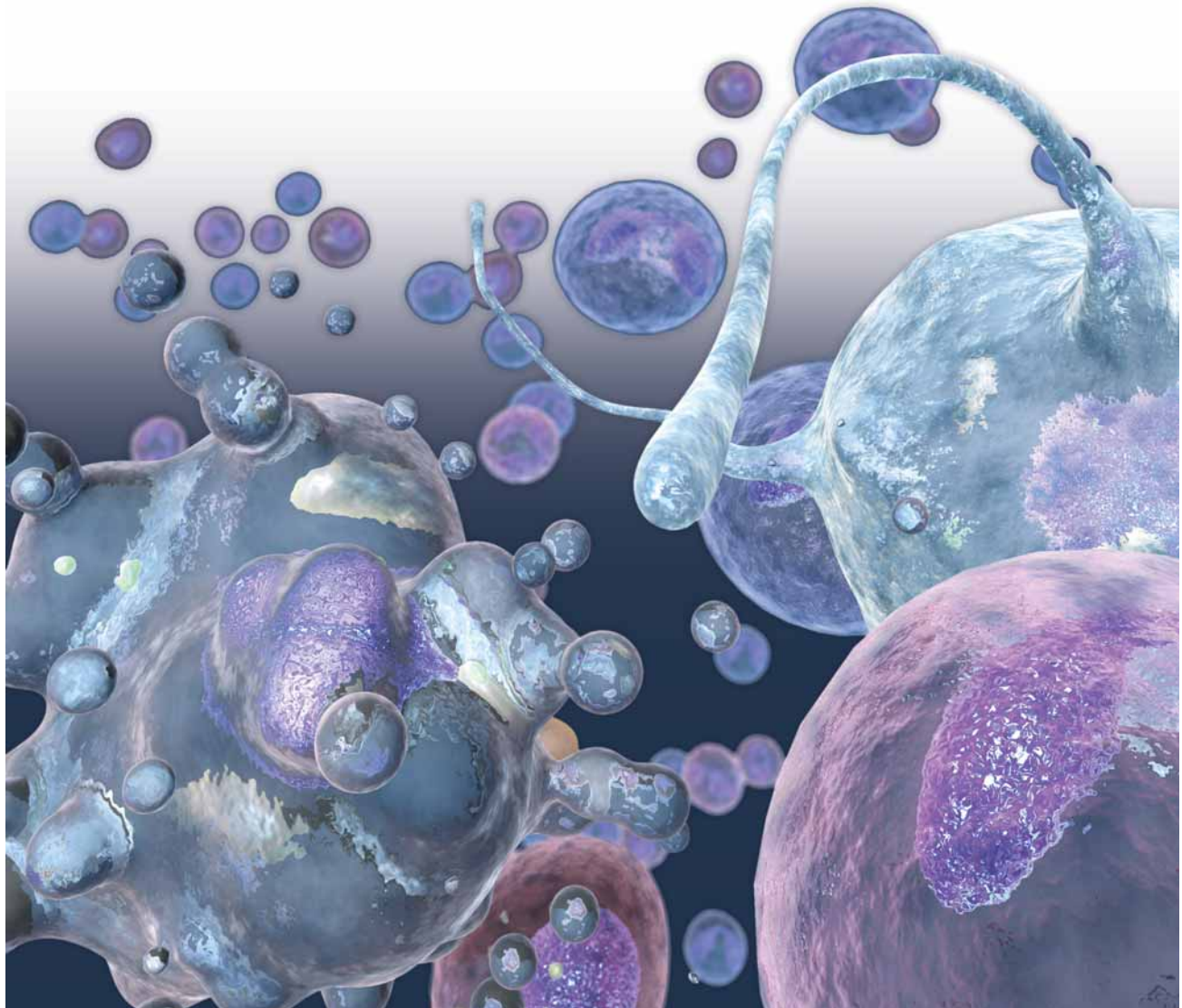


PART

//

The Flow of Genetic Information



CHAPTER 5 ■ The Organization and Sequences of Cellular Genomes

CHAPTER 6 ■ Replication, Maintenance, and Rearrangements of Genomic DNA

CHAPTER 7 ■ RNA Synthesis and Processing

CHAPTER 8 ■ Protein Synthesis, Processing, and Regulation

CHAPTER 5

The Organization and Sequences of Cellular Genomes

- **The Complexity of Eukaryotic Genomes** 155
- **Chromosomes and Chromatin** 166
- **The Sequences of Complete Genomes** 176
- **Bioinformatics and Systems Biology** 192
- **KEY EXPERIMENT:**
The Discovery of Introns 158
- **KEY EXPERIMENT:**
The Human Genome 188

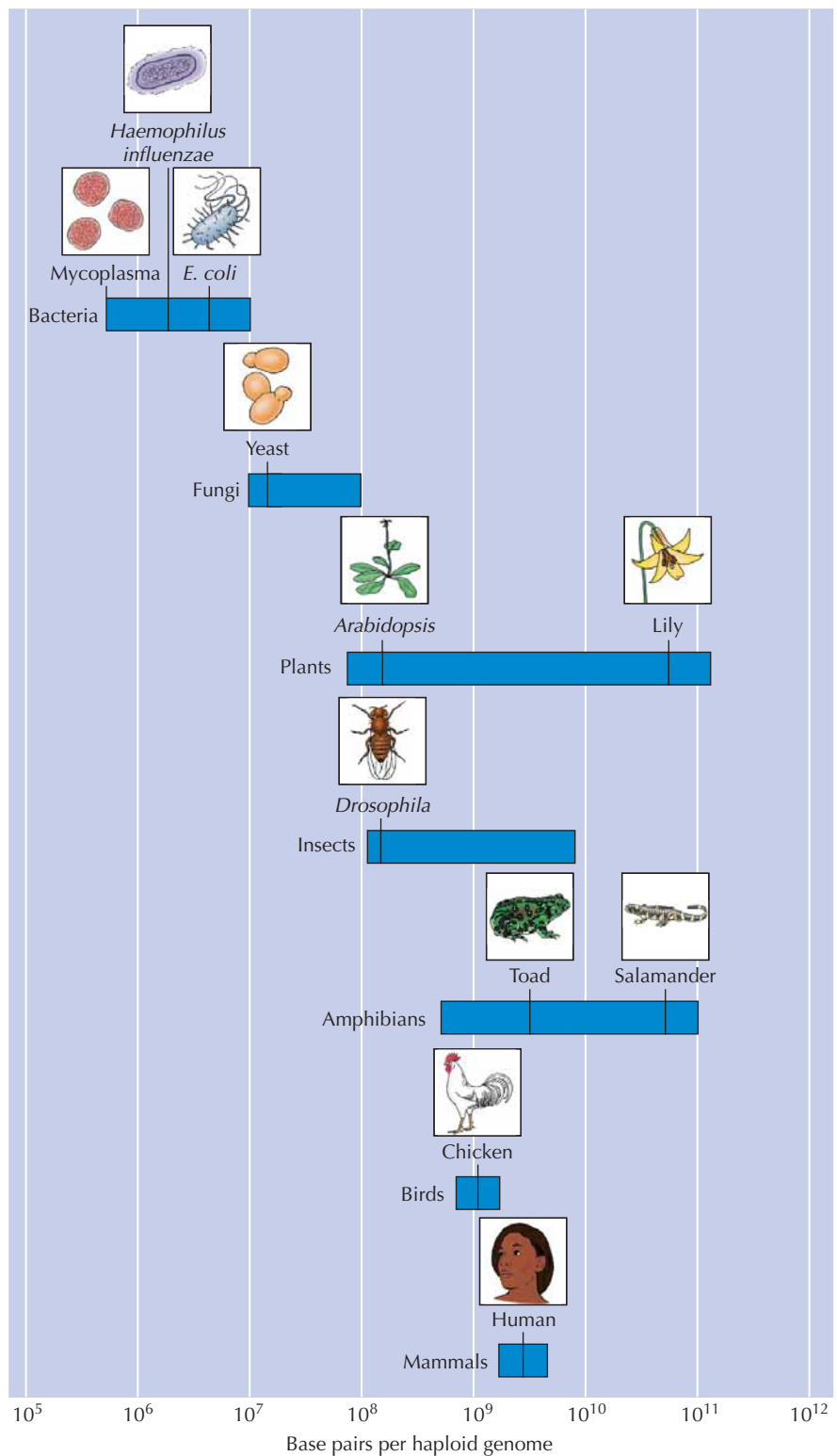
AS THE GENETIC MATERIAL, DNA PROVIDES A BLUEPRINT that directs all cellular activities and specifies the developmental plan of multicellular organisms. An understanding of gene structure and function is therefore fundamental to an appreciation of the molecular biology of cells. The development of gene cloning represented a major step toward this goal, enabling scientists to dissect complex eukaryotic genomes and probe the functions of eukaryotic genes. Continuing advances in recombinant DNA technology have now brought us to the exciting point of determining the sequences of entire genomes, providing a new approach to deciphering the genetic basis of cell behavior.

As reviewed in Chapter 4, the initial applications of recombinant DNA were directed toward the isolation and analysis of individual genes. More recently, large scale sequencing projects have yielded the complete genome sequences of many bacteria, of yeast, and of several species of plants and animals, including humans. The sequences of these complete cellular genomes provide a rich harvest of information, enabling the identification of many hitherto unknown genes and regulatory sequences. The results of these genome sequencing projects can be expected to stimulate many years of future research in molecular and cellular biology, and to have a profound impact on our understanding and treatment of human disease.

The Complexity of Eukaryotic Genomes

The genomes of most eukaryotes are larger and more complex than those of prokaryotes (Figure 5.1). This larger size of eukaryotic genomes is not inherently surprising, since one would expect to find more genes in organisms that are more complex. However, the genome size of many eukaryotes does not appear to be related to genetic complexity. For example, the genomes of salamanders and lilies contain more than ten times the

FIGURE 5.1 Genome size The range of sizes of the genomes of representative groups of organisms is shown on a logarithmic scale.



amount of DNA that is in the human genome, yet these organisms are clearly not ten times more complex than humans.

This apparent paradox was resolved by the discovery that the genomes of most eukaryotic cells contain not only functional genes but also large amounts of DNA sequences that do not code for proteins. The difference in the sizes of the salamander and human genomes thus reflects larger

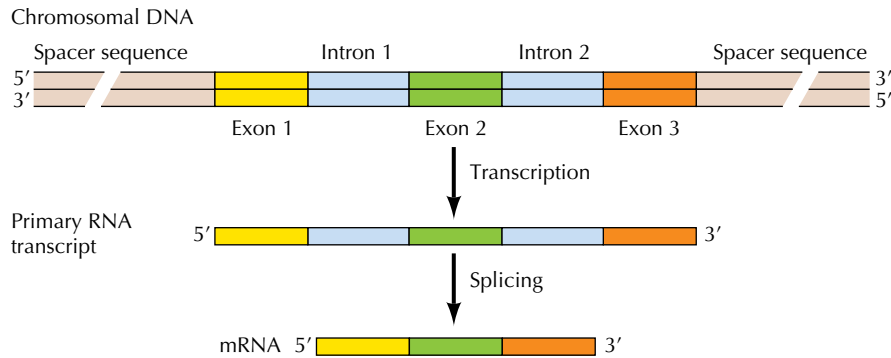


FIGURE 5.2 The structure of eukaryotic genes Most eukaryotic genes contain segments of coding sequences (exons) interrupted by noncoding sequences (introns). Both exons and introns are transcribed to yield a long primary RNA transcript. The introns are then removed by splicing to form the mature mRNA.

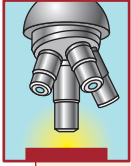
amounts of noncoding DNA, rather than more genes, in the genome of the salamander. The presence of large amounts of noncoding sequences is a general property of the genomes of complex eukaryotes. Thus the thousandfold greater size of the human genome compared to that of *E. coli* is not due solely to a larger number of human genes. The human genome is thought to contain 20,000–25,000 genes—only about 5 times more than *E. coli* has. Much of the complexity of eukaryotic genomes thus results from the abundance of several different types of noncoding sequences, which constitute most of the DNA of higher eukaryotic cells.

Introns and Exons

In molecular terms, a **gene** can be defined as a segment of DNA that is expressed to yield a functional product, which may be either an RNA (e.g., ribosomal and transfer RNAs) or a polypeptide. Some of the noncoding DNA in eukaryotes is accounted for by long DNA sequences that lie between genes (**spacer sequences**). However, large amounts of noncoding DNA are also found within most eukaryotic genes. Such genes have a split structure in which segments of coding sequence (called **exons**) are separated by noncoding sequences (intervening sequences, or **introns**) (Figure 5.2). The entire gene is transcribed to yield a long RNA molecule and the introns are then removed by splicing, so only exons are included in the mRNA. Although most introns have no known function, they account for a substantial fraction of DNA in the genomes of higher eukaryotes.

Introns were first discovered in 1977, independently in the laboratories of Phillip Sharp and Richard Roberts, during studies of the replication of adenovirus in cultured human cells. Adenovirus is a useful model for studies of gene expression, both because the viral genome is only about 3.5×10^4 base pairs long and because adenovirus mRNAs are produced at high levels in infected cells. One approach used to characterize the adenovirus mRNAs was to determine the locations of the corresponding viral genes by examination of RNA-DNA hybrids in the electron microscope. Because RNA-DNA hybrids are distinguishable from single-stranded DNA, the positions of RNA transcripts on a DNA molecule can be determined. Surprisingly, such experiments revealed that adenovirus mRNAs do not hybridize to only a single region of viral DNA (Figure 5.3). Instead, a single mRNA molecule hybridizes to several separated regions of the viral genome. Thus the adenovirus mRNA does not correspond to an uninterrupted transcript of the template DNA; rather the mRNA is assembled from several distinct blocks of sequences that originated from different parts of the viral DNA. This was subsequently shown to occur by **RNA splicing**, which will be discussed in detail in Chapter 7.

KEY EXPERIMENT

The Discovery of Introns**Spliced Segments at the 5' Terminus of Adenovirus 2 Late mRNA**

Susan M. Berget, Claire Moore, and Phillip A. Sharp
Massachusetts Institute of Technology, Cambridge, Massachusetts
Proceedings of the National Academy of Sciences USA, Volume 74, 1977,
 pages 3171–3175



Phillip Sharp



Richard Roberts

The Context

Prior to molecular cloning, little was known about mRNA synthesis in eukaryotic cells. However, it was clear that this process is more complex in eukaryotes than in bacteria. The synthesis of eukaryotic mRNAs appeared to require not only transcription but also processing reactions that modify the structure of primary transcripts. Most notably, eukaryotic mRNAs appeared to be synthesized as long primary transcripts, found in the nucleus, which were then cleaved to yield much shorter mRNA molecules that were exported to the cytoplasm.

These processing steps were generally assumed to involve the removal of sequences from the 5' and 3' ends of the primary transcripts. In this model, mRNAs embedded within long primary transcripts would be encoded by uninterrupted DNA sequences. This view of eukaryotic mRNA was changed radically by the discovery of splicing, made independently by Berget, Moore, and Sharp, and by Louise Chow, Richard Gelinas, Tom Broker, and Richard Roberts (An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA, 1977. *Cell* 12: 1–8).

The Experiments

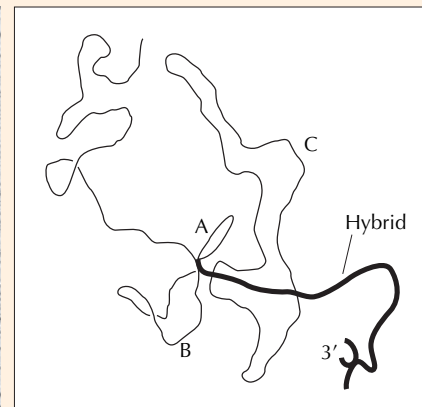
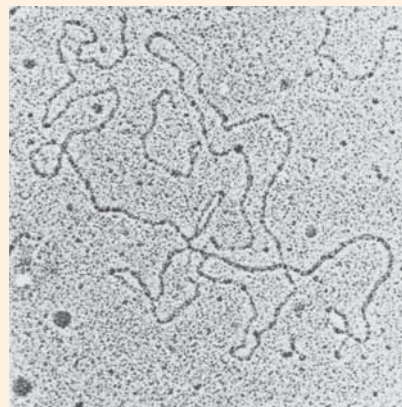
Both of the research groups that discovered splicing used adenovirus 2 to investigate mRNA synthesis in human cells. The major advantage of the virus is that it provides a model that is much simpler than the host cell. Viral DNA can be isolated directly from virus particles, and mRNAs encoding the viral structural proteins are present in such high

amounts that they can be purified directly from infected cells. Berget, Moore, and Sharp focused their experiments on an abundant mRNA that encodes a viral structural polypeptide known as the hexon.

To map the hexon mRNA on the viral genome, purified mRNA was hybridized to adenovirus DNA and the hybrid molecules were examined by electron microscopy. As expected, the body of the hexon mRNA formed hybrids with restriction fragments of adenovirus DNA that had previously been shown to contain the hexon gene. Surprisingly, however, sequences at the 5' end of hexon mRNA failed to hybridize to DNA sequences adjacent to those encoding the body of the message, suggesting that the 5' end of the mRNA had arisen from sequences located elsewhere in the viral genome.

This possibility was tested by hybridization of hexon mRNA to a restriction fragment extending

upstream of the hexon gene. The mRNA-DNA hybrids formed in this experiment displayed a complex loop structure (see figure). The body of the mRNA formed a long hybrid region with the previously identified hexon DNA sequences. Strikingly, the 5' end of the hexon mRNA hybridized to three short upstream regions of DNA, which were separated from each other and from the body of the message by large single-stranded DNA loops. The sequences at the 5' end of hexon mRNA thus appeared to be transcribed from three separate regions of the viral genome, which were spliced to the body of the mRNA during the processing of a long primary transcript.



An electron micrograph and tracing of hexon mRNA hybridized to adenovirus DNA. The single-stranded loops designated A, B, and C, correspond to introns.

KEY EXPERIMENT

The Impact

The discovery of splicing in adenovirus mRNA was quickly followed by similar experiments with cellular mRNAs, demonstrating that eukaryotic genes had a previously unexpected structure. Rather than being continuous, their coding sequences were interrupted by introns, which were removed from primary transcripts by splicing. Introns are now

known to account for much of the DNA in eukaryotic genomes, and the roles of introns in the evolution and regulation of gene expression continue to be active areas of investigation. The discovery of splicing also stimulated intense interest in the mechanism of this unexpected RNA processing reaction. As discussed in Chapter 7, these studies have not only illuminated new mechanisms of regulating gene expres-

sion; they have also revealed novel catalytic activities of RNA and provided critical evidence supporting the hypothesis that early evolution was based on self-replicating RNA molecules. The unexpected structure of adenovirus mRNAs has thus had a major impact on diverse areas of cellular and molecular biology.

Soon after the discovery of introns in adenovirus, similar observations were made on cloned genes of eukaryotic cells. For example, electron microscopic analysis of RNA-DNA hybrids and subsequent nucleotide sequencing of cloned genomic DNAs and cDNAs indicated that the coding region of the mouse β -globin gene (which encodes the β subunit of hemoglobin) is interrupted by two introns that are removed from the mRNA by splicing (Figure 5.4). The intron-exon structure of many eukaryotic genes is quite complicated, and the amount of DNA in the intron sequences is often

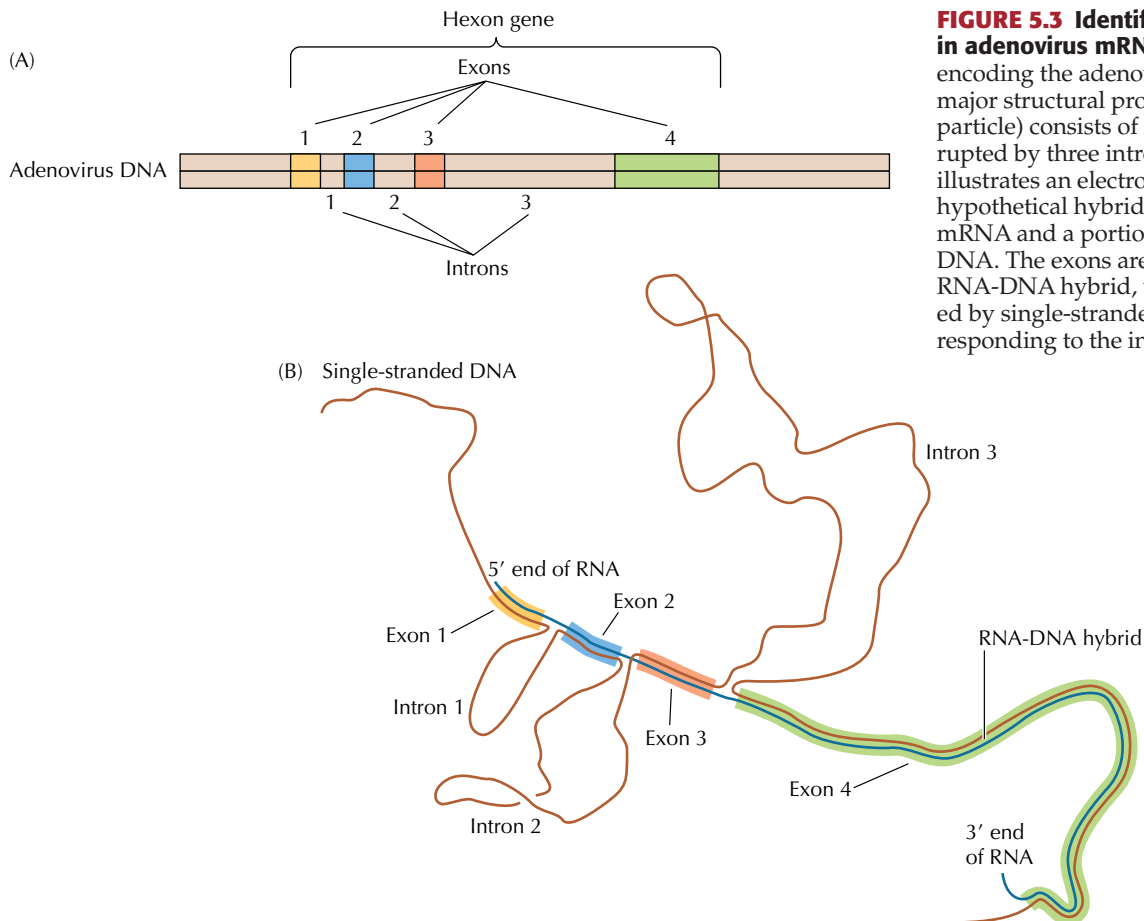
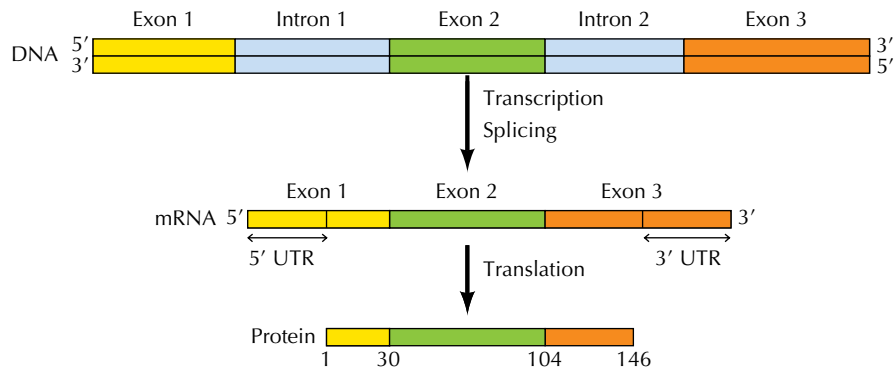


FIGURE 5.3 Identification of introns in adenovirus mRNA (A) The gene encoding the adenovirus hexon (a major structural protein of the viral particle) consists of four exons, interrupted by three introns. (B) This tracing illustrates an electron micrograph of a hypothetical hybrid between hexon mRNA and a portion of adenovirus DNA. The exons are seen as regions of RNA-DNA hybrid, which are separated by single-stranded DNA loops corresponding to the introns.

FIGURE 5.4 The mouse β -globin gene

This gene contains two introns, which divide the coding region among three exons. Exon 1 encodes amino acids 1 to 30, exon 2 encodes amino acids 31 to 104, and exon 3 encodes amino acids 105 to 146. Exons 1 and 3 also contain untranslated regions (UTRs) at the 5' and 3' ends of the mRNA, respectively.



greater than that in the exons. For example, an average human gene contains approximately 9 exons, interrupted by 8 introns and distributed over approximately 30,000 base pairs (30 **kilobases**, or **kb**) of genomic DNA (Table 5.1). The exons generally total only about 2.5 kb, including regions at both the 5' and 3' ends of the mRNA that are not translated into protein (5' and 3' untranslated regions or UTRs). Introns thus comprise more than 90% of the average human gene.

Introns are present in most genes of complex eukaryotes, although they are not universal. Almost all histone genes, for example, lack introns, so introns are clearly not required for gene function in eukaryotic cells. In addition, introns are not found in most genes of simple eukaryotes, such as yeasts. Conversely, introns *are* present in rare genes of prokaryotes. The presence or absence of introns is therefore not an absolute distinction between prokaryotic and eukaryotic genes, although introns are much more prevalent in higher eukaryotes (both plants and animals), where they account for a substantial amount of total genomic DNA. Many introns are conserved in genes of both plants and animals, indicating that they arose early in evolution, prior to the plant-animal divergence.

Most introns do not specify the synthesis of a cellular product, although a few do encode functional RNAs or proteins. However, introns play important roles in controlling gene expression. For example, the presence of introns allows the exons of a gene to be joined in different combinations, resulting in the synthesis of different proteins from the same gene. This process, called **alternative splicing** (Figure 5.5), occurs frequently in the genes of complex eukaryotes and is thought to be important in extending the functional repertoire of the 20,000–25,000 genes of the human genome.

Introns are also thought to have played an important role in evolution by facilitating recombination between protein-coding regions (exons) of differ-

TABLE 5.1 Characteristics of the Average Human Gene

Number of exons	9
Number of introns	8
Exon Sequence:	
5' untranslated region	300 base pairs
coding sequence	1400 base pairs
3' untranslated region	800 base pairs
TOTAL	2500 base pairs
Intron Sequence:	
	27,000 base pairs

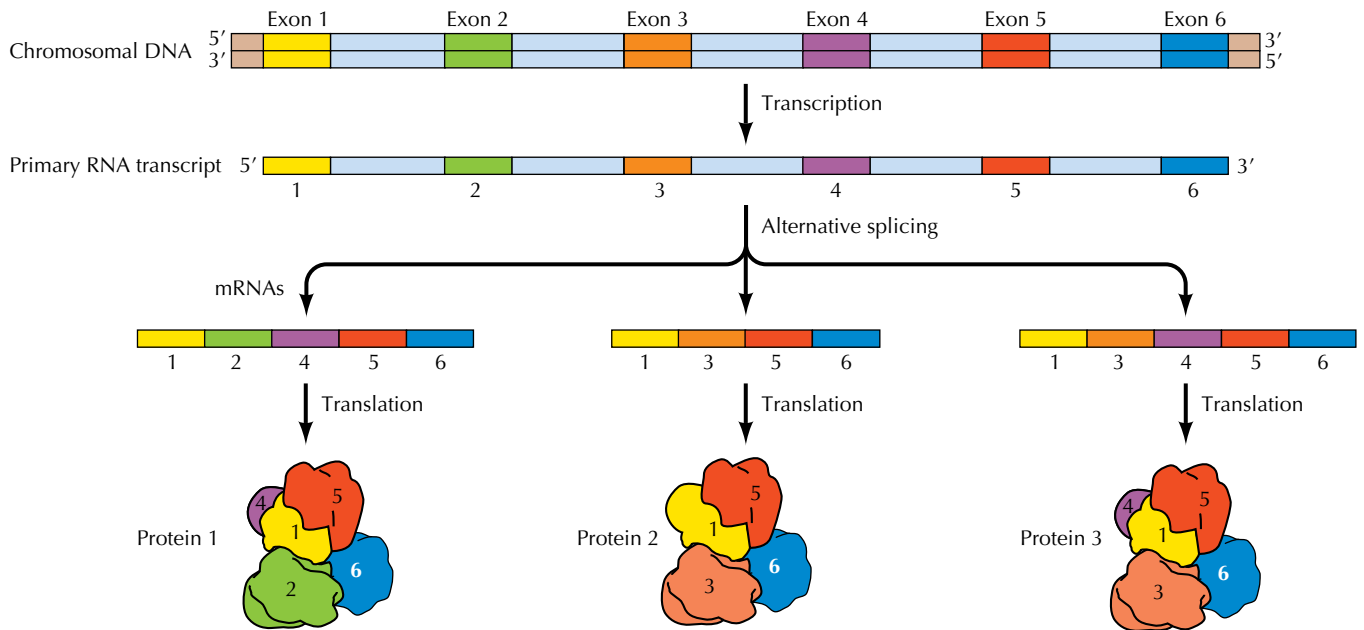


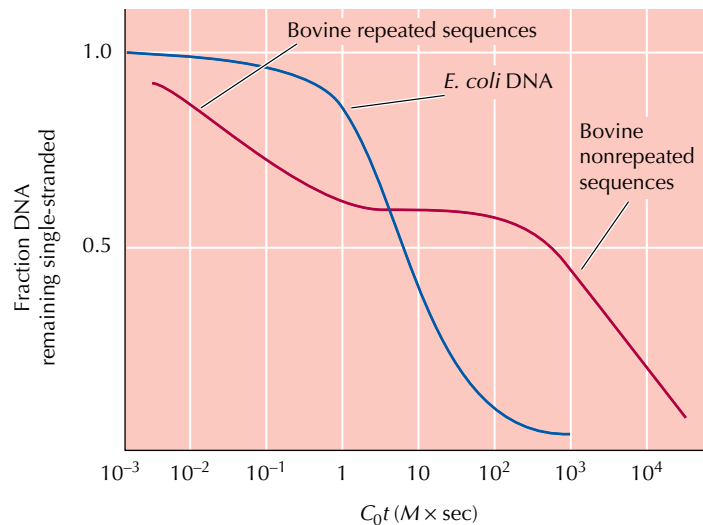
FIGURE 5.5 Alternative splicing The gene illustrated contains six exons, separated by five introns. Alternative splicing allows these exons to be joined in different combinations, resulting in the formation of three distinct mRNAs and proteins from the single primary transcript.

ent genes—a process known as exon shuffling. Exons frequently encode functionally distinct protein domains, so recombination between introns of different genes would result in new genes containing novel combinations of protein-coding sequences. As predicted by this hypothesis, DNA sequencing studies have demonstrated that some genes are chimeras of exons derived from several other genes, providing direct evidence that new genes can be formed by recombination between intron sequences.

Repetitive DNA Sequences

Introns make a substantial contribution to the large size of higher eukaryotic genomes. In humans, for example, introns account for approximately 20% of the total genomic DNA. However, an even larger portion of complex eukaryotic genomes consists of highly repeated noncoding DNA sequences. These sequences, sometimes present in hundreds of thousands of copies per genome, were first demonstrated by Roy Britten and David Kohne during studies of the rates of reassociation of denatured fragments of cellular DNAs (Figure 5.6). Denatured strands of DNA hybridize to each other (reassociate), re-forming double-stranded molecules (see Figure 4.24). Since DNA reassociation is a bimolecular reaction (two separated strands of denatured DNA must collide with each other in order to hybridize), the rate of reassociation depends on the concentration of DNA strands. When fragments of *E. coli* DNA were denatured and allowed to hybridize with each other, all of the DNA reassociated at the same rate, as expected if each DNA sequence were represented once per genome. However, reassociation of fragments of DNA extracted from mammalian cells showed a very different pattern. Approximately 50% of the DNA fragments reassociated at the rate expected for sequences present once per genome, but the remainder reasso-

FIGURE 5.6 Identification of repetitive sequences by DNA reassociation The kinetics of the reassociation of fragments of *E. coli* and bovine DNAs are illustrated as a function of C_0t , which is the initial concentration of DNA multiplied by the time of incubation. The *E. coli* DNA reassociates at a uniform rate, consistent with each fragment of DNA being represented once in a genome of 4.6×10^6 base pairs. In contrast, the bovine DNA fragments exhibit two distinct steps in their reassociation. About 60% of the DNA fragments (the nonrepeated sequences) reassociate more slowly than *E. coli* DNA, as expected for sequences represented as single copies in the larger bovine genome (3×10^9 base pairs). However, the other 40% of the bovine DNA fragments (the repeated sequences) reassociate more rapidly than *E. coli* DNA, indicating that multiple copies of these sequences are present.



ciated much more rapidly than expected. The interpretation of these results was that some sequences were present in multiple copies and therefore reassociated more rapidly than those sequences that were represented only once per genome. In particular, these experiments indicated that approximately 50% of mammalian DNA consists of highly repetitive sequences, some of which are repeated 10^5 to 10^6 times.

Further analysis, culminating in the sequencing of complete genomes, has identified several types of these highly repeated sequences (Table 5.2). One class (called **simple-sequence repeats**) consists of tandem arrays of up to thousands of copies of short sequences, ranging from 1 to 500 nucleotides. For example, one type of simple-sequence repeat in *Drosophila* consists of tandem repeats of the seven nucleotide unit ACAAAC. Because of their distinct base compositions, many simple-sequence DNAs can be separated from the rest of the genomic DNA by equilibrium centrifugation in CsCl density gradients. The density of DNA is determined by its base composition, with AT-rich sequences being less dense than GC-rich sequences. Therefore an AT-rich simple-sequence DNA bands in CsCl gradients at a lower density than the bulk of *Drosophila* genomic DNA (Figure 5.7). Since such repeat-sequence DNAs band as “satellites” separate from the main band of DNA, they are frequently referred to as **satellite DNAs**. These sequences are repeated millions of times per genome, accounting for about 10% of the DNA of most higher eukaryotes. Simple-sequence DNAs are not transcribed and do not convey functional genetic information.

TABLE 5.2 Repetitive Sequences in the Human Genome

Type of sequence	Number of copies	Fraction of genome
Simple-sequence repeats ^a	>1,000,000	~10%
Retrotransposons		
LINEs	850,000	21%
SINEs	1,500,000	13%
Retrovirus-like elements	450,000	8%
DNA transposons	300,000	3%

^aThe content of simple-sequence repeats is estimated from the fraction of heterochromatin in the human genome.

FIGURE 5.7 Satellite DNA Equilibrium centrifugation of *Drosophila* DNA in a CsCl gradient separates satellite DNAs (designated I–IV) with buoyant densities (in g/cm³) of 1.672, 1.687, and 1.705 from the main band of genomic DNA (buoyant density 1.701).

Some, however, play important roles in chromosome structure, as discussed in the next section of this chapter.

Other repetitive DNA sequences are scattered throughout the genome rather than being clustered as tandem repeats. These interspersed repetitive elements are a major contributor to genome size, accounting for approximately 45% of human genomic DNA. The two most prevalent classes of these sequences are called **SINEs** (short interspersed elements) and **LINES** (long interspersed elements). SINEs are 100–300 base pairs long. About 1.5 million such sequences are dispersed throughout the genome, accounting for approximately 13% of the total human DNA. Although SINEs are transcribed into RNA, they do not encode proteins and their function is unknown. The major human LINES are 4–6 kb long, although many repeated sequences derived from LINES are shorter, with an average size of about 1 kb. There are approximately 850,000 repeats of LINE sequences in the genome, accounting for about 21% of human DNA. LINES are transcribed and at least some encode proteins, but like SINEs, they have no known function in cell physiology.

Both SINEs and LINES are examples of transposable elements, which are capable of moving to different sites in genomic DNA. As discussed in detail in Chapter 6, both SINEs and LINES are **retrotransposons**, meaning that their transposition is mediated by reverse transcription (Figure 5.8). An RNA copy of a SINE or LINE is converted to DNA by reverse transcriptase within the cell, and the new DNA copy is integrated at a new site in the genome. A third class of interspersed repetitive sequences, which closely resemble retroviruses and are called **retrovirus-like elements**, also move within the genome by reverse transcription. Human retrovirus-like elements range from approximately 2–10 kb in length. There are approximately 450,000 retrovirus-like elements in the human genome, accounting for approximately 8% of human DNA. In contrast, the fourth class of inter-

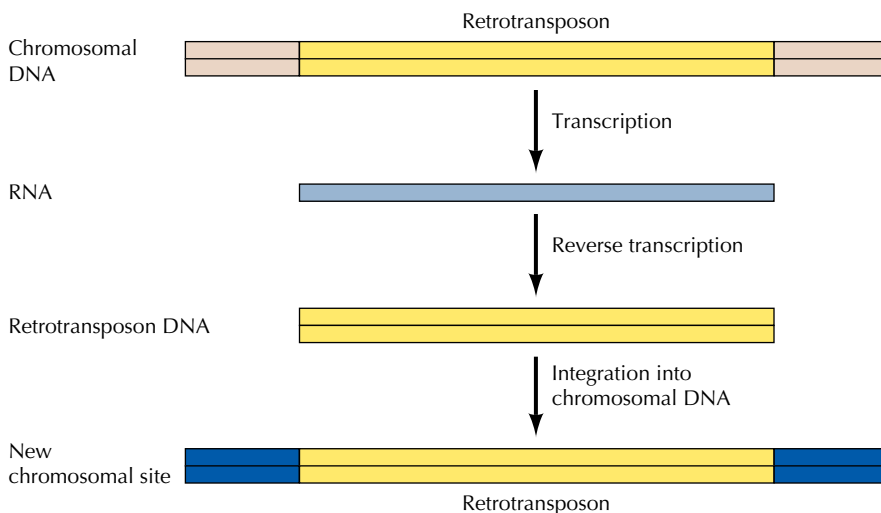
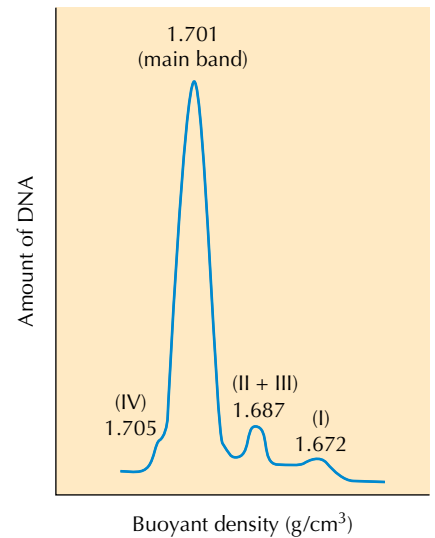


FIGURE 5.8 Movement of retrotransposons A retrotransposon present at one site in chromosomal DNA is transcribed into RNA, and then converted back into DNA by reverse transcription. The retrotransposon DNA can then integrate into a new chromosomal site.

scattered repetitive elements (**DNA transposons**) moves through the genome by being copied and reinserted as DNA sequences, rather than moving by reverse transcription. In the human genome, there are about 300,000 copies of DNA transposons, ranging from 80–3000 base pairs in length and accounting for approximately 3% of human DNA.

Nearly half of the human genome thus consists of interspersed repetitive elements that have replicated and moved through the genome by either RNA or DNA intermediates. It is noteworthy that the vast majority of these elements transpose via RNA intermediates, so reverse transcription has been responsible for generating more than 40% of the human genome. Some of these sequences may help regulate gene expression, but most interspersed repetitive sequences appear not to make a useful contribution to the cell. Instead, they appear to represent “selfish DNA elements” that have been selected for their own ability to replicate within the genome rather than conferring a selective advantage to their host. In some cases, however, transposable elements have played important evolutionary roles by stimulating gene rearrangements and contributing to the generation of genetic diversity.

Gene Duplication and Pseudogenes

Another factor contributing to the large size of eukaryotic genomes is that many genes are present in multiple copies, some of which are frequently nonfunctional. In some cases, multiple copies of genes are needed to produce RNAs or proteins required in large quantities, such as ribosomal RNAs or histones. In other cases, distinct members of a group of related genes (called a **gene family**) may be transcribed in different tissues or at different stages of development. For example, the α and β subunits of hemoglobin are both encoded by gene families in the human genome, with different members of these families being expressed in embryonic, fetal, and adult tissues (Figure 5.9). Members of many gene families (e.g., the globin genes) are clustered within a region of DNA; members of other gene families are dispersed to different chromosomes.

Gene families are thought to have arisen by duplication of an original ancestral gene, with different members of the family then diverging as a consequence of mutations during evolution. Such divergence can lead to the evolution of related proteins that are optimized to function in different tissues or at different stages of development. For example, fetal globins

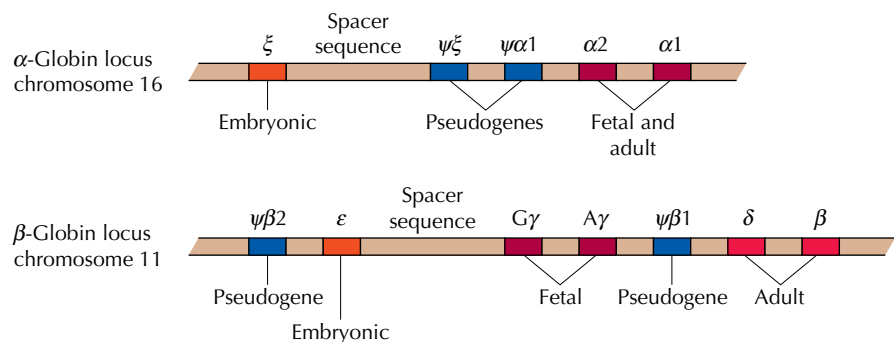


FIGURE 5.9 Globin gene families Members of the human α - and β -globin gene families are clustered on chromosomes 16 and 11, respectively. Each family contains genes that are specifically expressed in embryonic, fetal, and adult tissues, in addition to nonfunctional gene copies (pseudogenes).

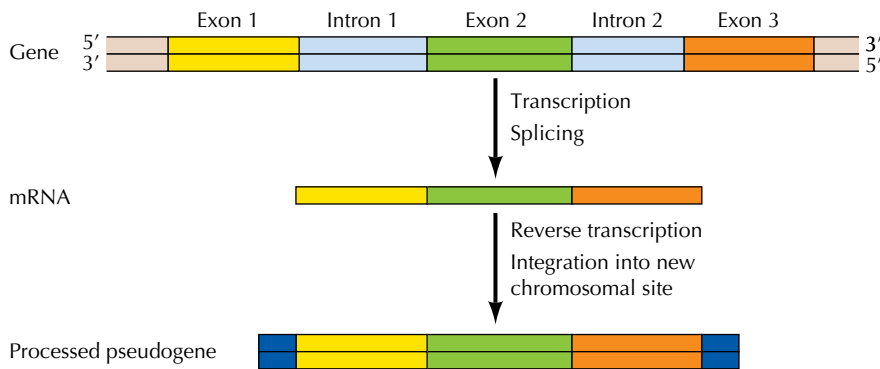


FIGURE 5.10 Formation of a processed pseudogene A gene is transcribed and spliced to yield an mRNA from which the introns have been removed. The mRNA is copied by reverse transcriptase, yielding a cDNA copy lacking introns. Integration into chromosomal DNA results in formation of a processed pseudogene.

have a higher affinity for O_2 than do adult globins—a difference that allows the fetus to obtain O_2 from the maternal circulation.

As might be expected, however, not all mutations enhance gene function. Some gene copies have instead sustained mutations that result in their loss of ability to produce a functional gene product. For example, the human α - and β -globin gene families each contain two genes that have been inactivated by mutations. Such nonfunctional gene copies (called **pseudogenes**) represent evolutionary relics that increase the size of eukaryotic genomes without making a functional genetic contribution. Recent studies have identified more than 20,000 pseudogenes in the human genome. Since this is generally assumed to be an underestimate, it is likely that our genome contains many more pseudogenes than functional genes.

Gene duplications can arise by two distinct mechanisms. The first is duplication of a segment of DNA, which can result in the transfer of a block of DNA sequence to a new location in the genome. Such duplications of DNA segments ranging from 1 kb to more than 50 kb are estimated to account for approximately 5% of the human genome. Alternatively, genes can be duplicated by reverse transcription of an mRNA, followed by integration of the cDNA copy into a new chromosomal site (Figure 5.10). This mode of gene duplication, analogous to the transposition of repetitive elements that move via RNA intermediates, results in the formation of gene copies that lack introns and also lack the normal chromosomal sequences that direct transcription of the gene into mRNA. As a result, duplication of a gene by reverse transcription usually yields an inactive gene copy called a **processed pseudogene**. Processed pseudogenes account for about two-thirds of the pseudogenes that have been identified in the human genome.

The Composition of Higher Eukaryotic Genomes

Having discussed several kinds of noncoding DNA that contribute to the genomic complexity of higher eukaryotes, it is of interest to overview the composition of cell genomes. In bacterial genomes, most of the DNA encodes proteins. For example, the genome of *E. coli* is approximately 4.6×10^6 base pairs long and contains about 4000 genes, with nearly 90% of the DNA used as protein-coding sequence. The yeast genome, which consists of 12×10^6 base pairs, is about 2.5 times the size of the genome of *E. coli*, but is still extremely compact. Only 4% of the genes of *Saccharomyces cerevisiae* contain introns, and these usually have only a single small intron near the start of the coding sequence. Approximately 70% of the yeast genome is used as protein-coding sequence, specifying a total of about 6000 proteins.

The relatively simple animal genomes of *C. elegans* and *Drosophila* are about 10 times larger than the yeast genome, but contain only 2–3 times more genes. Instead, these simple animal genomes contain more introns and more repetitive sequence, so that protein-coding sequences correspond to only about 25% of the *C. elegans* genome and about 13% of the genome of *Drosophila*. The genome of the model plant *Arabidopsis* contains a similar number of genes, with approximately 26% of the genome corresponding to protein-coding sequence.

The genomes of higher animals (such as humans) are approximately 20–30 times larger than those of *C. elegans* and *Drosophila*. However, a major surprise from deciphering the human genome sequence was the discovery that the human genome contains only 20,000 to 25,000 genes. It appears that only about 1.2% of the human genome consists of protein-coding sequence. Approximately 20% of the genome consists of introns, and more than 60% is composed of various types of repetitive and duplicated DNA sequences, with the remainder corresponding to pseudogenes, to nonrepetitive spacer sequences between genes, and to exon sequences that are present at the 5' and 3' ends of mRNAs but are not translated into protein. The increased size of the genomes of higher eukaryotes is thus due far more to the presence of large amounts of repetitive sequences and introns than to an increased number of genes.

Chromosomes and Chromatin

Not only are the genomes of most eukaryotes much more complex than those of prokaryotes, but the DNA of eukaryotic cells is also organized differently from that of prokaryotic cells. The genomes of prokaryotes are contained in single chromosomes, which are usually circular DNA molecules. In contrast, the genomes of eukaryotes are composed of multiple chromosomes, each containing a linear molecule of DNA. Although the numbers and sizes of chromosomes vary considerably between different species (Table 5.3), their basic structure is the same in all eukaryotes. The DNA of eukaryotic cells is tightly bound to small basic proteins (histones) that package the DNA in an orderly way in the cell nucleus. This task is substantial, given the DNA content of most eukaryotes. For example, the total extended length of DNA in a human cell is nearly 2 meters, but this DNA must fit into a nucleus with a diameter of only 5 to 10 μm .

Chromatin

The complexes between eukaryotic DNA and proteins are called **chromatin**, which typically contains about twice as much protein as DNA. The major proteins of chromatin are the **histones**—small proteins containing a high proportion of basic amino acids (arginine and lysine) that facilitate binding to the negatively charged DNA molecule. There are five major types of histones—called H1, H2A, H2B, H3, and H4—which are very similar among different species of eukaryotes (Table 5.4). The histones are extremely abundant proteins in eukaryotic cells; together their mass is approximately equal to that of the cell's DNA. In addition, chromatin contains an approximately equal mass of a wide variety of nonhistone chromosomal proteins. There are more than a thousand different types of these proteins, which are involved in a range of activities, including DNA replication and gene expression.

5.1

WEBSITE ANIMATION

Chromatin and Chromosomes

In a eukaryotic cell, DNA is wrapped tightly around histone proteins (forming chromatin), and when a cell prepares for division, the chromatin coils upon itself multiple times to form compact chromosomes.

TABLE 5.3 Chromosome Numbers of Eukaryotic Cells

Organism	Genome size (Mb) ^a	Chromosome number ^a
Yeast (<i>Saccharomyces cerevisiae</i>)	12	16
Slime mold (<i>Dictyostelium</i>)	70	7
<i>Arabidopsis thaliana</i>	125	5
Corn	5000	10
Onion	15,000	8
Lily	50,000	12
Nematode (<i>Caenorhabditis elegans</i>)	97	6
Fruit fly (<i>Drosophila</i>)	180	4
Toad (<i>Xenopus laevis</i>)	3000	18
Lungfish	50,000	17
Chicken	1200	39
Mouse	3000	20
Cow	3000	30
Dog	3000	39
Human	3000	23

^a Both genome size and chromosome number are for haploid cells.
Mb = millions of base pairs.

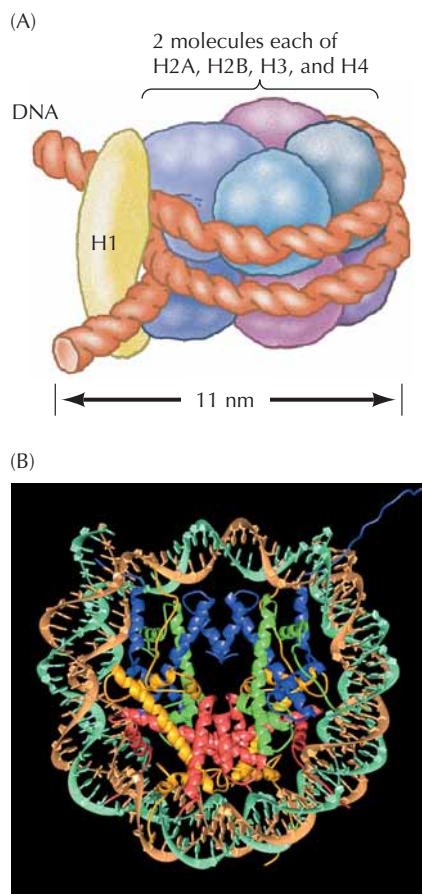
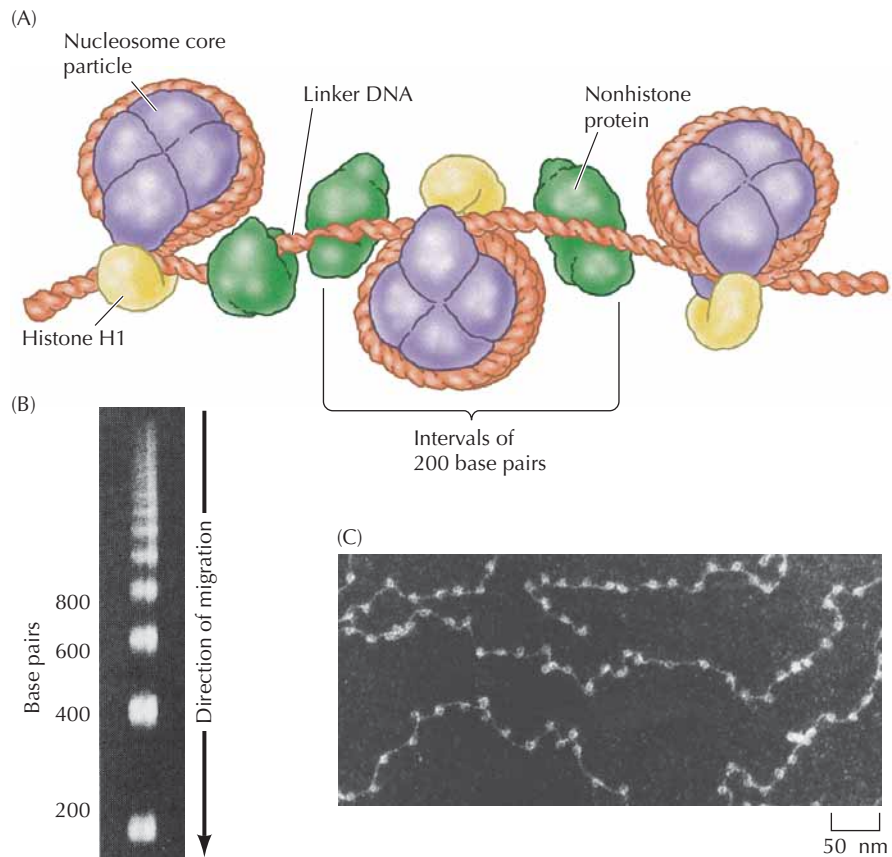
The basic structural unit of chromatin, the **nucleosome**, was described by Roger Kornberg in 1974 (Figure 5.11). Two types of experiments led to Kornberg's proposal of the nucleosome model. First, partial digestion of chromatin with micrococcal nuclease (an enzyme that degrades DNA) was found to yield DNA fragments approximately 200 base pairs long. In contrast, a similar digestion of naked DNA (not associated with proteins) yielded a continuous smear of randomly sized fragments. These results suggested that the binding of proteins to DNA in chromatin protects regions of the DNA from nuclease digestion, so that the enzyme can attack DNA only at sites separated by approximately 200 base pairs. Consistent with this notion, electron microscopy revealed that chromatin fibers have a beaded appearance, with the beads spaced at intervals of approximately 200 base pairs. Thus both the nuclease digestion and the electron microscopic studies suggested that chromatin is composed of repeating 200-base-pair units, which were called nucleosomes.

TABLE 5.4 The Major Histone Proteins

Histone ^a	Molecular weight	Number of amino acids	Percentage lysine + arginine
H1	22,500	244	30.8
H2A	13,960	129	20.2
H2B	13,774	125	22.4
H3	15,273	135	22.9
H4	11,236	102	24.5

^a Data are for rabbit (H1) and bovine histones.

FIGURE 5.11 The organization of chromatin in nucleosomes (A) The DNA is wrapped around histones in nucleosome core particles and sealed by histone H1. Nonhistone proteins bind to the linker DNA between nucleosome core particles. (B) Gel electrophoresis of DNA fragments obtained by partial digestion of chromatin with micrococcal nuclease. The linker DNA between the nucleosome core particles is preferentially sensitive, so limited digestion of chromatin yields fragments corresponding to multiples of 200 base pairs. (C) An electron micrograph of an extended chromatin fiber, illustrating its beaded appearance. (B, courtesy of Roger Kornberg, Stanford University; C, courtesy of Ada L. Olins and Donald E. Olins, Oak Ridge National Laboratory.)



More extensive digestion of chromatin with micrococcal nuclease was found to yield particles (called **nucleosome core particles**) that correspond to the beads visible by electron microscopy. Detailed analysis of these particles has shown that they contain 147 base pairs of DNA wrapped 1.67 times around a histone core consisting of two molecules each of H2A, H2B, H3, and H4 (the core histones) (Figure 5.12). One molecule of the fifth histone, H1, is bound to the DNA as it enters each nucleosome core particle. This forms a chromatin subunit known as a **chromatosome**, which consists of 166 base pairs of DNA wrapped around the histone core and held in place by H1 (a linker histone).

The packaging of DNA with histones yields a chromatin fiber approximately 10 nm in diameter that is composed of chromatosomes separated by linker DNA segments averaging about 50 base pairs in length (Figure 5.13). In the electron microscope, this 10-nm fiber has the beaded appearance that suggested the nucleosome model. Packaging of DNA into such a 10-nm chromatin fiber shortens its length approximately sixfold. The chromatin

FIGURE 5.12 Structure of a nucleosome (A) The nucleosome core particle consists of 147 base pairs of DNA wrapped 1.67 turns around a histone octamer consisting of two molecules each of H2A, H2B, H3, and H4. A chromatosome contains two full turns of DNA (166 base pairs) locked in place by one molecule of H1. (B) Model of the nucleosome core particle. The DNA backbones are shown in brown and turquoise. The histones are shown in blue (H3), green (H4), yellow (H2A), and red (H2B). (B, from K. Luger et al., 1997. *Nature* 389: 251.)

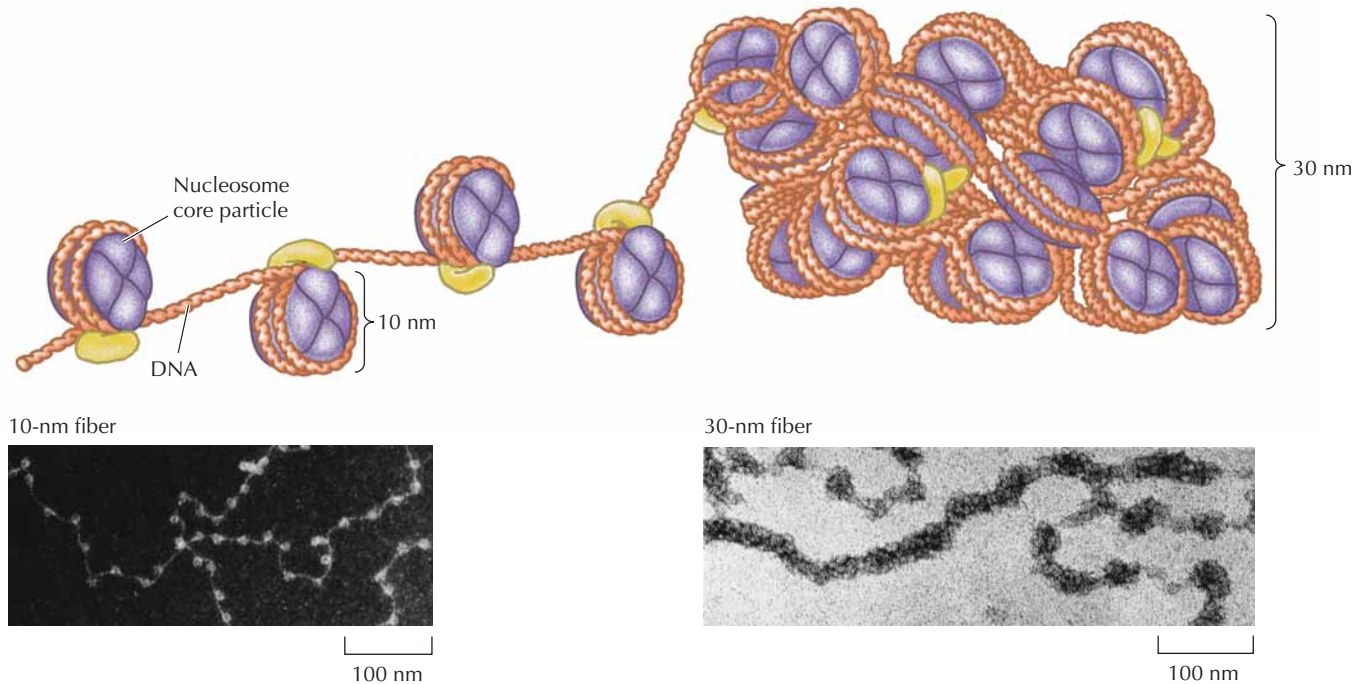
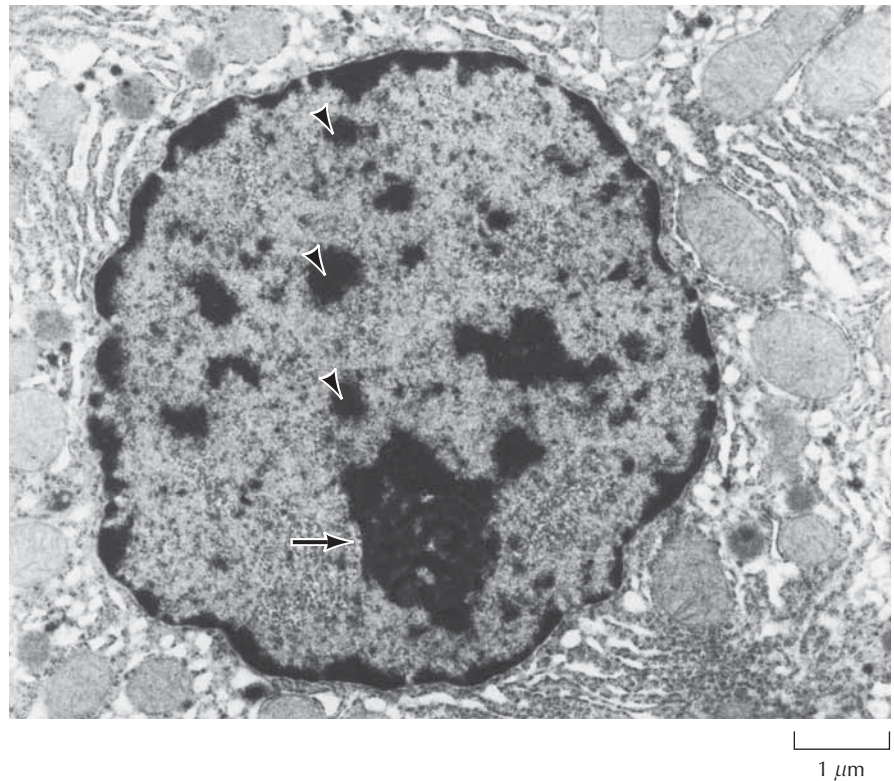


FIGURE 5.13 Chromatin fibers The packaging of DNA into nucleosomes yields a chromatin fiber approximately 10 nm in diameter. The chromatin is further condensed by coiling into a 30-nm fiber, containing about six nucleosomes per turn. (Photographs courtesy of Ada L. Olins and Donald E. Olins, Oak Ridge National Laboratory.)

can then be further condensed by coiling into 30-nm fibers, resulting in a total condensation of about fiftyfold. Interactions between histone H1 molecules appear to play an important role in this stage of chromatin condensation, which is critical to determining the accessibility of chromosomal DNA for processes such as DNA replication and transcription. Despite its importance, the structure of the 30-nm fiber remained unknown until 2005, when X-ray studies by Timothy Richmond and his colleagues revealed that the fiber is formed by two stacks of nucleosomes, with linker DNA zigzagging back and forth between them. Folding of 30-nm fibers upon themselves can lead to further condensation of chromatin within the cell.

The extent of chromatin condensation varies during the life cycle of the cell and plays an important role in regulating gene expression, as will be discussed in Chapter 7. In interphase (nondividing) cells, most of the chromatin (called **euchromatin**) is relatively decondensed and distributed throughout the nucleus (**Figure 5.14**). During this period of the cell cycle, genes are transcribed and the DNA is replicated in preparation for cell division. Most of the euchromatin in interphase nuclei appears to be in the form of 30-nm, or somewhat more condensed 60- to 130-nm, chromatin fibers. Genes that are actively transcribed are in a more decondensed state that makes the DNA accessible to the transcription machinery. In contrast to euchromatin, about 10% of interphase chromatin (called **heterochromatin**) is in a very highly condensed state that resembles the chromatin of cells undergoing mitosis. Heterochromatin is transcriptionally inactive and contains highly repeated DNA sequences, such as those present at centromeres and telomeres.

FIGURE 5.14 Interphase chromatin
Electron micrograph of an interphase nucleus. The euchromatin is distributed throughout the nucleus. The heterochromatin is indicated by arrowheads and the nucleolus by an arrow. (Courtesy of Ada L. Olins and Donald E. Olins, Oak Ridge National Laboratory.)



As cells enter mitosis, their chromosomes become highly condensed so that they can be distributed to daughter cells. Loops of 30-nm chromatin fibers are thought to fold upon themselves to form the compact metaphase chromosomes of mitotic cells in which the DNA has been condensed nearly ten thousandfold (Figure 5.15). Such condensed chromatin can no longer be used as a template for RNA synthesis, so transcription ceases during mitosis. Electron micrographs indicate that the DNA in metaphase chromosomes is organized into large loops attached to a protein scaffold (Figure 5.16), but we currently understand neither the detailed structure of this highly condensed chromatin nor the mechanism of chromatin condensation.

Metaphase chromosomes are so highly condensed that their morphology can be studied using the light microscope (Figure 5.17). Several staining techniques yield characteristic patterns of alternating light and dark chromosome bands, which result from the preferential binding of stains or fluores-

FIGURE 5.15 Chromatin condensation during mitosis Scanning electron micrograph of metaphase chromosomes. Artificial color has been added. (Biophoto Associates/Photo Researchers Inc.)

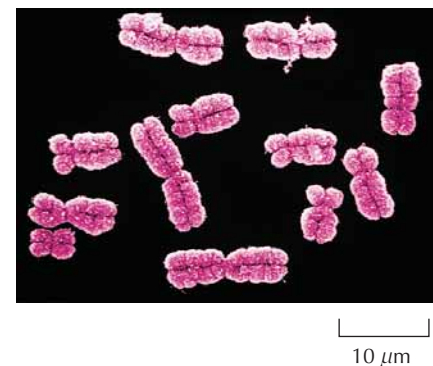


FIGURE 5.16 Structure of metaphase chromosomes An electron micrograph of DNA loops attached to the protein scaffold of metaphase chromosomes that have been depleted of histones. (From J. R. Paulson and U. K. Laemmli, 1977. *Cell* 12: 817.)

cent dyes to AT-rich versus GC-rich DNA sequences. These bands are specific for each chromosome and appear to represent distinct chromosome regions. Genes can be localized to specific chromosome bands by *in situ* hybridization, indicating that the packaging of DNA into metaphase chromosomes is a highly ordered and reproducible process.

Centromeres

The **centromere** is a specialized region of the chromosome that plays a critical role in ensuring the correct distribution of duplicated chromosomes to daughter cells during mitosis (Figure 5.18). The cellular DNA is replicated during interphase, resulting in the formation of two copies of each chromosome prior to the beginning of mitosis. As the cell enters mitosis, chromatin condensation leads to the formation of metaphase chromosomes consisting of two identical sister chromatids. These sister chromatids are held together at the centromere, which is seen as a constricted chromosomal region. As mitosis proceeds, microtubules of the mitotic spindle attach to the centromere, and the two sister chromatids separate and move to opposite poles of the spindle. At the end of mitosis, nuclear membranes re-form and the chromosomes decondense, resulting in the formation of daughter nuclei containing one copy of each parental chromosome.

The centromeres thus serve both as the sites of association of sister chromatids and as the attachment sites for microtubules of the mitotic spindle. They consist of specific DNA sequences to which a number of centromere-associated proteins bind, forming a specialized structure called the **kinetochore** (Figure 5.19). The binding of microtubules to kinetochore proteins

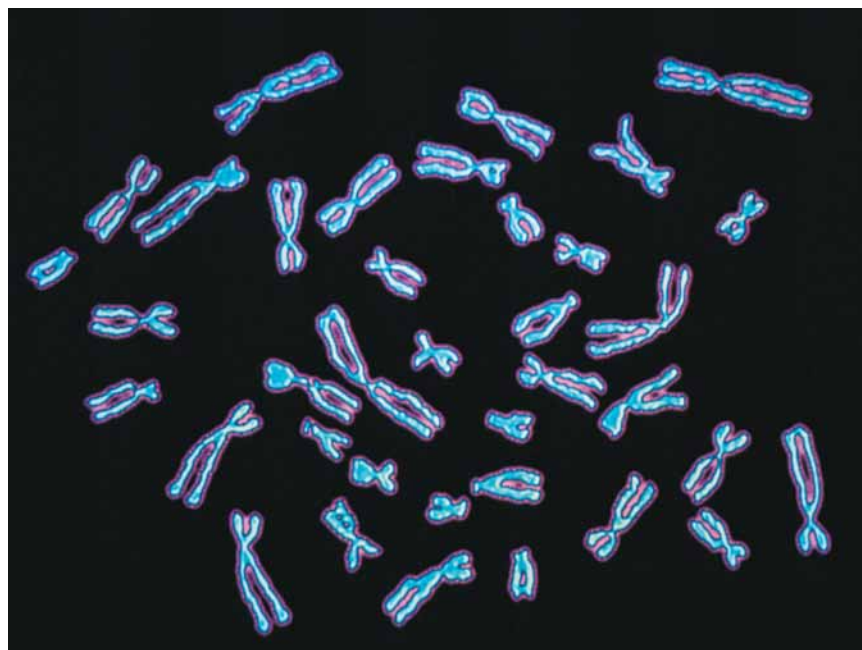
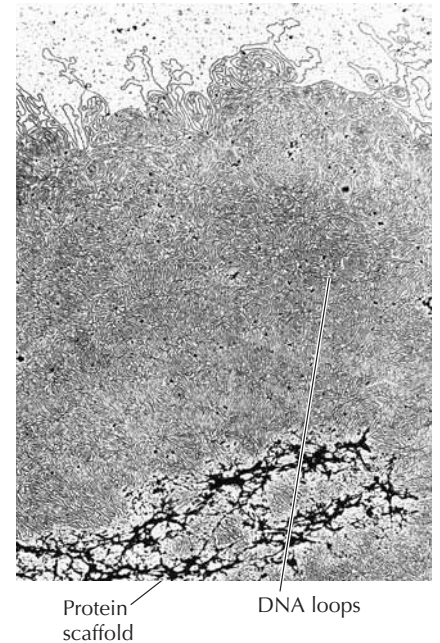


FIGURE 5.17 Human metaphase chromosomes A micrograph of human chromosomes spread from a metaphase cell. (Leonard Lessin/Peter Arnold, Inc.)

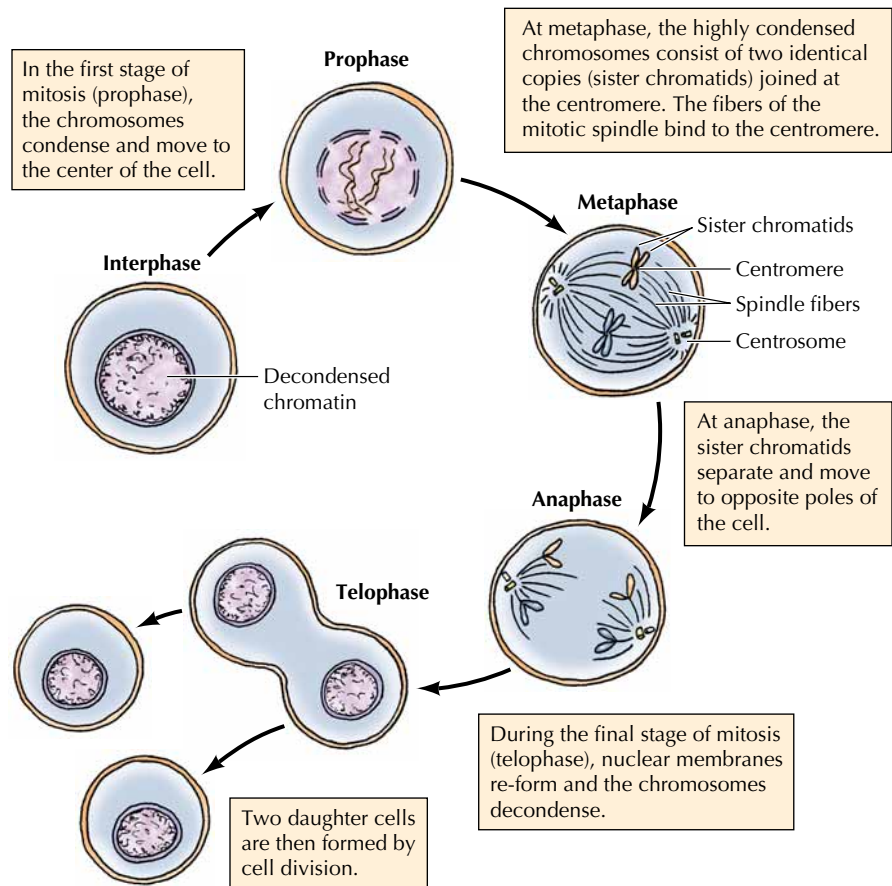
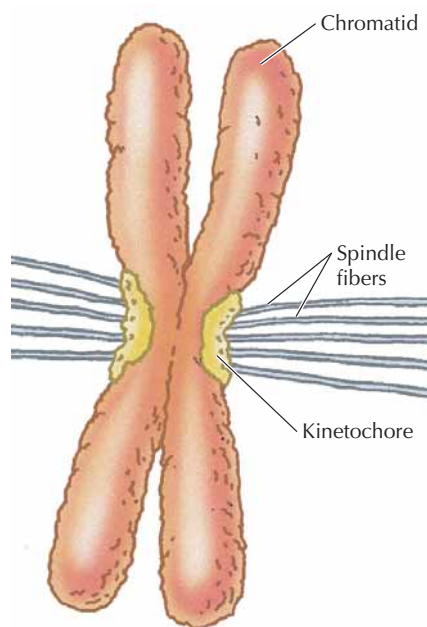


FIGURE 5.18 Chromosomes during mitosis Since DNA replicates during interphase, the cell contains two identical duplicated copies of each chromosome prior to entering mitosis.



mediates the attachment of chromosomes to the mitotic spindle. Proteins associated with the kinetochore then act as “molecular motors” that drive the movement of chromosomes along the spindle fibers, segregating the chromosomes to daughter nuclei.

Centromeric DNA sequences were initially defined in yeasts, where their function can be assayed by following the segregation of plasmids at mitosis (Figure 5.20). Plasmids that contain functional centromeres segregate like chromosomes and are equally distributed to daughter cells following mitosis. In the absence of a functional centromere, however, the plasmid does not segregate properly, and many daughter cells fail to inherit plasmid DNA. Assays of this type have enabled determination of the sequences required for centromere function. Such experiments first showed that the centromere sequences of the well-studied yeast *Saccharomyces cerevisiae* are contained in approximately 125 base pairs consisting of three sequence ele-

FIGURE 5.19 The centromere of a metaphase chromosome The centromere is the region at which the two sister chromatids remain attached at metaphase. Specific proteins bind to centromeric DNA, forming the kinetochore, which is the site of spindle fiber attachment.

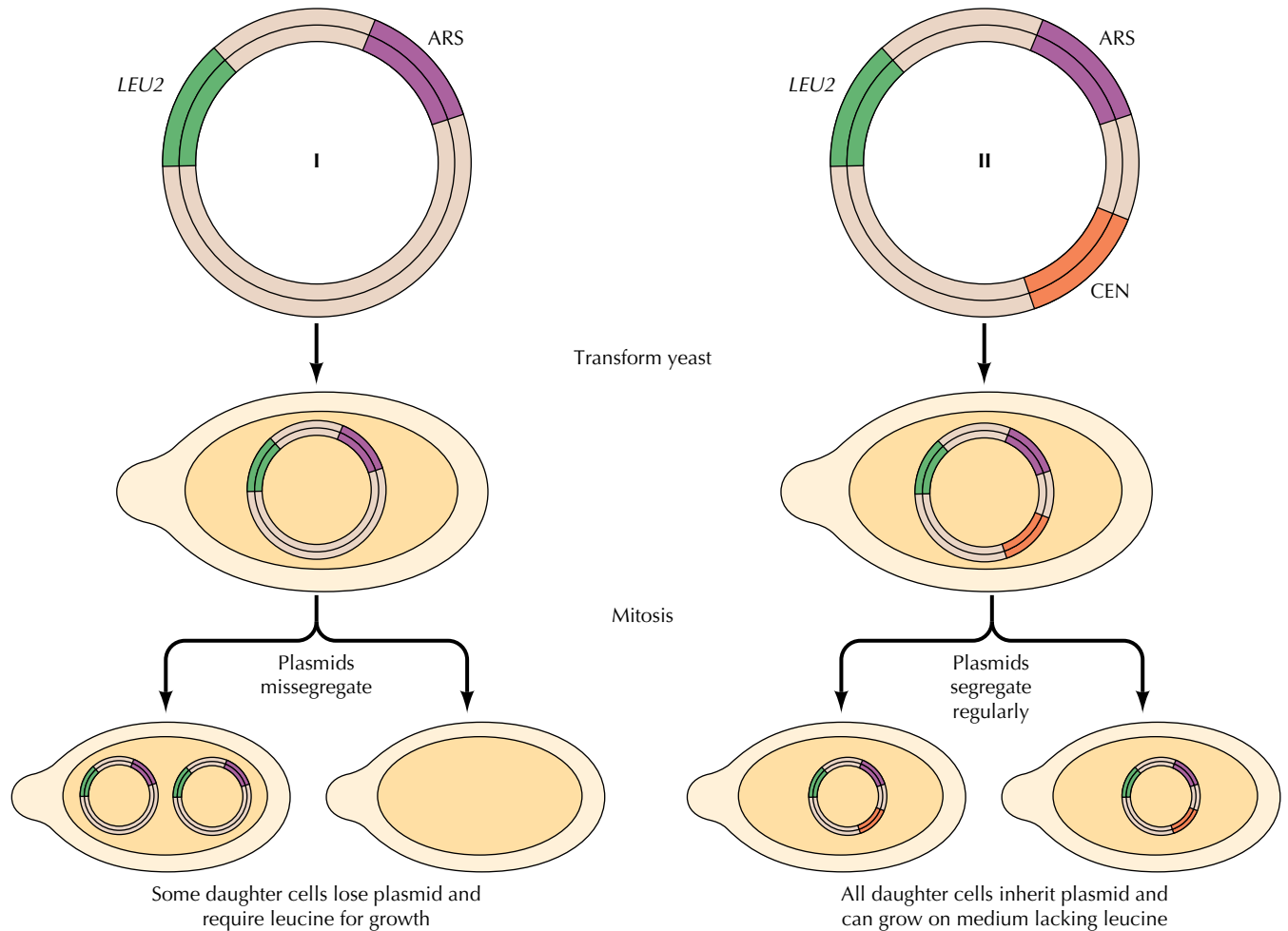


FIGURE 5.20 Assay of a centromere in yeast Both plasmids shown contain a selectable marker (*LEU2*) and DNA sequences that serve as origins of replication in yeast (ARS, which stands for autonomously replicating sequence). However, plasmid I lacks a centromere and is therefore frequently lost as a result of missegregation during mitosis. In contrast, the presence of a centromere (CEN) in plasmid II ensures its regular transmission to daughter cells.

ments: two short sequences of 8 and 25 base pairs separated by 78 to 86 base pairs of very AT-rich DNA (Figure 5.21A).

The short centromere sequences defined in *S. cerevisiae*, however, do not appear to reflect the situation in other eukaryotes. More recent studies have defined the centromeres of the fission yeast *Schizosaccharomyces pombe* by a similar functional approach. Although *S. cerevisiae* and *S. pombe* are both yeasts, they appear to be as divergent from each other as either is from humans and are quite different in many aspects of their cell biology. These two yeast species thus provide complementary models for simple and easily studied eukaryotic cells. The centromeres of *S. pombe* span 40 to 100 kb of DNA; they are approximately a thousand times larger than those of *S. cerevisiae*. They consist of a central core of 4 to 7 kb of single-copy DNA flanked by repetitive sequences (Figure 5.21B). Not only the central core but also the flanking repeated sequences are required for centromere function, so the

studied and CenH3-containing nucleosomes are required for assembly of the other kinetochore proteins needed for centromere function. It thus appears that chromatin structure rather than a specific DNA sequence may be the primary determinant of the identity and function of centromeres. However, we still do not understand how centromeric chromatin is specified and stably maintained following cell division, so fundamental questions about the nature of centromeres in higher eukaryotes remain to be answered.

Telomeres

The sequences at the ends of eukaryotic chromosomes, called **telomeres**, play critical roles in chromosome replication and maintenance. Telomeres were initially recognized as distinct structures because broken chromosomes were highly unstable in eukaryotic cells, implying that specific sequences are required at normal chromosomal termini. This was subsequently demonstrated by experiments in which telomeres from the protozoan *Tetrahymena* were added to the ends of linear molecules of yeast plasmid DNA. The addition of these telomeric DNA sequences allowed these plasmids to replicate as linear chromosome-like molecules in yeasts, demonstrating directly that telomeres are required for the replication of linear DNA molecules.

The telomere DNA sequences of a variety of eukaryotes are similar, consisting of repeats of a simple-sequence DNA containing clusters of G residues on one strand (Table 5.5). For example, the sequence of telomere repeats in humans and other mammals is AGGGTT, and the telomere repeat in *Tetrahymena* is GGGGTT. These sequences are repeated hundreds or thousands of times and terminate with a 3' overhang of single-stranded DNA. The repeated sequences of telomere DNA of some organisms (including humans) form loops at the ends of chromosomes as well as binding a number of proteins that protect the chromosome termini from degradation or from being joined together (Figure 5.22).

TABLE 5.5 Telomeric DNAs

Organism	Telomeric repeat sequence
Yeasts	
<i>Saccharomyces cerevisiae</i>	G ₁₋₃ T
<i>Schizosaccharomyces pombe</i>	G ₂₋₅ TTAC
Protozoans	
<i>Tetrahymena</i>	GGGGTT
<i>Dictyostelium</i>	G ₁₋₈ A
Plant	
<i>Arabidopsis</i>	AGGGTTT
Mammal	
Human	AGGGTT

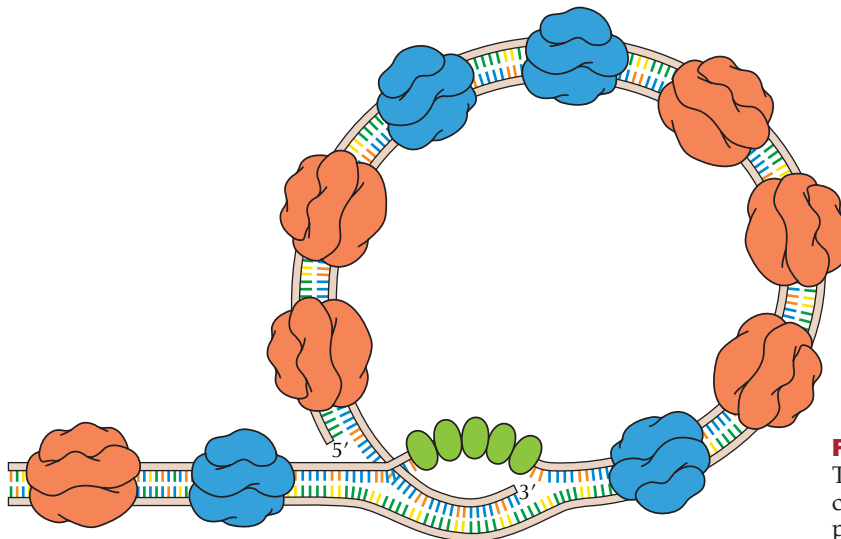


FIGURE 5.22 Structure of a telomere

Telomere DNA loops back on itself to form a circular structure and associates with a number of proteins that protect the ends of chromosomes.

■ **Cancer cells have high levels of telomerase, allowing them to maintain the ends of their chromosomes through indefinite divisions. Since normal somatic cells lack telomerase activity and do not divide indefinitely, drugs that inhibit telomerase are being developed as anti-cancer agents.**

Telomeres play a critical role in replication of the ends of linear DNA molecules (see Chapter 6). DNA polymerase is able to extend a growing DNA chain but cannot initiate synthesis of a new chain at the terminus of a linear DNA molecule. Consequently, the ends of linear chromosomes cannot be replicated by the normal action of DNA polymerase. This problem has been solved by the evolution of a special enzyme, **telomerase**, which uses reverse transcriptase activity to replicate telomeric DNA sequences. Maintenance of telomeres appears to be an important factor in determining the lifespan and reproductive capacity of cells, so studies of telomeres and telomerase have the promise of providing new insights into conditions such as aging and cancer.

The Sequences of Complete Genomes

Some of the most exciting recent advances in molecular biology have been the results of analyzing the complete nucleotide sequences of both the human genome and the genomes of several model organisms, including *E. coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila*, *Arabidopsis*, and the mouse (Table 5.6). The results of whole genome sequencing have taken us beyond the characterization of individual genes to a global view of the organization and gene content of entire genomes. In principle, this approach has the potential of identifying all the genes in an organism, which then become accessible for investigations of their structure and function. Moreover, the availability of complete genome sequences opens the exciting possibility of identifying the sequences that regulate gene expression by genome-wide analysis. While much remains to be learned, the available genome sequences have provided scientists with a unique data-

TABLE 5.6 Representative Sequenced Genomes

Organism	Genome size (Mb) ^a	Number of genes	Protein-coding sequence
Bacteria			
<i>Mycoplasma genitalium</i>	0.58	470	88%
<i>H. influenzae</i>	1.8	1743	89%
<i>E. coli</i>	4.6	4288	88%
Yeasts			
<i>S. cerevisiae</i>	12	6000	70%
<i>S. pombe</i>	12	4800	60%
Invertebrates			
<i>C. elegans</i>	97	19,000	25%
<i>Drosophila</i>	180	13,600	13%
Plants			
<i>Arabidopsis thaliana</i>	125	26,000	25%
Rice	390	37,000	12%
Fish			
Pufferfish	370	20,000–23,000	10%
Birds			
Chicken	1000	20,000–23,000	3%
Mammals			
Human	3200	20,000–25,000	1.2%

^aMb = millions of base pairs

base, consisting of the nucleotide sequences of complete sets of genes and their regulatory sequences. Since many of these genes have not been previously identified, determination of their functions will form the basis of many future studies in cell biology.

Prokaryotic Genomes

We now know the complete genome sequences of more than 100 different bacteria, and still more are in the process of being determined. The first complete sequence of a cellular genome, reported in 1995 by a team of researchers led by Craig Venter, was that of the bacterium *Haemophilus influenzae*, a common inhabitant of the human respiratory tract. The genome of *H. influenzae* is approximately 1.8×10^6 base pairs (1.8 megabases, or Mb), slightly less than half the size of the *E. coli* genome. The complete nucleotide sequence indicated that the *H. influenzae* genome is a circular molecule containing 1,830,137 base pairs of DNA. The sequence was then analyzed to identify the genes encoding rRNAs, tRNAs, and proteins. Potential protein-coding regions were identified by computer analysis of the DNA sequence to detect **open-reading frames**—long stretches of nucleotide sequence that can encode polypeptides because they contain none of the three chain-terminating codons (UAA, UAG, and UGA). Since these chain-terminating codons occur randomly once in every 21 codons (3 chain-terminating codons out of 64 total), open-reading frames that extend for more than a hundred codons usually represent functional genes.

This analysis identified six copies of rRNA genes, 54 different tRNA genes, and 1743 potential protein-coding regions in the *H. influenzae* genome (Figure 5.23). More than a thousand of these could be assigned a biological role (e.g., an enzyme of the citric acid cycle) on the basis of their relationships to known protein sequences, but the others represent genes of

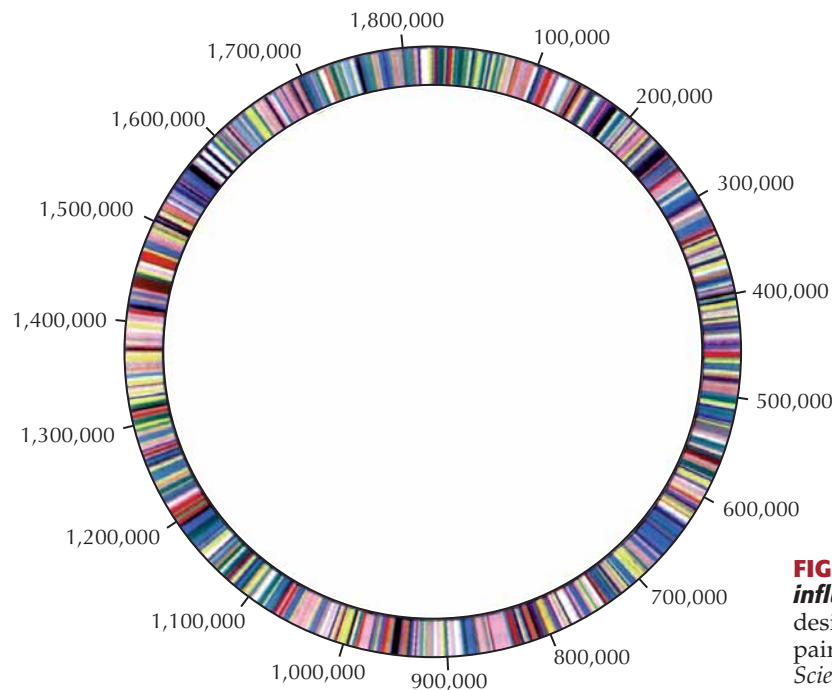


FIGURE 5.23 The genome of *Haemophilus influenzae* Predicted protein-coding regions are designated by colored bars. Numbers indicate base pairs of DNA. (From R. D. Fleischmann et al., 1995. *Science* 269: 496.)

unknown function. The predicted coding sequences have an average size of approximately 900 base pairs, so they cover about 1.6 Mb of DNA, corresponding to nearly 90% of the genome of *H. influenzae*.

The complete sequence of the genome of *Mycoplasma genitalium* is of particular interest because mycoplasmas are the simplest present-day bacteria and contain the smallest genomes of all known cells. The genome of *M. genitalium* is only 580 kb (0.58 Mb) long and may represent the minimal set of genes required to maintain a self-replicating organism. Analysis of its DNA sequence indicates that *M. genitalium* contains only 470 predicted protein-coding sequences, which correspond to approximately 88% of genomic DNA. Many of these sequences were identified as genes encoding proteins involved in DNA replication, transcription, translation, membrane transport, and energy metabolism. However, *M. genitalium* contains many fewer genes for metabolic enzymes than does *H. influenzae*, reflecting its more limited metabolism. For example, many genes known to encode components of biosynthetic pathways are lacking in the genome of *M. genitalium*, consistent with its need to obtain amino acids and nucleotide precursors from a host organism. Interestingly, the *Mycoplasma* genome also includes approximately 150 genes of currently unknown function. Thus, even in the simplest of cells, the biological roles of many genes remain to be determined.

The sequence of the genome of the archaebacterium *Methanococcus jannaschii*, reported in 1996, provided major insights into the evolutionary relationships between the archaebacteria, eubacteria, and eukaryotes. The genome of *M. jannaschii* is 1.7 Mb and contains 1738 predicted protein-coding sequences—similar in size to the genome of *H. influenzae*. However, only about one-third of the protein-coding sequences identified in *M. jannaschii* were related to known genes of either eubacteria or eukaryotes, indicating the distinct genetic composition of the archaebacteria. The genes of *M. jannaschii* encoding proteins involved in energy production and biosynthesis of cell constituents are related to those of eubacteria, suggesting that basic metabolic processes evolved in a common ancestor of both the archaebacteria and the eubacteria. Importantly, however, the *M. jannaschii* genes encoding proteins involved in DNA replication, transcription, and translation are more closely related to those of eukaryotes than to those of eubacteria. Genomic sequencing of this archaebacterium thus indicates that the archaebacteria and eukaryotes are as closely related to each other as either is to the eubacteria (see Figure 1.7).

Although the relative simplicity and facile genetics of *E. coli* have made it a favored organism of molecular biologists, the 4.6-Mb *E. coli* genome was not completely sequenced until 1997. Analysis of the *E. coli* sequence revealed a total of 4288 genes, with protein-coding sequences accounting for approximately 88% of the *E. coli* genome. Of the 4288 genes revealed by sequencing, 1835 had been previously identified and the functions of an additional 821 could be deduced by comparisons to the sequences of characterized genes of other organisms. However, the functions of 1632 *E. coli* genes (nearly 40% of the genome) could not be determined. Thus, even for an organism as thoroughly studied as *E. coli*, genomic sequencing demonstrates that a great deal remains to be learned about prokaryotic cell biology.

The Yeast Genome

As noted already, the simplest eukaryotic genome (1.2×10^7 base pairs of DNA) is found in the yeast *Saccharomyces cerevisiae*. Moreover, yeasts grow rapidly and are subject to simple genetic manipulations. Thus in many

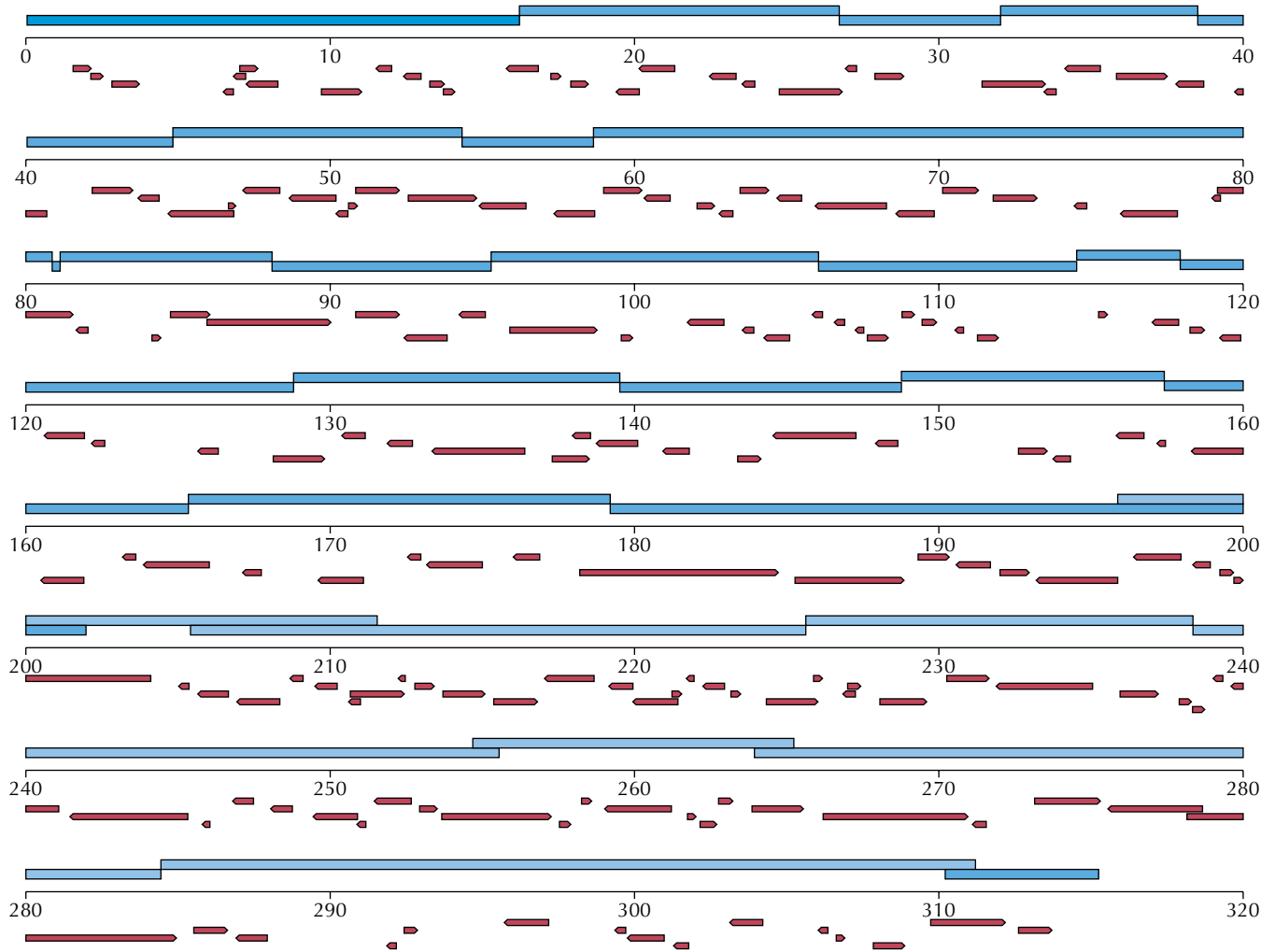


FIGURE 5.24 Yeast chromosome III
 The upper blue bars designate the clones used for DNA sequencing. Open-reading frames are indicated by arrows. (From S. G. Oliver et al., 1992. *Nature* 357: 38.)

ways yeasts are model eukaryotic cells that can be studied much more readily than the cells of mammals or other higher eukaryotes. Consequently, the complete sequencing of an entire yeast chromosome in 1992 (Figure 5.24), followed by determination of the sequence of the complete *S. cerevisiae* genome in 1996, were major steps in understanding the molecular biology of eukaryotic cells.

The *S. cerevisiae* genome contains about 6000 genes, including 5885 predicted protein-coding sequences, 140 ribosomal RNA genes, 275 transfer RNA genes, and 40 genes encoding small nuclear RNAs involved in RNA processing (discussed in Chapter 7). Yeasts thus have a high density of protein-coding sequences, similar to bacterial genomes, with protein-coding sequences accounting for approximately 70% of total yeast DNA. Consistent with this, only 4% of yeast genes were found to contain introns. Moreover, those *S. cerevisiae* genes that do contain introns usually have only a single small intron near the beginning of the gene.

Computer analysis was able to assign a predicted function to approximately 3000 of the *S. cerevisiae* protein-coding sequences based on similarities to the sequences of known genes. Based on analysis of these genes, it appears that approximately 11% of yeast proteins function in metabolism,

3% in the production and storage of metabolic energy, 3% in DNA replication, repair, and recombination, 7% in transcription, 6% in translation, and 14% in protein sorting and transport. However, the functions of many of these genes are only known in general terms (such as “transcription factor”), so their precise roles within the cell still need to be determined. Moreover, since half of the proteins encoded by the yeast genome were unrelated to previously described genes, the functions of an additional 3000 unknown proteins remain to be elucidated by genetic and biochemical analyses.

The sequence of the *S. cerevisiae* genome has been more recently followed by the sequence of the genome of the fission yeast, *S. pombe*, as well as the genomes of several other yeast and fungi. As discussed earlier in this chapter, *S. cerevisiae* and *S. pombe* are quite divergent and differ in many aspects of their biology, including the structure of their centromeres (see Figure 5.21). Interestingly, their genomes also display considerable differences. Although both *S. cerevisiae* and *S. pombe* have approximately the same amount of unique sequence DNA (12.5 Mb), *S. pombe* appears to contain only about 4800 genes. Introns are much more prevalent in *S. pombe* than in *S. cerevisiae*. Approximately 43% of *S. pombe* genes contain introns and the introns in *S. pombe* are larger than those in *S. cerevisiae*, so protein-coding sequence accounts for only about 60% of the *S. pombe* genome. The majority of *S. pombe* genes have homologs in the *S. cerevisiae* genome, but approximately 700 genes are unique to *S. pombe*.

Now that yeast genome sequences have been completed, determination of the functions of the many new genes described in both *S. cerevisiae* and *S. pombe* is a major goal. Fortunately, yeasts are particularly amenable to functional analyses of unknown genes because of the facility with which normal chromosomal loci can be inactivated by homologous recombination with cloned sequences (discussed in Chapter 4). Therefore direct functional analysis of yeast genes that were initially identified only on the basis of their nucleotide sequence can be systematically undertaken. Sequencing the yeast genomes has thus opened the door to studying many new areas of the biology of a simple eukaryotic cell. Such studies are expected to reveal the functions of many new genes that are not restricted to yeasts but are common to all eukaryotes, including humans.

The Genomes of *Caenorhabditis elegans* and *Drosophila melanogaster*

The genomes of *C. elegans* and *Drosophila* are relatively simple animal genomes, intermediate in size and complexity between those of yeasts and humans. Distinctive features of each of these organisms make them important models for genome analysis: *C. elegans* is widely used for studies of animal development, and *Drosophila* has been especially well analyzed genetically. The genomes of these organisms, however, are about tenfold larger than those of yeasts, introducing a new order of difficulty in genome mapping and sequencing. Determination of the sequence of *C. elegans* in 1998 therefore represented an important milestone in genome analysis, which extended genome sequencing from unicellular organisms (bacteria and yeast) to a multicellular organism recognized as an important model for animal development.

The initial phases of analysis of the *C. elegans* genome used DNA fragments cloned in cosmids, which accommodate DNA inserts of approximately 30–45 kb (see Table 4.3). This approach, however, was unable to cover the complete genome, which was accomplished by the cloning of much larger pieces of DNA in **yeast artificial chromosome (YAC)** vectors.

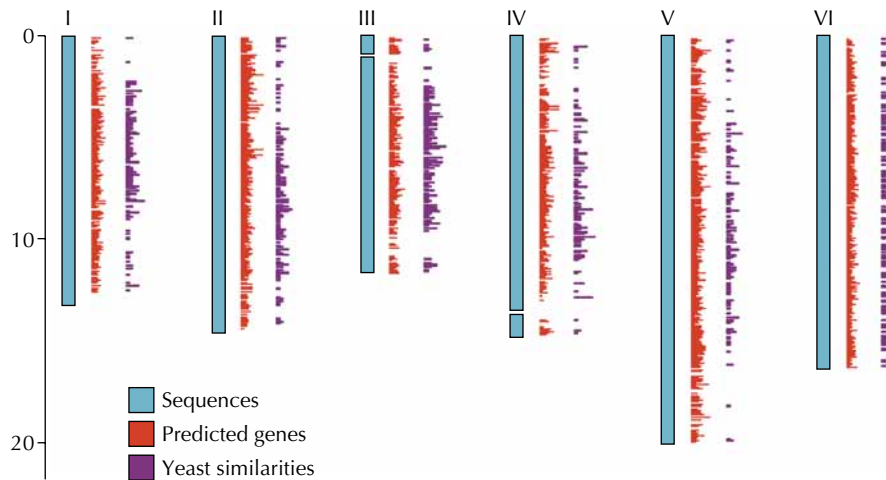


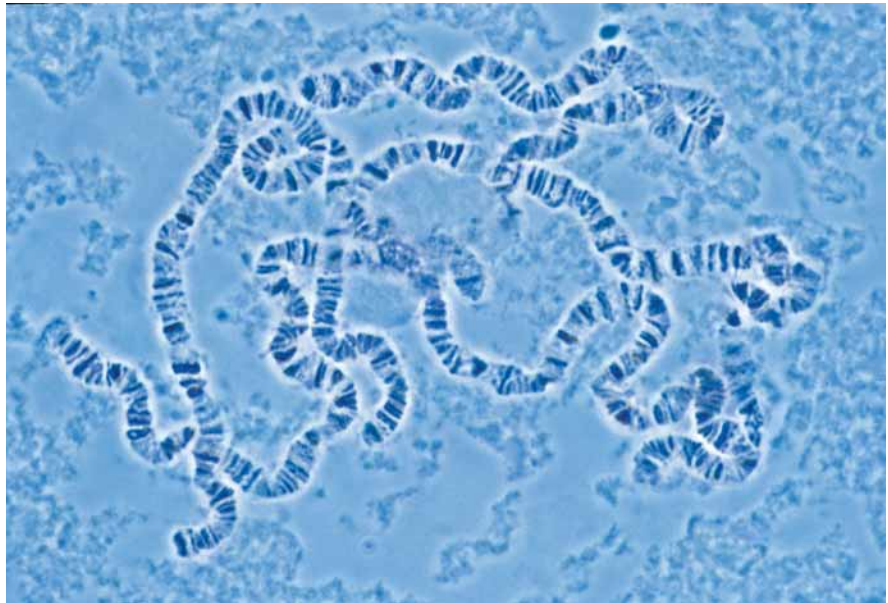
FIGURE 5.25 The *C. elegans* genome The positions of the predicted genes of *C. elegans* on each chromosome are indicated by red bars. Those that are similar to genes of yeast are indicated by purple. (From The *C. elegans* Sequencing Consortium, 1998. *Science* 282: 2012.)

As noted in Chapter 4, the unique feature of YACs is that they contain centromeres and telomeres, allowing them to replicate as linear chromosome-like molecules in yeasts. They can therefore be used to clone DNA fragments the size of yeast chromosomal DNAs, up to thousands of kilobases in length. The large DNA inserts that can be cloned in YACs and other high-capacity vectors are critically important for analysis of complex genomes.

The *C. elegans* genome is 97×10^6 base pairs and contains about 19,000 predicted protein-coding sequences—approximately three times the number of genes in yeast (Figure 5.25). In contrast to the compact genome organization of yeast, genes in *C. elegans* span about 5 kilobases and contain an average of five introns. Protein-coding sequences thus account for only about 25% of the *C. elegans* genome, as compared to 60–70% of *S. pombe* and *S. cerevisiae* and nearly 90% of bacterial genomes.

Approximately 40% of the predicted *C. elegans* proteins displayed significant similarity to known proteins of other organisms. As expected, there are substantially more similarities between the proteins of *C. elegans* and humans than between *C. elegans* and either yeast or bacteria. Proteins that are common between *C. elegans* and yeast may function in the basic cellular processes shared by these organisms, such as metabolism, DNA replication, transcription, translation, and protein sorting. These core biological processes appear to be carried out by a similar number of genes in both organisms, and it is likely that these genes will be shared by all eukaryotic cells. In contrast, the majority of *C. elegans* genes are not found in yeast and may function in the more intricate regulatory activities required for the development of multicellular organisms. Elucidating the functions of these genes is likely to be particularly exciting in terms of understanding animal development. Although adult *C. elegans* contain only 959 somatic cells in the entire body, they have all of the specialized cell types found in more complicated animals. Moreover, the complete pattern of cell divisions leading to *C. elegans* development has been described, including analysis of the connections made by all 302 neurons in the adult animal. Many of the genes involved in *C. elegans* development and differentiation have already been found to be related to genes involved in controlling the proliferation and differentiation of mammalian cells, substantiating the validity of *C. elegans* as a model for more complex animals. With little doubt, many more critical developmental control genes will be uncovered from studies of the *C. elegans* genomic sequence.

FIGURE 5.26 Polytene chromosomes of *Drosophila* A light micrograph of stained salivary gland chromosomes. The four chromosomes (X, 2, 3, and 4) are joined at their centromeres. (Peter J. Bryant/Biological Photo Service.)



Drosophila is another key model for animal development, which has been particularly well-characterized genetically. The advantages of *Drosophila* for genetic analysis include its relatively simple genome and the fact that it can be easily maintained and bred in the laboratory. In addition, a special tool for genetic analysis in *Drosophila* is provided by the giant **polytene chromosomes** that are found in some tissues, such as the salivary glands of larvae. These chromosomes arise in nondividing cells as a consequence of repeated replication of DNA strands that fail to separate from each other. Thus each polytene chromosome contains hundreds of identical DNA molecules aligned in parallel. Because of their size, these polytene chromosomes are visible in the light microscope, and appropriate staining procedures reveal a distinct banding pattern (Figure 5.26). The banding of polytene chromosomes provides a much greater degree of resolution than that achieved with metaphase chromosomes (e.g., see Figure 5.17). The polytene chromosomes are decondensed interphase chromosomes that contain actively expressed genes. More than 5000 bands are visible, each corresponding to an average length of approximately 20 kb of DNA. In contrast, the bands identified in human metaphase chromosomes contain several megabases of DNA.

The banding pattern of polytene chromosomes thus provides a high-resolution physical map of the *Drosophila* genome. Gene deletions can often be correlated with the loss of a specific chromosomal band, thereby defining the physical location of the gene on the chromosome. In addition, cloned DNAs can be mapped by *in situ* hybridization to polytene chromosomes, often with sufficient resolution to localize cloned genes to specific bands (Figure 5.27). Thus the map positions of cosmid or YAC clones (which span many bands) can readily be determined, providing the base for genomic sequence analysis.

Because of the power of *Drosophila* genetics, the sequencing of the *Drosophila* genome early in 2000 was an important advance in genomic analysis. The genome of *Drosophila* consists of approximately 180×10^6 base pairs, of which about one-third is heterochromatin. The heterochromatin consists principally of simple sequence satellite repeats, in addition to inter-

spersed transposable elements, and was not included in the genomic sequence. The remaining 120×10^6 base pairs of euchromatin was sequenced using a combination of **bacterial artificial chromosome (BAC)** clones, which carry large inserts of DNA (see Table 4.3), and a shotgun approach in which small fragments of DNA were randomly cloned and sequenced in plasmid vectors. The sequences of these small fragments of DNA were then assembled into a large contiguous sequence by identification of overlaps between fragments, and these sequence assemblies were aligned with the BAC clones to yield a complete sequence of the euchromatic portion of the *Drosophila* genome.

The *Drosophila* genome contains approximately 13,600 genes; surprisingly fewer than the number of genes in *C. elegans*, even though *Drosophila* is a more complex organism. However, it is important to note that this difference in gene number does not correspond to a difference in genetic complexity, because many genes are duplicated in both *Drosophila* and *C. elegans*. When these duplications are taken into account, it appears that both *Drosophila* and *C. elegans* contain a similar number of distinct genes, estimated between 10,000 and 15,000. Like *C. elegans*, *Drosophila* genes contain an average of 4 introns, and the total amount of intron sequence is similar to the amount of exon sequence. In total, protein-coding sequence accounts for about 13% of the *Drosophila* genome.

It is especially striking that a complex animal like *Drosophila* has only about twice the number of unique genes found in yeast, which appears to be a much simpler organism. Apparently, the complexity of multicellular organisms is not simply related to a greater number of genes. Part of the increased biological complexity of *Drosophila* and *C. elegans* may arise from the fact that their proteins are generally larger and contain more functional domains than the proteins of yeast. Further studies and functional analysis of the genes that have been uncovered by sequencing the *Drosophila* and *C. elegans* genomes will undoubtedly play a major role in understanding the ways in which these genes act to direct the complex process of animal development.

Plant Genomes

The completion of the genome sequence of *Arabidopsis thaliana* in 2000 extended genome sequencing from animals to plants, and was thus a major event in plant biology. *Arabidopsis thaliana* is a simple flowering plant, which has been widely used as a model for studies of plant molecular biology and development. Its advantages as a model organism for molecular biology and genetics include its relatively small genome of approximately 125×10^6 base pairs, similar in size to the genomes of *C. elegans* and *Drosophila*. Like the *Drosophila* genome, the *Arabidopsis* genome was sequenced principally using BAC vectors to accommodate large DNA inserts.

Surprisingly, analysis of the *Arabidopsis* genome indicated that it contained approximately 26,000 protein-coding genes—significantly more genes than were found in either *C. elegans* or *Drosophila*. However, this unexpectedly large number of genes does not reflect a greater diversity of proteins encoded by the *Arabidopsis* genome. Instead, it appears that the large number of genes in *Arabidopsis* is the result of duplications of large segments of the *Arabidopsis* genome. These duplications involve approximately 60% of the genome, so the number of distinct protein-coding genes in *Arabidopsis* is estimated to be about 15,000—similar to the number of genes in *C. elegans* or *Drosophila*.

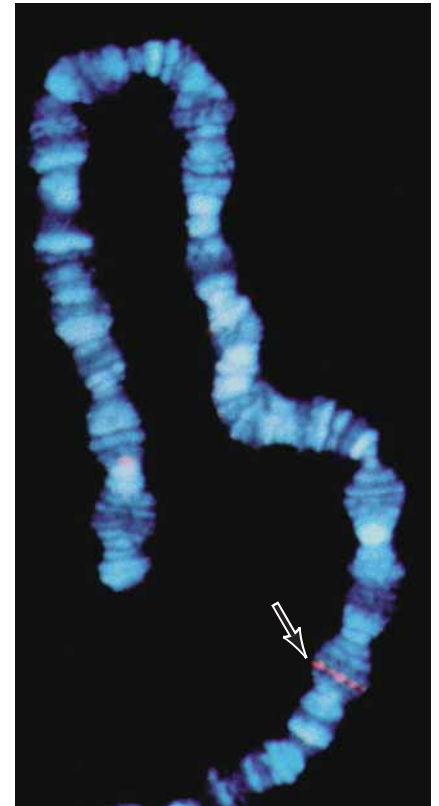
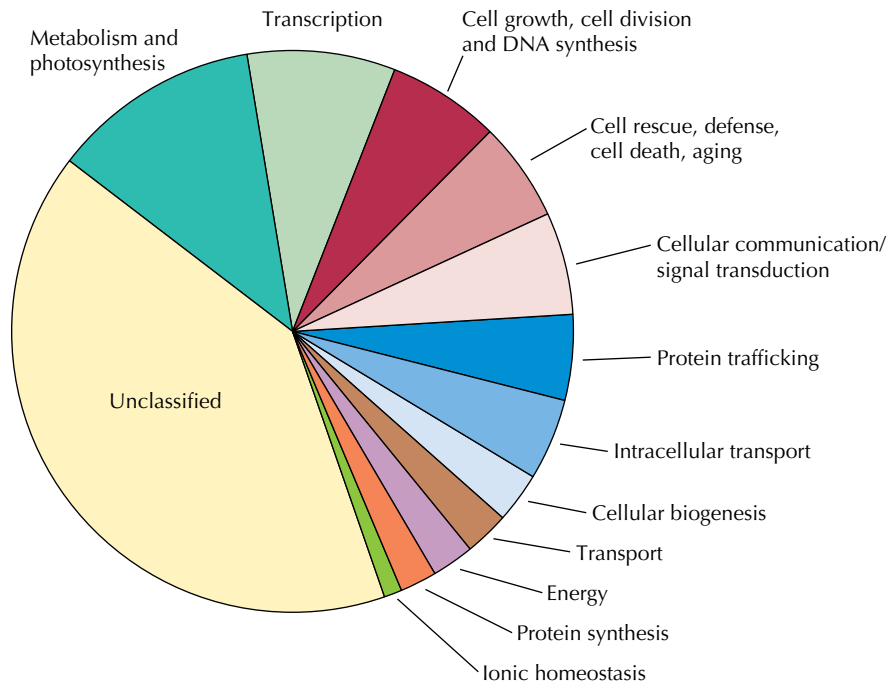


FIGURE 5.27 *In situ* hybridization to a *Drosophila* polytene chromosome. Hybridization of a YAC clone to a polytene chromosome is illustrated. The region of hybridization is indicated by an arrow. (Courtesy of Daniel L. Hartl, Harvard University.)

FIGURE 5.28 Functions of predicted genes of *Arabidopsis thaliana* The chart illustrates the proportion of *Arabidopsis* in different functional categories. (From The *Arabidopsis* Genome Initiative, 2000. *Nature* 408: 796.)



The gene density in *Arabidopsis* is also similar to that of *C. elegans*, with protein-coding sequences accounting for about 25% of the *Arabidopsis* genome. On the average, *Arabidopsis* genes have approximately 4 introns, and the total length of intron sequences is about the same as the total length of exon sequences. Transposable elements account for about 10% of the *Arabidopsis* genome. As in *Drosophila*, transposable element repeats are clustered at the centromeres together with satellite repetitive sequences.

Comparative analysis of the functions of the *Arabidopsis* genes has revealed both interesting similarities and differences between the genes of plants and animals. *Arabidopsis* genes involved in fundamental cellular processes such as DNA replication, repair, transcription, translation, and protein trafficking are similar to those in yeast, *C. elegans*, and *Drosophila*, reflecting the common evolutionary origins of all eukaryotic cells. In contrast, the *Arabidopsis* genes encoding proteins involved in processes such as cell signaling and membrane transport are quite different from those in animals, consistent with the major differences in physiology and development between plants and animals. About one-third of all *Arabidopsis* genes appear unique to plants, as they are not found in yeast or animal genomes. The largest functional group of *Arabidopsis* genes, corresponding to 22% of the genome, encodes proteins involved in metabolism and photosynthesis (Figure 5.28). Another large group of genes (12% of the genome) encodes proteins involved in plant defense. It is also noteworthy that *Arabidopsis* encodes more than 3000 proteins that regulate transcription (accounting for approximately 17% of the genome). This number of gene regulatory proteins (transcription factors) is two or three times more than are found in *Drosophila* and *C. elegans*, respectively. Many of the *Arabidopsis* transcription factors are unique to plants, presumably reflecting distinct features of gene expression in plant development and in the response of plants to the environment.

The sequence of *Arabidopsis* was followed in 2002 by publication of two draft sequences of the rice genome. Rice is of major importance as a cereal crop and is the staple food for more than half the world's population, so sequencing the rice genome has the potential of leading to very significant applications in agriculture and biotechnology. Two groups of researchers reported draft sequences of the genomes of two subspecies of rice: the *indica* subspecies, which is the most widely cultivated subspecies in China and most of the rest of Asia; and the *japonica* subspecies, which is the variety preferred in Japan. These initial draft sequences of the rice genome were followed by a high quality complete sequence of the *japonica* subspecies in 2005.

The rice genome consists of about 390×10^6 base pairs of DNA—about 3 times larger than the genome of *Arabidopsis*. At least 35% of the rice genome consists of transposable elements, in part accounting for its larger size. In addition, rice contain a surprisingly high number of predicted protein-coding genes, estimated at approximately 37,000. Like *Arabidopsis*, rice contains many duplicated genes, which have arisen as a result of duplication of large segments (approximately 60%) of the genome. Nonetheless, the rice genome contains more genes than either *Arabidopsis* or humans, underscoring the fact that gene number does not directly correlate with biological complexity in eukaryotes. Interestingly, approximately 70% of the genes predicted in rice are also found in *Arabidopsis*, and almost 90% of the genes that have been identified in *Arabidopsis* are found in rice. Most of the genes shared between *Arabidopsis* and rice are not found in yeast or animal genomes and therefore appear to be specific for plants.

The Human Genome

For many scientists, the ultimate goal of genome analysis was determination of the complete nucleotide sequence of the human genome: approximately 3×10^9 base pairs of DNA. To understand the magnitude of this undertaking, recall that the human genome is more than ten times larger than that of *Drosophila*; that the smallest human chromosome is several times larger than the entire yeast genome; and that the extended length of DNA that makes up the human genome is about 1 m long. From all of these perspectives, determination of the human genome sequence was a phenomenal undertaking, and its publication in draft form in 2001 was heralded as a scientific achievement of historic magnitude.

The human genome is distributed among 24 chromosomes (22 autosomes and the 2 sex chromosomes), each containing between 45 and 280 Mb of DNA (Figure 5.29). Prior to determination of the genome sequence, several thousand human genes had been identified and mapped to positions on the human chromosomes. One commonly used method to localize genes is *in situ* hybridization of probes labeled with fluorescent dyes to chromosomes—a method generally referred to as **fluorescence *in situ* hybridization**, or **FISH** (Figure 5.30). *In situ* hybridization to metaphase chromosomes allows the mapping of a cloned gene to a locus defined by a chromosome band. Because each band of human metaphase chromosomes contains thousands of kilobases of DNA, *in situ* hybridization to human metaphase chromosomes does not provide the detailed mapping information obtained by hybridization to the polytene chromosomes of *Drosophila*, which allows the localization of genes to interphase chromosome bands containing only 10 to 20 kb of DNA. Higher resolution can be obtained, however, by hybridization to more extended human chromosomes from

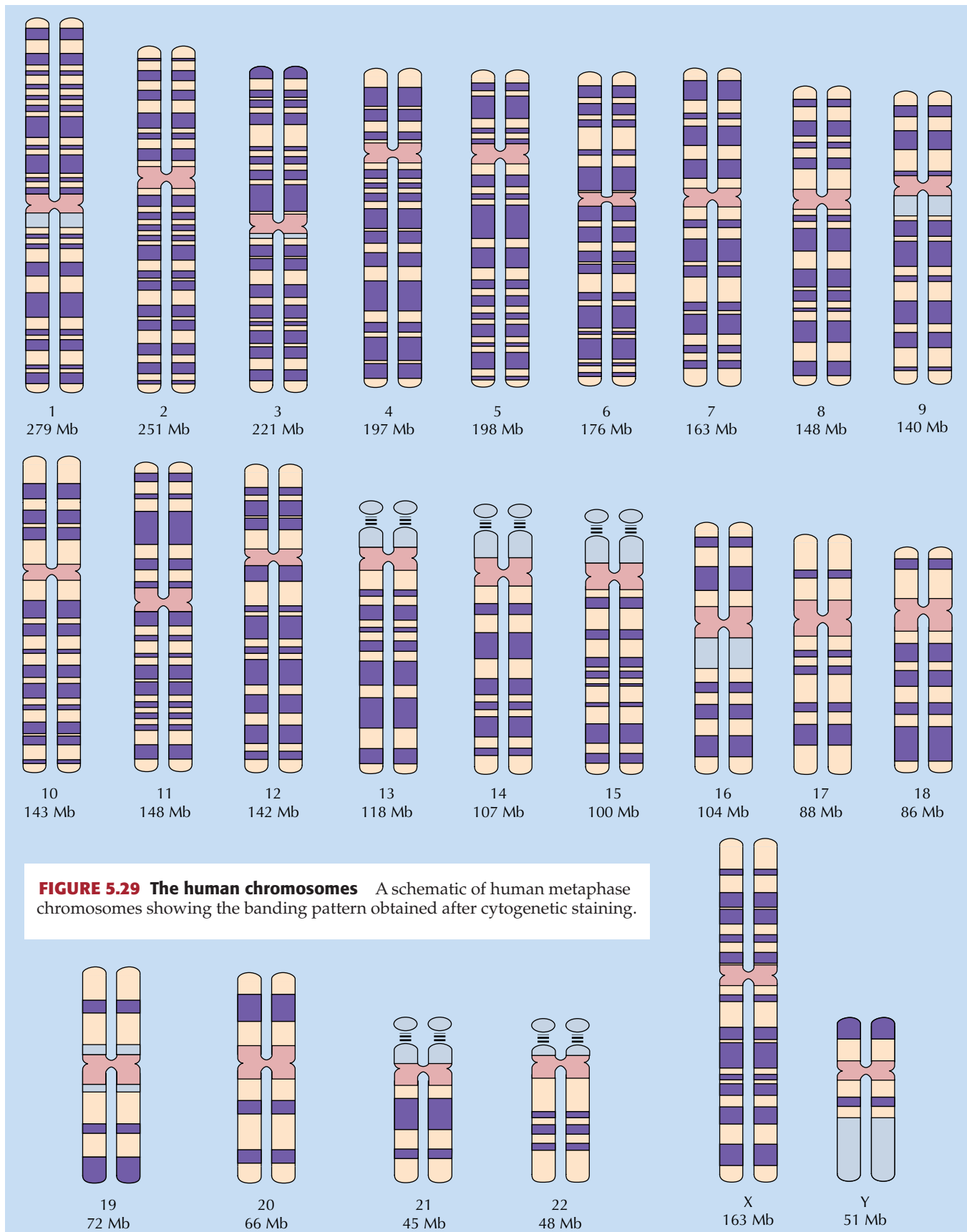


FIGURE 5.29 The human chromosomes A schematic of human metaphase chromosomes showing the banding pattern obtained after cytogenetic staining.

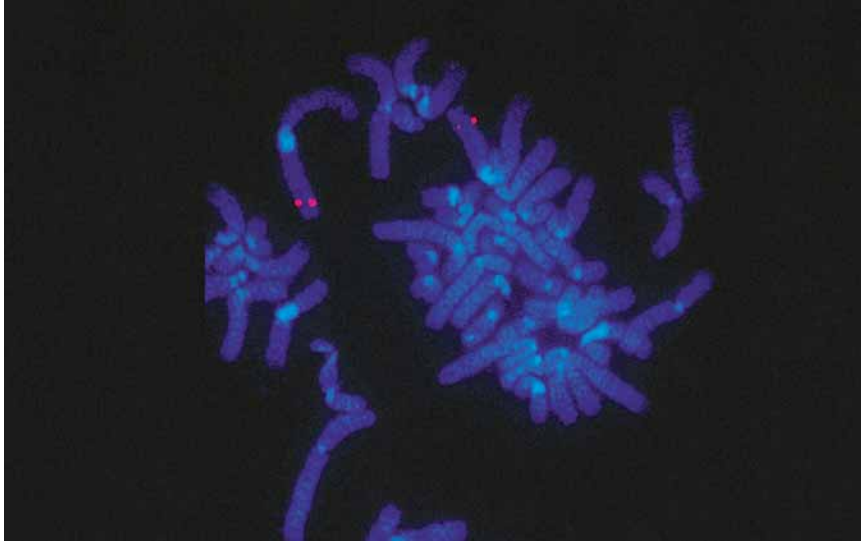


FIGURE 5.30 Fluorescence *in situ* hybridization A fluorescent probe for the gene encoding lamin B receptor is hybridized to stained human metaphase chromosomes (blue). Single gene hybridization signals are detected as red fluorescence. (Courtesy of K. L. Wydner and J. B. Lawrence, University of Massachusetts Medical Center.)

prometaphase or interphase cells, allowing the use of fluorescence *in situ* hybridization to map cloned genes to regions of about 100 kb. In addition to FISH, genetic linkage analysis and the physical mapping of cloned genomic and cDNA sequences were used to establish physical and genetic maps of the human genome, which provided a background for genomic sequencing.

The draft sequences of the human genome published in 2001 were produced by two independent teams of researchers, who used different approaches. One research team, The International Human Genome Sequencing Consortium, used BAC clones that had been mapped to sites on the human chromosomes as the substrates for sequencing. The other team, led by Craig Venter of Celera Genomics, used a shotgun approach in which small fragments were cloned and sequenced, and overlaps between the sequences of these fragments were then used to assemble the sequence of the genome. Both of these sequences were initially incomplete drafts in which approximately 90% of the euchromatin portion of the genome had been sequenced and assembled. Continuing efforts have closed the gaps and improved the accuracy of the draft sequences, leading to publication of a high-quality human genome sequence in 2004.

The sequenced euchromatin portion of the genome encompasses approximately 2.9×10^6 kb of DNA (Figure 5.31). The total size of the genome is approximately 3.2×10^6 kb, with the remaining 10% of the genome (0.3×10^6

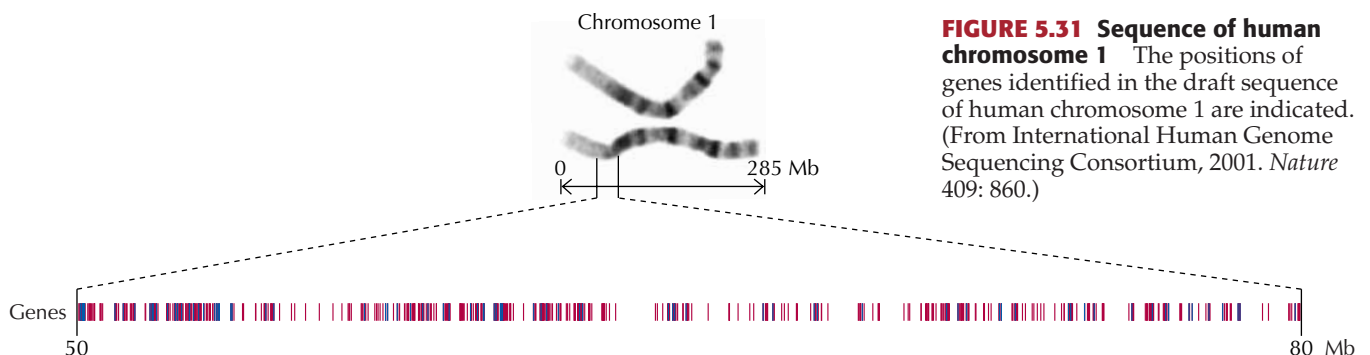
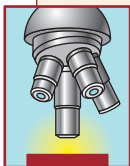


FIGURE 5.31 Sequence of human chromosome 1 The positions of genes identified in the draft sequence of human chromosome 1 are indicated. (From International Human Genome Sequencing Consortium, 2001. *Nature* 409: 860.)

KEY EXPERIMENT

The Human Genome



Initial Sequencing and Analysis of the Human Genome

International Human Genome Sequencing Consortium
Nature, Volume 409, 2001, pages 860–921

The Sequence of the Human Genome

J. Craig Venter and 273 others
Science, Volume 291, 2001, pages 1304–1351

The Context

The idea of sequencing the entire human genome was first conceived in the mid-1980s. It was initially met with broad skepticism among biologists, most of whom felt it was simply not a feasible undertaking. At the time, the largest genome that had been completely sequenced was that of Epstein-Barr virus, which totaled approximately 180,000 base pairs of DNA. From this perspective, sequencing the human genome, which was almost 20,000 times larger, seemed inconceivable to many. However, the idea of such a massive project in biology captivated the imagination of others, including Charles DeLisi who was then head of the Office of Health and Environmental Research at the Department of Energy. In 1986 DeLisi succeeded in launching the Human Genome Initiative as a project within the Department of Energy.

The project gained broader support in 1988 when it was endorsed by a committee of the National Research Council. This committee recommended a broader effort, including sequencing the genomes of several model organisms and the parallel development of detailed genetic and physical maps of the human chromosomes. This effort was centered at the National Institutes of Health, initially under the direction of James Watson (codiscoverer of the structure of DNA), and then under the leadership of Frances Collins.

The first complete genome to be sequenced was that of the bacterium *Haemophilus influenzae*, reported by Craig Venter and colleagues in 1995. Venter had been part of the genome

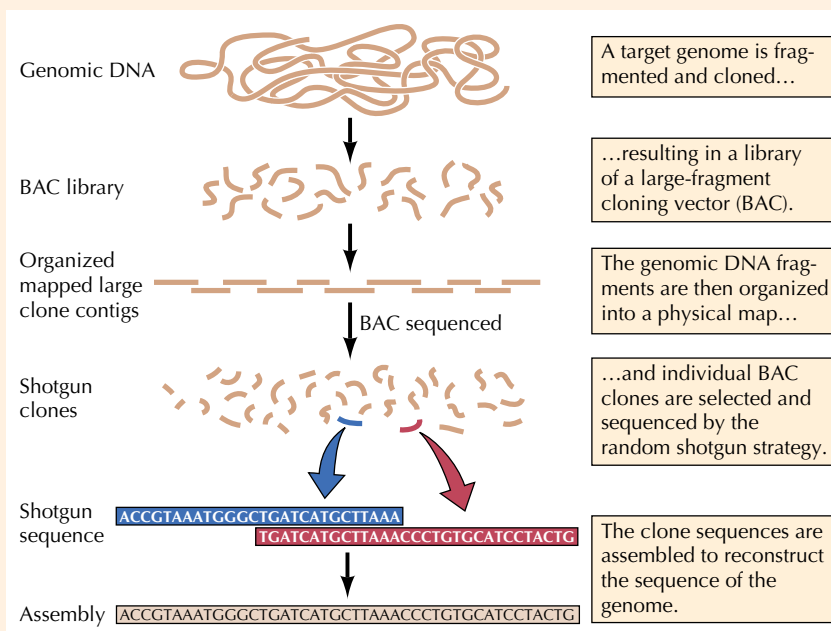
sequencing effort at the National Institutes of Health but had left to head a nonprofit company, The Institute for Genomic Research, in 1991. In the meantime, considerable progress had been made in mapping the human genome, and the initial sequence of *H. influenzae* was followed by the sequences of other bacteria, yeast, and *C. elegans* in 1998.

In 1998 Venter formed a new company, Celera Genomics, and announced plans to use advanced sequencing technologies to obtain the entire human genome sequence in 3 years. Collins and other leaders of the publicly funded genome project responded by accelerating their efforts, resulting in a

race that eventually led to the publication of two draft sequences of the human genome in February, 2001.

The Experiments

The two groups of scientists used different approaches to obtain the human genome sequence. The publicly funded team, The International Human Genome Sequencing Consortium, headed by Eric Lander, sequenced DNA fragments derived from BAC clones that had been previously mapped to human chromosomes, similar to the approach used to determine the sequence of the yeast and *C. elegans* genomes (see figure). In contrast, the Celera Genomics team used a whole-genome shotgun sequencing approach that Venter and colleagues had first used to sequence the genome of *H. influenzae*. In this approach, DNA fragments were sequenced at random, and overlaps between fragments were then used to reassemble a complete genome sequence. Both sequences covered only the euchromatin portion of the human genome—approximately 2900



Strategy for genome sequencing using BAC clones that had been organized into overlapping clusters (contigs) and mapped to human chromosomes.

KEY EXPERIMENT

Mb of DNA—with the heterochromatin repeat-rich portion of the genome (approximately 300 Mb) remaining unsequenced.

Both of these initially published versions were draft, rather than completed, sequences. Subsequent efforts completed the sequence, leading to publication of a highly accurate sequence of the human genome in 2004.

The Impact

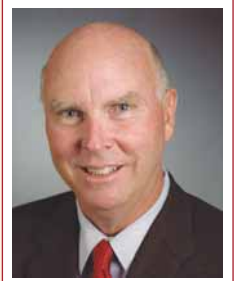
Several important conclusions immediately emerged from the human genome sequences. First, the number of human genes was surprisingly small and appears to be between 20,000 and 25,000 in the completed sequence. Interestingly, however, alternative splicing appears to be common in the human genome, so

many genes may encode more than 1 protein. Introns account for about 20% of the human genome and repetitive sequences for about 60%. It is noteworthy that over 40% of human DNA is composed of sequences derived by reverse transcription, emphasizing the importance of this mode of information transfer in shaping our genome.

Beyond these immediate conclusions, the sequence of the human genome, together with the genome sequences of other organisms, will provide a new basis for biology and medicine in the years to come. The impact of the genome sequence will be felt in discovering new genes and their functions, understanding gene regulation, elucidating the basis of human diseases, and developing new strategies for prevention and treatment



Eric Lander



Craig Venter

based on the genetic makeup of individuals. Knowledge of the human genome may ultimately contribute to meeting what Venter and colleagues refer to as “The real challenge of human biology...to explain how our minds have come to organize thoughts sufficiently well to investigate our own existence.”

kb) corresponding to highly repetitive sequences in heterochromatin. As discussed earlier in this chapter, interspersed repetitive sequences, the majority of which are transposable elements that have moved throughout the genome by reverse transcription of RNA intermediates, account for approximately 45% of the human euchromatin sequence. Another 5% of the genome consists of duplicated segments of DNA, so about 60% of the human genome consists of repetitive DNA sequences.

A major surprise from the genome sequence is the unexpectedly low number of human genes. The human genome consists of only 20,000–25,000 genes, which is not much larger than the number of genes in simpler animals like *C. elegans* and *Drosophila*. In fact, humans have fewer genes than rice, emphasizing one of the major conclusions that has emerged from the results of genome sequencing: the biological complexity of an organism is not simply a function of the number of genes in its genome. On the other hand, there appears to be a significant amount of alternative splicing in human genes, allowing a single gene to specify more than one protein (see Figure 5.5). Although the extent of alternative splicing in humans is not yet clear, it may substantially expand the number of proteins that can be encoded by the human genome.

Human genes are spread over much larger distances and contain more intron sequence than genes in *Drosophila* or *C. elegans*. The average protein-coding sequence in human genes is approximately 1400 base pairs, similar to that in *Drosophila* and *C. elegans*. However, the average human gene spans about 30 kb of DNA, with more than 90% of the gene corresponding to introns. Approximately 20% of the genome thus consists of introns, with only about 1.2% of the human genome corresponding to protein-coding sequences.

■ For many years, scientists generally accepted an estimate of approximately 100,000 genes in the human genome. On publication of the draft genome sequence in 2001, the number was drastically reduced to between 30,000 and 40,000. Current estimates, based on the high quality sequence published in 2004 and using improved computational tools to identify genes, reduce the number of human genes still further, to approximately 20,000 to 25,000.

Over 40% of the predicted human proteins are related to proteins in other sequenced organisms, including *Drosophila* and *C. elegans*. Many of these conserved proteins function in basic cellular processes, such as metabolism, DNA replication and repair, transcription, translation, and protein trafficking. Most of the proteins that are unique to humans are made up of protein domains that are also found in other organisms, but these domains are arranged in novel combinations to yield distinct proteins in humans. Compared to *Drosophila* and *C. elegans*, the human genome contains expanded numbers of genes involved in functions related to the greater complexity of vertebrates, such as the immune response, the nervous system, and blood clotting, as well as increased numbers of genes involved in development, cell signaling, and the regulation of transcription.

The Genomes of Other Vertebrates

In addition to the human genome, a large and growing number of vertebrate genomes have been sequenced in the last few years, including the genomes of fish, chickens, and other mammals (Figure 5.32). These sequences provide interesting comparisons to that of the human genome and are expected to facilitate the identification of a variety of different types of functional sequences, including regulatory elements that control gene expression.

■ Pufferfish contain a very powerful neurotoxin, called tetrodotoxin, in some of their tissues. In Japan, pufferfish are considered a delicacy and prepared by specially trained chefs in licensed restaurants.

The genome of the pufferfish *Fugu rubripes* was chosen for sequencing because it is unusually compact for a vertebrate genome. Consisting of only 3.7×10^8 base pairs of DNA, the pufferfish genome is only about one-eighth the size of the human genome. Although the pufferfish and human genomes contain a similar number of genes, the pufferfish has far less repetitive sequence and smaller introns. In particular, repetitive sequences account for only about 15% of the pufferfish genome (corresponding to approximately 50 million base pairs of DNA) as compared to about 60% of the human genome (approximately 2 billion base pairs). Because of this reduced amount of repetitive sequence, genes are more closely packed in the pufferfish and occupy about one-third of its genome. Pufferfish and human genes contain similar numbers of introns, but introns are shorter in the pufferfish, so that protein coding sequence corresponds to approximately one-third of the average gene or about 10% of the pufferfish genome (as compared to 1.2% of the human genome). The pufferfish thus provides a compact model of a vertebrate genome in which genes and critical regulatory sequences are highly concentrated, facilitating efforts to focus continuing studies on these functional genomic elements.

The chicken is intermediate between the pufferfish and mammals, both in evolutionary divergence and in the size of its genome. Consisting of approximately 10^9 base pairs, the chicken genome is about one-third the size of the human genome. However, it is estimated to contain 20,000 to 23,000 genes, similar to the gene content of humans. The smaller size of the chicken genome is largely the result of a substantial reduction in the amount of repetitive sequences and pseudogenes compared to mammalian genomes.

The mammalian genomes that have been sequenced, in addition to the human genome, include the genomes of the mouse, rat, dog, and chimpanzee. These genomes are all similar in size to the human genome and contain similar numbers of genes. However, each offers particular advantages for further understanding gene regulation and function. As discussed in earlier chapters, the mouse is the key model system for experimental studies of mammalian genetics and development, so the availability of the

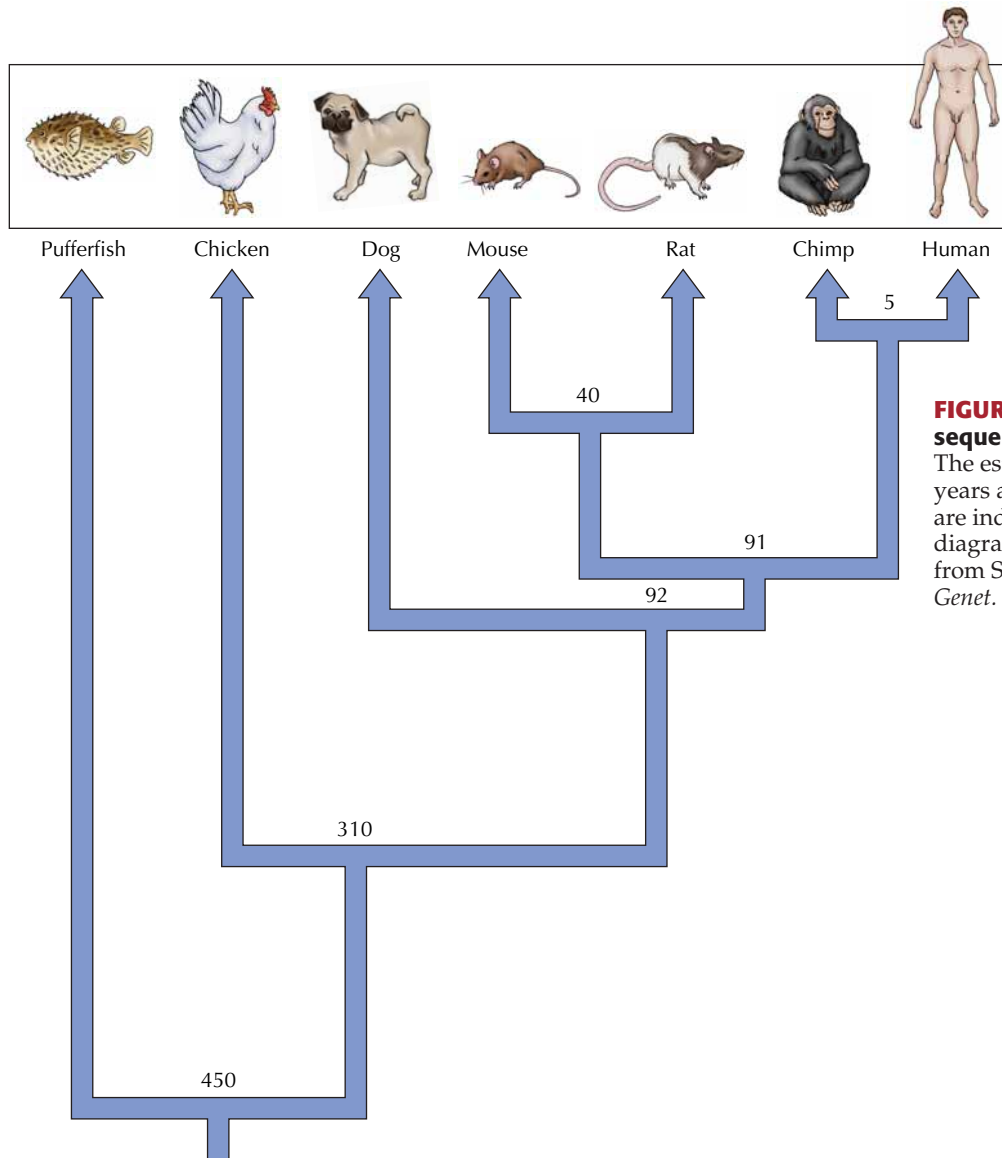


FIGURE 5.32 Evolution of sequenced vertebrates

The estimated times (millions of years ago) when species diverged are indicated at branch points in the diagram. (Times of divergence are from S. B. Hedges, 2002. *Nature Rev. Genet.* 3: 838.)

mouse genome sequence provides an essential database for research in these areas. Likewise, the rat is an important model for human physiology and medicine, and these studies will be facilitated by the availability of the rat genome sequence. Mice, rats, and humans have 90% of their genes in common, providing a clear genetic foundation for the use of the mouse and rat as models for human development and disease.

The many distinct breeds of pet dogs make the sequence of the dog genome particularly important in understanding the genetic basis of morphology, behavior, and a variety of complex diseases that afflict both dogs and humans. There are approximately 300 breeds of dogs, which differ in their physical and behavioral characteristics as well as in their susceptibility to a variety of diseases, including several types of cancer, blindness, deafness, and metabolic disorders. Susceptibility to particular diseases is a highly specific property of different breeds, greatly facilitating identification of the responsible genes. Since many of these diseases are common to both dogs and humans, genetic studies in dogs can be expected to impact

human health as well as veterinary medicine. An interesting example is provided by studies of sleep disorders in which the gene responsible for a rare inherited form of narcolepsy was identified in Doberman pinschers. Subsequent studies implicated related defects in human narcolepsy and possibly other sleep disorders. Similar types of genetic analysis are underway to understand the genetic basis of other complex diseases, such as hip dysplasia and rheumatoid arthritis, that are common in some breeds of dogs, and the results of these studies will undoubtedly benefit both dogs and humans. In the future, we can also expect genetic analysis of behavior in dogs. Since many canine behaviors, such as separation anxiety, are also common in humans, psychologists may have much to learn from the species that has been our closest companion for thousands of years.

The sequence of the genome of the chimpanzee, our nearest evolutionary relative, is expected to help pinpoint the unique features of our genome that distinguish humans from other primates. Interestingly however, comparison of the chimpanzee and human genome sequences does not suggest an easy answer to the question of what makes us human. The nucleotide sequences of the chimpanzee and human genomes are nearly 99% identical. The difference between the sequences of these closely related species (approximately 1 nucleotide in 100) is about ten times greater than the difference between the genomes of individual humans (approximately 1 nucleotide in 1000). Perhaps surprisingly, the sequence differences between humans and chimpanzees are not restricted to noncoding sequences. Instead, they frequently alter the coding sequences of genes, leading to changes in the amino acid sequences of most of the proteins encoded by chimpanzees and humans. Although many of these amino acid changes may not affect protein function, it appears that there are changes in the structure as well as in the expression of thousands of genes between chimpanzees and humans, so identifying those differences that are key to the origin of humans will not be a simple task.

Bioinformatics and Systems Biology

The human genome sequence, together with the sequences of other genomes, provides a wealth of information that forms a new framework for studies of cell and molecular biology and opens new possibilities in medical practice. In addition, the genome sequencing projects have raised new questions and substantially changed the way in which many problems in biology are being approached. Traditionally, molecular biologists have studied one or a few genes or proteins at a time. This has been changed by the genome sequencing projects, which introduced new large-scale experimental approaches in which vast amounts of data were generated. Handling the enormous amounts of data generated by whole genome sequencing required sophisticated computational analysis and spawned the new field of **bioinformatics**, which lies at the interface between biology and computer science and is focused on developing the computational methods needed to analyze and extract useful biological information from the sequence of billions of bases of DNA. The development of such computational methods has also led to other types of large-scale biological experimentation, including simultaneous analysis of the expression of thousands of mRNAs or proteins and the development of high-throughput methods to determine gene function using RNA interference. These large-scale experimental approaches form the basis of the new field of **systems biology**,

which seeks a quantitative understanding of the integrated dynamic behavior of complex biological systems and processes. Systems biology thus combines large-scale biological experimentation with quantitative analysis and the development of testable models for complex biological processes. The global analysis of cell proteins (proteomics), discussed in Chapter 2, is one example of these new large-scale experimental/computational approaches. Some of the additional research areas that are amenable to large-scale experimentation, bioinformatics, and systems biology are discussed below.

Systematic Screens of Gene Function

The identification of all of the genes in an organism opens the possibility for a large-scale systematic analysis of gene function. One approach is to systematically inactivate (or knockout) each gene in the genome by homologous recombination with an inactive mutant allele (see Figure 4.39). As noted in Chapter 4, this has been done in yeast to produce a collection of yeast strains with mutations in all known genes, which can then be analyzed to determine which genes are involved in any biological property of interest. Alternatively, large-scale screens based on RNA interference (RNAi) are being used to systematically dissect gene function in a variety of organisms, including *Drosophila*, *C. elegans*, and mammalian cells in culture.

In RNAi screens, double-stranded RNAs are used to induce degradation of the homologous mRNAs in cells (see Figure 4.42). With the availability of complete genome sequences, libraries of double-stranded RNAs can be designed and used in genome-wide screens to identify all of the genes involved in any biological process that can be assayed in a high-throughput manner. For example, genome-wide RNAi analysis has been used to identify genes required for the growth and viability of *Drosophila* cells in culture (Figure 5.33). Individual double-stranded RNAs from the genome-wide library are tested in microwells in a high-throughput format to identify those that interfere with the growth of cultured cells, thereby characterizing the entire set of genes in the *Drosophila* genome that are required for cell growth or survival. Similar RNAi screens have been used to identify genes involved in a variety of biological processes, including cell signaling path-

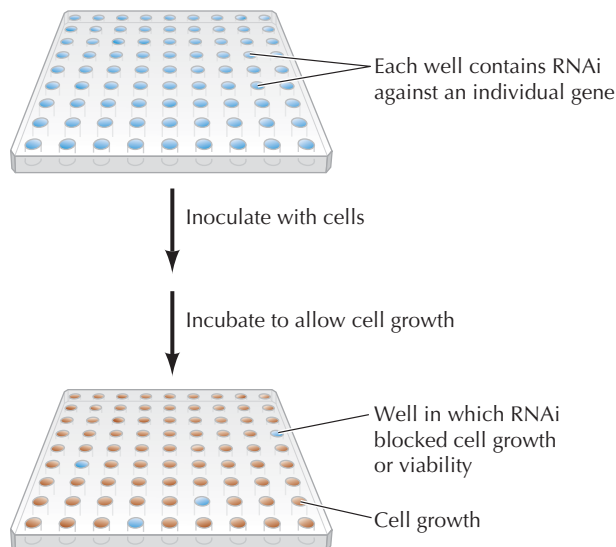


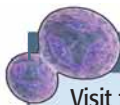
FIGURE 5.33 Genome-wide RNAi screen for cell growth and viability Each microwell contains RNAi corresponding to an individual gene in the *Drosophila* genome. *Drosophila* tissue culture cells are added to each well and incubated to allow cell growth. Those wells in which cells fail to grow identify genes required for cell growth or viability.

ways, protein degradation, and transmission at synapses in the nervous system.

Regulation of Gene Expression

Genome sequences can in principle reveal not only the protein-coding sequences of genes, but also the regulatory elements that control gene expression. As discussed in subsequent chapters, regulation of gene expression is critical to many aspects of cell function, including the development of complex multicellular organisms. Understanding the mechanisms that control gene expression is therefore a central undertaking in contemporary cell and molecular biology, and it is expected that the availability of genome sequences will contribute substantially to this task. Unfortunately, it is far more difficult to identify gene regulatory sequences than it is to identify protein-coding sequences. Most regulatory elements are short sequences of DNA, typically spanning only about 10 base pairs. Consequently, sequences resembling regulatory elements occur frequently by chance in genomic DNA, so physiologically significant elements can not be identified from DNA sequence alone. The identification of functional regulatory elements and elucidation of the signaling networks that control gene expression therefore represent major challenges in bioinformatics and systems biology.

The availability of genomic sequences has enabled scientists to undertake global studies of gene expression in which the expression levels of all genes in a cell can be assayed simultaneously. These experiments employ DNA microarrays in which each gene is represented by an oligonucleotide corresponding to a small dot on a slide (see Figure 4.27). Hybridization of fluorescent-labeled cDNA copies of cellular mRNAs to such a microarray allows simultaneous determination of the mRNA levels of all cellular genes. This approach has been particularly valuable in revealing global changes in gene regulation associated with discrete cell behaviors, such as cell differentiation or the response of cells to a particular hormone or growth factor. Since genes that are coordinately regulated within a cell may be controlled by similar mechanisms, analyzing changes in the expression of multiple genes can help to pinpoint shared regulatory elements.



COMPANION WEBSITE
Visit the website that accompanies **The Cell** (www.sinauer.com/cooper) for animations, videos, quizzes, problems, and other review material.

Human	CTGCCT---AAGTAGCCTAGACGCTCCCGTGC---CCCGGGCGGG---TAG
Mouse	CGCCGC---CTGCATTATTCAC-----
Rat	CTGCTC---ATGCATAATTCAC-----
Dog	CTGCTTCAAACAGTGGGGCAGACGGTCCCGCGGCCCAAGGCAGGCCG

	Err- α
Human	GCCTGGCCGAAAATCTCTCCCGCGCGCTGACCTTGGGTGCCCCAGCCA
Mouse	-----AAGCCTGTGGCGCGC---CGTACCTTGGGTGCCCCAGGGC
Rat	-----AAGTTTCT---CTGC---CTGACCTTGGGTGCCCCAGGGC
Dog	GGCTGC---AGACCTGCCCTGAGGGAAAGACCTTGGCGGCCCGAGGGC

Human	GGCTGCGGGCCCGAGACCCCG-----GGCCTCCCT
Mouse	GGCTGCAGGCTCACCACCC-----GTCTTTTCT
Rat	AG---GCATACACCCCGCCT-----TTTTTTTT
Dog	GGCGCGGGCCAGGCCCCCTCCCTCCCTCCCTCCCTCCCTCCCT

FIGURE 5.34 Conservation of functional gene regulatory elements Human, mouse, rat, and dog sequences near the transcription start site of a gene contain a functional regulatory element that binds the transcriptional regulatory protein Err- α . These sequences (highlighted in yellow) are conserved in all four genomes, whereas the surrounding sequences are not. (From X. Xie et al., 2005. *Nature* 434: 338.)

A variety of computational approaches are also being used to characterize functional regulatory elements. One approach is comparative analysis of the genome sequences of related organisms. This is based on the assumption that functionally important sequences are conserved in evolution, whereas nonfunctional segments of DNA diverge more rapidly. Computational analysis based on this approach has recently identified gene regulatory sequences that are conserved between the mouse, rat, dog, and human genomes (Figure 5.34). In addition, functional regulatory elements often occur in clusters, reflecting the fact that genes are generally regulated by the interactions of multiple transcription factors (see Chapter 7). Computer algorithms designed to detect clusters of transcription factor binding sites in genomic DNA have also proven useful in identifying sequences that regulate gene expression.

The combination of large-scale experimental methods and computational analysis has been successful in providing at least an initial indication of the transcriptional regulatory elements that govern expression of genes in yeast. However, extending these approaches to the far more complicated genomes of humans and other mammals remains a major challenge for future research.

Variation among Individuals and Genomic Medicine

Comparisons of genome sequences of related species is helpful in understanding the basis of differences between species, as well as in identifying genes and regulatory sequences that have been conserved in evolution. A different type of information can be gained by comparing the genome sequences of different individuals. Variations between individual genomes underlie differences in physical and mental characteristics, including susceptibility to many diseases. One of the major applications of the human genome sequence will be helping to uncover new genes involved in many of the diseases that afflict mankind, including cancer, heart disease, and degenerative diseases of the nervous system such as Parkinson's and Alzheimer's disease. In addition, understanding our unique genetic makeup as individuals is expected to lead to the development of new tailor-made strategies for disease prevention and treatment.

The genomes of two unrelated people differ in about one of every thousand bases. Most of this variation is in the form of single base changes, known as single nucleotide polymorphisms (SNPs), which are found at about 10 million positions in the genome. Over a million commonly occurring SNPs have been mapped in the human genome. These SNPs are distributed relatively uniformly in genomic DNA (Figure 5.35), and it is noteworthy that more than 90% of protein-coding genes contain at least one SNP. It is likely that these SNPs are responsible for most genetic differences in individual characteristics, so a substantial effort is being directed towards using SNPs to map the genes responsible for inherited differences in disease susceptibility. Analysis of these variations among individuals will not only allow specific genes to be associated with susceptibility to different diseases but will also enable physicians to tailor strategies for disease prevention and treatment to match the genetic makeup of individual patients. Comparisons between the genomes of different individuals may also help to elucidate the contribution of our genes to other unique characteristics, such as athletic ability or intelligence, and to better understand the interactions between genes and environment that lead to complex human behaviors.

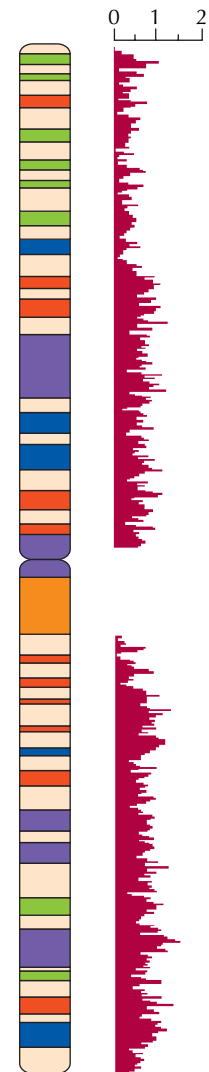


FIGURE 5.35 Single nucleotide polymorphisms (SNPs) in human chromosome 1 The distribution of SNPs (frequency per kilobase) is indicated. (From D. A. Hinds et al., 2005. *Science* 307: 1072.)

■ Current efforts aim to develop technologies that would be capable of sequencing the genome of individuals at very low cost. Such affordable 'personal genome projects' might provide better health care options tailored to the needs of individual patients.

KEY TERMS

gene, spacer sequence, exon, intron, RNA splicing, kilobase (kb), alternative splicing

simple-sequence repeat, satellite DNA, SINE, LINE, retrotransposon, DNA transposon, retrovirus-like element

gene family, pseudogene, processed pseudogene

chromatin, histone, nucleosome, nucleosome core particle, chromatosome, euchromatin, heterochromatin

centromere, kinetochore

telomere, telomerase

megabase (Mb), open-reading frame

SUMMARY

THE COMPLEXITY OF EUKARYOTIC GENOMES

Introns and Exons: Most eukaryotic genes have a split structure in which segments of coding sequence (exons) are interrupted by noncoding sequences (introns). In complex eukaryotes, introns account for more than ten times as much DNA as exons.

Repetitive DNA Sequences: Over 50% of mammalian DNA consists of highly repetitive DNA sequences, some of which are present in 10^5 to 10^6 copies per genome. These sequences include simple-sequence repeats as well as repetitive elements that have moved throughout the genome by either RNA or DNA intermediates.

Gene Duplications and Pseudogenes: Many eukaryotic genes are present in multiple copies, called gene families, which have arisen by duplication of ancestral genes. Some members of gene families function in different tissues or at different stages of development. Other members of gene families (pseudogenes) have been inactivated by mutations and no longer represent functional genes. Gene duplications can occur either by duplication of a segment of DNA or by reverse transcription of an mRNA, giving rise to a processed pseudogene. Approximately 5% of the human genome consists of duplicated DNA segments. In addition, there are more than 10,000 processed pseudogenes in the human genome.

The Composition of Higher Eukaryotic Genomes: Only a small fraction of the genome in complex eukaryotes corresponds to protein-coding sequences. The human genome is estimated to contain 20,000–25,000 genes, with protein-coding sequence corresponding to only about 1.2% of the DNA. Approximately 20% of the human genome consists of introns, and more than 60% is composed of repetitive and duplicated DNA sequences.

CHROMOSOMES AND CHROMATIN

Chromatin: The DNA of eukaryotic cells is wrapped around histones to form nucleosomes. Chromatin can be further compacted by the folding of nucleosomes into higher-order structures, including the highly condensed metaphase chromosomes of cells undergoing mitosis.

Centromeres: Centromeres are specialized regions of eukaryotic chromosomes that serve as the sites of association between sister chromatids and the sites of spindle fiber attachment during mitosis.

Telomeres: Telomeres are specialized sequences required to maintain the ends of eukaryotic chromosomes.

THE SEQUENCES OF COMPLETE GENOMES

Prokaryotic Genomes: The genomes of more than 100 different bacteria, including *E. coli*, have been completely sequenced. The *E. coli* genome contains 4288 genes, with protein-coding sequences accounting for nearly 90% of the DNA.

SUMMARY

The Yeast Genome: The first eukaryotic genome to be sequenced was that of the yeast *S. cerevisiae*. The *S. cerevisiae* genome contains about 6000 genes, and protein-coding sequences account for approximately 70% of the genome. The genome of the fission yeast *S. pombe* contains fewer genes (about 5000) and more introns than *S. cerevisiae*, with protein-coding sequence corresponding to about 60% of the *S. pombe* genome.

The Genomes of *Caenorhabditis elegans* and *Drosophila melanogaster*: The genome of *C. elegans* was the first sequenced genome of a multicellular organism. The *C. elegans* genome contains about 19,000 protein-coding sequences, which account for only about 25% of the genome. The genome of *Drosophila* contains approximately 14,000 genes, with protein-coding sequences accounting for about 13% of the genome. Although *Drosophila* contains fewer genes than *C. elegans*, many genes in both species are duplicated, and it appears that both species contain 10,000–15,000 unique genes. Some of these genes are shared between *Drosophila*, *C. elegans*, and yeast—these genes may encode proteins with common functions in all eukaryotic cells. However, the majority of *Drosophila* and *C. elegans* genes are not found in yeast and are likely to function in the regulation and development of multicellular animals.

Plant Genomes: The genome of the small flowering plant *Arabidopsis thaliana* contains approximately 26,000 genes—surprisingly more genes than were found in either *Drosophila* or *C. elegans*. However, many of these genes are the result of duplications of large segments of the *Arabidopsis* genome, so the number of unique genes in *Arabidopsis* is about 15,000. Many of these genes are unique to plants, including genes involved in plant physiology, development, and defense. The sequence of the rice genome is of particular agricultural interest because rice is the staple food for more than half the world's population. The draft sequence of the rice genome is estimated to contain approximately 37,000 genes, many of which are duplicated and may have arisen by duplication of large genome segments.

The Human Genome: The human genome appears to contain 20,000–25,000 genes—not much more than the number of genes found in simpler animals like *Drosophila* and *C. elegans*. Over 40% of the predicted human proteins are related to proteins found in other sequenced organisms, including *Drosophila* and *C. elegans*. In addition, the human genome contains expanded numbers of genes involved in the nervous system, the immune system, blood clotting, development, cell signaling, and the regulation of gene expression.

The Genomes of Other Vertebrates: The genomes of fish, chickens, mice, rats, dogs, and chimpanzees provide important comparisons to the human genome. All of these vertebrates contain similar numbers of genes but in some cases differ substantially in their content of repetitive sequences.

KEY TERMS

yeast artificial chromosome (YAC), polytene chromosome, bacterial artificial chromosome (BAC)

fluorescence *in situ* hybridization (FISH)

KEY TERMS

bioinformatics, systems biology

SUMMARY

BIOINFORMATICS AND SYSTEMS BIOLOGY

Systematic Screens of Gene Function: The genome sequencing projects have introduced large-scale experimental and computational approaches to research in cell and molecular biology. Genome-wide screens using RNA interference can systematically identify all of the genes in an organism that are involved in any biological process that can be assayed in a high-throughput format.

Regulation of Gene Expression: The identification of gene regulatory sequences and elucidation of the signaling networks that control gene expression are major challenges in bioinformatics and systems biology. These problems are being approached by genome-wide studies of gene expression combined with the development of computational approaches to identify functional regulatory elements.

Variation among Individuals and Genomic Medicine: Variations in our genomes are responsible for the characteristics of individual people, including susceptibility to many diseases. Analysis of these variations will allow the identification of genes responsible for disease susceptibility and enable the development of new strategies for disease prevention and treatment that match the genetic makeup of different individuals.

Questions

1. Many eukaryotic organisms have genome sizes that are much larger than their complexity would seem to require. Explain this paradox.
2. How were introns discovered during studies of adenovirus mRNAs?
3. How do intron sequences in the human genome increase the diversity of proteins expressed from the limited number of 20,000–25,000 genes?
4. How can simple-sequence repetitive DNA be separated from the bulk of the nuclear DNA?
5. Yeast (*S. cerevisiae*) centromeres form a kinetochore that attaches to a single microtubule, whereas multiple microtubules are attached to the kinetochores of most animal cells. How does the structure of *S. cerevisiae* centromeres reflect this difference?
6. When circular plasmids are provided with a centromere sequence and inserted into yeast cells, they reproduce and segregate normally each cell division. However, if a linear chromosome is generated by cutting the plasmid at a single site with a restriction endonuclease, the plasmid genes are quickly lost from the yeast. Explain. What additional experiment could you perform to test your hypothesized explanation?
7. What is the average distance between genes in the human genome?
8. Approximately how many molecules of histone H1 are bound to yeast genomic DNA?
9. What is the average length of an intron in a human gene?
10. You have made a library in a plasmid vector containing complete human cDNAs. What is the expected average size of an insert?
11. How was the approach used by Celera Genomics to sequence the human genome different from that used by the International Human Genome Sequencing Consortium?
12. Why is it more difficult to identify regulatory sequences than it is to identify protein coding sequences? What are the different approaches used to identify functional regulatory sequences?
13. What is a SNP? What results are expected from the study of SNPs?

References and Further Reading

The Complexity of Eukaryotic Genomes

- Berget, S. M., C. Moore and P. A. Sharp. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* 74: 3171–3175. [P]
- Breathnach, R., J. L. Mandel and P. Chambon. 1977. Ovalbumin gene is split in chicken DNA. *Nature* 270: 314–319. [P]
- Britten, R. J. and D. E. Kohne. 1968. Repeated sequences in DNA. *Science* 161: 529–540. [P]
- Chow, L. T., R. E. Gelinas, T. R. Broker and R. J. Roberts. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12: 1–8. [P]
- Fritsch, E. F., R. M. Lawn and T. Maniatis. 1980. Molecular cloning and characterization of the human β -like globin gene cluster. *Cell* 19: 959–972. [P]
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921. [P]
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945. [P]
- Kazazian, H. H., Jr. 2004. Mobile elements: Drivers of genome evolution. *Science* 303: 1626–1632. [R]
- Little, P. F. R. 1982. Globin pseudogenes. *Cell* 28: 683–684. [R]
- Maniatis, T. and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418: 236–243. [R]
- Roy, S. W. and W. Gilbert. 2005. Complex early genes. *Proc. Natl. Acad. Sci. USA* 102: 1986–1991. [P]
- Stoltzfus, A., D. F. Spencer, M. Zuker, J. M. Logsdon, Jr. and W. F. Doolittle. 1994. Testing the exon theory of genes: The evidence from protein structure. *Science* 265: 202–207. [R]
- Tilghman, S. M., P. J. Curtis, D. C. Tiemeier, P. Leder and C. Weissmann. 1978. The intervening sequence of a mouse β -globin gene is transcribed within the 15S β -globin mRNA precursor. *Proc. Natl. Acad. Sci. USA* 75: 1309–1313. [P]
- Venter, J. C. and 273 others. 2001. The sequence of the human genome. *Science* 291: 1304–1351. [P]
- Zhang, Z. and M. Gerstein. 2004. Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.* 14: 328–335. [R]

Chromosomes and Chromatin

- Blackburn, E. H. 2001. Switching and signaling at the telomere. *Cell* 106: 661–673. [R]
- Blackburn, E. H. 2005. Telomeres and telomerase: Their mechanisms of action and the effects of altering their functions. *FEBS Letters* 579: 859–862. [R]
- Blasco, M. A. 2005. Telomeres and human disease: Ageing, cancer and beyond. *Nature Rev. Genet.* 6: 611–622. [R]
- Carbon, J. 1984. Yeast centromeres: Structure and function. *Cell* 37: 351–353. [R]
- Clarke, L. 1990. Centromeres of budding and fission yeasts. *Trends Genet.* 6: 150–154. [R]
- Dorigo, B., T. Schalch, A. Kulangara, S. Duda, R. R. Schroeder and T. J. Richmond. 2004. Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* 306: 1571–1573. [P]
- Felsenfeld, G. and M. Groudine. 2003. Controlling the double helix. *Nature* 421: 448–453. [R]
- Ferreira, M. G., K. M. Miller and J. P. Cooper. 2004. Indecent exposure: When telomeres become uncapped. *Mol. Cell* 13: 7–18. [R]
- Greider, C. W. 1999. Telomeres do D-loop-T-loop. *Cell* 97: 419–422. [R]
- Henikoff, S., and Y. Dalal. 2005. Centromeric chromatin: What makes it unique? *Curr. Opin. Genet. Dev.* 15: 177–184. [R]
- Kornberg, R. D. 1974. Chromatin structure: A repeating unit of histones and DNA. *Science* 184: 868–871. [P]
- Koshland, D. and A. Strunnikov. 1996. Mitotic chromosome condensation. *Ann. Rev. Cell Biol.* 12: 305–333. [R]
- Luger, K., A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389: 251–260. [P]
- Pardue, M. L. and J. G. Gall. 1970. Chromosomal localization of mouse satellite DNA. *Science* 168: 1356–1358. [P]
- Paulson, J. R. and U. K. Laemmli. 1977. The structure of histone-depleted metaphase chromosomes. *Cell* 12: 817–828. [P]
- Richmond, T. J., J. T. Finch, B. Rushton, D. Rhodes and A. Klug. 1984. Structure of the nucleosome core particle at 7 Å resolution. *Nature* 311: 532–537. [P]
- Schalch, T., S. Duda, D. F. Sargent and T. J. Richmond. 2005. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* 436: 138–141. [P]

- Schueler, M. G., A. W. Higgins, M. Katharine Rudd, K. Gustashaw and H. F. Willard. 2001. Genomic and genetic definition of a functional human centromere. *Science* 294: 109–115. [P]
- Sun, X., J. Wahlstrom and G. Karpen. 1997. Molecular structure of a functional *Drosophila* centromere. *Cell* 91: 1007–1019. [P]
- Szostak, J. W. and E. H. Blackburn. 1982. Cloning yeast telomeres on linear plasmid vectors. *Cell* 29: 245–255. [P]
- Zakian, V. A. 1995. Telomeres: Beginning to understand the end. *Science* 270: 1601–1607. [R]

The Sequences of Complete Genomes

- Adams, M. D. and 194 others. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195. [P]
- Aparicio, S. and 40 others. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297: 1301–1310. [P]
- Baltimore, D. 2001. Our genome unveiled. *Nature* 409: 814–816. [R]
- Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1462. [P]
- Bult, C. J. and 39 others. 1996. Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. *Science* 273: 1058–1073. [P]
- Chakravarti, A. 2001. Single nucleotide polymorphisms: To a future of genetic medicine. *Nature* 409: 822–823. [R]
- Chervitz, S. A., L. Aravind, G. Sherlock, C. A. Ball, E. V. Koonin, S. S. Dwight, M. A. Harris, K. Dolinski, S. Mohr, T. Smith, S. Weng, J. M. Cherry and D. Botstein. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* 282: 2022–2028. [R]
- Ellegren, H. 2005. The dog has its day. *Nature* 438: 745–746. [R]
- Fleischmann, R. D., and 39 others. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512. [P]
- Fraser, C. M. and 28 others. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403. [P]

- Goff, S. A. and 54 others. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Japonica*). *Science* 296: 92–100. [P]
- Goffeau, A. and 15 others. 1996. Life with 6000 genes. *Science* 274: 546–567. [P]
- Goldstein, D. B. and G. L. Cavalleri. 2005. Understanding human diversity. *Nature* 437: 1241–1242. [R]
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695–715. [P]
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921. [P]
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945. [P]
- International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* 436: 793–800. [P]
- Kirkness, E. F., V. Bafna, A. L. Halpern, S. Levy, K. Remington, D. B. Rusch, A. L. Delcher, M. Pop, W. Wang, C. M. Fraser and J. C. Venter. 2003. The dog genome: Survey sequencing and comparative analysis. *Science* 301: 1898–1903. [P]
- Lindblad-Toh, K. and 46 others. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438: 803–819. [P]
- Martienssen, R. and W. R. McCombie. 2001. The first plant genome. *Cell* 105: 571–574. [R]
- Mouse Genome Sequencing Consortium, 2002. Initial sequence and comparative analysis of the mouse genome. *Nature* 420: 520–562. [P]
- Oliver, S. G. and 146 others. 1992. The complete DNA sequence of yeast chromosome III. *Nature* 357: 38–46. [P]
- Peltonen, L. and V. A. McKusick. 2001. Dissecting human disease in the postgenomic era. *Science* 291: 1224–1229. [R]
- Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428: 493–521. [P]
- Rubin, G. M. 2001. The draft sequences: Comparing species. *Nature* 409: 820–821. [R]
- Rubin, G. M. and 54 others. 2000. Comparative genomics of the eukaryotes. *Science* 287: 2204–2215. [R]
- Sutter, N. B. and E. A. Ostrander. 2004. Dog star rising: The canine genetic system. *Nature Rev. Genet.* 5: 900–910. [R]
- The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815. [P]
- The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282: 2012–2018. [P]
- The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69–87. [P]
- The International Chimpanzee Chromosome 22 Consortium. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429: 382–388. [P]
- Venter, J. C. and 273 others. 2001. The sequence of the human genome. *Science* 291: 1304–1351. [P]
- Walbot, V. 2000. A green chapter in the book of life. *Nature* 408: 794–795. [R]
- Wood, V. and 132 others. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415: 871–880. [P]
- Yu, J. and 99 others. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indica*). *Science* 296: 79–92. [P]
- Bioinformatics and Systems Biology**
- Bell, J. 2004. Predicting disease using genomics. *Nature* 429: 453–456. [R]
- Ehrenberg, M., J. Elf, E. Aurell, R. Sandberg and J. Tegner. 2003. Systems biology is taking off. *Genome Res.* 13: 2377–2380. [R]
- Friedman, A. and N. Perrimon. 2004. Genome-wide high-throughput screens in functional genomics. *Curr. Opin. Genet. Dev.* 14: 470–476. [R]
- Ge, H., A. J. M. Walhout and M. Vidal. 2003. Integrating ‘omic’ information: A bridge between genomics and systems biology. *Trends Genet.* 19: 551–559. [R]
- Guttmacher, A. E. and F. S. Collins. 2002. Genomic medicine—a primer. *N. Engl. J. Med.* 347: 1512–1520. [R]
- Harbison, C. T. and 19 others. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431: 99–104. [P]
- Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer and D. R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* 307: 1072–1079. [P]
- Kirschner, M. W. 2005. The meaning of systems biology. *Cell* 121: 503–504. [R]
- Kitano, H. 2002. Systems biology: A brief overview. *Science* 295: 1662–1664. [R]
- Sieburth, D., Q. Ch’ng, M. Dybbs, M. Tavazoie, S. Kennedy, D. Wang, D. Dupuy, J.-F. Rual, D. E. Hill, M. Vidal, G. Ruvkun and J. M. Kaplan. 2005. Systematic analysis of genes required for synapse structure and function. *Nature* 436: 510–517. [P]
- Vavouri, T. and G. Elgar. 2005. Prediction of *cis*-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. *Curr. Opin. Genet. Dev.* 15: 395–402. [R]
- Wasserman, W. W. and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet.* 5: 276–287. [R]
- Xie, X., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals. *Nature* 434: 338–345. [P]