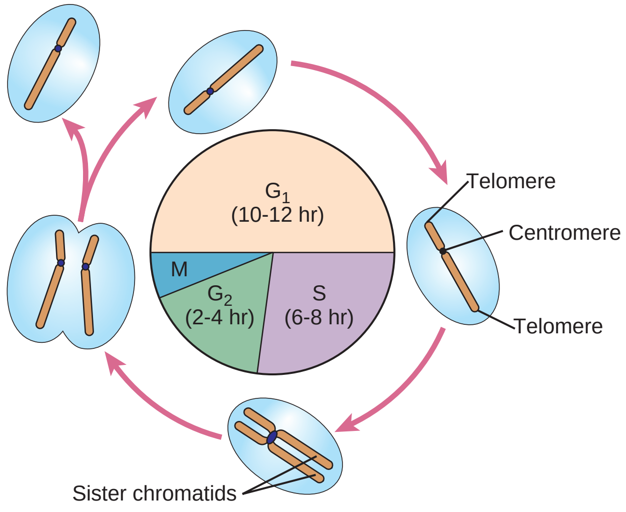

Figure 2-8 **A typical mitotic cell cycle, described in the text.** The telomeres, the centromere, and sister chromatids are indicated.

addition, checkpoints monitor and control the accuracy of DNA synthesis as well as the assembly and attachment of an elaborate network of microtubules that facilitate chromosome movement. If damage to the genome is detected, these mitotic checkpoints halt cell cycle progression until repairs are made or, if the damage is excessive, until the cell is instructed to die by programmed cell death (a process called **apoptosis**).

During G_1, each cell contains one diploid copy of the genome. As the process of cell division begins, the cell enters **S phase,** the stage of programmed DNA synthesis, ultimately leading to the precise replication of each chromosome's DNA. During this stage, each chromosome, which in G_1 has been a single DNA molecule, is duplicated and consists of two **sister chromatids** (see Fig. 2-8), each of which contains an identical copy of the original linear DNA double helix. The two sister chromatids are held together physically at the **centromere,** a region of DNA that associates with a number of specific proteins to form the **kinetochore**. This complex structure serves to attach each chromosome to the microtubules of the **mitotic spindle** and to govern chromosome movement during mitosis. DNA synthesis during S phase is not synchronous throughout all chromosomes or even within a single chromosome; rather, along each chromosome, it begins at hundreds to thousands of sites, called **origins of DNA replication**. Individual chromosome segments have their own characteristic time of replication during the 6- to 8-hour S phase. The ends of each chromosome (or chromatid) are marked by **telomeres**, which consist of specialized repetitive DNA sequences that ensure the integrity of the chromosome during cell division. Correct maintenance of the ends of chromosomes requires a special enzyme called **telomerase,** which ensures that the very ends of each chromosome are replicated.

The essential nature of these structural elements of chromosomes and their role in ensuring genome integrity is illustrated by a range of clinical conditions that result from defects in elements of the telomere or kinetochore or cell cycle machinery or from inaccurate replication of even small portions of the genome (see Box). Some of these conditions will be presented in greater detail in subsequent chapters.

CLINICAL CONSEQUENCES OF ABNORMALITIES AND VARIATION IN CHROMOSOME STRUCTURE AND MECHANICS

Medically relevant conditions arising from abnormal structure or function of chromosomal elements during cell division include the following:

- A broad spectrum of congenital abnormalities in children with inherited defects in genes encoding key components of the mitotic spindle checkpoint at the kinetochore
- A range of **birth defects and developmental disorders** due to anomalous segregation of chromosomes with multiple or missing centromeres (see Chapter 6)
- A variety of cancers associated with overreplication (amplification) or altered timing of replication of specific regions of the genome in S phase (see Chapter 15)
- **Roberts syndrome** of growth retardation, limb shortening, and microcephaly in children with abnormalities of a gene required for proper sister chromatid alignment and cohesion in S phase
- **Premature ovarian failure** as a major cause of female infertility due to mutation in a meiosis-specific gene required for correct sister chromatid cohesion
- The so-called **telomere syndromes,** a number of degenerative disorders presenting from childhood to adulthood in patients with abnormal telomere shortening due to defects in components of telomerase
- And, at the other end of the spectrum, common gene variants that correlate with the number of copies of the repeats at telomeres and with life expectancy and **longevity**

By the end of S phase, the DNA content of the cell has doubled, and each cell now contains two copies of the diploid genome. After S phase, the cell enters a brief stage called G_2. Throughout the whole cell cycle, the cell gradually enlarges, eventually doubling its total mass before the next mitosis. G_2 is ended by mitosis, which begins when individual chromosomes begin to condense and become visible under the microscope as thin, extended threads, a process that is considered in greater detail in the following section.

The G_1, S, and G_2 phases together constitute interphase. In typical dividing human cells, the three phases take a total of 16 to 24 hours, whereas mitosis lasts only 1 to 2 hours (see Fig. 2-8). There is great variation, however, in the length of the cell cycle, which ranges from a few hours in rapidly dividing cells, such as those of the dermis of the skin or the intestinal mucosa, to months in other cell types.

Mitosis

During the mitotic phase of the cell cycle, an elaborate apparatus ensures that each of the two daughter cells receives a complete set of genetic information. This result is achieved by a mechanism that distributes one chromatid of each chromosome to each daughter cell (Fig. 2-9). The process of distributing a copy of each chromosome to each daughter cell is called **chromosome segregation.** The importance of this process for normal cell growth is illustrated by the observation that many tumors are invariably characterized by a state of genetic imbalance resulting from mitotic errors in the distribution of chromosomes to daughter cells.

The process of mitosis is continuous, but five stages, illustrated in Figure 2-9, are distinguished: prophase, prometaphase, metaphase, anaphase, and telophase.

- ***Prophase.*** This stage is marked by gradual condensation of the chromosomes, formation of the mitotic spindle, and formation of a pair of **centrosomes,** from which microtubules radiate and eventually take up positions at the poles of the cell.
- ***Prometaphase.*** Here, the nuclear membrane dissolves, allowing the chromosomes to disperse within the cell and to attach, by their kinetochores, to microtubules of the mitotic spindle.
- ***Metaphase.*** At this stage, the chromosomes are maximally condensed and line up at the equatorial plane of the cell.
- ***Anaphase.*** The chromosomes separate at the centromere, and the sister chromatids of each chromosome now become independent **daughter chromosomes,** which move to opposite poles of the cell.
- ***Telophase.*** Now, the chromosomes begin to decondense from their highly contracted state, and a nuclear membrane begins to re-form around each of the two daughter nuclei, which resume their interphase appearance. To complete the process of cell division, the cytoplasm cleaves by a process known as **cytokinesis.**

There is an important difference between a cell entering mitosis and one that has just completed the process. A cell in G_2 has a fully replicated genome (i.e., a 4n complement of DNA), and each chromosome consists of a pair of sister chromatids. In contrast, after mitosis, the chromosomes of each daughter cell have only one copy of the genome. This copy will not be duplicated until a daughter cell in its turn reaches the S phase of the next cell cycle (see Fig. 2-8). The entire process of mitosis thus ensures the orderly duplication and distribution of the genome through successive cell divisions.

The Human Karyotype

The condensed chromosomes of a dividing human cell are most readily analyzed at metaphase or prometaphase. At these stages, the chromosomes are visible under the microscope as a so-called **chromosome spread;** each chromosome consists of its sister chromatids, although in most chromosome preparations, the two chromatids are held together so tightly that they are rarely visible as separate entities.

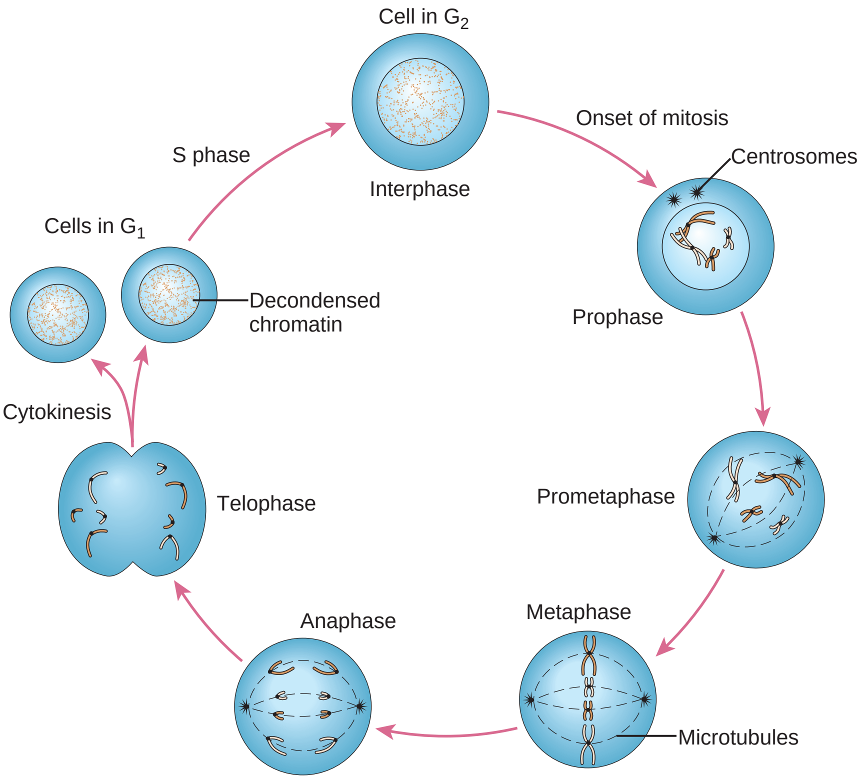

Figure 2-9 **Mitosis.** Only two chromosome pairs are shown. For details, see text.

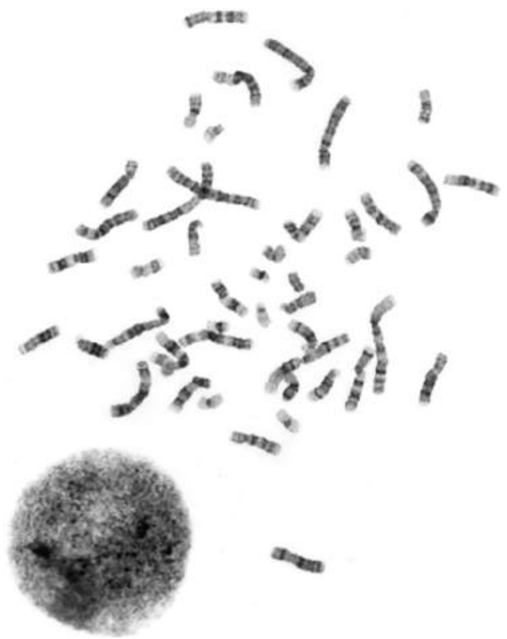

Figure 2-10 A chromosome spread prepared from a lymphocyte culture that has been stained by the Giemsa-banding (G-banding) technique. The darkly stained nucleus adjacent to the chromosomes is from a different cell in interphase, when chromosomal material is diffuse throughout the nucleus. *See Sources & Acknowledgments.*

As stated earlier, there are 24 different types of human chromosome, each of which can be distinguished cytologically by a combination of overall length, location of the centromere, and sequence content, the latter reflected by various staining methods. The centromere is apparent as a **primary constriction,** a narrowing or pinching-in of the sister chromatids due to formation of the kinetochore. This is a recognizable cytogenetic landmark, dividing the chromosome into two **arms,** a short arm designated **p** (for *petit*) and a long arm designated **q.**

Figure 2-10 shows a prometaphase cell in which the chromosomes have been stained by the Giemsa-staining (**G-banding**) method (also see Chapter 5). Each chromosome pair stains in a characteristic pattern of alternating light and dark bands (G bands) that correlates roughly with features of the underlying DNA sequence, such as base composition (i.e., the percentage of base pairs that are GC or AT) and the distribution of repetitive DNA elements. With such banding techniques, all of the chromosomes can be individually distinguished, and the nature of many structural or numerical abnormalities can be determined, as we examine in greater detail in Chapters 5 and 6.

Although experts can often analyze metaphase chromosomes directly under the microscope, a common procedure is to cut out the chromosomes from a digital image or photomicrograph and arrange them in pairs in a standard classification (Fig. 2-11). The completed picture is called a **karyotype.** The word *karyotype* is also used to refer to the standard chromosome set of an individual ("a normal male karyotype") or of a species ("the human karyotype") and, as a verb, to the process of preparing such a standard figure ("to karyotype").

Unlike the chromosomes seen in stained preparations under the microscope or in photographs, the chromosomes of living cells are fluid and dynamic structures. During mitosis, the chromatin of each interphase chromosome condenses substantially (Fig. 2-12). When maximally condensed at metaphase, DNA in chromosomes is approximately 1/10,000 of its fully extended state. When chromosomes are prepared to reveal bands (as in Figs. 2-10 and 2-11), as many as 1000 or more bands can be recognized in stained preparations of all the chromosomes. Each cytogenetic band therefore contains as many as 50 or more genes, although the density of genes in the genome, as mentioned previously, is variable.

Meiosis

Meiosis, the process by which diploid cells give rise to haploid gametes, involves a type of cell division that is unique to germ cells. In contrast to mitosis, meiosis consists of one round of DNA replication followed by *two* rounds of chromosome segregation and cell division (see meiosis I and meiosis II in Fig. 2-13). As outlined here and illustrated in Figure 2-14, the overall sequence of events in male and female meiosis is the same; however, the timing of gametogenesis is very different in the two sexes, as we will describe more fully later in this chapter.

Meiosis I is also known as the **reduction division** because it is the division in which the chromosome number is reduced by half through the pairing of homologues in prophase and by their segregation to different cells at anaphase of meiosis I. Meiosis I is also notable because it is the stage at which genetic **recombination** (also called **meiotic crossing over**) occurs. In this process, as shown for one pair of chromosomes in Figure 2-14, homologous segments of DNA are exchanged between nonsister chromatids of each pair of homologous chromosomes, thus ensuring that none of the gametes produced by meiosis will be identical to another. The conceptual and practical consequences of recombination for many aspects of human genetics and genomics are substantial and are outlined in the Box at the end of this section.

Prophase of meiosis I differs in a number of ways from mitotic prophase, with important genetic consequences, because homologous chromosomes need to pair and exchange genetic information. The most critical early stage is called **zygotene,** when homologous chromosomes begin to align along their entire length. The process of meiotic pairing—called **synapsis**—is normally precise, bringing corresponding DNA sequences into alignment along the length of the entire chromosome pair. The paired homologues—now called **bivalents**—are held together by a ribbon-like proteinaceous structure

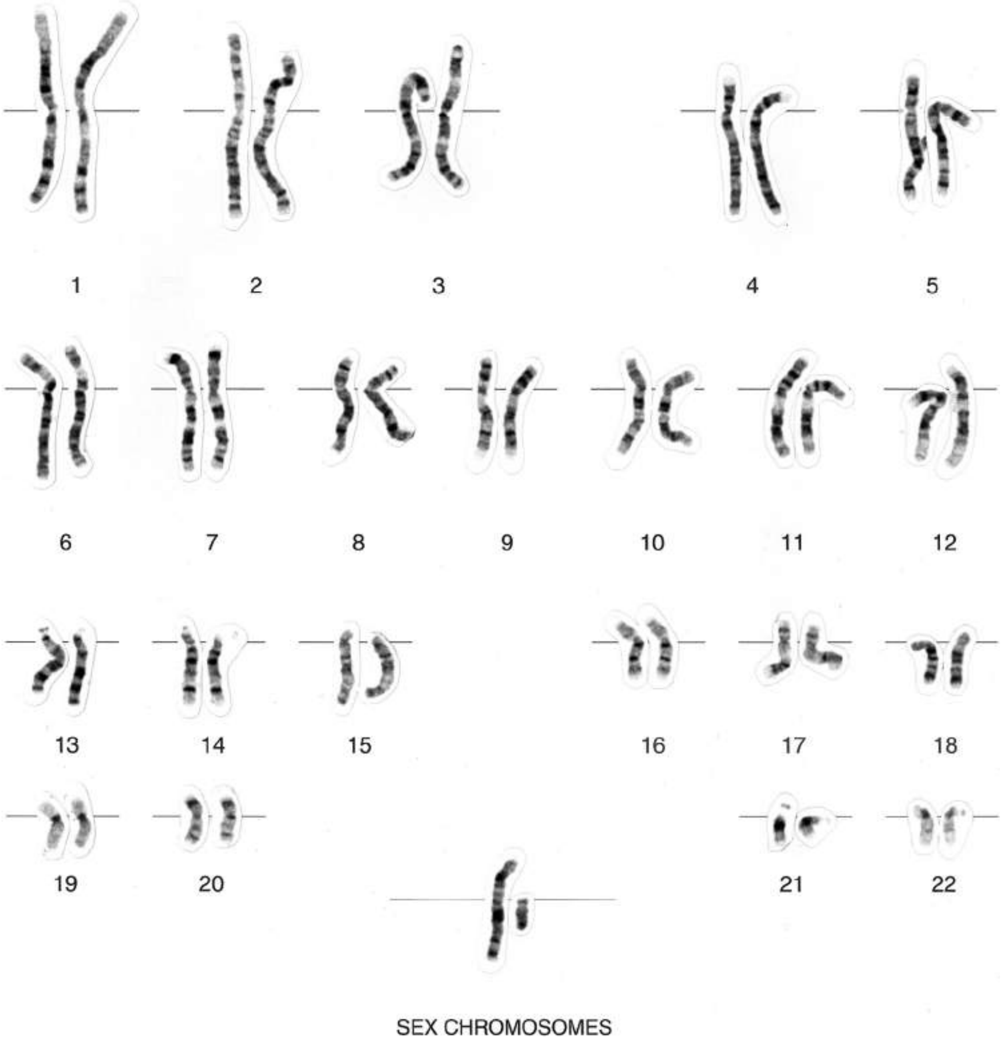

Figure 2-11 **A human male karyotype with Giemsa banding (G banding).** The chromosomes are at the prometaphase stage of mitosis and are arranged in a standard classification, numbered 1 to 22 in order of length, with the X and Y chromosomes shown separately. *See Sources & Acknowledgments.*

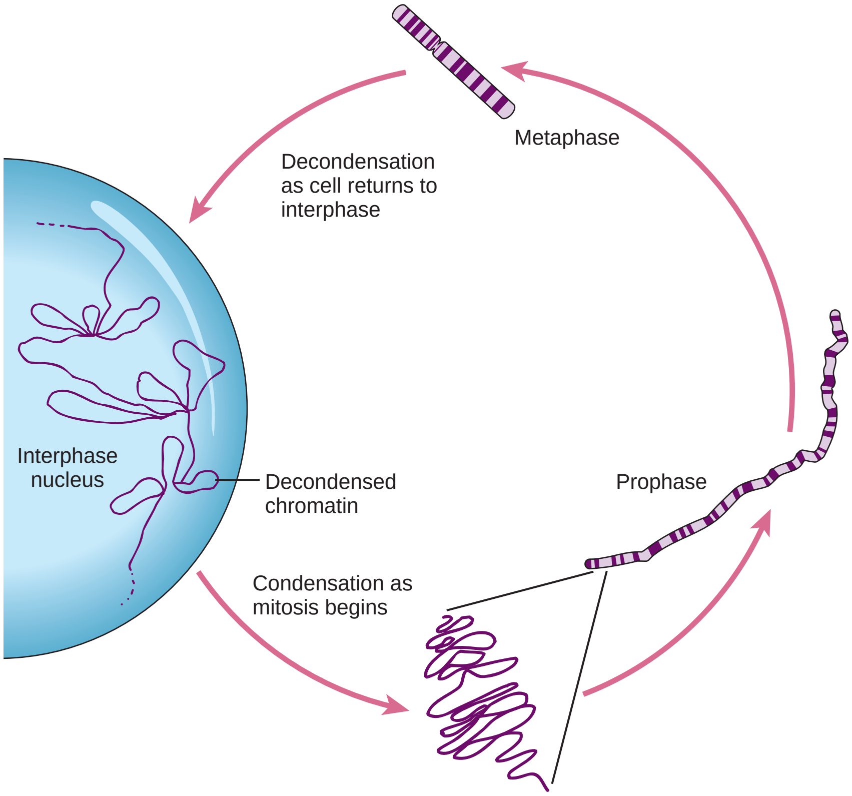
Metaphase
Decondensation as cell returns to interphase
Interphase nucleus
Decondensed chromatin
Prophase
Condensation as mitosis begins

Figure 2-12 Cycle of condensation and decondensation as a chromosome proceeds through the cell cycle.

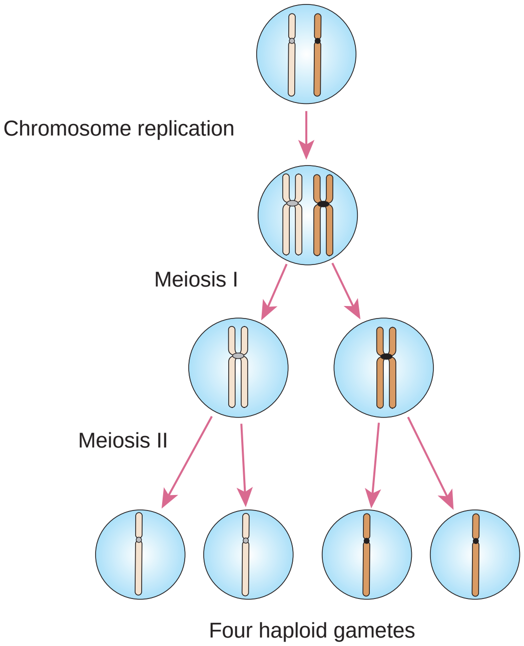

Figure 2-13 A simplified representation of the essential steps in meiosis, consisting of one round of DNA replication followed by two rounds of chromosome segregation, meiosis I and meiosis II.

called the **synaptonemal complex,** which is essential to the process of recombination. After synapsis is complete, meiotic crossing over takes place during **pachytene,** after which the synaptonemal complex breaks down.

Metaphase I begins, as in mitosis, when the nuclear membrane disappears. A spindle forms, and the paired chromosomes align themselves on the equatorial plane with their centromeres oriented toward different poles (see Fig. 2-14).

Anaphase of meiosis I again differs substantially from the corresponding stage of mitosis. Here, it is the two members of each bivalent that move apart, not the sister chromatids (contrast Fig. 2-14 with Fig. 2-9). The homologous centromeres (with their attached sister

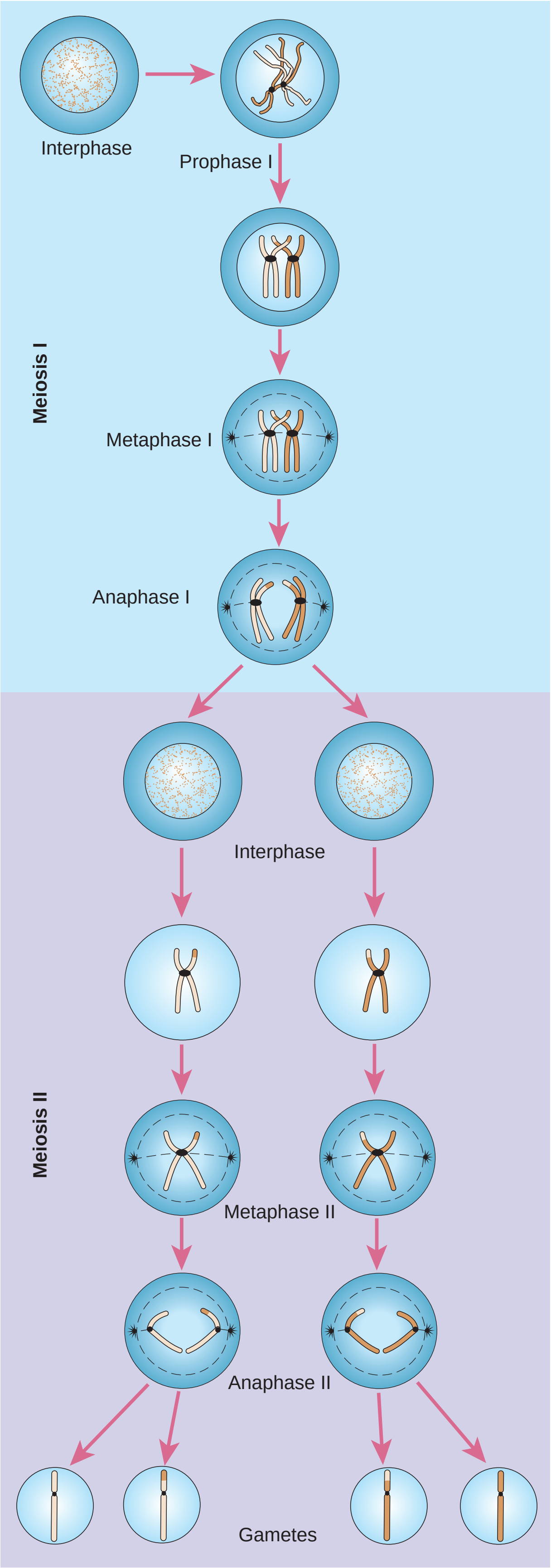

Figure 2-14 **Meiosis and its consequences.** A single chromosome pair and a single crossover are shown, leading to formation of four distinct gametes. The chromosomes replicate during interphase and begin to condense as the cell enters prophase of meiosis I. In meiosis I, the chromosomes synapse and recombine. A crossover is visible as the homologues align at metaphase I, with the centromeres oriented toward opposite poles. In anaphase I, the exchange of DNA between the homologues is apparent as the chromosomes are pulled to opposite poles. After completion of meiosis I and cytokinesis, meiosis II proceeds with a mitosis-like division. The sister kinetochores separate and move to opposite poles in anaphase II, yielding four haploid products.

chromatids) are drawn to opposite poles of the cell, a process termed **disjunction.** Thus the chromosome number is halved, and each cellular product of meiosis I has the haploid chromosome number. The 23 pairs of homologous chromosomes assort independently of one another, and as a result, the original paternal and maternal chromosome sets are sorted into random combinations. The possible number of combinations of the 23 chromosome pairs that can be present in the gametes is 2^{23} (more than 8 million). Owing to the process of crossing over, however, the variation in the genetic material that is transmitted from parent to child is actually much greater than this. As a result, each chromatid typically contains segments derived from each member of the original parental chromosome pair, as illustrated schematically in Figure 2-14. For example, at this stage, a typical large human chromosome would be composed of three to five segments, alternately paternal and maternal in origin, as inferred from DNA sequence variants that distinguish the respective parental genomes (Fig. 2-15).

After telophase of meiosis I, the two haploid daughter cells enter meiotic interphase. In contrast to mitosis, this interphase is brief, and meiosis II begins. The notable point that distinguishes meiotic and mitotic interphase is that there is no S phase (i.e., no DNA

GENETIC CONSEQUENCES AND MEDICAL RELEVANCE OF HOMOLOGOUS RECOMBINATION

The take-home lesson of this portion of the chapter is a simple one: **the genetic content of each gamete is unique,** because of random assortment of the parental chromosomes to shuffle the combination of sequence variants *between* chromosomes and because of homologous recombination to shuffle the combination of sequence variants *within* each and every chromosome. This has significant consequences for patterns of genomic variation among and between different populations around the globe and for diagnosis and counseling of many common conditions with complex patterns of inheritance (see Chapters 8 and 10).

The **amounts and patterns of meiotic recombination** are determined by sequence variants in specific genes and at specific "hot spots" and differ between individuals, between the sexes, between families, and between populations (see Chapter 10).

Because recombination involves the physical intertwining of the two homologues until the appropriate point during meiosis I, it is also critical for ensuring proper chromosome segregation during meiosis. Failure to recombine properly can lead to **chromosome missegregation (nondisjunction)** in meiosis I and is a frequent cause of pregnancy loss and of chromosome abnormalities like Down syndrome (see Chapters 5 and 6).

Major ongoing efforts to **identify genes and their variants responsible for various medical conditions** rely on tracking the inheritance of millions of sequence differences within families or the sharing of variants within groups of even unrelated individuals affected with a particular condition. The utility of this approach, which has uncovered thousands of gene-disease associations to date, depends on patterns of homologous recombination in meiosis (see Chapter 10).

Although homologous recombination is normally precise, areas of repetitive DNA in the genome and genes of variable copy number in the population are prone to occasional **unequal crossing over** during meiosis, leading to variations in clinically relevant traits such as drug response, to common disorders such as the thalassemias or autism, or to abnormalities of sexual differentiation (see Chapters 6, 8, and 11).

Although homologous recombination is a normal and essential part of meiosis, it also occurs, albeit more rarely, in somatic cells. Anomalies in **somatic recombination** are one of the causes of **genome instability in cancer** (see Chapter 15).

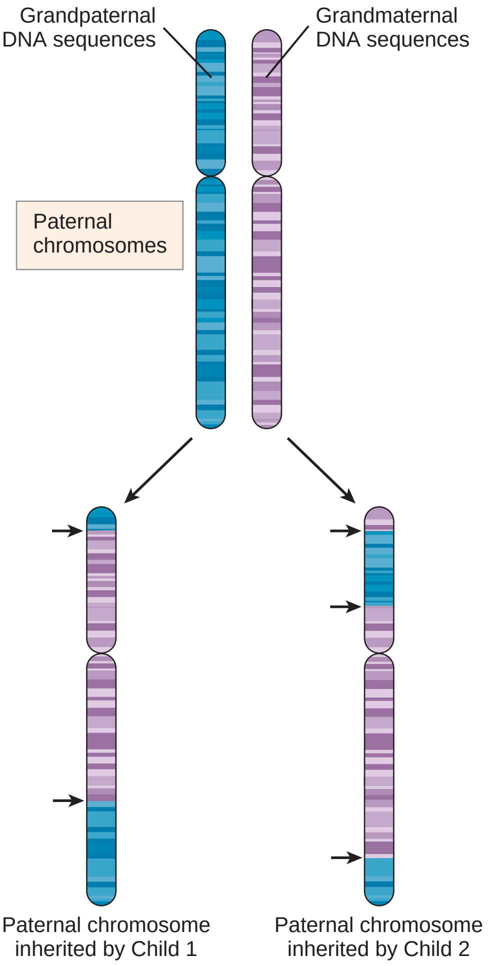
Grandpaternal DNA sequences
Grandmaternal DNA sequences
Paternal chromosomes
Paternal chromosome inherited by Child 1
Paternal chromosome inherited by Child 2

Figure 2-15 The effect of homologous recombination in meiosis. In this example, representing the inheritance of sequences on a typical large chromosome, an individual has distinctive homologues, one containing sequences inherited from his father (*blue*) and one containing homologous sequences from his mother (*purple*). After meiosis in spermatogenesis, he transmits a single complete copy of that chromosome to his two offspring. However, as a result of crossing over (*arrows*), the copy he transmits to each child consists of alternating segments of the two grandparental sequences. Child 1 inherits a copy after two crossovers, whereas child 2 inherits a copy with three crossovers.

synthesis and duplication of the genome) between the first and second meiotic divisions.

Meiosis II is similar to an ordinary mitosis, except that the chromosome number is 23 instead of 46; the chromatids of each of the 23 chromosomes separate, and one chromatid of each chromosome passes to each daughter cell (see Fig. 2-14). However, as mentioned earlier, because of crossing over in meiosis I, the chromosomes of the resulting gametes are not identical (see Fig. 2-15).

HUMAN GAMETOGENESIS AND FERTILIZATION

The cells in the germline that undergo meiosis, primary spermatocytes or primary oocytes, are derived from the zygote by a long series of mitoses before the onset of meiosis. Male and female gametes have different histories, marked by different patterns of gene expression that reflect their developmental origin as an XY or XX embryo. The human primordial germ cells are recognizable by the fourth week of development outside the embryo proper, in the endoderm of the yolk sac. From there, they migrate during the sixth week to the genital ridges and associate with somatic cells to form the primitive gonads, which soon differentiate into testes or ovaries, depending on the cells' sex chromosome constitution (XY or XX), as we examine in greater detail in Chapter 6. Both spermatogenesis and oogenesis require meiosis but have important differences in detail and timing that may have clinical and genetic consequences for the offspring. Female meiosis is initiated once, early during fetal life, in a limited number of cells. In contrast, male meiosis is initiated continuously in many cells from a dividing cell population throughout the adult life of a male.

In the female, successive stages of meiosis take place over several decades—in the fetal ovary before the female in question is even born, in the oocyte near the time of ovulation in the sexually mature female, and after fertilization of the egg that can become that female's offspring. Although postfertilization stages can be studied in vitro, access to the earlier stages is limited. Testicular material for the study of male meiosis is less difficult to obtain, inasmuch as testicular biopsy is included in the assessment of many men attending infertility clinics. Much remains to be learned about the cytogenetic, biochemical, and molecular mechanisms involved in normal meiosis and about the causes and consequences of meiotic irregularities.

Spermatogenesis

The stages of spermatogenesis are shown in Figure 2-16. The seminiferous tubules of the testes are lined with **spermatogonia,** which develop from the primordial

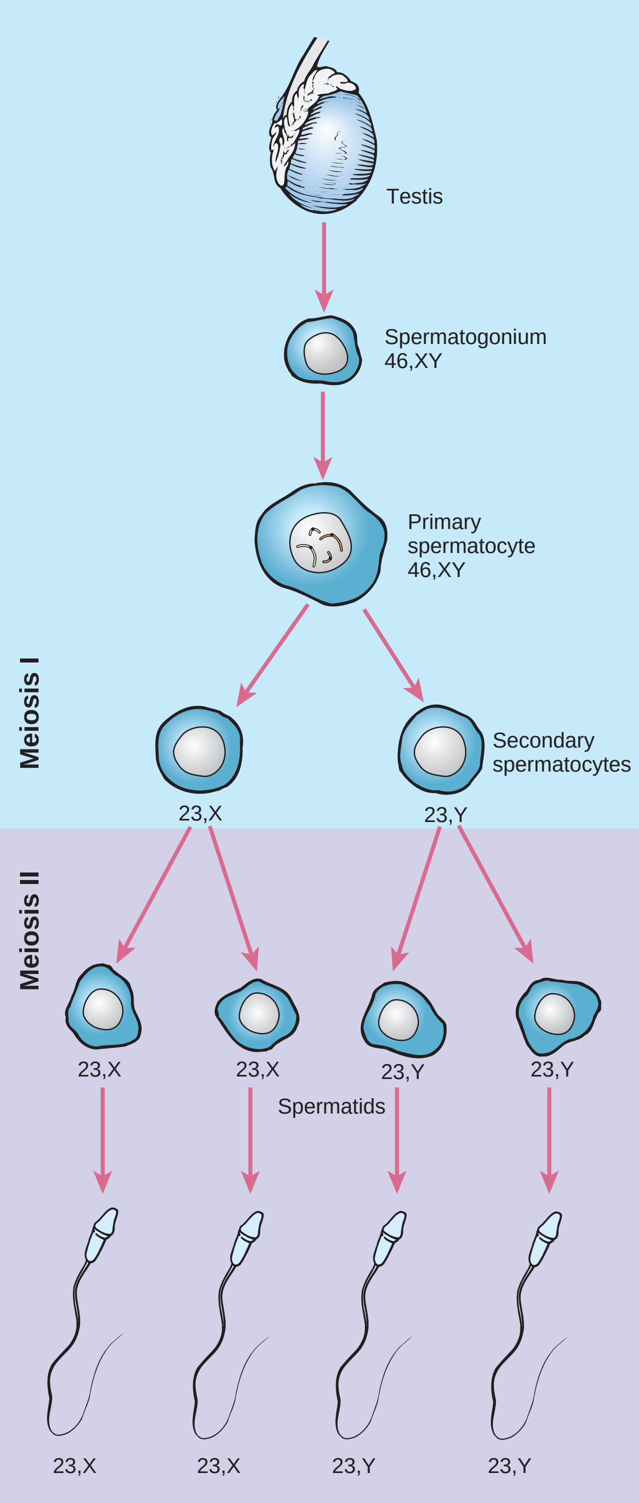

Figure 2-16 Human spermatogenesis in relation to the two meiotic divisions. The sequence of events begins at puberty and takes approximately 64 days to be completed. The chromosome number (46 or 23) and the sex chromosome constitution (X or Y) of each cell are shown. *See Sources & Acknowledgments.*

germ cells by a long series of mitoses and which are in different stages of differentiation. **Sperm** (spermatozoa) are formed only after sexual maturity is reached. The last cell type in the developmental sequence is the **primary spermatocyte,** a diploid germ cell that undergoes meiosis I to form two haploid **secondary spermatocytes**. Secondary spermatocytes rapidly enter meiosis II, each forming two **spermatids,** which differentiate without further division into sperm. In humans, the entire process takes approximately 64 days. The enormous number of sperm produced, typically approximately 200 million per ejaculate and an estimated 10^{12} in a lifetime, requires several hundred successive mitoses.

As discussed earlier, normal meiosis requires pairing of homologous chromosomes followed by recombination. The autosomes and the X chromosomes in females present no unusual difficulties in this regard; but what of the X and Y chromosomes during spermatogenesis? Although the X and Y chromosomes are different and are not homologues in a strict sense, they do have relatively short identical segments at the ends of their respective short arms (Xp and Yp) and long arms (Xq and Yq) (see Chapter 6). Pairing and crossing over occurs in both regions during meiosis I. These homologous segments are called **pseudoautosomal** to reflect their autosome-like pairing and recombination behavior, despite being on different sex chromosomes.

Oogenesis

Whereas spermatogenesis is initiated only at the time of puberty, oogenesis begins during a female's development as a fetus (Fig. 2-17). The **ova** develop from **oogonia,** cells in the ovarian cortex that have descended from the primordial germ cells by a series of approximately 20 mitoses. Each oogonium is the central cell in a developing follicle. By approximately the third month of fetal development, the oogonia of the embryo have begun to develop into **primary oocytes,** most of which have already entered prophase of meiosis I. The process of oogenesis is not synchronized, and both early and late stages coexist in the fetal ovary. Although there are several million oocytes at the time of birth, most of these degenerate; the others remain arrested in prophase I (see Fig. 2-14) for decades. Only approximately 400 eventually mature and are ovulated as part of a woman's menstrual cycle.

After a woman reaches sexual maturity, individual follicles begin to grow and mature, and a few (on average one per month) are ovulated. Just before ovulation, the oocyte rapidly completes meiosis I, dividing in such a way that one cell becomes the secondary oocyte (an egg or **ovum**), containing most of the cytoplasm with its organelles; the other cell becomes the first polar body (see Fig. 2-17). Meiosis II begins promptly and proceeds to the metaphase stage during ovulation, where it halts again, only to be completed if fertilization occurs.

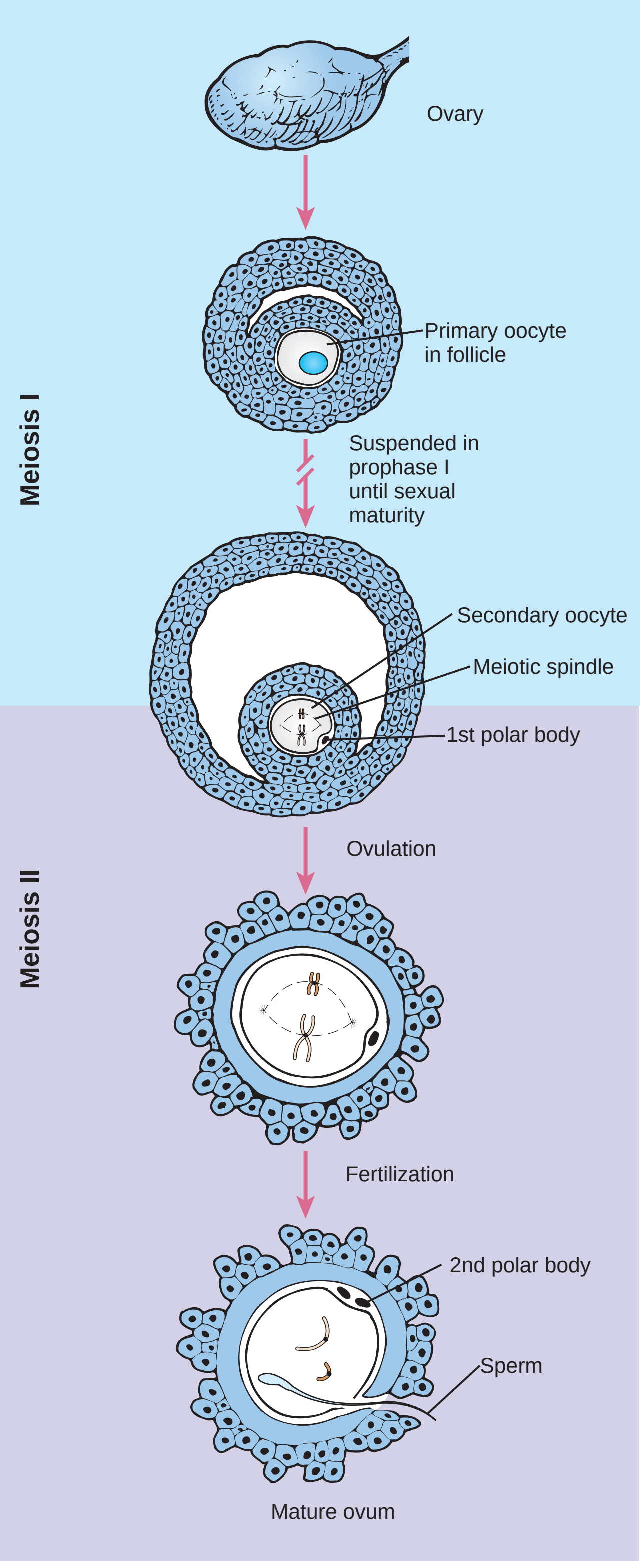

Figure 2-17 Human oogenesis and fertilization in relation to the two meiotic divisions. The primary oocytes are formed prenatally and remain suspended in prophase of meiosis I for years until the onset of puberty. An oocyte completes meiosis I as its follicle matures, resulting in a secondary oocyte and the first polar body. After ovulation, each oocyte continues to metaphase of meiosis II. Meiosis II is completed only if fertilization occurs, resulting in a fertilized mature ovum and the second polar body.

Fertilization

Fertilization of the egg usually takes place in the fallopian tube within a day or so of ovulation. Although many sperm may be present, the penetration of a single sperm into the ovum sets up a series of biochemical events that usually prevent the entry of other sperm.

Fertilization is followed by the completion of meiosis II, with the formation of the second polar body (see Fig. 2-17). The chromosomes of the now-fertilized egg and sperm form **pronuclei,** each surrounded by its own nuclear membrane. It is only upon replication of the parental genomes after fertilization that the two haploid genomes become one diploid genome within a shared nucleus. The diploid **zygote** divides by mitosis to form two diploid daughter cells, the first in the series of cell divisions that initiate the process of embryonic development (see Chapter 14).

Although development begins at the time of conception, with the formation of the zygote, in clinical medicine the stage and duration of pregnancy are usually measured as the "menstrual age," dating from the beginning of the mother's last menstrual period, typically approximately 14 days before conception.

MEDICAL RELEVANCE OF MITOSIS AND MEIOSIS

The biological significance of mitosis and meiosis lies in ensuring the constancy of chromosome number—and thus the integrity of the genome—from one cell to its progeny and from one generation to the next. The medical relevance of these processes lies in errors of one or the other mechanism of cell division, leading to the formation of an individual or of a cell lineage with an abnormal number of chromosomes and thus an abnormal dosage of genomic material.

As we see in detail in Chapter 5, meiotic nondisjunction, particularly in oogenesis, is the most common mutational mechanism in our species, responsible for chromosomally abnormal fetuses in at least several percent of all recognized pregnancies. Among pregnancies that survive to term, chromosome abnormalities are a leading cause of developmental defects, failure to thrive in the newborn period, and intellectual disability.

Mitotic nondisjunction in somatic cells also contributes to genetic disease. Nondisjunction soon after fertilization, either in the developing embryo or in extraembryonic tissues like the placenta, leads to chromosomal mosaicism that can underlie some medical conditions, such as a proportion of patients with Down syndrome. Further, abnormal chromosome segregation in rapidly dividing tissues, such as in cells of the colon, is frequently a step in the development of chromosomally abnormal tumors, and thus evaluation of chromosome and genome balance is an important diagnostic and prognostic test in many cancers.

GENERAL REFERENCES

Green ED, Guyer MS, National Human Genome Research Institute: Charting a course for genomic medicine from base pairs to bedside, *Nature* 470:204–213, 2011.

Lander ES: Initial impact of the sequencing of the human genome, *Nature* 470:187–197, 2011.

Moore KL, Presaud TVN, Torchia MG: *The developing human: clinically oriented embryology*, ed 9, Philadelphia, 2013, WB Saunders.

REFERENCES FOR SPECIFIC TOPICS

Deininger P: Alu elements: know the SINES, *Genome Biol* 12:236, 2011.

Frazer KA: Decoding the human genome, *Genome Res* 22:1599–1601, 2012.

International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome, *Nature* 409:860–921, 2001.

International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome, *Nature* 431:931–945, 2004.

Venter J, Adams M, Myers E, et al: The sequence of the human genome, *Science* 291:1304–1351, 2001.

PROBLEMS

1. At a certain locus, a person has two alleles, *A* and *a*.
   a. What alleles will be present in this person's gametes?
   b. When do *A* and *a* segregate (1) if there is no crossing over between the locus and the centromere of the chromosome? (2) if there is a single crossover between the locus and the centromere?

2. What is the main cause of numerical chromosome abnormalities in humans?

3. Disregarding crossing over, which increases the amount of genetic variability, estimate the probability that all your chromosomes have come to you from your father's mother and your mother's mother. Would you be male or female?

4. A chromosome entering meiosis is composed of two sister chromatids, each of which is a single DNA molecule.
   a. In our species, at the end of meiosis I, how many chromosomes are there per cell? How many chromatids?
   b. At the end of meiosis II, how many chromosomes are there per cell? How many chromatids?
   c. When is the diploid chromosome number restored? When is the two-chromatid structure of a typical metaphase chromosome restored?

5. From Figure 2-7, estimate the number of genes per million base pairs on chromosomes 1, 13, 18, 19, 21, and 22. Would a chromosome abnormality of equal size on chromosome 18 or 19 be expected to have greater clinical impact? On chromosome 21 or 22?

CHAPTER 3

The Human Genome: Gene Structure and Function

Over the past three decades, remarkable progress has been made in our understanding of the structure and function of genes and chromosomes. These advances have been aided by the applications of molecular genetics and genomics to many clinical problems, thereby providing the tools for a distinctive new approach to medical genetics. In this chapter, we present an overview of gene structure and function and the aspects of molecular genetics required for an understanding of the genetic and genomic approach to medicine. To supplement the information discussed here and in subsequent chapters, we provide additional material online to detail many of the experimental approaches of modern genetics and genomics that are becoming critical to the practice and understanding of human and medical genetics.

The increased knowledge of genes and of their organization in the genome has had an enormous impact on medicine and on our perception of human physiology. As 1980 Nobel laureate Paul Berg stated presciently at the dawn of this new era:

> *Just as our present knowledge and practice of medicine relies on a sophisticated knowledge of human anatomy, physiology, and biochemistry, so will dealing with disease in the future demand a detailed understanding of the molecular anatomy, physiology, and biochemistry of the human genome.... We shall need a more detailed knowledge of how human genes are organized and how they function and are regulated. We shall also have to have physicians who are as conversant with the molecular anatomy and physiology of chromosomes and genes as the cardiac surgeon is with the structure and workings of the heart.*

INFORMATION CONTENT OF THE HUMAN GENOME

How does the 3-billion-letter digital code of the human genome guide the intricacies of human anatomy, physiology, and biochemistry to which Berg referred? The answer lies in the enormous amplification and integration of information content that occurs as one moves from genes in the genome to their products in the cell and to the observable expression of that genetic information as cellular, morphological, clinical, or biochemical traits—what is termed the **phenotype** of the individual. This hierarchical expansion of information from the genome to phenotype includes a wide range of structural and regulatory RNA products, as well as protein products that orchestrate the many functions of cells, organs, and the entire organism, in addition to their interactions with the environment. Even with the essentially complete sequence of the human genome in hand, we still do not know the precise number of genes in the genome. Current estimates are that the genome contains approximately 20,000 **protein-coding genes** (see Box in Chapter 2), but this figure only begins to hint at the levels of complexity that emerge from the decoding of this digital information (Fig. 3-1).

As introduced briefly in Chapter 2, the product of protein-coding genes is a protein whose structure ultimately determines its particular functions in the cell. But if there were a simple one-to-one correspondence between genes and proteins, we could have at most approximately 20,000 different proteins. This number seems insufficient to account for the vast array of functions that occur in human cells over the life span. The answer to this dilemma is found in two features of gene structure and function. First, many genes are capable of generating multiple different products, not just one (see Fig. 3-1). This process, discussed later in this chapter, is accomplished through the use of alternative coding segments in genes and through the subsequent biochemical modification of the encoded protein; these two features of complex genomes result in a substantial amplification of information content. Indeed, it has been estimated that in this way, these 20,000 human genes can encode many hundreds of thousands of different proteins, collectively referred to as the **proteome**. Second, individual proteins do not function by themselves. They form elaborate networks, involving many different proteins and regulatory RNAs that respond in a coordinated and integrated fashion to many different genetic, developmental, or environmental signals. The combinatorial nature of protein networks results in an even greater diversity of possible cellular functions.

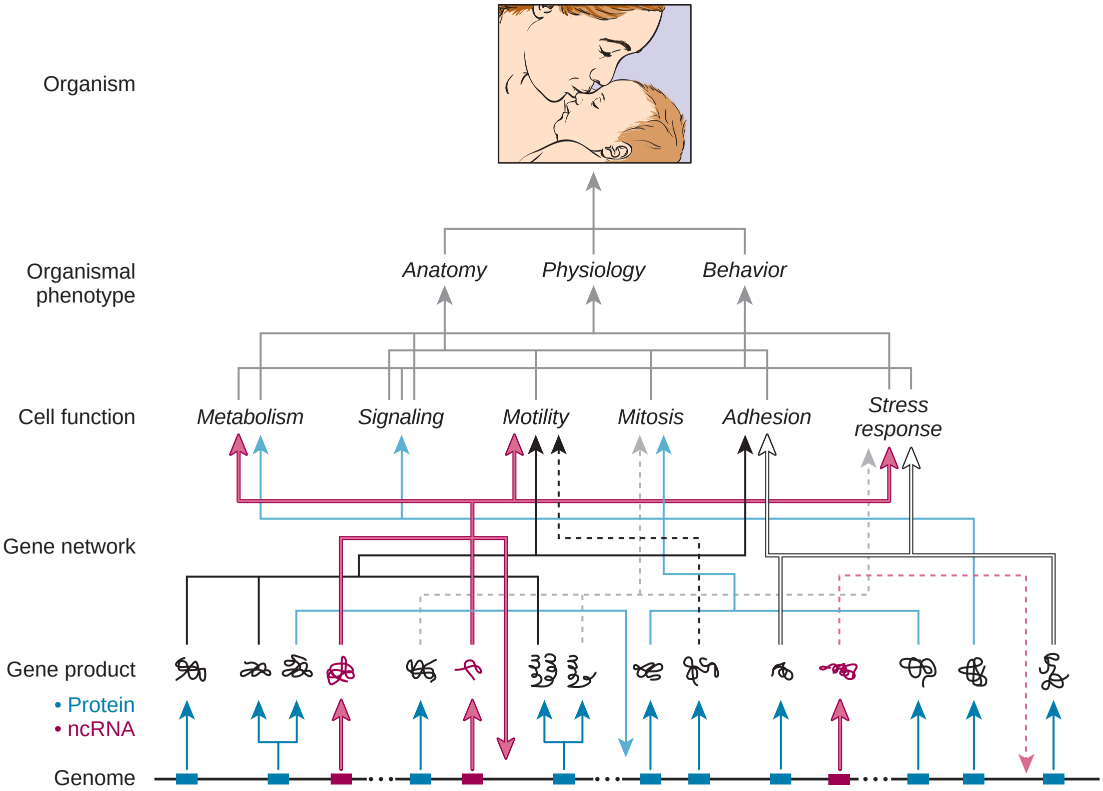

Figure 3-1 The amplification of genetic information from genome to gene products to gene networks and ultimately to cellular function and phenotype. The genome contains both protein-coding genes (*blue*) and noncoding RNA (ncRNA) genes (*red*). Many genes in the genome use alternative coding information to generate multiple different products. Both small and large ncRNAs participate in gene regulation. Many proteins participate in multigene networks that respond to cellular signals in a coordinated and combinatorial manner, thus further expanding the range of cellular functions that underlie organismal phenotypes.

Genes are located throughout the genome but tend to cluster in particular regions on particular chromosomes and to be relatively sparse in other regions or on other chromosomes. For example, chromosome 11, an approximately 135 million-bp (megabase pairs [Mb]) chromosome, is relatively gene-rich with approximately 1300 protein-coding genes (see Fig. 2-7). These genes are not distributed randomly along the chromosome, and their localization is particularly enriched in two chromosomal regions with gene density as high as one gene every 10 kb (Fig. 3-2). Some of the genes belong to families of related genes, as we will describe more fully later in this chapter. Other regions are gene-poor, and there are several so-called gene deserts of a million base pairs or more without any known protein-coding genes. Two caveats here: first, the process of gene identification and genome annotation remains very much an ongoing challenge; despite the apparent robustness of recent estimates, it is virtually certain that there are some genes, including clinically relevant genes, that are currently undetected or that display characteristics that we do not currently recognize as being associated with genes. And second, as mentioned in Chapter 2, many genes are not protein-coding; their products are functional RNA molecules (**noncoding RNAs** or ncRNAs; see Fig. 3-1) that play a variety of roles in the cell, many of which are only just being uncovered.

For genes located on the autosomes, there are two copies of each gene, one on the chromosome inherited from the mother and one on the chromosome inherited from the father. For most autosomal genes, both copies are expressed and generate a product. There are, however, a growing number of genes in the genome that are exceptions to this general rule and are expressed at characteristically different levels from the two copies, including some that, at the extreme, are expressed from only one of the two homologues. These examples of **allelic imbalance** are discussed in greater detail later in this chapter, as well as in Chapters 6 and 7.

THE CENTRAL DOGMA: DNA → RNA → PROTEIN

How does the genome specify the functional complexity and diversity evident in Figure 3-1? As we saw in the previous chapter, genetic information is contained in DNA in the chromosomes within the cell nucleus. However, protein synthesis, the process through which

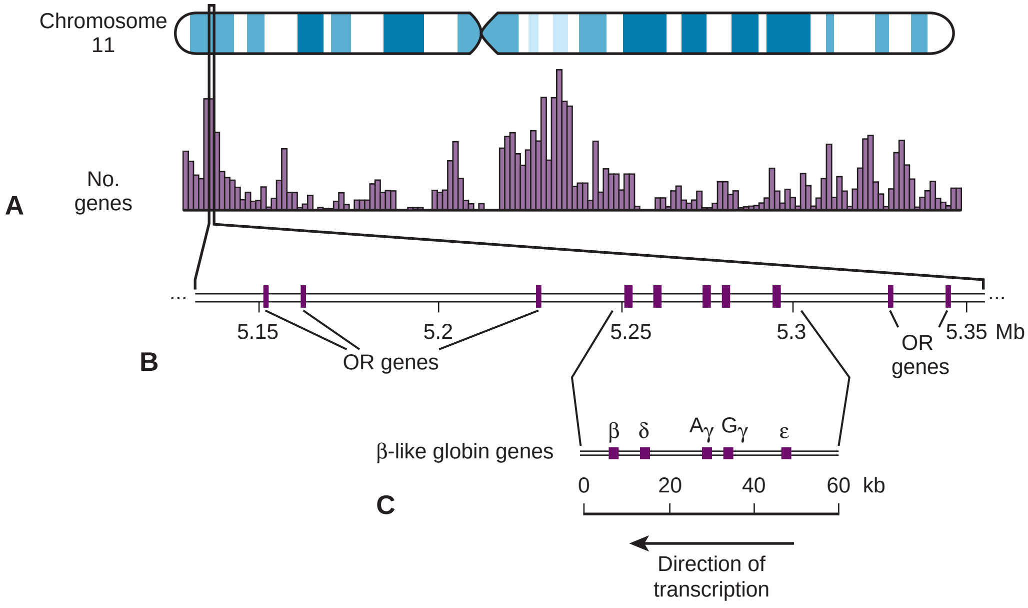

Figure 3-2 **Gene content on chromosome 11, which consists of 135 Mb of DNA. A,** The distribution of genes is indicated along the chromosome and is high in two regions of the chromosome and low in other regions. **B,** An expanded region from 5.15 to 5.35 Mb (measured from the short-arm telomere), which contains 10 known protein-coding genes, five belonging to the olfactory receptor (OR) gene family and five belonging to the globin gene family. **C,** The five β-like globin genes expanded further. *See Sources & Acknowledgments.*

information encoded in the genome is actually used to specify cellular functions, takes place in the cytoplasm. This compartmentalization reflects the fact that the human organism is a **eukaryote**. This means that human cells have a nucleus containing the genome, which is separated by a nuclear membrane from the cytoplasm. In contrast, in prokaryotes like the intestinal bacterium *Escherichia coli,* DNA is not enclosed within a nucleus. Because of the compartmentalization of eukaryotic cells, information transfer from the nucleus to the cytoplasm is a complex process that has been a focus of much attention among molecular and cellular biologists.

The molecular link between these two related types of information—the DNA code of genes and the amino acid code of protein—is **ribonucleic acid** (**RNA**). The chemical structure of RNA is similar to that of DNA, except that each nucleotide in RNA has a ribose sugar component instead of a deoxyribose; in addition, uracil (U) replaces thymine as one of the pyrimidine bases of RNA (Fig. 3-3). An additional difference between RNA and DNA is that RNA in most organisms exists as a single-stranded molecule, whereas DNA, as we saw in Chapter 2, exists as a double helix.

The informational relationships among DNA, RNA, and protein are intertwined: genomic DNA directs the synthesis and sequence of RNA, RNA directs the synthesis and sequence of polypeptides, and specific proteins are involved in the synthesis and metabolism of DNA and RNA. This flow of information is referred to as the **central dogma** of molecular biology.

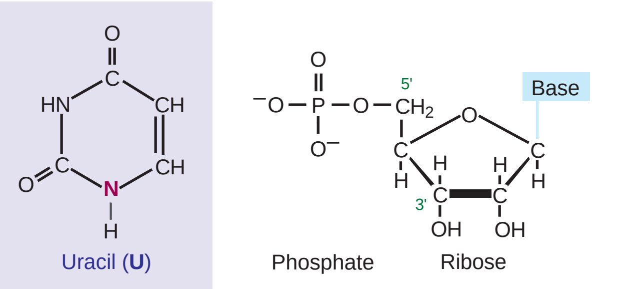

Figure 3-3 **The pyrimidine uracil and the structure of a nucleotide in RNA.** Note that the sugar ribose replaces the sugar deoxyribose of DNA. Compare with Figure 2-2.

Genetic information is stored in the DNA of the genome by means of a code (the **genetic code,** discussed later) in which the sequence of adjacent bases ultimately determines the sequence of amino acids in the encoded polypeptide. First, RNA is synthesized from the DNA template through a process known as **transcription**. The RNA, carrying the coded information in a form called **messenger RNA** (**mRNA**)**,** is then transported from the nucleus to the cytoplasm, where the RNA sequence is decoded, or translated, to determine the sequence of amino acids in the protein being synthesized. The process of **translation** occurs on **ribosomes,** which are cytoplasmic organelles with binding sites for all of the interacting molecules, including the mRNA, involved in protein synthesis. Ribosomes are themselves made up of many different structural proteins in association with specialized types of RNA known as **ribosomal RNA (rRNA)**. Translation involves yet a third type of RNA, **transfer RNA (tRNA),** which provides the molecular link between the code contained in the base sequence of each mRNA and the amino acid sequence of the protein encoded by that mRNA.

Because of the interdependent flow of information represented by the central dogma, one can begin discussion of the molecular genetics of gene expression at any of its three informational levels: DNA, RNA, or protein. We begin by examining the structure of genes in the genome as a foundation for discussion of the genetic code, transcription, and translation.

GENE ORGANIZATION AND STRUCTURE

In its simplest form, a protein-coding gene can be visualized as a segment of a DNA molecule containing the code for the amino acid sequence of a polypeptide chain and the regulatory sequences necessary for its expression. This description, however, is inadequate for genes in the human genome (and indeed in most eukaryotic genomes) because few genes exist as continuous coding sequences. Rather, in the majority of genes, the coding sequences are interrupted by one or more noncoding regions (Fig. 3-4). These intervening sequences, called **introns,** are initially transcribed into RNA in the

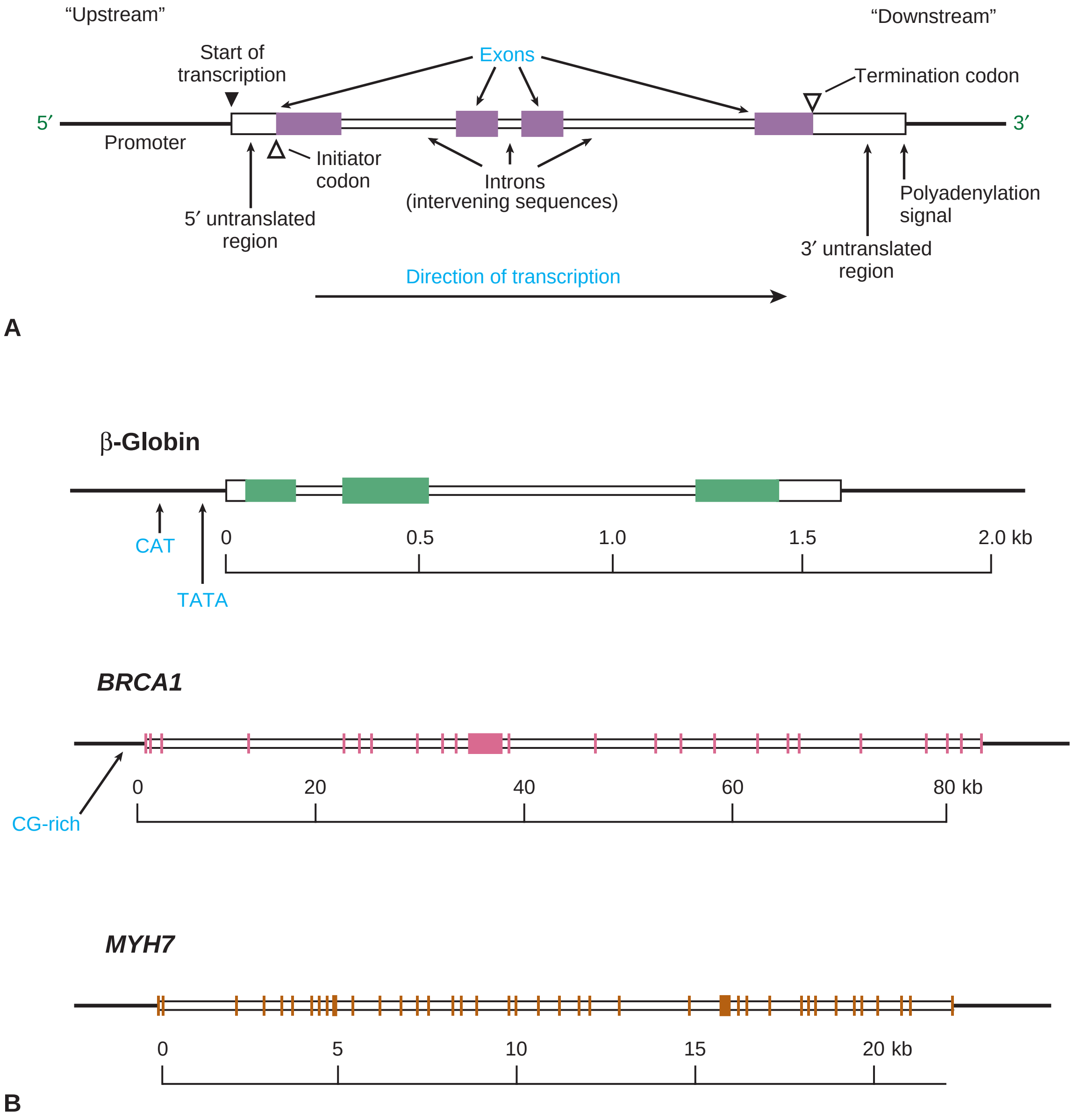

Figure 3-4 **A,** General structure of a typical human gene. Individual labeled features are discussed in the text. **B,** Examples of three medically important human genes. Different mutations in the β-globin gene, with three exons, cause a variety of important disorders of hemoglobin (Cases 42 and 44). Mutations in the *BRCA1* gene (24 exons) are responsible for many cases of inherited breast or breast and ovarian cancer (Case 7). Mutations in the β-myosin heavy chain (*MYH7*) gene (40 exons) lead to inherited hypertrophic cardiomyopathy.

nucleus but are not present in the mature mRNA in the cytoplasm, because they are removed ("spliced out") by a process we will discuss later. Thus information from the intronic sequences is not normally represented in the final protein product. Introns alternate with **exons,** the segments of genes that ultimately determine the amino acid sequence of the protein. In addition, the collection of coding exons in any particular gene is flanked by additional sequences that are transcribed but untranslated, called the 5′ and 3′ untranslated regions (see Fig. 3-4). Although a few genes in the human genome have no introns, most genes contain at least one and usually several introns. In many genes, the cumulative length of the introns makes up a far greater proportion of a gene's total length than do the exons. Whereas some genes are only a few kilobase pairs in length, others stretch on for hundreds of kilobase pairs. Also, few genes are exceptionally large; for example, the dystrophin gene on the X chromosome (mutations in which lead to Duchenne muscular dystrophy [Case 14]) spans more than 2 Mb, of which, remarkably, less than 1% consists of coding exons.

Structural Features of a Typical Human Gene

A range of features characterize human genes (see Fig. 3-4). In Chapters 1 and 2, we briefly defined *gene* in general terms. At this point, we can provide a molecular definition of a gene as *a sequence of DNA that specifies production of a functional product,* be it a polypeptide or a functional RNA molecule. A gene includes not only the actual coding sequences but also adjacent nucleotide sequences required for the proper expression of the gene—that is, for the production of normal mRNA or other RNA molecules in the correct amount, in the correct place, and at the correct time during development or during the cell cycle.

The adjacent nucleotide sequences provide the molecular "start" and "stop" signals for the synthesis of mRNA transcribed from the gene. Because the primary RNA transcript is synthesized in a 5′ to 3′ direction, the transcriptional start is referred to as the 5′ end of the transcribed portion of a gene (see Fig. 3-4). By convention, the genomic DNA that precedes the transcriptional start site in the 5′ direction is referred to as the "upstream" sequence, whereas DNA sequence located in the 3′ direction past the end of a gene is referred to as the "downstream" sequence. At the 5′ end of each gene lies a **promoter** region that includes sequences responsible for the proper initiation of transcription. Within this region are several DNA elements whose sequence is often conserved among many different genes; this conservation, together with functional studies of gene expression, indicates that these particular sequences play an important role in gene regulation. Only a subset of genes in the genome is expressed in any given tissue or at any given time during development. Several different types of promoter are found in the human genome, with different regulatory properties that specify the patterns as well as the levels of expression of a particular gene in different tissues and cell types, both during development and throughout the life span. Some of these properties are encoded in the genome, whereas others are specified by features of chromatin associated with those sequences, as discussed later in this chapter. Both promoters and other **regulatory elements** (located either 5′ or 3′ of a gene or in its introns) can be sites of mutation in genetic disease that can interfere with the normal expression of a gene. These regulatory elements, including **enhancers, insulators,** and **locus control regions,** are discussed more fully later in this chapter. Some of these elements lie a significant distance away from the coding portion of a gene, thus reinforcing the concept that the genomic environment in which a gene resides is an important feature of its evolution and regulation.

The 3′ untranslated region contains a signal for the addition of a sequence of adenosine residues (the so-called polyA tail) to the end of the mature RNA. Although it is generally accepted that such closely neighboring regulatory sequences are part of what is called a gene, the precise dimensions of any particular gene will remain somewhat uncertain until the potential functions of more distant sequences are fully characterized.

Gene Families

Many genes belong to gene families, which share closely related DNA sequences and encode polypeptides with closely related amino acid sequences.

Members of two such gene families are located within a small region on chromosome 11 (see Fig. 3-2) and illustrate a number of features that characterize gene families in general. One small and medically important gene family is composed of genes that encode the protein chains found in hemoglobins. The β-globin gene cluster on chromosome 11 and the related α-globin gene cluster on chromosome 16 are believed to have arisen by duplication of a primitive precursor gene approximately 500 million years ago. These two clusters contain multiple genes coding for closely related globin chains expressed at different developmental stages, from embryo to adult. Each cluster is believed to have evolved by a series of sequential gene duplication events within the past 100 million years. The exon-intron patterns of the functional globin genes have been remarkably conserved during evolution; each of the functional globin genes has two introns at similar locations (see the β-globin gene in Fig. 3-4), although the sequences contained within the introns have accumulated far more nucleotide base changes over time than have the coding sequences of each gene. The control of expression of the various globin genes, in the normal state as well as in the many inherited disorders of hemoglobin, is considered in more detail both later in this chapter and in Chapter 11.

The second gene family shown in Figure 3-2 is the family of olfactory receptor (OR) genes. There are estimated to be as many as 1000 OR genes in the genome. ORs are responsible for our acute sense of smell that can recognize and distinguish thousands of structurally diverse chemicals. OR genes are found throughout the genome on nearly every chromosome, although more than half are found on chromosome 11, including a number of family members near the β-globin cluster.

Pseudogenes

Within both the β-globin and OR gene families are sequences that are related to the functional globin and OR genes but that do not produce any functional RNA or protein product. DNA sequences that closely resemble known genes but are nonfunctional are called **pseudogenes,** and there are tens of thousands of pseudogenes related to many different genes and gene families located all around the genome. Pseudogenes are of two general types, processed and nonprocessed. **Nonprocessed pseudogenes** are thought to be byproducts of evolution, representing "dead" genes that were once functional but are now vestigial, having been inactivated by mutations in critical coding or regulatory sequences. In contrast to nonprocessed pseudogenes, **processed pseudogenes** are pseudogenes that have been formed, not by mutation, but by a process called **retrotransposition,** which involves transcription, generation of a DNA copy of the mRNA (a so-called **cDNA**) by reverse transcription, and finally integration of such DNA copies back into the genome at a location usually quite distant from the original gene. Because such pseudogenes are created by retrotransposition of a DNA copy of processed mRNA, they lack introns and are not necessarily or usually on the same chromosome (or chromosomal region) as their progenitor gene. In many gene families, there are as many or even more pseudogenes as there are functional gene members.

Noncoding RNA Genes

As just discussed, many genes are protein coding and are transcribed into mRNAs that are ultimately translated into their respective proteins; their products comprise the enzymes, structural proteins, receptors, and regulatory proteins that are found in various human tissues and cell types. However, as introduced briefly in Chapter 2, there are additional genes whose functional product appears to be the RNA itself (see Fig. 3-1). These so-called **noncoding RNAs** (**ncRNAs**) have a range of functions in the cell, although many do not as yet have any identified function. By current estimates, there are some 20,000 to 25,000 ncRNA genes in addition to the approximately 20,000 protein-coding genes that we introduced earlier. Thus the collection of ncRNAs represents approximately half of all identified human genes. Chromosome 11, for example, in addition to its 1300 protein-coding genes, has an estimated 1000 ncRNA genes.

Some of the types of ncRNA play largely generic roles in cellular infrastructure, including the tRNAs and rRNAs involved in translation of mRNAs on ribosomes, other RNAs involved in control of RNA splicing, and small nucleolar RNAs (snoRNAs) involved in modifying rRNAs. Additional ncRNAs can be quite long (thus sometimes called long ncRNAs, or **lncRNAs**) and play roles in gene regulation, gene silencing, and human disease, as we explore in more detail later in this chapter.

A particular class of small RNAs of growing importance are the **microRNAs** (**miRNAs**), ncRNAs of only approximately 22 bases in length that suppress translation of target genes by binding to their respective mRNAs and regulating protein production from the target transcript(s). Well over 1000 miRNA genes have been identified in the human genome; some are evolutionarily conserved, whereas others appear to be of quite recent origin during evolution. Some miRNAs have been shown to down-regulate hundreds of mRNAs each, with different combinations of target RNAs in

NONCODING RNAS AND DISEASE

The importance of various types of ncRNAs for medicine is underscored by their roles in a range of human diseases, from early developmental syndromes to adult-onset disorders.

- Deletion of a cluster of miRNA genes on chromosome 13 leads to a form of **Feingold syndrome,** a developmental syndrome of skeletal and growth defects, including microcephaly, short stature, and digital anomalies.
- Mutations in the miRNA gene *MIR96*, in the region of the gene critical for the specificity of recognition of its target mRNA(s), can result in **progressive hearing loss** in adults.
- Aberrant levels of certain classes of miRNAs have been reported in a wide variety of cancers, central nervous system disorders, and cardiovascular disease (see Chapter 15).
- Deletion of clusters of snoRNA genes on chromosome 15 results in **Prader-Willi syndrome,** a disorder characterized by obesity, hypogonadism, and cognitive impairment (see Chapter 6).
- Abnormal expression of a specific lncRNA on chromosome 12 has been reported in patients with a pregnancy-associated disease called **HELLP syndrome**.
- Deletion, abnormal expression, and/or structural abnormalities in different lncRNAs with roles in long-range regulation of gene expression and genome function underlie a variety of disorders involving telomere length maintenance, monoallelic expression of genes in specific regions of the genome, and X chromosome dosage (see Chapter 6).

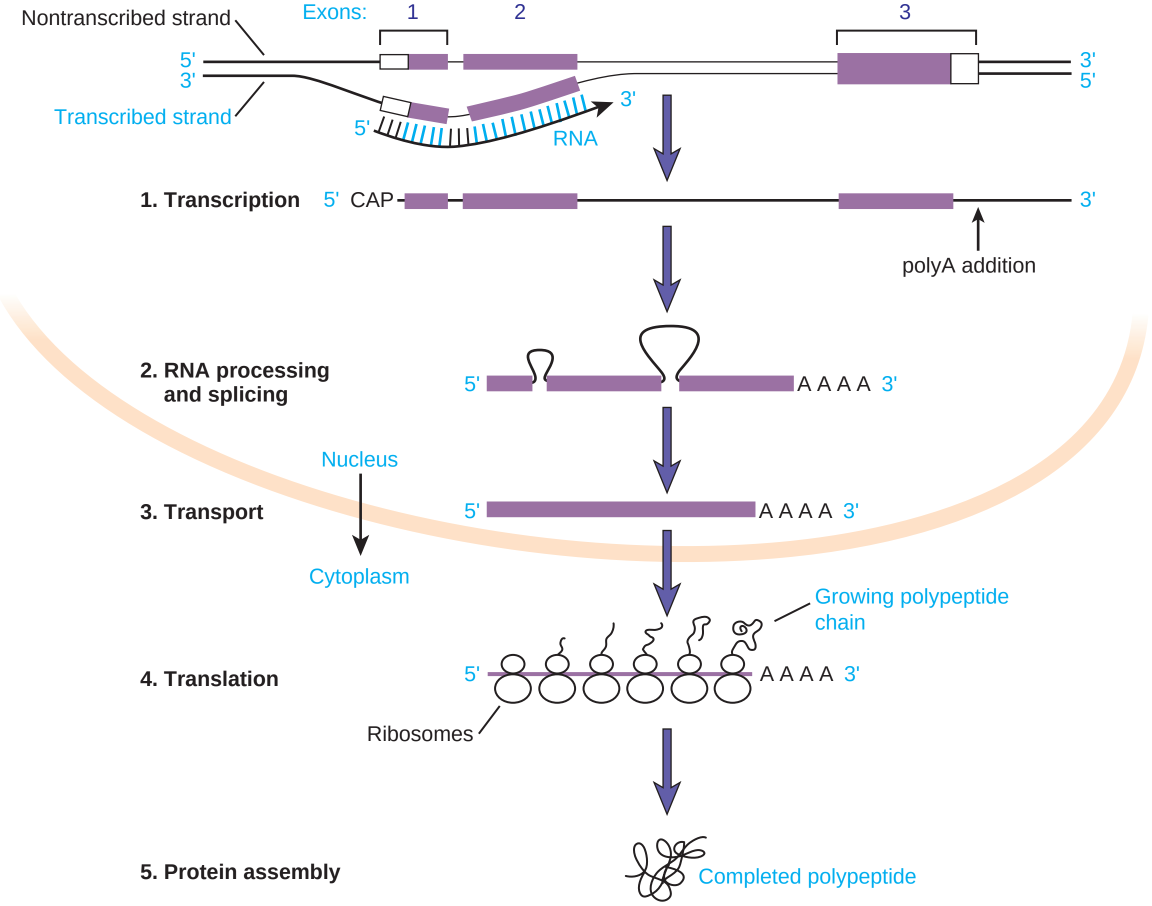

Figure 3-5 **Flow of information from DNA to RNA to protein for a hypothetical gene with three exons and two introns.** Within the exons, *purple* indicates the coding sequences. Steps include transcription, RNA processing and splicing, RNA transport from the nucleus to the cytoplasm, and translation.

different tissues; combined, the miRNAs are thus predicted to control the activity of as many as 30% of all protein-coding genes in the genome.

Although this is a fast-moving area of genome biology, mutations in several ncRNA genes have already been implicated in human diseases, including cancer, developmental disorders, and various diseases of both early and adult onset (see Box).

FUNDAMENTALS OF GENE EXPRESSION

For genes that encode proteins, the flow of information from gene to polypeptide involves several steps (Fig. 3-5). Initiation of transcription of a gene is under the influence of promoters and other regulatory elements, as well as specific proteins known as **transcription factors,** which interact with specific sequences within these regions and determine the spatial and temporal pattern of expression of a gene. Transcription of a gene is initiated at the transcriptional "start" site on chromosomal DNA at the beginning of a 5′ transcribed but *un*translated *r*egion (called the 5′ UTR), just upstream from the coding sequences, and continues along the chromosome for anywhere from several hundred base pairs to more than a million base pairs, through both introns and exons and past the end of the coding sequences. After modification at both the 5′ and 3′ ends of the primary RNA transcript, the portions corresponding to introns are removed, and the segments corresponding to exons are spliced together, a process called **RNA splicing**. After splicing, the resulting mRNA (containing a central segment that is now colinear with the coding portions of the gene) is transported from the nucleus to the cytoplasm, where the mRNA is finally translated into the amino acid sequence of the encoded polypeptide. Each of the steps in this complex pathway is subject to error, and mutations that interfere with the individual steps have been implicated in a number of inherited disorders (see Chapters 11 and 12).

Transcription

Transcription of protein-coding genes by RNA polymerase II (one of several classes of RNA polymerases) is initiated at the transcriptional start site, the point in the 5′ UTR that corresponds to the 5′ end of the final

RNA product (see Figs. 3-4 and 3-5). Synthesis of the primary RNA transcript proceeds in a 5′ to 3′ direction, whereas the strand of the gene that is transcribed and that serves as the template for RNA synthesis is actually read in a 3′ to 5′ direction with respect to the direction of the deoxyribose phosphodiester backbone (see Fig. 2-3). Because the RNA synthesized corresponds both in polarity and in base sequence (substituting U for T) to the 5′ to 3′ strand of DNA, this 5′ to 3′ strand of non-transcribed DNA is sometimes called the *coding*, or **sense,** DNA strand. The 3′ to 5′ strand of DNA that is used as a template for transcription is then referred to as the *noncoding*, or **antisense,** strand. Transcription continues through both intronic and exonic portions of the gene, beyond the position on the chromosome that eventually corresponds to the 3′ end of the mature mRNA. Whether transcription ends at a predetermined 3′ termination point is unknown.

The primary RNA transcript is processed by addition of a chemical "cap" structure to the 5′ end of the RNA and cleavage of the 3′ end at a specific point downstream from the end of the coding information. This cleavage is followed by addition of a polyA tail to the 3′ end of the RNA; the polyA tail appears to increase the stability of the resulting polyadenylated RNA. The location of the polyadenylation point is specified in part by the sequence AAUAAA (or a variant of this), usually found in the 3′ untranslated portion of the RNA transcript. All of these post-transcriptional modifications take place in the nucleus, as does the process of RNA splicing. The fully processed RNA, now called mRNA, is then transported to the cytoplasm, where translation takes place (see Fig. 3-5).

Translation and the Genetic Code

In the cytoplasm, mRNA is translated into protein by the action of a variety of short RNA adaptor molecules, the tRNAs, each specific for a particular amino acid. These remarkable molecules, each only 70 to 100 nucleotides long, have the job of bringing the correct amino acids into position along the mRNA template, to be added to the growing polypeptide chain. Protein synthesis occurs on ribosomes, macromolecular complexes made up of rRNA (encoded by the 18S and 28S rRNA genes), and several dozen ribosomal proteins (see Fig. 3-5).

The key to translation is a code that relates specific amino acids to combinations of three adjacent bases along the mRNA. Each set of three bases constitutes a **codon,** specific for a particular amino acid (Table 3-1). In theory, almost infinite variations are possible in the

TABLE 3-1 The Genetic Code

				Second Base					
First Base	U		C		A		G		Third Base
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC	phe	UCC	ser	UAC	tyr	UGC	cys	C
	UUA	leu	UCA	ser	UAA	stop	UGA	stop	A
	UUG	leu	UCG	ser	UAG	stop	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC	leu	CCC	pro	CAC	his	CGC	arg	C
	CUA	leu	CCA	pro	CAA	gln	CGA	arg	A
	CUG	leu	CCG	pro	CAG	gln	CGG	arg	G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC	ile	ACC	thr	AAC	asn	AGC	ser	C
	AUA	ile	ACA	thr	AAA	lys	AGA	arg	A
	AUG	met	ACG	thr	AAG	lys	AGG	arg	G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC	val	GCC	ala	GAC	asp	GGC	gly	C
	GUA	val	GCA	ala	GAA	glu	GGA	gly	A
	GUG	val	GCG	ala	GAG	glu	GGG	gly	G

Abbreviations for Amino Acids			
ala (A)	alanine	leu (L)	leucine
arg (R)	arginine	lys (K)	lysine
asn (N)	asparagine	met (M)	methionine
asp (D)	aspartic acid	phe (F)	phenylalanine
cys (C)	cysteine	pro (P)	proline
gln (Q)	glutamine	ser (S)	serine
glu (E)	glutamic acid	thr (T)	threonine
his (H)	glycine	trp (W)	tryptophan
gly (G)	histidine	tyr (Y)	tyrosine
ile (I)	isoleucine	val (V)	valine

Stop, Termination codon.
Codons are shown in terms of mRNA, which are complementary to the corresponding DNA codons.

arrangement of the bases along a polynucleotide chain. At any one position, there are four possibilities (A, T, C, or G); thus, for three bases, there are 4^3, or 64, possible triplet combinations. These 64 codons constitute the **genetic code**.

Because there are only 20 amino acids and 64 possible codons, most amino acids are specified by more than one codon; hence the code is said to be **degenerate**. For instance, the base in the third position of the triplet can often be either purine (A or G) or either pyrimidine (T or C) or, in some cases, any one of the four bases, without altering the coded message (see Table 3-1). Leucine and arginine are each specified by six codons. Only methionine and tryptophan are each specified by a single, unique codon. Three of the codons are called **stop** (or **nonsense**) **codons** because they designate termination of translation of the mRNA at that point.

Translation of a processed mRNA is always initiated at a codon specifying methionine. Methionine is therefore the first encoded (amino-terminal) amino acid of each polypeptide chain, although it is usually removed before protein synthesis is completed. The codon for methionine (the **initiator codon,** AUG) establishes the **reading frame** of the mRNA; each subsequent codon is read in turn to predict the amino acid sequence of the protein.

The molecular links between codons and amino acids are the specific tRNA molecules. A particular site on each tRNA forms a three-base **anticodon** that is complementary to a specific codon on the mRNA. Bonding between the codon and anticodon brings the appropriate amino acid into the next position on the ribosome for attachment, by formation of a peptide bond, to the carboxyl end of the growing polypeptide chain. The ribosome then slides along the mRNA exactly three bases, bringing the next codon into line for recognition by another tRNA with the next amino acid. Thus proteins are synthesized from the amino terminus to the carboxyl terminus, which corresponds to translation of the mRNA in a 5′ to 3′ direction.

As mentioned earlier, translation ends when a stop codon (UGA, UAA, or UAG) is encountered in the same reading frame as the initiator codon. (Stop codons in either of the other unused reading frames are not read, and therefore have no effect on translation.) The completed polypeptide is then released from the ribosome, which becomes available to begin synthesis of another protein.

INCREASING FUNCTIONAL DIVERSITY OF PROTEINS

Many proteins undergo extensive post-translational packaging and processing as they adopt their final functional state (see Chapter 12). The polypeptide chain that is the primary translation product folds on itself and forms intramolecular bonds to create a specific three-dimensional structure that is determined by the amino acid sequence itself. Two or more polypeptide chains, products of the same gene or of different genes, may combine to form a single multiprotein complex. For example, two α-globin chains and two β-globin chains associate noncovalently to form a tetrameric hemoglobin molecule (see Chapter 11). The protein products may also be modified chemically by, for example, addition of methyl groups, phosphates, or carbohydrates at specific sites. These modifications can have significant influence on the function or abundance of the modified protein. Other modifications may involve cleavage of the protein, either to remove specific amino-terminal sequences after they have functioned to direct a protein to its correct location within the cell (e.g., proteins that function within mitochondria) or to split the molecule into smaller polypeptide chains. For example, the two chains that make up mature insulin, one 21 and the other 30 amino acids long, are originally part of an 82–amino acid primary translation product called proinsulin.

Transcription of the Mitochondrial Genome

The previous sections described fundamentals of gene expression for genes contained in the nuclear genome. The mitochondrial genome has its own transcription and protein-synthesis system. A specialized RNA polymerase, encoded in the nuclear genome, is used to transcribe the 16-kb mitochondrial genome, which contains two related promoter sequences, one for each strand of the circular genome. Each strand is transcribed in its entirety, and the mitochondrial transcripts are then processed to generate the various individual mitochondrial mRNAs, tRNAs, and rRNAs.

GENE EXPRESSION IN ACTION

The flow of information outlined in the preceding sections can best be appreciated by reference to a particular well-studied gene, the β-globin gene. The β-globin chain is a 146–amino acid polypeptide, encoded by a gene that occupies approximately 1.6 kb on the short arm of chromosome 11. The gene has three exons and two introns (see Fig. 3-4). The β-globin gene, as well as the other genes in the β-globin cluster (see Fig. 3-2), is transcribed in a centromere-to-telomere direction. The orientation, however, is different for different genes in the genome and depends on which strand of the chromosomal double helix is the coding strand for a particular gene.

DNA sequences required for accurate initiation of transcription of the β-globin gene are located in the promoter within approximately 200 bp upstream from the transcription start site. The double-stranded DNA sequence of this region of the β-globin gene, the corresponding RNA sequence, and the translated sequence of the first 10 amino acids are depicted in Figure 3-6 to illustrate the relationships among these three

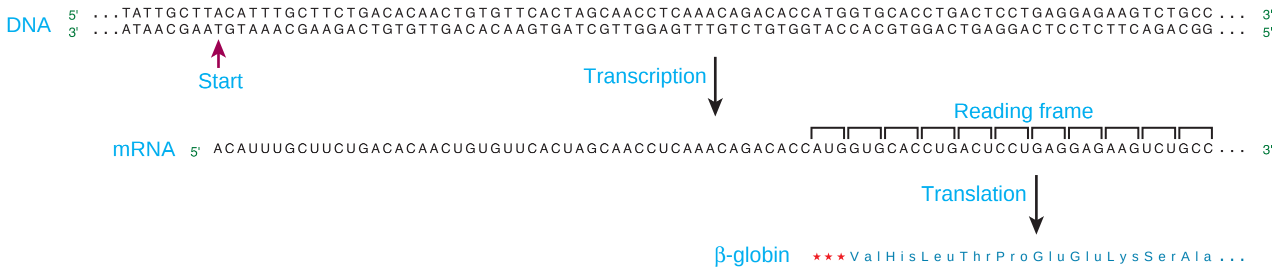
DNA
5' ...TATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCC... 3'
3' ...ATAACGAATGTAAACGAAGACTGTGTTGACACAAGTGATCGTTGGAGTTTGTCTGTGGTACCACGTGGACTGAGGACTCCTCTTCAGACGG... 5'
Start
Transcription
Reading frame
mRNA
5' ACAUUUGCUUCUGACACAACUGUGUUCACUAGCAACCUCAAACAGACACCAUGGUGCACCUGACUCCUGAGGAGAAGUCUGCC... 3'
Translation
β-globin
★★★ValHisLeuThrProGluGluLysSerAla...

Figure 3-6 Structure and nucleotide sequence of the 5′ end of the human β-globin gene on the short arm of chromosome 11. Transcription of the 3′ to 5′ (*lower*) strand begins at the indicated start site to produce β-globin messenger RNA (mRNA). The translational reading frame is determined by the AUG initiator codon (★★★); subsequent codons specifying amino acids are indicated in *blue*. The other two potential frames are not used.

information levels. As mentioned previously, it is the 3′ to 5′ strand of the DNA that serves as the template and is actually transcribed, but it is the 5′ to 3′ strand of DNA that directly corresponds to the 5′ to 3′ sequence of the mRNA (and, in fact, is identical to it except that U is substituted for T). Because of this correspondence, the 5′ to 3′ DNA strand of a gene (i.e., the strand that is *not* transcribed) is the strand generally reported in the scientific literature or in databases.

In accordance with this convention, the complete sequence of approximately 2.0 kb of chromosome 11 that includes the β-globin gene is shown in Figure 3-7. (It is sobering to reflect that a printout of the entire human genome at this scale would require over 300 books the size of this textbook!) Within these 2.0 kb are contained most, but not all, of the sequence elements required to encode and regulate the expression of this gene. Indicated in Figure 3-7 are many of the important structural features of the β-globin gene, including conserved promoter sequence elements, intron and exon boundaries, 5′ and 3′ UTRs, RNA splice sites, the initiator and termination codons, and the polyadenylation signal, all of which are known to be mutated in various inherited defects of the β-globin gene (see Chapter 11).

Initiation of Transcription

The β-globin promoter, like many other gene promoters, consists of a series of relatively short functional elements that interact with specific regulatory proteins (generically called **transcription factors**) that control transcription, including, in the case of the globin genes, those proteins that restrict expression of these genes to erythroid cells, the cells in which hemoglobin is produced. There are well over a thousand sequence-specific, DNA-binding transcription factors in the genome, some of which are ubiquitous in their expression, whereas others are cell type– or tissue-specific.

One important promoter sequence found in many, but not all, genes is the **TATA box,** a conserved region rich in adenines and thymines that is approximately 25 to 30 bp upstream of the start site of transcription (see Figs. 3-4 and 3-7). The TATA box appears to be important for determining the position of the start of transcription, which in the β-globin gene is approximately 50 bp upstream from the translation initiation site (see Fig. 3-6). Thus in this gene, there are approximately 50 bp of sequence at the 5′ end that are transcribed but are not translated; in other genes, the 5′ UTR can be much longer and can even be interrupted by one or more introns. A second conserved region, the so-called CAT box (actually CCAAT), is a few dozen base pairs farther upstream (see Fig. 3-7). Both experimentally induced and naturally occurring mutations in either of these sequence elements, as well as in other regulatory sequences even farther upstream, lead to a sharp reduction in the level of transcription, thereby demonstrating the importance of these elements for normal gene expression. Many mutations in these regulatory elements have been identified in patients with the hemoglobin disorder β-thalassemia (see Chapter 11).

Not all gene promoters contain the two specific elements just described. In particular, genes that are constitutively expressed in most or all tissues (so-called housekeeping genes) often lack the CAT and TATA boxes, which are more typical of tissue-specific genes. Promoters of many housekeeping genes contain a high proportion of cytosines and guanines in relation to the surrounding DNA (see the promoter of the *BRCA1* breast cancer gene in Fig. 3-4). Such CG-rich promoters are often located in regions of the genome called **CpG islands,** so named because of the unusually high concentration of the dinucleotide 5′-CpG-3′ (the *p* representing the phosphate group between adjacent bases; see Fig. 2-3) that stands out from the more general AT-rich genomic landscape. Some of the CG-rich sequence elements found in these promoters are thought to serve as binding sites for specific transcription factors. CpG islands are also important because they are targets for **DNA methylation**. Extensive DNA methylation at CpG islands is usually associated with repression of gene transcription, as we will discuss further later in the

```
            5' ....agccacaccctagggttggccaatctactcccaggagcagggagggcaggagccagggctgggcataaaa
                                                                                      ***
           gtcagggcagagccatctattgcttACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATG

           ValHisLeuThrProGluGluLysSerAlaValThrAlaLeuTrpGlyLysValAsnValAspGluValGlyGlyGlu
Exon 1     GTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAG

           AlaLeuGlyAr-
           GCCCTGGGCAGgttggtatcaaggttacaagacaggtttaaggagaccaatagaaactgggcatgtggagacagagaag

                                                                        -gLeuLeuValValTyr
Intron 1   actcttgggtttctgataggcactgactctctctgcctattggtctattttcccacccttagGCTGCTGGTGGTCTAC

           ProTrpThrGlnArgPhePheGluSerPheGlyAspLeuSerThrProAspAlaValMetGlyAsnProLysValLys
           CCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAG

Exon 2     AlaHisGlyLysLysValLeuGlyAlaPheSerAspGlyLeuAlaHisLeuAspAsnLeuLysGlyThrPheAlaThr
           GCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACA

           LeuSerGluLeuHisCysAspLysLeuHisValAspProGluAsnPheArg
           CTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGgtgagtctatgggacccttgatgtttt

           ctttccccttcttttctatggttaagttcatgtcataggaaggggagaagtaacagggtacagtttagaatgggaaac

           agacgaatgattgcatcagtgtggaagtctcaggatcgttttagtttcttttatttgctgttcataacaattgttttc

           ttttgtttaattcttgctttcttttttttcttctccgcaatttttactattatacttaatgccttaacattgtgtat

Intron 2   aacaaaaggaaatatctctgagatacattaagtaacttaaaaaaaaactttacacagtctgcctagtacattactatt

           tggaatatatgtgtgcttatttgcatattcataatgtccctactttattttcttttatttttaattgatacataatca

           ttatacatatttatgggttaaagtgtaatgttttaatatgtgtacacatattgaccaaatcagggtaattttgcatt

           tgtaattttaaaaaatgctttcttcttttaatatacttttttgtttatcttatttctaatactttccctaatctcttt

           ctttcagggcaataatgatacaatgtatcatgcctctttgcaccattctaaagaataacagtgataatttctgggtta

           aggcaatagcaatatttctgcatataaatatttctgcatataaattgtaactgatgtaagaggtttcatattgctaa

           tagcagctacaatccagctaccattctgcttttattttatggttgggataaggctggattattctgagtccaagctag

                                                       LeuLeuGlyAsnValLeuValCysValLeuAla
           gccctttgctaatcatgttcatacctcttatcttcctcccacagCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCC

           HisHisPheGlyLysGluPheThrProProValGlnAlaAlaTryGlnLysValValAlaGlyValAlaAsnAlaLeu
           CATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTG

           AlaHisLysTyrHisTer
Exon 3     GCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTAC

           TAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCaatgat

           gtatttaaattatttctgaatattttactaaaagggaatgtgggaggtcagtgcatttaaaacataaagaaatgatg

           agctgttcaaaccttgggaaaatacactatatcttaaactccatgaaagaaggtgaggctgcaaccagctaatgcaca

           ttggcaacagcccctgatgcctatgccttattcatccctcagaaaaggattcttgtagaggcttga....   3'
```

Figure 3-7 Nucleotide sequence of the complete human β-globin gene. The sequence of the 5′ to 3′ strand of the gene is shown. *Tan* areas with capital letters represent exonic sequences corresponding to mature mRNA. Lowercase letters indicate introns and flanking sequences. The CAT and TATA box sequences in the 5′ flanking region are indicated in *brown*. The GT and AG dinucleotides important for RNA splicing at the intron-exon junctions and the AATAAA signal important for addition of a polyA tail also are highlighted. The ATG initiator codon (AUG in mRNA) and the TAA stop codon (UAA in mRNA) are shown in *red* letters. The amino acid sequence of β-globin is shown above the coding sequence; the three-letter abbreviations in Table 3-1 are used here. *See Sources & Acknowledgments.*

context of chromatin and its role in the control of gene expression.

Transcription by RNA polymerase II (RNA pol II) is subject to regulation at multiple levels, including binding to the promoter, initiation of transcription, unwinding of the DNA double helix to expose the template strand, and elongation as RNA pol II moves along the DNA. Although some silenced genes are devoid of RNA pol II binding altogether, consistent with their inability to be transcribed in a given cell type, others have RNA pol II poised bidirectionally at the transcriptional start site, perhaps as a means of fine-tuning transcription in response to particular cellular signals.

In addition to the sequences that constitute a promoter itself, there are other sequence elements that can markedly alter the efficiency of transcription. The best characterized of these “activating” sequences are called **enhancers.** Enhancers are sequence elements that can act at a distance from a gene (often several or even hundreds of kilobases away) to stimulate transcription. Unlike promoters, enhancers are both position and orientation independent and can be located either 5′ or 3′ of the transcription start site. Specific enhancer elements function only in certain cell types and thus appear to be involved in establishing the tissue specificity or level of expression of many genes, in concert with one or more transcription factors. In the case of the β-globin gene, several tissue-specific enhancers are present both within the gene itself and in its flanking regions. The interaction of enhancers with specific regulatory proteins leads to increased levels of transcription.

Normal expression of the β-globin gene during development also requires more distant sequences called the **locus control region (LCR),** located upstream of the ε-globin gene (see Fig. 3-2), which is required for establishing the proper chromatin context needed for appropriate high-level expression. As expected, mutations that disrupt or delete either enhancer or LCR sequences interfere with or prevent β-globin gene expression (see Chapter 11).

RNA Splicing

The primary RNA transcript of the β-globin gene contains two introns, approximately 100 and 850 bp in length, that need to be removed and the remaining RNA segments joined together to form the mature mRNA. The process of **RNA splicing,** described generally earlier, is typically an exact and highly efficient one; 95% of β-globin transcripts are thought to be accurately spliced to yield functional globin mRNA. The splicing reactions are guided by specific sequences in the primary RNA transcript at both the 5′ and the 3′ ends of introns. The 5′ sequence consists of nine nucleotides, of which two (the dinucleotide GT [GU in the RNA transcript] located in the intron immediately adjacent to the splice site) are virtually invariant among splice sites in different genes (see Fig. 3-7). The 3′ sequence consists of approximately a dozen nucleotides, of which, again, two—the AG located immediately 5′ to the intron-exon boundary—are obligatory for normal splicing. The splice sites themselves are unrelated to the reading frame of the particular mRNA. In some instances, as in the case of intron 1 of the β-globin gene, the intron actually splits a specific codon (see Fig. 3-7).

The medical significance of RNA splicing is illustrated by the fact that mutations within the conserved sequences at the intron-exon boundaries commonly impair RNA splicing, with a concomitant reduction in the amount of normal, mature β-globin mRNA; mutations in the GT or AG dinucleotides mentioned earlier invariably eliminate normal splicing of the intron containing the mutation. Representative splice site mutations identified in patients with β-thalassemia are discussed in detail in Chapter 11.

Alternative Splicing

As just discussed, when introns are removed from the primary RNA transcript by RNA splicing, the remaining exons are spliced together to generate the final, mature mRNA. However, for most genes, the primary transcript can follow multiple alternative splicing pathways, leading to the synthesis of multiple related but different mRNAs, each of which can be subsequently translated to generate different protein products (see Fig. 3-1). Some of these alternative events are highly tissue- or cell type–specific, and, to the extent that such events are determined by primary sequence, they are subject to allelic variation between different individuals. Nearly all human genes undergo alternative splicing to some degree, and it has been estimated that there are an average of two or three alternative transcripts per gene in the human genome, thus greatly expanding the information content of the human genome beyond the approximately 20,000 protein-coding genes. The regulation of alternative splicing appears to play a particularly impressive role during neuronal development, where it may contribute to generating the high levels of functional diversity needed in the nervous system. Consistent with this, susceptibility to a number of neuropsychiatric conditions has been associated with shifts or disruption of alternative splicing patterns.

Polyadenylation

The mature β-globin mRNA contains approximately 130 bp of 3′ untranslated material (the 3′ UTR) between the stop codon and the location of the polyA tail (see Fig. 3-7). As in other genes, cleavage of the 3′ end of the mRNA and addition of the polyA tail is controlled, at least in part, by an AAUAAA sequence approximately 20 bp before the polyadenylation site. Mutations in this polyadenylation signal in patients with β-thalassemia document the importance of this signal for proper 3′ cleavage and polyadenylation (see Chapter 11). The 3′ UTR of some genes can be up to several kb in length. Other genes have a number of alternative polyadenylation sites, selection among which may influence the stability of the resulting mRNA and thus the steady-state level of each mRNA.

RNA Editing and RNA-DNA Sequence Differences

Recent findings suggest that the conceptual principle underlying the central dogma—that RNA and protein sequences reflect the underlying genomic sequence—may not always hold true. RNA editing to change the nucleotide sequence of the mRNA has been demonstrated in a number of organisms, including humans. This process involves deamination of adenosine at particular sites, converting an A in the DNA sequence to an inosine in the resulting RNA; this is then read by the translational machinery as a G, leading to changes in gene expression and protein function, especially in the nervous system. More widespread RNA-DNA differences involving other bases (with corresponding changes in the encoded amino acid sequence) have also been reported, at levels that vary among individuals. Although the mechanism(s) and clinical relevance of these events remain controversial, they illustrate the existence of a range of processes capable of increasing transcript and proteome diversity.

EPIGENETIC AND EPIGENOMIC ASPECTS OF GENE EXPRESSION

Given the range of functions and fates that different cells in any organism must adopt over its lifetime, it is apparent that not all genes in the genome can be actively expressed in every cell at all times. As important as completion of the Human Genome Project has been for contributing to our understanding of human biology and disease, identifying the genomic sequences and features that direct developmental, spatial, and temporal aspects of gene expression remains a formidable challenge. Several decades of work in molecular biology have defined critical regulatory elements for many individual genes, as we saw in the previous section, and more recent attention has been directed toward performing such studies on a genome-wide scale.

In Chapter 2, we introduced general aspects of chromatin that package the genome and its genes in all cells. Here, we explore the specific characteristics of chromatin that are associated with active or repressed genes as a step toward identifying the regulatory code for expression of the human genome. Such studies focus on reversible changes in the chromatin landscape as determinants of gene function rather than on changes to the genome sequence itself and are thus called *epi*genetic or, when considered in the context of the entire genome, *epig*enomic (Greek *epi-*, over or upon).

The field of **epigenetics** is growing rapidly and is the study of heritable changes in cellular function or gene expression that can be transmitted from cell to cell (and even generation to generation) as a result of chromatin-based molecular signals (Fig. 3-8). Complex epigenetic states can be established, maintained, and transmitted by a variety of mechanisms: modifications to the DNA, such as **DNA methylation;** numerous **histone modifications** that alter chromatin packaging or access; and substitution of specialized **histone variants** that mark chromatin associated with particular sequences or regions in the genome. These chromatin changes can be highly dynamic and transient, capable of responding rapidly and sensitively to changing needs in the cell, or they can be long lasting, capable of being transmitted through multiple cell divisions or even to subsequent generations. In either instance, the key concept is that epigenetic mechanisms do *not* alter the underlying DNA sequence, and this distinguishes them from genetic mechanisms, which are sequence based. Together, the epigenetic marks and the DNA sequence make up the set of signals that guide the genome to express its genes at the right time, in the right place, and in the right amounts.

Increasing evidence points to a role for epigenetic changes in human disease in response to environmental or lifestyle influences. The dynamic and reversible nature of epigenetic changes permits a level of adaptability or plasticity that greatly exceeds the capacity of DNA sequence alone and thus is relevant both to the origins and potential treatment of disease. A number of large-scale epigenomics projects (akin to the original Human Genome Project) have been initiated to catalogue DNA methylation sites genome-wide (the so-called methylome), to evaluate CpG landscapes across the genome, to discover new histone variants and modification patterns in various tissues, and to document positioning of nucleosomes around the genome in different cell types, and in samples from both asymptomatic individuals and those with cancer or other diseases. These analyses are part of a broad effort (called the **ENCODE Project,** for *Enc*yclopedia *o*f *D*NA *E*lements) to explore epigenetic patterns in chromatin genome-wide in order to better understand control of gene expression in different tissues or disease states.

DNA Methylation

DNA methylation involves the modification of cytosine bases by methylation of the carbon at the fifth position in the pyrimidine ring (Fig. 3-9). Extensive DNA methylation is a mark of repressed genes and is a widespread mechanism associated with the establishment of specific programs of gene expression during cell differentiation and development. Typically, DNA methylation occurs on the C of CpG dinucleotides (see Fig. 3-8) and inhibits gene expression by recruitment of specific methyl-CpG–binding proteins that, in turn, recruit chromatin-modifying enzymes to silence transcription. The presence of 5-methylcytosine (5-mC) is considered to be a stable epigenetic mark that can be faithfully transmitted through cell division; however, altered methylation states are frequently observed in cancer, with hypomethylation of large genomic segments or with regional hypermethylation (particularly at CpG islands) in others (see Chapter 15).

Extensive demethylation occurs during germ cell development and in the early stages of embryonic development,

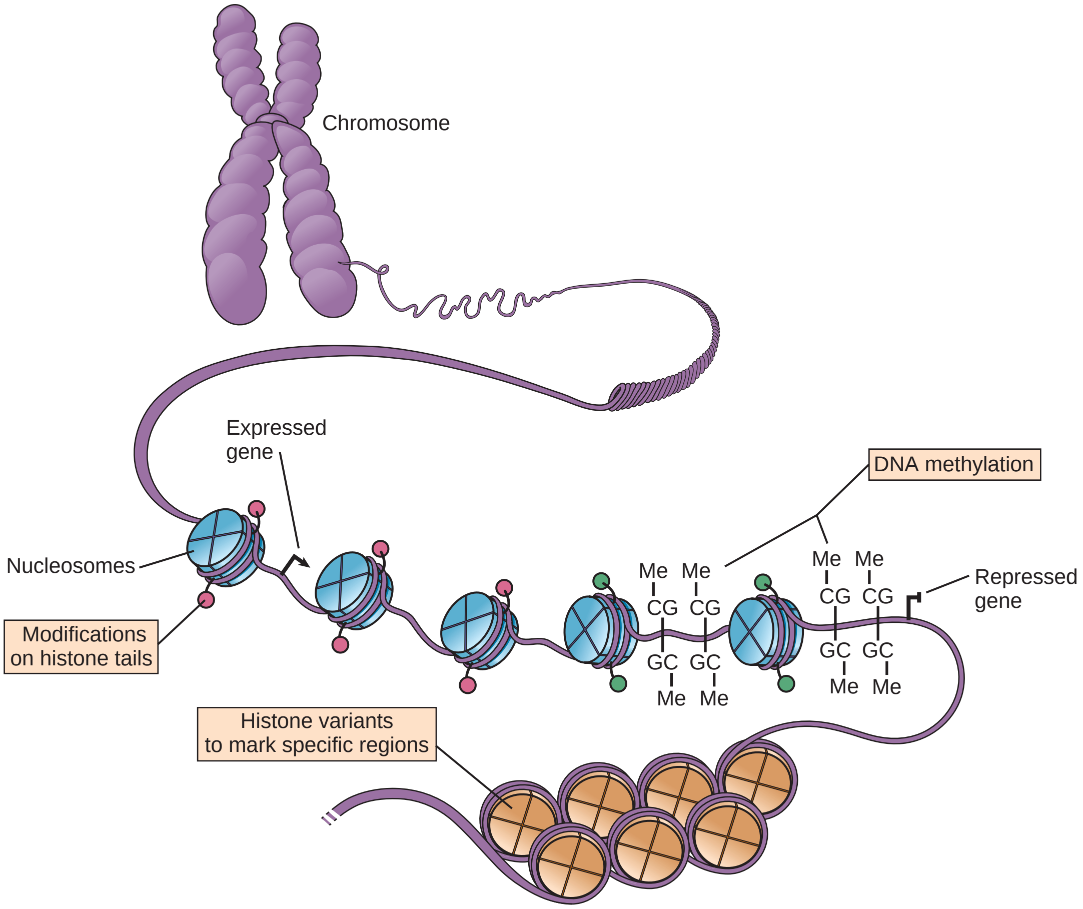

Figure 3-8 Schematic representation of chromatin and three major epigenetic mechanisms: DNA methylation at CpG dinucleotides, associated with gene repression; various modifications (indicated by different colors) on histone tails, associated with either gene expression or repression; and various histone variants that mark specific regions of the genome, associated with specific functions required for chromosome stability or genome integrity. Not to scale.

consistent with the need to "re-set" the chromatin environment and restore totipotency or pluripotency of the zygote and of various stem cell populations. Although the details are still incompletely understood, these reprogramming steps appear to involve the enzymatic conversion of 5-mC to 5-hydroxymethylcytosine (5-hmC; see Fig. 3-9), as a likely intermediate in the demethylation of DNA. Overall, 5-mC levels are stable across adult tissues (approximately 5% of all cytosines), whereas 5-hmC levels are much lower and much more variable (0.1% to 1% of all cytosines). Interestingly, although 5-hmC is widespread in the genome, its highest levels are found in known regulatory regions, suggesting a possible role in the regulation of specific promoters and enhancers.

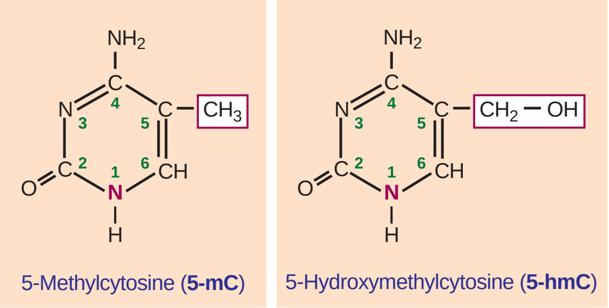

Figure 3-9 The modified DNA bases, 5-methylcytosine and 5-hydroxymethylcytosine. Compare to the structure of cytosine in Figure 2-2. The added methyl and hydroxymethyl groups are boxed in *purple*. The atoms in the pyrimidine rings are numbered 1 to 6 to indicate the 5-carbon.

Histone Modifications

A second class of epigenetic signals consists of an extensive inventory of modifications to any of the core histone types, H2A, H2B, H3, and H4 (see Chapter 2). Such modifications include histone methylation, phosphorylation, acetylation, and others at specific amino acid residues, mostly located on the N-terminal "tails" of histones that extend out from the core nucleosome itself (see Fig. 3-8). These epigenetic modifications are believed to influence gene expression by affecting chromatin compaction or accessibility and by signaling protein complexes that—depending on the nature of the signal—activate or silence gene expression at that site. There are dozens of modified sites that can be experimentally queried genome-wide by using antibodies that recognize specifically modified sites—for example, histone H3 methylated at lysine position 9 (H3K9 methylation, using the one-letter abbreviation K for lysine; see Table 3-1) or histone H3 acetylated at lysine position 27 (H3K27 acetylation). The former is a repressive mark associated with silent regions of the genome, whereas the latter is a mark for activating regulatory regions.

Specific patterns of different histone modifications are associated with promoters, enhancers, or the body of genes in different tissues and cell types. The ENCODE Project, introduced earlier, examined 12 of the most common modifications in nearly 50 different cell types and integrated the individual chromatin profiles to assign putative functional attributes to well over half of the human genome. This finding implies that much more of the genome plays a role, directly or indirectly, in determining the varied patterns of gene expression that distinguish cell types than previously inferred from the fact that less than 2% of the genome is "coding" in a traditional sense.

Histone Variants

The histone modifications just discussed involve modification of the core histones themselves, which are all encoded by multigene clusters in a few locations in the genome. In contrast, the many dozens of histone variants are products of entirely different genes located elsewhere in the genome, and their amino acid sequences are distinct from, although related to, those of the canonical histones.

Different histone variants are associated with different functions, and they replace—all or in part—the related member of the core histones found in typical nucleosomes to generate specialized chromatin structures (see Fig. 3-8). Some variants mark specific regions or loci in the genome with highly specialized functions; for example, the CENP-A histone is a histone H3-related variant that is found exclusively at functional centromeres in the genome and contributes to essential features of centromeric chromatin that mark the location of kinetochores along the chromosome fiber. Other variants are more transient and mark regions of the genome with particular attributes; for example, H2A.X is a histone H2A variant involved in the response to DNA damage to mark regions of the genome that require DNA repair.

Chromatin Architecture

In contrast to the impression one gets from viewing the genome as a linear string of sequence (see Fig. 3-7), the genome adopts a highly ordered and dynamic arrangement within the space of the nucleus, correlated with and likely guided by the epigenetic and epigenomic signals just discussed. This three-dimensional landscape is highly predictive of the map of all expressed sequences in any given cell type (the **transcriptome**) and reflects dynamic changes in chromatin architecture at different levels (Fig. 3-10). First, large chromosomal domains (up to millions of base pairs in size) can exhibit coordinated patterns of gene expression at the chromosome level, involving dynamic interactions between different intrachromosomal and interchromosomal points of contact within the nucleus. At a finer level, technical advances to map and sequence points of contact around the genome in the context of three-dimensional space have pointed to ordered loops of chromatin that position and orient genes precisely, exposing or blocking critical regulatory regions for access by RNA pol II, transcription factors, and other regulators. Lastly, specific and dynamic patterns of nucleosome positioning differ among cell types and tissues in the face of changing environmental and developmental cues (see Fig. 3-10). The biophysical, epigenomic, and/or genomic properties that facilitate or specify the orderly and dynamic packaging of each chromosome during each cell cycle, without reducing the genome to a disordered tangle within the nucleus, remain a marvel of landscape engineering.

GENE EXPRESSION AS THE INTEGRATION OF GENOMIC AND EPIGENOMIC SIGNALS

The gene expression program of a cell encompasses the specific subset of the approximately 20,000 protein-coding genes in the genome that are actively transcribed and translated into their respective functional products, the subset of the estimated 20,000 to 25,000 ncRNA genes that are transcribed, the amount of products produced, and the particular sequence (alleles) of those products. The gene expression profile of any particular cell or cell type in a given individual at a given time (whether in the context of the cell cycle, early development, or one's entire life span) and under a given set of circumstances (as influenced by environment, lifestyle, or disease) is thus the integrated sum of several different but interrelated effects, including the following:

- The primary sequence of genes, their allelic variants, and their encoded products
- Regulatory sequences and their epigenetic positioning in chromatin
- Interactions with the thousands of transcriptional factors, ncRNAs, and other proteins involved in the control of transcription, splicing, translation, and post-translational modification
- Organization of the genome into subchromosomal domains

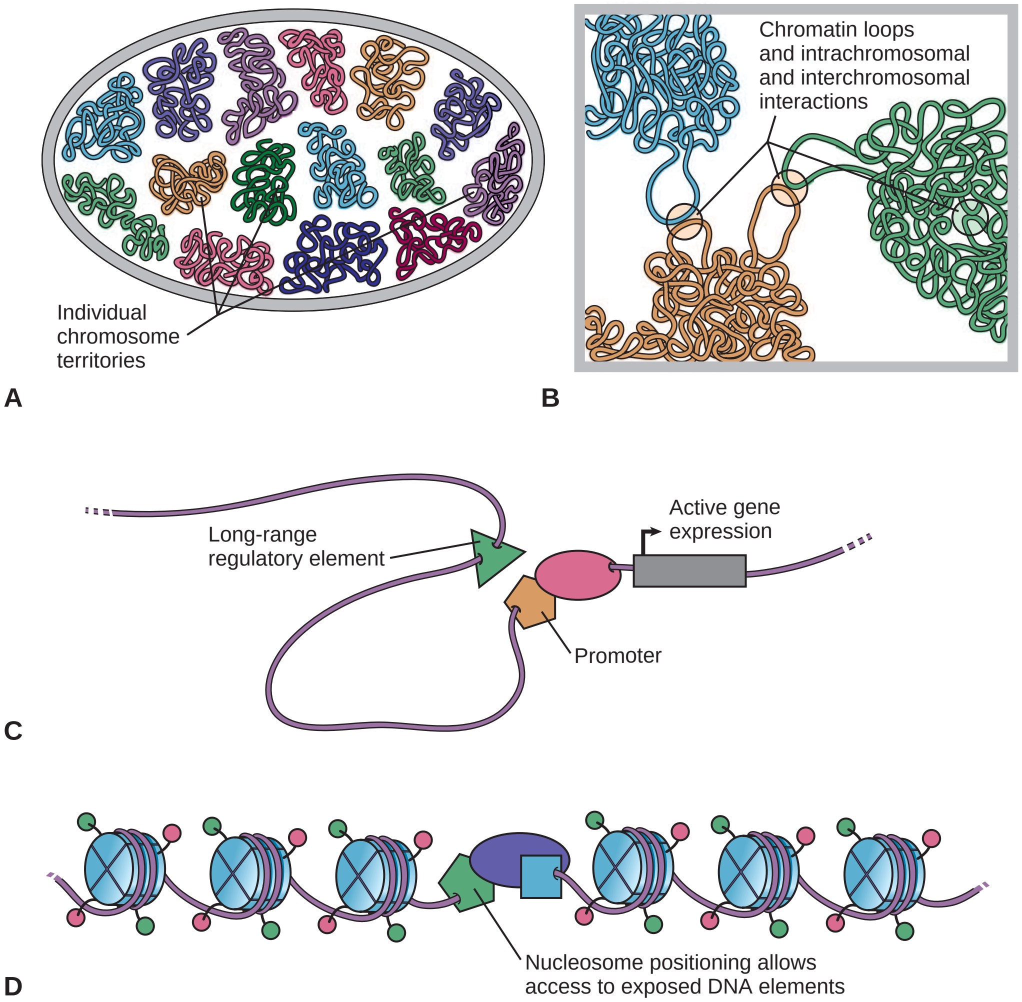

Figure 3-10 **Three-dimensional architecture and dynamic packaging of the genome, viewed at increasing levels of resolution. A,** Within interphase nuclei, each chromosome occupies a particular territory, represented by the different colors. **B,** Chromatin is organized into large subchromosomal domains within each territory, with loops that bring certain sequences and genes into proximity with each other, with detectable intrachromosomal and interchromosomal interactions. **C,** Loops bring long-range regulatory elements (e.g., enhancers or locus-control regions) into association with promoters, leading to active transcription and gene expression. **D,** Positioning of nucleosomes along the chromatin fiber provides access to specific DNA sequences for binding by transcription factors and other regulatory proteins.

- Programmed interactions between different parts of the genome
- Dynamic three-dimensional chromatin packaging in the nucleus

All of these orchestrate in an efficient, hierarchical, and highly programmed fashion. Disruption of any one—due to genetic variation, to epigenetic changes, and/or to disease-related processes—would be expected to alter the overall cellular program and its functional output (see Box).

ALLELIC IMBALANCE IN GENE EXPRESSION

It was once assumed that genes present in two copies in the genome would be expressed from both homologues at comparable levels. However, it has become increasingly evident that there can be extensive imbalance between alleles, reflecting both the amount of sequence variation in the genome and the interplay between genome sequence and epigenetic patterns that were just discussed.

THE EPIGENETIC LANDSCAPE OF THE GENOME AND MEDICINE

- Different chromosomes and chromosomal regions occupy characteristic territories within the nucleus. The probability of physical proximity influences the incidence of specific chromosome abnormalities (see Chapters 5 and 6).
- The genome is organized into megabase-sized domains with locally shared characteristics of base pair composition (i.e., GC rich or AT rich), gene density, timing of replication in the S phase, and presence of particular histone modifications (see Chapter 5).
- Modules of coexpressed genes correspond to distinct anatomical or developmental stages in, for example, the human brain or the hematopoietic lineage. Such coexpression networks are revealed by shared regulatory networks and epigenetic signals, by clustering within genomic domains, and by overlapping patterns of altered gene expression in various disease states.
- Although monozygotic twins share virtually identical genomes, they can be quite discordant for certain traits, including susceptibility to common diseases. Significant changes in DNA methylation occur during the lifetime of such twins, implicating epigenetic regulation of gene expression as a source of diversity.
- The epigenetic landscape can integrate genomic and environmental contributions to disease. For example, differential DNA methylation levels correlate with underlying sequence variation at specific loci in the genome and thereby modulate genetic risk for rheumatoid arthritis.

In Chapter 2, we introduced the general finding that any individual genome carries two different alleles at a minimum of 3 to 5 million positions around the genome, thus distinguishing by sequence the maternally and paternally inherited copies of that sequence position (see Fig. 2-6). Here, we explore ways in which those sequence differences reveal allelic imbalance in gene expression, both at autosomal loci and at X chromosome loci in females.

By determining the sequences of all the RNA products—the transcriptome—in a population of cells, one can quantify the relative level of transcription of all the genes (both protein-coding and noncoding) that are transcriptionally active in those cells. Consider, for example, the collection of protein-coding genes. Although an average cell might contain approximately 300,000 copies of mRNA in total, the abundance of specific mRNAs can differ over many orders of magnitude; among genes that are active, most are expressed at low levels (estimated to be < 10 copies of that gene's mRNA per cell), whereas others are expressed at much higher levels (several hundred to a few thousand copies of that mRNA per cell). Only in highly specialized cell types are particular genes expressed at very high levels (many tens of thousands of copies) that account for a significant proportion of all mRNA in those cells.

Now consider an expressed gene with a sequence variant that allows one to distinguish between the RNA products (whether mRNA or ncRNA) transcribed from each of two alleles, one allele with a T that is transcribed to yield RNA with an A and the other allele with a C that is transcribed to yield RNA with a G (Fig. 3-11). By sequencing individual RNA molecules and comparing the number of sequences generated that contain an A or G at that position, one can infer the ratio of transcripts from the two alleles in that sample. Although most genes show essentially equivalent levels of biallelic expression, recent analyses of this type have demonstrated widespread unequal allelic expression for 5% to 20% of autosomal genes in the genome (Table 3-2). For most of these genes, the extent of imbalance is twofold or less, although up to tenfold differences have been observed for some genes. This allelic imbalance may reflect interactions between genome sequence and gene regulation; for example, sequence changes can alter the relative binding of various transcription factors or other transcriptional regulators to the two alleles or the extent of DNA methylation observed at the two alleles (see Table 3-2).

Monoallelic Gene Expression

Some genes, however, show a much more complete form of allelic imbalance, resulting in monoallelic gene expression (see Fig. 3-11). Several different mechanisms have been shown to account for allelic imbalance of this type for particular subsets of genes in the genome: DNA rearrangement, random monoallelic expression, parent-of-origin imprinting, and, for genes on the X chromosome in females, X chromosome inactivation. Their distinguishing characteristics are summarized in Table 3-2.

Somatic Rearrangement

A highly specialized form of monoallelic gene expression is observed in the genes encoding **immunoglobulins** and **T-cell receptors**, expressed in B cells and T cells, respectively, as part of the immune response. Antibodies are encoded in the germline by a relatively small number of genes that, during B-cell development, undergo a unique process of somatic rearrangement that involves the cutting and pasting of DNA sequences in lymphocyte precursor cells (but *not* in any other cell lineages) to rearrange genes in somatic cells to generate enormous antibody diversity. The highly orchestrated DNA rearrangements occur across many hundreds of kilobases but involve only one of the two alleles, which is chosen randomly in any given B cell (see Table 3-2). Thus expression of mature mRNAs for the immunoglobulin heavy or light chain subunits is exclusively monoallelic.

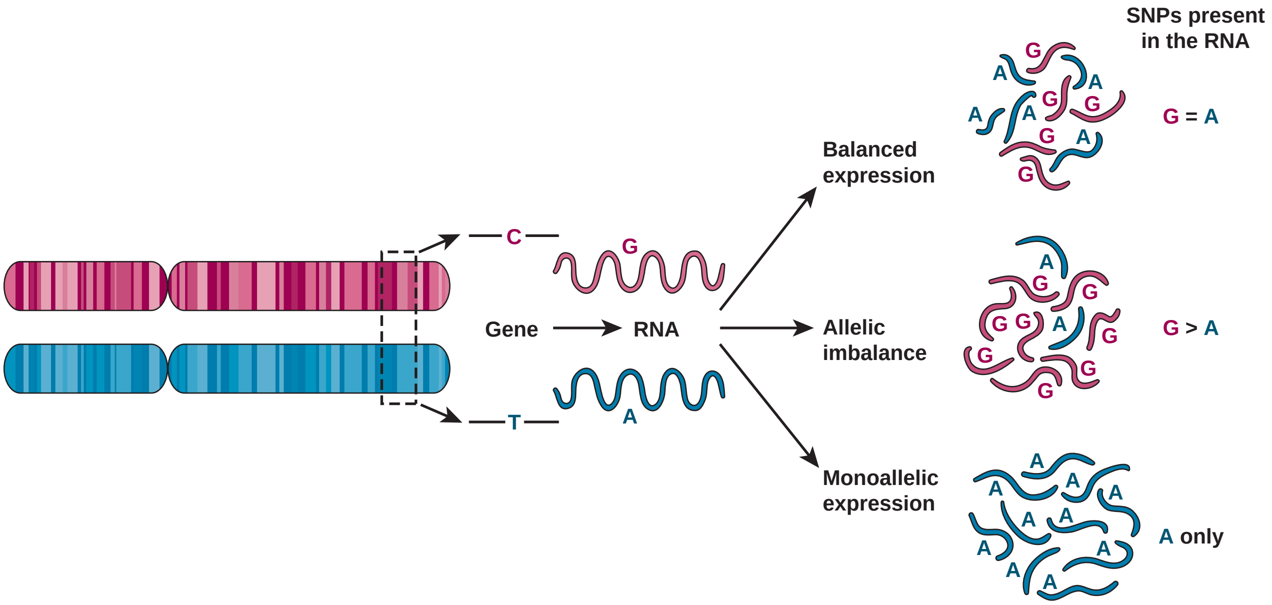

Figure 3-11 Allelic expression patterns for a gene sequence with a transcribed DNA variant (here, a C or a T) to distinguish the alleles. As described in the text, the relative abundance of RNA transcripts from the two alleles (here, carrying a G or an A) demonstrates whether the gene shows balanced expression *(top)*, allelic imbalance *(center)*, or exclusively monoallelic expression *(bottom)*. Different underlying mechanisms for allelic imbalance are compared in Table 3-2. SNP, Single nucleotide polymorphism.

TABLE 3-2 Allelic Imbalance in Gene Expression

Type	Characteristics	Genes Affected	Basis	Developmental Origin
Unbalanced expression	Unequal RNA abundance from two alleles due to DNA variants and associated epigenetic changes; usually < twofold difference in expression	5%-20% of autosomal genes	Sequence variants cause different levels of expression at the two alleles	Early embryogenesis
Monoallelic expression				
• Somatic rearrangement	Changes in DNA organization to produce functional gene at one allele, but not other	Immunoglobulin genes, T-cell receptor genes	Random choice of one allele	B- and T-cell lineages
• Random allelic silencing or activation	Expression from only one allele at a locus, due to differential epigenetic packaging at locus	Olfactory receptor genes in sensory neurons; other chemosensory or immune system genes; up to 10% of all genes in other cell types	Random choice of one allele	Specific cell types
• Genomic imprinting	Epigenetic silencing of allele(s) in imprinted region	>100 genes with functions in development	Imprinted region marked epigenetically according to parent of origin	Parental germline
• X chromosome inactivation	Epigenetic silencing of alleles on one X chromosome in females	Most X-linked genes in females	Random choice of one X chromosome	Early embryogenesis

This mechanism of somatic rearrangement and random monoallelic gene expression is also observed at the T-cell receptor genes in the T-cell lineage. However, such behavior is unique to these gene families and cell lineages; the rest of the genome remains highly stable throughout development and differentiation.

Random Monoallelic Expression

In contrast to this highly specialized form of DNA rearrangement, monoallelic expression typically results from differential epigenetic regulation of the two alleles. One well-studied example of random monoallelic expression involves the OR gene family described earlier (see Fig. 3-2). In this case, only a single allele of one OR gene is expressed in each olfactory sensory neuron; the many hundred other copies of the OR family remain repressed in that cell. Other genes with chemosensory or immune system functions also show random monoallelic expression, suggesting that this mechanism may be a general one for increasing the diversity of responses for cells that interact with the outside world. However, this mechanism is apparently not restricted to the immune and sensory systems, because a substantial subset of all human genes (5% to 10% in different cell types) has been shown to undergo random allelic silencing; these genes are broadly distributed on all autosomes, have a wide range of functions, and vary in terms of the cell types and tissues in which monoallelic expression is observed.

Parent-of-Origin Imprinting

For the examples just described, the choice of which allele is expressed is not dependent on parental origin; either the maternal or paternal copy can be expressed in different cells and their clonal descendants. This distinguishes *random* forms of monoallelic expression from **genomic imprinting,** in which the choice of the allele to be expressed is *nonrandom* and is determined solely by parental origin. Imprinting is a normal process involving the introduction of epigenetic marks (see Fig. 3-8) in the germline of one parent, but not the other, at specific locations in the genome. These lead to monoallelic expression of a gene or, in some cases, of multiple genes within the imprinted region.

Imprinting takes place during gametogenesis, before fertilization, and marks certain genes as having come from the mother or father (Fig. 3-12). After conception, the parent-of-origin imprint is maintained in some or all of the somatic tissues of the embryo and silences gene expression on allele(s) within the imprinted region; whereas some imprinted genes show monoallelic expression throughout the embryo, others show tissue-specific imprinting, especially in the placenta, with biallelic expression in other tissues. The imprinted state persists postnatally into adulthood through hundreds of cell divisions so that only the maternal or paternal copy of the gene is expressed. Yet, imprinting must be reversible: a paternally derived allele, when it is inherited by a female, must be converted in her germline so that she can then pass it on with a maternal imprint to her offspring. Likewise, an imprinted maternally derived allele, when it is inherited by a male, must be converted in his germline so that he can pass it on as a paternally imprinted allele to his offspring (see Fig. 3-12). Control over this conversion process appears to be governed by specific DNA elements called **imprinting control regions** or **imprinting centers** that are located within imprinted regions throughout the genome; although their precise mechanism of action is not known, many appear to involve ncRNAs that initiate the epigenetic change in chromatin, which then spreads outward along the chromosome over the imprinted region. Notably, although the imprinted region can encompass more than a single gene, this form of monoallelic expression is confined to a delimited genomic segment, typically a few hundred kilobase pairs to a few megabases in overall size; this distinguishes genomic imprinting both from the more general form of random monoallelic expression described earlier (which appears to involve individual genes under locus-specific control) and from X chromosome inactivation, described in the next section (which involves genes along the entire chromosome).

To date, approximately 100 imprinted genes have been identified on many different autosomes. The involvement of these genes in various chromosomal disorders is described more fully in Chapter 6. For clinical conditions due to a single imprinted gene, such as **Prader-Willi syndrome** (Case 38) and **Beckwith-Wiedemann syndrome** (Case 6), the effect of genomic imprinting on inheritance patterns in pedigrees is discussed in Chapter 7.

X Chromosome Inactivation

The chromosomal basis for sex determination, introduced in Chapter 2 and discussed in more detail in Chapter 6, results in a dosage difference between typical males and females with respect to genes on the X chromosome. Here we discuss the chromosomal and molecular mechanisms of X chromosome inactivation, the most extensive example of random monoallelic expression in the genome and a mechanism of **dosage compensation** that results in the epigenetic silencing of most genes on one of the two X chromosomes in females.

In normal female cells, the choice of which X chromosome is to be inactivated is a random one that is then maintained in each clonal lineage. Thus females are mosaic with respect to X-linked gene expression; some cells express alleles on the paternally inherited X but not the maternally inherited X, whereas other cells do the opposite (Fig. 3-13). This mosaic pattern of gene expression distinguishes most X-linked genes from imprinted genes, whose expression, as we just noted, is determined strictly by parental origin.

Although the inactive X chromosome was first identified cytologically by the presence of a heterochromatic mass (called the **Barr body**) in interphase cells, many epigenetic features distinguish the active and inactive X chromosomes, including DNA methylation, histone modifications, and a specific histone variant, macroH2A, that is particularly enriched in chromatin on the inactive X. As well as providing insights into the mechanisms of X inactivation, these features can be useful diagnostically for identifying inactive X chromosomes in clinical material, as we will see in Chapter 6.

Although X inactivation is clearly a chromosomal phenomenon, not all genes on the X chromosome show monoallelic expression in female cells. Extensive

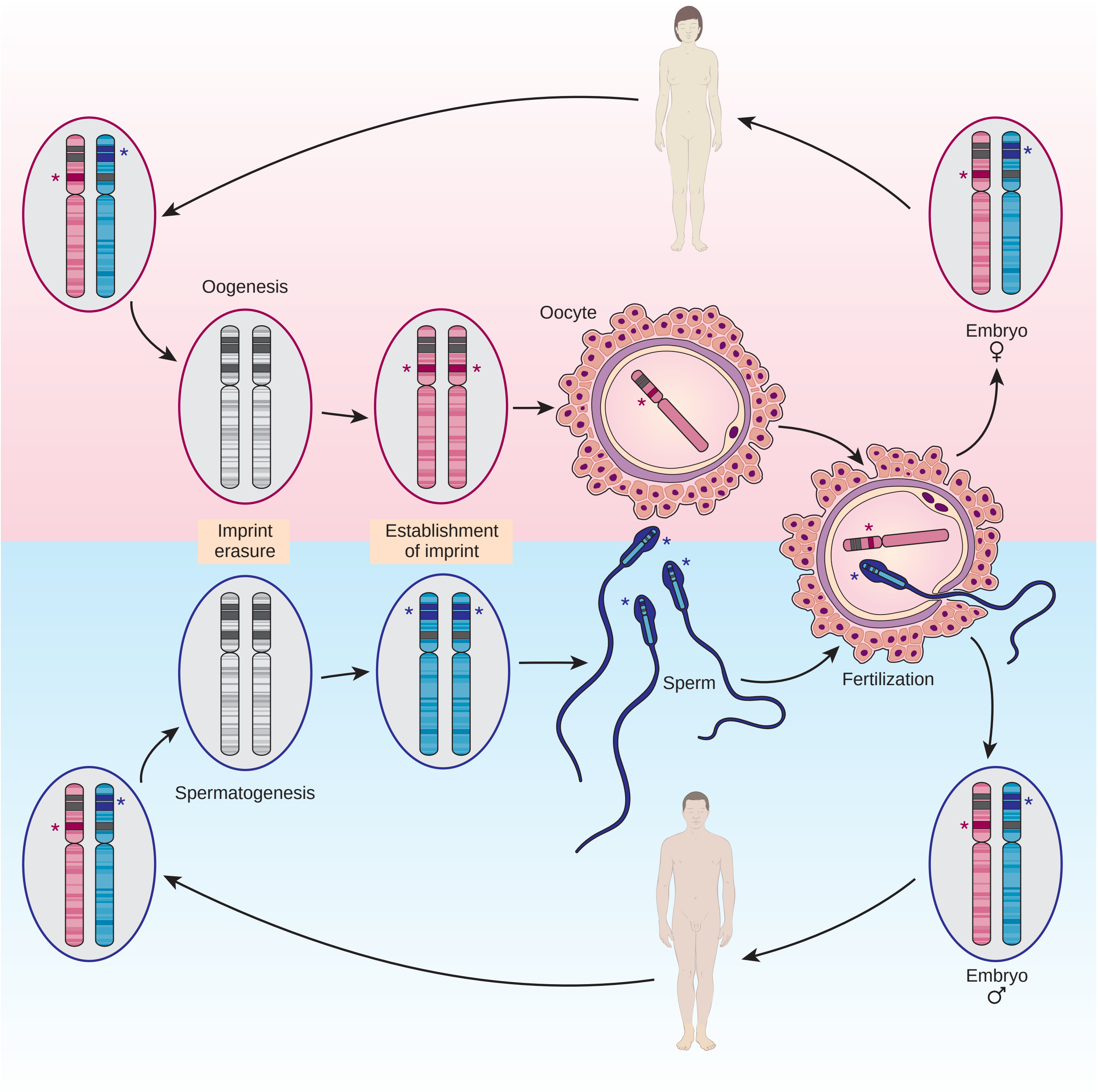

Figure 3-12 Genomic imprinting and conversion of maternal and paternal imprints during passage through male or female gametogenesis. Within a hypothetical imprinted region on an pair of homologous autosomes, paternally imprinted genes are indicated in *blue*, whereas a maternally imprinted gene is indicated in *red*. After fertilization, both male and female embryos have one copy of the chromosome carrying a paternal imprint and one copy carrying a maternal imprint. During oogenesis *(top)* and spermatogenesis *(bottom)*, the imprints are erased by removal of epigenetic marks, and new imprints determined by the sex of the parent are established within the imprinted region. Gametes thus carry a monoallelic imprint appropriate to the parent of origin, whereas somatic cells in both sexes carry one chromosome of each imprinted type.

analysis of expression of nearly all X-linked genes has demonstrated that at least 15% of the genes show biallelic expression and are expressed from both active and inactive X chromosomes, at least to some extent; a proportion of these show significantly higher levels of mRNA production in female cells compared to male cells and are interesting candidates for a role in explaining sexually dimorphic traits.

A special subset of genes is located in the pseudoautosomal segments, which are essentially identical on

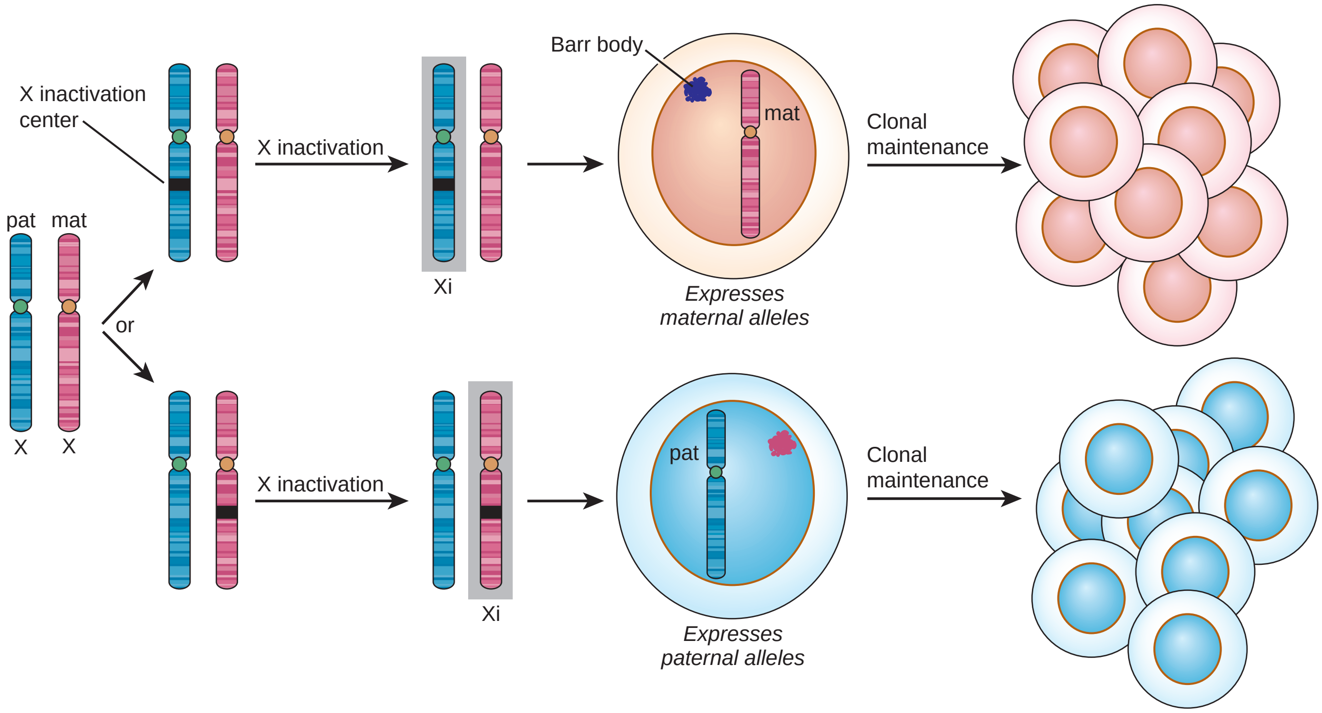

Figure 3-13 **Random X chromosome inactivation early in female development.** Shortly after conception of a female embryo, both the paternally and maternally inherited X chromosomes (pat and mat, respectively) are active. Within the first week of embryogenesis, one or the other X is chosen at random to become the future inactive X, through a series of events involving the X inactivation center *(black box)*. That X then becomes the inactive X (Xi, indicated by the *shading*) in that cell and its progeny and forms the Barr body in interphase nuclei. The resulting female embryo is thus a clonal mosaic of two epigenetically determined cell types: one expresses alleles from the maternal X (*pink* cells), whereas the other expresses alleles from the paternal X (*blue* cells). The ratio of the two cell types is determined randomly but varies among normal females and among females who are carriers of X-linked disease alleles (see Chapters 6 and 7).

the X and Y chromosomes and undergo recombination during spermatogenesis (see Chapter 2). These genes have two copies in both females (two X-linked copies) and males (one X-linked and one Y-linked copy) and thus do not undergo X inactivation; as expected, these genes show balanced biallelic expression, as one sees for most autosomal genes.

The X Inactivation Center and the* XIST *Gene. X inactivation occurs very early in female embryonic development, and determination of which X will be designated the inactive X in any given cell in the embryo is a random choice under the control of a complex locus called the **X inactivation center**. This region contains an unusual ncRNA gene, ***XIST,*** that appears to be a key master regulatory locus for X inactivation. *XIST (*an acronym for inactive X [*Xi*]–specific *t*ranscripts) has the novel feature that it is expressed only from the allele on the inactive X; it is transcriptionally silent on the active X in both male and female cells. Although the exact mode of action of *XIST* is unknown, X inactivation cannot occur in its absence. The product of *XIST* is a long ncRNA that stays in the nucleus in close association with the inactive X chromosome.

Additional aspects and consequences of X chromosome inactivation will be discussed in Chapter 6, in the context of individuals with structurally abnormal X chromosomes or an abnormal number of X chromosomes, and in Chapter 7, in the case of females carrying deleterious mutant alleles for X-linked disease.

VARIATION IN GENE EXPRESSION AND ITS RELEVANCE TO MEDICINE

The regulated expression of genes in the human genome involves a set of complex interrelationships among different levels of control, including proper gene dosage (controlled by mechanisms of chromosome replication and segregation), gene structure, chromatin packaging and epigenetic regulation, transcription, RNA splicing, and, for protein-coding loci, mRNA stability, translation, protein processing, and protein degradation. For some genes, fluctuations in the level of functional gene product, due either to inherited variation in the structure of a particular gene or to changes induced by nongenetic factors such as diet or the environment, are of relatively little importance. For other genes, even relatively minor changes in the level of expression can have dire clinical consequences, reflecting the importance of those gene products in particular biological pathways. The nature of inherited variation in the structure and function of chromosomes, genes, and the genome, combined with the influence of this variation on the expression of specific traits, is the very essence of medical and molecular genetics and is dealt with in subsequent chapters.

GENERAL REFERENCES

Brown TA: *Genomes*, ed 3, New York, 2007, Garland Science.
Lodish H, Berk A, Kaiser CA, et al: *Molecular cell biology*, ed 7, New York, 2012, WH Freeman.
Strachan T, Read A: *Human molecular genetics*, ed 4, New York, 2010, Garland Science.

REFERENCES FOR SPECIFIC TOPICS

Bartolomei MS, Ferguson-Smith AC: Mammalian genomic imprinting, *Cold Spring Harbor Perspect Biol* 3:1002592, 2011.
Beck CR, Garcia-Perez JL, Badge RM, et al: LINE-1 elements in structural variation and disease, *Annu Rev Genomics Hum Genet* 12:187–215, 2011.
Berg P: Dissections and reconstructions of genes and chromosomes (Nobel Prize lecture), *Science* 213:296–303, 1981.
Chess A: Mechanisms and consequences of widespread random monoallelic expression, *Nat Rev Genet* 13:421–428, 2012.
Dekker J: Gene regulation in the third dimension, *Science* 319:1793–1794, 2008.
Djebali S, Davis CA, Merkel A, et al: Landscape of transcription in human cells, *Nature* 489:101–108, 2012.
ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome, *Nature* 489:57–74, 2012.
Gerstein MB, Bruce C, Rozowsky JS, et al: What is a gene, post-ENCODE? *Genome Res* 17:669–681, 2007.
Guil S, Esteller M: Cis-acting noncoding RNAs: friends and foes, *Nat Struct Mol Biol* 19:1068–1074, 2012.
Heyn H, Esteller M: DNA methylation profiling in the clinic: applications and challenges, *Nature Rev Genet* 13:679–692, 2012.
Hubner MR, Spector DL: Chromatin dynamics, *Annu Rev Biophys* 39:471–489, 2010.
Li M, Wang IX, Li Y, et al: Widespread RNA and DNA sequence differences in the human transcriptome, *Science* 333:53–58, 2011.
Nagano T, Fraser P: No-nonsense functions for long noncoding RNAs, *Cell* 145:178–181, 2011.
Willard HF: The human genome: a window on human genetics, biology and medicine. In Ginsburg GS, Willard HF, editors: *Genomic and personalized medicine*, ed 2, New York, 2013, Elsevier.
Zhou VW, Goren A, Bernstein BE: Charting histone modifications and the functional organization of mammalian genomes, *Nat Rev Genet* 12:7–18, 2012.

PROBLEMS

1. The following amino acid sequence represents part of a protein. The normal sequence and four mutant forms are shown. By consulting Table 3-1, determine the double-stranded sequence of the corresponding section of the normal gene. Which strand is the strand that RNA polymerase "reads"? What would the sequence of the resulting mRNA be? What kind of mutation is each mutant protein most likely to represent?

Normal	-lys-arg-his-his-tyr-leu-
Mutant 1	-lys-arg-his-his-cys-leu-
Mutant 2	-lys-arg-ile-ile-ile-
Mutant 3	-lys-glu-thr-ser-leu-ser-
Mutant 4	-asn-tyr-leu-

2. The following items are related to each other in a hierarchical fashion: chromosome, base pair, nucleosome, kilobase pair, intron, gene, exon, chromatin, codon, nucleotide, promoter. What are these relationships?

3. Describe how mutation in each of the following might be expected to alter or interfere with normal gene function and thus cause human disease: promoter, initiator codon, splice sites at intron-exon junctions, one base pair deletion in the coding sequence, stop codon.

4. Most of the human genome consists of sequences that are not transcribed and do not directly encode gene products. For each of the following, consider ways in which these genome elements might contribute to human disease: introns, *Alu* or LINE repetitive sequences, locus control regions, pseudogenes.

5. Contrast the mechanisms and consequences of RNA splicing and somatic rearrangement.

6. Consider different ways in which mutations or variation in the following might lead to human disease: epigenetic modifications, DNA methylation, miRNA genes, lncRNA genes.

7. Contrast the mechanisms and consequences of genomic imprinting and X chromosome inactivation.