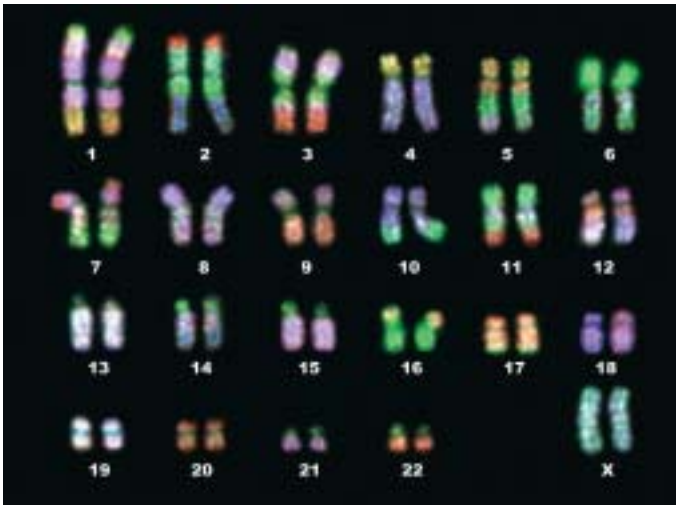
10

MOLECULAR STRUCTURE OF GENES AND CHROMOSOMES



These brightly colored RxFISH-painted chromosomes are both beautiful and useful in revealing chromosome anomalies and in comparing karyotypes of different species. [© Department of Clinical Cytogenetics, Addenbrookes Hospital/Photo Researchers, Inc.]

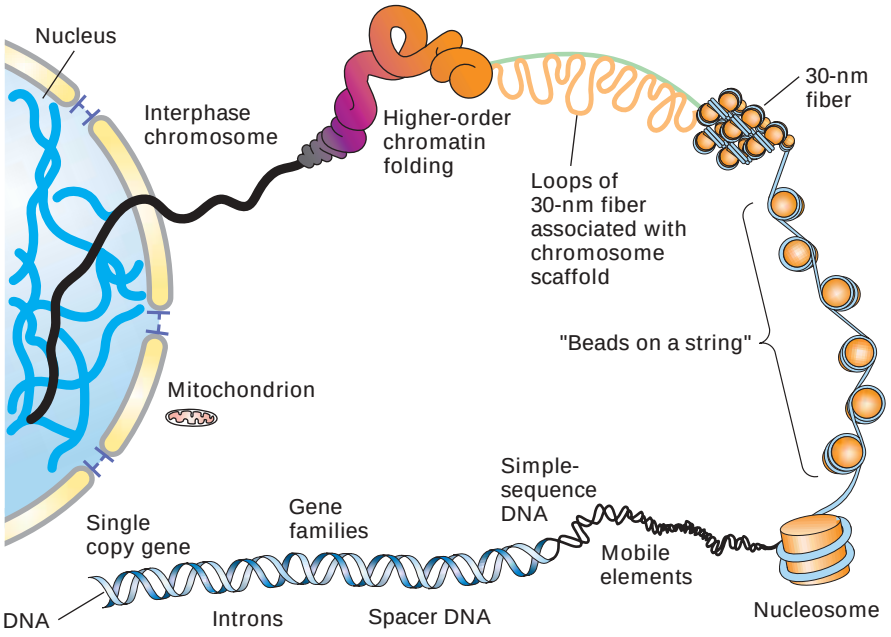
By the beginning of the twenty-first century, molecular biologists had completed sequencing the entire genomes of hundreds of viruses, scores of bacteria, and the budding yeast *S. cerevisiae,* a unicellular eukaryote. In addition, the vast majority of the genome sequence is now known for several multicellular eukaryotes including the roundworm *C. elegans,* the fruit fly *D. melanogaster,* and humans (see Figure 9-34). Detailed analysis of these sequencing data has revealed that a large portion of the genomes of higher eukaryotes does not encode mRNAs or any other RNAs required by the organism. Remarkably, such noncoding DNA constitutes more than 95 percent of human chromosomal DNA.

The noncoding DNA in multicellular organisms contains many regions that are similar but not identical. Variations within some stretches of this *repetitious DNA* are so great that each single person can be distinguished by a DNA "fingerprint" based on these sequence variations. Moreover, some repetitious DNA sequences are not found in constant positions in the DNA of individuals of the same species. Such "mobile" DNA elements, which are present in both prokaryotic and eukaryotic organisms, can cause mutations when they move to new sites in the genome. Even though they generally have no function in the life cycle of an individual organism, mobile elements probably have played an important role in evolution.

In higher eukaryotes, DNA regions encoding proteins—that is, **genes**—lie amidst this expanse of apparently nonfunctional DNA. In addition to the nonfunctional DNA *between* genes, noncoding **introns** are common *within* genes of multicellular plants and animals. Introns are less common, but sometimes present, in single-celled eukaryotes and very rare in bacteria. Sequencing of the same protein-coding gene in a variety of eukaryotic species has shown that evolutionary pressure selects for maintenance of relatively similar sequences in the coding regions, or **exons.** In contrast, wide sequence variation, even including total loss, occurs among introns, suggesting that most intron sequences have little functional significance.

The sheer length of cellular DNA is a significant problem with which cells must contend. The DNA in a single human cell, which measures about 2 meters in total length, must be contained within cells with diameters of less than 10 μm, a compaction ratio of greater than 10^5. Specialized eukaryotic proteins associated with nuclear DNA fold and organize it into the structures of DNA and protein visualized as individual **chromosomes** during mitosis. Mitochondria and

OUTLINE



▶ FIGURE 10-1 **Overview of the structure of genes and chromosomes.** DNA of higher eukaryotes consists of unique and repeated sequences. Only ~5% of human DNA encodes proteins and functional RNAs and the regulatory sequences that control their expression; the remainder is merely spacer DNA between genes and introns within genes. Much of this DNA, ~50% in humans, is derived from mobile DNA elements, genetic symbiots that have contributed to the evolution of contemporary genomes. Each chromosome consists of a single, long molecule of DNA up to ~280 Mb in humans, organized into increasing levels of condensation by the histone and nonhistone proteins with which it is intricately complexed. Much smaller DNA molecules are localized in mitochondria and chloroplasts.

chloroplasts also contain DNA, probably evolutionary remnants of their origins, that encode essential components of these vital organelles.

In this chapter we first present a molecular definition of genes and the complexities that arise in higher organisms from the processing of mRNA precursors into alternatively spliced mRNAs. Next we discuss the main classes of eukaryotic DNA and the special properties of mobile DNA. We also consider the packaging of DNA and proteins into compact complexes, the large-scale structure of chromosomes, and the functional elements required for chromosome duplication and segregation. In the final section, we consider organelle DNA and how it differs from nuclear DNA. Figure 10-1 provides an overview of these interrelated subjects.

10.1 Molecular Definition of a Gene

In molecular terms, a gene commonly is defined as *the entire nucleic acid sequence that is necessary for the synthesis of a functional gene product (polypeptide or RNA).* According to this definition, a gene includes more than the nucleotides encoding the amino acid sequence of a protein, referred to as the *coding region*. A gene also includes all the DNA sequences required for synthesis of a particular RNA transcript. In eukaryotic genes, transcription-control regions known as **enhancers** can lie 50 kb or more from the coding region. Other critical noncoding regions in eukaryotic genes are the sequences that specify 3′ cleavage and polyadenylation, known as *poly(A) sites,* and splicing of primary RNA transcripts, known as *splice sites* (see Figure 4-14). Mutations in these RNA-processing signals prevent expression of a functional mRNA and thus of the encoded polypeptide.

Although most genes are transcribed into mRNAs, which encode proteins, clearly some DNA sequences are transcribed into RNAs that do not encode proteins (e.g., tRNAs and rRNAs). However, because the DNA that encodes tRNAs and rRNAs can cause specific phenotypes when it is mutated, these DNA regions generally are referred to as tRNA and rRNA *genes,* even though the final products of these genes are RNA molecules and not proteins. Many other RNA molecules described in later chapters also are transcribed from non-protein-coding genes.

Most Eukaryotic Genes Produce Monocistronic mRNAs and Contain Lengthy Introns

As discussed in Chapter 4, many bacterial mRNAs are *polycistronic;* that is, a single mRNA molecule (e.g., the mRNA encoded by the *trp* operon) includes the coding region for several proteins that function together in a biological process. In contrast, most eukaryotic mRNAs are *monocistronic;* that is, each mRNA molecule encodes a single protein. This difference between polycistronic and monocistronic mRNAs correlates with a fundamental difference in their translation.

Within a bacterial polycistronic mRNA a ribosome-binding site is located near the start site for each of the protein-coding regions, or cistrons, in the mRNA. Translation initiation can begin at any of these multiple internal sites, producing multiple proteins (see Figure 4-12a). In most eukaryotic mRNAs, however, the 5′-cap structure directs ribosome binding, and translation begins at the closest AUG start codon (see Figure 4-12b). As a result, translation begins only at this site. In many cases, the primary transcripts of eukaryotic protein-coding genes are processed into a single type of

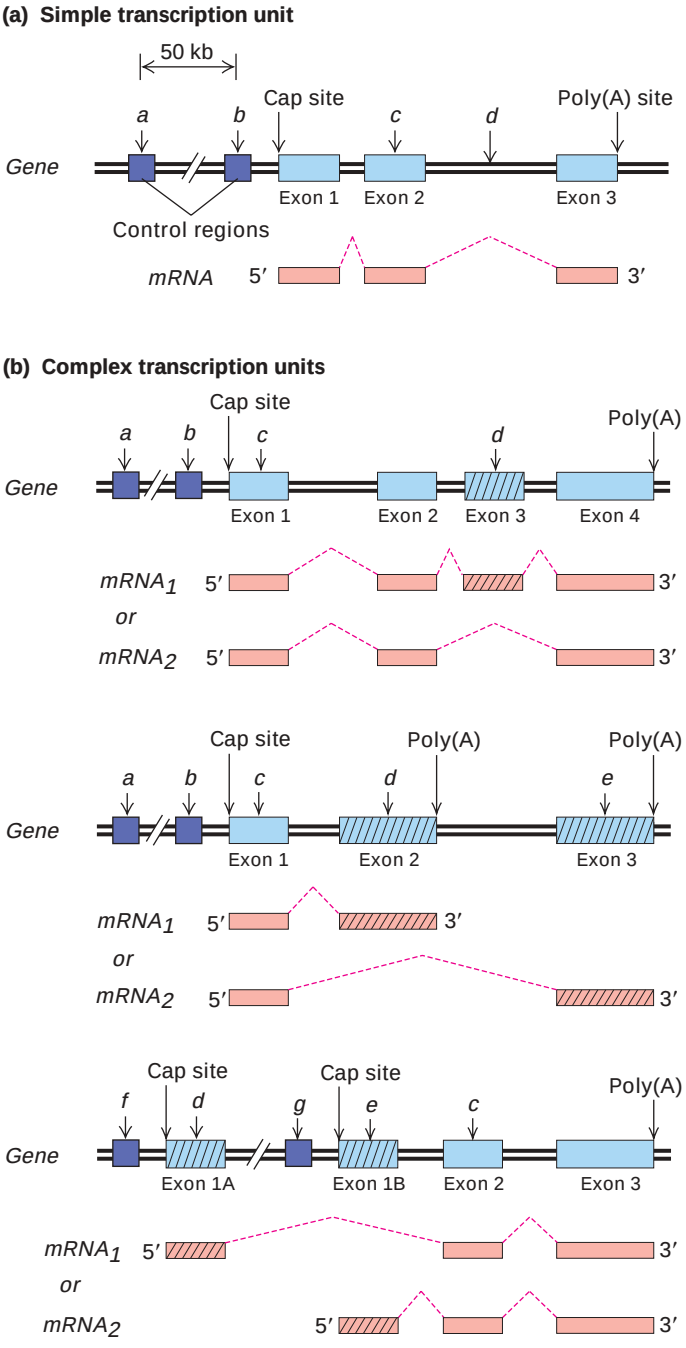
mRNA, which is translated to give a single type of polypeptide (see Figure 4-14).

Unlike bacterial and yeast genes, which generally lack introns, most genes in multicellular animals and plants contain introns, which are removed during RNA processing. In many cases, the introns in a gene are considerably longer than the exons. For instance, of the ≈50,000 base pairs composing many human genes encoding average-size proteins, more than 95 percent are present in introns and noncoding 5′ and 3′ regions. Many large proteins in higher organisms have repeated domains and are encoded by genes consisting of repeats of similar exons separated by introns of variable length. An example of this is fibronectin, a component of the extracellular matrix that is encoded by a gene containing multiple copies of three types of exons (see Figure 4-15).

Simple and Complex Transcription Units Are Found in Eukaryotic Genomes

The cluster of genes that form a bacterial operon comprises a single **transcription unit,** which is transcribed from a particular promoter into a single primary transcript. In other words, genes and transcription units often are distinguishable in prokaryotes. In contrast, most eukaryotic genes and transcription units generally are identical, and the two terms commonly are used interchangeably. Eukaryotic transcription units, however, are classified into two types, depending on the fate of the primary transcript.

The primary transcript produced from a *simple* transcription unit is processed to yield a single type of mRNA, encoding a single protein. Mutations in exons, introns, and transcription-control regions all may influence expression of the protein encoded by a simple transcription unit (Figure 10-2a).



▶ FIGURE 10-2 **Comparison of simple and complex eukaryotic transcription units.** (a) A simple transcription unit includes a region that encodes one protein, extending from the 5′ cap site to the 3′ poly(A) site, and associated control regions. Introns lie between exons (blue rectangles) and are removed during processing of the primary transcripts (dashed red lines); they thus do not occur in the functional monocistronic mRNA. Mutations in a transcription-control region (*a, b*) may reduce or prevent transcription, thus reducing or eliminating synthesis of the encoded protein. A mutation within an exon (*c*) may result in an abnormal protein with diminished activity. A mutation within an intron (*d*) that introduces a new splice site results in an abnormally spliced mRNA encoding a nonfunctional protein. (b) Complex transcription units produce primary transcripts that can be processed in alternative ways. (*Top*) If a primary transcript contains alternative splice sites, it can be processed into mRNAs with the same 5′ and 3′ exons but different internal exons. (*Middle*) If a primary transcript has two poly(A) sites, it can be processed into mRNAs with alternative 3′ exons. (*Bottom*) If alternative promoters (*f* or *g*) are active in different cell types, $mRNA_1$, produced in a cell type in which *f* is activated, has a different exon (1A) than $mRNA_2$ has, which is produced in a cell type in which *g* is activated (and where exon 1B is used). Mutations in control regions (*a* and *b*) and those designated *c* within exons shared by the alternative mRNAs affect the proteins encoded by both alternatively processed mRNAs. In contrast, mutations (designated *d* and *e*) within exons unique to one of the alternatively processed mRNAs affect only the protein translated from that mRNA. For genes that are transcribed from different promoters in different cell types *(bottom)*, mutations in different control regions (*f* and *g*) affect expression only in the cell type in which that control region is active.

In the case of *complex* transcription units, which are quite common in multicellular organisms, the primary RNA transcript can be processed in more than one way, leading to formation of mRNAs containing different exons. Each mRNA, however, is monocistronic, being translated into a single polypeptide, with translation usually initiating at the first AUG in the mRNA. Multiple mRNAs can arise from a primary transcript in three ways (Figure 10-2b):

1. Use of different splice sites, producing mRNAs with the same 5′ and 3′ exons but different internal exons. Figure 10-2b (*top*) shows one example of this type of alternative RNA processing, *exon skipping.*

2. Use of alternative poly(A) sites, producing mRNAs that share the same 5′ exons but have different 3′ exons (Figure 10-2b [*middle*]).

3. Use of alternative promoters, producing mRNAs that have different 5′ exons and common 3′ exons. A gene expressed selectively in two or more types of cells is often transcribed from distinct cell-type-specific promoters (Figure 10-2b [*bottom*]).

Examples of all three types of alternative RNA processing occur during sexual differentiation in *Drosophila* (see Figure 12-14). Commonly, one mRNA is produced from a complex transcription unit in some cell types, and an alternative mRNA is made in other cell types. For example, differences in RNA splicing of the primary fibronectin transcript in fibroblasts and hepatocytes determines whether or not the secreted protein includes domains that adhere to cell surfaces (see Figure 4-15).

The relationship between a mutation and a gene is not always straightforward when it comes to complex transcription units. A mutation in the control region or in an exon shared by alternative mRNAs will affect all the alternative proteins encoded by a given complex transcription unit. On the other hand, mutations in an exon present in only one of the alternative mRNAs will affect only the protein encoded by that mRNA. As explained in Chapter 9, **genetic complementation** tests commonly are used to determine if two mutations are in the same or different genes (see Figure 9-7). However, in the complex transcription unit shown in Figure 10-2b (*middle*), mutations *d* and *e* would complement each other in a genetic complementation test, even though they occur in the same gene. This is because a chromosome with mutation *d* can express a normal protein encoded by $mRNA_2$ and a chromosome with mutation *e* can express a normal protein encoded by $mRNA_1$. However, a chromosome with mutation *c* in an exon common to both mRNAs would not complement either mutation *d* or *e*. In other words, mutation *c* would be in the same complementation groups as mutations *d* and *e*, even though *d* and *e* themselves would not be in the same complementation group!

KEY CONCEPTS OF SECTION 10.1

Molecular Definition of a Gene

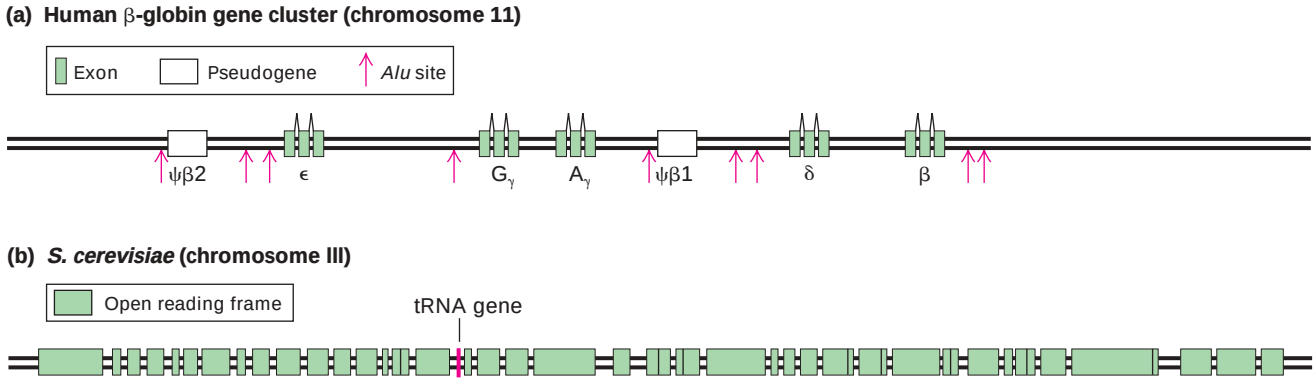
- In molecular terms, a gene is the entire DNA sequence required for synthesis of a functional protein or RNA molecule. In addition to the coding regions (exons), a gene includes control regions and sometimes introns.
- Most bacterial and yeast genes lack introns, whereas most genes in multicellular organisms contain introns. The total length of intron sequences often is much longer than that of exon sequences.
- A simple eukaryotic transcription unit produces a single monocistronic mRNA, which is translated into a single protein.
- A complex eukaryotic transcription unit is transcribed into a primary transcript that can be processed into two or more different monocistronic mRNAs depending on the choice of splice sites or polyadenylation sites (see Figure 10-2b).
- Many complex transcription units (e.g., the fibronectin gene) express one mRNA in one cell type and an alternative mRNA in a different cell type.

10.2 Chromosomal Organization of Genes and Noncoding DNA

Having reviewed the relation between transcription units and genes, we now consider the organization of genes on chromosomes and the relationship of noncoding DNA sequences to coding sequences.

Genomes of Many Organisms Contain Much Nonfunctional DNA

Comparisons of the total chromosomal DNA per cell in various species first suggested that much of the DNA in certain organisms does not encode RNA or have any apparent regulatory or structural function. For example, yeasts, fruit flies, chickens, and humans have successively more DNA in their haploid chromosome sets (12; 180; 1300; and 3300 Mb, respectively), in keeping with what we perceive to be the increasing complexity of these organisms. Yet the vertebrates with the greatest amount of DNA per cell are amphibians, which are surely less complex than humans in their structure and behavior. Even more surprising, the unicellular protozoal species *Amoeba dubia* has 200 times more DNA per cell than humans. Many plant species also have considerably more DNA per cell than humans have. For example, tulips have 10 times as much DNA per cell as humans. The DNA content per cell also varies considerably between closely related species. All insects or all amphibians would appear to be similarly complex, but the amount of haploid DNA in



▲ FIGURE 10-3 **Comparative density of genes in ≈80-kb regions of genomic DNA from humans and the yeast *S. cerevisiae*.** (a) In the diagram of the β-globin gene cluster on human chromosome 11, the green boxes represent exons of β-globin–related genes. Exons spliced together to form one mRNA are connected by caret-like spikes. The human β-globin gene cluster contains two pseudogenes (white); these regions are related to the functional globin-type genes but are not transcribed. Each red arrow indicates the location of an *Alu* sequence, an ≈300-bp noncoding repeated sequence that is abundant in the human genome. (b) In the diagram of yeast DNA from chromosome III, the green boxes indicate open reading frames. Most of these potential protein-coding sequences probably are functional genes without introns. Note the much higher proportion of noncoding-to-coding sequences in the human DNA than in the yeast DNA. [Part (a), see F. S. Collins and S. M. Weissman, 1984, *Prog. Nucl. Acid Res. Mol. Biol.* **31**:315; part (b), see S. G. Oliver et al., 1992, *Nature* **357**:28.]

species within each of these phylogenetic classes varies by a factor of 100.

Detailed sequencing and identification of exons in chromosomal DNA have provided direct evidence that the genomes of higher eukaryotes contain large amounts of noncoding DNA. For instance, only a small portion of the β-globin gene cluster of humans, about 80 kb long, encodes protein (Figure 10-3a). Moreover, compared with other regions of vertebrate DNA, the β-globin gene cluster is unusually rich in protein-coding sequences, and the introns in globin genes are considerably shorter than those in many human genes. In contrast, a typical 80-kb stretch of DNA from the yeast *S. cerevisiae*, a single-celled eukaryote (Figure 10-3b) contains many closely spaced protein-coding sequences without introns and relatively much less noncoding DNA.

The density of genes varies greatly in different regions of human chromosomal DNA, from "gene-rich" regions, such as the β-globin cluster, to large gene-poor "deserts." Of the 94 percent of human genomic DNA that has been sequenced, only ≈1.5 percent corresponds to protein-coding sequences (exons). Most human exons contain 50–200 base pairs, although the 3′ exon in many transcription units is much longer. Human introns vary in length considerably. Although many are ≈90 bp long, some are much longer; their median length is 3.3 kb. Approximately one-third of human genomic DNA is thought to be transcribed into pre-mRNA precursors, but some 95 percent of these sequences are in introns, which are removed by RNA splicing.

Different selective pressures during evolution may account, at least in part, for the remarkable difference in the amount of nonfunctional DNA in unicellular and multicellular organisms. For example, microorganisms must compete for limited amounts of nutrients in their environment, and metabolic economy thus is a critical characteristic. Since synthesis of nonfunctional (i.e., noncoding) DNA requires time and energy, presumably there was selective pressure to lose nonfunctional DNA during the evolution of microorganisms. On the other hand, natural selection in vertebrates depends largely on their behavior. The energy invested in DNA synthesis is trivial compared with the metabolic energy required for the movement of muscles; thus there was little selective pressure to eliminate nonfunctional DNA in vertebrates.

Protein-Coding Genes May Be Solitary or Belong to a Gene Family

The nucleotide sequences within chromosomal DNA can be classified on the basis of structure and function, as shown in Table 10-1. We will examine the properties of each class, beginning with protein-coding genes, which comprise two groups.

In multicellular organisms, roughly 25–50 percent of the protein-coding genes are represented only once in the haploid genome and thus are termed *solitary* genes. A well-studied example of a solitary protein-coding gene is the chicken lysozyme gene. The 15-kb DNA sequence encoding chicken lysozyme constitutes a simple transcription unit containing four exons and three introns. The flanking regions, extending for about 20 kb upstream and downstream from the transcription unit, do not encode any detectable mRNAs. Lysozyme, an enzyme that cleaves the polysaccharides in bacterial cell walls, is an abundant

TABLE 10-1 Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	≈15* (0.8)†
Duplicated or diverged genes in gene families	Variable	2–≈1000	≈15* (0.8)†
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINEs	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes	Variable	1–≈100	≈0.4
Unclassified spacer DNA	Variable	n.a.‡	≈25

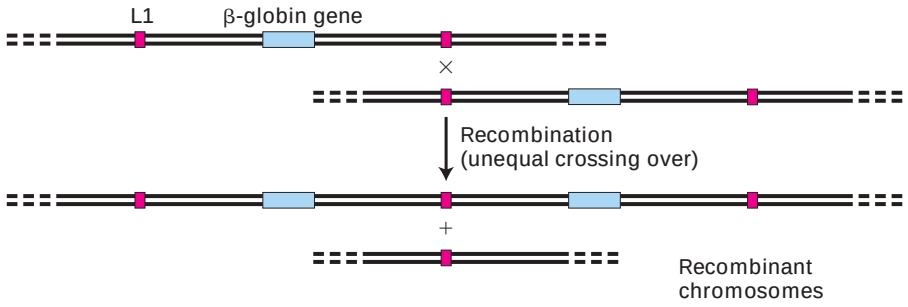
*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.

SOURCE: E. S. Lander et al., 2001, *Nature* **409**:860.

component of chicken egg-white protein and also is found in human tears. Its activity helps to keep the surface of the eye and the chicken egg sterile.

Duplicated genes constitute the second group of protein-coding genes. These are genes with close but nonidentical sequences that generally are located within 5–50 kb of one another. In vertebrate genomes, duplicated genes probably constitute half the protein-coding DNA sequences. A set of duplicated genes that encode proteins with similar but nonidentical amino acid sequences is called a **gene family;** the encoded, closely related, homologous proteins constitute a **protein family.** A few protein families, such as protein kinases, transcription factors, and vertebrate immunoglobulins, include hundreds of members. Most protein families, however, include from just a few to 30 or so members; common examples are cytoskeletal proteins, 70-kDa heat-shock proteins, the myosin heavy chain, chicken ovalbumin, and the α- and β-globins in vertebrates.

The genes encoding the β-like globins are a good example of a gene family. As shown in Figure 10-3a, the β-like globin gene family contains five functional genes designated β, δ, A_γ, G_γ, and ϵ; the encoded polypeptides are similarly designated. Two identical β-like globin polypeptides combine with two identical α-globin polypeptides (encoded by another gene family) and four small heme groups to form a hemoglobin molecule (see Figure 3-10). All the hemoglobins formed from the different β-like globins carry oxygen in the blood, but they exhibit somewhat different properties that are suited to specific roles in human physiology. For example, hemoglobins containing either the A_γ or G_γ polypeptides are expressed only during fetal life. Because these fetal hemoglobins have a higher affinity for oxygen than adult hemoglobins, they can effectively extract oxygen from the maternal circulation in the placenta. The lower oxygen affinity of adult hemoglobins, which are expressed after birth, permits better release of oxygen to the tissues, espe-



▲ FIGURE 10-4 **Gene duplication resulting from unequal crossing over.** Each parental chromosome (*top*) contains one ancestral β-globin gene containing three exons and two introns. Homologous noncoding L1 repeated sequences lie 5′ and 3′ of the β-globin gene. The parental chromosomes are shown displaced relative to each other, so that the L1 sequences are aligned. Homologous recombination between L1 sequences as shown would generate one recombinant chromosome with two copies of the β-globin gene and one chromosome with a deletion of the β-globin gene. Subsequent independent mutations in the duplicated genes could lead to slight changes in sequence that might result in slightly different functional properties of the encoded proteins. Unequal crossing over also can result from rare recombinations between unrelated sequences. [See D. H. A. Fitch et al., 1991, *Proc. Nat'l. Acad. Sci. USA* **88**:7396.]

cially muscles, which have a high demand for oxygen during exercise.

The different β-globin genes probably arose by duplication of an ancestral gene, most likely as the result of an "unequal crossover" during meiotic recombination in a developing germ cell (egg or sperm) (Figure 10-4). Over evolutionary time the two copies of the gene that resulted accumulated random mutations; beneficial mutations that conferred some refinement in the basic oxygen-carrying function of hemoglobin were retained by natural selection, resulting in *sequence drift*. Repeated gene duplications and subsequent sequence drift are thought to have generated the contemporary globin-like genes observed in humans and other complex species today.

Two regions in the human β-like globin gene cluster contain nonfunctional sequences, called **pseudogenes,** similar to those of the functional β-like globin genes (see Figure 10-3a). Sequence analysis shows that these pseudogenes have the same apparent exon-intron structure as the functional β-like globin genes, suggesting that they also arose by duplication of the same ancestral gene. However, sequence drift during evolution generated sequences that either terminate translation or block mRNA processing, rendering such regions nonfunctional even if they were transcribed into RNA. Because such pseudogenes are not deleterious, they remain in the genome and mark the location of a gene duplication that occurred in one of our ancestors. As discussed in a later section, other nonfunctional gene copies can arise by reverse transcription of mRNA into cDNA and integration of this intron-less DNA into a chromosome.

Several different gene families encode the various proteins that make up the cytoskeleton. These proteins are present in varying amounts in almost all cells. In vertebrates, the major cytoskeletal proteins are the actins, tubulins, and intermediate filament proteins like the keratins. We examined the origin of one such family, the tubulin family, in the last chapter (see Figure 9-32). Although the physiological rationale for the cytoskeletal protein families is not as obvious as it is for the globins, the different members of a family probably have similar but subtly different functions suited to the particular type of cell in which they are expressed.

Tandemly Repeated Genes Encode rRNAs, tRNAs, and Histones

In vertebrates and invertebrates, the genes encoding rRNAs and some other noncoding RNAs such as some of the snRNAs involved in RNA splicing occur as *tandemly repeated arrays.* These are distinguished from the duplicated genes of gene families in that the multiple tandemly repeated genes encode identical or nearly identical proteins or functional RNAs. Most often copies of a sequence appear one after the other, in a head-to-tail fashion, over a long stretch of DNA. Within a tandem array of rRNA genes, each copy is exactly, or almost exactly, like all the others. Although the transcribed portions of rRNA genes are the same in a given individual, the nontranscribed spacer regions between the transcribed regions can vary.

The tandemly repeated rRNA, tRNA, and histone genes are needed to meet the great cellular demand for their transcripts. To understand why, consider that a fixed maximal number of RNA copies can be produced from a single gene during one cell generation when the gene is fully loaded with RNA polymerase molecules. If more RNA is required than can be transcribed from one gene, multiple copies of the gene are necessary. For example, during early embryonic development in humans, many embryonic cells have a doubling time of ≈24 hours and contain 5–10 million ribosomes. To produce enough rRNA to form this many ribosomes, an embryonic human cell needs at least 100 copies of the large and small subunit rRNA genes, and most of these must be close

TABLE 10-2 Effect of Gene Copy Number and RNA Polymerase Loading on Rate of Pre-rRNA Synthesis in Humans

Copies of Pre-rRNA Gene	RNA Polymerase Molecules per Gene	Molecules of Pre-rRNA Produced in 24 Hours
1	1	288
1	≈250	≈70,000
100	≈250	≈7,000,000

to maximally active for the cell to divide every 24 hours (Table 10-2). That is, multiple RNA polymerases must be loaded onto and transcribing each rRNA gene at the same time (see Figure 12-32).

All eukaryotes, including yeasts, contain 100 or more copies of the genes encoding 5S rRNA and the large and small subunit rRNAs. The importance of repeated rRNA genes is illustrated by *Drosophila* mutants called *bobbed* (because they have stubby wings), which lack a full complement of the tandemly repeated **pre-rRNA** genes. A *bobbed* mutation that reduces the number of pre-rRNA genes to less than ≈50 is a recessive lethal mutation.

Multiple copies of tRNA and histone genes also occur, often in clusters, but generally not in tandem arrays.

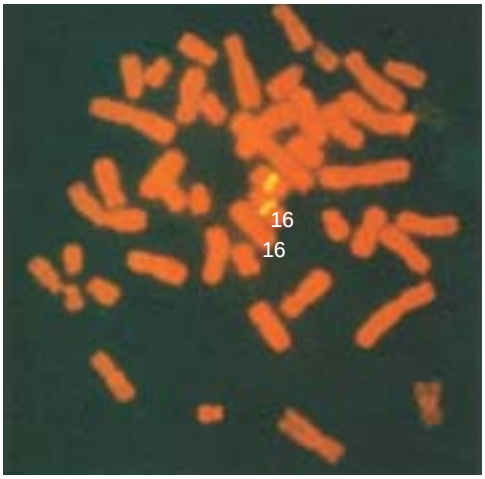
Most Simple-Sequence DNAs Are Concentrated in Specific Chromosomal Locations

Besides duplicated protein-coding genes and tandemly repeated genes, eukaryotic cells contain multiple copies of other DNA sequences in the genome, generally referred to as *repetitious DNA* (see Table 10-1). Of the two main types of repetitious DNA, the less prevalent is **simple-sequence DNA**, which constitutes about 3 percent of the human genome and is composed of perfect or nearly perfect repeats of relatively short sequences. The more common type of repetitious DNA, composed of much longer sequences, is discussed in Section 10.3.

Simple-sequence DNA is commonly called *satellite* DNA because in early studies of DNAs from higher organisms using equilibrium buoyant-density ultracentrifugation some simple-sequence DNAs banded at a different position from the bulk of cellular DNA. These were called *satellite bands* to distinguish them from the main band of DNA in the buoyant-density gradient. Simple-sequence DNAs in which the repeats contain 1–13 base pairs are often called *microsatellites*. Most have repeat lengths of 1–4 base pairs and usually occur in tandem repeats of 150 base pairs or fewer. Microsatellites are thought to have originated by "backward slippage" of a daughter strand on its template strand during DNA replication so that the same short sequence is copied twice.

MEDICINE

Microsatellites occasionally occur within transcription units. Some individuals are born with a larger number of repeats in specific genes than observed in the general population, presumably because of daughter-strand slippage during DNA replication in a germ cell from which they developed. Such expanded microsatellites have been found to cause at least 14 different types of neuromuscular diseases, depending on the gene in which they occur. In some cases expanded microsatellites behave like a recessive mutation because they interfere with the function or expression of the encoded gene. But in the more common types of diseases associated with expanded microsatellite repeats, *myotonic dystrophy* and *spinocerebellar ataxia*, the expanded repeats behave like dominant mutations because they interfere with RNA processing in general in the neurons where the affected genes are expressed. ▮



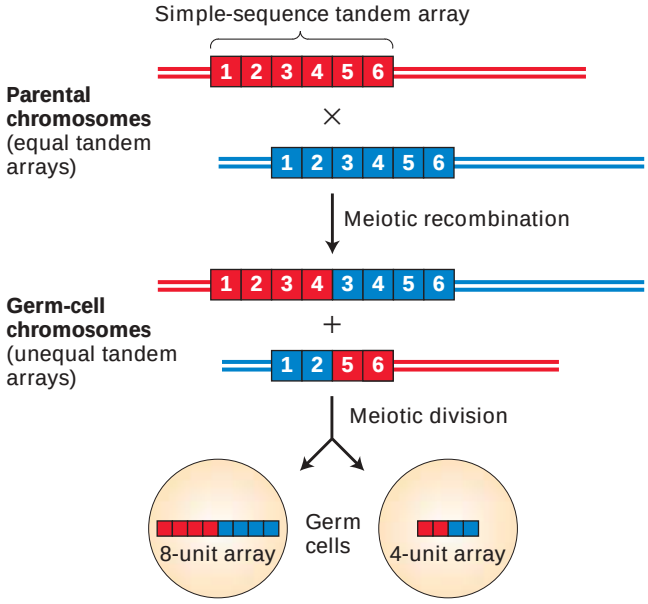
▲ EXPERIMENTAL FIGURE 10-5 **Simple-sequence DNAs are useful chromosomal markers.** Human metaphase chromosomes stained with a fluorescent dye were hybridized in situ with a particular simple-sequence DNA labeled with a fluorescent biotin derivative. When viewed under the appropriate wavelength of light, the DNA appears red and the hybridized simple-sequence DNA appears as a yellow band on chromosome 16, thus locating this particular simple sequence to one site in the genome. [See R. K. Moyzis et al., 1987, *Chromosoma* **95**:378; courtesy of R. K. Moyzis.]

Most satellite DNA is composed of repeats of 14–500 base pairs in tandem repeats of 20–100 kb. In situ hybridization studies with metaphase chromosomes have localized these satellite DNAs to specific chromosomal regions. In most mammals, much of this satellite DNA lies near **centromeres,** the discrete chromosomal regions that attach to spindle microtubules during mitosis and meiosis. Satellite DNA is also located at **telomeres,** the ends of chromosomes, and at specific locations within chromosome arms in some organisms. These latter sequences can be useful for identifying particular chromosomes by **fluorescence in situ hybridization (FISH)**, as illustrated in Figure 10-5.

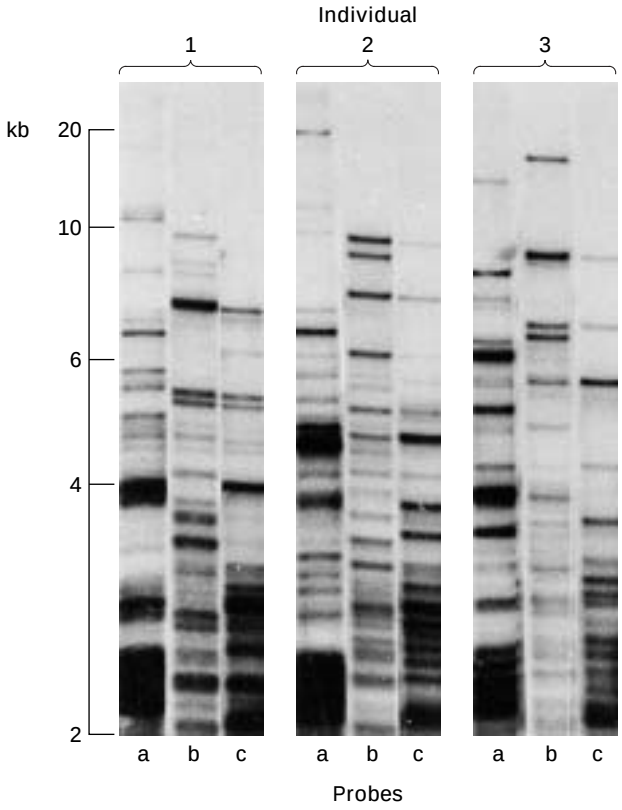
Simple-sequence DNA located at centromeres may assist in attaching chromosomes to spindle microtubules during mitosis. As yet, however, there is little clear-cut experimental evidence demonstrating any function for most simple-sequence DNA, with the exception of the short repeats at the very ends of chromosomes discussed in a later section.

DNA Fingerprinting Depends on Differences in Length of Simple-Sequence DNAs

Within a species, the nucleotide sequences of the repeat units composing simple-sequence DNA tandem arrays are highly conserved among individuals. In contrast, differences in the *number* of repeats, and thus in the length of simple-sequence tandem arrays containing the same repeat unit, are quite common among individuals. These differences in length are thought to result from unequal crossing over within regions of simple-sequence DNA during meiosis (Figure 10-6). As a consequence of this unequal crossing over, the lengths of some tandem arrays are unique in each individual.



▲ FIGURE 10-6 **Generation of differences in lengths of a simple-sequence DNA by unequal crossing over during meiosis.** In this example, unequal crossing over within a stretch of DNA containing six copies (1–6) of a particular simple-sequence repeat unit yields germ cells containing either an eight-unit or a four-unit tandem array.



▲ EXPERIMENTAL FIGURE 10-7 **Probes for minisatellite DNA can reveal unique restriction fragments (DNA fingerprints) that distinguish individuals.** DNA samples from three individuals (1, 2, and 3) were subjected to Southern blot analysis using the restriction enzyme *Hinf*1 and three different labeled minisatellites as probes (lanes a, b, and c). DNA from each individual produced a unique band pattern with each probe. Conditions of electrophoresis can be adjusted so that for each person at least 50 bands can be resolved with this restriction enzyme. The nonidentity of these three samples is easily distinguished. [From A. J. Jeffreys et al., 1985, *Nature* **316**:76; courtesy of A. J. Jeffreys.]

In humans and other mammals, some of the satellite DNA exists in relatively short 1- to 5-kb regions made up of 20–50 repeat units, each containing 15 to about 100 base pairs. These regions are called *minisatellites* to distinguish them from the more common regions of tandemly repeated satellite DNA, which are ≈20–100 kb in length. They differ from microsatellites mentioned earlier, which have very short repeat units. Even slight differences in the total lengths of various minisatellites from different individuals can be detected by **Southern blotting** of cellular DNA treated with a restriction enzyme that cuts outside the repeat sequence (Figure 10-7). The polymerase chain reaction (PCR), using

primers that hybridize to the unique sequences flanking each minisatellite, also can detect differences in minisatellite lengths between individuals. These DNA polymorphisms form the basis of *DNA fingerprinting*, which is superior to conventional fingerprinting for identifying individuals.

KEY CONCEPTS OF SECTION 10.2

Chromosomal Organization of Genes and Noncoding DNA

- In the genomes of prokaryotes and most lower eukaryotes, which contain few nonfunctional sequences, coding regions are densely arrayed along the genomic DNA.
- In contrast, vertebrate genomes contain many sequences that do not code for RNAs or have any structural or regulatory function. Much of this nonfunctional DNA is composed of repeated sequences. In humans, only about 1.5 percent of total DNA (the exons) actually encodes proteins or functional RNAs.
- Variation in the amount of nonfunctional DNA in the genomes of various species is largely responsible for the lack of a consistent relationship between the amount of DNA in the haploid chromosomes of an animal or plant and its phylogenetic complexity.
- Eukaryotic genomic DNA consists of three major classes of sequences: genes encoding proteins and functional RNAs, including gene families and tandemly repeated genes; repetitious DNA; and spacer DNA (see Table 10-1).
- About half the protein-coding genes in vertebrate genomic DNA are solitary genes, each occurring only once in the haploid genome. The remainder are duplicated genes, which arose by duplication of an ancestral gene and subsequent independent mutations (see Figure 10-4).
- Duplicated genes encode closely related proteins and generally appear as a cluster in a particular region of DNA. The proteins encoded by a gene family have homologous but nonidentical amino acid sequences and exhibit similar but slightly different properties.
- In invertebrates and vertebrates, rRNAs are encoded by multiple copies of genes located in tandem arrays in genomic DNA. Multiple copies of tRNA and histone genes also occur, often in clusters, but not generally in tandem arrays.
- Simple-sequence DNA, which consists largely of quite short sequences repeated in long tandem arrays, is preferentially located in centromeres, telomeres, and specific locations within the arms of particular chromosomes.
- The length of a particular simple-sequence tandem array is quite variable between individuals in a species, probably because of unequal crossing over during meiosis (see Figure 10-6). Differences in the lengths of some simple-sequence tandem arrays form the basis for DNA fingerprinting.

10.3 Mobile DNA

The second type of repetitious DNA in eukaryotic genomes, termed *interspersed repeats* (also known as *moderately repeated DNA*, or *intermediate-repeat DNA*) is composed of a very large number of copies of relatively few sequence families (see Table 10-1). These sequences, which are interspersed throughout mammalian genomes, make up ≈25–50 percent of mammalian DNA (≈45 percent of human DNA).

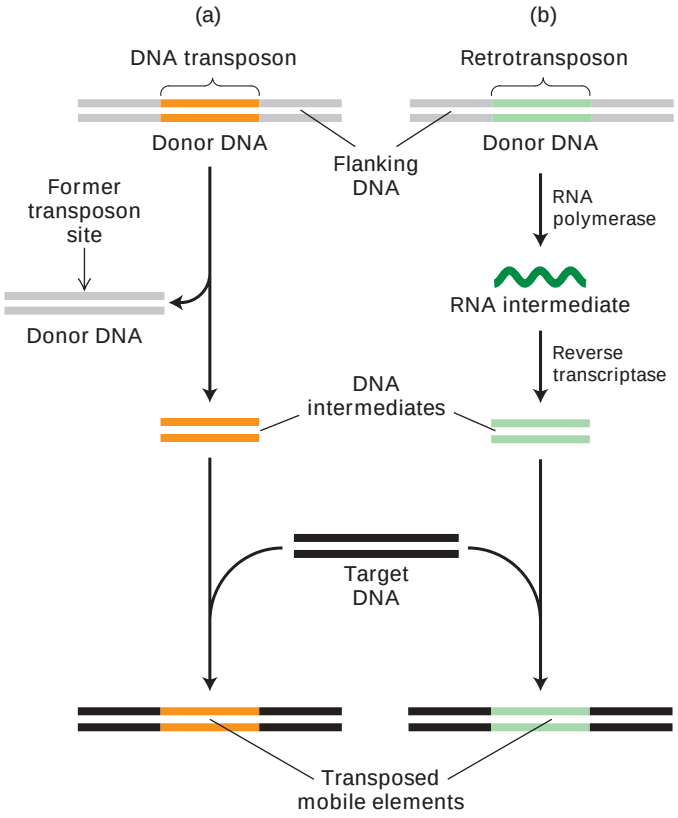
Because moderately repeated DNA sequences have the unique ability to "move" in the genome, they are called **mobile DNA elements** (or *transposable elements*). Although mobile DNA elements, ranging from hundreds to a few thousand base pairs in length, originally were discovered in eukaryotes, they also are found in prokaryotes. The process by which these sequences are copied and inserted into a new site in the genome is called **transposition.** Mobile DNA elements (or simply *mobile elements*) are essentially molecular symbiots that in most cases appear to have no specific function in the biology of their host organisms, but exist only to maintain themselves. For this reason, Francis Crick referred to them as "selfish DNA."

When transposition of eukaryotic mobile elements occurs in germ cells, the transposed sequences at their new sites can be passed on to succeeding generations. In this way, mobile elements have multiplied and slowly accumulated in eukaryotic genomes over evolutionary time. Since mobile elements are eliminated from eukaryotic genomes very slowly, they now constitute a significant portion of the genomes of many eukaryotes. Transposition also may occur within a somatic cell; in this case the transposed sequence is transmitted only to the daughter cells derived from that cell. In rare cases, this may lead to a somatic-cell mutation with detrimental phenotypic effects, for example, the inactivation of a tumor-suppressor gene (Chapter 23).

Movement of Mobile Elements Involves a DNA or an RNA Intermediate

Barbara McClintock discovered the first mobile elements while doing classical genetic experiments in maize (corn) during the 1940s. She characterized genetic entities that could move into and back out of genes, changing the phenotype of corn kernels. Her theories were very controversial until similar mobile elements were discovered in bacteria, where they were characterized as specific DNA sequences, and the molecular basis of their transposition was deciphered.

As research on mobile elements progressed, they were found to fall into two categories: (1) those that transpose directly as DNA and (2) those that transpose via an RNA intermediate transcribed from the mobile element by an RNA polymerase and then converted back into double-stranded DNA by a **reverse transcriptase** (Figure 10-8). Mobile elements that transpose through a DNA intermediate are generally referred to as **DNA transposons.** Mobile elements that transpose to new sites in the genome via an RNA intermedi-



▲ FIGURE 10-8 **Classification of mobile elements into two major classes.** (a) Eukaryotic DNA transposons (orange) move via a DNA intermediate, which is excised from the donor site. (b) Retrotransposons (green) are first transcribed into an RNA molecule, which then is reverse-transcribed into double-stranded DNA. In both cases, the double-stranded DNA intermediate is integrated into the target-site DNA to complete movement. Thus DNA transposons move by a cut-and-paste mechanism, whereas retrotransposons move by a copy-and-paste mechanism.

ate are called **retrotransposons** because their movement is analogous to the infectious process of retroviruses. Indeed, retroviruses can be thought of as retrotransposons that evolved genes encoding viral coats, thus allowing them to transpose between cells. Retrotransposons can be further classified on the basis of their specific mechanism of transposition. We describe the structure and movement of the major types of mobile elements and then consider their likely role in evolution.

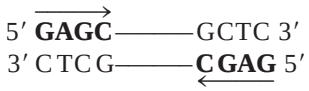
Mobile Elements That Move as DNA Are Present in Prokaryotes and Eukaryotes

Most mobile elements in bacteria transpose directly as DNA. In contrast, most mobile elements in eukaryotes are retrotransposons, but eukaryotic DNA transposons also occur. Indeed, the original mobile elements discovered by Barbara McClintock are DNA transposons.

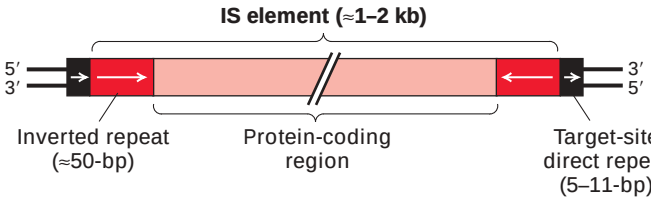
Bacterial Insertion Sequences The first molecular understanding of mobile elements came from the study of certain *E. coli* mutations caused by the spontaneous insertion of a DNA sequence, ≈1–2 kb long, into the middle of a gene. These inserted stretches of DNA are called *insertion sequences*, or *IS elements*. So far, more than 20 different IS elements have been found in *E. coli* and other bacteria.

Transposition of an IS element is a very rare event, occurring in only one in 10^5–10^7 cells per generation, depending on the IS element. Many transpositions inactivate essential genes, killing the host cell and the IS elements it carries. Therefore, higher rates of transposition would probably result in too great a mutation rate for the host cell to survive. However, since IS elements transpose more or less randomly, some transposed sequences enter nonessential regions of the genome (e.g., regions between genes), allowing the cell to survive. At a very low rate of transposition, most host cells survive and therefore propagate the symbiotic IS element. IS elements also can insert into plasmids or lysogenic viruses, and thus be transferred to other cells. When this happens, IS elements can transpose into the chromosomes of virgin cells.

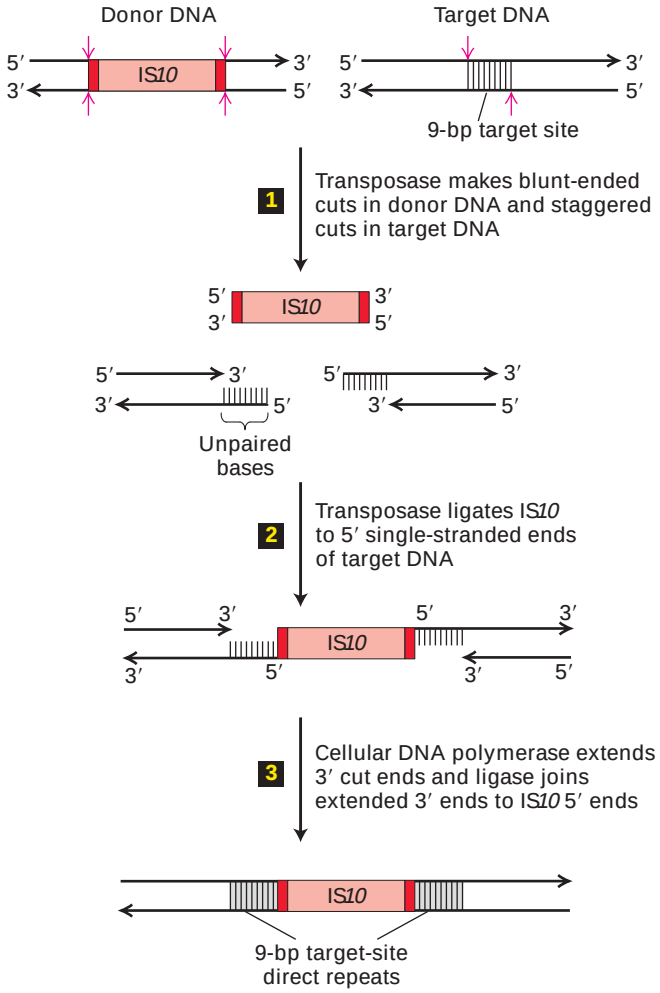
The general structure of IS elements is diagrammed in Figure 10-9. An *inverted repeat*, usually containing ≈50 base pairs, invariably is present at each end of an insertion sequence. In an inverted repeat the 5′ → 3′sequence on one strand is repeated on the other strand, as:



Between the inverted repeats is a region that encodes *transposase*, an enzyme required for transposition of the IS ele-



▲ FIGURE 10-9 **General structure of bacterial IS elements.** The relatively large central region of an IS element, which encodes one or two enzymes required for transposition, is flanked by an inverted repeat at each end. The sequences of the inverted repeats are nearly identical, but they are oriented in opposite directions. The sequence is characteristic of a particular IS element. The 5′ and 3′ short *direct* (as opposed to *inverted*) repeats are not transposed with the insertion element; rather, they are insertion-site sequences that become duplicated, with one copy at each end, during insertion of a mobile element. The length of the direct repeats is constant for a given IS element, but their sequence depends on the site of insertion and therefore varies with each transposition of the IS element. Arrows indicate sequence orientation. The regions in this diagram are not to scale; the coding region makes up most of the length of an IS element.



▲ FIGURE 10-10 **Model for transposition of bacterial insertion sequences.** Step 1: Transposase, which is encoded by the IS element (IS*10* in this example), cleaves both strands of the donor DNA next to the inverted repeats (dark red), excising the IS*10* element. At a largely random target site, transposase makes staggered cuts in the target DNA. In the case of IS*10,* the two cuts are 9 bp apart. Step 2: Ligation of the 3′ ends of the excised IS element to the staggered sites in the target DNA also is catalyzed by transposase. Step 3: The 9-bp gaps of single-stranded DNA left in the resulting intermediate are filled in by a cellular DNA polymerase; finally cellular DNA ligase forms the 3′→5′ phosphodiester bonds between the 3′ ends of the extended target DNA strands and the 5′ ends of the IS*10* strands. This process results in duplication of the target-site sequence on each side of the inserted IS element. Note that the length of the target site and IS*10* are not to scale. [See H. W. Benjamin and N. Kleckner, 1989, *Cell* **59**:373, and 1992, *Proc. Nat'l. Acad. Sci. USA* **89**:4648.]

ment to a new site. The transposase is expressed at a very low rate, accounting for the very low frequency of transposition. An important hallmark of IS elements is the presence of a short *direct-repeat sequence,* containing 5–11 base pairs, immediately adjacent to both ends of the inserted element. The *length* of the direct repeat is characteristic of each type of IS element, but its *sequence* depends on the target site where a particular copy of the IS element is inserted. When the sequence of a mutated gene containing an IS element is compared with the sequence of the wild-type gene before insertion, only one copy of the short direct-repeat sequence is found in the wild-type gene. Duplication of this target-site sequence to create the second direct repeat adjacent to an IS element occurs during the insertion process.

As depicted in Figure 10-10, transposition of an IS element is similar to a "cut-and-paste" operation in a word-processing program. Transposase performs three functions in this process: it (1) precisely excises the IS element in the donor DNA, (2) makes staggered cuts in a short sequence in the target DNA, and (3) ligates the 3′ termini of the IS element to the 5′ ends of the cut donor DNA. Finally, a host-cell DNA polymerase fills in the single-stranded gaps, generating the short direct repeats that flank IS elements, and DNA ligase joins the free ends.

Eukaryotic DNA Transposons

McClintock's original discovery of mobile elements came from observation of certain spontaneous mutations in maize that affect production of any of the several enzymes required to make anthocyanin, a purple pigment in maize kernels. Mutant kernels are white, and wild-type kernels are purple. One class of these mutations is revertible at high frequency, whereas a second class of mutations does not revert unless they occur in the presence of the first class of mutations. McClintock called the agent responsible for the first class of mutations the *activator (Ac) element* and those responsible for the second class *dissociation (Ds) elements* because they also tended to be associated with chromosome breaks.

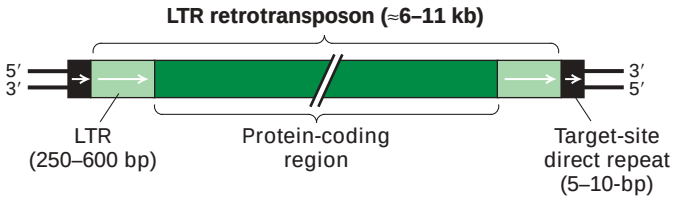
Many years after McClintock's pioneering discoveries, cloning and sequencing revealed that Ac elements are equivalent to bacterial IS elements. Like IS elements, they contain inverted terminal repeat sequences that flank the coding region for a transposase, which recognizes the terminal repeats and catalyzes transposition to a new site in DNA. Ds elements are deleted forms of the Ac element in which a portion of the sequence encoding transposase is missing. Because it does not encode a functional transposase, a Ds element cannot move by itself. However, in plants that carry the Ac element, and thus express a functional transposase, Ds elements can move.

Since McClintock's early work on mobile elements in corn, transposons have been identified in other eukaryotes. For instance, approximately half of all the spontaneous mutations observed in *Drosophila* are due to the insertion of mobile elements. Although most of the mobile elements in *Drosophila* function as retrotransposons, at least one—the **P element**—functions as a DNA transposon, moving by a cut-and-paste mechanism similar to that used by bacterial insertion sequences. Current methods for constructing transgenic *Drosophila* depend on engineered, high-level expression of the P-element transposase and use of the P-element inverted terminal repeats as targets for transposition.

DNA transposition by the cut-and-paste mechanism can result in an increase in the copy number of a transposon when it occurs during S phase, the period of the cell cycle when DNA synthesis occurs. This happens when the donor DNA is from one of the two daughter DNA molecules in a region of a chromosome that has replicated and the target DNA is in the region that has not yet replicated. When DNA replication is complete at the end of the S phase, the target DNA in its new location is also replicated. This results in a net increase by one in the total number of these transposons in the cell. When this occurs during the S phase preceding meiosis, two of the four germ cells produced have the extra copy. Repetition of this process over evolutionary time has resulted in the accumulation of large numbers of DNA transposons in the genomes of some organisms. Human DNA contains about 300,000 copies of full-length and deleted DNA tranposons, amounting to ≈3 percent of human DNA.

Some Retrotransposons Contain LTRs and Behave Like Intracellular Retroviruses

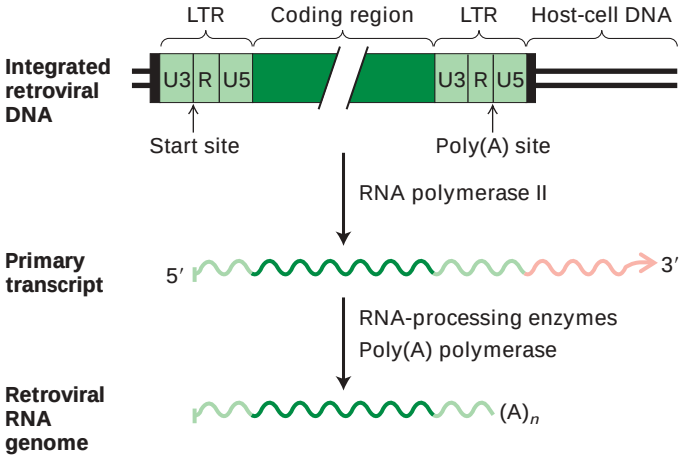
The genomes of all eukaryotes studied from yeast to humans contain retrotransposons, mobile DNA elements that transpose through an RNA intermediate utilizing a reverse transcriptase (see Figure 10-8b). These mobile elements are divided into two major categories, those containing and those lacking **long terminal repeats (LTRs).** LTR retrotransposons, which we discuss in this section, are common in yeast (e.g., Ty elements) and in *Drosophila* (e.g., *copia* elements). Although less abundant in mammals than non-LTR retrotransposons, LTR retrotransposons nonetheless constitute ≈8 percent of human genomic DNA. Because they exhibit some similarities with retroviruses, these mobile elements sometimes are called *viral retrotransposons*. In mammals, retrotransposons lacking LTRs are the most common type of mobile element; these are described in the next section.



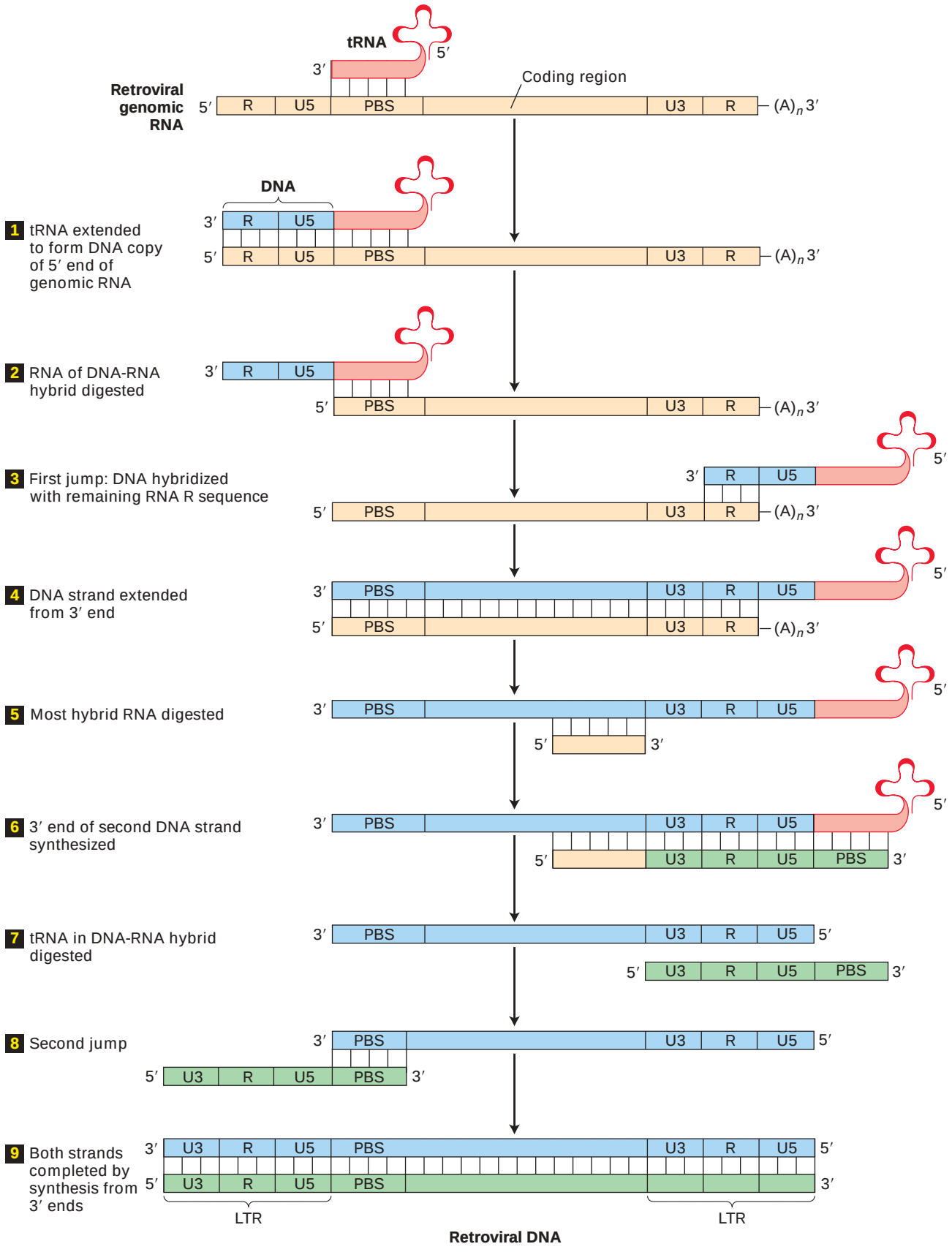
▲ FIGURE 10-11 **General structure of eukaryotic LTR retrotransposons.** The central protein-coding region is flanked by two long terminal repeats (LTRs), which are element-specific direct repeats. Like other mobile elements, integrated retrotransposons have short target-site direct repeats at each end. Note that the different regions are not drawn to scale. The protein-coding region constitutes 80 percent or more of a retrotransposon and encodes reverse transcriptase, integrase, and other retroviral proteins.

The general structure of LTR retrotransposons found in eukaryotes is depicted in Figure 10-11. In addition to short 5′ and 3′ direct repeats typical of all mobile elements, these retrotransposons are marked by the presence of LTRs flanking the central protein-coding region. These *l*ong direct *t*erminal *r*epeats, containing ≈250–600 base pairs, are characteristic of integrated retroviral DNA and are critical to the life cycle of retroviruses. In addition to sharing LTRs with retroviruses, LTR retrotransposons encode all the proteins of the most common type of retroviruses, except for the envelope proteins. Lacking these envelope proteins, LTR retrotransposons cannot bud from their host cell and infect other cells; however, they can transpose to new sites in the DNA of their host cell.

A key step in the retroviral life cycle is formation of retroviral genomic RNA from integrated retroviral DNA (see Figure 4-43). This process serves as a model for generation of the RNA intermediate during transposition of LTR retrotransposons. As depicted in Figure 10-12, the leftward retroviral LTR functions as a promoter that directs host-cell RNA polymerase II to initiate transcription at the 5′ nucleotide of the R sequence. After the entire downstream retroviral DNA has been transcribed, the RNA sequence corresponding to the rightward LTR directs host-cell RNA-processing enzymes to cleave the primary transcript and add a poly(A) tail at the 3′ end of the R sequence.

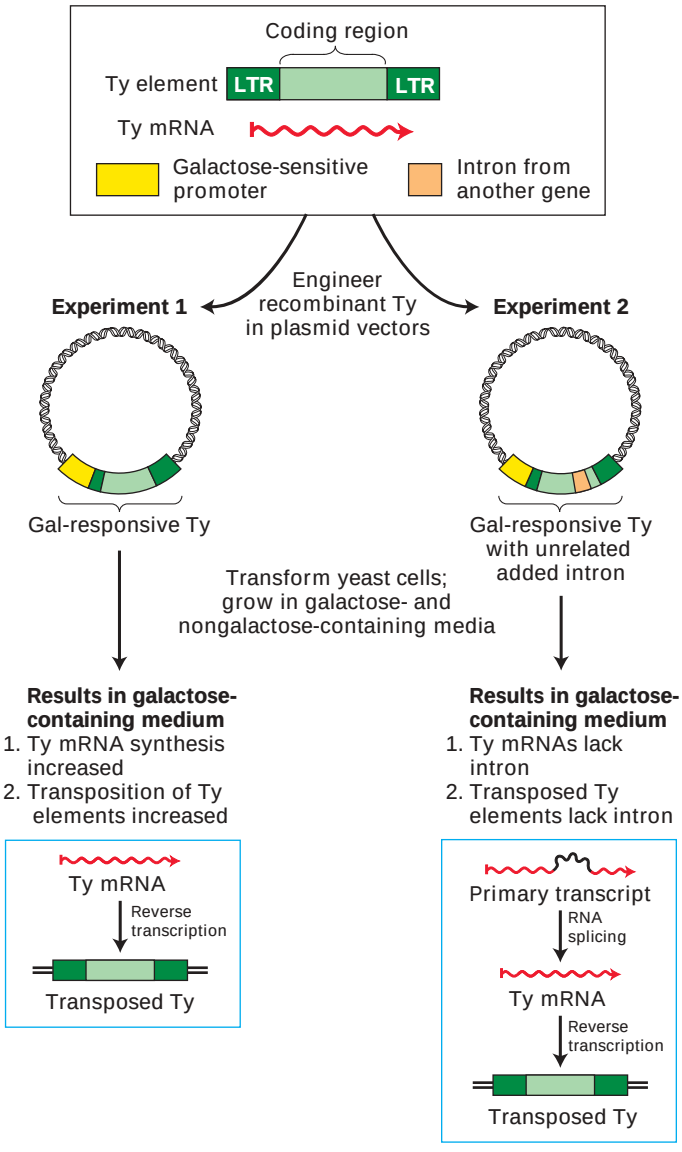


▲ FIGURE 10-12 **Generation of retroviral genomic RNA from integrated retroviral DNA.** The left LTR directs cellular RNA polymerase II to initiate transcription at the first nucleotide of the left R region. The resulting primary transcript extends beyond the right LTR. The right LTR, now present in the RNA primary transcript, directs cellular enzymes to cleave the primary transcript at the last nucleotide of the right R region and to add a poly(A) tail, yielding a retroviral RNA genome with the structure shown at the top of Figure 10-13. A similar mechanism is thought to generate the RNA intermediate during transposition of retrotransposons. The short direct-repeat sequences (black) of target-site DNA are generated during integration of the retroviral DNA into the host-cell genome.



MEDIA CONNECTIONS
Focus Animation: Retroviral Reverse Transcription

◀ FIGURE 10-13 **Model for reverse transcription of retroviral genomic RNA into DNA.** In this model, a complicated series of nine events generates a double-stranded DNA copy of the single-stranded RNA genome of a retrovirus (*top*). The genomic RNA is packaged in the virion with a retrovirus-specific cellular tRNA hybridized to a complementary sequence near its 5′ end called the *primer-binding site* (PBS). The retroviral RNA has a short direct-repeat terminal sequence (R) at each end. The overall reaction is carried out by reverse transcriptase, which catalyzes polymerization of deoxyribonucleotides. RNaseH digests the RNA strand in a DNA-RNA hybrid. The entire process yields a double-stranded DNA molecule that is longer than the template RNA and has a long terminal repeat (LTR) at each end. The different regions are not shown to scale. The PBS and R regions are actually much shorter than the U5 and U3 regions, and the central coding region is very much longer than the other regions. [See E. Gilboa et al., 1979, *Cell* **18**:93.]



The resulting retroviral RNA genome, which lacks a complete LTR, is packaged into a virion that buds from the host cell.

After a retrovirus infects a cell, reverse transcription of its RNA genome by the retrovirus-encoded reverse transcriptase yields a double-stranded DNA containing complete LTRs (Figure 10-13). Integrase, another enzyme encoded by retroviruses that is closely related to the transposase of some DNA transposons, uses a similar mechanism to insert the double-stranded retroviral DNA into the host-cell genome. In this process, short direct repeats of the target-site sequence are generated at either end of the inserted viral DNA sequence.

As noted above, LTR retrotransposons encode reverse transcriptase and integrase. By analogy with retroviruses, these mobile elements are thought to move by a "copy-and-paste" mechanism whereby reverse transcriptase converts an RNA copy of a donor-site element into DNA, which is inserted into a target site by integrase. The experiments depicted in Figure 10-14 provided strong evidence for the role of an RNA intermediate in transposition of Ty elements.

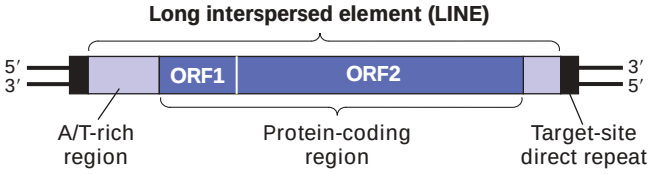
Sequencing of the human genome has revealed that the most common LTR retrotransposon–related sequences in humans are derived from endogenous retroviruses (ERVs). Most of the 443,000 ERV-related DNA sequences in the human genome consist only of isolated LTRs. These are derived from full-length proviral DNA by homologous recombination between the two LTRs, resulting in deletion of the internal retroviral sequences.

◀ EXPERIMENTAL FIGURE 10-14 **Recombinant plasmids demonstrate that the yeast Ty element transposes through an RNA intermediate.** When yeast cells are transformed with a Ty-containing plasmid, the Ty element can transpose to new sites, although normally this occurs at a low rate. Using the elements diagrammed at the top, researchers engineered two different plasmid vectors containing recombinant Ty elements adjacent to a galactose-sensitive promoter. These plasmids were transformed into yeast cells, which were grown in a galactose-containing and a nongalactose medium. In experiment 1, growth of cells in galactose-containing medium resulted in many more transpositions than in nongalactose medium, indicating that transcription into an mRNA intermediate is required for Ty transposition. In experiment 2, an intron from an unrelated yeast gene was inserted into the putative protein-coding region of the recombinant galactose-responsive Ty element. The observed absence of the intron in transposed Ty elements is strong evidence that transposition involves an mRNA intermediate from which the intron was removed by RNA splicing, as depicted in the box on the right. In contrast, eukaryotic DNA transposons, like the Ac element of maize, contain introns within the transposase gene, indicating that they do not transpose via an RNA intermediate. [See J. Boeke et al., 1985, *Cell* **40**:491]

Retrotransposons That Lack LTRs Move by a Distinct Mechanism

The most abundant mobile elements in mammals are retrotransposons that lack LTRs, sometimes called *nonviral retrotransposons*. These moderately repeated DNA sequences form two classes in mammalian genomes: *long interspersed elements (LINEs)* and *short interspersed elements (SINEs)*. In humans, full-length LINEs are ≈6 kb long, and SINEs are ≈300 bp long. Repeated sequences with the characteristics of LINEs have been observed in protozoans, insects, and plants, but for unknown reasons they are particularly abundant in the genomes of mammals. SINEs also are found primarily in mammalian DNA. Large numbers of LINEs and SINEs in higher eukaryotes have accumulated over evolutionary time by repeated copying of sequences at a few positions in the genome and insertion of the copies into new positions. Although these mobile elements do not contain LTRs, the available evidence indicates that they transpose through an RNA intermediate.

LINEs Human DNA contains three major families of LINE sequences that are similar in their mechanism of transposition, but differ in their sequences: L1, L2, and L3. Only members of the L1 family transpose in the contemporary human genome. LINE sequences are present at ≈900,000 sites in the human genome, accounting for a staggering 21 percent of total human DNA. The general structure of a complete LINE is diagrammed in Figure 10-15. LINEs usually are flanked by short direct repeats, the hallmark of mobile elements, and contain two long open reading frames (ORFs). ORF1, ≈1 kb long, encodes an RNA-binding protein. ORF2, ≈4 kb long, encodes a protein that has a long region of homology with the reverse transcriptases of retroviruses and viral retrotransposons, but also exhibits DNA endonuclease activity.



▲ FIGURE 10-15 **General structure of a LINE, one of the two classes of non-LTR retrotransposons in mammalian DNA.** The length of the target-site direct repeats varies among copies of the element at different sites in the genome. Although the full-length L1 sequence is ≈6 kb long, variable amounts of the left end are absent at over 90 percent of the sites where this mobile element is found. The shorter open reading frame (ORF1), ≈1 kb in length, encodes an RNA-binding protein. The longer ORF2, ≈4 kb in length, encodes a bifunctional protein with reverse transcriptase and DNA endonuclease activity.

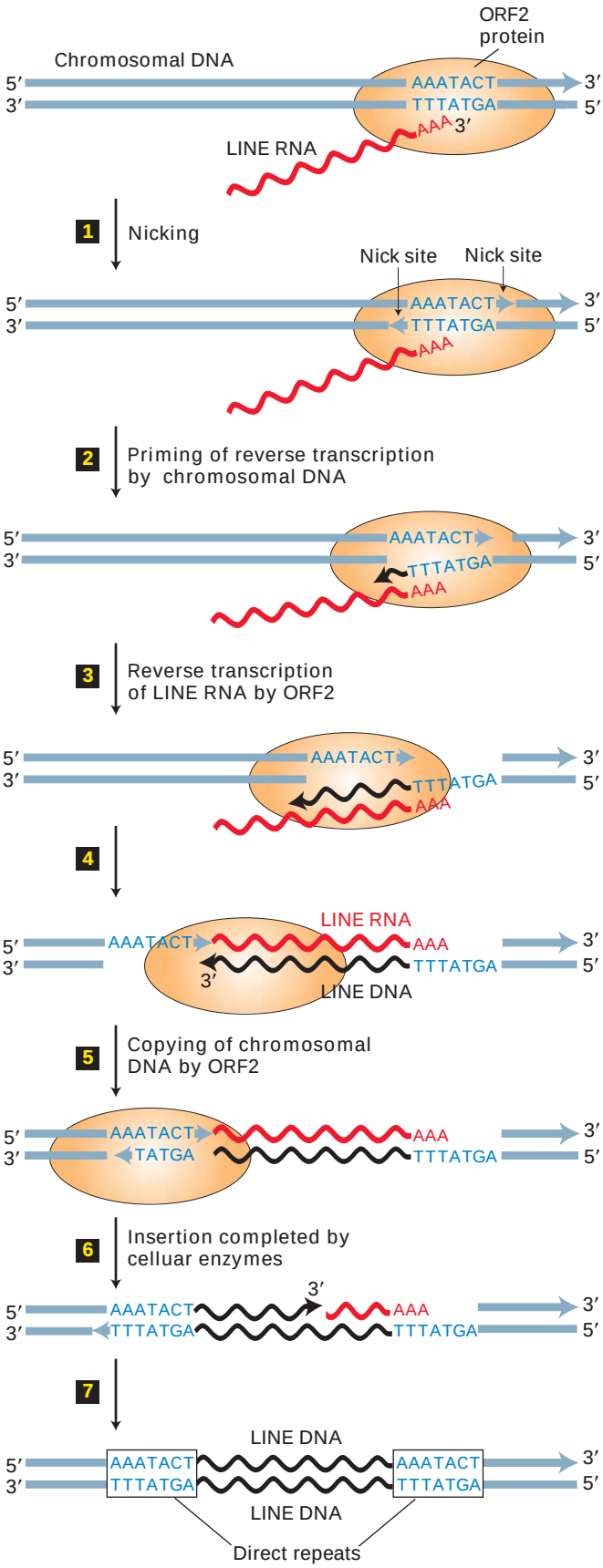
MEDICINE

Evidence for the mobility of L1 elements first came from analysis of DNA cloned from humans with certain genetic diseases. DNA from these patients was found to carry mutations resulting from insertion of an L1 element into a gene, whereas no such element occurred within this gene in either parent. About 1 in 600 mutations that cause significant disease in humans are due to L1 transpositions or SINE transpositions that are catalyzed by L1-encoded proteins. Later experiments similar to those just described with yeast Ty elements (see Figure 10-14) confirmed that L1 elements transpose through an RNA intermediate. In these experiments, an intron was introduced into a cloned mouse L1 element, and the recombinant L1 element was stably transformed into cultured hamster cells. After several cell doublings, a PCR-amplified fragment corresponding to the L1 element but lacking the inserted intron was detected in the cells. This finding strongly suggests that over time the recombinant L1 element containing the inserted intron had transposed to new sites in the hamster genome through an RNA intermediate that underwent RNA splicing to remove the intron. ▮

Since LINEs do not contain LTRs, their mechanism of transposition through an RNA intermediate differs from that of LTR retrotransposons. ORF1 and ORF2 proteins are translated from a LINE RNA. In vitro studies indicate that transcription by RNA polymerase II is directed by promoter sequences at the left end of integrated LINE DNA. LINE RNA is polyadenylated by the same post-transcriptional mechanism that polyadenylates other mRNAs. The LINE RNA then is transported into the cytoplasm, where it is translated into ORF1 and ORF2 proteins. Multiple copies of ORF1 protein then bind to the LINE RNA, and ORF2 protein binds to the poly(A) tail.

The LINE RNA is then transported back into the nucleus as a complex with ORF1 and ORF2. ORF2 then makes staggered nicks in chromosomal DNA on either side of any A/T-rich sequence in the genome (Figure 10-16, step 1). Reverse transcription of LINE RNA by ORF2 is primed by the single-stranded T-rich sequence generated by the nick in the bottom strand, which hybridizes to the LINE poly(A) tail (step 2). ORF2 then reverse-transcribes the LINE RNA (step 3) and then continues this new DNA strand, switching to the single-stranded region of the upper chromosomal strand as a template (steps 4 and 5). Cellular enzymes then hydrolyze the RNA and extend the 3′ end of the chromosomal DNA top strand, replacing the LINE RNA strand with DNA (step 6).

▶ FIGURE 10-16 **Proposed mechanism of LINE reverse transcription and integration.** Only ORF2 protein is represented. Newly synthesized LINE DNA is shown in black. See the text for explanation. [Adapted from D. D. Luan et al., 1993, *Cell* **72**:595.]



Finally, 5′ and 3′ ends of DNA strands are ligated, completing the insertion (step 7). These last steps (6 and 7) probably are catalyzed by the same cellular enzymes that remove RNA primers and ligate Okazaki fragments during DNA replication (see Figure 4-33). The complete process results in insertion of a copy of the original LINE retrotransposon into a new site in chromosomal DNA. A short direct repeat is generated at the insertion site because of the initial staggered cleavage of the two chromosomal DNA strands (step 1).

The vast majority of LINEs in the human genome are truncated at their 5′ end, suggesting that reverse transcription terminated before completion and the resulting fragments extending variable distances from the poly(A) tail were inserted. Because of this shortening, the average size of LINE elements is only about 900 base pairs, whereas the full-length sequence is ≈6 kb long. In addition, nearly all the full-length elements contain stop codons and frameshift mutations in ORF1 and ORF2; these mutations probably have accumulated in most LINE sequences over evolutionary time. As a result of truncation and mutation, only ≈0.01 percent of the LINE sequences in the human genome are full-length with intact open reading frames for ORF1 and ORF2, ≈60–100 in total.

SINEs The second most abundant class of mobile elements in the human genome, SINEs constitute ≈13 percent of total human DNA. Varying in length from about 100 to 400 base pairs, these retrotransposons do not encode protein, but most contain a 3′ A/T-rich sequence similar to that in LINEs. SINEs are transcribed by RNA polymerase III, the same nuclear RNA polymerase that transcribes genes encoding tRNAs, 5S rRNAs, and other small stable RNAs (Chapter 11). Most likely, the ORF1 and ORF2 proteins expressed from full-length LINEs mediate transposition of SINEs by the retrotransposition mechanism depicted in Figure 10-16.

SINEs occur at about 1.6 million sites in the human genome. Of these, ≈1.1 million are *Alu elements*, so named because most of them contain a single recognition site for the restriction enzyme *Alu*I. *Alu* elements exhibit considerable sequence homology with and may have evolved from 7SL RNA, a component of the signal-recognition particle. This abundant cytosolic ribonucleoprotein particle aids in targeting certain polypeptides, as they are being synthesized, to the membranes of the endoplasmic reticulum (Chapter 16). *Alu* elements are scattered throughout the human genome at sites where their insertion has not disrupted gene expression: between genes, within introns, and in the 3′ untranslated regions of some mRNAs. For instance, nine *Alu* elements are located within the human β-globin gene cluster (see Figure 10-3b). The overall frequency of L1 and SINE retrotranspositions in humans is estimated to be about one new retrotransposition in very eight individuals, with ≈40 percent being L1 and 60 percent SINEs, of which ≈90 percent are *Alu* elements.

Similar to other mobile elements, most SINEs have accumulated mutations from the time of their insertion in the germ line of an ancient ancestor of modern humans. Like LINEs, many SINEs also are truncated at their 5′ end. Table 10-1 summarizes the major types of interspersed repeats derived from mobile elements in the human genome.

In addition to the mobile elements listed in Table 10-1, DNA copies of a wide variety of mRNAs appear to have integrated into chromosomal DNA. Since these sequences lack introns and do not have flanking sequences similar to those of the functional gene copies, they clearly are not simply duplicated genes that have drifted into nonfunctionality and become pseudogenes, as discussed earlier (Figure 10-3a). Instead, these DNA segments appear to be retrotransposed copies of spliced and polyadenylated (processed) mRNA. Compared with normal genes encoding mRNAs, these inserted segments generally contain multiple mutations, which are thought to have accumulated since their mRNAs were first reverse-transcribed and randomly integrated into the genome of a germ cell in an ancient ancestor. These nonfunctional genomic copies of mRNAs are referred to as *processed pseudogenes*. Most processed pseudogenes are flanked by short direct repeats, supporting the hypothesis that they were generated by rare retrotransposition events involving cellular mRNAs.

Other moderately repetitive sequences representing partial or mutant copies of genes encoding small nuclear RNAs (snRNAs) and tRNAs are found in mammalian genomes. Like processed pseudogenes derived from mRNAs, these nonfunctional copies of small RNA genes are flanked by short direct repeats and most likely result from rare retrotransposition events that have accumulated through the course of evolution. Enzymes expressed from a LINE are thought to have carried out all these retrotransposition events involving mRNAs, snRNAs, and tRNAs.

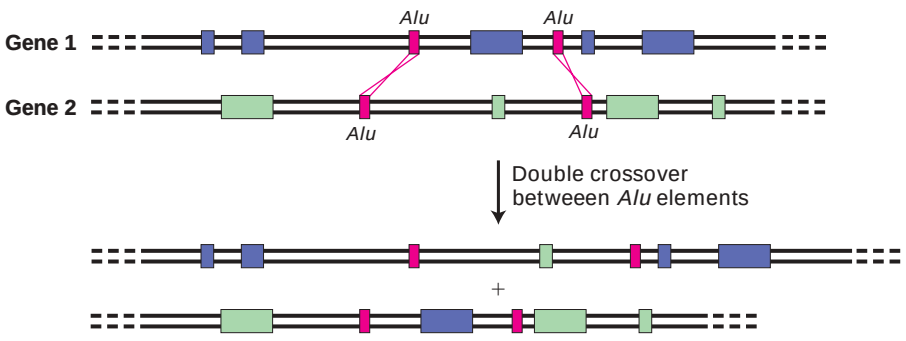
Mobile DNA Elements Probably Had a Significant Influence on Evolution

Although mobile DNA elements appear to have no direct function other than to maintain their own existence, their presence probably had a profound impact on the evolution of modern-day organisms. As mentioned earlier, about half the spontaneous mutations in *Drosophila* result from insertion of a mobile DNA element into or near a transcription unit. In mammals, however, mobile elements cause a much smaller proportion of spontaneous mutations: ≈10 percent in mice and only 0.1–0.2 percent in humans. Still, mobile elements have been found in mutant alleles associated with several human genetic diseases.

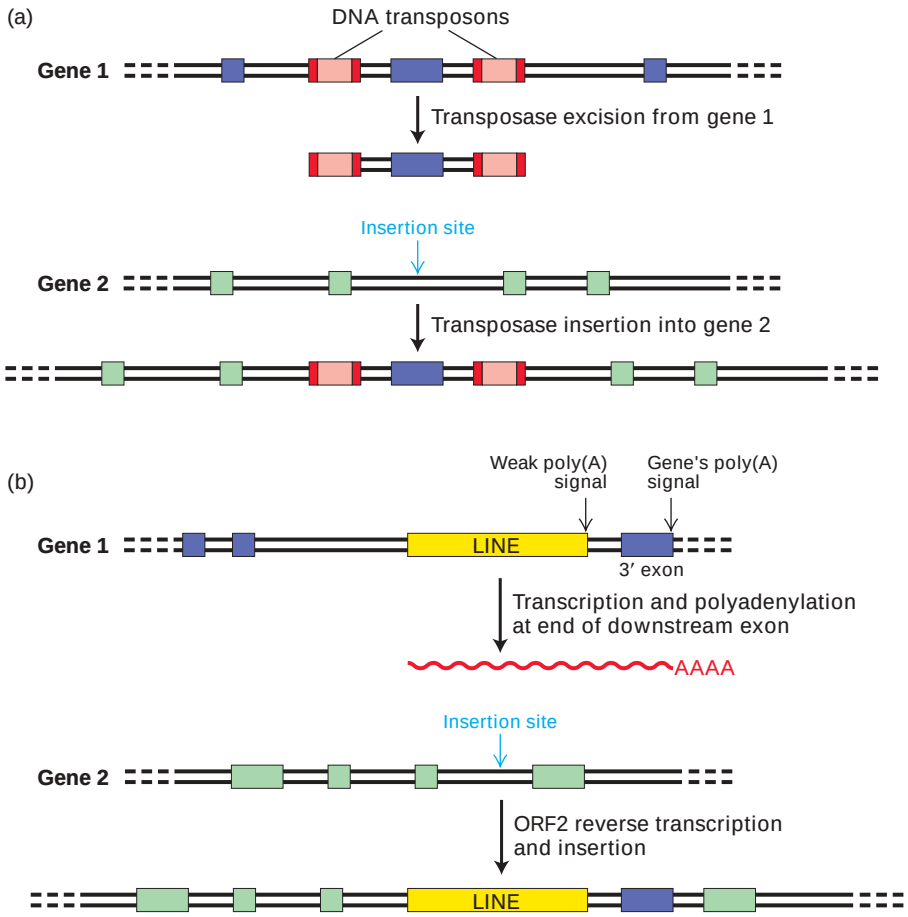
In lineages leading to higher eukaryotes, homologous recombination between mobile DNA elements dispersed throughout ancestral genomes may have generated gene duplications and other DNA rearrangements during evolution (see Figure 10-4). For instance, cloning and sequencing of the β-globin gene cluster from various primate species has provided strong evidence that the human G_γ and A_γ genes arose from an unequal homologous crossover between two L1 sequences flanking an ancestral globin gene. Subsequent divergence of such duplicated genes could lead to acquisition of distinct, beneficial functions associated with each member of a gene family. Unequal crossing over between mobile elements located within introns of a particular gene could lead to the duplication of exons within that gene. This process most likely influenced the evolution of genes that contain multiple copies of similar exons encoding similar protein domains, such as the fibronectin gene (see Figure 4-15).

Some evidence suggests that during the evolution of higher eukaryotes, recombination between interspersed repeats in introns of *two separate* genes also occurred, generating new genes made from novel combinations of preexisting exons (Figure 10-17). This evolutionary process, termed **exon shuffling,** may have occurred during evolution of the genes encoding tissue plasminogen activator, the Neu receptor, and epidermal growth factor, which all contain an EGF domain (see Figure 3-8). In this case, exon shuffling presumably resulted in insertion of an EGF domain–encoding exon into an intron of the ancestral form of each of these genes.

Both DNA transposons and LINE retrotransposons have been shown to occasionally carry unrelated flanking sequences when they insert into new sites by the mechanisms diagrammed in Figure 10-18. These mechanisms likely also contributed to exon shuffling during the evolution of contemporary genes.



▶ FIGURE 10-17 **Exon shuffling via recombination between homologous interspersed repeats.** Recombination between interspersed repeats in the introns of separate genes produces transcription units with a new combination of exons. In the example shown here, a double crossover between two sets of *Alu* repeats results in an exchange of exons between the two genes.



◀ FIGURE 10-18 **Exon shuffling by transposition.** (a) Transposition of an exon flanked by homologous DNA transposons into an intron on a second gene. As we saw in Figure 10-10, step 1, transposase can recognize and cleave the DNA at the ends of the transposon inverted repeats. In gene 1, if the transposase cleaves at the left end of the transposon on the left and at the right end of the transposon on the right, it can transpose all the intervening DNA, including the exon from gene 1, to a new site in an intron of gene 2. The net result is an insertion of the exon from gene 1 into gene 2. (b) Integration of an exon into another gene via LINE transposition. Some LINEs have weak poly(A) signals. If such a LINE is in the 3′-most intron of gene 1, during transposition its transcription may continue beyond its own poly(A) signals and extend into the 3′ exon, transcribing the cleavage and polyadenylation signals of gene 1 itself. This RNA can then be reverse-transcribed and integrated by the LINE ORF2 protein (Figure 10-16) into an intron on gene 2, introducing a new 3′ exon (from gene 1) into gene 2.

In addition to causing changes in coding sequences in the genome, recombination between mobile elements and transposition of DNA adjacent to DNA transposons and retrotransposons likely played a significant role in the evolution of regulatory sequences that control gene expression. As noted earlier, eukaryotic genes have transcription-control regions, called enhancers, that can operate over distances of tens of thousands of base pairs. Transcription of many genes is controlled through the combined effects of several enhancer elements. Insertion of mobile elements near such transcription-control regions probably contributed to the evolution of new combinations of enhancer sequences. These in turn control which specific genes are expressed in particular cell types and the amount of the encoded protein produced in modern organisms, as we discuss in the next chapter.

These considerations suggest that the early view of mobile DNA elements as completely selfish molecular parasites misses the mark. Rather, they have likely contributed profoundly to the evolution of higher organisms by promoting (1) the generation of gene families via gene duplication, (2) the creation of new genes via shuffling of preexisting exons, and (3) formation of more complex regulatory regions that provide multifaceted control of gene expression.

KEY CONCEPTS OF SECTION 10.3

Mobile DNA

- Mobile DNA elements are moderately repeated DNA sequences interspersed at multiple sites throughout the genomes of higher eukaryotes. They are present less frequently in prokaryotic genomes.
- Mobile DNA elements that transpose to new sites directly as DNA are called *DNA transposons;* those that first are transcribed into an RNA copy of the element, which then is reverse-transcribed into DNA, are called *retrotransposons* (see Figure 10-8).
- A common feature of all mobile elements is the presence of short direct repeats flanking the sequence.
- Enzymes encoded by mobile elements themselves catalyze insertion of these sequences at new sites in genomic DNA.
- Although DNA transposons, similar in structure to bacterial IS elements, occur in eukaryotes (e.g., the *Drosophila* P element), retrotransposons generally are much more abundant, especially in vertebrates.
- LTR retrotransposons are flanked by long terminal repeats (LTRs), similar to those in retroviral DNA; like

retroviruses, they encode reverse transcriptase and integrase. They move in the genome by being transcribed into RNA, which then undergoes reverse transcription and integration into the host-cell chromosome (see Figure 10-13).

- Long interspersed elements (LINEs) and short interspersed elements (SINEs) are the most abundant mobile elements in the human genome. LINEs account for ≈21 percent of human DNA; SINEs, for ≈13 percent.
- Both LINEs and SINEs lack LTRs and have an A/T-rich stretch at one end. They are thought to move by a nonviral retrotransposition mechanism mediated by LINE-encoded proteins involving priming by chromosomal DNA (see Figure 10-16).
- SINE sequences exhibit extensive homology with small cellular RNAs transcribed by RNA polymerase III. *Alu* elements, the most common SINEs in humans, are ≈300-bp sequences found scattered throughout the human genome.
- Some moderately repeated DNA sequences are derived from cellular RNAs that were reverse-transcribed and inserted into genomic DNA at some time in evolutionary history. Those derived from mRNAs, called *processed pseudogenes*, lack introns, a feature that distinguishes them from pseudogenes, which arose by sequence drift of duplicated genes.
- Mobile DNA elements most likely influenced evolution significantly by serving as recombination sites and by mobilizing adjacent DNA sequences.

10.4 Structural Organization of Eukaryotic Chromosomes

We turn now to the question of how DNA molecules are organized within eukaryotic cells. Because the total length of cellular DNA is up to a hundred thousand times a cell's length, the packing of DNA is crucial to cell architecture. During interphase, when cells are not dividing, the genetic material exists as a nucleoprotein complex called **chromatin,** which is dispersed through much of the nucleus. Further folding and compaction of chromatin during mitosis produces the visible **metaphase chromosomes,** whose morphology and staining characteristics were detailed by early cytogeneticists. In this section, we consider the properties of chromatin and its organization into chromosomes. Important features of chromosomes in their entirety are covered in the next section.

Eukaryotic Nuclear DNA Associates with Histone Proteins to Form Chromatin

When the DNA from eukaryotic nuclei is isolated in isotonic buffers (i.e., buffers with the same salt concentration found in cells, ≈0.15 M KCl), it is associated with an equal mass of protein as chromatin. The general structure of chromatin has been found to be remarkably similar in the cells of all eukaryotes, including fungi, plants, and animals.

The most abundant proteins associated with eukaryotic DNA are **histones,** a family of small, basic proteins present in all eukaryotic nuclei. The five major types of histone proteins—termed *H1*, *H2A*, *H2B*, *H3*, and *H4*—are rich in positively charged basic amino acids, which interact with the negatively charged phosphate groups in DNA.

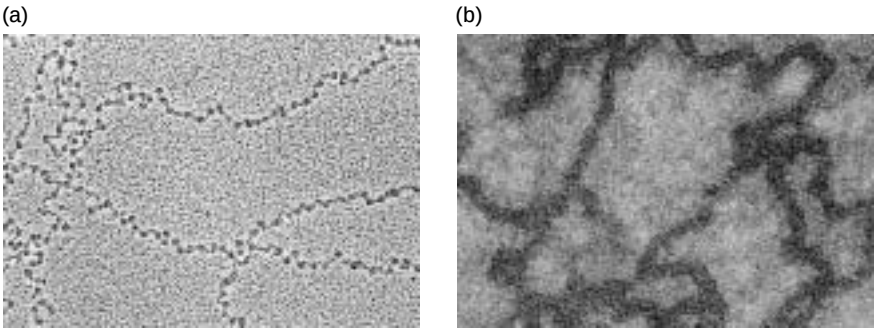
The amino acid sequences of four histones (H2A, H2B, H3, and H4) are remarkably similar among distantly related species. For example, the sequences of histone H3 from sea urchin tissue and calf thymus differ by only a single amino acid, and H3 from the garden pea and calf thymus differ only in four amino acids. Minor histone variants encoded by genes that differ from the highly conserved major types also exist, particularly in vertebrates.

The amino acid sequence of H1 varies more from organism to organism than do the sequences of the other major histones. In certain tissues, H1 is replaced by special histones. For example, in the nucleated red blood cells of birds, a histone termed *H5* is present in place of H1. The similarity in sequence among histones from all eukaryotes suggests that they fold into very similar three-dimensional conformations, which were optimized for histone function early in evolution in a common ancestor of all modern eukaryotes.

Chromatin Exists in Extended and Condensed Forms

When chromatin is extracted from nuclei and examined in the electron microscope, its appearance depends on the salt concentration to which it is exposed. At low salt concentration in the absence of divalent cations such as Mg^{+2}, isolated chromatin resembles "beads on a string" (Figure 10-19a). In this extended form, the string is composed of free DNA called "linker" DNA connecting the beadlike structures termed **nucleosomes.** Composed of DNA and histones, nucleosomes are about 10 nm in diameter and are the primary structural units of chromatin. If chromatin is isolated at physiological salt concentration (≈0.15 M KCl, 0.004 M Mg^{+2}), it assumes a more condensed fiberlike form that is 30 nm in diameter (Figure 10-19b).

Structure of Nucleosomes The DNA component of nucleosomes is much less susceptible to nuclease digestion than is the linker DNA between them. If nuclease treatment is carefully controlled, all the linker DNA can be digested, releasing individual nucleosomes with their DNA component. A nucleosome consists of a protein core with DNA wound around its surface like thread around a spool. The core is an octamer containing two copies each of histones H2A, H2B, H3, and H4. X-ray crystallography has shown that the octameric histone core is a roughly disk-shaped molecule made of interlocking histone subunits



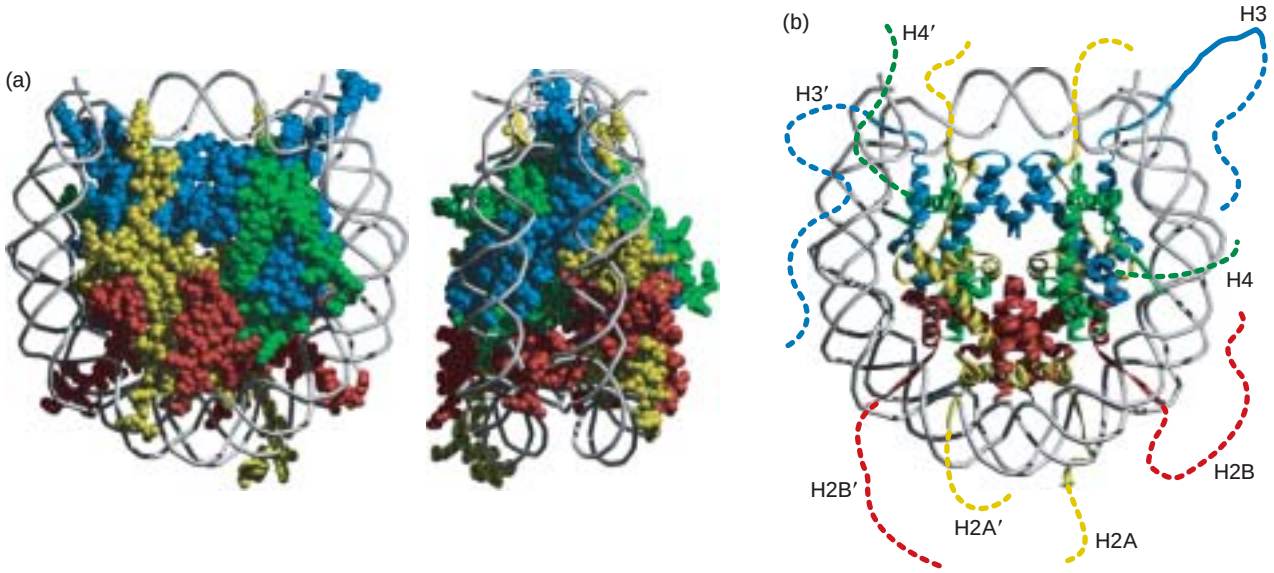
▲ EXPERIMENTAL FIGURE 10-19 **The extended and condensed forms of extracted chromatin have very different appearances in electron micrographs.** (a) Chromatin isolated in low-ionic-strength buffer has an extended "beads-on-a-string" appearance. The "beads" are nucleosomes (10-nm diameter) and the "string" is connecting (linker) DNA. (b) Chromatin isolated in buffer with a physiological ionic strength (0.15 M KCl) appears as a condensed fiber 30 nm in diameter. [Part (a) courtesy of S. McKnight and O. Miller, Jr.; part (b) courtesy of B. Hamkalo and J. B. Rattner.]

(Figure 10-20). Nucleosomes from all eukaryotes contain 147 base pairs of DNA wrapped slightly less than two turns around the protein core. The length of the linker DNA is more variable among species, ranging from about 15 to 55 base pairs.

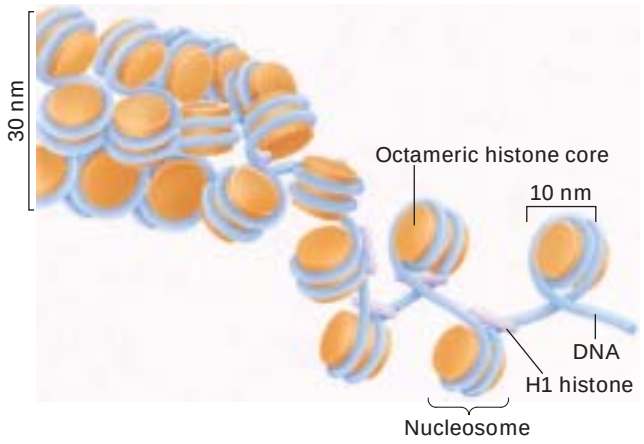
In cells, newly replicated DNA is assembled into nucleosomes shortly after the replication fork passes, but when isolated histones are added to DNA in vitro at physiological salt concentration, nucleosomes do not spontaneously form. However, nuclear proteins that bind histones and assemble them with DNA into nucleosomes in vitro have been characterized. Proteins of this type are thought to assemble histones and newly replicated DNA into nucleosomes in vivo as well.

Structure of Condensed Chromatin

When extracted from cells in isotonic buffers, most chromatin appears as fibers ≈30 nm in diameter (see Figure 10-19b). In these condensed



▲ FIGURE 10-20 **Structure of the nucleosome based on x-ray crystallography.** (a) Space-filling model shown from the front (*left*) and from the side (*right*, rotated clockwise 90°). H2A is yellow; H2B is red; H3 is blue; H4 is green. The sugar-phosphate backbone of the DNA strands is shown as white tubes. The N-terminal tails of the eight histones and the two H2A C-terminal tails, involved in condensation of the chromatin, are not visible because they are disordered in the crystal. (b) Ribbon diagram of the histones showing the lengths of the histone tails (dotted lines) not visible in the crystal structure. H2A N-terminal tails at the bottom, C-terminal tails at the top. [Part (a) after K. Luger et al., 1997, *Nature* **389**:251; part (b) K. Luger and T. J. Richmond, 1998, *Curr. Opin. Gen. Dev.* **8**:140.]



▲ FIGURE 10-21 **Solenoid model of the 30-nm condensed chromatin fiber in a side view.** The octameric histone core (see Figure 10-20) is shown as an orange disk. Each nucleosome associates with one H1 molecule, and the fiber coils into a solenoid structure with a diameter of 30 nm. [Adapted from M. Grunstein, 1992, *Sci. Am.* **267**:68.]

fibers, nucleosomes are thought to be packed into an irregular spiral or solenoid arrangement, with approximately six nucleosomes per turn (Figure 10-21). H1, the fifth major histone, is bound to the DNA on the inside of the solenoid, with one H1 molecule associated with each nucleosome. Recent electron microscopic studies suggest that the 30-nm fiber is less uniform than a perfect solenoid. Condensed chromatin may in fact be quite dynamic, with regions occasionally partially unfolding and then refolding into a solenoid structure.

The chromatin in chromosomal regions that are not being transcribed exists predominantly in the condensed, 30-nm fiber form and in higher-order folded structures whose detailed conformation is not currently understood. The regions of chromatin actively being transcribed are thought to assume the extended beads-on-a-string form.

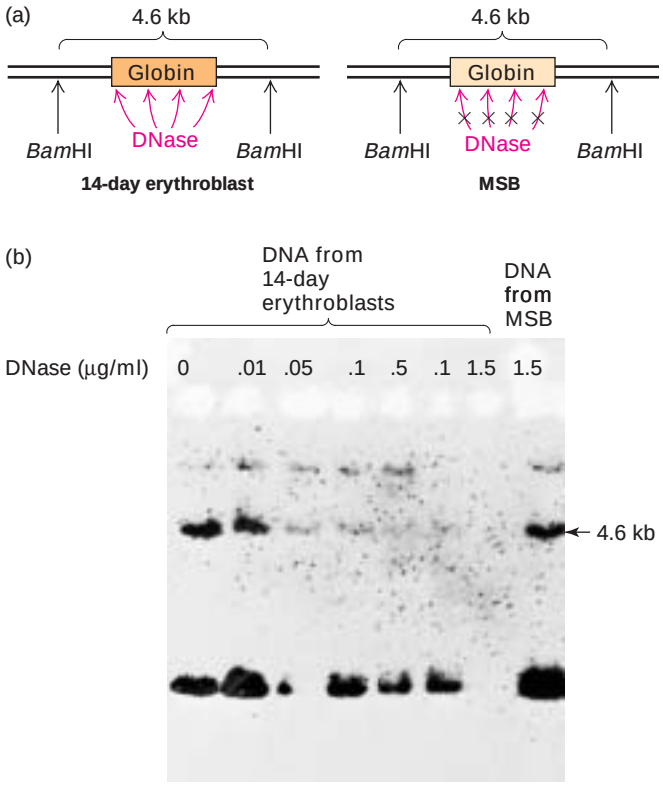
Modification of Histone Tails Controls Chromatin Condensation

Each of the histone proteins making up the nucleosome core contains a flexible amino terminus of 11–37 residues extending from the fixed structure of the nucleosome; these termini are called *histone tails*. Each H2A also contains a flexible C-terminal tail (see Figure 10-20b). The histone tails are required for chromatin to condense from the beads-on-a-string conformation into the 30-nm fiber. Several positively charged lysine side chains in the histone tails may interact with linker DNA, and the tails of one nucleosome likely interact with neighboring nucleosomes. The histone tail lysines, especially those in H3 and H4, undergo reversible acetylation and deacetylation by enzymes that act on specific lysines in the N-termini. In the acetylated form, the positive charge of the lysine ϵ-amino group is neutralized, thereby eliminating its interaction with a DNA phosphate group. Thus the greater the acetylation of histone N-termini, the less likely chromatin is to form condensed 30-nm fibers and possibly higher-order folded structures.

The histone tails can also bind to other proteins associated with chromatin that influence chromatin structure and processes such as transcription and DNA replication. The interaction of histone tails with these proteins can be regulated by a variety of covalent modifications of histone tail amino acid side chains. These include acetylation of lysine ϵ-amino groups, as mentioned earlier, as well as methylation of these groups, a process that prevents acetylation, thus maintaining their positive charge. Arginine side chains can also be methylated. Serine and threonine side chains can be phosphorylated, introducing a negative charge. Finally, a single 76-amino-acid ubiquitin molecule can be added to some lysines. Recall that addition of multiple linked ubiquitin molecules to a protein can mark it for degradation by the proteasome (Chapter 3). In this case, the addition of a single ubiquitin does not affect the stability of a histone, but influences chromatin structure. In summary, multiple types of covalent modifications of histone tails can influence chromatin structure by altering histone-DNA interactions and interactions between nucleosomes and by controlling interactions with additional proteins that participate in the regulation of transcription, as discussed in the next chapter.

The extent of histone acetylation is correlated with the relative resistance of chromatin DNA to digestion by nucleases. This phenomenon can be demonstrated by digesting isolated nuclei with DNase I. Following digestion, the DNA is completely separated from chromatin protein, digested to completion with a restriction enzyme, and analyzed by Southern blotting (see Figure 9-26). An intact gene treated with a restriction enzyme yields characteristic fragments. When a gene is exposed first to DNase, it is cleaved at random sites within the boundaries of the restriction enzyme cut sites. Consequently, any Southern blot bands normally seen with that gene will be lost. This method has been used to show that the transcriptionally inactive β-globin gene in nonerythroid cells, where it is associated with relatively unacetylated histones, is much more resistant to DNase I than is the active, transcribed β-globin gene in erythroid precursor cells, where it is associated with acetylated histones (Figure 10-22). These results indicate that the chromatin structure of nontranscribed DNA is more condensed, and therefore more protected from DNase digestion, than that of transcribed DNA. In condensed chromatin, the DNA is largely inaccessible to DNase I because of its close association with histones and other less abundant chromatin proteins. In contrast, actively transcribed DNA is much more accessible to DNase I digestion because it is present in the extended, beads-on-a-string form of chromatin.

Genetic studies in yeast indicate that specific histone acetylases are required for the full activation of transcription of a number of genes. Consequently, as discussed in

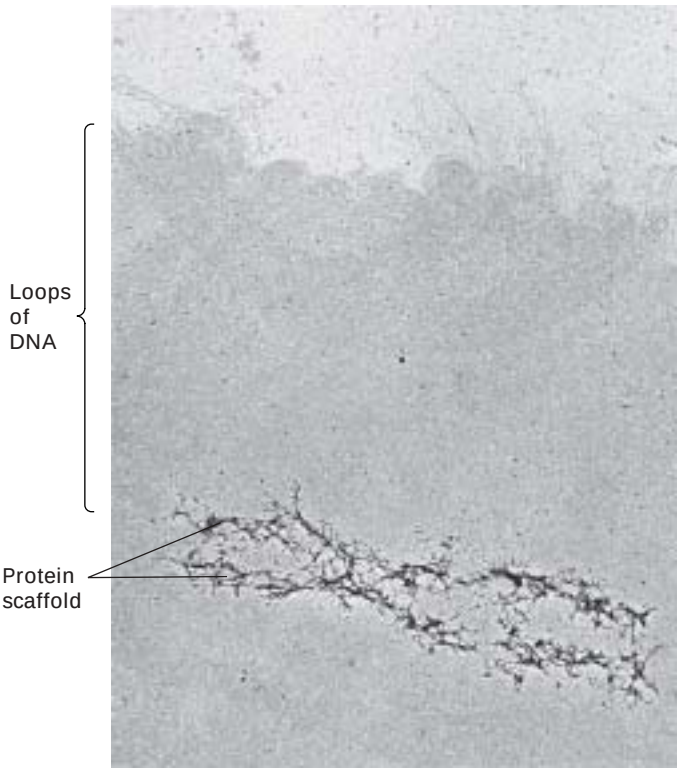


▲ EXPERIMENTAL FIGURE 10-22 **Nontranscribed genes are less susceptible to DNase I digestion than active genes.** Chick embryo erythroblasts at 14 days actively synthesize globin, whereas cultured undifferentiated MSB cells do not. (a) Nuclei from each type of cell were isolated and exposed to increasing concentrations of DNase I. The nuclear DNA was then extracted and treated with the restriction enzyme *Bam*HI, which cleaves the DNA around the globin sequence and normally releases a 4.6-kb globin fragment. (b) The DNase I- and *Bam*HI-digested DNA was subjected to Southern blot analysis with a probe of labeled cloned adult globin DNA, which hybridizes to the 4.6-kb *Bam*HI fragment. If the globin gene is susceptible to the initial DNase digestion, it would be cleaved repeatedly and would not be expected to show this fragment. As seen in the Southern blot, the transcriptionally active DNA from the 14-day globin-synthesizing cells was sensitive to DNase I digestion, indicated by the absence of the 4.6-kb band at higher nuclease concentrations. In contrast, the inactive DNA from MSB cells was resistant to digestion. These results suggest that the inactive DNA is in a more condensed form of chromatin in which the globin gene is shielded from DNase digestion. [See J. Stalder et al., 1980, *Cell* **19**:973; photograph courtesy of H. Weintraub.]

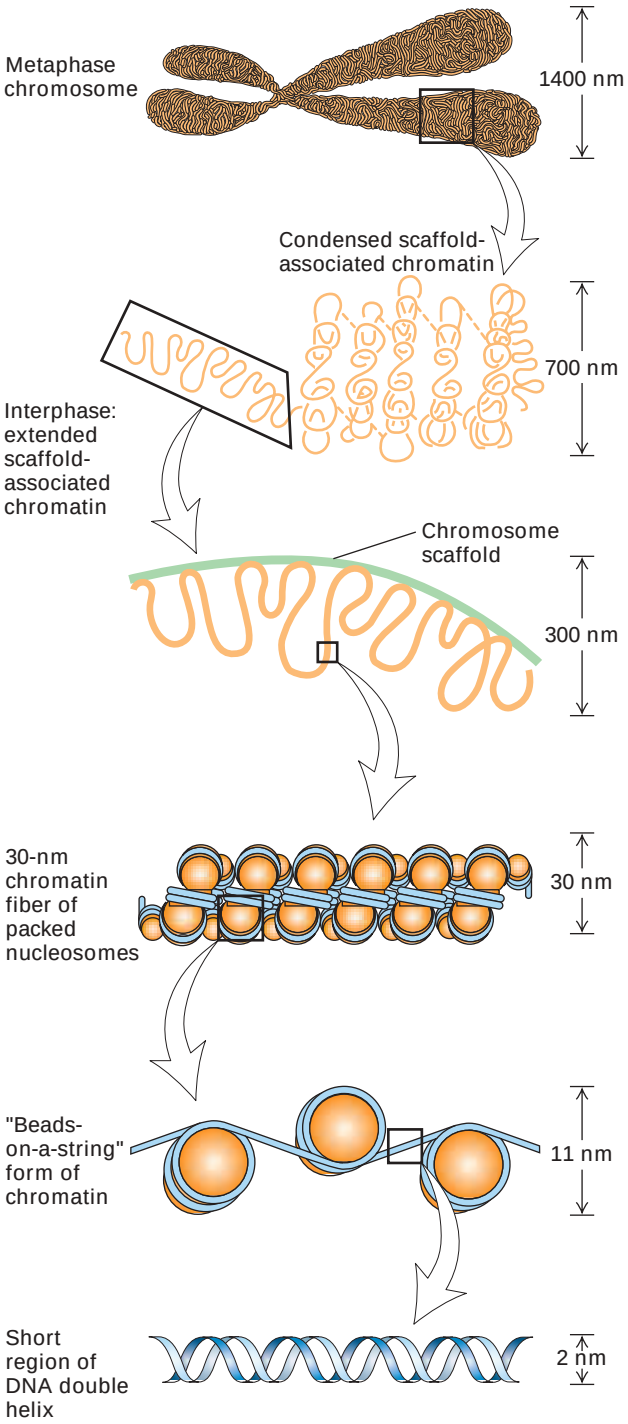
Chapter 11, the control of acetylation of histone N-termini in specific chromosomal regions is thought to contribute to gene control by regulating the strength of the interaction of histones with DNA and the folding of chromatin into condensed structures. Genes in condensed, folded regions of chromatin are inaccessible to RNA polymerase and other proteins required for transcription.

Nonhistone Proteins Provide a Structural Scaffold for Long Chromatin Loops

Although histones are the predominant proteins in chromosomes, nonhistone proteins are also involved in organizing chromosome structure. Electron micrographs of histone-depleted metaphase chromosomes from HeLa cells reveal long loops of DNA anchored to a *chromosome scaffold* composed of nonhistone proteins (Figure 10-23). This scaffold has the shape of the metaphase chromosome and persists even when the DNA is digested by nucleases. As depicted schematically in Figure 10-24, loops of the 30-nm chromatin fiber a few megabases in length have been proposed to associate with a flexible chromosome scaffold, yielding an extended form characteristic of chromosomes during interphase. Folding of the scaffold has been proposed to produce the highly condensed structure characteristic of metaphase chromosomes. But the geometry of scaffold folding in metaphase chromosomes has not yet been determined.



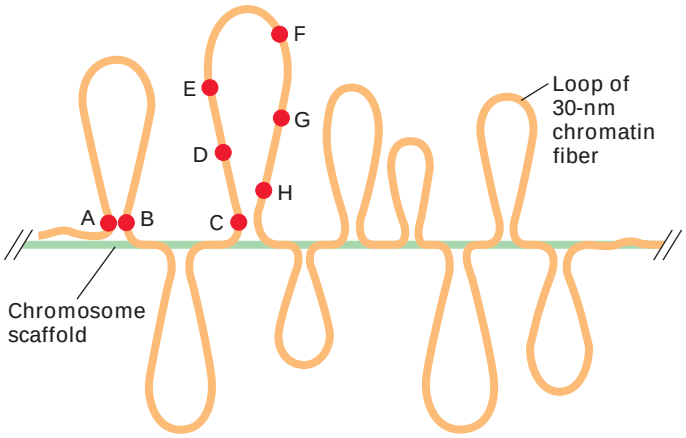
▲ EXPERIMENTAL FIGURE 10-23 **An electron micrograph of a histone-depleted metaphase chromosome reveals the scaffold around which the DNA is organized.** The long loops of DNA are visible extending from the nonhistone protein scaffold (the dark structure). The scaffold shape reflects that of the metaphase chromosome itself. The chromosome was prepared from HeLa cells by treatment with a mild detergent. [From J. R. Paulson and U. K. Laemmli, 1977, *Cell* **12**:817. Copyright 1977 MIT.]



▲ FIGURE 10-24 **Model for the packing of chromatin and the chromosome scaffold in metaphase chromosomes.** In interphase chromosomes, long stretches of 30-nm chromatin loop out from extended scaffolds. In metaphase chromosomes, the scaffold is folded further into a highly compacted structure, whose precise geometry has not been determined.

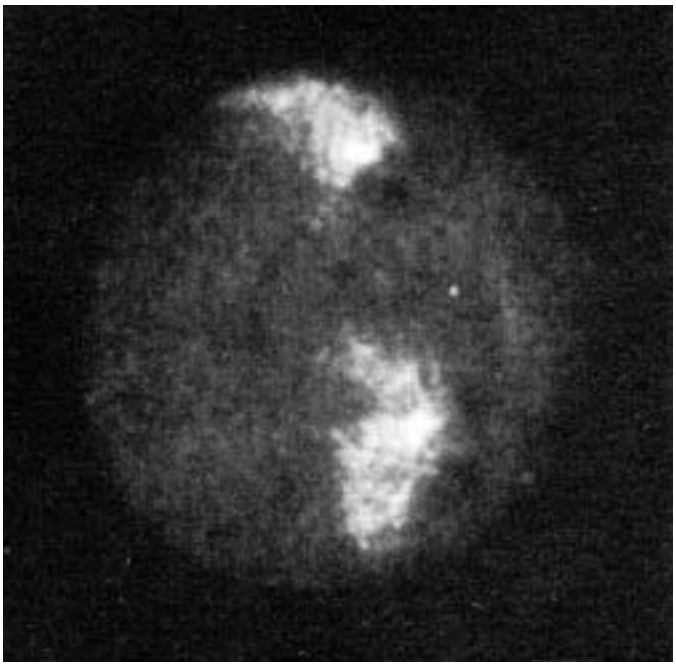
In situ hybridization experiments with several different fluorescent-labeled probes to DNA in human interphase cells support the loop model shown in Figure 10-24. In these experiments, some probe sequences separated by millions of base pairs in linear DNA appeared reproducibly very close to one another in interphase nuclei from different cells (Figure 10-25). These closely spaced probe sites are postulated to lie close to specific sequences in the DNA, called *scaffold-associated regions (SARs)* or *matrix-attachment regions (MARs)*, that are bound to the chromosome scaffold. SARs have been mapped by digesting histone-depleted chromosomes with restriction enzymes and then recovering the fragments that are bound to scaffold proteins.

In general, SARs are found between transcription units. In other words, genes are located primarily within chromatin loops, which are attached at their bases to a chromosome scaffold. Experiments with transgenic mice indicate that in some cases SARs are required for transcription of neighboring genes. In *Drosophila*, some SARs can insulate transcription units from each other, so that proteins regulating transcription of one gene do not influence the transcription of a neighboring gene separated by a SAR.



▲ EXPERIMENTAL FIGURE 10-25 **Fluorescent-labeled probes hybridized to interphase chromosomes demonstrate chromatin loops and permit their measurement.** In situ hybridization of interphase cells was carried out with several different probes specific for sequences separated by known distances in linear, cloned DNA. Lettered circles represent probes. Measurement of the distances between different hybridized probes, which could be distinguished by their color, showed that some sequences (e.g., A, B, and C), separated from one another by millions of base pairs, appear located near one another within nuclei. For some sets of sequences, the measured distances in nuclei between one probe (e.g., C) and sequences successively farther away initially appear to increase (e.g., D, E, and F) and then appear to decrease (e.g., G and H). The measured distances between probes are consistent with loops ranging in size from 1 million to 4 million base pairs.
[Adapted from H. Yokota et al., 1995, *J. Cell Biol.* **130**:1239.]

MEDIA CONNECTIONS
Focus Animation: Three-Dimensional Packing of Nuclear Chromosomes



▲ EXPERIMENTAL FIGURE 10-26 **During interphase human chromosomes remain in specific domains in the nucleus.** Fixed interphase human lymphocytes were hybridized in situ to biotin-labeled probes specific for sequences along the full length of human chromosome 7 and visualized with fluorescently labeled avidin. In the diploid cell shown here, each of the two chromosome 7s is restricted to a territory or domain within the nucleus, rather than stretching throughout the entire nucleus. [From P. Lichter et al., 1988, *Hum. Genet.* **80**:224.]

Individual interphase chromosomes, which are less condensed than metaphase chromosomes, cannot be resolved by standard microscopy or electron microscopy. Nonetheless, the chromatin of interphase cells is associated with extended scaffolds and is further organized into specific domains. This can be demonstrated by the in situ hybridization of interphase nuclei with a large mixture of fluorescent-labeled probes specific for sequences along the length of a particular chromosome. As illustrated in Figure 10-26, the bound probes are visualized within restricted regions or domains of the nucleus rather than appearing throughout the nucleus. Use of probes specific for different chromosomes shows that there is little overlap between chromosomes in interphase nuclei. However, the precise positions of chromosomes are not reproducible between cells.

Chromatin Contains Small Amounts of Other Proteins in Addition to Histones and Scaffold Proteins

The total mass of the histones associated with DNA in chromatin is about equal to that of the DNA. Interphase chromatin and metaphase chromosomes also contain small amounts of a complex set of other proteins. For instance, a growing list of DNA-binding **transcription factors** have been identified associated with interphase chromatin. The structure and function of these critical nonhistone proteins, which help regulate transcription, are examined in Chapter 11. Other low-abundance nonhistone proteins associated with chromatin regulate DNA replication during the eukaryotic cell cycle (Chapter 21).

A few other nonhistone DNA-binding proteins are present in much larger amounts than the transcription or replication factors. Some of these exhibit high mobility during electrophoretic separation and thus have been designated *HMG (high-mobility group) proteins.* When genes encoding the most abundant HMG proteins are deleted from yeast cells, normal transcription is disturbed in most other genes examined. Some HMG proteins have been found to bind to DNA cooperatively with transcription factors that bind to specific DNA sequences, stabilizing multiprotein complexes that regulate transcription of a neighboring gene.

Eukaryotic Chromosomes Contain One Linear DNA Molecule

In lower eukaryotes, the sizes of the largest DNA molecules that can be extracted indicate that each chromosome contains a single DNA molecule. For example, the DNA from each of the quite small chromosomes of *S. cerevisiae* (2.3×10^5 to 1.5×10^6 base pairs) can be separated and individually identified by pulsed-field gel electrophoresis. Physical analysis of the largest DNA molecules extracted from several genetically different *Drosophila* species and strains shows that they are from 6×10^7 to 1×10^8 base pairs long. These sizes match the DNA content of single stained metaphase chromosomes of these *Drosophila* species, as measured by the amount of DNA-specific stain absorbed. The longest DNA molecules in human chromosomes are too large (2.8×10^8 base pairs, or almost 10 cm long) to extract without breaking. Nonetheless, the observations with lower eukaryotes support the conclusion that each chromosome visualized during mitosis contains a single DNA molecule.

To summarize our discussion so far, we have seen that the eukaryotic chromosome is a linear structure composed of an immensely long, single DNA molecule that is wound around histone octamers about every 200 bp, forming strings of closely packed nucleosomes. Nucleosomes fold to form a 30-nm chromatin fiber, which is attached to a flexible protein scaffold at intervals of millions of base pairs, resulting in long loops of chromatin extending from the scaffold (see Figure 10-24). In addition to this general chromosomal structure, a complex set of thousands of low-abundance regulatory proteins are associated with specific sequences in chromosomal DNA.

KEY CONCEPTS OF SECTION 10.4

Structural Organization of Eukaryotic Chromosomes

- In eukaryotic cells, DNA is associated with about an equal mass of histone proteins in a highly condensed nucleoprotein complex called *chromatin*. The building block of chromatin is the nucleosome, consisting of a histone octamer around which is wrapped 147 bp of DNA (see Figure 10-21).
- The chromatin in transcriptionally inactive regions of DNA within cells is thought to exist in a condensed, 30-nm fiber form and higher-order structures built from it (see Figure 10-19).
- The chromatin in transcriptionally active regions of DNA within cells is thought to exist in an open, extended form.
- The reversible acetylation and deacetylation of lysine residues in the N-termini of histones H2A, H2B, H3, and H4 controls how tightly DNA is bound by the histone octamer and affects the assembly of nucleosomes into the condensed forms of chromatin (see Figure 10-20b).
- Histone tails can also be modified by methylation, phosphorylation, and monoubiquitination. These modifications influence chromatin structure by regulating the binding of histone tails to other less abundant chromatin-associated proteins.
- Hypoacetylated, transcriptionally inactive chromatin assumes a more condensed structure and is more resistant to DNase I than hyperacetylated, transcriptionally active chromatin.
- Each eukaryotic chromosome contains a single DNA molecule packaged into nucleosomes and folded into a 30-nm chromatin fiber, which is attached to a protein scaffold at specific sites (see Figure 10-24). Additional folding of the scaffold further compacts the structure into the highly condensed form of metaphase chromosomes.

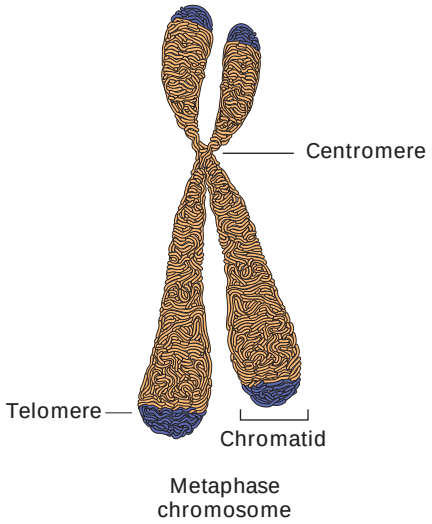
10.5 Morphology and Functional Elements of Eukaryotic Chromosomes

Having examined the detailed structural organization of chromosomes in the previous section, we now view them from a more global perspective. Early microscopic observations on the number and size of chromosomes and their staining patterns led to the discovery of many important general characteristics of chromosome structure. Researchers subsequently identified specific chromosomal regions critical to their replication and segregation to daughter cells during cell division.

Chromosome Number, Size, and Shape at Metaphase Are Species-Specific

As noted previously, in nondividing cells individual chromosomes are not visible, even with the aid of histologic stains for DNA (e.g., Feulgen or Giemsa stains) or electron microscopy. During mitosis and meiosis, however, the chromosomes condense and become visible in the light microscope. Therefore, almost all cytogenetic work (i.e., studies of chromosome morphology) has been done with condensed metaphase chromosomes obtained from dividing cells—either somatic cells in mitosis or dividing gametes during meiosis.

The condensation of metaphase chromosomes probably results from several orders of folding and coiling of 30-nm chromatin fibers (see Figure 10-24). At the time of mitosis, cells have already progressed through the S phase of the cell cycle and have replicated their DNA. Consequently, the chromosomes that become visible during metaphase are duplicated structures. Each metaphase chromosome consists of two sister **chromatids,** which are attached at the centromere (Figure 10-27). The number, sizes, and shapes of the metaphase chromosomes constitute the **karyotype,** which is distinctive for each species. In most organisms, all cells have the same karyotype. However, species that appear quite similar can have very different karyotypes, indicating that similar genetic potential can be organized on chromosomes in very different ways. For example, two species of small deer—the Indian muntjac and Reeves muntjac—contain about the same total amount of genomic DNA. In one species, this



▲ FIGURE 10-27 **Microscopic appearance of typical metaphase chromosome.** Each chromosome has replicated and comprises two chromatids, each containing one of two identical DNA molecules. The centromere, where the chromatids are attached, is required for their separation late in mitosis. Special telomere sequences at the ends function in preventing chromosome shortening.

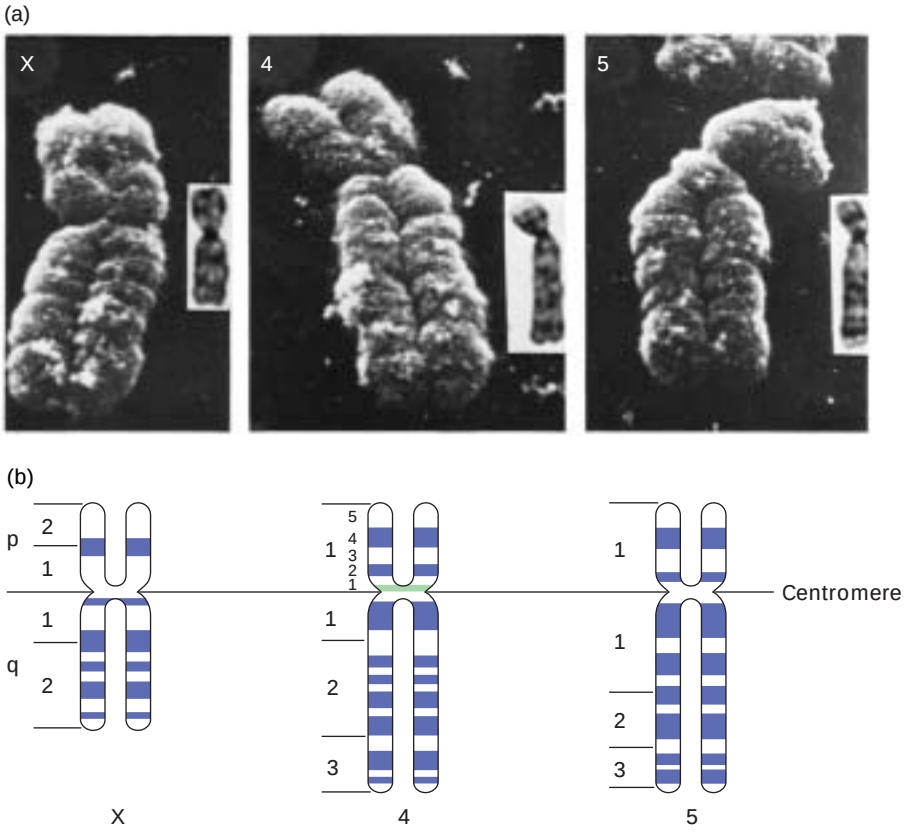
DNA is organized into 22 pairs of homologous **autosomes** and two physically separate sex chromosomes. In contrast, the other species contains only three pairs of autosomes; one sex chromosome is physically separate, but the other is joined to the end of one autosome.

During Metaphase, Chromosomes Can Be Distinguished by Banding Patterns and Chromosome Painting

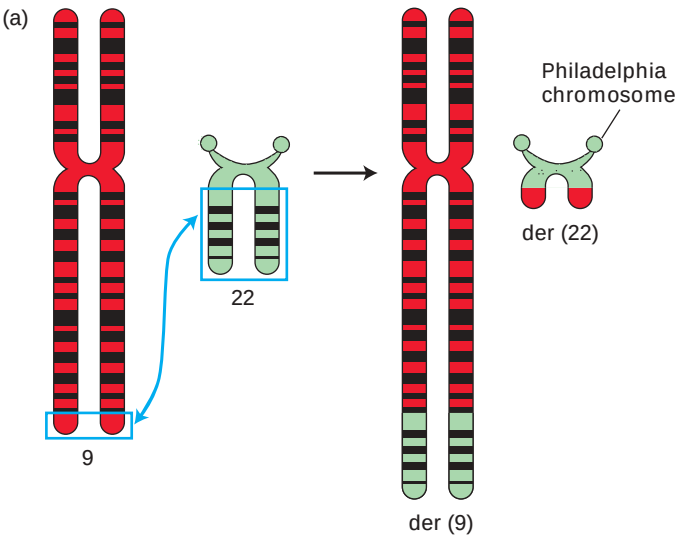
Certain dyes selectively stain some regions of metaphase chromosomes more intensely than other regions, producing characteristic banding patterns that are specific for individual chromosomes. Although the molecular basis for the regularity of chromosomal bands remains unknown, they serve as useful visible landmarks along the length of each chromosome and can help to distinguish chromosomes of similar size and shape.

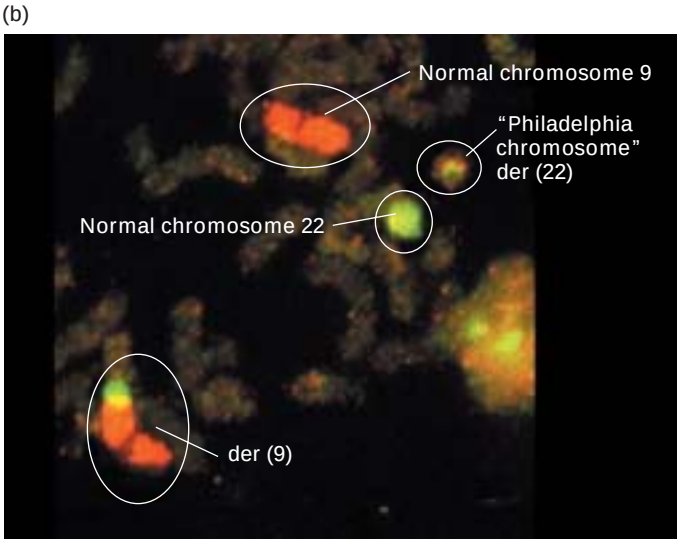
G bands are produced when metaphase chromosomes are subjected briefly to mild heat or proteolysis and then stained with Giemsa reagent, a permanent DNA dye (Figure 10-28). G bands correspond to large regions of the human genome that have an unusually low G+C content. Treatment of chromosomes with a hot alkaline solution before staining with Giemsa reagent produces *R bands* in a pattern that is approximately the reverse of the G-band pattern. The distinctiveness of these banding patterns permits cytologists to identify specific parts of a chromosome and to locate the sites of chromosomal breaks and translocations (Figure 10-29a). In addition, cloned DNA probes that have hybridized to specific sequences in the chromosomes can be located in particular bands.

A recently developed method for visualizing each of the human chromosomes in distinct, bright colors, called *chromosome painting*, greatly simplifies differentiating chromosomes of similar size and shape. This technique, a variation of fluorescence in situ hybridization, makes use of probes specific for sites scattered along the length of each chromosome. The probes are labeled with one of two dyes that fluoresce at different wavelengths. Probes specific for each chromosome are labeled with a predetermined fraction of each of the two dyes. After the probes are hybridized to chromosomes and the excess removed, the sample is observed with a fluorescent microscope in which a detector determines the fraction of each dye present at each fluorescing position in the microscopic field. This information is conveyed to a computer, and a special program assigns a false color image to each type of chromosome. A related technique called *multicolor FISH* can detect chromosomal translocations (Figure 10-29b). The much more detailed analysis possible with this technique permits detection of chromosomal translocations that banding analysis does not reveal.

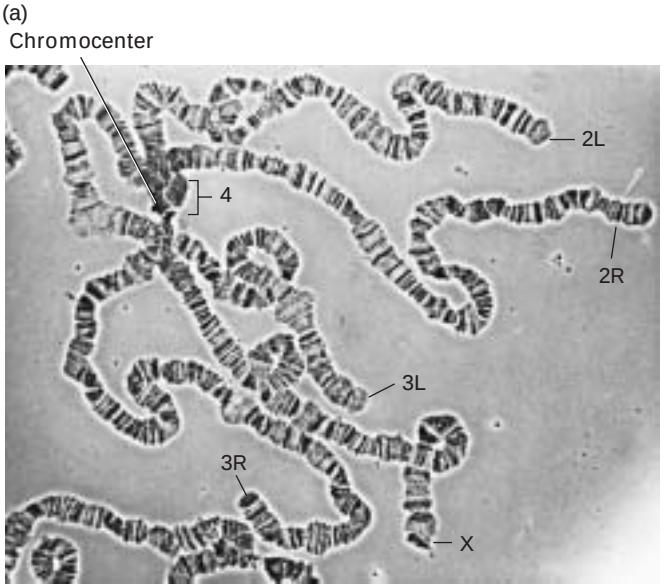


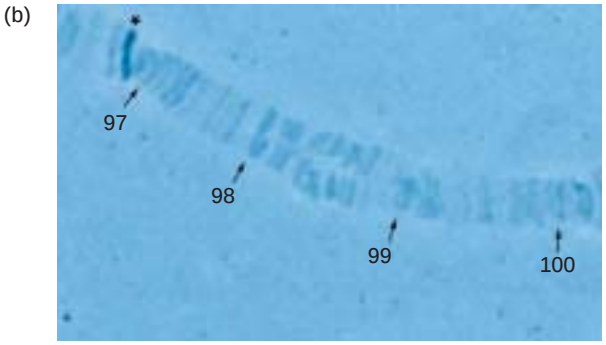
◀ EXPERIMENTAL FIGURE 10-28
G bands produced with Giemsa stains are useful markers for identifying specific chromosomes. The chromosomes are subjected to brief proteolytic treatment and then stained with Giemsa reagent, producing distinctive bands at characteristic places. (a) Scanning electron micrographs of chromosomes X, 4, and 5 show constrictions at the sites where bands appear in light micrographs (*insets*). (b) Standard diagrams of G bands (purple). Regions of variable length (green) are most common in the centromere region (e.g., on chromosome 4). By convention, p denotes the short arm; q, the long arm. Each arm is divided into major sections (1, 2, etc.) and subsections. The short arm of chromosome 4, for example, has five subsections. DNA located in the fourth subsection would be said to be in p14. [Part (a) from C. J. Harrison et al., 1981, *Exp. Cell Res.* **134**:141; courtesy of C. J. Harrison.]





▲ EXPERIMENTAL FIGURE 10-29 **Chromosomal translocations can be analyzed using banding patterns and multicolor FISH.** Characteristic chromosomal translocations are associated with certain genetic disorders and specific types of cancers. For example, in nearly all patients with chronic myelogenous leukemia, the leukemic cells contain the Philadelphia chromosome, a shortened chromosome 22 [der (22)], and an abnormally long chromosome 9 [der (9)]. These result from a translocation between normal chromosomes 9 and 22. This translocation can be detected by classical banding analysis (a) and by multicolor FISH (b). [Part (a) from J. Kuby, 1997, *Immunology,* 3d ed., W. H. Freeman and Company, p. 578; part (b) courtesy of J. Rowley and R. Espinosa.]



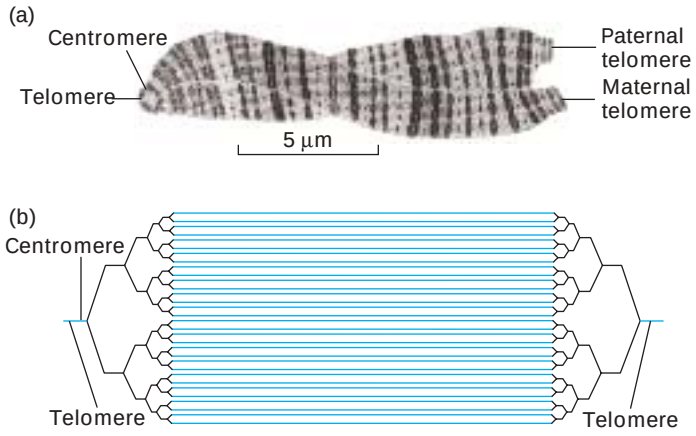


▶ EXPERIMENTAL FIGURE 10-30 **Banding on *Drosophila* polytene salivary gland chromosomes and in situ hybridization are used together to localize gene sequences.** (a) In this light micrograph of *Drosophila melanogaster* larval salivary gland chromosomes, four chromosomes can be observed (X, 2, 3, and 4), with a total of approximately 5000 distinguishable bands. The centromeres of all four chromosomes often appear fused at the chromocenter. The tips of chromosomes 2 and 3 are labeled (L = left arm; R = right arm), as is the tip of the X chromosome. (b) A particular DNA sequence can be mapped on *Drosophila* salivary gland chromosomes by in situ hybridization. This photomicrograph shows a portion of a chromosome that was hybridized with a cloned DNA sequence labeled with biotin-derivatized nucleotides. Hybridization is detected with the biotin-binding protein avidin that is covalently bound to the enzyme alkaline phosphatase. On addition of a soluble substrate, the enzyme catalyzes a reaction that results in formation of an insoluble colored precipitate at the site of hybridization (asterisk). Since the very reproducible banding patterns are characteristic of each *Drosophila* polytene chromosome, the hybridized sequence can be located on a particular chromosome. The numbers indicate major bands. Bands between those indicated are designated with numbers and letters (not shown). [Part (a) courtesy of J. Gall; part (b) courtesy of F. Pignoni.]

Interphase Polytene Chromosomes Arise by DNA Amplification

The larval salivary glands of *Drosophila* species and other dipteran insects contain enlarged interphase chromosomes that are visible in the light microscope. When fixed and stained, these **polytene chromosomes** are characterized by a large number of reproducible, well-demarcated bands that have been assigned standardized numbers (Figure 10-30a). The highly reproducible banding pattern seen in *Drosophila* salivary gland chromosomes provides an extremely powerful method for locating specific DNA sequences along the lengths of the chromosomes in this species. For example, the chromosomal location of a cloned DNA sequence can be accurately determined by hybridizing a labeled sample of the cloned DNA to polytene chromosomes prepared from larval salivary glands (Figure 10-30b).

A generalized amplification of DNA gives rise to the polytene chromosomes found in the salivary glands of *Drosophila.* This process, termed *polytenization,* occurs when the DNA repeatedly replicates, but the daughter chromosomes do not separate. The result is an enlarged chromosome composed of many parallel copies of itself (Figure 10-31). The amplification of chromosomal DNA greatly increases gene copy number, presumably to supply sufficient mRNA for protein synthesis in the massive salivary gland cells. Although the bands seen in human metaphase chromosomes probably represent very long folded or compacted stretches of DNA containing about 10^7 base pairs, the bands in *Drosophila* polytene chromosomes represent much shorter stretches of only 50,000–100,000 base pairs.



▲ FIGURE 10-31 **Amplification of DNA along the polytene fourth chromosome of *Drosophila melanogaster.*** (a) Morphology of the stained polytene fourth chromosome as it appears in the light microscope at high magnification. The fourth chromosome is by far the smallest (see Figure 10-30). The homologous chromatids (paternal and maternal) are paired. The banding pattern results from reproducible packing of DNA and protein within each amplified site along the chromosome. Dark bands are regions of more highly compacted chromatin. (b) The pattern of amplification of one of the two homologous chromatids during five replications. Double-stranded DNA is represented by a single line. Telomere and centromere DNA are not amplified. In salivary gland polytene chromosomes, each parental chromosome undergoes ≈10 replications (2^{10} = 1024 strands). [Part (a) from C. Bridges, 1935, *J. Hered.* **26**:60; part (b) adapted from C. D. Laird et al., 1973, *Cold Spring Harbor Symp. Quant. Biol.* **38**:311.]

Heterochromatin Consists of Chromosome Regions That Do Not Uncoil

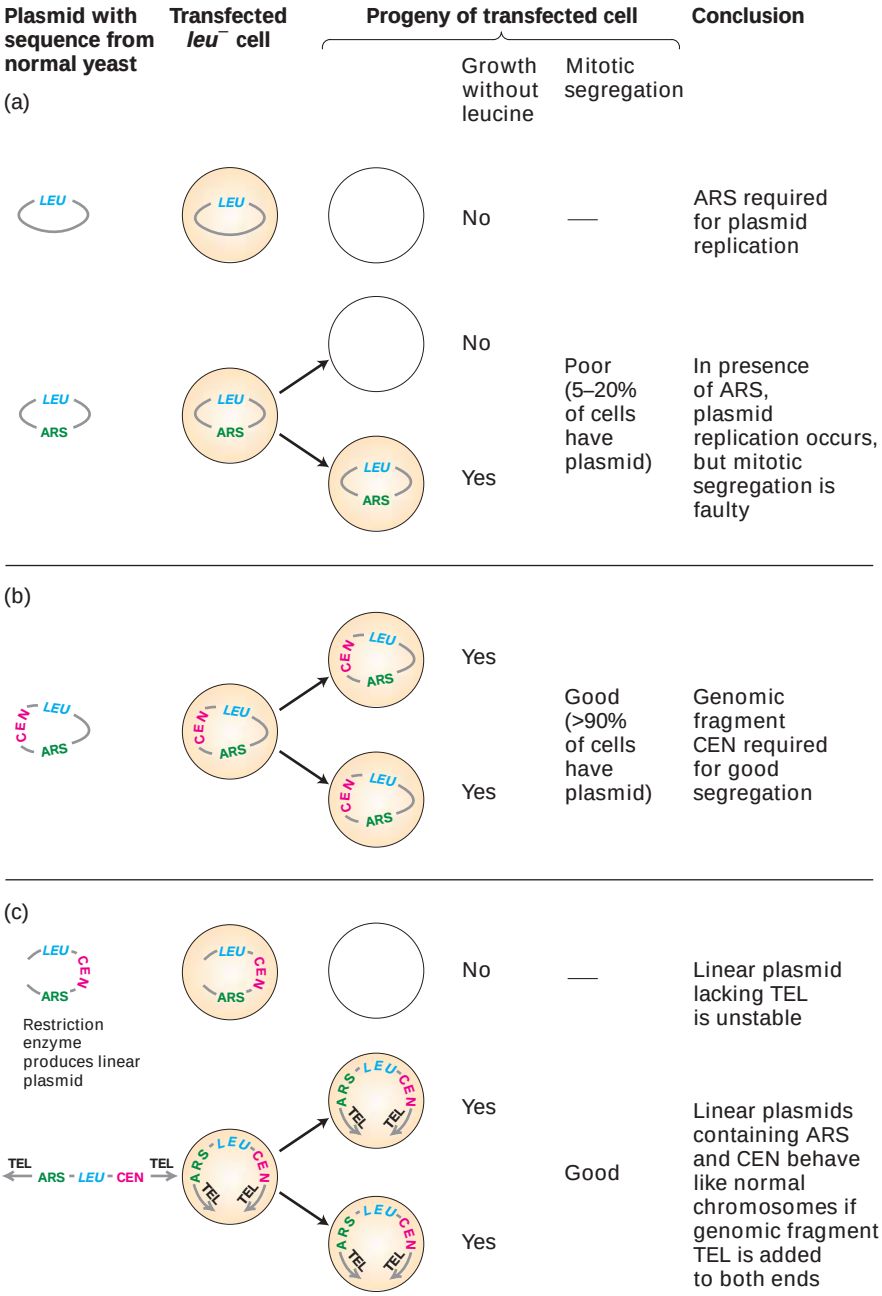
As cells exit from mitosis and the condensed chromosomes uncoil, certain sections of the chromosomes remain dark-staining. The dark-staining areas, termed **heterochromatin,** are regions of condensed chromatin. The light-staining, less condensed portions of chromatin are called **euchromatin.** Heterochromatin appears most frequently—but not exclusively—at the centromere and telomeres of chromosomes and is mostly simple-sequence DNA.

In mammalian cells, heterochromatin appears as darkly staining regions of the nucleus, often associated with the nuclear envelope. Pulse labeling with ^{3}H-uridine and autoradiography have shown that most transcription occurs in regions of euchromatin and the nucleolus. Because of this and because heterochromatic regions apparently remain condensed throughout the life cycle of the cell, they have been regarded as sites of inactive genes. However, some transcribed genes have been located in regions of heterochromatin. Also, not all inactive genes and nontranscribed regions of DNA are visible as heterochromatin.

Three Functional Elements Are Required for Replication and Stable Inheritance of Chromosomes

Although chromosomes differ in length and number between species, cytogenetic studies have shown that they all behave similarly at the time of cell division. Moreover, any eukaryotic chromosome must contain three functional elements in order to replicate and segregate correctly: (1) **replication origins** at which DNA polymerases and other proteins initiate synthesis of DNA (see Figures 4-34 and 4-36), (2) the centromere, and (3) the two ends, or telomeres. The yeast transformation studies depicted in Figure 10-32 demonstrated the functions of these three chromosomal elements and established their importance for chromosome function.

As discussed in Chapter 4, replication of DNA begins from sites that are scattered throughout eukaryotic chromosomes. The yeast genome contains many ≈100-bp sequences, called *autonomously replicating sequences (ARSs),* that act as replication origins. The observation that insertion of an ARS into a circular plasmid allows the plasmid to replicate in yeast cells provided the first functional identification of origin sequences in eukaryotic DNA (see Figure 10-32a).



▲ EXPERIMENTAL FIGURE 10-32 **Yeast transfection experiments identify the functional chromosomal elements necessary for normal chromosome replication and segregation.** In these experiments, plasmids containing the *LEU* gene from normal yeast cells are constructed and introduced into *leu*$^-$ cells by transfection. If the plasmid is maintained in the *leu*$^-$ cells, they are transformed to *LEU*$^+$ by the *LEU* gene on the plasmid and can form colonies on medium lacking leucine. (a) Sequences that allow autonomous replication (ARS) of a plasmid were identified because their insertion into a plasmid vector containing a cloned *LEU* gene resulted in a high frequency of transformation to *LEU*$^+$. However, even plasmids with ARS exhibit poor segregation during mitosis, and therefore do not appear in each of the daughter cells. (b) When randomly broken pieces of genomic yeast DNA are inserted into plasmids containing ARS and *LEU,* some of the subsequently transfected cells produce large colonies, indicating that a high rate of mitotic segregation among their plasmids is facilitating the continuous growth of daughter cells. The DNA recovered from plasmids in these large colonies contains yeast centromere (CEN) sequences. (c) When *leu*$^-$ yeast cells are transfected with linearized plasmids containing *LEU,* ARS, and CEN, no colonies grow. Addition of telomere (TEL) sequences to the ends of the linear DNA gives the linearized plasmids the ability to replicate as new chromosomes that behave very much like a normal chromosome in both mitosis and meiosis. [See A. W. Murray and J. W. Szostak, 1983, *Nature* **305**:89, and L. Clarke and J. Carbon, 1985, *Ann. Rev. Genet.* **19**:29.]

Even though circular ARS-containing plasmids can replicate in yeast cells, only about 5–20 percent of progeny cells contain the plasmid because mitotic segregation of the plasmids is faulty. However, plasmids that also carry a CEN sequence, derived from the centromeres of yeast chromosomes, segregate equally or nearly so to both mother and daughter cells during mitosis (see Figure 10-32b).

If circular plasmids containing an ARS and CEN sequence are cut once with a restriction enzyme, the resulting linear plasmids do not produce LEU^+ colonies unless they contain special telomeric (TEL) sequences ligated to their ends (see Figure 10-32c). The first successful experiments involving transfection of yeast cells with linear plasmids were achieved by using the ends of a DNA molecule that was known to replicate as a linear molecule in the ciliated protozoan *Tetrahymena*. During part of the life cycle of *Tetrahymena*, much of the nuclear DNA is repeatedly copied in short pieces to form a so-called *macronucleus*. One of these repeated fragments was identified as a dimer of ribosomal DNA, the ends of which contained a repeated sequence $(G_4T_2)_n$. When a section of this repeated TEL sequence was ligated to the ends of linear yeast plasmids containing ARS and CEN, replication and good segregation of the linear plasmids occurred.

Centromere Sequences Vary Greatly in Length

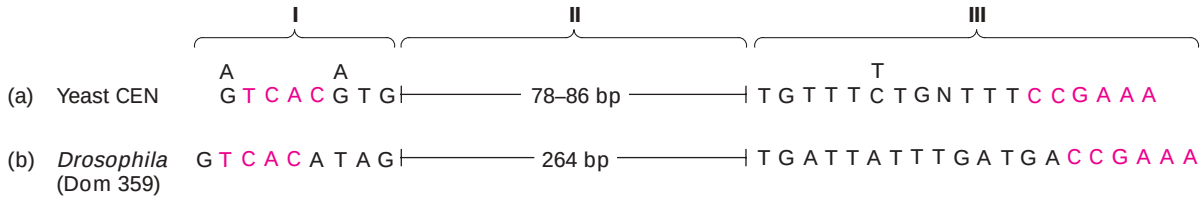
Once the yeast centromere regions that confer mitotic segregation were cloned, their sequences could be determined and compared, revealing three regions (I, II, and III) conserved between them (Figure 10-33a). Short, fairly well conserved nucleotide sequences are present in regions I and III. Although region II seems to have a fairly constant length, it contains no definite consensus sequence; however, it is rich in A and T residues. Regions I and III are bound by proteins that interact with a set of more than 30 proteins that bind the short *S. cerevisiae* chromosome to one microtubule of the spindle apparatus during mitosis. Region II is bound to a nucleosome that has a variant form of histone H3 replacing the usual H3. Centromeres from all eukaryotes similarly are bound by nucleosomes with a specialized, centromere-specific form of histone H3 called *CENP-A* in humans. *S. cerevisiae* has by far the simplest centromere sequence known in nature.

In the fission yeast *S. pombe*, centromeres are ≈40 kb in length and are composed of repeated copies of sequences similar those in *S. cerevisiae* centromeres. Multiple copies of proteins homologous to those that interact with the *S. cerevisiae* centromeres bind to these complex *S. pombe* centromeres and in turn bind the much longer *S. pombe* chromosomes to several microtubules of the mitotic spindle apparatus. In plants and animals, centromeres are megabases in length and are composed of multiple repeats of simple-sequence DNA. One *Drosophila* simple-sequence DNA, which comes from a centromeric region, has a repeat unit that bears some similarity to yeast CEN regions I and III (see Figure 10-33b). In humans, centromeres contain 2- to 4-megabase arrays of a 171-bp simple-sequence DNA called *alphoid* DNA that is bound by nucleosomes with the CENP-A histone H3 variant, as well as other repeated simple-sequence DNA.

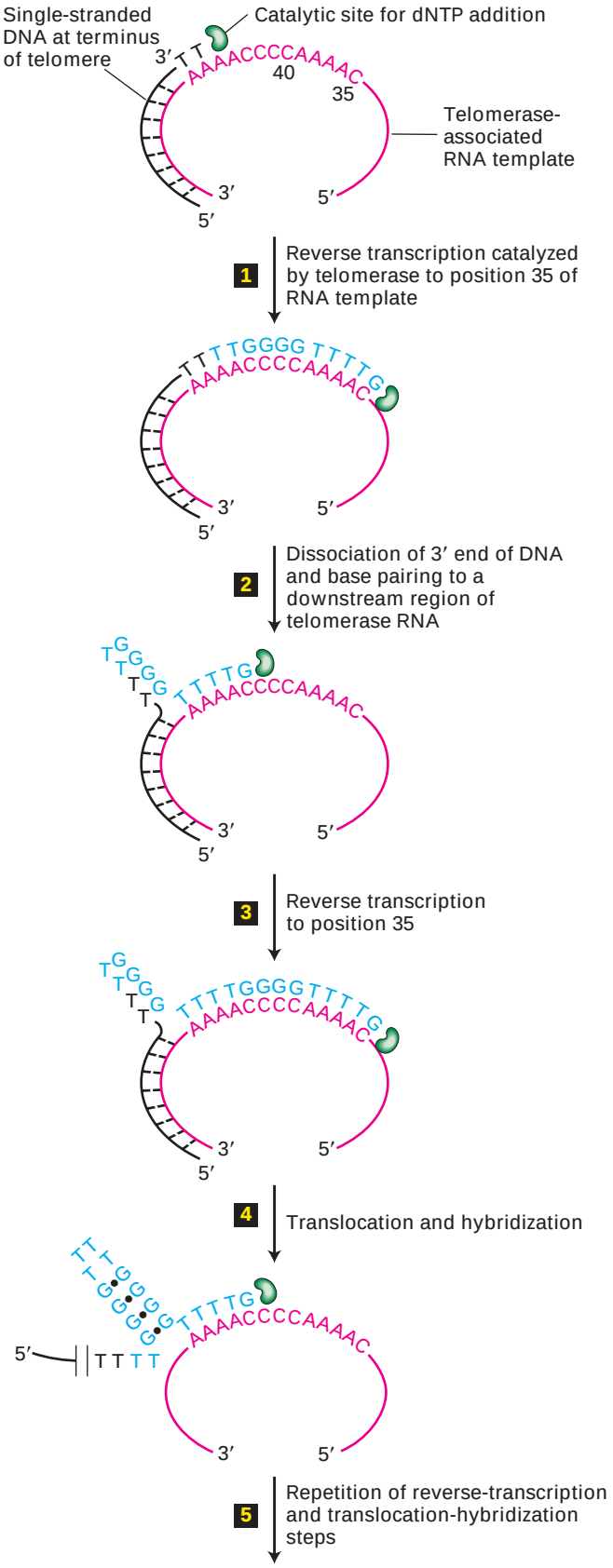
In higher eukaryotes, a complex protein structure called the **kinetochore** assembles at centromeres and associates with multiple mitotic spindle fibers during mitosis. Homologs of most of the centromeric proteins found in the yeasts occur in humans and other higher eukaryotes and are thought to be components of kinetochores. The role of the centromere and proteins that bind to it in the segregation of sister chromatids during mitosis is described in Chapters 20 and 21.

Addition of Telomeric Sequences by Telomerase Prevents Shortening of Chromosomes

Sequencing of telomeres from a dozen or so organisms, including humans, has shown that most are repetitive oligomers with a high G content in the strand with its 3′ end at the end of the chromosome. The telomere repeat sequence in humans and other vertebrates is TTAGGG. These simple sequences are repeated at the very termini of chromosomes for a total of a few hundred base pairs in yeasts and protozoans and a few thousand base pairs in vertebrates. The 3′ end of the G-rich strand extends 12–16 nucleotides beyond the 5′ end of the complementary C-rich strand. This region is bound by specific proteins that both protect the ends of



▲ FIGURE 10-33 **Comparison of yeast CEN sequence and *Drosophila* simple-sequence DNA.** (a) Consensus yeast CEN sequence, based on analysis of 10 yeast centromeres, includes three conserved regions. Region II, although variable in sequence, is fairly constant in length and is rich in A and T residues. (b) One *Drosophila* simple-sequence DNA located near the centromere has a repeat unit with some homology to the yeast consensus CEN, including two identical 4-bp and 6-bp stretches (red). [See L. Clarke and J. Carbon, 1985, *Ann. Rev. Genet.* **19**:29.]



linear chromosomes from attack by exonucleases and associate telomeres in specific domains within the nucleus.

The need for a specialized region at the ends of eukaryotic chromosomes is apparent when we consider that all known DNA polymerases elongate DNA chains at the 3′ end, and all require an RNA or DNA primer. As the growing fork approaches the end of a linear chromosome, synthesis of the leading strand continues to the end of the DNA template strand, completing one daughter DNA double helix. However, because the lagging-strand template is copied in a discontinuous fashion, it cannot be replicated in its entirety (see Figure 4-33). When the final RNA primer is removed, there is no upstream strand onto which DNA polymerase can build to fill the resulting gap. Without some special mechanism, the daughter DNA strand resulting from lagging-strand synthesis would be shortened at each cell division.

The problem of telomere shortening is solved by an enzyme that adds telomeric sequences to the ends of each chromosome. The enzyme is a protein and RNA complex called *telomere terminal transferase,* or *telomerase*. Because the sequence of the telomerase-associated RNA, as we will see, serves as the template for addition of deoxyribonucleotides to the ends of telomeres, the source of the enzyme and not the source of the telomeric DNA primer determines the sequence added. This was proved by transforming *Tetrahymena* with a mutated form of the gene encoding the

◀ FIGURE 10-34 **Mechanism of action of telomerase.** The single-stranded 3′ terminus of a telomere is extended by telomerase, counteracting the inability of the DNA replication mechanism to synthesize the extreme terminus of linear DNA. Telomerase elongates this single-stranded end by a reiterative reverse-transcription mechanism. The action of the telomerase from the protozoan *Oxytricha,* which adds a T_4G_4 repeat unit, is depicted; other telomerases add slightly different sequences. The telomerase contains an RNA template (red) that base-pairs to the 3′ end of the lagging-strand template. The telomerase catalytic site (green) then adds deoxyribonucleotides (blue) using the RNA molecule as a template; this reverse transcription proceeds to position 35 of the RNA template (step 1). The strands of the resulting DNA-RNA duplex are then thought to slip relative to each other, leading to displacement of a single-stranded region of the telomeric DNA strand and to uncovering of part of the RNA template sequence (step 2). The lagging-strand telomeric sequence is again extended to position 35 by telomerase, and the DNA-RNA duplex undergoes translocation and hybridization as before (steps 3 and 4). The slippage mechanism is thought to be facilitated by the unusual base pairing (black dots) between the displaced G residues, which is less stable than Watson-Crick base pairing. Telomerases can add multiple repeats by repetition of steps 3 and 4. DNA polymerase α-primase can prime synthesis of new Okazaki fragments on this extended template strand. The net result prevents shortening of the lagging strand at each cycle of DNA replication [Adapted from D. Shippen-Lentz and E. H. Blackburn, 1990, *Nature* **247**:550.]

MEDIA CONNECTIONS
Focus Animation: Telomere Replication

telomerase-associated RNA. The resulting telomerase added a DNA sequence complementary to the mutated RNA sequence to the ends of telomeric primers. Thus telomerase is a specialized form of a reverse transcriptase that carries its own internal RNA template to direct DNA synthesis.

Figure 10-34 depicts how telomerase, by reverse transcription of its associated RNA, elongates the 3′ end of the single-stranded DNA at the end of the G-rich strand mentioned above. Cells from knockout mice that cannot produce the telomerase-associated RNA exhibit no telomerase activity, and their telomeres shorten successively with each cell generation. Such mice can breed and reproduce normally for three generations before the long telomere repeats become substantially eroded. Then, the absence of telomere DNA results in adverse effects, including fusion of chromosome termini and chromosomal loss. By the fourth generation, the reproductive potential of these knockout mice declines, and they cannot produce offspring after the sixth generation.

The human genes expressing the telomerase protein and the telomerase-associated RNA are active in germ cells but are turned off in most adult tissues, in which cells replicate only rarely. However, these genes are reactivated in most human cancer cells, where telomerase is required for the multiple cell divisions necessary to form a tumor. This phenomenon has stimulated a search for inhibitors of human telomerase as potential therapeutic agents for treating cancer. ∎

While telomerase prevents telomere shortening in most eukaryotes, some organisms use alternative strategies. *Drosophila* species maintain telomere lengths by the regulated insertion of non-LTR retrotransposons into telomeres. This is one of the few instances in which a mobile element has a specific function in its host organism.

Yeast Artificial Chromosomes Can Be Used to Clone Megabase DNA Fragments

The research on circular and linear plasmids in yeast identified all the basic components of a *yeast artificial chromosome (YAC).* To construct YACs, TEL sequences from yeast cells or from the protozoan *Tetrahymena* are combined with yeast CEN and ARS sequences; to these are added DNA with selectable yeast genes and enough DNA from any source to make a total of more than 50 kb. (Smaller DNA segments do not work as well.) Such artificial chromosomes replicate in yeast cells and segregate almost perfectly, with only 1 daughter cell in 1000 to 10,000 failing to receive an artificial chromosome. During meiosis, the two sister chromatids of the artificial chromosome separate correctly to produce haploid spores. The successful propagation of YACs and studies such as those depicted in Figure 10-32 strongly support the conclusion that yeast chromosomes, and probably all eukaryotic chromosomes, are linear, double-stranded DNA molecules containing special regions that ensure replication and proper segregation.

YACs are valuable experimental tools that can be used to clone very long chromosomal pieces from other species. For instance, YACs have been used extensively for cloning fragments of human DNA up to 1000 kb (1 megabase) in length, permitting the isolation of overlapping YAC clones that collectively encompass nearly the entire length of individual human chromosomes. Currently, research is in progress to construct human artificial chromosomes using the much longer centromeric DNA of human chromosomes and the longer human telomeres. Such human artificial chromosomes might be useful in gene therapy to treat human genetic diseases such as sickle-cell anemia, cystic fibrosis, and inborn errors of metabolism. ∎

KEY CONCEPTS OF SECTION 10.5

Morphology and Functional Elements of Eukaryotic Chromosomes

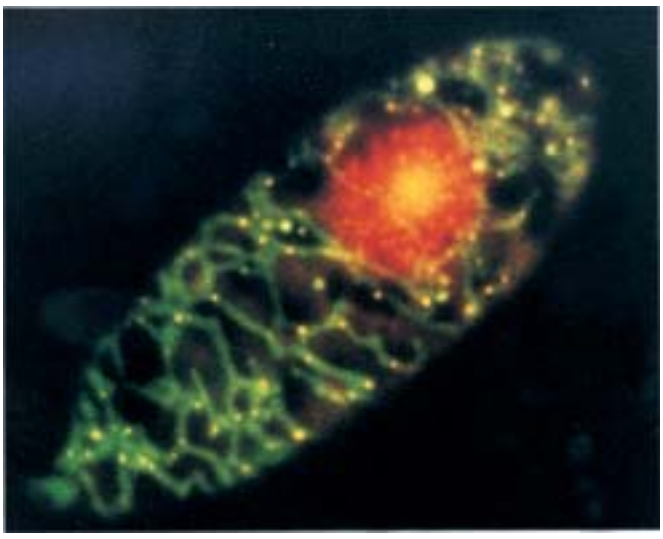
- During metaphase, eukaryotic chromosomes become sufficiently condensed that they can be visualized individually in the light microscope.
- The karyotype, the set of metaphase chromosomes, is characteristic of each species. Closely related species can have dramatically different karyotypes, indicating that similar genetic information can be organized on chromosomes in different ways.
- Banding analysis and chromosome painting, a more precise method, are used to identify the different human metaphase chromosomes and to detect translocations and deletions (see Figure 10-29).
- In certain cells of the fruit fly *Drosophila melanogaster* and related insects, interphase chromosomes are reduplicated 10 times, generating polytene chromosomes that are visible in the light microscope (see Figure 10-31).
- The highly reproducible banding patterns of polytene chromosomes make it possible to localize cloned *Drosophila* DNA on a *Drosophila* chromosome by in situ hybridization (Figure 10-30) and to visualize chromosomal deletions and rearrangements as changes in the normal pattern of bands.
- When metaphase chromosomes decondense during interphase, certain regions, termed *heterochromatin,* remain much more condensed than the bulk of chromatin, called *euchromatin.*
- Three types of DNA sequences are required for a long linear DNA molecule to function as a chromosome: a replication origin, called *ARS* in yeast; a centromere (CEN) sequence; and two telomere (TEL) sequences at the ends of the DNA (see Figure 10-32).
- DNA fragments up to 10^6 base pairs long can be cloned in yeast artificial chromosome (YAC) vectors.

10.6 Organelle DNAs

Although the vast majority of DNA in most eukaryotes is found in the nucleus, some DNA is present within the mitochondria of animals, plants, and fungi and within the chloroplasts of plants. These organelles are the main cellular sites for ATP formation, during oxidative phosphorylation in mitochondria and photosynthesis in chloroplasts (Chapter 8). Many lines of evidence indicate that mitochondria and chloroplasts evolved from bacteria that were endocytosed into ancestral cells containing a eukaryotic nucleus, forming *endosymbionts.* Over evolutionary time, most of the bacterial genes encoding components of the present-day organelles were transferred to the nucleus. However, mitochondria and chloroplasts in today's eukaryotes retain DNAs encoding proteins essential for organellar function as well as the ribosomal and transfer RNAs required for their translation. Thus eukaryotic cells have multiple genetic systems: a predominant nuclear system and secondary systems with their own DNA in the mitochondria and chloroplasts.

Mitochondria Contain Multiple mtDNA Molecules

Individual mitochondria are large enough to be seen under the light microscope, and even the mitochondrial DNA (mtDNA) can be detected by fluorescence microscopy. The mtDNA is located in the interior of the mitochondrion, the region known as the matrix (see Figure 5-26). As judged by the number of yellow fluorescent "dots" of mtDNA, a *Euglena gracilis* cell contains at least 30 mtDNA molecules (Figure 10-35).



▲ EXPERIMENTAL FIGURE 10-35 **Dual staining reveals the multiple mitochondrial DNA molecules in a growing *Euglena gracilis* cell.** Cells were treated with a mixture of two dyes: ethidium bromide, which binds to DNA and emits a red fluorescence, and DiOC6, which is incorporated specifically into mitochondria and emits a green fluorescence. Thus the nucleus emits a red fluorescence, and areas rich in mitochondrial DNA fluoresce yellow—a combination of red DNA and green mitochondrial fluorescence. [From Y. Hayashi and K. Ueda, 1989, *J. Cell Sci.* **93**:565.]

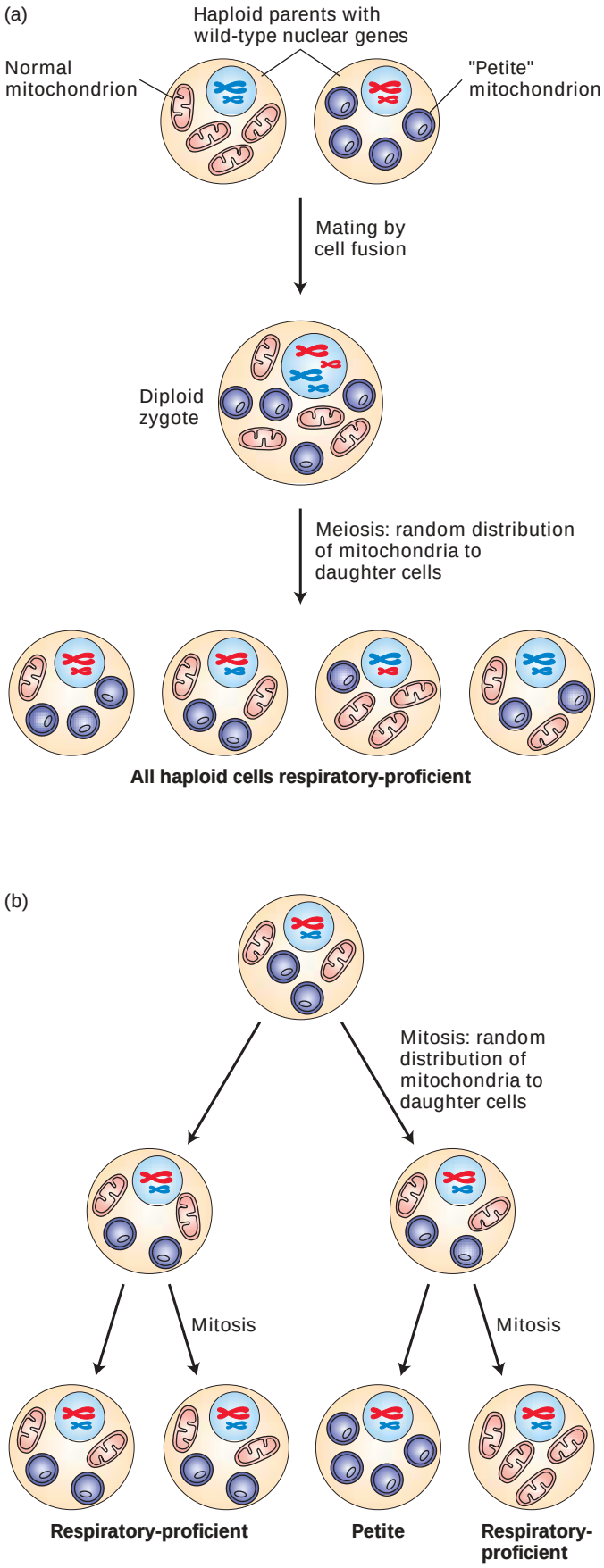
Since the dyes used to visualize nuclear and mitochondrial DNA do not affect cell growth or division, replication of mtDNA and division of the mitochondrial network can be followed in living cells using time-lapse microscopy. Such studies show that in most organisms mtDNA replicates throughout interphase. At mitosis each daughter cell receives approximately the same number of mitochondria, but since there is no mechanism for apportioning exactly equal numbers of mitochondria to the daughter cells, some cells contain more mtDNA than others. By isolating mitochondria from cells and analyzing the DNA extracted from them, it can be seen that each mitochondrion contains multiple mtDNA molecules. Thus the total amount of mtDNA in a cell depends on the number of mitochondria, the size of the mtDNA, and the number of mtDNA molecules per mitochondrion. Each of these parameters varies greatly between different cell types.

mtDNA Is Inherited Cytoplasmically and Encodes rRNAs, tRNAs, and Some Mitochondrial Proteins

Studies of mutants in yeasts and other single-celled organisms first indicated that mitochondria exhibit *cytoplasmic inheritance* and thus must contain their own genetic system (Figure 10-36). For instance, *petite* yeast mutants exhibit structurally abnormal mitochondria and are incapable of oxidative phosphorylation. As a result, petite cells grow more slowly than wild-type yeasts and form smaller colonies (hence the name "petite"). Genetic crosses between different (haploid) yeast strains showed that the *petite* mutation does not segregate with any known nuclear gene or chromosome. In later studies, most petite mutants were found to contain deletions of mtDNA.

In the mating by fusion of haploid yeast cells, both parents contribute equally to the cytoplasm of the resulting diploid; thus inheritance of mitochondria is biparental (see Figure 10-36a). In mammals and most other multicellular organisms, however, the sperm contributes little (if any) cytoplasm to the zygote, and virtually all the mitochondria in the embryo are derived from those in the egg, not the sperm. Studies in mice have shown that 99.99 percent of mtDNA is maternally inherited, but a small part (0.01 percent) is inherited from the male parent. In higher plants, mtDNA is inherited exclusively in a uniparental fashion through the female parent (egg), not the male (pollen).

The entire mitochondrial genome from a number of different organisms has now been cloned and sequenced, and



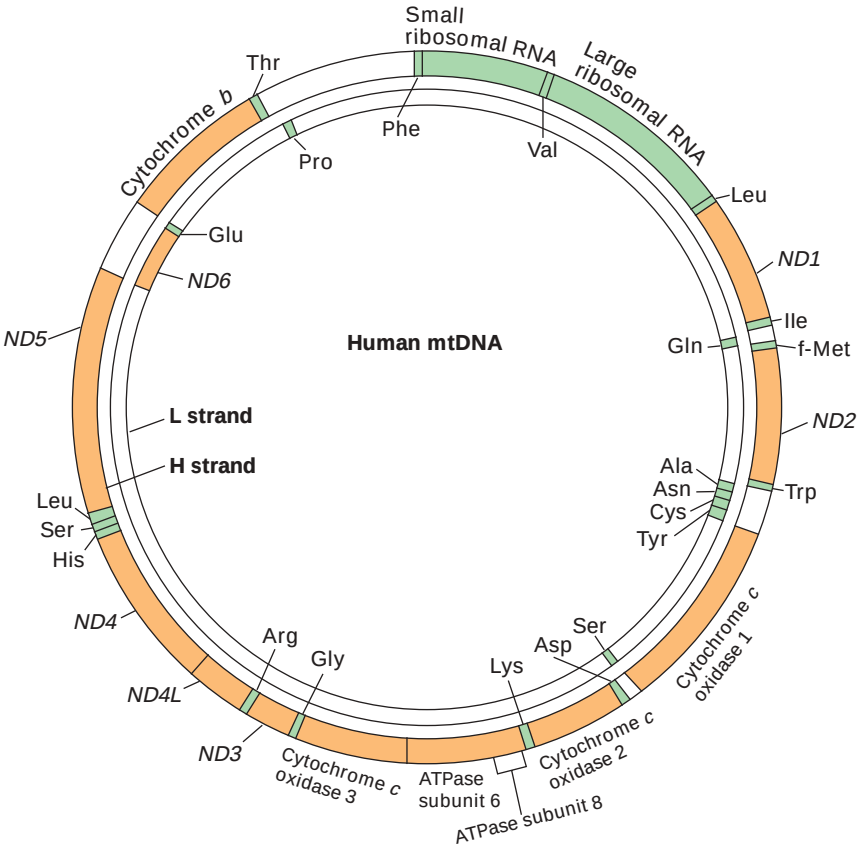
mtDNAs from all these sources have been found to encode rRNAs, tRNAs, and essential mitochondrial proteins. All proteins encoded by mtDNA are synthesized on mitochondrial ribosomes. All mitochondrially synthesized polypeptides identified thus far (with one possible exception) are not complete enzymes but subunits of multimeric complexes used in electron transport or ATP synthesis. Most proteins localized in mitochondria, such as the mitochondrial RNA and DNA polymerases, are synthesized on cytosolic ribosomes and are imported into the organelle by processes discussed in Chapter 16.

The Size and Coding Capacity of mtDNA Vary Considerably in Different Organisms

Surprisingly, the size of the mtDNA, the number and nature of the proteins it encodes, and even the mitochondrial genetic code itself vary greatly between different organisms. Human mtDNA, a circular molecule that has been completely sequenced, is among the smallest known mtDNAs, containing 16,569 base pairs (Figure 10-37). It encodes the two rRNAs found in mitochondrial ribosomes and the 22 tRNAs used to translate mitochondrial mRNAs. Human mtDNA has 13 sequences that begin with an ATG (methionine) codon, end with a stop codon, and are long enough to encode a polypeptide of more than 50 amino acids; all the possible proteins encoded by these open reading frames have been identified. Mammalian mtDNA, in contrast to nuclear DNA, lacks introns and contains no long noncoding sequences.

The mtDNA from most multicellular animals (metazoans) is about the same size as human mtDNA and encodes similar gene products. In contrast, yeast mtDNA is almost five times as large (≈78,000 bp). The mtDNAs from yeast and other fungi encode many of the same gene products as mammalian mtDNA, as well as others whose genes are found in the nuclei of metazoan cells.

◀ FIGURE 10-36 **Cytoplasmic inheritance of the *petite* mutation in yeast.** Petite-strain mitochondria are defective in oxidative phosphorylation owing to a deletion in mtDNA. (a) Haploid cells fuse to produce a diploid cell that undergoes meiosis, during which random segregation of parental chromosomes and mitochondria containing mtDNA occurs. Since yeast normally contain ≈50 mtDNA molecules per cell, all products of meiosis usually contain both normal and petite mtDNAs and are capable of respiration. (b) As these haploid cells grow and divide mitotically, the cytoplasm (including the mitochondria) is randomly distributed to the daughter cells. Occasionally, a cell is generated that contains only defective petite mtDNA and yields a petite colony. Thus formation of such petite cells is independent of any nuclear genetic marker.



▶ FIGURE 10-37 **The coding capacity of human mitochondrial DNA (mtDNA).** Proteins and RNAs encoded by each of the two strands are shown separately. Transcription of the outer (H) strand occurs in the clockwise direction and of the inner (L) strand in the counterclockwise direction. The abbreviations for amino acids denote the corresponding tRNA genes. *ND1*, *ND2*, etc., denote genes encoding subunits of the NADH-CoQ reductase complex. The 207-bp gene encoding F_0 ATPase subunit 8 overlaps, out of frame, with the N-terminal portion of the segment encoding F_0 ATPase subunit 6. Mammalian mtDNA genes do not contain introns, although intervening DNA lies between some genes. [See D. A. Clayton, 1991, *Ann. Rev. Cell Biol.* **7**:453.]

In contrast to other eukaryotes, which contain a single type of mtDNA, plants contain several types of mtDNA that appear to recombine with one another. Plant mtDNAs are much larger and more variable in size than the mtDNAs of other organisms. Even in a single family of plants, mtDNAs can vary as much as eightfold in size (watermelon = 330 kb; muskmelon = 2500 kb). Unlike animal, yeast, and fungal mtDNAs, plant mtDNAs contain genes encoding a 5S mitochondrial rRNA, which is present only in the mitochondrial ribosomes of plants, and the α subunit of the F_1 ATPase. The mitochondrial rRNAs of plants are also considerably larger than those of other eukaryotes. Long, noncoding regions and duplicated sequences are largely responsible for the greater length of plant mtDNAs. ▮

Differences in the size and coding capacity of mtDNA from various organisms most likely reflect the movement of DNA between mitochondria and the nucleus during evolution. Direct evidence for this movement comes from the observation that several proteins encoded by mtDNA in some species are encoded by nuclear DNA in others. It thus appears that entire genes moved from the mitochondrion to the nucleus, or vice versa, during evolution.

The most striking example of this phenomenon involves the gene *cox II*, which encodes subunit 2 of cytochrome *c* oxidase. This gene is found in mtDNA in all organisms studied except for one species of legume, the mung bean: in this organism only, the *cox II* gene is nuclear. Many RNA transcripts of plant mitochondrial genes are edited, mainly by the enzyme-catalyzed conversion of selected C residues to U, and occasionally U to C. (RNA editing is discussed in Chapter 12.) The nuclear *cox II* gene of mung bean corresponds more closely to the edited cox II RNA transcripts than to the mitochondrial *cox II* genes found in other legumes. These observations are strong evidence that the *cox II* gene moved from the mitochondrion to the nucleus during mung bean evolution by a process that involved an RNA intermediate. Presumably this movement involved a reverse-transcription mechanism similar to that by which processed pseudogenes are generated in the nuclear genome from nuclear-encoded mRNAs.

Products of Mitochondrial Genes Are Not Exported

As far as is known, all RNA transcripts of mtDNA and their translation products remain in the mitochondrion, and all mtDNA-encoded proteins are synthesized on mitochondrial ribosomes. Mitochondria encode the rRNAs that form mitochondrial ribosomes, although all but one or two of the ribosomal proteins (depending on the species) are imported from the cytosol. In most eukaryotes, all the tRNAs used for protein synthesis in mitochondria are encoded by mtDNAs. However, in wheat, in the parasitic protozoan *Trypanosoma brucei* (the cause of African sleeping sickness), and in ciliated

protozoans, most mitochondrial tRNAs are encoded by the nuclear DNA and imported into the mitochondrion.

Reflecting the bacterial ancestry of mitochondria, mitochondrial ribosomes resemble bacterial ribosomes and differ from eukaryotic cytosolic ribosomes in their RNA and protein compositions, their size, and their sensitivity to certain antibiotics (see Figure 4-24). For instance, chloramphenicol blocks protein synthesis by bacterial and mitochondrial ribosomes from most organisms, but cycloheximide does not. This sensitivity of mitochondrial ribosomes to the important aminoglycoside class of antibiotics is the main cause of the toxicity that these antibiotics can cause. Conversely, cytosolic ribosomes are sensitive to cycloheximide and resistant to chloramphenicol. In cultured mammalian cells the only proteins synthesized in the presence of cycloheximide are encoded by mtDNA and produced by mitochondrial ribosomes. ▮

Mitochondrial Genetic Codes Differ from the Standard Nuclear Code

The genetic code used in animal and fungal mitochondria is different from the standard code used in all prokaryotic and eukaryotic nuclear genes; remarkably, the code even differs in mitochondria from different species (Table 10-3). Why and how these differences arose during evolution is mysterious. UGA, for example, is normally a stop codon, but is read as tryptophan by human and fungal mitochondrial translation systems; however, in plant mitochondria, UGA is still recognized as a stop codon. AGA and AGG, the standard nuclear codons for arginine, also code for arginine in fungal and plant mtDNA, but they are stop codons in mammalian mtDNA and serine codons in *Drosophila* mtDNA.

As shown in Table 10-3, plant mitochondria appear to utilize the standard genetic code. However, comparisons of the amino acid sequences of plant mitochondrial proteins with the nucleotide sequences of plant mtDNAs suggested that CGG could code for *either* arginine (the "standard" amino acid) or tryptophan. This apparent nonspecificity of the plant mitochondrial code is explained by editing of mitochondrial RNA transcripts, which can convert cytosine residues to uracil residues. If a CGG sequence is edited to UGG, the codon specifies tryptophan, the standard amino acid for UGG, whereas unedited CGG codons encode the standard arginine. Thus the translation system in plant mitochondria does utilize the standard genetic code. ▮

Mutations in Mitochondrial DNA Cause Several Genetic Diseases in Humans

The severity of disease caused by a mutation in mtDNA depends on the nature of the mutation and on the proportion of mutant and wild-type mtDNAs present in a particular cell type. Generally, when mutations in mtDNA are found, cells contain mixtures of wild-type and mutant mtDNAs—a condition known as *heteroplasmy*. Each time a mammalian somatic or germ-line cell divides, the mutant and wild-type mtDNAs will segregate randomly into the daughter cells, as occurs in yeast cells (see Figure 10-36). Thus, the mtDNA genotype, which fluctuates from one generation and from one cell division to the next, can drift toward predominantly wild-type or predominantly mutant mtDNAs. Since all enzymes for the replication and growth of mitochondria, such as DNA and RNA polymerases, are imported from the cytosol, a mutant mtDNA should not be at a "replication disadvantage"; mutants that involve large

TABLE 10-3 Alterations in the Standard Genetic Code in Mitochondria

Codon	Standard Code: Nuclear-Encoded Proteins	Mammals	Mitochondria			
			Drosophila	*Neurospora*	Yeasts	Plants
UGA	Stop	Trp	Trp	Trp	Trp	Stop
AGA, AGG	Arg	Stop	Ser	Arg	Arg	Arg
AUA	Ile	Met	Met	Ile	Met	Ile
AUU	Ile	Met	Met	Met	Met	Ile
CUU, CUC, CUA, CUG	Leu	Leu	Leu	Leu	Thr	Leu

SOURCES: S. Anderson et al., 1981, *Nature* **290**:457; P. Borst, in *International Cell Biology 1980–1981*, H. G. Schweiger, ed., Springer-Verlag, p. 239; C. Breitenberger and U. L. Raj Bhandary, 1985, *Trends Biochem. Sci.* **10**:478–483; V. K. Eckenrode and C. S. Levings, 1986, *In Vitro Cell Dev. Biol.* **22**:169–176; J. M. Gualber et al., 1989, *Nature* **341**:660–662; and P. S. Covello and M. W. Gray, 1989, *Nature* **341**:662–666.

deletions of mtDNA might even be at a selective advantage in replication because they can replicate faster.

All cells have mitochondria, yet mutations in mtDNA affect only some tissues. Those most usually affected are tissues that have a high requirement for ATP produced by oxidative phosphorylation and tissues that require most of or all the mtDNA in the cell to synthesize sufficient amounts of functional mitochondrial proteins. *Leber's hereditary optic neuropathy* (degeneration of the optic nerve, accompanied by increasing blindness), for instance, is caused by a missense mutation in the mtDNA gene encoding subunit 4 of the NADH-CoQ reductase. Any of several large deletions in mtDNA causes another set of diseases including *chronic progressive external ophthalmoplegia* and *Kearns-Sayre syndrome*, which are characterized by eye defects and, in Kearns-Sayre syndrome, also by abnormal heartbeat and central nervous system degeneration. A third condition, causing "ragged" muscle fibers (with improperly assembled mitochondria) and associated uncontrolled jerky movements, is due to a single mutation in the TψCG loop of the mitochondrial lysine tRNA. As a result of this mutation, the translation of several mitochondrial proteins apparently is inhibited ▮

Chloroplasts Contain Large Circular DNAs Encoding More Than a Hundred Proteins

As we discuss in Chapter 8, the structure of chloroplasts is similar in many respects to that of mitochondria. Like mitochondria, chloroplasts contain multiple copies of the organellar DNA and ribosomes, which synthesize some chloroplast-encoded proteins using the "standard" genetic code. Other chloroplast proteins are fabricated on cytosolic ribosomes and are incorporated into the organelle after translation (Chapter 16).

Chloroplast DNAs are circular molecules of 120,000–160,000 bp, depending on the species. The complete sequences of several chloroplast DNAs have been determined. Of the ≈120 genes in chloroplast DNA, about 60 are involved in RNA transcription and translation, including genes for rRNAs, tRNAs, RNA polymerase subunits, and ribosomal proteins. About 20 genes encode subunits of the chloroplast photosynthetic electron-transport complexes and the F_0F_1 ATPase complex. Also encoded in the chloroplast genome is the larger of the two subunits of ribulose 1,5-bisphosphate carboxylase, which is involved in the fixation of carbon dioxide during photosynthesis.

Reflecting the endosymbiotic origin of chloroplasts, some regions of chloroplast DNA are strikingly similar to the DNA of present-day bacteria. For instance, chloroplast DNA encodes four subunits of RNA polymerase that are highly homologous to the subunits of *E. coli* RNA polymerase. One segment of chloroplast DNA encodes eight proteins that are homologous to eight *E. coli* ribosomal proteins; moreover, the order of these genes is the same in the two DNAs.

Although the overall organization of chloroplast DNAs from different species is similar, some differences in gene composition occur. For instance, liverwort chloroplast DNA has some genes that are not detected in the larger tobacco chloroplast DNA, and vice versa. Since chloroplasts in both species contain virtually the same set of proteins, these data suggest that some genes are present in the chloroplast DNA of one species and in the nuclear DNA of the other, indicating that some exchange of genes between chloroplast and nucleus occurred during evolution. ▮

Methods similar to those used for the transformation of yeast cells (Chapter 9) have been developed for stably introducing foreign DNA into the chloroplasts of higher plants. The large number of chloroplast DNA molecules per cell permits the introduction of thousands of copies of an engineered gene into each cell, resulting in extraordinarily high levels of foreign protein production. Chloroplast transformation has recently led to the engineering of plants that are resistant to bacterial and fungal infections, drought, and herbicides. The level of production of foreign proteins is comparable with that achieved with engineered bacteria, making it likely that chloroplast transformation will be used for the production of human pharmaceuticals and possibly for the engineering of food crops containing high levels of all the amino acids essential to humans. ▮

KEY CONCEPTS OF SECTION 10.6

Organelle DNAs

- Mitochondria and chloroplasts most likely evolved from bacteria that formed a symbiotic relationship with ancestral cells containing a eukaryotic nucleus. Most of the genes originally within these organelles moved to the nuclear genome over evolutionary time, leaving different gene sets in the mtDNAs of different organisms.
- Most mitochondrial and chloroplast DNAs are circular molecules, reflecting their probable bacterial origin. Plant mtDNAs and chloroplast DNAs generally are longer than mtDNAs from other eukaryotes, largely because they contain more noncoding regions and repetitive sequences.
- All mtDNAs and chloroplast DNAs appear to encode rRNAs, tRNAs, and some of the proteins involved in mitochondrial or photosynthetic electron transport and ATP synthesis.
- Because most mtDNA is inherited from egg cells rather than sperm, mutations in mtDNA exhibit a maternal cytoplasmic pattern of inheritance.

- Mitochondrial ribosomes resemble bacterial ribosomes in their structure, sensitivity to chloramphenicol, and resistance to cycloheximide.
- The genetic code of animal and fungal mtDNAs differs slightly from that of bacteria and the nuclear genome and varies between different animals and plants (see Table 10-3). In contrast, plant mtDNAs and chloroplast DNAs appear to conform to the standard genetic code.
- Several human neuromuscular disorders result from mutations in mtDNA. Patients generally have a mixture of wild-type and mutant mtDNA in their cells (heteroplasmy): the higher the fraction of mutant mtDNA, the more severe the mutant phenotype.

PERSPECTIVES FOR THE FUTURE

The human genome sequence is a goldmine for new discoveries about molecular cell biology, new proteins that may be the basis of effective therapies of human diseases, and even for discoveries concerning early human history and evolution. However, because only ≈1.5 percent of the sequence encodes proteins, finding new genes is like finding a needle in a haystack. Identification of genes in bacterial genome sequences is relatively simple because of the scarcity of introns. Simply searching for long open reading frames free of stop codons identifies most genes. In contrast, the search for human genes is vastly complicated by the structure of human genes, most of which are composed of multiple, relatively short exons separated by much longer, noncoding introns. Some have estimated that only approximately half of all human genes have been identified. And the identification of complex transcription units by examining DNA sequence alone is even more challenging. Future improvements in bioinformatic methods for gene identification, and further characterization of cDNA copies of mRNAs isolated from the hundreds of human cell types, will likely lead to the discovery of new proteins and, through their study, to a new understanding of biological processes and applications to medicine and agriculture.

We have seen that although most transposons do not function directly in cellular processes, they have helped to shape modern genomes by promoting gene duplications, exon shuffling, the generation of new combinations of transcription control sequences, and other aspects of contemporary genomes. They also have the potential today to teach us about our own history and origins. This is because L1 and *Alu* retrotransposons have inserted into new sites in individuals throughout our history. Large numbers of these interspersed repeats are polymorphic in the population—they occur at a particular site in some individuals and not others. Individuals sharing an insertion at a particular site descended from the same common ancestor that developed from an egg or sperm in which that insertion occurred. The time elapsed from the initial insertion can be estimated by the differences in sequences of the element among individuals that carry them because these differences arose from the accumulation of random mutations. Further studies of these retrotransposon polymorphisms will undoubtedly add immensely to our understanding both of human migrations since *Homo sapiens* first evolved and of the history of contemporary populations.

KEY TERMS

centromere *413*
chromatid *430*
chromatin *424*
chromosome scaffold *427*
cytoplasmic inheritance *438*
euchromatin *433*
exon shuffling *422*
fluorescence in situ hybridization (FISH) *413*
G bands *431*
gene family *410*
heterochromatin *433*
histones *424*
karyotype *430*
LINEs *420*
long terminal repeats (LTRs) *417*
mobile DNA elements *414*
nucleosome *424*
P element *416*
polytene chromosomes *433*
pseudogene *411*
retrotransposons *415*
scaffold-associated regions (SARs) *428*
simple-sequence DNA *412*
SINEs *420*
telomere *413*
transcription unit *407*
transposition *414*
transposons *414*

REVIEW THE CONCEPTS

1. Genes can be transcribed into mRNA for protein-coding genes or RNA for genes such as ribosomal or transfer RNAs. Define a gene. Describe how a complex transcription unit can be alternatively processed to generate a variety of mRNAs and ultimately proteins.

2. Sequencing of the human genome has revealed much about the organization of genes. Describe the differences between single genes, gene families, pseudogenes, and tandemly repeated genes.

3. Much of the human genome consists of repetitive DNA. Describe the difference between microsatellite and minisatellite DNA. How is this repetitive DNA useful for identifying individuals by the technique of DNA fingerprinting?

4. Mobile DNA elements that can move or transpose to a new site directly as DNA are called DNA transposons. Describe the mechanism by which a bacterial insertion sequence can transpose.

5. Retrotransposons are a class of mobile elements that use a RNA intermediate. Contrast the mechanism of transposition between retrotransposons that contain and those that lack long terminal repeats (LTRs).

6. Describe the role that mobile DNA elements may have played in the evolution of modern organisms. What is the process known as exon shuffling, and what role do mobile DNA elements play in this process?

7. DNA in a cell associates with proteins to form chromatin. What is a nucleosome? What role do histones play in nucleosomes? How are nucleosomes arranged in condensed 30-nm fibers?

8. Describe the general organization of a eukaryotic chromosome. What structural role do scaffold associated regions (SARs) or matrix attachment regions (MARs) play? Where are genes primarily located relative to chromosome structure?

9. Metaphase chromosomes can be identified by characteristic banding patterns. What are G bands and R bands? What is chromosome painting, and how is this technique useful?

10. Replication and segregation of eukaryotic chromosomes require three functional elements: replication origins, a centromere, and telomeres. Describe how these three elements function. What is the role of telomerase in maintaining chromosome structure? What is a yeast artificial chromosome (YAC)?

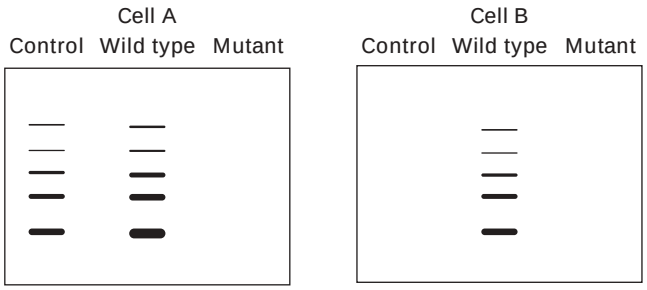
11. Mitochondria contain their own DNA molecules. Describe the types of genes encoded in the mitochondrial genome. How do the mitochondrial genomes of plants, fungi, and animals differ?

12. Mitochondria and chloroplasts are thought to have evolved from symbiotic bacteria present in nucleated cells. Review the experimental evidence that supports this hypothesis.

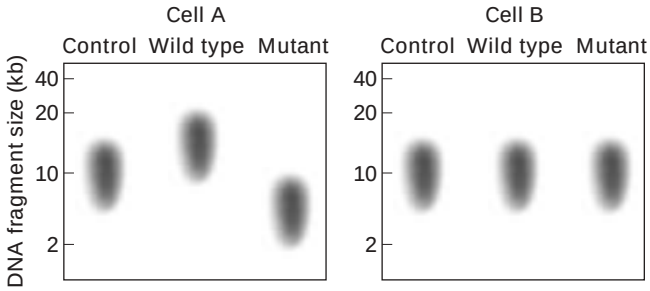
ANALYZE THE DATA

In culture, normal human cells undergo a finite number of cell divisions until they no longer proliferate and enter a state known as replicative senescence. The inability to maintain normal telomere length is thought to play an important role in this process. Telomerase is a ribonucleoprotein complex that regenerates the ends of telomeres lost during each round of DNA replication. Human telomerase consists of a template containing RNA subunit and a catalytic protein subunit known as human telomerase reverse transcriptase (hTERT). Most normal cells do not express telomerase; most cancer cells do express telomerase. Thus, telomerase is proposed to play a key role in the transformation of cells from a normal to a malignant state.

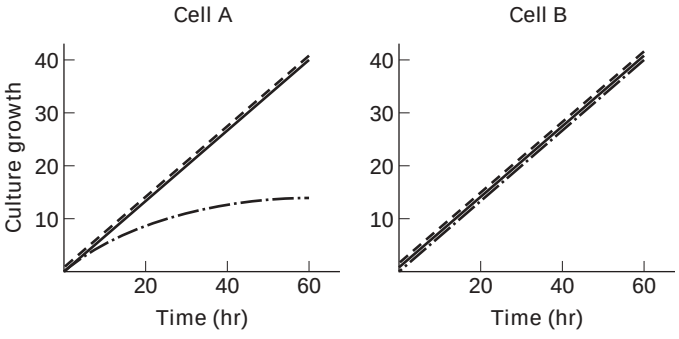
a. In the following experiments, the role of telomerase in the growth of human cancer cells was investigated (see W. C. Hahn et al., 1999, *Nature Medicine* **5**:1164–1170). Immortal, telomerase-positive cells (cell A) and immortal, telomerase-negative cells (cell B) were transfected with a plasmid expressing a wild-type or mutated hTERT. Telomerase activity in cell extracts was measured using the telomeric repeat amplification protocol (TRAP) assay, a PCR-based assay that measures the addition of telomere repeat units onto a DNA fragment. A six-base-pair ladder pattern is typically seen. Control indicates transfection of cells with just the plasmid. Wild type and mutant indicate transfection with a plasmid expressing a wild-type hTERT or the mutated hTERT, respectively. What do you conclude about the effect of the mutant hTERT on telomerase activity in the transfected cells? What type of mutation would this represent?



b. Telomere length in these transfected cells was examined by Southern blot analysis of total genomic DNA digested with a restriction enzyme and probed with a telomere-specific DNA sequence. What do you conclude about the lengths of the telomeres in cell A and cell B after transfection with the wild-type or mutant hTERT? By what mechanism do cells A and B maintain their telomere lengths?



c. The proliferation of transfected cells A and B was assayed by measuring the number of cells versus time in culture. What do you conclude about the effect of the mutant hTERT on cell proliferation in transfected cells A and B?



d. Do the results of the three assays support or refute the proposed role of telomerase in long-term survival of cells in culture? What is the value of comparing transfection of cells A with cells B?

e. Tumorigenic cells, which can replicate continuously in culture, can cause tumors when injected into mice that lack a functional immune system (nude mice). From the above assays, what would you predict would happen if wild-type or mutant hTERT transfected cells A were injected into nude mice?

REFERENCES

Molecular Definition of a Gene

Ayoubi, T. A., and W. J. Van De Ven. 1996. Regulation of gene expression by alternative promoters. *FASEB J.* **10**:453–460.

Maniatis, T., and B. Tasic. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**:236–243.

Mathe, C., et al. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucl. Acids Res.* **30**:4103–4117.

Chromosomal Organization of Genes and Noncoding DNA

Cummings, C. J., and H. Y. Zoghbi. 2000. Trinucleotide repeats: mechanisms and pathophysiology. *Ann. Rev. Genomics Hum. Genet.* **1**:281–328.

Goff, S. A., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**:92–100.

Lander, E. S., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

Ranum, L. P., and J. W. Day. 2002. Dominantly inherited, noncoding microsatellite expansion disorders. *Curr. Opin. Genet. Dev.* **12**:266–271.

Mobile DNA

Feschotte, C., N. Jiang, and S. R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**:329–341.

Gray, Y. H. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet.* **16**:461–468.

Kazazian, H. H., Jr. 1999. An estimated frequency of endogenous insertional mutations in humans. *Nature Genet.* **22**:130.

Ostertag, E. M., and H. H. Kazazian, Jr. 2001. Biology of mammalian L1 retrotransposons. *Ann. Rev. Genet.* **35**:501–538.

Structural Organization of Eukaryotic Chromosomes

Horn, P. J., and C. L. Peterson. 2002. Molecular biology. Chromatin higher order folding—wrapping up transcription. *Science* **297**:1824–1827.

Luger, K., and T. J. Richmond. 1998. The histone tails of the nucleosome. *Curr. Opin. Genet. Dev.* **8**:140–146.

Morphology and Functional Elements of Eukaryotic Chromosomes

Cheeseman, I. M., D. G. Drubin, and G. Barnes. 2002. Simple centromere, complex kinetochore: linking spindle microtubules and centromeric DNA in budding yeast. *J. Cell Biol.* **157**:199–203.

Kelleher, C., et al. 2002. Telomerase: biochemical considerations for enzyme and substrate. *Trends Biochem. Sci.* **27**:572–579.

Kitagawa, K., and P. Hieter. 2001. Evolutionary conservation between budding yeast and human kinetochores. *Nature Rev. Mol. Cell Biol.* **2**:678–687.

Larin, Z., and J. E. Mejia. 2002. Advances in human artificial chromosome technology. *Trends Genet.* **18**:313–319.

McEachern, M. J., A. Krauskopf, and E. H. Blackburn. 2000. Telomeres and their control. *Ann. Rev. Genet.* **34**:331–358.

Organelle DNA

Clayton, D. A. 2000. Transcription and replication of mitochondrial DNA. *Hum. Reprod.* **2**(Suppl.):11–17.

Dahl, H.-H. M., and D. R. Thorburn. 2001. Mitochondrial diseases: beyond the magic circle. *Am. J. Med. Genet.* **106**:1–3.

Daniell, H., M. S. Khan, and L. Allison. 2002. Milestones in chloroplast genetic engineering: an environmentally friendly era in biotechnology. *Trends Plant Sci.* **7**:84–91.

Gray, M. W. 1999. Evolution of organellar genomes. *Curr. Opin. Genet. Dev.* **9**:678–687.

Palmer, J. D., et al. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Nat'l. Acad. Sci. USA* **97**:6960–6966.

Sugiura, M., T. Hirose, and M. Sugita. 1998. Evolution and mechanism of translation in chloroplasts. *Ann. Rev. Genet.* **32**:437–459.