DNA, Chromosomes, and Genomes

4

Life depends on the ability of cells to store, retrieve, and translate the genetic instructions required to make and maintain a living organism. This *hereditary* information is passed on from a cell to its daughter cells at cell division, and from one generation of an organism to the next through the organism's reproductive cells. These instructions are stored within every living cell as its **genes**, the information-containing elements that determine the characteristics of a species as a whole and of the individuals within it.

As soon as genetics emerged as a science at the beginning of the twentieth century, scientists became intrigued by the chemical structure of genes. The information in genes is copied and transmitted from cell to daughter cell millions of times during the life of a multicellular organism, and it survives the process essentially unchanged. What form of molecule could be capable of such accurate and almost unlimited replication and also be able to direct the development of an organism and the daily life of a cell? What kind of instructions does the genetic information contain? How can the enormous amount of information required for the development and maintenance of an organism fit within the tiny space of a cell?

The answers to several of these questions began to emerge in the 1940s. At this time, researchers discovered, from studies in simple fungi, that genetic information consists primarily of instructions for making proteins. Proteins are the macromolecules that perform most cell functions: they serve as building blocks for cell structures and form the enzymes that catalyze the cell's chemical reactions (Chapter 3), they regulate gene expression (Chapter 7), and they enable cells to communicate with each other (Chapter 15) and to move (Chapter 16). The properties and functions of a cell are determined largely by the proteins that it is able to make. With hindsight, it is hard to imagine what other type of instructions the genetic information could have contained.

Painstaking observations of cells and embryos in the late 19th century had led to the recognition that the hereditary information is carried on *chromosomes*, threadlike structures in the nucleus of a eucaryotic cell that become visible by light microscopy as the cell begins to divide (**Figure 4–1**). Later, as biochemical analysis became possible, chromosomes were found to consist of both deoxyribonucleic acid (DNA) and protein. For many decades, the DNA was thought to be merely a structural element. However, the other crucial advance made in the 1940s was the identification of DNA as the likely carrier of genetic information. This breakthrough in our understanding of cells came from studies

In This Chapter

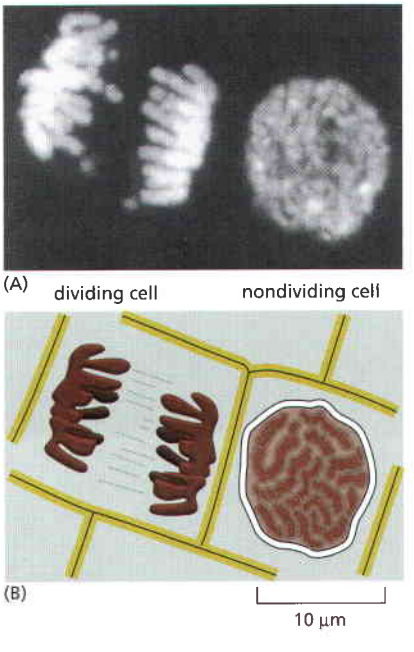


Figure 4–1 Chromosomes in cells. (A) Two adjacent plant cells photographed through a light microscope. The DNA has been stained with a fluorescent dye (DAPI) that binds to it. The DNA is present in chromosomes, which become visible as distinct structures in the light microscope only when they become compact, sausage-shaped structures in preparation for cell division, as shown on the left. The cell on the right, which is not dividing, contains identical chromosomes, but they cannot be clearly distinguished in the light microscope at this phase in the cell's life cycle, because they are in a more extended conformation. (B) Schematic diagram of the outlines of the two cells along with their chromosomes. (A, courtesy of Peter Shaw.)

of inheritance in bacteria (Figure 4–2). But as the 1950s began, both how proteins could be specified by instructions in the DNA and how this information might be copied for transmission from cell to cell seemed completely mysterious. The mystery was suddenly solved in 1953, when the structure of DNA was correctly predicted by James Watson and Francis Crick. As outlined in Chapter 1, the double-helical structure of DNA immediately solved the problem of how the information in this molecule might be copied, or *replicated*. It also provided the first clues as to how a molecule of DNA might use the sequence of its subunits to encode the instructions for making proteins. Today, the fact that DNA is the genetic material is so fundamental to biological thought that it is difficult to appreciate the enormous intellectual gap that was filled.

In this chapter we begin by describing the structure of DNA. We see how, despite its chemical simplicity, the structure and chemical properties of DNA make it ideally suited as the raw material of genes. We then consider how the many proteins in chromosomes arrange and package this DNA. The packing has to be done in an orderly fashion so that the chromosomes can be replicated and apportioned correctly between the two daughter cells at each cell division. It must also allow access to chromosomal DNA for the enzymes that repair it when it is damaged and for the specialized proteins that direct the expression of its many genes. We shall also see how the packaging of DNA differs along the length of each chromosome in eucaryotes, and how it can store a valuable record of the cell's developmental history.

In the past two decades, there has been a revolution in our ability to determine the exact sequence of subunits in DNA molecules. As a result, we now know the order of the 3 billion DNA subunits that provide the information for producing a human adult from a fertilized egg, as well as the DNA sequences of thousands of other organisms. Detailed analyses of these sequences have provided exciting insights into the process of evolution, and it is with this subject that the chapter ends.

This is the first of four chapters that deal with basic genetic mechanisms—the ways in which the cell maintains, replicates, expresses, and occasionally improves the genetic information carried in its DNA. This chapter presents a broad overview of DNA and how it is packaged into chromosomes. In the following chapter (Chapter 5) we discuss the mechanisms by which the cell accurately replicates and repairs DNA; we also describe how DNA sequences can be

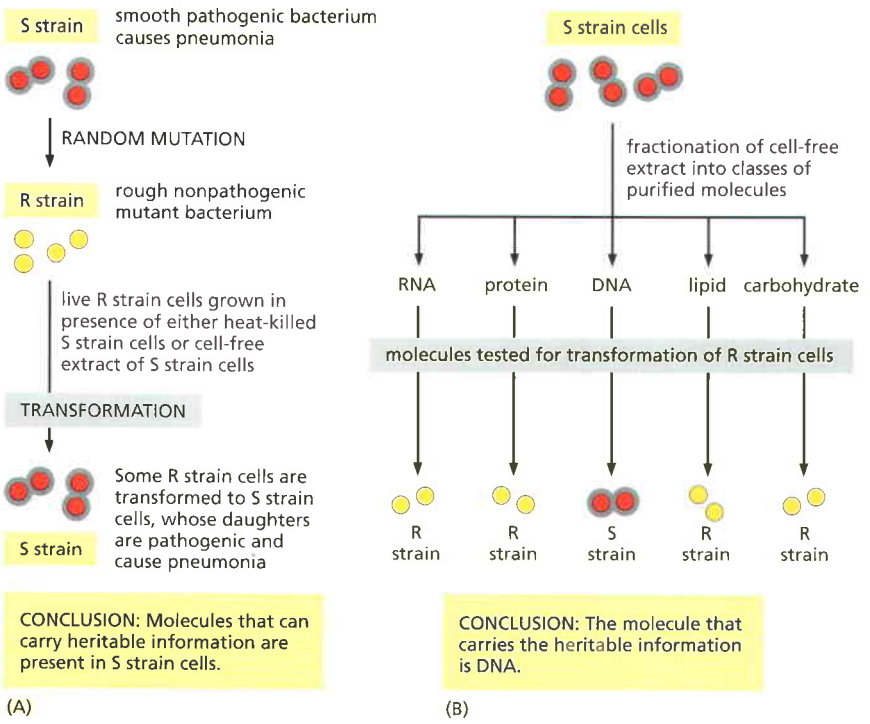


Figure 4–2 The first experimental demonstration that DNA is the genetic material. These experiments, carried out in the 1940s, showed that adding purified DNA to a bacterium changed its properties and that this change was faithfully passed on to subsequent generations. Two closely related strains of the bacterium *Streptococcus pneumoniae* differ from each other in both their appearance under the microscope and their pathogenicity. One strain appears smooth (S) and causes death when injected into mice, and the other appears rough (R) and is nonlethal. (A) An initial experiment shows that a substance present in the S strain can change (or transform) the R strain into the S strain and that this change is inherited by subsequent generations of bacteria. (B) This experiment, in which the R strain has been incubated with various classes of biological molecules purified from the S strain, identifies the substance as DNA.

rearranged through the process of genetic recombination. Gene expression—the process through which the information encoded in DNA is interpreted by the cell to guide the synthesis of proteins—is the main topic of Chapter 6. In Chapter 7, we describe how this gene expression is controlled by the cell to ensure that each of the many thousands of proteins and RNA molecules encrypted in its DNA are manufactured only at the proper time and place in the life of the cell.

THE STRUCTURE AND FUNCTION OF DNA

Biologists in the 1940s had difficulty in conceiving how DNA could be the genetic material because of the apparent simplicity of its chemistry. DNA was known to be a long polymer composed of only four types of subunits, which resemble one another chemically. Early in the 1950s, DNA was examined by x-ray diffraction analysis, a technique for determining the three-dimensional atomic structure of a molecule (discussed in Chapter 8). The early x-ray diffraction results indicated that DNA was composed of two strands of the polymer wound into a helix. The observation that DNA was double-stranded was of crucial significance and provided one of the major clues that led to the Watson–Crick model for DNA structure. But only when this model was proposed in 1953 did DNA's potential for replication and information encoding become apparent. In this section we examine the structure of the DNA molecule and explain in general terms how it is able to store hereditary information.

A DNA Molecule Consists of Two Complementary Chains of Nucleotides

A **deoxyribonucleic acid** (**DNA**) molecule consists of two long polynucleotide chains composed of four types of nucleotide subunits. Each of these chains is known as a *DNA chain*, or a *DNA strand*. *Hydrogen bonds* between the base portions of the nucleotides hold the two chains together (**Figure 4–3**). As we saw in Chapter 2 (Panel 2–6, pp. 116–117), nucleotides are composed of a five-carbon sugar to which are attached one or more phosphate groups and a nitrogen-containing base. In the case of the nucleotides in DNA, the sugar is deoxyribose attached to a single phosphate group (hence the name deoxyribonucleic acid), and the base may be either *adenine (A)*, *cytosine (C)*, *guanine (G)*, or *thymine (T)*. The nucleotides are covalently linked together in a chain through the sugars and phosphates, which thus form a "backbone" of alternating sugar–phosphate–sugar–phosphate. Because only the base differs in each of the four types of subunits, each polynucleotide chain in DNA is analogous to a necklace (the backbone) strung with four types of beads (the four bases A, C, G, and T). These same symbols (A, C, G, and T) are also commonly used to denote the four different nucleotides—that is, the bases with their attached sugar and phosphate groups.

The way in which the nucleotide subunits are linked together gives a DNA strand a chemical polarity. If we think of each sugar as a block with a protruding knob (the 5′ phosphate) on one side and a hole (the 3′ hydroxyl) on the other (see Figure 4–3), each completed chain, formed by interlocking knobs with holes, will have all of its subunits lined up in the same orientation. Moreover, the two ends of the chain will be easily distinguishable, as one has a hole (the 3′ hydroxyl) and the other a knob (the 5′ phosphate) at its terminus. This polarity in a DNA chain is indicated by referring to one end as the *3′ end* and the other as the *5′ end*.

The three-dimensional structure of DNA—the **double helix**—arises from the chemical and structural features of its two polynucleotide chains. Because these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside (see Figure 4–3). In each case, a bulkier two-ring base (a purine; see Panel 2–6, pp. 116–117) is paired with a single-ring base (a pyrimidine); A always pairs with T, and G with C (**Figure**

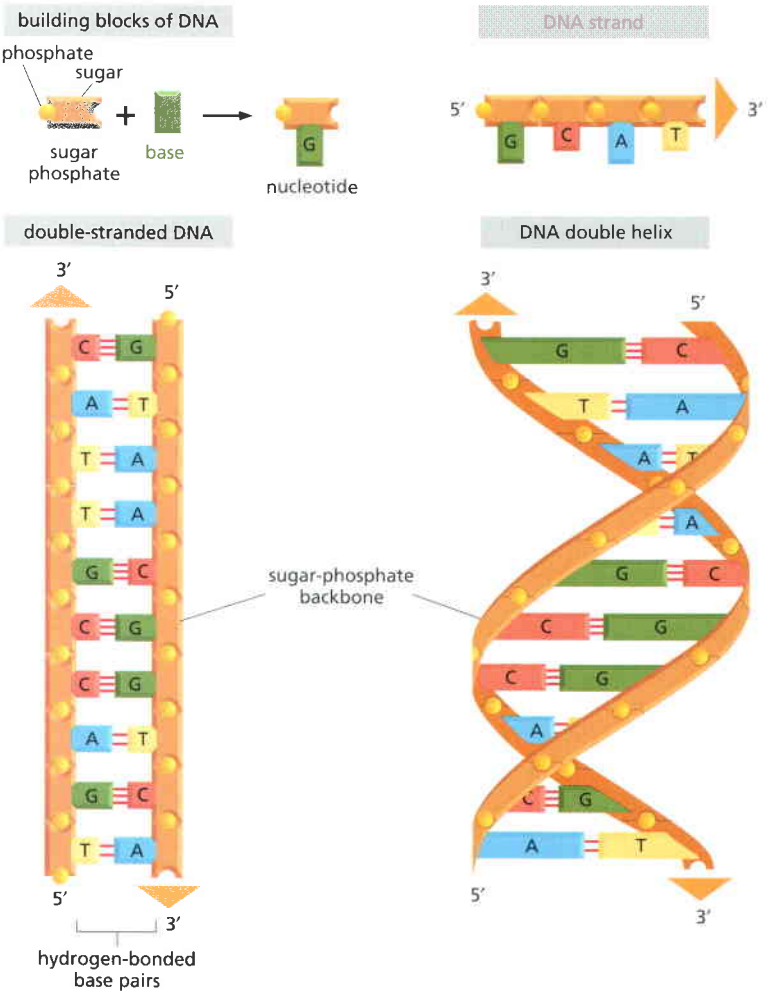


Figure 4–3 DNA and its building blocks. <CAGA> DNA is made of four types of nucleotides, which are linked covalently into a polynucleotide chain (a DNA strand) with a sugar-phosphate backbone from which the bases (A, C, G, and T) extend. A DNA molecule is composed of two DNA strands held together by hydrogen bonds between the paired bases. The *arrowheads* at the ends of the DNA strands indicate the polarities of the two strands, which run antiparallel to each other in the DNA molecule. In the diagram at the bottom left of the figure, the DNA molecule is shown straightened out; in reality, it is twisted into a double helix, as shown on the right. For details, see Figure 4–5.

4–4). This *complementary base-pairing* enables the **base pairs** to be packed in the energetically most favorable arrangement in the interior of the double helix. In this arrangement, each base pair is of similar width, thus holding the sugar-phosphate backbones an equal distance apart along the DNA molecule. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones

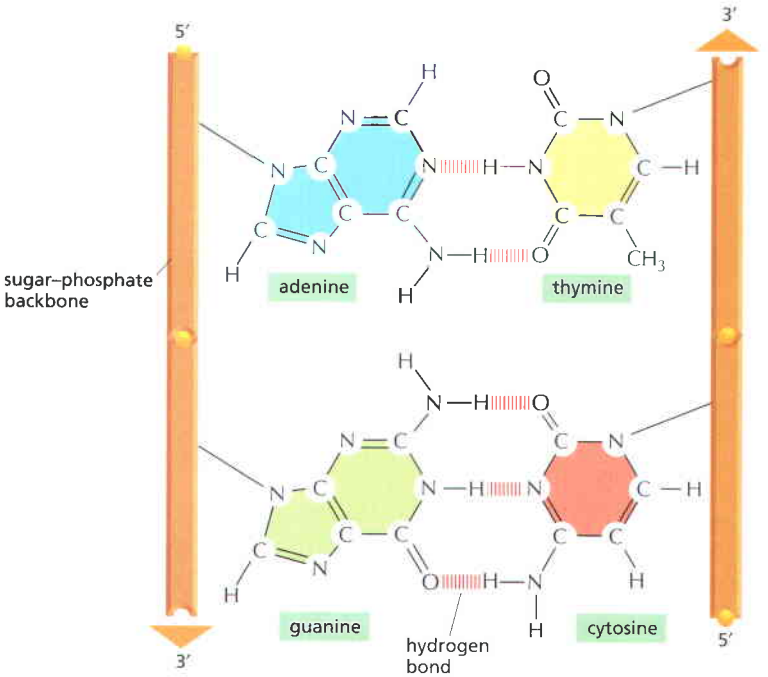


Figure 4–4 Complementary base pairs in the DNA double helix. The shapes and chemical structure of the bases allow hydrogen bonds to form efficiently only between A and T and between G and C, where atoms that are able to form hydrogen bonds (see Panel 2–3, pp. 110–111) can be brought close together without distorting the double helix. As indicated, two hydrogen bonds form between A and T, while three form between G and C. The bases can pair in this way only if the two polynucleotide chains that contain them are antiparallel to each other.

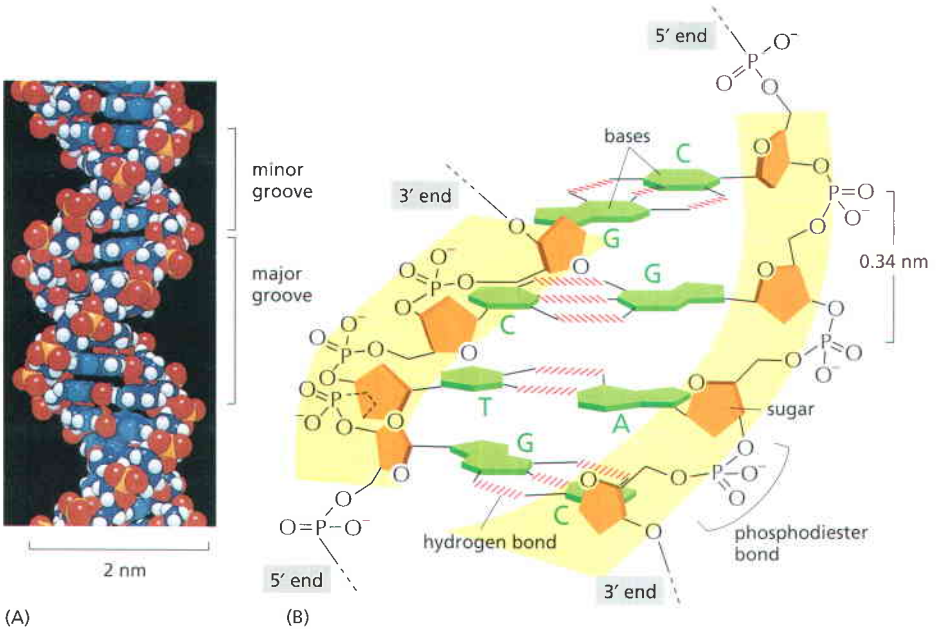


Figure 4–5 The DNA double helix. (A) A space-filling model of 1.5 turns of the DNA double helix. Each turn of DNA is made up of 10.4 nucleotide pairs, and the center-to-center distance between adjacent nucleotide pairs is 3.4 nm. The coiling of the two strands around each other creates two grooves in the double helix: the wider groove is called the major groove, and the smaller the minor groove. (B) A short section of the double helix viewed from its side, showing four base pairs. The nucleotides are linked together covalently by phosphodiester bonds that join the 3′-hydroxyl (–OH) group of one sugar to the 5′-hydroxyl group of the next sugar. Thus, each polynucleotide strand has a chemical polarity; that is, its two ends are chemically different. The 5′ end of the DNA polymer is by convention often illustrated carrying a phosphate group, while the 3′-end is shown with a hydroxyl.

wind around each other to form a double helix, with one complete turn every ten base pairs (**Figure 4–5**).

The members of each base pair can fit together within the double helix only if the two strands of the helix are **antiparallel**—that is, only if the polarity of one strand is oriented opposite to that of the other strand (see Figures 4–3 and 4–4). A consequence of these base-pairing requirements is that each strand of a DNA molecule contains a sequence of nucleotides that is exactly **complementary** to the nucleotide sequence of its partner strand.

The Structure of DNA Provides a Mechanism for Heredity

Genes carry biological information that must be copied accurately for transmission to the next generation each time a cell divides to form two daughter cells. Two central biological questions arise from these requirements: how can the information for specifying an organism be carried in chemical form, and how is it accurately copied? The discovery of the structure of the DNA double helix was a landmark in twentieth-century biology because it immediately suggested answers to both questions, thereby providing a molecular explanation for the problem of heredity. We discuss these answers briefly in this section, and we shall examine them in much more detail in subsequent chapters.

DNA encodes information through the order, or sequence, of the nucleotides along each strand. Each base—A, C, T, or G—can be considered as a letter in a four-letter alphabet that spells out biological messages in the chemical structure of the DNA. As we saw in Chapter 1, organisms differ from one another because their respective DNA molecules have different nucleotide sequences and, consequently, carry different biological messages. But how is the nucleotide alphabet used to make messages, and what do they spell out?

As discussed above, it was known well before the structure of DNA was determined that genes contain the instructions for producing proteins. The DNA messages must therefore somehow encode proteins (**Figure 4–6**). This relationship immediately makes the problem easier to understand. As discussed in Chapter 3, the properties of a protein, which are responsible for its biological function, are determined by its three-dimensional structure. This structure is determined in turn by the linear sequence of the amino acids of which it is composed. The linear sequence of nucleotides in a gene must therefore somehow spell out the linear sequence of amino acids in a protein. The exact correspondence between the four-letter nucleotide alphabet of DNA and the twenty-letter amino acid alphabet of proteins—the genetic code—is not obvious from the DNA structure, and it took over a decade after the discovery of the double helix

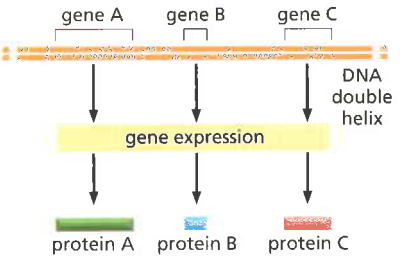


Figure 4–6 The relationship between genetic information carried in DNA and proteins (discussed in Chapter 1).

before it was worked out. In Chapter 6 we will describe this code in detail in the course of elaborating the process, known as *gene expression*, through which a cell converts the nucleotide sequence of a gene first into the nucleotide sequence of an RNA molecule, and then into the amino acid sequence of a protein.

The complete set of information in an organism's DNA is called its **genome**, and it carries the information for all the proteins and RNA molecules that the organism will ever synthesize. (The term genome is also used to describe the DNA that carries this information.) The amount of information contained in genomes is staggering: for example, a typical human diploid cell contains 2 meters of DNA double helix. Written out in the four-letter nucleotide alphabet, the nucleotide sequence of a very small human gene occupies a quarter of a page of text (**Figure 4–7**), while the complete sequence of nucleotides in the human genome would fill more than a thousand books the size of this one. In addition to other critical information, it carries the instructions for roughly 24,000 distinct proteins.

At each cell division, the cell must copy its genome to pass it to both daughter cells. The discovery of the structure of DNA also revealed the principle that makes this copying possible: because each strand of DNA contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of its partner strand, each strand can act as a **template**, or mold, for the synthesis of a new complementary strand. In other words, if we designate the two DNA strands as S and S′, strand S can serve as a template for making a new strand S′, while strand S′ can serve as a template for making a new strand S (**Figure 4–8**). Thus, the genetic information in DNA can be accurately copied by the beautifully simple process in which strand S separates from strand S′, and each separated strand then serves as a template for the production of a new complementary partner strand that is identical to its former partner.

The ability of each strand of a DNA molecule to act as a template for producing a complementary strand enables a cell to copy, or *replicate*, its genome before passing it on to its descendants. In the next chapter we shall describe the elegant machinery the cell uses to perform this enormous task.

In Eucaryotes, DNA Is Enclosed in a Cell Nucleus

As described in Chapter 1, nearly all the DNA in a eucaryotic cell is sequestered in a nucleus, which in many cells occupies about 10% of the total cell volume. This compartment is delimited by a *nuclear envelope* formed by two concentric lipid bilayer membranes (**Figure 4–9**). These membranes are punctured at intervals by large nuclear pores, which transport molecules between the nucleus and the cytosol. The nuclear envelope is directly connected to the extensive membranes of the endoplasmic reticulum, which extend out from it into the cytoplasm. And it is mechanically supported by a network of intermediate filaments called the *nuclear lamina*, which forms a thin sheetlike meshwork just beneath the inner nuclear membrane (see Figure 4–9B).

The nuclear envelope allows the many proteins that act on DNA to be concentrated where they are needed in the cell, and, as we see in subsequent

CCCTGTGGAGCCACACCCTAGGGTTGGCCA
ATCTACTCCCAGGAGCAGGGAGGGCAGGAG
CCAGGGCTGGGCATAAAAGTCAGGGCAGAG
CCATCTATTGCTTACATTTGCTTCTGACAC
AACTGTGTTCACTAGCAACTCAAACAGACA
CCATGGTGCACCTGACTCCTGAGGAGAAGT
CTGCCGTTACTGCCCTGTGGGGCAAGGTGA
ACGTGGATGAAGTTGGTGGTGAGGCCCTGG
GCAGGTTGGTATCAAGGTTACAAGACAGGT
TTAAGGAGACCAATAGAAACTGGGCATGTG
GAGACAGAGAAGACTCTTGGGTTTCTGATA
GGCACTGACTCTCTCTGCCTATTGGTCTAT
TTTCCCACCCTTAGGCTGCTGGTGGTCTAC
CCTTGGACCCAGAGGTTCTTTGAGTCCTTT
GGGGATCTGTCCACTCCTGATGCTGTTATG
GGCAACCCTAAGGTGAAGGCTCATGGCAAG
AAAGTGCTCGGTGCCTTTAGTGATGGCCTG
GCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAG
CTGCACGTGGATCCTGAGAACTTCAGGGTG
AGTCTATGGGACCCTTGATGTTTTCTTTCC
CCTTCTTTTCTATGGTTAAGTTCATGTCAT
AGGAAGGGGAGAAGTAACAGGGTACAGTTT
AGAATGGGAAACAGACGAATGATTGCATCA
GTGTGGAAGTCTCAGGATCGTTTTAGTTTC
TTTTATTTGCTGTTCATAACAATTGTTTTC
TTTTGTTTAATTCTTGCTTTCTTTTTTTTT
CTTCTCCGCAATTTTTACTATTATACTTAA
TGCCTTAACATTGTGTATAACAAAAGGAAA
TATCTCTGAGATACATTAAGTAACTTAAAA
AAAAACTTTACACAGTCTGCCTAGTACATT
ACTATTTGGAATATATGTGTGCTTATTTGC
ATATTCATAATCTCCCTACTTTATTTTCTT
TTATTTTTAATTGATACATAATCATTATAC
ATATTTATGGGTTAAAGTGTAATGTTTTAA
TATGTGTACACATATTGACCAAATCAGGGT
AATTTTGCATTTGTAATTTTAAAAAATGCT
TTCTTCTTTTAATATACTTTTTTGTTTATC
TTATTTCTAATACTTTCCCTAATCTCTTTC
TTTCAGGGCAATAATGATACAATGTATCAT
GCCTCTTTGCACCATTCTAAAGAATAACAG
TGATAATTTCTGGGTTAAGGCAATAGCAAT
ATTTCTGCATATAAATATTTCTGCATATAA
ATTGTAACTGATGTAAGAGGTTTCATATTG
CTAATAGCAGCTACAATCCAGCTACCATTC
TGCTTTTATTTTATGGTTGGGATAAGGCTG
GATTATTCTGAGTCCAAGCTAGGCCCTTTT
GCTAATCATGTTCATACCTCTTATCTTCCT
CCCACAGCTCCTGGGCAACGTGCTGGTCTG
TGTGCTGGCCCATCACTTTGGCAAAGAATT
CACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGC
CCACAAGTATCACTAAGCTCGCTTTCTTGC
TGTCCAATTTCTATTAAAGGTTCCTTTGTT
CCCTAAGTCCAACTACTAAACTGGGGGATA
TTATGAAGGGCCTTGAGCATCTGGATTCTG
CCTAATAAAAAACATTTATTTTCATTGCAA
TGATGTATTTAAATTATTTCTGAATATTTT
ACTAAAAAGGGAATGTGGGAGGTCAGTGCA
TTTAAAACATAAAGAAATGATGAGCTGTTC
AAACCTTGGGAAAATACACTATATCTTAAA
CTCCATGAAAGAAGGTGAGGCTGCAACCAG
CTAATGCACATTGGCAACAGCCCCTGATGC
CTATGCCTTATTCATCCCTCAGAAAAGGAT
TCTTGTAGAGGCTTGATTTGCAGGTTAAAG
TTTTGCTATGCTGTATTTTACATTACTTAT
TGTTTTAGCTGTCCTCATGAATGTCTTTTC

Figure 4–7 **The nucleotide sequence of the human β-globin gene.** By convention, a nucleotide sequence is written from its 5′ end to its 3′ end, and it should be read from left to right in successive lines down the page as though it were normal English text. This gene carries the information for the amino acid sequence of one of the two types of subunits of the hemoglobin molecule, the protein that carries oxygen in the blood. A different gene, the α-globin gene, carries the information for the other type of hemoglobin subunit (a hemoglobin molecule has four subunits, two of each type). Only one of the two strands of the DNA double helix containing the β-globin gene is shown; the other strand has the exact complementary sequence. The DNA sequences highlighted in *yellow* show the three regions of the gene that specify the amino acid sequence for the β-globin protein. We shall see in Chapter 6 how the cell splices these three sequences together at the level of messenger RNA in order to synthesize a full-length β-globin protein.

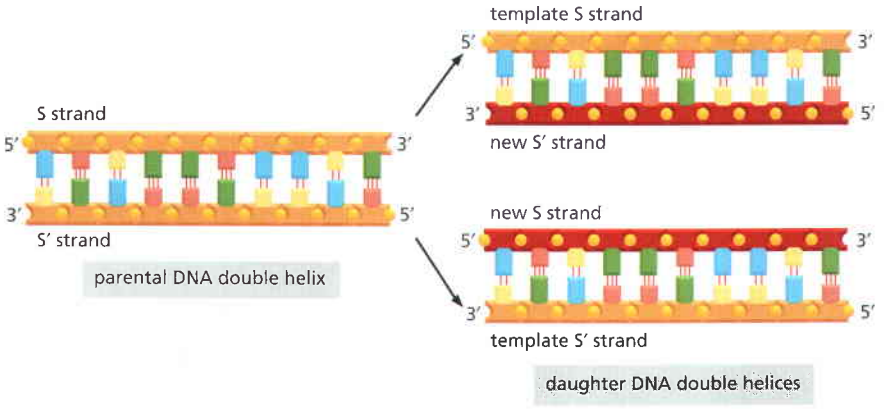


Figure 4–8 DNA as a template for its own duplication. As the nucleotide A successfully pairs only with T, and G with C, each strand of DNA can act as a template to specify the sequence of nucleotides in its complementary strand. In this way, double-helical DNA can be copied precisely, with each parental DNA helix producing two identical daughter DNA helices.

chapters, it also keeps nuclear and cytosolic enzymes separate, a feature that is crucial for the proper functioning of eucaryotic cells. Compartmentalization, of which the nucleus is an example, is an important principle of biology; it serves to establish an environment in which biochemical reactions are facilitated by the high concentration of both substrates and the enzymes that act on them. Compartmentalization also prevents enzymes needed in one part of the cell from interfering with the orderly biochemical pathways in another.

Summary

Genetic information is carried in the linear sequence of nucleotides in DNA. Each molecule of DNA is a double helix formed from two complementary strands of nucleotides held together by hydrogen bonds between G-C and A-T base pairs. Duplication of the genetic information occurs by the use of one DNA strand as a template for the formation of a complementary strand. The genetic information stored in an organism's DNA contains the instructions for all the proteins the organism will ever synthesize and is said to comprise its genome. In eucaryotes, DNA is contained in the cell nucleus, a large membrane-bound compartment.

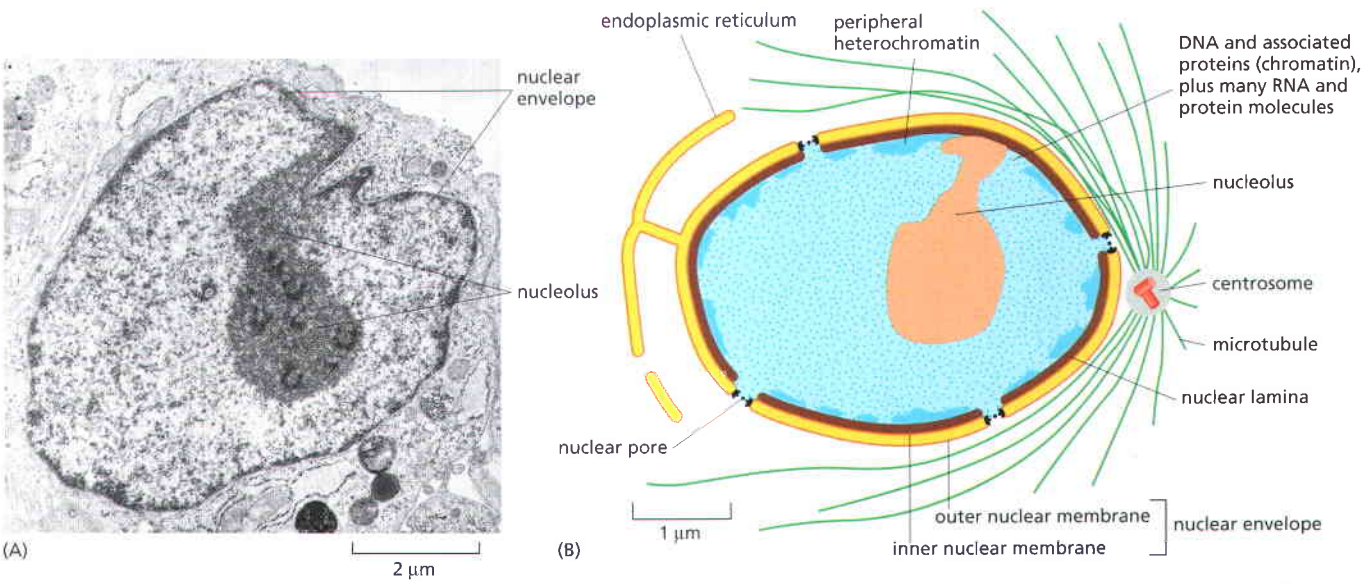


Figure 4–9 A cross-sectional view of a typical cell nucleus. (A) Electron micrograph of a thin section through the nucleus of a human fibroblast. (B) Schematic drawing, showing that the nuclear envelope consists of two membranes, the outer one being continuous with the endoplasmic reticulum membrane (see also Figure 12–8). The space inside the endoplasmic reticulum (the ER lumen) is colored *yellow;* it is continuous with the space between the two nuclear membranes. The lipid bilayers of the inner and outer nuclear membranes are connected at each nuclear pore. A sheet-like network of intermediate filaments *(brown)* inside the nucleus provides mechanical support for the nuclear envelope, forming a special supporting structure called the nuclear lamina (for details, see Chapter 12). The heterochromatin near the lamina contains specially condensed regions of DNA that will be discussed later.

CHROMOSOMAL DNA AND ITS PACKAGING IN THE CHROMATIN FIBER

The most important function of DNA is to carry genes, the information that specifies all the proteins and RNA molecules that make up an organism—including information about when, in what types of cells, and in what quantity each protein is to be made. The genomes of eucaryotes are divided up into chromosomes, and in this section we see how genes are typically arranged on each chromosome. In addition, we describe the specialized DNA sequences that are required for a chromosome to be accurately duplicated and passed on from one generation to the next.

We also confront the serious challenge of DNA packaging. If the double helices comprising all 46 chromosomes in a human cell could be laid end-to-end, they would reach approximately 2 meters; yet the nucleus, which contains the DNA, is only about 6 μm in diameter. This is geometrically equivalent to packing 40 km (24 miles) of extremely fine thread into a tennis ball! The complex task of packaging DNA is accomplished by specialized proteins that bind to and fold the DNA, generating a series of coils and loops that provide increasingly higher levels of organization, preventing the DNA from becoming an unmanageable tangle. Amazingly, although the DNA is very tightly folded, it is compacted in a way that keeps it available to the many enzymes in the cell that replicate it, repair it, and use its genes to produce RNA molecules and proteins.

Eucaryotic DNA Is Packaged into a Set of Chromosomes

In eucaryotes, the DNA in the nucleus is divided between a set of different **chromosomes**. For example, the human genome—approximately 3.2×10^9 nucleotides—is distributed over 24 different chromosomes. Each chromosome consists of a single, enormously long linear DNA molecule associated with proteins that fold and pack the fine DNA thread into a more compact structure. The complex of DNA and protein is called *chromatin* (from the Greek *chroma*, "color," because of its staining properties). In addition to the proteins involved in packaging the DNA, chromosomes are also associated with many proteins and RNA molecules required for the processes of gene expression, DNA replication, and DNA repair.

Bacteria carry their genes on a single DNA molecule, which is often circular (see Figure 1–29). This DNA is associated with proteins that package and condense the DNA, but they are different from the proteins that perform these functions in eucaryotes. Although often called the bacterial "chromosome," it does not have the same structure as eucaryotic chromosomes, and less is known about how the bacterial DNA is packaged. Therefore, our discussion of chromosome structure will focus almost entirely on eucaryotic chromosomes.

With the exception of the germ cells (discussed in Chapter 21) and a few highly specialized cell types that cannot multiply and lack DNA altogether (for example, red blood cells), each human cell contains two copies of each chromosome, one inherited from the mother and one from the father. The maternal and paternal chromosomes of a pair are called **homologous chromosomes** (**homologs**). The only nonhomologous chromosome pairs are the sex chromosomes in males, where a *Y chromosome* is inherited from the father and an *X chromosome* from the mother. Thus, each human cell contains a total of 46 chromosomes—22 pairs common to both males and females, plus two so-called sex chromosomes (X and Y in males, two Xs in females). *DNA hybridization* is a technique in which a labeled nucleic acid strand serves as a "probe" that localizes a complementary strand, as will be described in detail in Chapter 8. This technique can be used to distinguish these human chromosomes by "painting" each one a different color (**Figure 4–10**). Chromosome painting is typically done at the stage in the cell cycle called mitosis, when chromosomes are especially compacted and easy to visualize (see below).

Another more traditional way to distinguish one chromosome from another

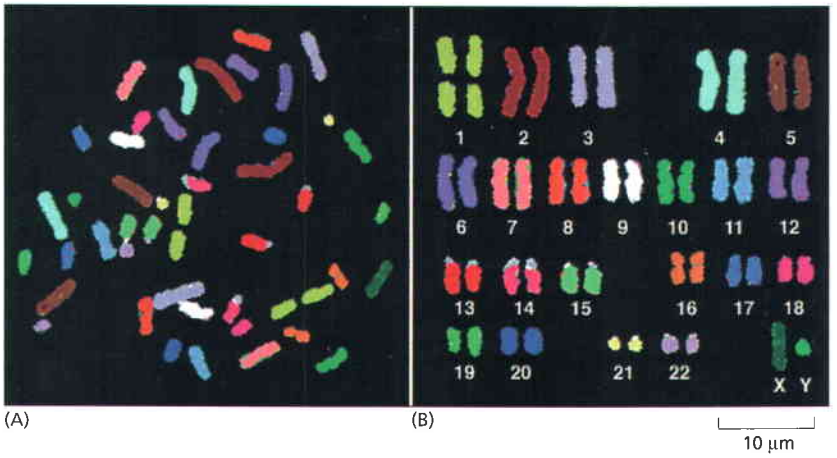


Figure 4–10 The complete set of human chromosomes. These chromosomes, from a male, were isolated from a cell undergoing nuclear division (mitosis) and are therefore highly compacted. Each chromosome has been "painted" a different color to permit its unambiguous identification under the light microscope. Chromosome painting is performed by exposing the chromosomes to a collection of human DNA molecules that have been coupled to a combination of fluorescent dyes. For example, DNA molecules derived from chromosome 1 are labeled with one specific dye combination, those from chromosome 2 with another, and so on. Because the labeled DNA can form base pairs, or hybridize, only to the chromosome from which it was derived (discussed in Chapter 8), each chromosome is differently labeled. For such experiments, the chromosomes are subjected to treatments that separate the double-helical DNA into individual strands, designed to permit base-pairing with the single-stranded labeled DNA while keeping the chromosome structure relatively intact. (A) The chromosomes visualized as they originally spilled from the lysed cell. (B) The same chromosomes artificially lined up in their numerical order. This arrangement of the full chromosome set is called a karyotype. (From E. Schröck et al., *Science* 273:494–497, 1996. With permission from AAAS.)

along each mitotic chromosome (**Figure 4–11**). The structural bases for these banding patterns are not well understood. Nevertheless, the pattern of bands on each type of chromosome is unique, and it is these patterns that initially allowed each human chromosome to be identified and numbered.

The display of the 46 human chromosomes at mitosis is called the human **karyotype**. If parts of chromosomes are lost or are switched between chromosomes, these changes can be detected by changes in the banding patterns or by changes in the pattern of chromosome painting (**Figure 4–12**). Cytogeneticists use these alterations to detect chromosome abnormalities that are associated with inherited defects, as well as to characterize cancers that are associated with specific chromosome rearrangements in somatic cells (discussed in Chapter 20).

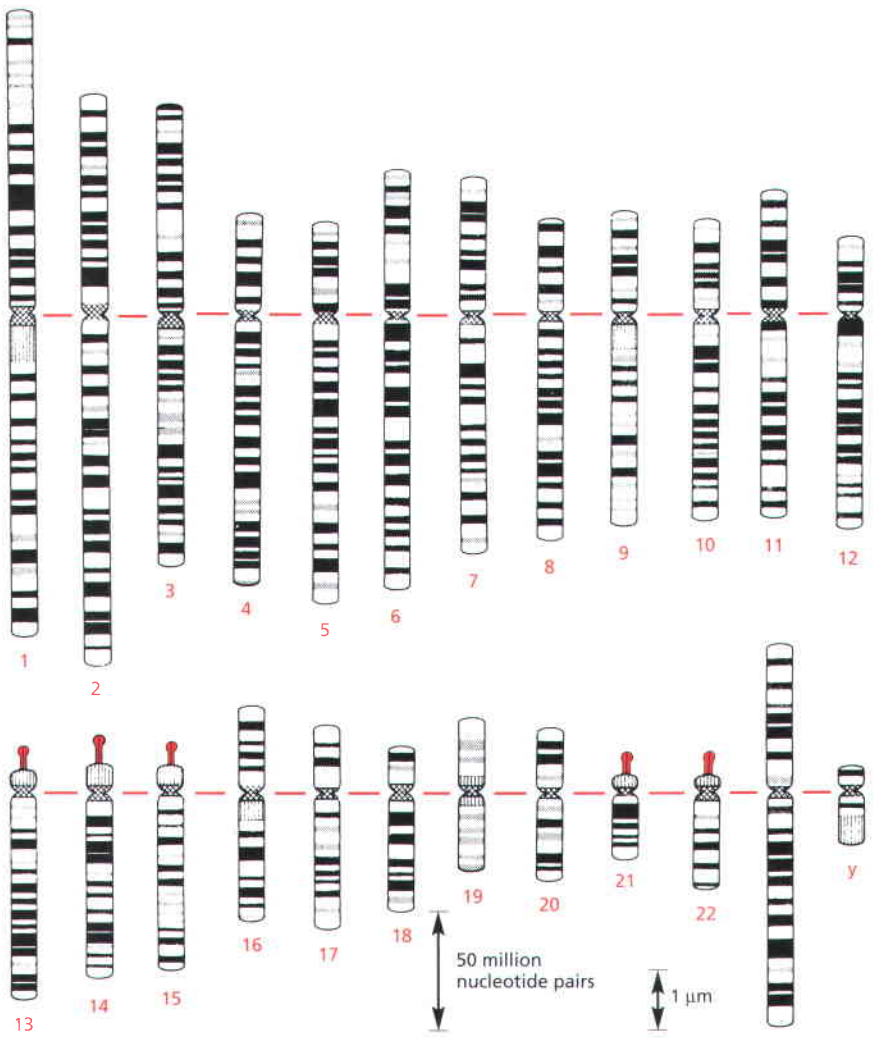


Figure 4–11 The banding patterns of human chromosomes. Chromosomes 1–22 are numbered in approximate order of size. A typical human somatic (non-germ-line) cell contains two of each of these chromosomes, plus two sex chromosomes—two X chromosomes in a female, one X and one Y chromosome in a male. The chromosomes used to make these maps were stained at an early stage in mitosis, when the chromosomes are incompletely compacted. The *horizontal red line* represents the position of the centromere (see Figure 4–21), which appears as a constriction on mitotic chromosomes. The red knobs on chromosomes 13, 14, 15, 21, and 22 indicate the positions of genes that code for the large ribosomal RNAs (discussed in Chapter 6). These patterns are obtained by staining chromosomes with Giemsa stain, and they can be observed under the light microscope. (For micrographs, see Figure 21–18; adapted from U. Franke, *Cytogenet. Cell Genet.* 31:24–32, 1981. With

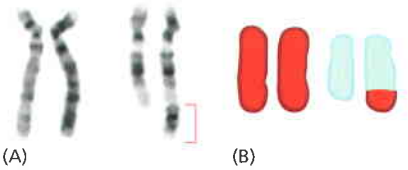


Figure 4–12 An aberrant human chromosome. (A) Two pairs of chromosomes, stained with Giemsa (see Figure 4–11), from a patient with ataxia, a disease characterized by progressive deterioration of motor skills. The patient has a normal pair of chromosome 4s *(left-hand pair),* but one normal chromosome 12 and one aberrant chromosome 12, as seen by its greater length *(right-hand pair).* The additional material contained on the aberrant chromosome 12 (*red bracket*) was deduced, from its pattern of bands, as a copy of part of chromosome 4 that had become attached to chromosome 12 through an abnormal recombination event, called a chromosomal translocation. (B) Drawing of the same two chromosome pairs, "painted" *red* for chromosome 4 DNA and *blue* for chromosome 12 DNA. The two techniques give rise to the same conclusion regarding the nature of the aberrant chromosome 12, but chromosome painting provides better resolution, allowing the clear identification of even short pieces of chromosomes that have become translocated. However, Giemsa staining is easier to perform. (Adapted from E. Schröck et al., *Science* 273:494–497, 1996. With permission from AAAS.)

Chromosomes Contain Long Strings of Genes

Chromosomes carry genes—the functional units of heredity. A gene is usually defined as a segment of DNA that contains the instructions for making a particular protein (or a set of closely related proteins). Although this definition holds for the majority of genes, several percent of genes produce an RNA molecule, instead of a protein, as their final product. Like proteins, these RNA molecules perform a diverse set of structural and catalytic functions in the cell, and we discuss them in detail in subsequent chapters.

As might be expected, some correlation exists between the complexity of an organism and the number of genes in its genome (see Table 1–1, p. 18). For example, some simple bacteria have only 500 genes, compared to about 25,000 for humans. Bacteria and some single-celled eucaryotes, such as yeast, have especially concise genomes; the complete nucleotide sequence of their genomes reveals that the DNA molecules that make up their chromosomes are little more than strings of closely packed genes (**Figure 4–13**). However, chromosomes from many eucaryotes (including humans) contain, in addition to genes, a large excess of interspersed DNA that does not seem to carry critical information. Sometimes called "junk DNA" to signify that its usefulness to the cell has not been demonstrated, the particular nucleotide sequence of most of this DNA may not be important. However, some of this DNA is crucial for the proper expression of certain genes, as we discuss elsewhere.

Because of differences in the amount of DNA interspersed between genes, genome sizes can vary widely (see Figure 1–37). For example, the human genome is 200 times larger than that of the yeast *S. cerevisiae,* but 30 times smaller than that of some plants and amphibians and 200 times smaller than that of a species of amoeba. Moreover, because of differences in the amount of excess DNA, the genomes of similar organisms (bony fish, for example) can vary several hundredfold in their DNA content, even though they contain roughly the same number of genes. Whatever the excess DNA may do, it seems clear that it is not a great handicap for a eucaryotic cell to carry a large amount of it.

How the genome is divided into chromosomes also differs from one eucaryotic species to the next. For example, compared with 46 for humans, somatic cells from a species of small deer contain only 6 chromosomes, while those from a species of carp contain over 100. Even closely related species with similar genome sizes can have very different numbers and sizes of chromosomes (**Figure 4–14**). Thus, there is no simple relationship between chromosome number,

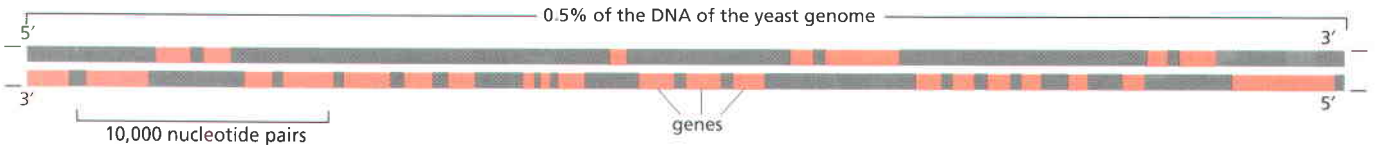


Figure 4–13 The arrangement of genes in the genome of *S. cerevisiae*. *S. cerevisiae* is a budding yeast widely used for brewing and baking. The genome of this yeast cell is distributed over 16 chromosomes. A small region of one chromosome has been arbitrarily selected to show the high density of genes characteristic of this species. As indicated by the *light red shading,* some genes are transcribed from the lower strand, while others are transcribed from the upper strand. There are about 6300 genes in the complete genome, which contains somewhat more than 12 million nucleotide pairs. (For the closely packed genes of a bacterium whose genome is 4.6 million nucleotides long, see Figure 1–29).

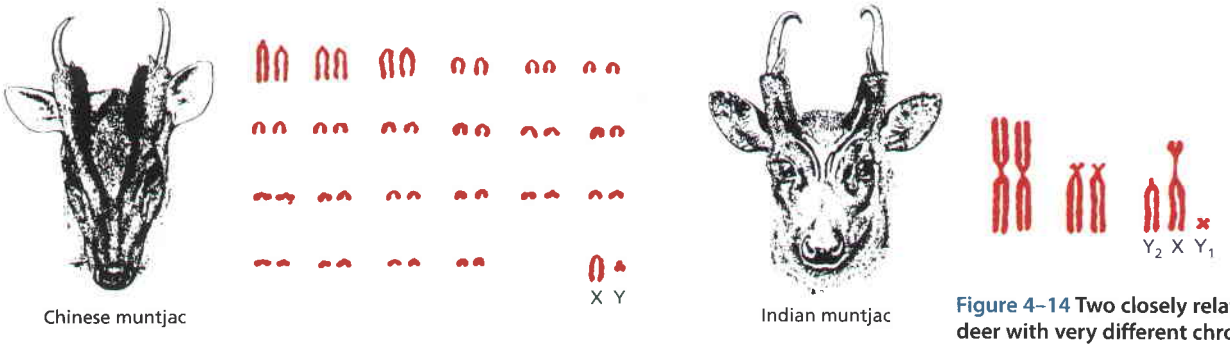


Figure 4–14 Two closely related species of deer with very different chromosome numbers. In the evolution of the Indian muntjac, initially separate chromosomes fused, without having a major effect on the animal. These two species contain a similar number of genes. (Adapted from M.W. Strickberger, Evolution, 3rd ed. Sudbury, MA: Jones & Bartlett Publishers, 2000.)

species complexity, and total genome size. Rather, the genomes and chromosomes of modern-day species have each been shaped by a unique history of seemingly random genetic events, acted on by selection pressures over long evolutionary times.

The Nucleotide Sequence of the Human Genome Shows How Our Genes Are Arranged

In Chapter 1 we discussed, in general terms, how the information in DNA is read out and used, through RNA intermediates, to make proteins (see Figure 1–4). In 1999, it became possible for the first time to see exactly how genes are arranged along an entire vertebrate chromosome (**Figure 4–15**). Today, with the publication of the "first draft" of the entire human genome in 2001 and the "finished DNA sequence" in 2004, the genetic information in all human chromosomes is available. The sheer quantity of information provided by the Human Genome Project is staggering (**Figure 4–16** and **Table 4–1**). At its peak, the Project generated raw nucleotide sequences at a rate of 1000 nucleotides per second around the clock. It will be many decades before this information is fully analyzed, but it has already stimulated new experiments that have had major effects on the content of every chapter in this book.

The first striking feature of the human genome is how little of it (only a few percent) codes for proteins (**Figure 4–17**). Much of the remaining chromosomal

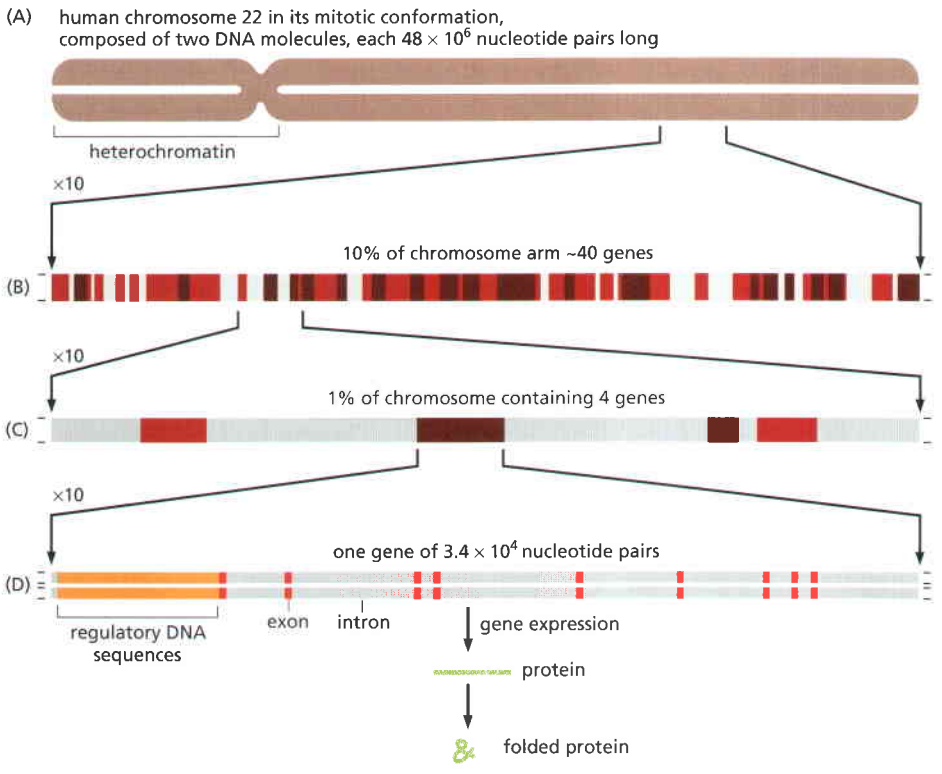


Figure 4–15 The organization of genes on a human chromosome. (A) Chromosome 22, one of the smallest human chromosomes, contains 48×10^6 nucleotide pairs and makes up approximately 1.5% of the entire human genome. Most of the left arm of chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin), which is discussed later in this chapter. (B) A tenfold expansion of a portion of chromosome 22, with about 40 genes indicated. Those in *dark brown* are known genes and those in *red* are predicted genes. (C) An expanded portion of (B) shows the entire length of several genes. (D) The intron–exon arrangement of a typical gene is shown after a further tenfold expansion. Each exon *(red)* codes for a portion of the protein, while the DNA sequence of the introns *(gray)* is relatively unimportant, as discussed in detail in Chapter 6.

The human genome (3.2×10^9 nucleotide pairs) is the totality of genetic information belonging to our species. Almost all of this genome is distributed over the 22 autosomes and 2 sex chromosomes (see Figures 4–10 and 4–11) found within the nucleus. A minute fraction of the human genome (16,569 nucleotide pairs—in multiple copies per cell) is found in the mitochondria (introduced in Chapter 1, and discussed in detail in Chapter 14). The term *human genome sequence* refers to the complete nucleotide sequence of DNA in the 24 nuclear chromosomes and the mitochondria. Being diploid, a human somatic cell nucleus contains roughly twice the haploid amount of DNA, or 6.4×10^9 nucleotide pairs when not duplicating its chromosomes in preparation for division. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

DNA is made up of short, mobile pieces of DNA that have gradually inserted themselves in the chromosome over evolutionary time. We discuss these *transposable elements* in detail in later chapters.

A second notable feature of the human genome is the large average gene size of 27,000 nucleotide pairs. As discussed above, a typical gene carries in its linear sequence of nucleotides the information for the linear sequence of the amino acids of a protein. Only about 1300 nucleotide pairs are required to encode a protein of average size (about 430 amino acids in humans). Most of the remaining DNA in a gene consists of long stretches of noncoding DNA that interrupt the relatively short segments of DNA that code for protein. As will be discussed in detail in Chapter 6, the coding sequences are called **exons**; the intervening (noncoding) sequences in genes are called **introns** (see Figure 4–15 and Table 4–1). The majority of human genes thus consist of a long string of alternating exons and introns, with most of the gene consisting of introns. In contrast, the majority of genes from organisms with concise genomes lack introns. This accounts for the much smaller size of their genes (about one-twentieth that of human genes), as well as for the much higher fraction of coding DNA in their chromosomes.

In addition to introns and exons, each gene is associated with *regulatory DNA sequences*, which are responsible for ensuring that the gene is turned on or off at the proper time, expressed at the appropriate level, and only in the proper type of cell. In humans, the regulatory sequences for a typical gene are spread out over tens of thousands of nucleotide pairs. As would be expected, these regulatory sequences are more compressed in organisms with concise genomes. We discuss in Chapter 7 how regulatory DNA sequences work.

Finally, the nucleotide sequence of the human genome has revealed that the critical information needed to produce a human seems to be in an alarming state of disarray. As one commentator described our genome, "In some ways it may resemble your garage/bedroom/refrigerator/life: highly individualistic, but unkempt; little evidence of organization; much accumulated clutter (referred to by the uninitiated as 'junk'); virtually nothing ever discarded; and the few patently valuable items indiscriminately, apparently carelessly, scattered throughout."

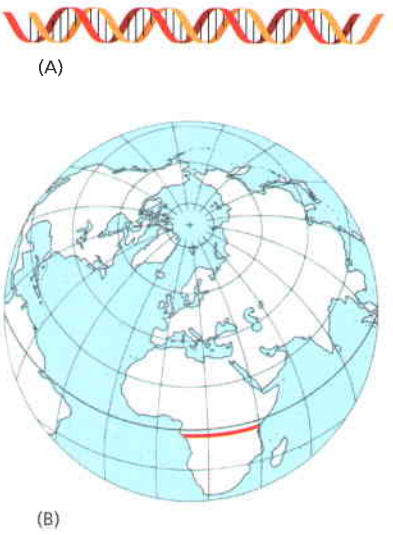


Figure 4–16 Scale of the human genome. If drawn with a 1 mm space between each nucleotide, as in (A), the human genome would extend 3200 km (approximately 2000 miles), far enough to stretch across the center of Africa, the site of our human origins *(red line* in B). At this scale, there would be, on average, a protein-coding gene every 130 m. An average gene would extend for 30 m, but the coding sequences in this gene would add up to only just over a meter.

Table 4–1 Some Vital Statistics for the Human Genome

	HUMAN GENOME
DNA length	3.2×10^9 nucleotide pairs*
Number of genes	approximately 25,000
Largest gene	2.4×10^6 nucleotide pairs
Mean gene size	27,000 nucleotide pairs
Smallest number of exons per gene	1
Largest number of exons per gene	178
Mean number of exons per gene	10.4
Largest exon size	17,106 nucleotide pairs
Mean exon size	145 nucleotide pairs
Number of pseudogenes**	more than 20,000
Percentage of DNA sequence in exons (protein coding sequences)	1.5%
Percentage of DNA in other highly conserved sequences***	3.5%
Percentage of DNA in high-copy repetitive elements	approximately 50%

* The sequence of 2.85 billion nucleotides is known precisely (error rate of only about one in 100,000 nucleotides). The remaining DNA primarily consists of short highly repeated sequences that are tandemly repeated, with repeat numbers differing from one individual to the next.

** A pseudogene is a nucleotide sequence of DNA closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.

*** Preserved functional regions; these include DNA encoding 5′ and 3′ UTRs (untranslated regions), structural and functional RNAs, and conserved protein-binding sites on the DNA.

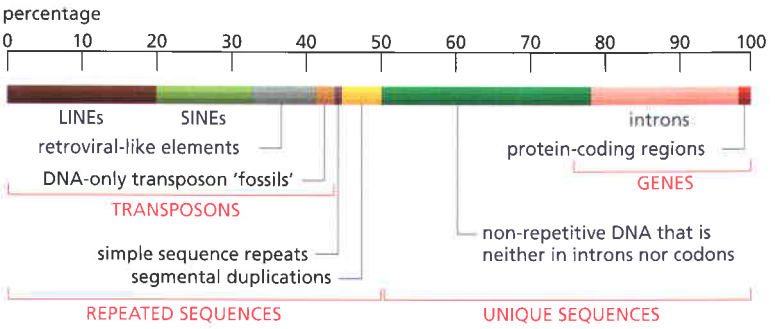


Figure 4–17 Representation of the nucleotide sequence content of the completely sequenced human genome. The LINEs, SINEs, retroviral-like elements, and DNA-only transposons are mobile genetic elements that have multiplied in our genome by replicating themselves and inserting the new copies in different positions. These mobile genetic elements are discussed in Chapter 5 (see Table 5–3, p. 318). Simple sequence repeats are short nucleotide sequences (less than 14 nucleotide pairs) that are repeated again and again for long stretches. Segmental duplications are large blocks of the genome (1000–200,000 nucleotide pairs) that are present at two or more locations in the genome. The most highly repeated blocks of DNA in heterochromatin have not yet been completely sequenced; therefore about 10% of human DNA sequences are not represented in this diagram. (Data courtesy of E. Margulies.)

Genome Comparisons Reveal Evolutionarily Conserved DNA Sequences

A major obstacle in interpreting the nucleotide sequences of human chromosomes is the fact that much of the sequence is probably unimportant. Moreover, the coding regions of the genome (the exons) are typically found in short segments (average size about 145 nucleotide pairs) floating in a sea of DNA whose exact nucleotide sequence is of little consequence. This arrangement makes it very difficult to identify all the exons in a stretch of DNA sequence. Even harder is the determination of where a gene begins and ends and exactly how many exons it spans.

Accurate gene identification requires approaches that extract information from the inherently low signal-to-noise ratio of the human genome. We shall describe some of them in Chapter 8. Here we discuss only one general approach, which is based on the observation that sequences that have a function are relatively conserved during evolution, whereas those without a function are free to mutate randomly. The strategy is therefore to compare the human sequence with that of the corresponding regions of a related genome, such as that of the mouse. Humans and mice are thought to have diverged from a common mammalian ancestor about 80×10^6 years ago, which is long enough for the majority of nucleotides in their genomes to have been changed by random mutational events. Consequently, the only regions that will have remained closely similar in the two genomes are those in which mutations would have impaired function and put the animals carrying them at a disadvantage, resulting in their elimination from the population by natural selection. Such closely similar regions are known as *conserved regions*. The conserved regions include both functionally important exons and regulatory DNA sequences. In contrast, *nonconserved regions* represent DNA whose sequence is unlikely to be critical for function.

The power of this method can be increased by comparing our genome with the genomes of additional animals whose genomes have been completely sequenced, including the rat, chicken, chimpanzee, and dog. By revealing in this way the results of a very long natural "experiment," lasting for hundreds of millions of years, such comparative DNA sequencing studies have highlighted the most interesting regions in these genomes. The comparisons reveal that roughly 5% of the human genome consists of "multi-species conserved sequences," as discussed in detail near the end of this chapter. Unexpectedly, only about one-third of these sequences code for proteins. Some of the conserved noncoding sequences correspond to clusters of protein-binding sites that are involved in gene regulation, while others produce RNA molecules that are not translated into protein. But the function of the majority of these sequences remains unknown. This unexpected discovery has led scientists to conclude that we understand much less about the cell biology of vertebrates than we had previously imagined. Certainly, there are enormous opportunities for new discoveries, and we should expect many surprises ahead.

Comparative studies have revealed not only that humans and other mammals share most of the same genes, but also that large blocks of our genomes contain these genes in the same order, a feature called *conserved synteny*. As a result, large blocks of our chromosomes can be recognized in other species. This allows the chromosome painting technique to be used to reconstruct the recent evolutionary history of human chromosomes (**Figure 4–18**).

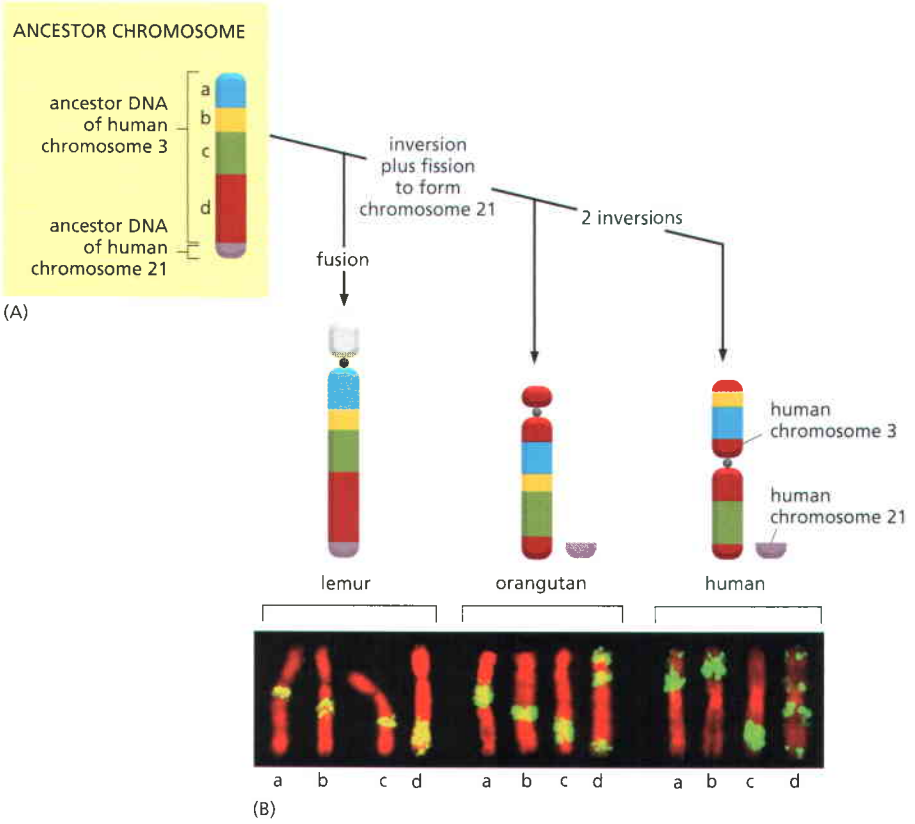


Figure 4–18 A proposed evolutionary history of human chromosome 3 and its relatives in other mammals. (A) The order of chromosome 3 segments hypothesized to be present on a chromosome of a mammalian ancestor is shown (*yellow* box). The minimum changes in this ancestral chromosome necessary to account for the appearance of each of the three modern chromosomes are indicated. (The present-day chromosomes of humans and African apes are identical at this resolution.) The *small circles* depicted in the modern chromosomes represent the positions of centromeres. A fission and inversion that leads to a change in chromosome organization is thought to occur once every $5–10 \times 10^6$ years in mammals. (B) Some of the chromosome painting experiments that led to the diagram in (A). Each image shows the chromosome most closely related to human chromosome 3, painted *green* by hybridization with different segments of DNA, lettered a, b, c, and d along the *bottom* of the figure. These letters correspond to the colored segments of the diagrams in (A), as indicated on the ancestral chromosome. (From S. Müller et al., *Proc. Natl Acad. Sci. U.S.A.* 97:206–211, 2000. With permission from National Academy of Sciences.)

Chromosomes Exist in Different States Throughout the Life of a Cell

We have seen how genes are arranged in chromosomes, but to form a functional chromosome, a DNA molecule must be able to do more than simply carry genes: it must be able to replicate, and the replicated copies must be separated and reliably partitioned into daughter cells at each cell division. This process occurs through an ordered series of stages, collectively known as the **cell cycle**, which provides for a temporal separation between the duplication of chromosomes and their segregation into two daughter cells. The cell cycle is briefly summarized in Figure 4–19, and it is discussed in detail in Chapter 17. Only certain

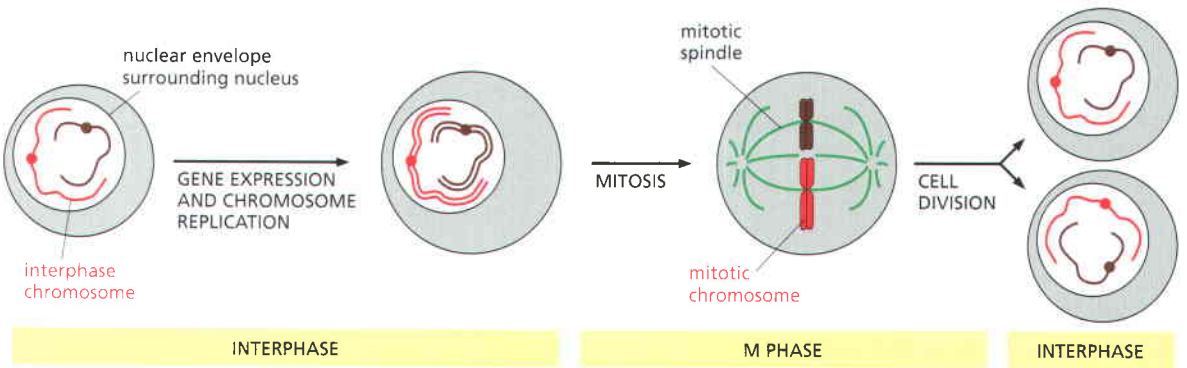


Figure 4–19 A simplified view of the eucaryotic cell cycle. During interphase, the cell is actively expressing its genes and is therefore synthesizing proteins. Also, during interphase and before cell division, the DNA is replicated and each chromosome is duplicated to produce two closely paired daughter chromosomes (a cell with only two chromosomes is illustrated here). Once DNA replication is complete, the cell can enter *M phase*, when mitosis occurs and the nucleus is divided into two daughter nuclei. During this stage, the chromosomes condense, the nuclear envelope breaks down, and the mitotic spindle forms from microtubules and other proteins. The condensed mitotic chromosomes are captured by the mitotic spindle, and one complete set of chromosomes is then pulled to each end of the cell by separating each daughter chromosome pair. A nuclear envelope re-forms around each chromosome set, and in the final step of M phase, the cell divides to produce two daughter cells. Most of the time in the cell cycle is spent in interphase; M phase is brief in comparison, occupying only about an hour in many mammalian cells.

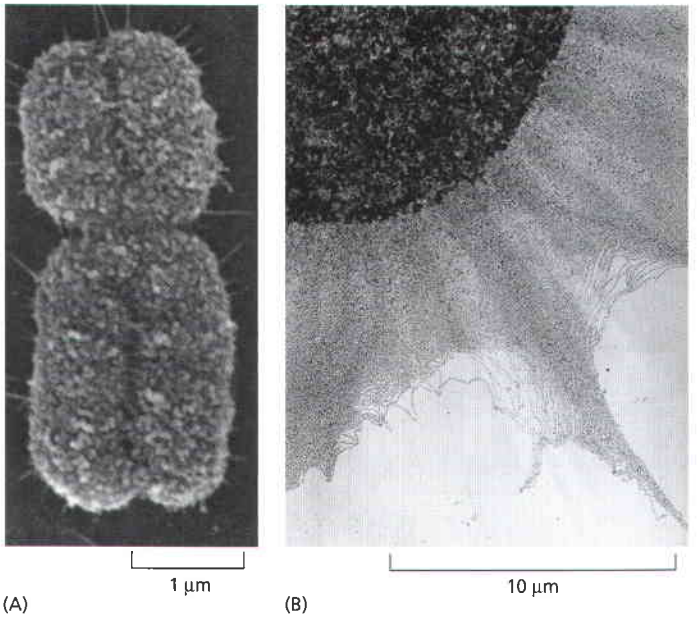


Figure 4–20 A comparison of extended interphase chromatin with the chromatin in a mitotic chromosome. (A) A scanning electron micrograph of a mitotic chromosome: a condensed duplicated chromosome in which the two new chromosomes are still linked together (see Figure 4–21). The constricted region indicates the position of the centromere, described in Figure 4–21. (B) An electron micrograph showing an enormous tangle of chromatin spilling out of a lysed interphase nucleus. Note the difference in scales. (A, courtesy of Terry D. Allen; B, courtesy of Victoria Foe.)

parts of the cycle concern us in this chapter. During *interphase* chromosomes are replicated, and during *mitosis* they become highly condensed and then are separated and distributed to the two daughter nuclei. The highly condensed chromosomes in a dividing cell are known as *mitotic chromosomes* (**Figure 4–20**A). This is the form in which chromosomes are most easily visualized; in fact, the images of chromosomes shown so far in the chapter are of chromosomes in mitosis. During cell division, this condensed state is important for the accurate separation of the duplicated chromosomes by the mitotic spindle, as discussed in Chapter 17.

During the portions of the cell cycle when the cell is not dividing, the chromosomes are extended and much of their chromatin exists as long, thin tangled threads in the nucleus so that individual chromosomes cannot be easily distinguished (Figure 4–20B). We shall refer to chromosomes in this extended state as *interphase chromosomes*. Since cells spend most of their time in interphase, and this is where their genetic information is being read out, chromosomes are of greatest interest to cell biologists when they are least visible.

Each DNA Molecule That Forms a Linear Chromosome Must Contain a Centromere, Two Telomeres, and Replication Origins

A chromosome operates as a distinct structural unit: for a copy to be passed on to each daughter cell at division, each chromosome must be able to replicate, and the newly replicated copies must subsequently be separated and partitioned correctly into the two daughter cells. These basic functions are controlled by three types of specialized nucleotide sequences in the DNA, each of which binds specific proteins that guide the machinery that replicates and segregates chromosomes (**Figure 4–21**).

Experiments in yeasts, whose chromosomes are relatively small and easy to manipulate, have identified the minimal DNA sequence elements responsible for each of these functions. One type of nucleotide sequence acts as a DNA **replication origin**, the location at which duplication of the DNA begins. Eucaryotic chromosomes contain many origins of replication to ensure that the entire chromosome can be replicated rapidly, as discussed in detail in Chapter 5.

After replication, the two daughter chromosomes remain attached to one another and, as the cell cycle proceeds, are condensed further to produce mitotic chromosomes. The presence of a second specialized DNA sequence, called a **centromere**, allows one copy of each duplicated and condensed chromosome to be pulled into each daughter cell when a cell divides. A protein

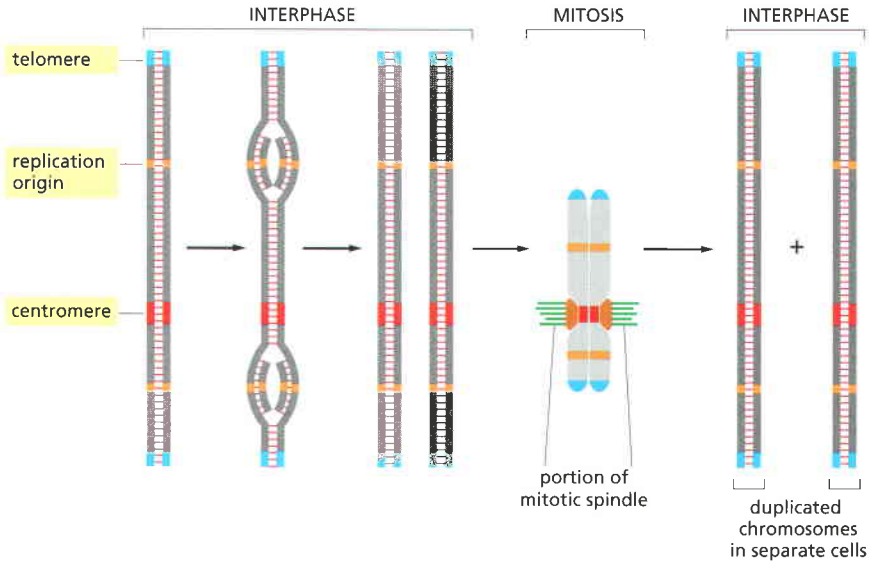


Figure 4–21 The three DNA sequences required to produce a eucaryotic chromosome that can be replicated and then segregated at mitosis. Each chromosome has multiple origins of replication, one centromere, and two telomeres. Shown here is the sequence of events that a typical chromosome follows during the cell cycle. The DNA replicates in interphase, beginning at the origins of replication and proceeding bidirectionally from the origins across the chromosome. In M phase, the centromere attaches the duplicated chromosomes to the mitotic spindle so that one copy is distributed to each daughter cell during mitosis. The centromere also helps to hold the duplicated chromosomes together until they are ready to be moved apart. The telomeres form special caps at each chromosome end.

complex called a *kinetochore* forms at the centromere and attaches the duplicated chromosomes to the mitotic spindle, allowing them to be pulled apart (discussed in Chapter 17).

The third specialized DNA sequence forms **telomeres**, the ends of a chromosome. Telomeres contain repeated nucleotide sequences that enable the ends of chromosomes to be efficiently replicated. Telomeres also perform another function: the repeated telomere DNA sequences, together with the regions adjoining them, form structures that protect the end of the chromosome from being mistaken by the cell for a broken DNA molecule in need of repair. We discuss both this type of repair and the structure and function of telomeres in Chapter 5.

In yeast cells, the three types of sequences required to propagate a chromosome are relatively short (typically less than 1000 base pairs each) and therefore use only a tiny fraction of the information-carrying capacity of a chromosome. Although telomere sequences are fairly simple and short in all eucaryotes, the DNA sequences that form centromeres and replication origins in more complex organisms are much longer than their yeast counterparts. For example, experiments suggest that human centromeres contain up to 100,000 nucleotide pairs and may not require a stretch of DNA with a defined nucleotide sequence. Instead, as we shall discuss later in this chapter, they seem to consist of a large, regularly repeating protein–nucleic acid structure that can be inherited when a chromosome replicates.

DNA Molecules Are Highly Condensed in Chromosomes

All eucaryotic organisms have special ways of packaging DNA into chromosomes. For example, if the 48 million nucleotide pairs of DNA in human chromosome 22 could be laid out as one long perfect double helix, the molecule would extend for about 1.5 cm if stretched out end to end. But chromosome 22 measures only about 2 μm in length in mitosis (see Figures 4–10 and 4–11), representing an end-to-end compaction ratio of nearly 10,000-fold. This remarkable feat of compression is performed by proteins that successively coil and fold the DNA into higher and higher levels of organization. Although much less condensed than mitotic chromosomes, the DNA of human interphase chromosomes is still tightly packed, with an overall compaction ratio of approximately 500-fold (the length of a chromosome's DNA helix divided by the end-to-end length of that chromosome).

In reading these sections it is important to keep in mind that chromosome structure is dynamic. We have seen that each chromosome condenses to an unusual degree in the M phase of the cell cycle. Much less visible, but of enormous interest and importance, specific regions of interphase chromosomes

decondense as the cells gain access to specific DNA sequences for gene expression, DNA repair, and replication—and then recondense when these processes are completed. The packaging of chromosomes is therefore accomplished in a way that allows rapid localized, on-demand access to the DNA. In the next sections we discuss the specialized proteins that make this type of packaging possible.

Nucleosomes Are a Basic Unit of Eucaryotic Chromosome Structure

The proteins that bind to the DNA to form eucaryotic chromosomes are traditionally divided into two general classes: the **histones** and the *nonhistone chromosomal proteins*. The complex of both classes of protein with the nuclear DNA of eucaryotic cells is known as **chromatin**. Histones are present in such enormous quantities in the cell (about 60 million molecules of each type per human cell) that their total mass in chromatin is about equal to that of the DNA.

Histones are responsible for the first and most basic level of chromosome packing, the **nucleosome**, a protein–DNA complex discovered in 1974. When interphase nuclei are broken open very gently and their contents examined under the electron microscope, most of the chromatin is in the form of a fiber with a diameter of about 30 nm (**Figure 4–22**A). If this chromatin is subjected to treatments that cause it to unfold partially, it can be seen under the electron microscope as a series of "beads on a string" (Figure 4–22B). The string is DNA, and each bead is a "nucleosome core particle" that consists of DNA wound around a protein core formed from histones. <ACTC>

The structural organization of nucleosomes was determined after first isolating them from unfolded chromatin by digestion with particular enzymes (called nucleases) that break down DNA by cutting between the nucleosomes. After digestion for a short period, the exposed DNA between the nucleosome core particles, the *linker DNA*, is degraded. Each individual nucleosome core particle consists of a complex of eight histone proteins—two molecules each of histones H2A, H2B, H3, and H4—and double-stranded DNA that is 147 nucleotide pairs long. The *histone octamer* forms a protein core around which the double-stranded DNA is wound (**Figure 4–23**).

Each nucleosome core particle is separated from the next by a region of linker DNA, which can vary in length from a few nucleotide pairs up to about 80. (The term *nucleosome* technically refers to a nucleosome core particle plus one of its adjacent DNA linkers, but it is often used synonymously with nucleosome core particle.) On average, therefore, nucleosomes repeat at intervals of about 200 nucleotide pairs. For example, a diploid human cell with 6.4×10^9 nucleotide pairs contains approximately 30 million nucleosomes. The formation of nucleosomes converts a DNA molecule into a chromatin thread about one-third of its initial length.

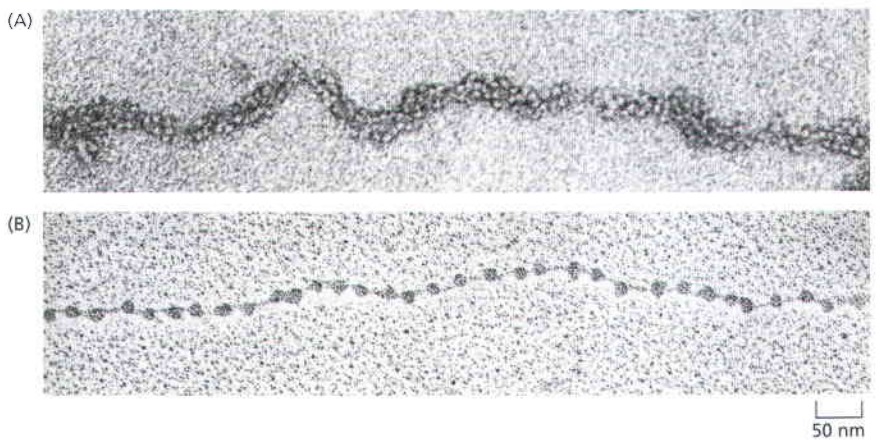


Figure 4–22 Nucleosomes as seen in the electron microscope. (A) Chromatin isolated directly from an interphase nucleus appears in the electron microscope as a thread 30 nm thick. (B) This electron micrograph shows a length of chromatin that has been experimentally unpacked, or decondensed, after isolation to show the nucleosomes. (A, courtesy of Barbara Hamkalo; B, courtesy of Victoria Foe.)

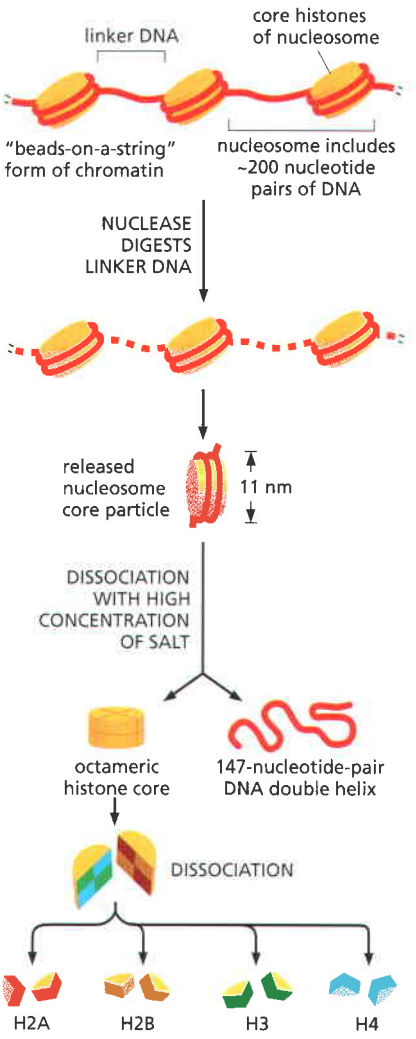


Figure 4–23 Structural organization of the nucleosome. A nucleosome contains a protein core made of eight histone molecules. In biochemical experiments, the nucleosome core particle can be released from isolated chromatin by digestion of the linker DNA with a nuclease, an enzyme that breaks down DNA. (The nuclease can degrade the exposed linker DNA but cannot attack the DNA wound tightly around the nucleosome core.) After dissociation of the isolated nucleosome into its protein core and DNA, the length of the DNA that was wound around the core can be determined. This length of 147 nucleotide pairs is sufficient to wrap 1.7 times around the histone core.

The Structure of the Nucleosome Core Particle Reveals How DNA Is Packaged

The high-resolution structure of a nucleosome core particle, solved in 1997, revealed a disc-shaped histone core around which the DNA was tightly wrapped 1.7 turns in a left-handed coil (**Figure 4–24**). All four of the histones that make up the core of the nucleosome are relatively small proteins (102–135 amino acids), and they share a structural motif, known as the *histone fold*, formed from three α helices connected by two loops (**Figure 4–25**). In assembling a nucleosome, the histone folds first bind to each other to form H3–H4 and H2A–H2B dimers, and the H3–H4 dimers combine to form tetramers. An H3–H4 tetramer then further combines with two H2A–H2B dimers to form the compact octamer core, around which the DNA is wound (**Figure 4–26**).

The interface between DNA and histone is extensive: 142 hydrogen bonds are formed between DNA and the histone core in each nucleosome. Nearly half of these bonds form between the amino acid backbone of the histones and the phosphodiester backbone of the DNA. Numerous hydrophobic interactions and salt linkages also hold DNA and protein together in the nucleosome. For example, more than one-fifth of the amino acids in each of the core histones are either lysine or arginine (two amino acids with basic side chains), and their positive

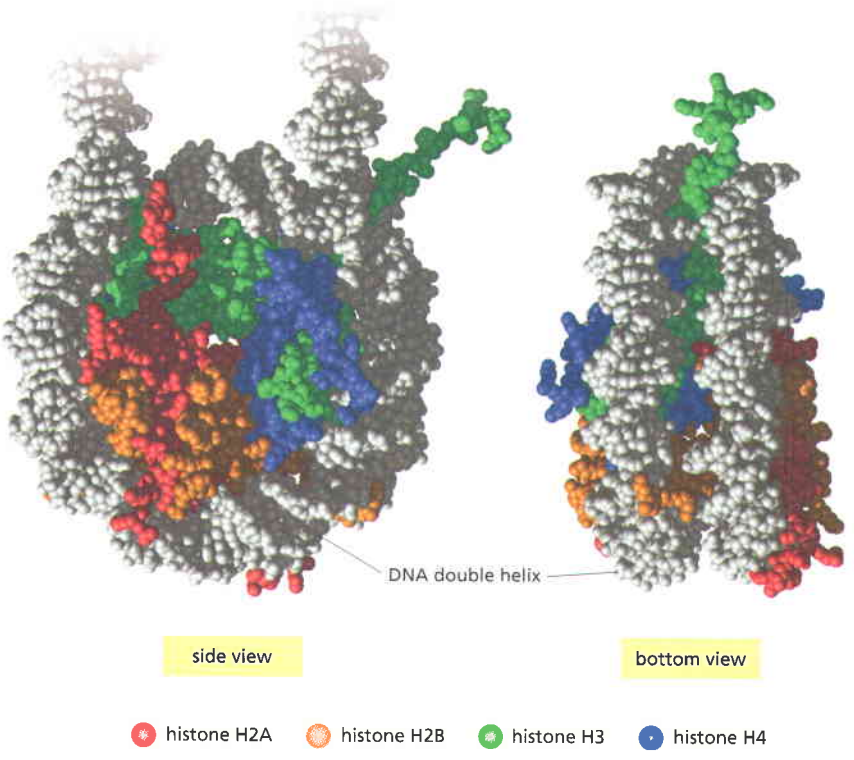


Figure 4–24 The structure of a nucleosome core particle, as determined by x-ray diffraction analyses of crystals. Each histone is colored according to the scheme in Figure 4–23, with the DNA double helix in *light gray*. (From K. Luger et al., *Nature* 389:251–260, 1997. With permission from Macmillan Publishers Ltd.)

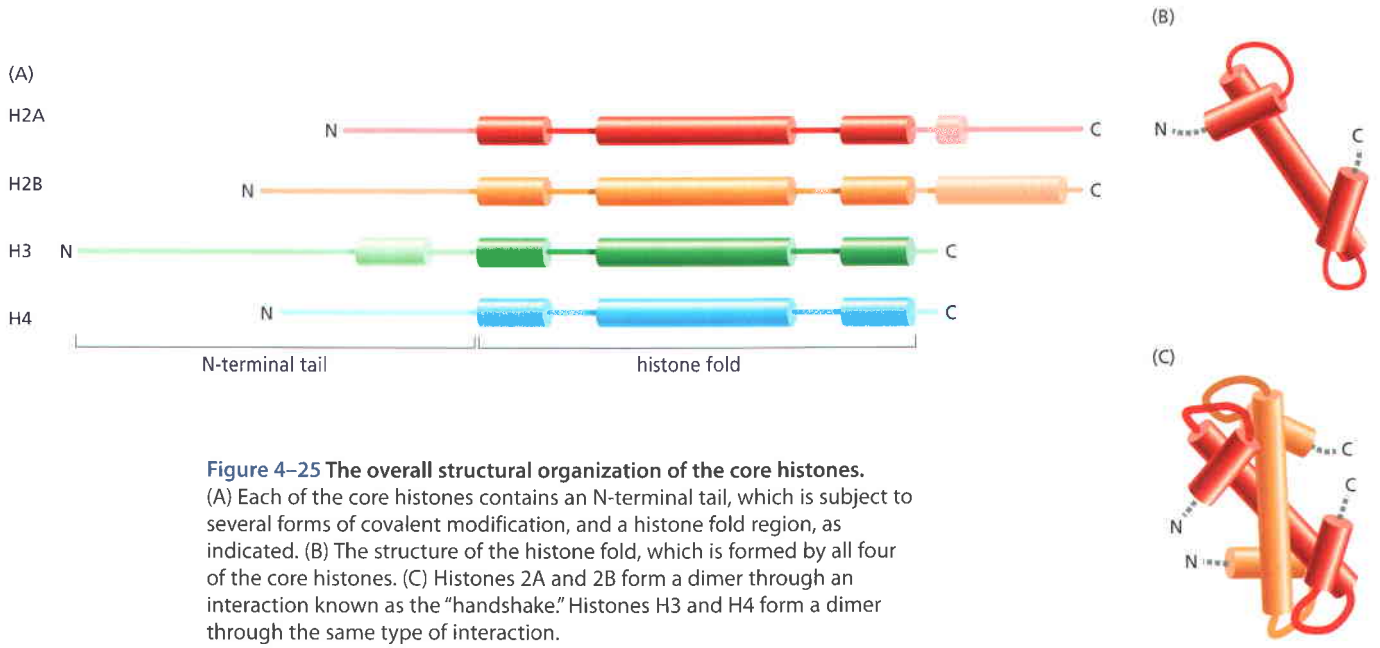


Figure 4–25 The overall structural organization of the core histones. (A) Each of the core histones contains an N-terminal tail, which is subject to several forms of covalent modification, and a histone fold region, as indicated. (B) The structure of the histone fold, which is formed by all four of the core histones. (C) Histones 2A and 2B form a dimer through an interaction known as the "handshake." Histones H3 and H4 form a dimer through the same type of interaction.

charges can effectively neutralize the negatively charged DNA backbone. These numerous interactions explain in part why DNA of virtually any sequence can be bound on a histone octamer core. The path of the DNA around the histone core is not smooth; rather, several kinks are seen in the DNA, as expected from the nonuniform surface of the core. The bending requires a substantial compression of the minor groove of the DNA helix. Certain dinucleotides in the minor groove are especially easy to compress, and some nucleotide sequences bind the nucleosome more tightly than others (**Figure 4–27**). This probably explains some striking, but unusual, cases of very precise positioning of nucleosomes along a stretch of DNA. For most of the DNA sequences found in chromosomes, however, the sequence preference of nucleosomes must be small enough to allow other factors to dominate, inasmuch as nucleosomes can occupy any one of a number of positions relative to the DNA sequence in most chromosomal regions.

In addition to its histone fold, each of the core histones has an N-terminal amino acid "tail", which extends out from the DNA–histone core (see Figure 4–26). These histone tails are subject to several different types of covalent modifications that in turn control critical aspects of chromatin structure and function, as we shall discuss shortly.

As a reflection of their fundamental role in DNA function through controlling chromatin structure, the histones are among the most highly conserved eucaryotic proteins. For example, the amino acid sequence of histone H4 from a pea and from a cow differ at only 2 of the 102 positions. This strong evolutionary conservation suggests that the functions of histones involve nearly all of their amino acids, so that a change in any position is deleterious to the cell. This suggestion has been tested directly in yeast cells, in which it is possible to mutate a given histone gene *in vitro* and introduce it into the yeast genome in place of the normal gene. As might be expected, most changes in histone sequences are lethal; the few that are not lethal cause changes in the normal pattern of gene expression, as well as other abnormalities.

Despite the high conservation of the core histones, eucaryotic organisms also produce smaller amounts of specialized variant core histones that differ in amino acid sequence from the main ones. As we shall see, these variants, combined with a surprisingly large variety of covalent modifications that can be added to the histones in nucleosomes, make possible the many different chromatin structures that are required for DNA function in higher eucaryotes.

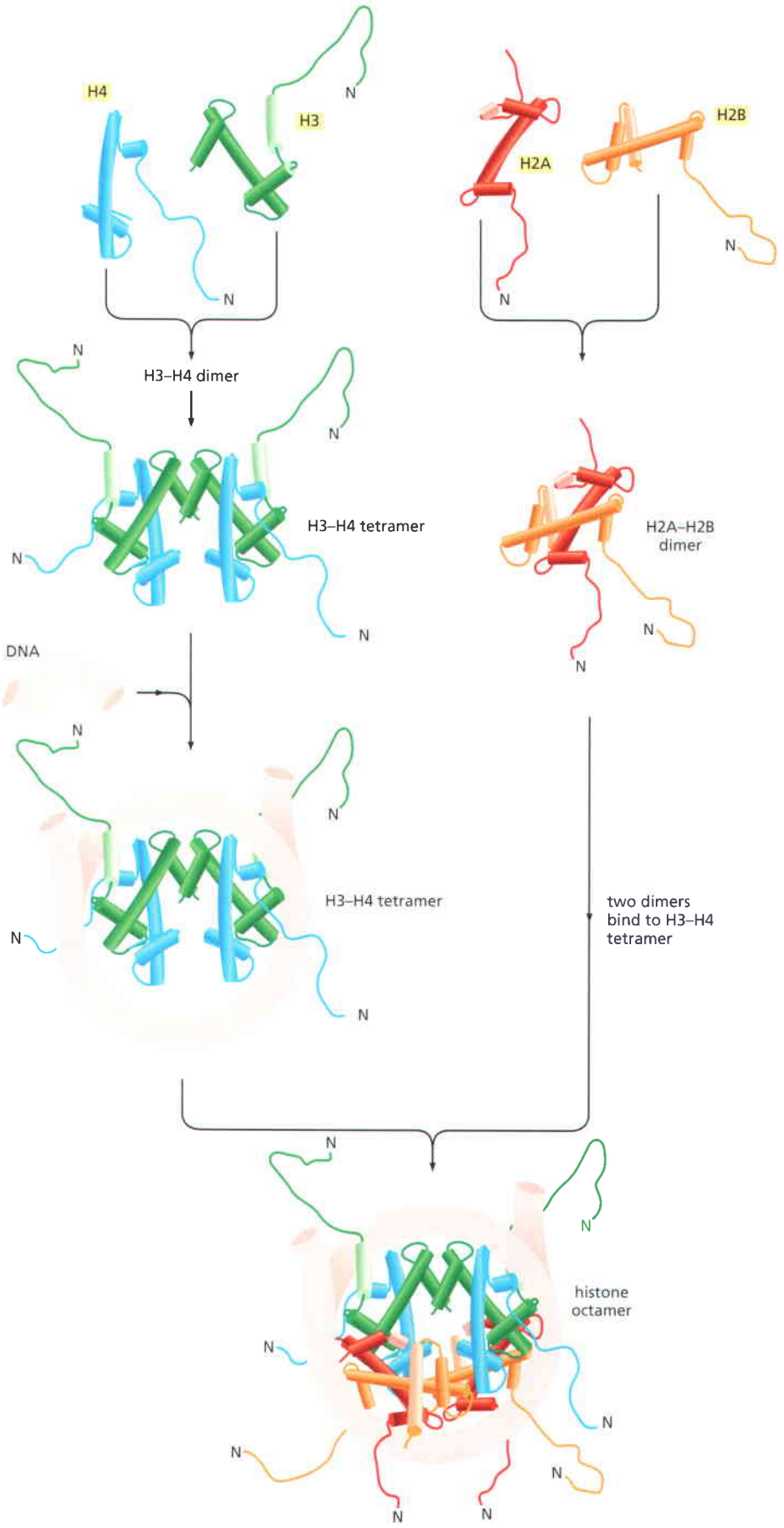


Figure 4–26 The assembly of a histone octamer on DNA. The histone H3–H4 dimer and the H2A–H2B dimer are formed from the handshake interaction. An H3–H4 tetramer forms and binds to the DNA. Two H2A–H2B dimers are then added, to complete the nucleosome. The histones are colored as in Figures 4–24 and 4–25. Note that all eight N-terminal tails of the histones protrude from the disc-shaped core structure. Their conformations are highly flexible.

Inside the cell, the nucleosome assembly reactions shown here are mediated by *histone chaperone* proteins, some specific for H3–H4 and others specific for H2A–H2B. (Adapted from figures by J. Waterborg.)

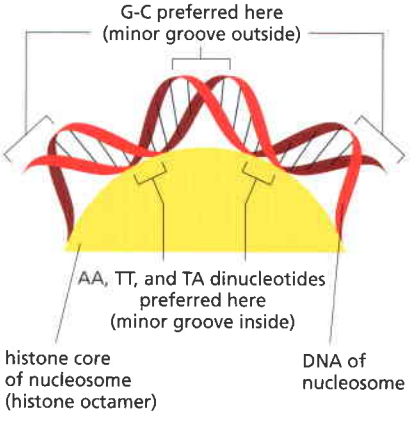


Figure 4–27 The bending of DNA in a nucleosome. The DNA helix makes 1.7 tight turns around the histone octamer. This diagram illustrates how the minor groove is compressed on the inside of the turn. Owing to certain structural features of the DNA molecule, the indicated dinucleotides are preferentially accommodated in such a narrow minor groove, which helps to explain why certain DNA sequences will bind more tightly than others to the nucleosome core.

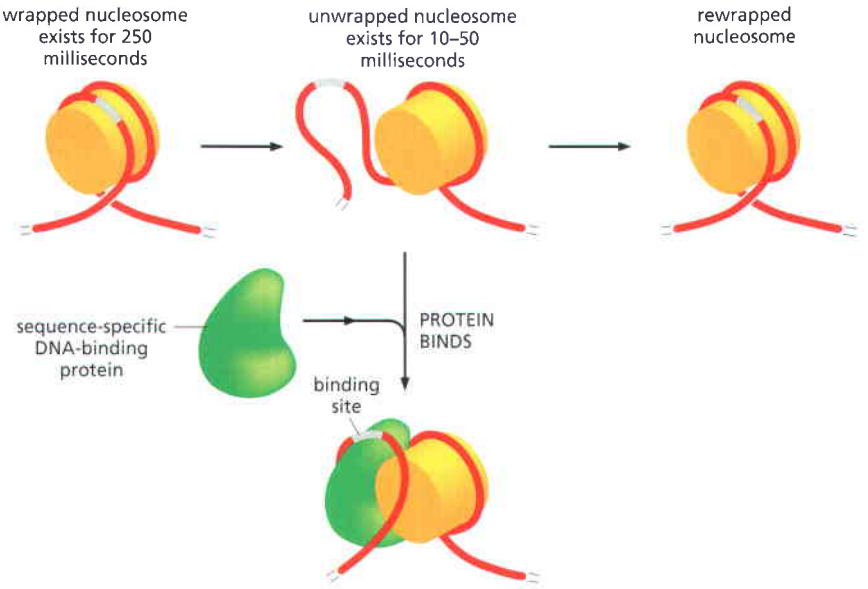


Figure 4–28 Dynamic nucleosomes. Kinetic measurements show that the DNA in an isolated nucleosome is surprisingly dynamic, rapidly uncoiling and then rewrapping around its nucleosome core. As indicated, this makes most of its bound DNA sequence accessible to other DNA-binding proteins. (Data from G. Li and J. Widom, *Nat. Struct. Mol. Biol.* 11:763–769, 2004. With permission from Macmillan Publishers Ltd.)

Nucleosomes Have a Dynamic Structure, and Are Frequently Subjected to Changes Catalyzed by ATP-Dependent Chromatin-Remodeling Complexes

For many years biologists thought that, once formed in a particular position on DNA, a nucleosome remains fixed in place because of the very tight association between its core histones and DNA. If true, this would pose problems for genetic readout mechanisms, which in principle require rapid access to many specific DNA sequences, as well as for the rapid passage of the DNA transcription and replication machinery through chromatin. But kinetic experiments show that the DNA in an isolated nucleosome unwraps from each end at rate of about 4 times per second, remaining exposed for 10 to 50 milliseconds before the partially unwrapped structure recloses. Thus, most of the DNA in an isolated nucleosome is in principle available for binding other proteins (**Figure 4–28**).

For the chromatin in a cell, a further loosening of DNA–histone contacts is clearly required, because eucaryotic cells contain a large variety of ATP-dependent *chromatin remodeling complexes*. The subunit in these complexes that hydrolyzes ATP is evolutionarily related to the DNA helicases (discussed in Chapter 5), and it binds both to the protein core of the nucleosome and to the double-stranded DNA that winds around it. By using the energy of ATP hydrolysis to move this DNA relative to the core, this subunit changes the structure of a nucleosome temporarily, making the DNA less tightly bound to the histone core. Through repeated cycles of ATP hydrolysis, the remodeling complexes can catalyze *nucleosome sliding*, and by pulling the nucleosome core along the DNA double helix in this way, they make the nucleosomal DNA available to other proteins in the cell (**Figure 4–29**). In addition, by cooperating with negatively

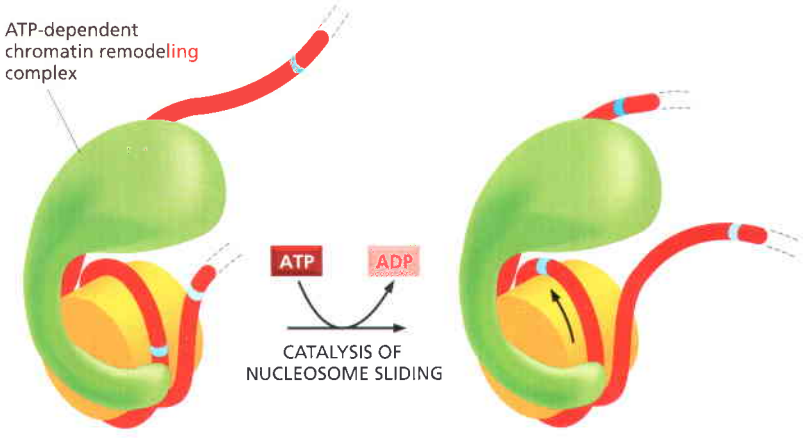


Figure 4–29 The nucleosome sliding catalyzed by ATP-dependent chromatin remodeling complexes. Using the energy of ATP hydrolysis, the remodeling complex is thought to push on the DNA of its bound nucleosome and loosen its attachment to the nucleosome core. Each cycle of ATP binding, ATP hydrolysis, and release of the ADP and P_i products thereby moves the DNA with respect to the histone octamer in the direction of the arrow in this diagram. It requires many such cycles to produce the nucleosome sliding shown. (See also Figure 4–46B.)

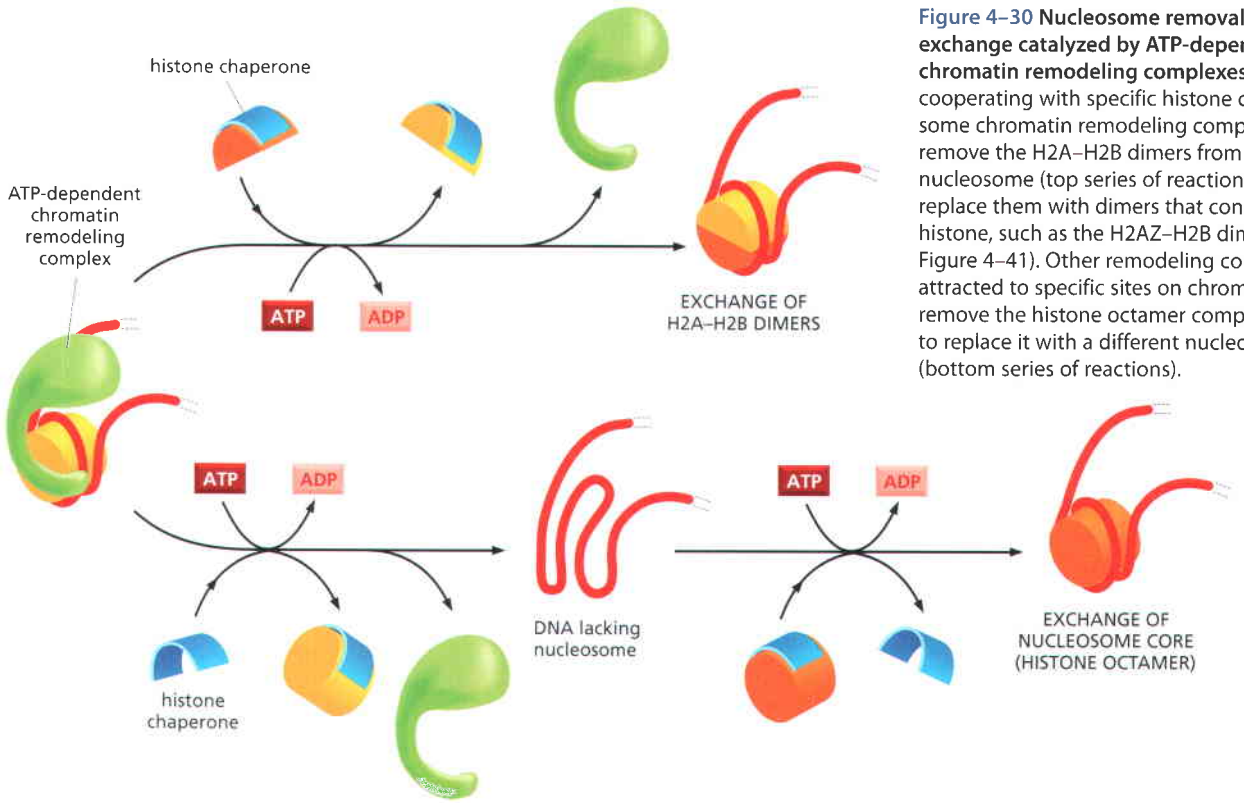


Figure 4–30 Nucleosome removal and histone exchange catalyzed by ATP-dependent chromatin remodeling complexes. By cooperating with specific histone chaperones, some chromatin remodeling complexes can remove the H2A–H2B dimers from a nucleosome (top series of reactions) and replace them with dimers that contain a variant histone, such as the H2AZ–H2B dimer (see Figure 4–41). Other remodeling complexes are attracted to specific sites on chromatin to remove the histone octamer completely and/or to replace it with a different nucleosome core (bottom series of reactions).

charged proteins that serve as histone chaperones, some remodeling complexes are able to remove either all or part of the nucleosome core from a nucleosome—catalyzing either an exchange of its H2A–H2B histones, or the complete removal of the octameric core from the DNA (**Figure 4–30**).

Cells contain dozens of different ATP-dependent chromatin remodeling complexes that are specialized for different roles. Most are large protein complexes that can contain 10 or more subunits. The activity of these complexes is carefully controlled by the cell. As genes are turned on and off, chromatin remodeling complexes are brought to specific regions of DNA where they act locally to influence chromatin structure (discussed in Chapter 7; see also Figure 4–46, below).

As pointed out previously, for most of the DNA sequences found in chromosomes, experiments show that a nucleosome can occupy any one of a number of positions relative to the DNA sequence. The most important influence on nucleosome positioning appears to be the presence of other tightly bound proteins on the DNA. Some bound proteins favor the formation of a nucleosome adjacent to them. Others create obstacles that force the nucleosomes to move to positions between them. The exact positions of nucleosomes along a stretch of DNA therefore depends mainly on the presence and nature of other proteins bound to the DNA. Due to the presence of ATP-dependent remodeling complexes, the arrangement of nucleosomes on DNA can be highly dynamic, changing rapidly according to the needs of the cell.

Nucleosomes Are Usually Packed Together into a Compact Chromatin Fiber

Although enormously long strings of nucleosomes form on the chromosomal DNA, chromatin in a living cell probably rarely adopts the extended "beads on a string" form. Instead, the nucleosomes are packed on top of one another, generating regular arrays in which the DNA is even more highly condensed. Thus, when nuclei are very gently lysed onto an electron microscope grid, most of the chromatin is seen to be in the form of a fiber with a diameter of about 30 nm, which is considerably wider than chromatin in the "beads on a string" form (see Figure 4–22).

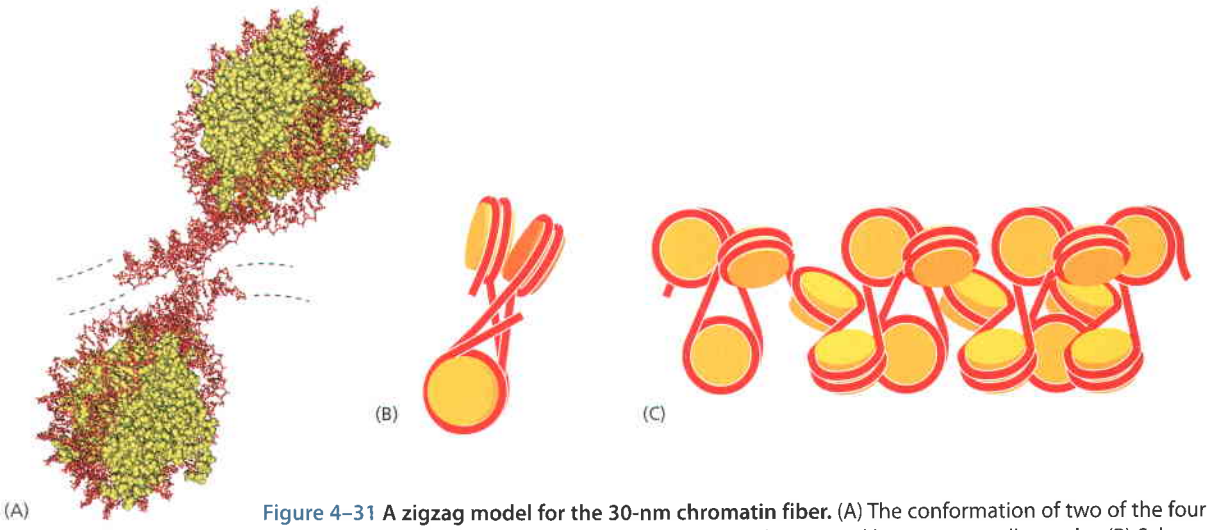


Figure 4–31 A zigzag model for the 30-nm chromatin fiber. (A) The conformation of two of the four nucleosomes in a tetranucleosome, from a structure determined by x-ray crystallography. (B) Schematic of the entire tetranucleosome; the fourth nucleosome is not visible, being stacked on the bottom nucleosome and behind it in this diagram. (C) Diagrammatic illustration of a possible zigzag structure that could account for the 30-nm chromatin fiber. (Adapted from C.L. Woodcock, *Nat. Struct. Mol. Biol.* 12:639–640, 2005. With permission from Macmillan Publishers Ltd.)

How are nucleosomes packed in the 30-nm chromatin fiber? This question has not yet been answered definitively, but important information concerning the structure has been obtained. In particular, high-resolution structural analyses have been performed on homogeneous short strings of nucleosomes, prepared from purified histones and purified DNA molecules. The structure of a tetranucleosome, obtained by X-ray crystallography, has been used to support a zigzag model for the stacking of nucleosomes in the 30-nm fiber (**Figure 4–31**). But cryoelectron microscopy of longer strings of nucleosomes supports a very different solenoidal structure with intercalated nucleosomes (**Figure 4–32**).

What causes the nucleosomes to stack so tightly on each other in a 30-nm fiber? The nucleosome to nucleosome linkages formed by histone tails, most notably the H4 tail (**Figure 4–33**) constitute one important factor. Another

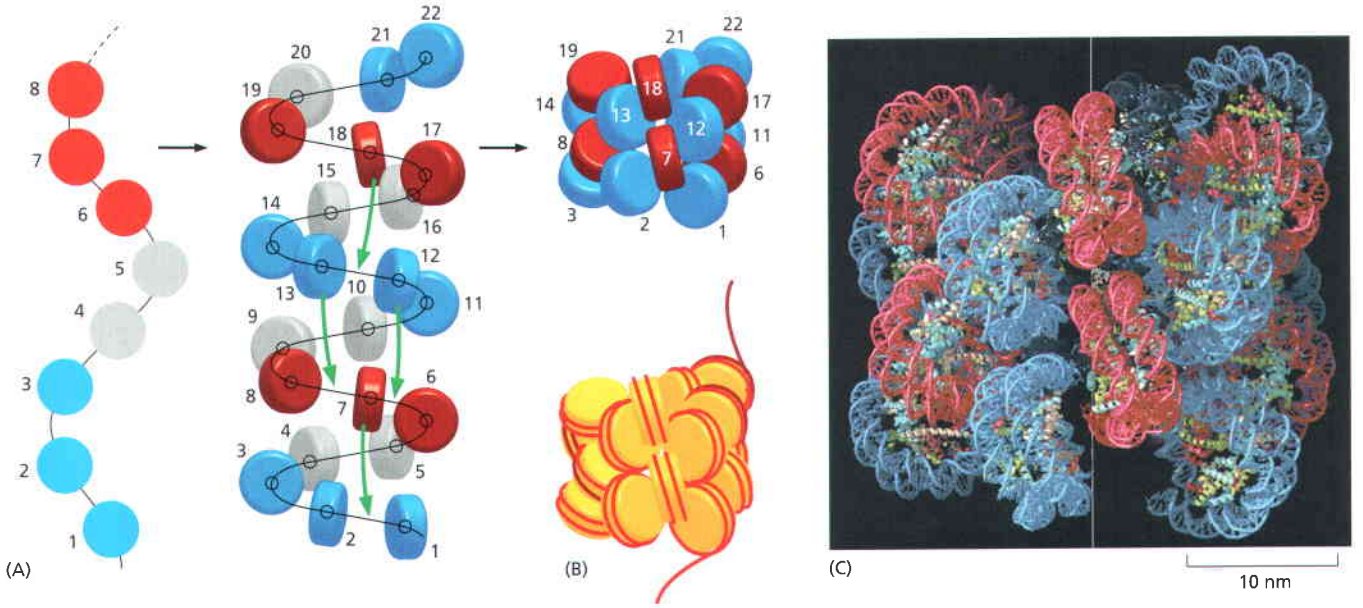
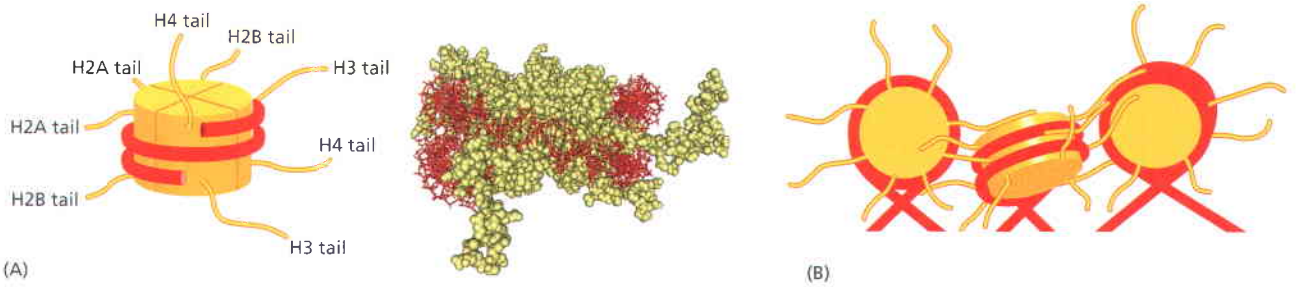


Figure 4–32 An interdigitated solenoid model for the 30-nm chromatin fiber. (A) Drawings in which strings of color-coded nucleosomes are used to illustrate how the solenoid is generated. (B) Schematic diagram of final structure in (A). (C) Structural model. The model is derived from high-resolution cryoelectron microscopy images of nucleosome arrays reconstituted from purified histones and DNA molecules of specific length and sequence. Both nucleosome octamers and a linker histone (discussed below) were used to produce regularly repeating arrays containing up to 72 nucleosomes. (Adapted from P. Robinson, L. Fairall, V. Huynh and D. Rhodes, *Proc. Natl Acad. Sci. U.S.A.* 103:6506–6511, 2006. With permission from National Academy of Sciences.)



important factor is an additional histone that is often present in a 1-to-1 ratio with nucleosome cores, known as **histone H1**. This so-called linker histone is larger than the individual core histones and it has been considerably less well conserved during evolution. A single histone H1 molecule binds to each nucleosome, contacting both DNA and protein, and changing the path of the DNA as it exits from the nucleosome. Although it is not understood in detail how H1 pulls nucleosomes together into the 30-nm fiber, a change in the exit path in DNA seems crucial for compacting nucleosomal DNA so that it interlocks to form the 30-nm fiber (**Figure 4–34**). Most eucaryotic organisms make several histone H1 proteins of related but quite distinct amino acid sequences.

It is possible that the 30-nm structure found in chromosomes is a fluid mosaic of several different variations. For example, a linker histone in the H1 family was present in the nucleosomal arrays studied in Figure 4–32 but was missing from the tetranucleosome in Figure 4–31. Moreover, we saw earlier that the linker DNA that connects adjacent nucleosomes can vary in length; these differences in linker length probably introduce local perturbations into the structure. And the presence of many other DNA-binding proteins, as well as proteins that bind directly to histones, will certainly add important additional features to any array of nucleosomes.

Figure 4–33 A speculative model for the role played by histone tails in the formation of the 30-nm fiber. (A) This schematic diagram shows the approximate exit points of the eight histone tails, one from each histone protein, that extend from each nucleosome. The actual structure is shown to its right. In the high-resolution structure of the nucleosome, the tails are largely unstructured, suggesting that they are highly flexible. (B) A speculative model showing how the histone tails may help to pack nucleosomes together into the 30-nm fiber. This model is based on (1) experimental evidence that histone tails aid in the formation of the 30-nm fiber, and (2) the x-ray crystal structure of the nucleosome, in which the tails of one nucleosome contact the histone core of an adjacent nucleosome in the crystal lattice.

Summary

A gene is a nucleotide sequence in a DNA molecule that acts as a functional unit for the production of a protein, a structural RNA, or a catalytic or regulatory RNA molecule. In eucaryotes, protein-coding genes are usually composed of a string of alternating introns and exons associated with regulatory regions of DNA. A chromosome is formed from a single, enormously long DNA molecule that contains a linear array of many genes. The human genome contains 3.2 × 10^9 DNA nucleotide pairs, divided between 22 different autosomes and 2 sex chromosomes. Only a small percentage of this DNA codes for proteins or functional RNA molecules. A chromosomal DNA molecule also contains three other types of functionally important nucleotide sequences: replication origins and telomeres allow the DNA molecule to be efficiently replicated, while a centromere attaches the daughter DNA molecules to the mitotic spindle, ensuring their accurate segregation to daughter cells during the M phase of the cell cycle.

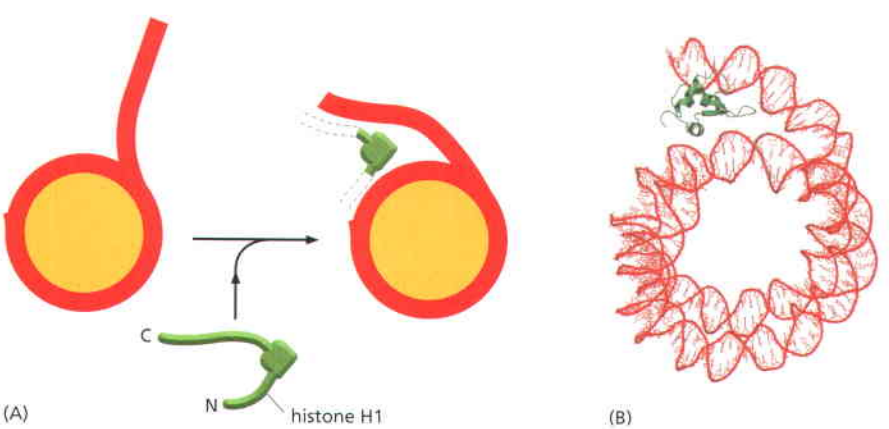


Figure 4–34 How the linker histone binds to the nucleosome. The position and structure of the globular region of histone H1 are shown. As indicated, this region constrains an additional 20 nucleotide pairs of DNA where it exits from the nucleosome core. This type of binding by H1 is thought to be important for forming the 30-nm chromatin fiber. The long C-terminal tail of histone H1 is also required for the high-affinity binding of H1 to chromatin, but neither its position or that of the N-terminal tail is known. (A) Schematic, (B) structure. (B, from D. Brown, T. Izard and T. Misteli, *Nat. Struct. Mol. Biol.* 13:250–255, 2006. With permission from Macmillan Publishers Ltd.)

The DNA in eucaryotes is tightly bound to an equal mass of histones, which form repeated arrays of DNA–protein particles called nucleosomes. The nucleosome is composed of an octameric core of histone proteins around which the DNA double helix is wrapped. Nucleosomes are spaced at intervals of about 200 nucleotide pairs, and they are usually packed together (with the aid of histone H1 molecules) into quasi-regular arrays to form a 30-nm chromatin fiber. Despite the high degree of compaction in chromatin, its structure must be highly dynamic to allow access to the DNA. There is some spontaneous DNA unwrapping and rewrapping in the nucleosome itself; however, the general strategy for reversibly changing local chromatin structure features ATP-driven chromatin remodeling complexes. Cells contain a large set of such complexes, which are targeted to specific regions of chromatin at appropriate times. The remodeling complexes collaborate with histone chaperones to allow nucleosome cores to be repositioned, reconstituted with different histones, or completely removed to expose the underlying DNA.

THE REGULATION OF CHROMATIN STRUCTURE

Having described how DNA is packaged into nucleosomes to create a chromatin fiber, we now turn to the mechanisms that create different chromatin structures in different regions of a cell's genome. We now know that mechanisms of this type are used to control many genes in eucaryotes. Most importantly, certain types of chromatin structure can be inherited; that is, the structure can be directly passed down from a cell to its decendents. Because the cell memory that results is based on an inherited protein structure rather than on a change in DNA sequence, this is a form of **epigenetic inheritance**. The prefix *epi* is Greek for "on"; this is appropriate, because epigenetics represents a form of inheritance that is superimposed on the genetic inheritance based on DNA (**Figure 4–35**).

In Chapter 7, we shall introduce the many different ways in which the expression of genes is regulated. There we discuss epigenetic inheritance in detail and present several distinct mechanisms that can produce it. Here, we are concerned with only one, that based on chromatin structure. We begin this section with an introduction to inherited chromatin structures and then describe the basis for them—the covalent modification of histones in nucleosomes. We shall see that these modifications serve as recognition sites for protein modules that bring specific protein complexes to the appropriate regions of chromatin, thereby producing specific effects on gene expression or inducing other biological functions. Through such mechanisms, chromatin structure plays a central role in the development, growth, and maintenance of eucaryotic organisms, including ourselves.

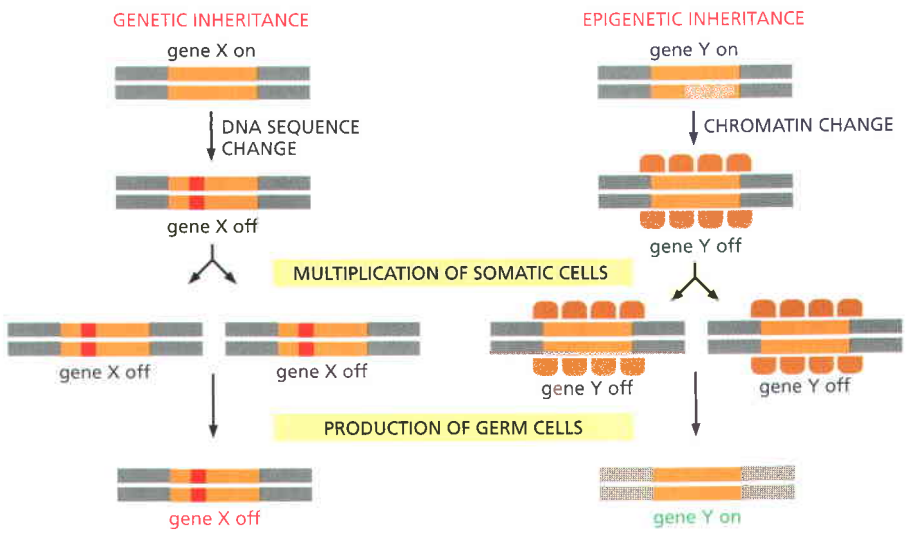


Figure 4–35 A comparison of genetic inheritance with an epigenetic inheritance based on chromatin structures. Genetic inheritance is based on the direct inheritance of DNA nucleotide sequences during DNA replication. DNA sequence changes are not only transmitted faithfully from a somatic cell to all of its descendents, but also through germ cells from one generation to the next. The field of genetics, reviewed in Chapter 8, is based on the inheritance of these changes between generations. The type of epigenetic inheritance shown here is based on other molecules bound to the DNA, and it is therefore less permanent than a change in DNA sequence; in particular, epigenetic information is usually (but not always) erased during the formation of eggs and sperm.

Only one epigenetic mechanism, that based on an inheritance of chromatin structures, is discussed in this chapter. Other epigenetic mechanisms are presented in Chapter 7, which focuses on the control of gene expression (see Figure 7–86).

Some Early Mysteries Concerning Chromatin Structure

Thirty years ago, histones were viewed as relatively uninteresting proteins. Nucleosomes were known to cover all of the DNA in chromosomes, and they were thought to exist to allow the enormous amounts of DNA in many eucaryotic cells to be packaged into compact chromosomes. Extrapolating from what was known in bacteria, many scientists believed that gene regulation in eucaryotes would simply bypass nucleosomes, treating them as uninvolved bystanders.

But there were reasons to challenge this view. Thus, for example, biochemists had determined that mammalian chromatin consists of an approximately equal mass of histone and non-histone proteins. This would mean that, *on average*, every 200 nucleotide pairs of DNA in our cells is associated with more than 1000 amino acids of non-histone proteins (that is, a mass of protein equivalent to the total mass of the histone octamer plus histone H1). We now know that many of these proteins bind to nucleosomes, and their abundance might suggest that histones are more than just packaging proteins.

A second reason to challenge the view that histones were inconsequential to gene regulation was based on the amazingly slow rate of evolutionary change in the sequences of the four core histones. The previously mentioned fact that there are only two amino acid differences in the sequence of mammalian and pea histone H4 implies that a change in almost any one of the 102 amino acids in H4 must be deleterious to these organisms. What type of process could make the life of an organism so sensitive to the exact structure of the nucleosome core that only two amino acids had changed in more than 500 million years of random variation followed by natural selection?

Last but not least, a combination of genetics and cytology had revealed that a particular form of chromatin silences the genes that it packages without regard to nucleotide sequence—and does so in a manner that is directly inherited by both daughter cells when a cell divides. It is to this subject that we turn next.

Heterochromatin Is Highly Organized and Unusually Resistant to Gene Expression

Light-microscope studies in the 1930s distinguished two types of chromatin in the interphase nuclei of many higher eucaryotic cells: a highly condensed form, called **heterochromatin**, and all the rest, which is less condensed, called **euchromatin**. Heterochromatin represents an especially compact form of chromatin (see Figure 4–9), and we are finally beginning to understand important aspects of its molecular properties. Although present in many locations along chromosomes, it is also highly concentrated in specific regions, most notably at the centromeres and telomeres introduced previously (see Figure 4–21). In a typical mammalian cell, more than ten percent of the genome is packaged in this way.

The DNA in heterochromatin contains very few genes, and those euchromatic genes that become packaged into heterochromatin are turned off by this type of packaging. However, we know now that the term *heterochromatin* encompasses several distinct types of chromatin structures whose common feature is an especially high degree of compaction. Thus, heterochromatin should not be thought of as encapsulating "dead" DNA, but rather as creating different types of compact chromatin with distinct features that make it highly resistant to gene expression for the vast majority of genes.

When a gene that is normally expressed in euchromatin is experimentally relocated into a region of heterochromatin, it ceases to be expressed, and the gene is said to be *silenced*. These differences in gene expression are examples of **position effects**, in which the activity of a gene depends on its position relative to a nearby region of heterochromatin on a chromosome. First recognized in *Drosophila*, position effects have now been observed in many eucaryotes, including yeasts, plants, and humans.

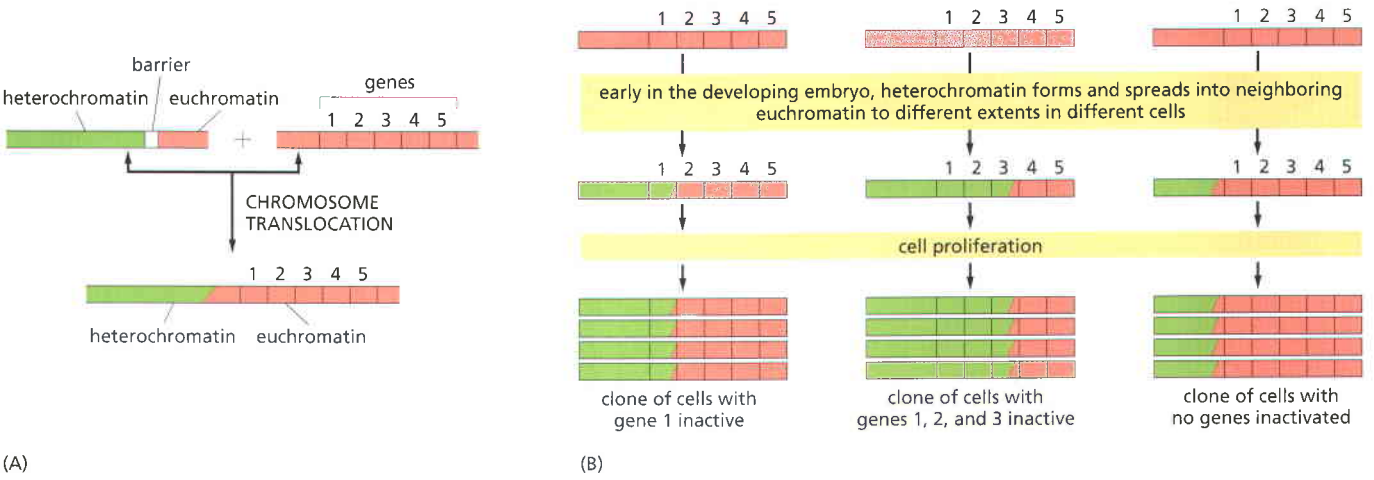


Figure 4–36 The cause of position effect variegation in *Drosophila*. (A) Heterochromatin *(green)* is normally prevented from spreading into adjacent regions of euchromatin *(red)* by special *barrier* DNA sequences, which we shall discuss shortly. In flies that inherit certain chromosomal rearrangements, however, this barrier is no longer present. (B) During the early development of such flies, heterochromatin can spread into neighboring chromosomal DNA, proceeding for different distances in different cells. This spreading soon stops, but the established pattern of heterochromatin is inherited, so that large clones of progeny cells are produced that have the same neighboring genes condensed into heterochromatin and thereby inactivated (hence the "variegated" appearance of some of these flies; see Figure 4–37). Although "spreading" is used to describe the formation of new heterochromatin close to previously existing heterochromatin, the term may not be wholly accurate. There is evidence that during expansion, heterochromatin can "skip over" some regions of chromatin, sparing the genes that lie within them from repressive effects.

The position effects associated with heterochromatin exhibit a feature called *position effect variegation*, which in retrospect provided critical clues concerning chromatin function. In *Drosophila*, chromosome breakage events that directly connect a region of heterochromatin to a region of euchromatin tend to inactivate the nearby euchromatic genes. The zone of inactivation spreads a different distance in different early cells in the fly embryo, but once the heterchromatic condition is established on a gene, it tends to be stably inherited by all of the cell's progeny (**Figure 4–36**). This remarkable phenomenon was first recognized through a detailed genetic analysis of the mottled loss of red pigment in the fly eye (**Figure 4–37**), but it shares many features with the extensive spread of heterochromatin that inactivates of one of the two X chromosomes in female mammals (see p. 473).

Extensive genetic screens have been carried out in *Drosophila*, as well as in fungi, in a search for gene products that either enhance or suppress the spread of heterochromatin and its stable inheritance—that is, for genes that when mutated serve as either enhancers or suppressors of position effect variegation. In this way, more than 50 genes have been identified that play a critical role in these processes. In recent years, the detailed characterization of the proteins produced by these genes has revealed that many are nonhistone chromosomal proteins that underlie a remarkable mechanism for eucaryotic gene control, one

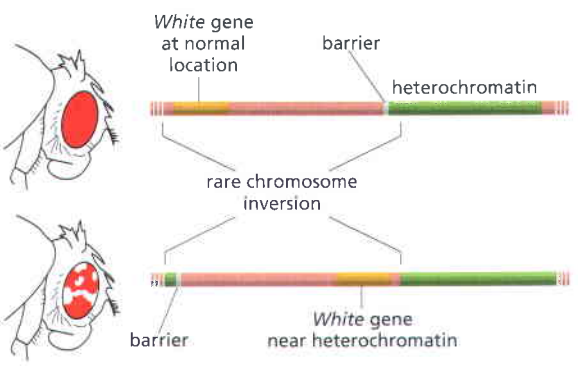


Figure 4–37 The discovery of position effects on gene expression. The *White* gene in the fruit fly *Drosophila* controls eye pigment production and is named after the mutation that first identified it. Wild-type flies with a normal *White* gene *(White$^+$)* have normal pigment production, which gives them *red* eyes, but if the *White* gene is mutated and inactivated, the mutant flies *(White$^-$)* make no pigment and have *white* eyes. In flies in which a normal *White$^+$* gene has been moved near a region of heterochromatin, the eyes are mottled, with both *red* and *white* patches. The *white* patches represent cell lineages in which the *White$^+$* gene has been silenced by the effects of the heterochromatin. In contrast, the *red* patches represent cell lineages in which the *White$^+$* gene is expressed. Early in development, when the heterochromatin is first formed, it spreads into neighboring euchromatin to different extents in different embryonic cells (see Figure 4–36). The presence of large patches of *red* and *white* cells reveals that the state of transcriptional activity, as determined by the packaging of this gene into chromatin in those ancestor cells, is inherited by all daughter cells.

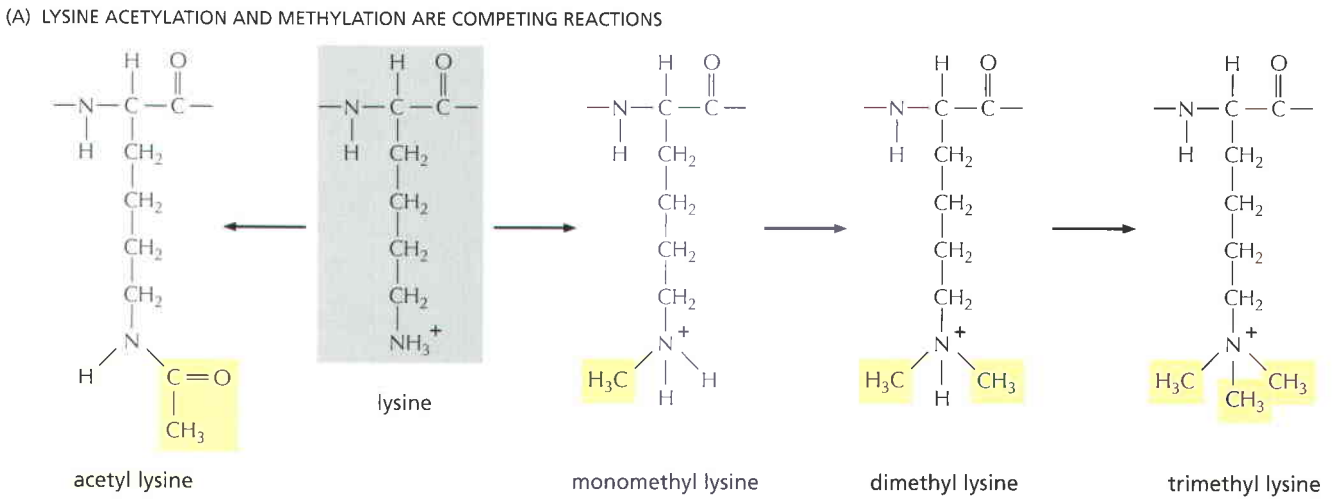
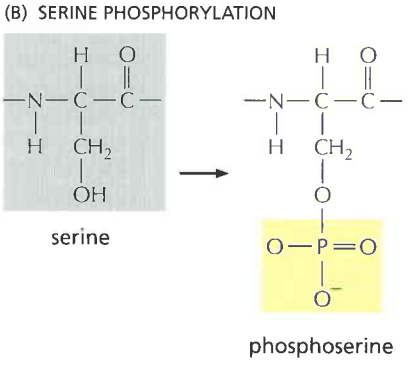


Figure 4–38 Some prominent types of covalent amino acid side-chain modifications found on nucleosomal histones. (A) Three different levels of lysine methylation are shown; each can be recognized by a different binding protein and thus each can have a different significance for the cell. Note that acetylation removes the plus charge on lysine, and that, most importantly, an acetylated lysine cannot be methylated, and *vice versa*. (B) Serine phosphorylation adds a negative charge to a histone. Modifications not shown here are the mono- or di-methylation of an arginine, the phosphorylation of a threonine, the addition of ADP-ribose to a glutamic acid, and the addition of a ubiquityl, sumoyl, or biotin group to a lysine.



that requires the precise amino acid sequences of the core histones. This mechanism of gene control therefore helps to explain the remarkably slow change in the histones over time.

The Core Histones Are Covalently Modified at Many Different Sites

The amino acid side chains of the four histones in the nucleosome core are subjected to a remarkable variety of covalent modifications, including the acetylation of lysines, the mono-, di-, and tri-methylation of lysines, and the phosphorylation of serines (**Figure 4–38**). A large number of these side-chain modifications occur on the eight relatively unstructured N-terminal "histone tails" that protrude from the nucleosome (**Figure 4–39**). However, there are also specific side-chain modifications on the nucleosome's globular core (**Figure 4–40**).

All of the above types of modifications are reversible. The modification of a particular amino acid side chain in a nucleosome is created by a specific enzyme, with most of these enzymes acting only on one or a few sites. A different enzyme is responsible for removing each side chain modification. Thus, for example, acetyl groups are added to specific lysines by a set of different histone acetyl transferases (HATs) and removed by a set of histone deacetylase complexes (HDACs). Likewise, methyl groups are added to lysine side chains by a set of different histone methyl transferases and removed by a set of histone demethylases. Each enzyme is recruited to specific sites on the chromatin at defined times in each cell's life history. For the most part, the initial recruitment of these enzymes depends on *gene regulatory proteins* that bind to specific DNA sequences along chromosomes, and these are produced at different times in the life of an organism, as described in Chapter 7. But in at least some cases, the covalent modifications on nucleosomes can persist long after the gene regulatory proteins that first induced them have disappeared, thereby carrying a memory in the cell of its developmental history. Very different patterns of covalent modifications are therefore found on different groups of nucleosomes, according to their exact position on a chromosome and the status of the cell.

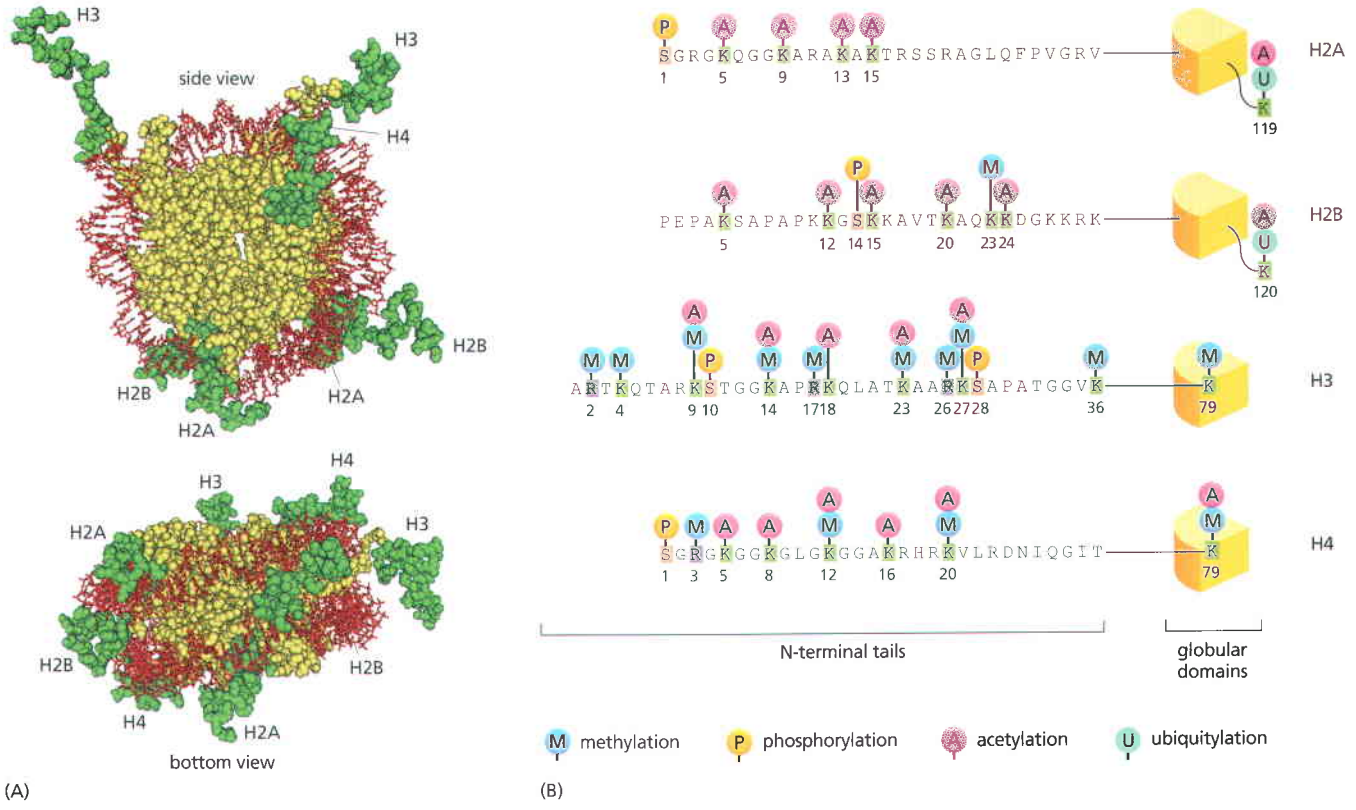


Figure 4–39 The covalent modification of core histone tails. (A) The structure of the nucleosome highlighting the location of the first 30 amino acids in each of its eight N-terminal histone tails *(green)*. (B) Well-documented modifications of the four histone core proteins are indicated. Although only a single symbol is used for methylation here (*M*), each lysine (*K*) or arginine (*R*) can be methylated in several different ways. Note also that some positions (e.g., lysine 9 of H3) can be modified either by methylation or by acetylation, but not both. Most of the modifications shown add a relatively small molecule onto the histone tails; the exception is ubiquitin, a 76 amino acid protein also used for other cell processes (see Figure 6–92). (Adapted from H. Santos-Rosa and C. Caldas, *Eur. J. Cancer* 41:2381–2402, 2005. With permission from Elsevier.)

The modifications of the histones are carefully controlled, and they have important consequences. The acetylation of lysines on the N-terminal tails tends to loosen chromatin structure, in part because adding an acetyl group to lysine removes its positive charge, thereby reducing the affinity of the tails for

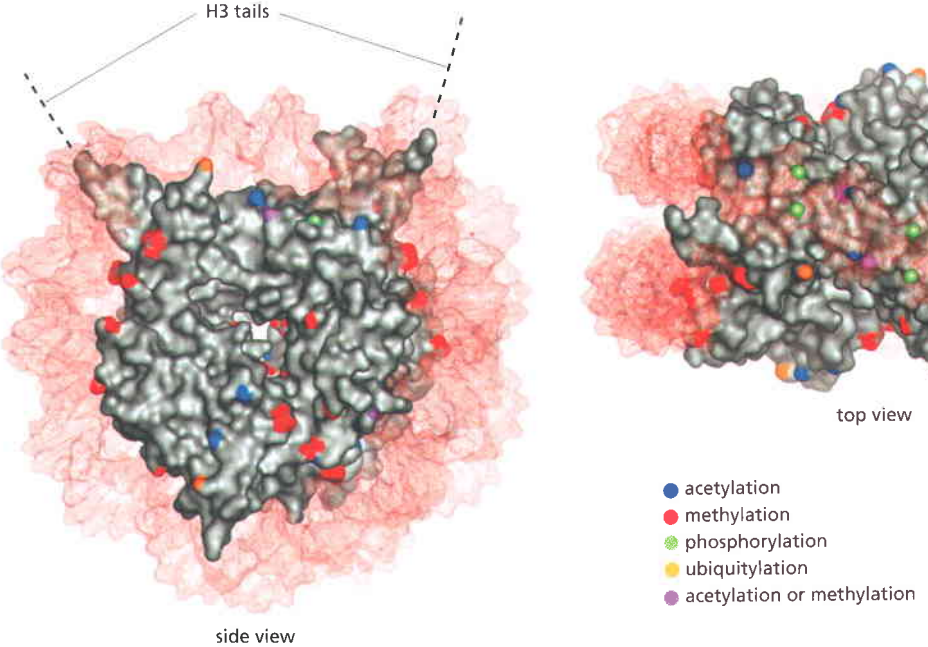


Figure 4–40 A map of histone modifications on the surface of the nucleosome core particle. As noted, the histone tails have been omitted here (compare with Figure 4–39). The functions of most of these core modifications are not yet known. (Adapted from M.S. Cosgrove, J.D. Boeke and C. Wolberger, *Nat. Struct. Mol. Biol.* 11:1037–1043, 2004. With permission from Macmillan Publishers Ltd.)

adjacent nucleosomes (see Figure 4–33). However, the most profound effect of the histone modifications is their ability to attract specific proteins to a stretch of chromatin that has been appropriately modified. These new proteins determine how and when genes will be expressed, as well as other biological functions. In this way, the precise structure of a domain of chromatin determines the expression of the genes packaged in it, and thereby the structure and function of the eucaryotic cell.

Chromatin Acquires Additional Variety Through the Site-Specific Insertion of a Small Set of Histone Variants

Despite the tight conservation of the amino acid sequences of the four core histones over hundreds of millions of years, eucaryotes also contain a few variant histones that assemble into nucleosomes. These histones are present in much smaller amounts than the major histones, and they have been less well conserved over long evolutionary times. Except for histone H4, variants exist for each of the core histones; some examples are shown in **Figure 4–41**.

The major histones are synthesized primarily during the S phase of the cell cycle (see Figure 17–4) and assembled into nucleosomes on the daughter DNA helices just behind the replication fork (see Figure 5–38). In contrast, most histone variants are synthesized throughout interphase. They are often inserted into already-formed chromatin, which requires a histone-exchange process catalyzed by the ATP-dependent chromatin remodeling complexes discussed previously. These remodeling complexes contain subunits that cause them to bind both to specific sites on chromatin and to histone chaperones that carry a particular variant. As a result, each histone variant is inserted into chromatin in a highly selective manner (see Figure 4–30).

The Covalent Modifications and the Histone Variants Act in Concert to Produce a "Histone Code" That Helps to Determine Biological Function

The number of possible distinct markings on an individual nucleosome is enormous. Even with the recognition that some of the covalent modifications are mutually exclusive (for example, it is not possible for a lysine to be both acetylated and methylated at the same time), and that other modifications are created together as a set, it is clear that thousands of combinations can exist. In addition, there is the further diversity created by nucleosomes that contain histone variants.

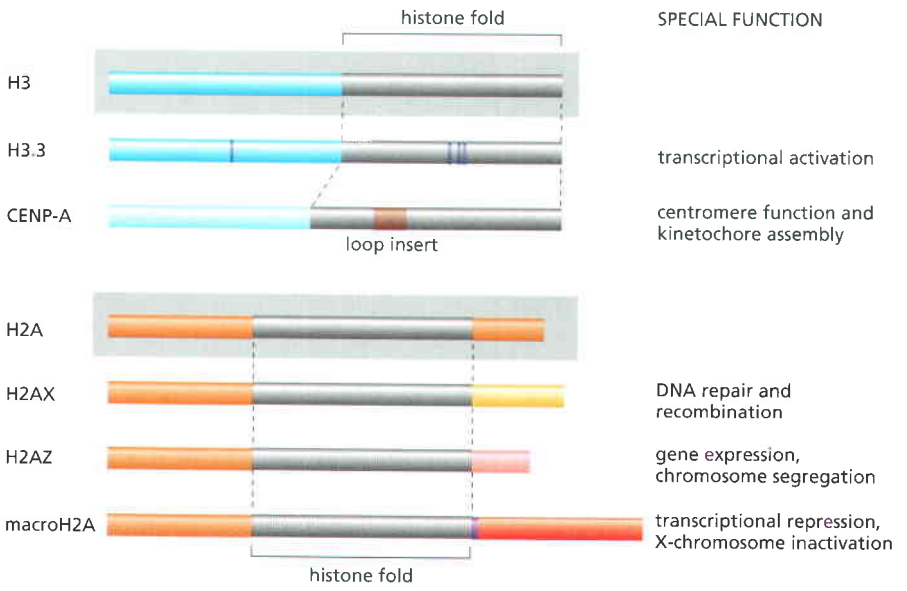
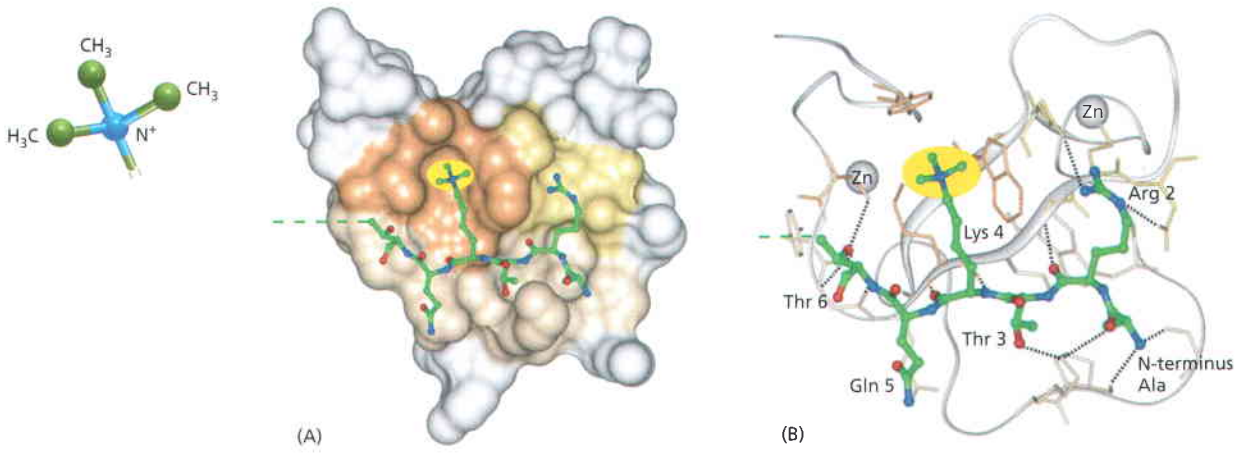


Figure 4–41 The structure of some histone variants compared with the major histone that they replace. These histones are inserted into nucleosomes at specific sites on chromosomes by ATP-dependent chromatin remodeling enzymes that act in concert with histone chaperones (see Figure 4–30). The CENP-A variant of histone H3 is discussed later in this chapter (see Figures 4–48 to 4–51); other variants are discussed in Chapter 7. The sequences that are colored differently in each variant are different from the corresponding sequence of the major histone. (Adapted from K. Sarma and D. Reinberg, *Nat. Rev. Mol. Cell. Biol.* 6:139–149, 2005. With permission from Macmillan Publishers Ltd.)



Many of the combinations appear to have a specific meaning for the cell because they determine how and when the DNA packaged in the nucleosomes is accessed, leading to the **histone code** hypothesis. For example, one type of marking signals that a stretch of chromatin has been newly replicated, another signals that the DNA in that chromatin has been damaged and needs repair, while many others signal when and how gene expression should take place. Small protein modules bind to specific marks, recognizing for example a trimethylated lysine 4 on histone H3 (**Figure 4–42**). These modules are thought to act in concert with other modules as part of a *code-reader complex*, so as to allow particular combinations of markings on chromatin to attract additional protein complexes that execute an appropriate biological function at the right time (**Figure 4–43**).

Figure 4–42 How each mark on a nucleosome is read. The structure of a protein module that specifically recognizes histone H3 trimethylated on lysine 4 is shown. (A) Space-filling model of an ING PHD domain bound to a histone tail (*green,* with the trimethyl group highlighted in *yellow).* (B) A ribbon model showing how the N-terminal six amino acids in the H3 tail are recognized. The *dashed lines* represent hydrogen bonds. This is one of many PHD domains that recognize methylated lysines on histones; different domains bind tightly to lysines located at different positions, and they can discriminate between a mono-, di-, and tri-methylated lysine. In a similar way, other small protein modules recognize specific histone side chains that have been marked with acetyl groups, phosphate groups, and so on. (Adapted from P.V. Pena et al., *Nature* 442:100–103, 2006. With permission from Macmillan Publishers Ltd.)

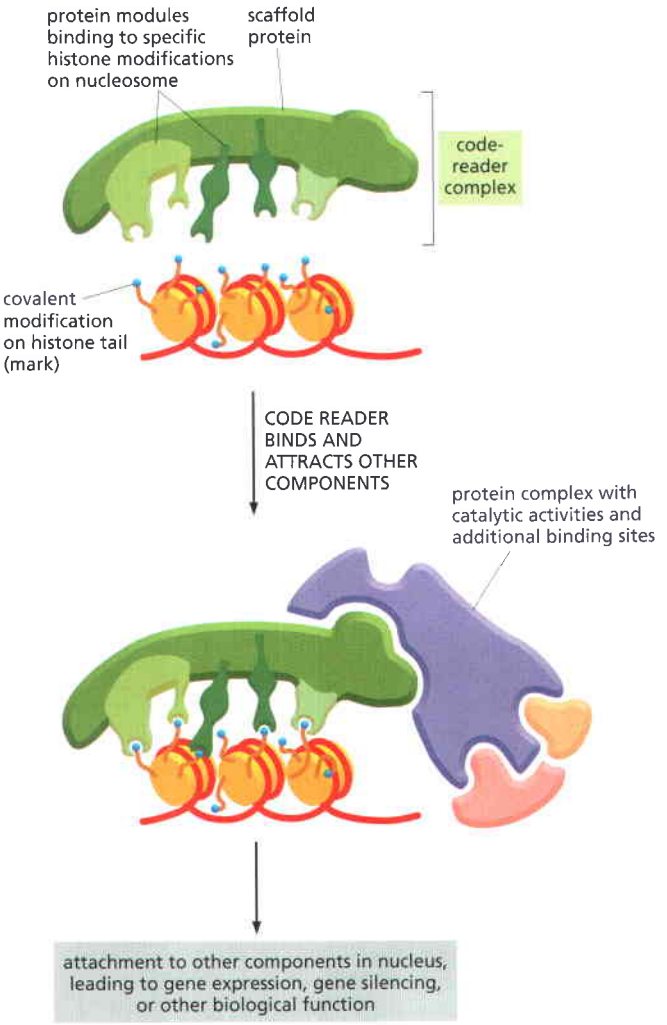


Figure 4–43 Schematic diagram showing how the histone code could be read by a code-reader complex. A large protein complex that contains a series of protein modules, each of which recognizes a specific histone mark, is schematically illustrated *(green).* This "code-reader complex" will bind tightly only to a region of chromatin that contains several of the different histone marks that it recognizes. Therefore, only a specific combination of marks will cause the complex to bind to chromatin and attract additional protein complexes *(purple)* that catalyze a biological function.

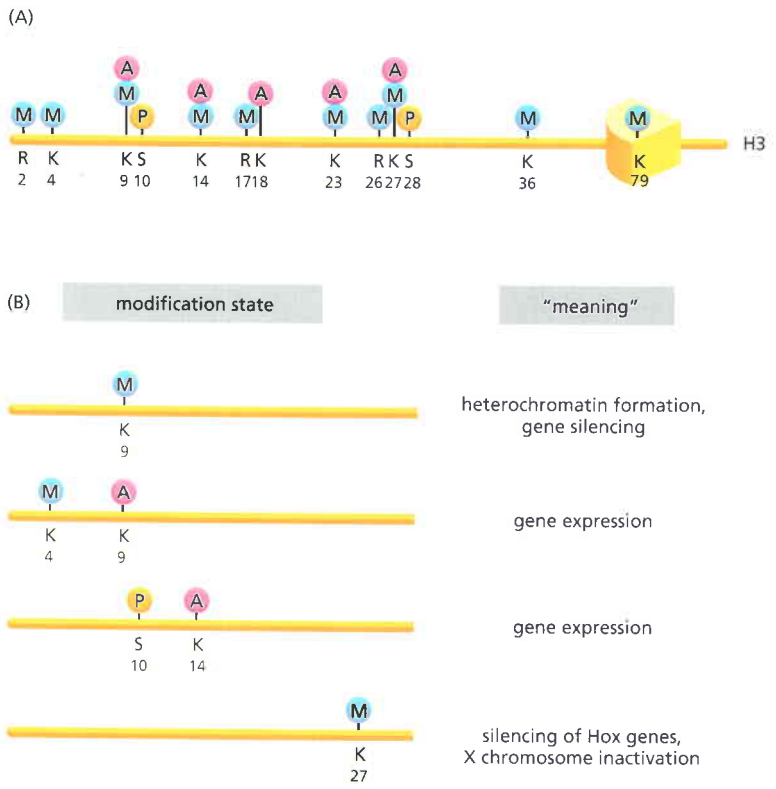


Figure 4–44 Some specific meanings of the histone code. (A) The modifications on the histone H3 N-terminal tail are shown, repeated from Figure 4–39. (B) The H3 tail can be marked by different combinations of modifications that convey a specific meaning to the stretch of chromatin where this combination occurs. Only a few of the meanings are known, including the four examples shown. To focus on just one example, the trimethylation of lysine 9 attracts the heterochromatin-specific protein HP1, which induces a spreading wave of further lysine 9 trimethylation followed by further HP1 binding, according to the general scheme that will be illustrated shortly (see Figure 4–46). Not shown is the fact that, as just implied (see Figure 4–43), reading the histone code generally involves the joint recognition of marks at other sites on the nucleosome along with the indicated H3 tail recognition. In addition, specific levels of methylation (mono-, di-, or tri-methyl groups) are required, as in Figure 4–42.

The marks on nucleosomes due to covalent additions to histones are dynamic, being constantly removed and added at rates that depend on their chromosomal locations. Because the histone tails extend outward from the nucleosome core and are likely to be accessible even when chromatin is condensed, they would seem to provide an especially suitable format for creating marks in a form that can be readily altered as a cell's needs change. Although much remains to be learned about the meaning of the many different histone code combinations, a few well-studied examples of the information that can be encoded in the histone H3 tail are listed in **Figure 4–44**.

A Complex of Code-reader and Code-writer Proteins Can Spread Specific Chromatin Modifications for Long Distances Along a Chromosome

The phenomenon of position effect variegation described previously requires that at least some modified forms of chromatin have the ability to spread for substantial distances along a chromosomal DNA molecule (see Figure 4–36). How is this possible?

The enzymes that modify (or remove modifications from) the histones in nucleosomes are part of multisubunit complexes. They can initially be brought to a particular region of chromatin by one of the sequence-specific DNA-binding proteins (gene regulatory proteins) discussed in Chapters 6 and 7 (for a specific example, see Figure 7–87). But after a modifying enzyme "writes" its mark on one or a few neighboring nucleosomes, events that resemble a chain reaction can ensue. In this case, the "code-writer" enzyme works in concert with a code-reader protein located in the same protein complex. This second protein contains a code-reader module that recognizes the mark and binds tightly to the newly modified nucleosome (see Figure 4–42), positioning its attached writer enzyme near an adjacent nucleosome. Through many such read–write cycles, the reader protein can carry the writer enzyme along the DNA—spreading the mark in a hand-over-hand manner along the chromosome (**Figure 4–45**).

In reality, the process is more complicated than the scheme just described. Both readers and writers are part of a protein complex that is likely to contain

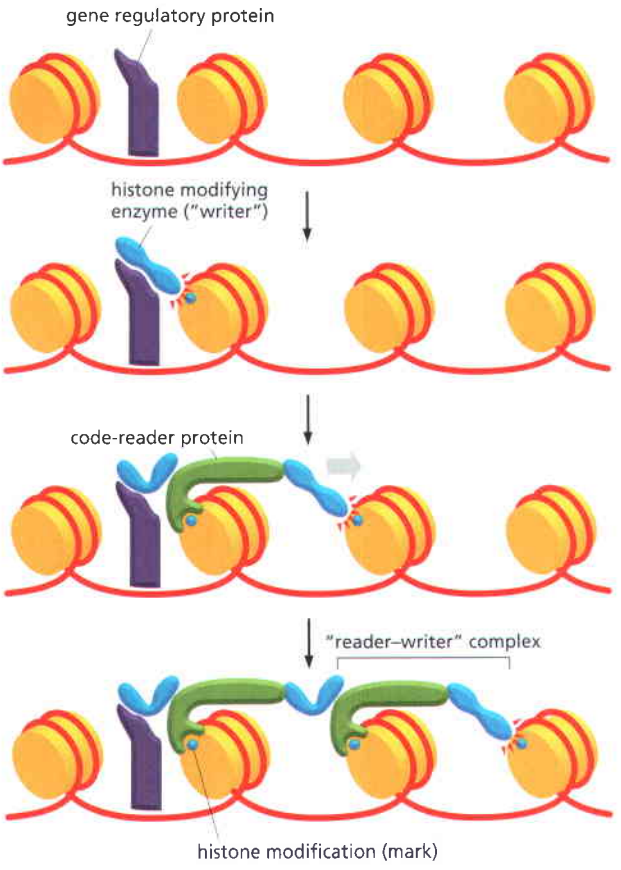


Figure 4–45 How the recruitment of a code-reader–writer complex can spread chromatin changes along a chromosome. The code-writer is an enzyme that creates a specific modification on one or more of the four nucleosomal histones. After its recruitment to a specific site on a chromosome by a gene regulatory protein, the writer collaborates with a code-reader protein to spread its mark from nucleosome to nucleosome by means of the indicated reader–writer complex. For this mechanism to work, the reader must recognize the same histone modification mark that the writer produces (see also Figure 4–43).

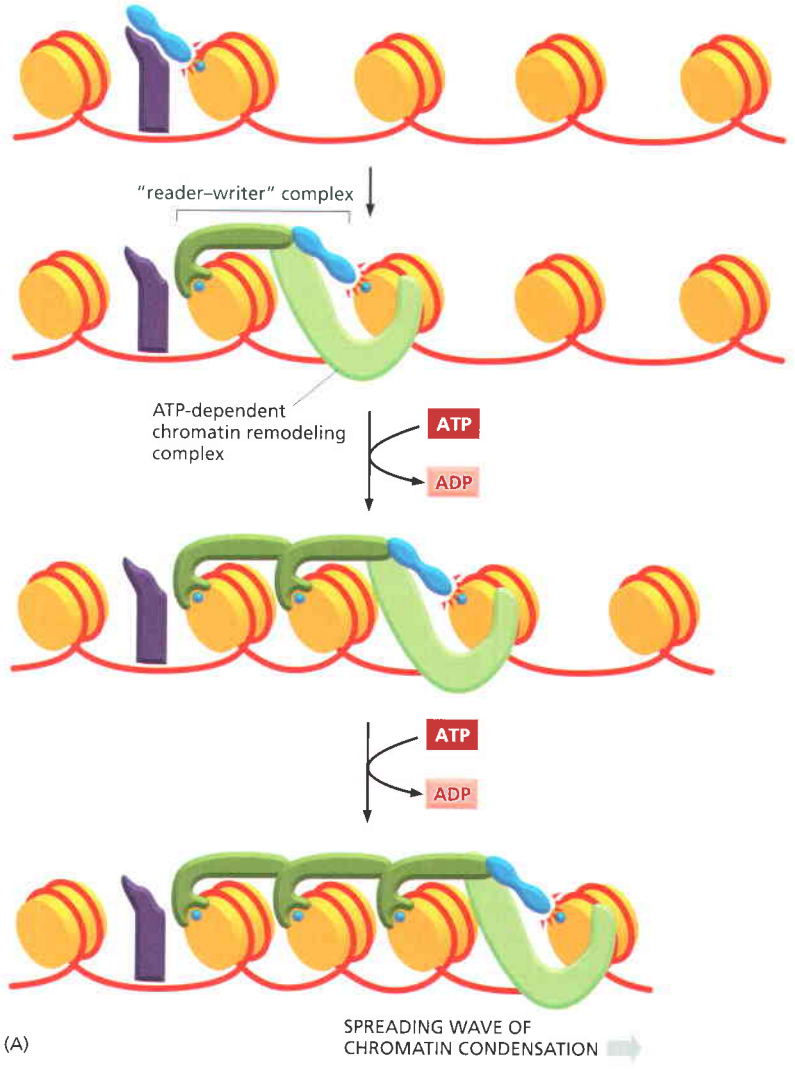
multiple readers and writers, and to require multiple marks on the nucleosome to spread. Moreover, many of these reader–writer complexes also contain an ATP-dependent chromatin remodeling protein, and the reader, writer, and remodeling proteins work in concert to either decondense or condense long stretches of chromatin as the reader moves progressively along the nucleosome-packaged DNA (**Figure 4–46**).

Some idea of the complexity of the processes just described can be derived from the results of genetic screens for mutant genes that either enhance or suppress the spreading and stability of heterochromatin in tests for position effect variegation in *Drosophila* (see Figure 4–37). As pointed out previously, more than 50 such genes are known, and most of them are likely to function as subunits in one or more reader–writer-remodeling protein complexes.

Barrier DNA Sequences Block the Spread of Reader–Writer Complexes and thereby Separate Neighboring Chromatin Domains

The above mechanism for spreading chromatin structures raises a potential problem. Inasmuch as each chromosome consists of one continuous, very long DNA molecule, what prevents a cacophony of confusing cross-talk between adjacent chromatin domains of different structure and function? Early studies of position effect variegation had suggested an answer: the existence of specific DNA sequences that separate one chromatin domain from another (see Figure 4–37). Several such *barrier* sequences have now been identified and characterized through the use of genetic engineering techniques that allow specific regions of DNA sequence to be deleted or added to chromosomes.

For example, a sequence called HS4 normally separates the active chromatin domain that contains the β-globin locus from an adjacent region of silenced, condensed chromatin in erythrocytes (see Figure 7–61). If this sequence is deleted, the β-globin locus is invaded by condensed chromatin. This chromatin silences the genes it covers, and it spreads to a different extent in different cells, causing a pattern of position effect variegation similar to that



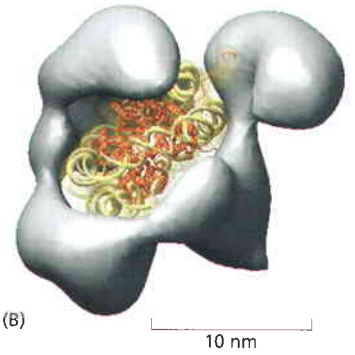


Figure 4–46 How a complex containing reader–writer and ATP-dependent chromatin remodeling proteins can spread chromatin changes along a chromosome. (A) A spreading wave of chromatin condensation. This mechanism is identical to that in Figure 4–45, except that the reader–writer complex collaborates with an ATP-dependent chromatin remodeling protein (see Figure 4–29) to reposition nucleosomes and pack them into highly condensed arrays. This is a highly simplified view of the mechanism known to be able to spread a major form of heterochromatin for long distances along chromosomes (see Figure 4–36). The heterochromatin-specific protein HP1 plays a major role in that process. HP1 binds to trimethyl lysine 9 on histone H3, and it remains associated with the condensed chromatin as one of the readers in a reader–writer-remodeling complex that, while incompletely understood, is considerably more intricate than that shown here. (B) The actual structure of a chromatin reader-remodeling complex, showing how it is thought to interact with a nucleosome. Modeled in *gray* is the yeast RSC complex, which contains 15 subunits—including an ATP-dependent chromatin remodeling protein and at least 4 subunits with code-reader domains. (B, from A.E. Leschziner et al., *Proc. Natl Acad. Sci. U.S.A.* 104:4913–4918, 2007. With permission from National Academy of Sciences.)

observed in *Drosophila*. As described in Chapter 7, this invasion has dire consequences: the globin genes are poorly expressed, and individuals who carry such a deletion have a severe form of anemia.

The HS4 sequence is often added to both ends of a gene that is experimentally inserted into a mammalian genome, in order to protect that gene from the silencing caused by spreading heterochromatin. Analysis of this barrier sequence reveals that it contains a cluster of binding sites for histone acetylase enzymes. Since the acetylation of a lysine side chain is incompatible with the methylation of the same side chain, histone acetylases and histone deacetylases are logical candidates for the formation of barriers on the DNA that block the spread of different forms of chromatin (**Figure 4–47**). However, several other types of chromatin modifications are known that can also protect genes from silencing.

The Chromatin in Centromeres Reveals How Histone Variants Can Create Special Structures

The presence of nucleosomes carrying histone variants is thought to produce marks in chromatin that are unusually long lasting. Consider, for example, the formation and inheritance of the chromatin that forms on centromeres, the DNA region of each chromosome required for the orderly segregation of the chromosomes into daughter cells each time a cell divides (see Figure 4–21). In many complex organisms, including humans, each centromere is embedded in a stretch of special *centric heterochromatin* that persists throughout interphase, even though the centromere-mediated movement of DNA occurs only during

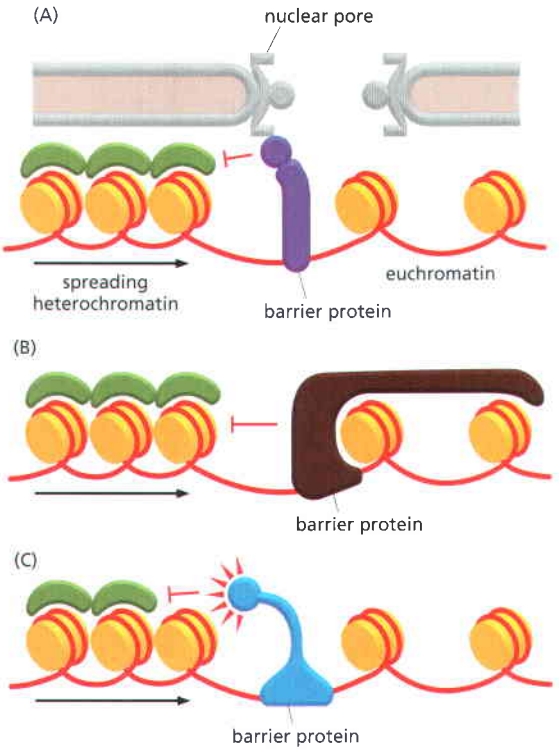


Figure 4–47 Some mechanisms of barrier action. These models are derived from different analyses of barrier action, and a combination of several of them may function at any one site. (A) The tethering of a region of chromatin to a large fixed site, such as the nuclear pore complex illustrated here, can form a barrier that stops the spread of heterochromatin. (B) The tight binding of barrier proteins to a group of nucleosomes can compete with heterochromatin spreading. (C) By recruiting a group of highly active histone-modifying enzymes, barriers can erase the histone marks that are required for heterochromatin to spread. For example, a potent acetylation of lysine 9 on histone H3 will compete with lysine 9 methylation, thereby preventing the HP1 protein binding needed to form some forms of heterochromatin (see Figure 4–46). (Based on A.G. West and P. Fraser, *Hum. Mol. Genet.* 14:R101–R111, 2005. With permission from Oxford University Press.)

mitosis. This chromatin contains a centromere-specific variant H3 histone, known as CENP-A (see Figure 4–41), plus additional proteins that pack the nucleosomes into particularly dense arrangements and form the kinetechore, the special structure required for attachment of the mitotic spindle.

A specific DNA sequence of approximately 125 nucleotide pairs is sufficient to serve as a centromere in the yeast *S. cerevisiae.* Despite its small size, more than a dozen different proteins assemble on this DNA sequence; the proteins include the CENP-A histone H3 variant, which, along with the three other core histones, forms a centromere-specific nucleosome. The additional proteins at the yeast centromere attach this nucleosome to a single microtubule from the yeast mitotic spindle (**Figure 4–48**).

The centromeres in more complex organisms are considerably larger than those in budding yeasts. For example, fly and human centromeres extend over hundreds of thousands of nucleotide pairs and do not seem to contain a centromere-specific DNA sequence. These centromeres largely consist of short, repeated DNA sequences, known as *alpha satellite DNA* in humans. But the same repeat sequences are also found at other (non-centromeric) positions on

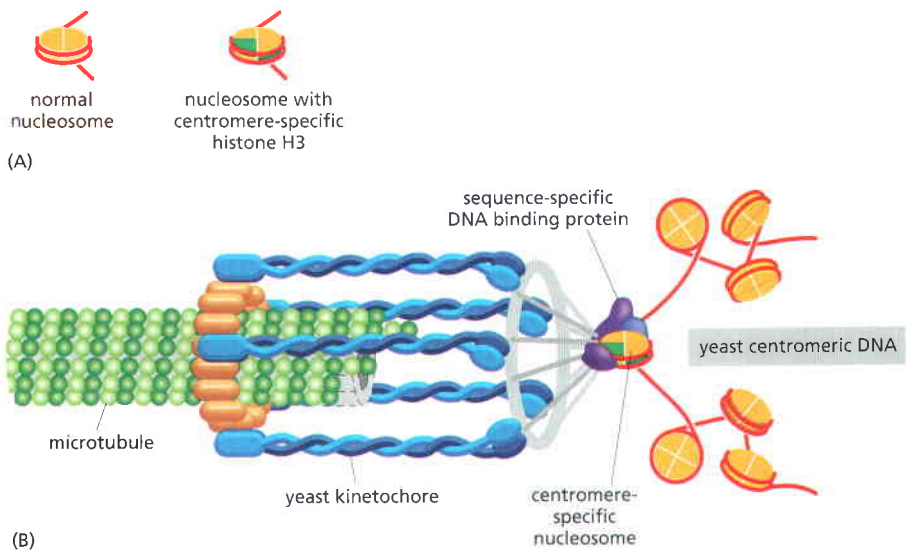


Figure 4–48 A model for the structure of a simple centromere. In the yeast *Saccharomyces cerevisiae,* a special centromeric DNA sequence assembles a single nucleosome in which two copies of an H3 variant histone (called CENP-A in most organisms) replaces the normal H3. Peptide sequences unique to this variant histone (see Figure 4–41) then help to assemble additional proteins, some of which form a kinetochore. This kinetochore is unusual in capturing only a single microtubule; humans have much larger centromeres and form kinetochores that can capture 20 or more microtubules (see Figure 4–50). The kinetochore is discussed in detail in Chapter 17. (Adapted from A. Joglekar et al., *Nat. Cell Biol.* 8:381–383, 2006. With permission from Macmillan Publishers Ltd.)

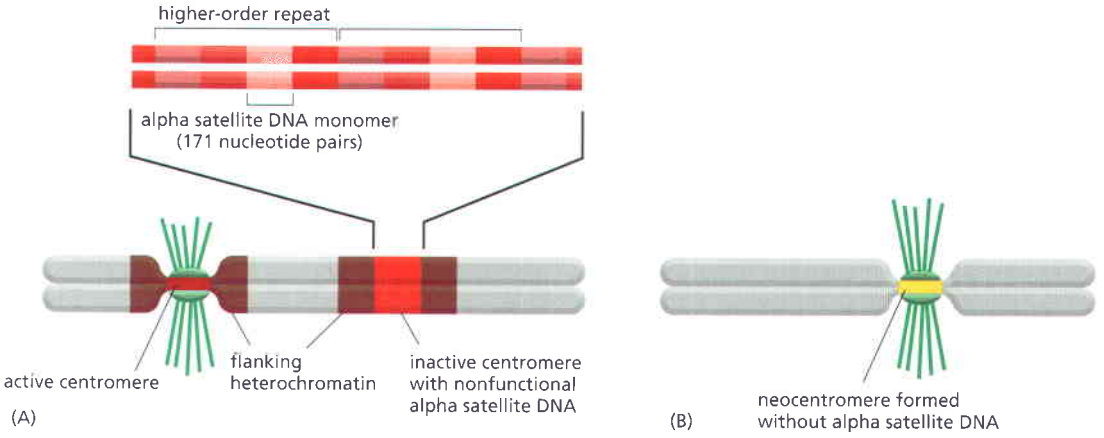


Figure 4–49 Evidence for the plasticity of human centromere formation. (A) A series of A-T-rich alpha satellite DNA sequences are repeated many thousands of times at each human centromere *(red)*, surrounded by pericentric heterochromatin *(brown)*. However, due to an ancient chromosome breakage and rejoining event, some human chromosomes contain two blocks of alpha satellite DNA, each of which presumably functioned as a centromere in its original chromosome. Usually, these dicentric chromosomes are not stably propagated because they attach improperly to the spindle and are broken apart during mitosis. In chromosomes that do survive, however, one of the centromeres has somehow inactivated, even though it contains all the necessary DNA sequences. This allows the chromosome to be stably propagated. (B) In a small fraction (1/2000) of human births, extra chromosomes are observed in cells of the offspring. Some of these extra chromosomes, which have formed from a breakage event, lack alpha satellite DNA altogether, yet new centromeres (neocentromeres) have arisen from what was originally euchromatic DNA.

chromosomes, indicating that they are not sufficient to direct centromere formation. Most strikingly, in some unusual cases, new human centromeres (called neocentromeres) have been observed to form spontaneously on fragmented chromosomes. Some of these new positions were originally euchromatic and lack alpha satellite DNA altogether (**Figure 4–49**).

It therefore seems that centromeres in complex organisms are defined by an assembly of proteins, instead of by a specific DNA sequence. When antibodies that stain specific modified nucleosomes are used to examine the stretched chromosome fibers from centromeres, one observes striking alternation of two modified forms of chromatin (**Figure 4–50**). It appears that this arrangement allows the centric heterochromatin to fold so as to position the CENP-A-containing nucleosomes on the outside of the mitotic chromosome, where they bind the set of proteins that form the kinetechore plates. These plates in turn capture a group of microtubules from the mitotic spindle in order to partition the chromosomes accurately, as described in Chapter 17.

Chromatin Structures Can Be Directly Inherited

To explain the above observations, it has been proposed that *de novo* centromere formation requires an initial seeding event, involving the formation of a specialized DNA–protein structure that contains nucleosomes formed with the CENP-A variant of histone H3. In humans, this seeding event happens more readily on arrays of alpha satellite DNA than on other DNA sequences. The H3–H4 tetramers from each nucleosome on the parental DNA helix are directly inherited by the daughter DNA helices at a replication fork (see Figure 5–38). Therefore, once a set of CENP-A-containing nucleosomes has been assembled on a stretch of DNA, it is easy to understand how a new centromere could be generated in the same place on both daughter chromosomes following each round of cell division (**Figure 4–51**).

The plasticity of centromeres may provide an important evolutionary advantage. We have seen that chromosomes evolve in part by breakage and rejoining events (see Figure 4–18). Many of these events produce chromosomes with two centromeres, or chromosome fragments with no centromeres at all. Although rare, both the inactivation of centromeres and their ability to be activated *de novo*

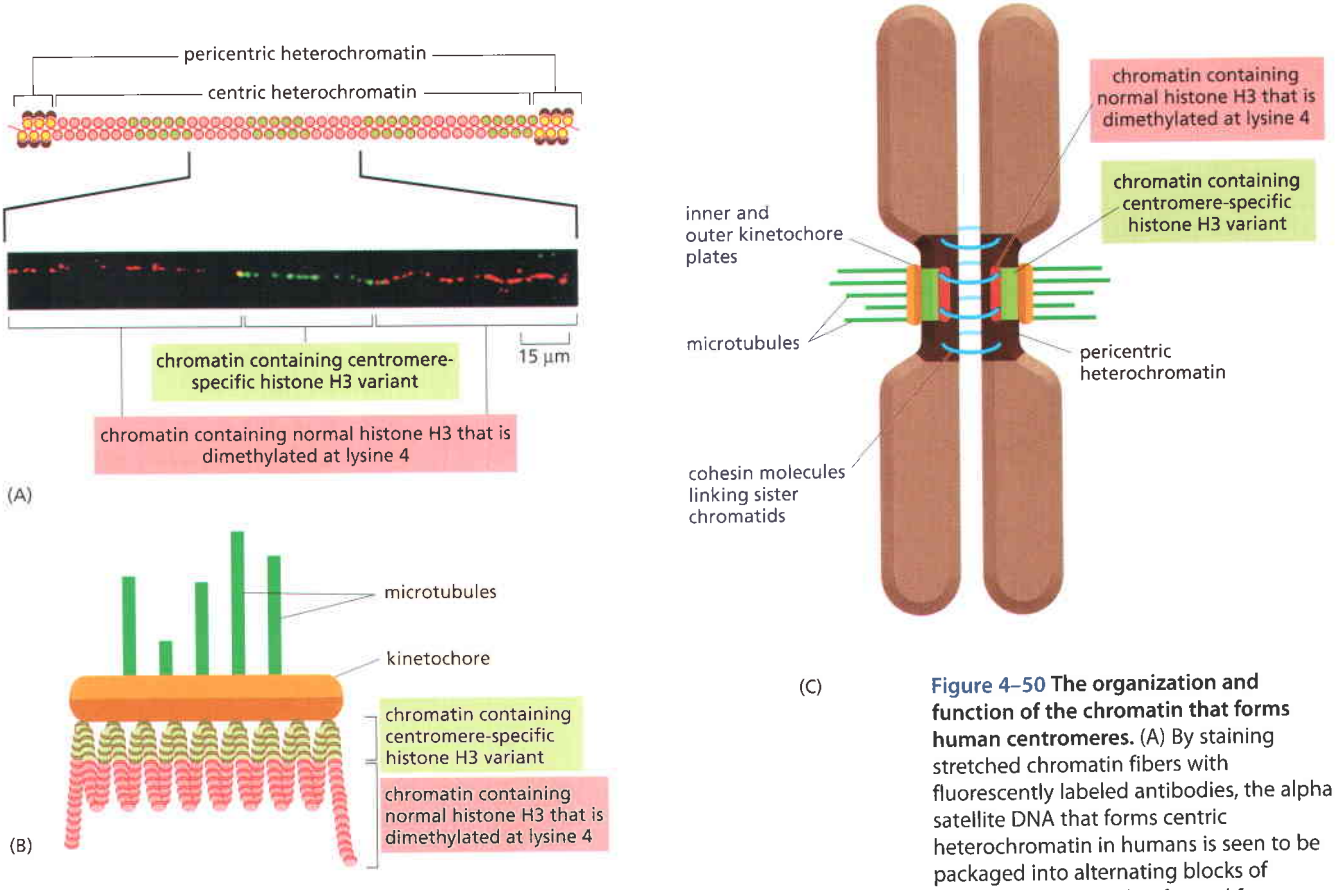


Figure 4–50 The organization and function of the chromatin that forms human centromeres. (A) By staining stretched chromatin fibers with fluorescently labeled antibodies, the alpha satellite DNA that forms centric heterochromatin in humans is seen to be packaged into alternating blocks of chromatin. One block is formed from a long string of nucleosomes containing the CENP-A H3 variant histone *(green)*; the other block contains nucleosomes that are specially marked with a dimethyl lysine 4 *(red)*. Each block is more than a thousand nucleosomes long. (B) A model for the organization of the two types of centric heterochromatin. As in yeast, the nucleosomes that contain the H3 variant histone form the kinetochore. (C) The arrangement of the centric and pericentric heterochromatin on a human metaphase chromosome, as determined by fluorescence microscopy using the same antibodies as in (A). (Adapted from B.A. Sullivan and G.H. Karpen, *Nat. Struct. Mol. Biol.* 11:1076–1083, 2004. With permission from Macmillan Publishers Ltd.)

may occasionally allow newly formed chromosomes to be maintained stably, thereby facilitating the process of chromosome evolution.

There are some striking similarities between the formation and maintenance of centromeres and the formation and maintenance of other regions of heterochromatin. In particular, the entire centromere forms as an all-or-none entity, suggesting a highly cooperative addition of proteins after a seeding event. Moreover, once formed, the structure seems to be directly inherited on the DNA as part of each round of chromosome replication.

Chromatin Structures Add Unique Features to Eucaryotic Chromosome Function

Although a great deal remains to be learned about the functions of different chromatin structures, the packaging of DNA into nucleosomes was probably crucial for the evolution of eucaryotes like ourselves. Complex multicellular organisms would appear to be possible only if the cells in different lineages can specialize by changing the accessibility and responsiveness of many hundreds of genes to genetic readout. As described in Chapter 22, each cell has a stored memory of its past developmental history in the regulatory circuits that control its many genes.

Although bacteria also require cell memory mechanisms, the complexity of the memory circuits required by higher eucaryotes is unprecedented. The packaging of selected regions of eucaryotic genomes into different forms of chromatin makes possible a type of cell memory mechanism that is not available to bacteria. The crucial feature of this uniquely eucaryotic form of gene regulation is the storage of the memory of the state of a gene on a gene-by-gene basis—in the form of local chromatin structures that can persist for various lengths of time. At one extreme are structures like centric heterochromatin that, once established, are stably inherited from one cell generation to the next (see Figure

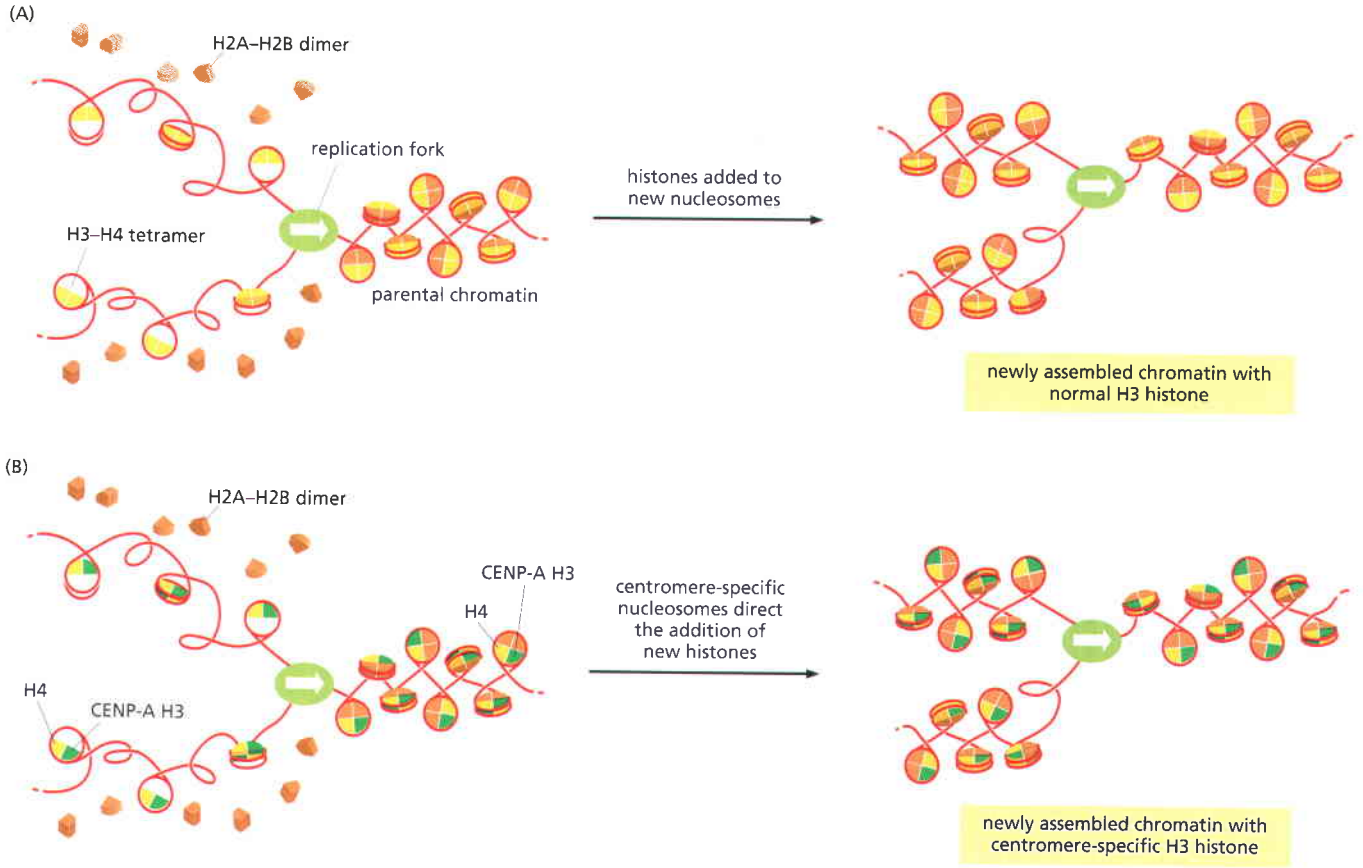


Figure 4–51 A model for the direct inheritance of centromeric heterochromatin. (A) The normal assembly of chromatin on the two daughter DNA helices produced at a replication fork requires the deposition of H2A–H2B dimers onto directly inherited H3–H4 tetramers, as well as the assembly of new histone octamers (see Figure 5–38 for details). (B) At a centromere, the inheritance of H3 variant–H4 tetramers seeds the formation of new histone octamers that likewise contain the variant H3 histone. A similar seeding process could cause the adjacent blocks of centric heterochromatin (containing H3 modified at dimethyl lysine 4; see Figure 4–50) to be inherited. Although the details are not known, the seeding process is likely to involve other centromeric proteins that are inherited along with the nucleosomes (see Figure 4–52).

4–51). Closely related mechanisms that are likewise based on the direct inheritance of parental forms of chromatin by the daughter DNA helices behind the replication fork are thought to be responsible for other types of condensed chromatin (**Figure 4–52**). For example, the permanently silenced, classical type

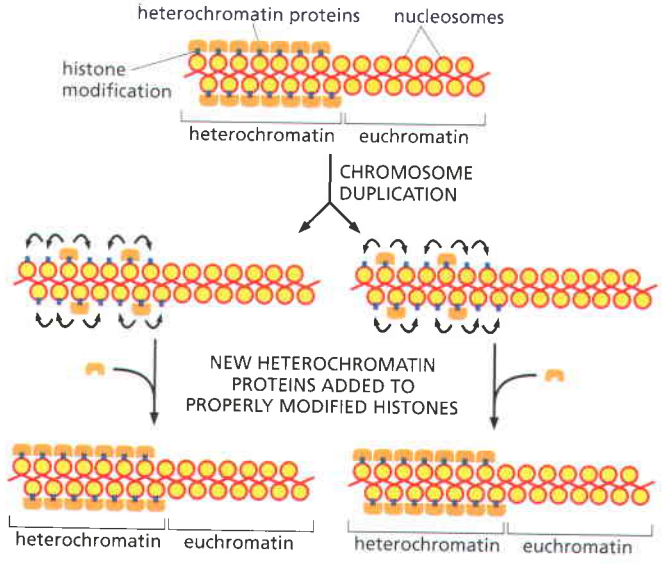


Figure 4–52 How the packaging of DNA in chromatin can be inherited during chromosome replication. In this model, some of the specialized chromatin components are distributed to each daughter chromosome after DNA duplication, along with the specially marked nucleosomes that they bind. After DNA replication, the inherited nucleosomes that are specially modified, acting in concert with the inherited chromatin components, change the pattern of histone modification on the newly formed daughter nucleosomes nearby. This creates new binding sites for the same chromatin components, which then assemble to complete the structure. The latter process is likely to involve code reader–writer-remodeling complexes operating in a manner similar to that previously illustrated in Figure 4–46.

of heterochromatin contains the HP1 protein, whereas the condensed chromatin that coats important developmental regulatory genes is maintained by the polycomb group of proteins. The latter type of heterochromatin silences a large number of genes that encode gene regulatory proteins early in embryonic development, covering a total of about 2 percent of the human genome, and it is removed only when each individual gene is needed by the developing organism (discussed in Chapter 22). Although other types of inherited chromatin structures exist, it is not yet clear how many different types there are: the number could certainly exceed 10 (see p. 238). The fundamental importance of this mechanism for distinguishing different genes is schematically represented in (**Figure 4–53**).

Other forms of chromatin can have a shorter lifetime, much less than the division time of the cell; however, many have a built-in persistence that helps to mediate biological function.

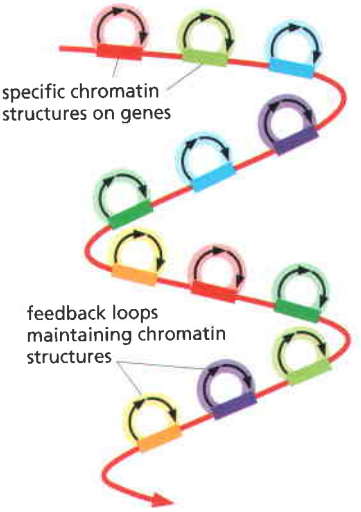


Figure 4–53 Schematic illustration of cell memory stored as chromatin-based epigenetic information in the genes of eucaryotes. Genes in eucaryotic cells can be packaged into a large variety of different chromatin structures, indicated here by different colors. At least some of these chromatin structures have a special effect on gene expression that can be directly inherited as epigenetic information when a cell divides. This allows some of the gene regulatory proteins that create different gene states to act only once, inasmuch as the state can be remembered after the regulatory protein is gone. Epigenetic information can also be stored in networks of signaling molecules that control gene expression (see Figure 7–86).

Summary

Despite the uniform assembly of chromosomal DNA into nucleosomes, a large variety of different chromatin structures are possible in eucaryotic organisms. This variety is based on a large set of reversible covalent modifications of the four histones in the nucleosome core. These modifications include the mono-, di-, and tri-methylation of many different lysine side chains, an important reaction that is incompatible with the acetylation of the same lysines. Specific combinations of the modifications mark each nucleosome with a histone code. The histone code is read when protein modules that are part of a larger protein complex bind to the modified nucleosomes in a region of chromatin. These code-reader proteins then attract additional proteins that catalyze biologically relevant functions.

Some code-reader protein complexes contain a histone-modifying enzyme, such as a histone methylase, that "writes" the same mark that the code-reader recognizes. A reader–writer-remodeling complex of this type can spread a specific form of chromatin for long distances along a chromosome. In particular, large regions of condensed heterochromatin are thought to be formed in this way. Heterochromatin is commonly found around centromeres and near telomeres, but it is also present at many other positions in chromosomes. The tight packaging of DNA into heterochromatin usually silences the genes within it.

The phenomenon of position effect variegation provides good evidence for the direct inheritance of condensed forms of chromatin by the daughter DNA helices formed at a replication fork, and a similar mechanism appears to be responsible for maintaining the specialized chromatin at centromeres. More generally, the ability to transmit specific chromatin structures from one cell generation to the next provides the basis for an epigenetic cell memory process that is likely to be critical for maintaining the complex set of different cell states required by complex multicellular organisms.

THE GLOBAL STRUCTURE OF CHROMOSOMES

Having discussed the DNA and protein molecules from which the 30-nm chromatin fiber is made, we now turn to the organization of the chromosome on a more global scale. As a 30-nm fiber, the typical human chromosome would still be 0.1 cm in length and able to span the nucleus more than 100 times. Clearly, there must be a still higher level of folding, even in interphase chromosomes. Although its molecular basis is still largely a mystery, this higher-order packaging almost certainly involves the folding of the 30-nm fiber into a series of loops and coils. This chromatin packing is fluid, frequently changing in response to the needs of the cell.

We shall begin by describing some unusual interphase chromosomes that can be easily visualized, inasmuch as certain features of these exceptional cases are thought to be representative of all interphase chromosomes. Moreover, they

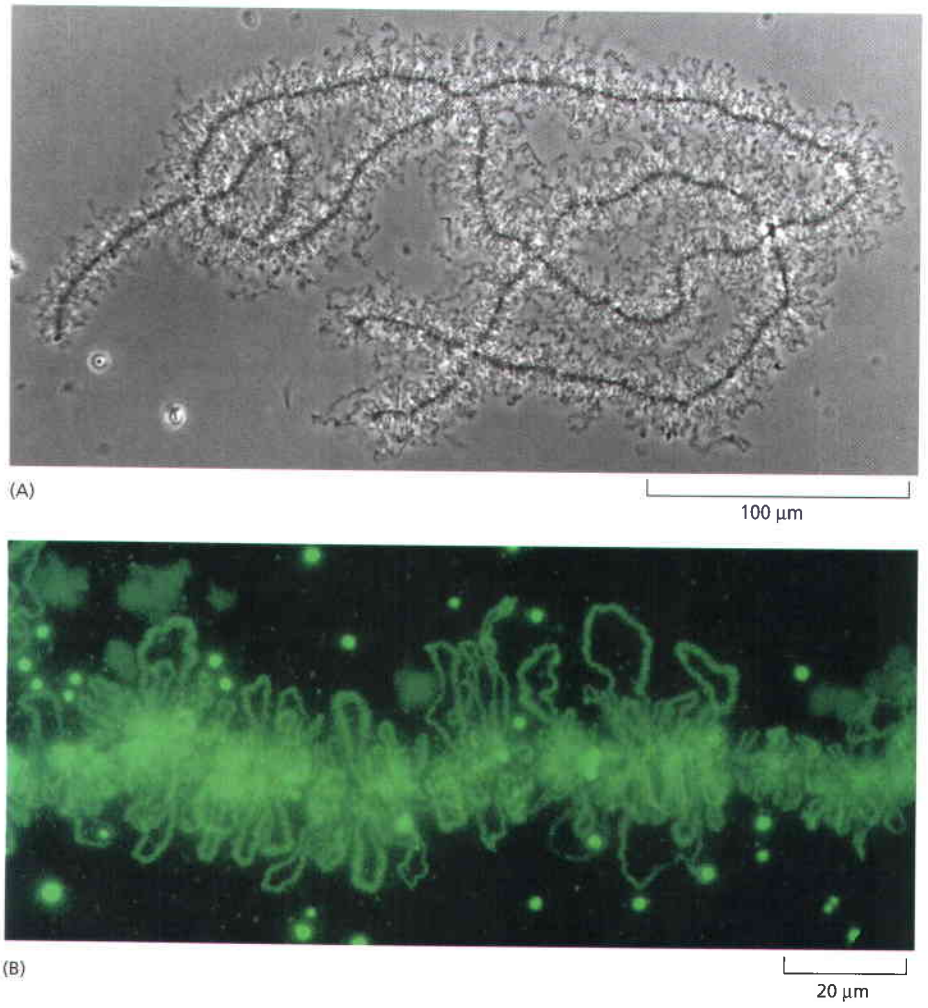


Figure 4–54 Lampbrush chromosomes. (A) A light micrograph of lampbrush chromosomes in an amphibian oocyte. Early in oocyte differentiation, each chromosome replicates to begin meiosis, and the homologous replicated chromosomes pair to form this highly extended structure containing a total of four replicated DNA molecules, or chromatids. The lampbrush chromosome stage persists for months or years, while the oocyte builds up a supply of materials required for its ultimate development into a new individual. (B) An enlarged region of a similar chromosome, stained with a fluorescent reagent that makes the loops active in RNA synthesis clearly visible. (Courtesy of Joseph G. Gall.)

provide a unique means for investigating some fundamental aspects of chromatin structure raised in the previous section. Next we describe how a typical interphase chromosome is arranged in the cell nucleus, focusing on human cells. Finally, we conclude by discussing the additional tenfold compaction that interphase chromosomes undergo during the process of mitosis.

Chromosomes Are Folded into Large Loops of Chromatin

Insight into the structure of the chromosomes in interphase cells has been obtained from studies of the stiff and extended meiotically paired chromosomes in growing amphibian oocytes (immature eggs). These very unusual **lampbrush chromosomes** (the largest chromosomes known) are clearly visible even in the light microscope, where they are seen to be organized into a series of large chromatin loops emanating from a linear chromosomal axis (**Figure 4–54**).

The organization of a lampbrush chromosome is illustrated in **Figure 4–55**. A given loop always contains the same DNA sequence, and it remains extended in the same manner as the oocyte grows. These chromosomes are producing large amounts of RNA for the oocyte, and most of the genes present in the DNA loops are being actively expressed. The majority of the DNA, however, is not in loops but remains highly condensed in the *chromomeres* on the axis, where genes are generally not expressed.

It is thought that the interphase chromosomes of all eucaryotes are similarly arranged in loops. Although these loops are normally too small and fragile to be easily observed in a light microscope, other methods can be used to infer their presence. For example, it has become possible to assess the frequency with

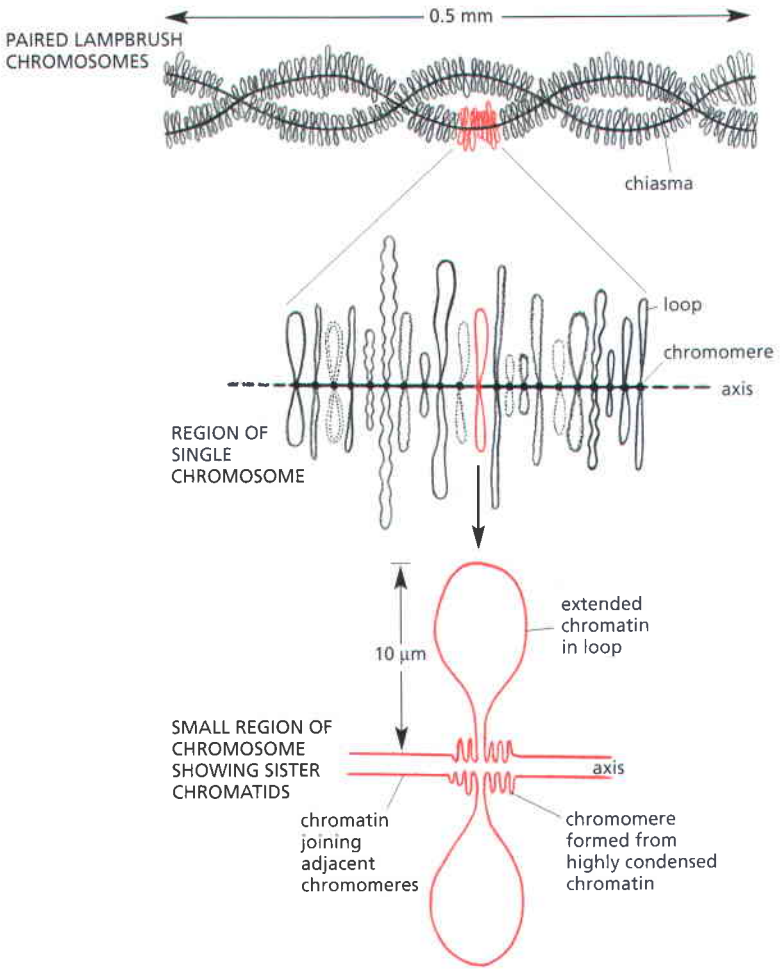


Figure 4–55 A model for the structure of a lampbrush chromosome. The set of lampbrush chromosomes in many amphibians contains a total of about 10,000 chromatin loops, although most of the DNA in each chromosome remains highly condensed in the chromomeres. Each loop corresponds to a particular DNA sequence. Four copies of each loop are present in each cell, since each of the two major units shown at the top consists of two closely apposed, newly replicated chromosomes. This four-stranded structure is characteristic of this stage of development of the oocyte, the diplotene stage of meiosis; see Figure 21–9.

which any two loci along an interphase chromosome are paired with each other, thus revealing likely candidates for the sites on chromatin that form the closely apposed bases of loop structures (**Figure 4–56**). These experiments and others suggest that the DNA in human chromosomes is organized into loops of

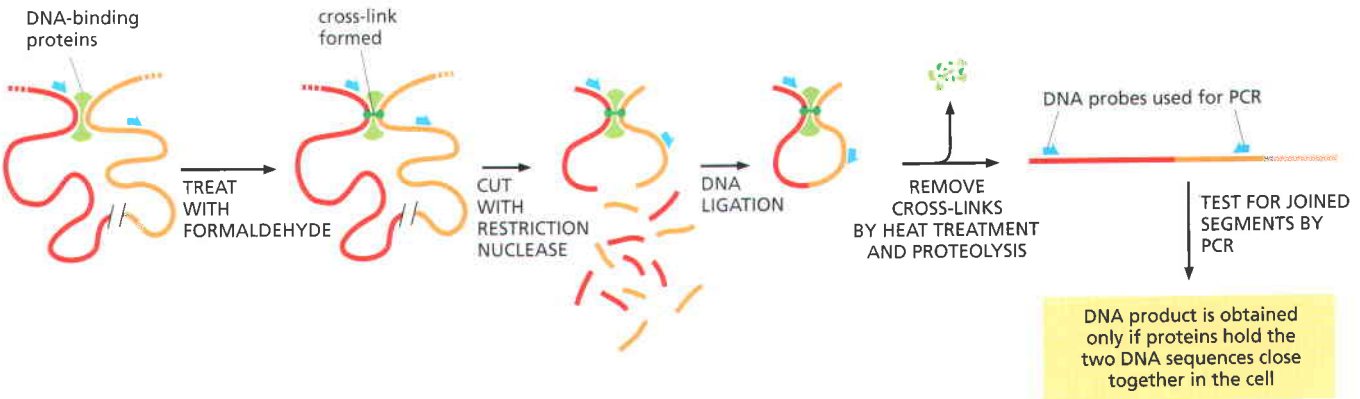


Figure 4–56 A method for determining the position of loops in interphase chromosomes. In this technique, known as the chromosome conformation capture (3C) method, cells are treated with formaldehyde to create the indicated covalent DNA–protein and DNA–DNA cross-links. The DNA is then treated with a restriction nuclease that chops the DNA into many pieces, cutting at strictly defined nucleotide sequences and forming sets of identical "cohesive ends" (see Figure 8–34). The cohesive ends can be made to join through their complementary base-pairing. Importantly, prior to the ligation step shown, the DNA is diluted so that the fragments that have been kept in close proximity to each other (through cross-linking) are the ones most likely to join. Finally, the cross-links are reversed and the newly ligated fragments of DNA are identified and quantified by PCR (the polymerase chain reaction, described in Chapter 8). By combining the frequency-of-association information generated by the 3C technique with DNA sequence information, structural models can be produced for the interphase conformation of chromosomes.

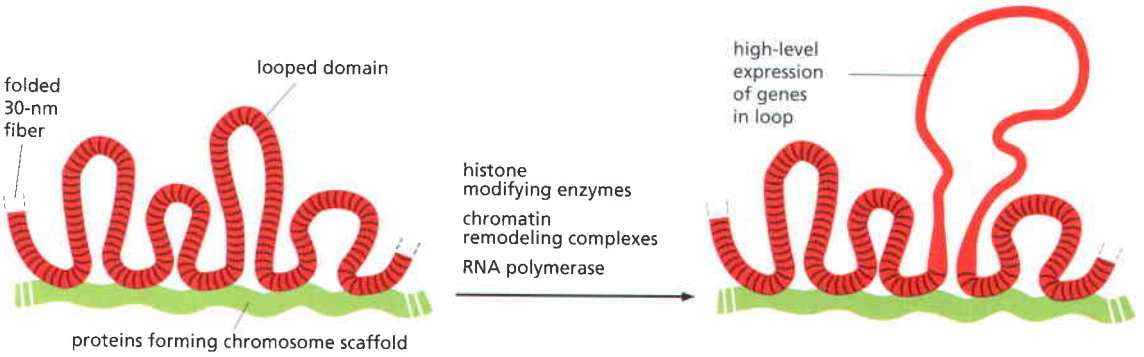


Figure 4–57 A model for the organization of an interphase chromosome. A section of an interphase chromosome is shown folded into a series of looped domains, each containing perhaps 50,000–200,000 nucleotide pairs of double-helical DNA condensed into a 30-nm fiber. The chromatin in each individual loop is further condensed through poorly understood folding processes that are reversed when the cell requires direct access to the DNA packaged in the loop. Neither the composition of the postulated chromosomal axis nor how the folded 30-nm fiber is anchored to it is clear. However, in mitotic chromosomes the bases of the chromosomal loops are enriched both in condensins and in DNA topoisomerase II enzymes, two proteins that may form much of the axis at metaphase (see Figure 4–74).

different lengths. A typical loop might contain between 50,000 and 200,000 nucleotide pairs of DNA, although loops of a million nucleotide pairs have also been suggested (**Figure 4–57**).

Polytene Chromosomes Are Uniquely Useful for Visualizing Chromatin Structures

Certain giant insect cells have grown to their enormous size through multiple cycles of DNA synthesis without cell division. Such cells with more than the normal DNA complement are said to be *polyploid* when they contain increased numbers of standard chromosomes. But in several types of secretory cells in fly larvae, all the homologous chromosome copies are held side by side, like drinking straws in a box, creating single large **polytene chromosomes**. Because polytene chromosomes can disperse to form a conventional polyploid cell in some cases, these two chromosomal states are closely related. The underlying structure of a polytene chromosome must therefore be similar to that of a normal chromosome.

Polyteny has been most studied in the larval salivary gland cells of the fruit fly *Drosophila*. When these polytene chromosomes are viewed in the light microscope, distinct alternating dark *bands* and light *interbands* are visible (**Figure 4–58**), each formed from a thousand identical DNA sequences arranged side-by-side in register. About 95% of the DNA in polytene chromosomes is in bands, and 5% is in interbands. A very thin band can contain 3000 nucleotide pairs, while a thick band may contain 200,000 nucleotide pairs in each of its chromatin strands. The chromatin in each band appears dark because the DNA is more condensed than the DNA in interbands; it may also contain a higher proportion of proteins (**Figure 4–59**).

There are approximately 3700 bands and 3700 interbands in the complete set of *Drosophila* polytene chromosomes. The bands can be recognized by their different thicknesses and spacings, and each one has been given a number to generate a chromosome "map" that has been indexed to the finished genome sequence of this fly.

The *Drosophila* polytene chromosomes provide a good starting point for examining how chromatin is organized on a large scale. In the previous section, we saw that there are many forms of chromatin, each of which contains nucleosomes with a different combination of modified histones. By reading this histone code, specific sets of non-histone proteins assemble on the nucleosomes to

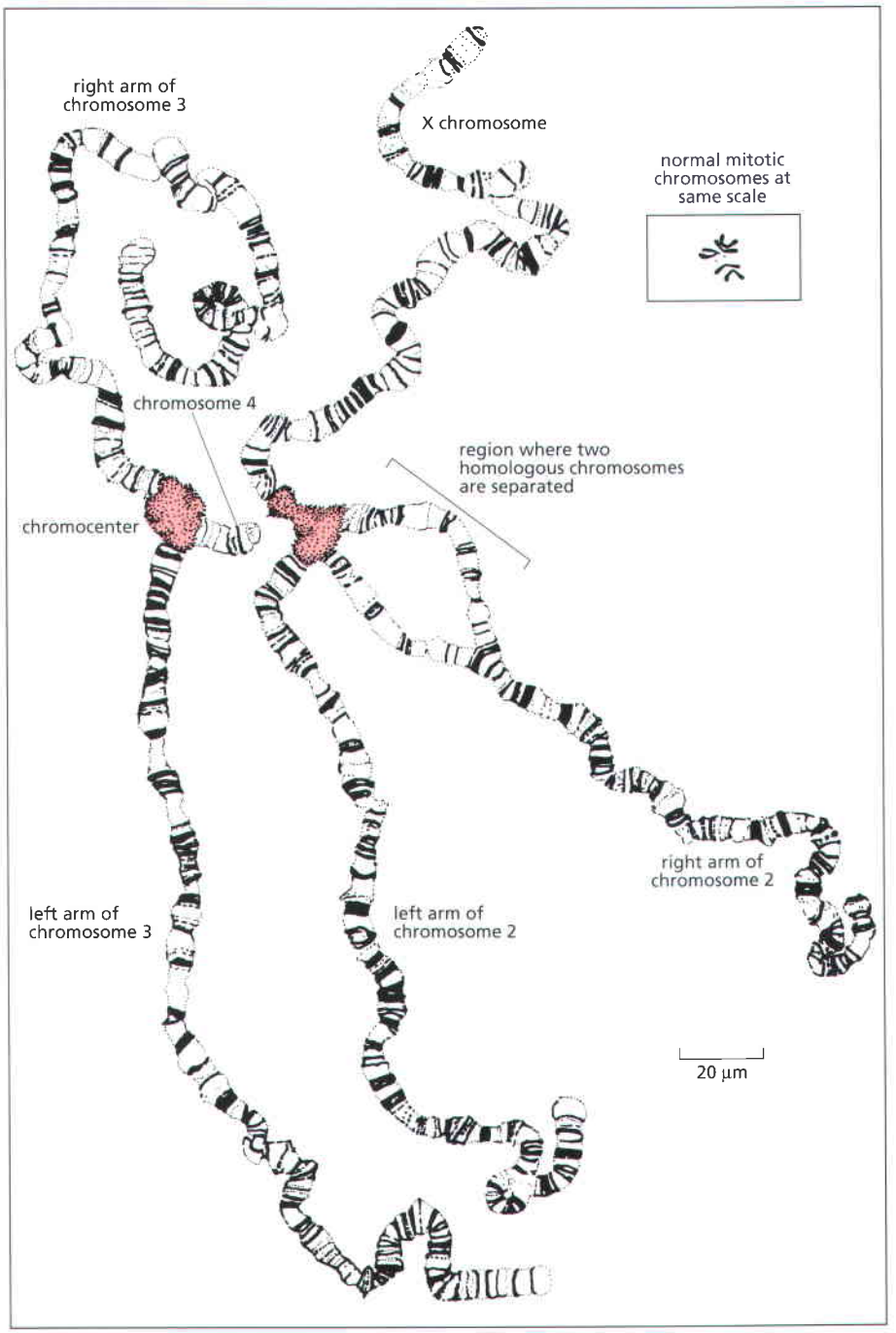


Figure 4–58 The entire set of polytene chromosomes in one *Drosophila* salivary cell. In this drawing of a light micrograph, the giant chromosomes have been spread out for viewing by squashing them against a microscope slide. *Drosophila* has four chromosomes, and there are four different chromosome pairs present. But each chromosome is tightly paired with its homolog (so that each pair appears as a single structure), which is not true in most nuclei (except in meiosis). Each chromosome has undergone multiple rounds of replication, and the homologues and all their duplicates have remained in exact register with each other, resulting in huge chromatin cables many DNA strands thick.

The four polytene chromosomes are normally linked together by heterochromatic regions near their centromeres that aggregate to create a single large chromocenter *(pink region)*. In this preparation, however, the chromocenter has been split into two halves by the squashing procedure used. (Adapted from T.S. Painter, *J. Hered.* 25:465–476, 1934. With permission from Oxford University Press.)

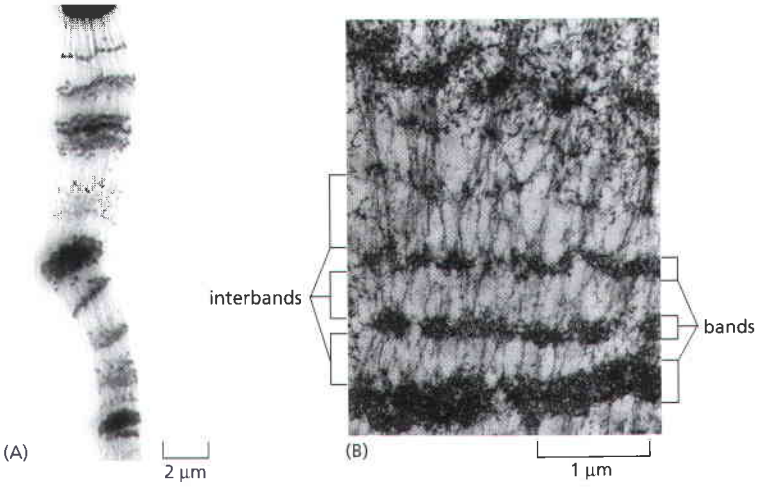


Figure 4–59 Micrographs of polytene chromosomes from *Drosophila* salivary glands. (A) Light micrograph of a portion of a chromosome. The DNA has been stained with a fluorescent dye, but a reverse image is presented here that renders the DNA *black* rather than *white*; the bands are clearly seen to be regions of increased DNA concentration. This chromosome has been processed by a high pressure treatment so as to show its distinct pattern of bands and interbands more clearly. (B) An electron micrograph of a small section of a *Drosophila* polytene chromosome seen in thin section. Bands of very different thickness can be readily distinguished, separated by interbands, which contain less condensed chromatin. (A, adapted from D.V. Novikov, I. Kireev and A.S. Belmont, *Nat. Methods* 4:483–485, 2007. With permission from Macmillan Publishers Ltd. B, courtesy of Veikko Sorsa.)

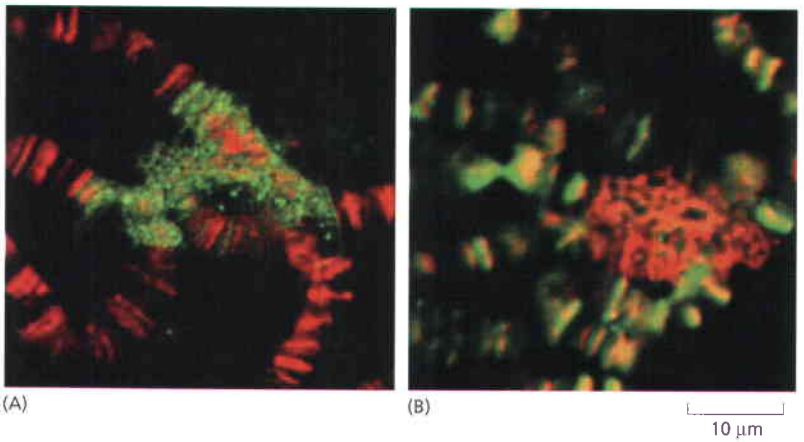


Figure 4–60 The pattern of histone modifications on *Drosophila* polytene chromosomes. Antibodies that specifically recognize different histone modifications can reveal where each modification is found with reference to the many bands and interbands on these chromosomes. In the two preparations shown here, the positions of two different markings on histone H3 tails are compared. In both cases, the antibody labeling the modified histone is *green*, and the DNA is stained *red*. Only a small region surrounding each chromocenter is shown. (A) Dimethyl Lys 9 *(green)* is a histone modification associated with the pericentric heterochromatin. It is seen to be associated with the chromocenter. (B) Acetylated Lys 9 *(green)* is a modification that is concentrated in histones associated with active genes. It is seen to be present in numerous bands on the chromosome arms, but not in the heterochromatic chromocenter. Similar experiments can be carried out to position many other modified histones, as well as the non-histone proteins (see, for example, Figure 22–45 for chromosomes stained for Polycomb). (Adapted from A. Ebert, S. Lein, G. Schotta and G. Reuter, *Chromosome Res.* 14:377–392, 2006. With permission from Springer.)

affect biological function in different ways. Some of these non-histone proteins can spread for long distances along the DNA, imparting a similar chromatin structure to contiguous regions of the genome (see Figure 4–46). Thus, in some regions, all of the chromatin has a similar structure and is separated from neighboring domains by barrier proteins (see Figure 4–47). At low resolution, the interphase chromosome can therefore be considered as a mosaic of chromatin structures, each containing particular nucleosome modifications associated with a particular set of non-histone proteins. (At a higher level of resolution one would also emphasize the many sequence-specific DNA-binding proteins that will be described in Chapter 7).

This view of an interphase chromosome helps us to interpret the results obtained from studies of polytene chromosomes. By staining with highly specific antibodies, one can show that differently modified histones (**Figure 4–60**), as well as distinct sets of non-histone proteins, are located on different polytene chromosome bands. This suggests a powerful general strategy. By employing combinations of antibodies that bind tightly to each of the many different histone modifications that create the histone code (see Figure 4–39), it may be possible to determine which combinations of modifications specify particular types of chromatin domains. And by carrying out similar experiments with antibodies that recognize each of the hundreds of different non-histone proteins in chromatin, one can attempt to decipher the many different meanings encoded in histone modifications.

There Are Multiple Forms of Heterochromatin

Molecular studies have led to a reevaluation of our view of heterochromatin. For many decades, heterochromatin was thought to be a single entity defined by its highly condensed structure and its ability to silence genes permanently. But if we define heterochromatin as a form of compact chromatin that can silence genes, be epigenetically inherited, and spread along chromosomes to cause position effect variegation (see Figure 4–36), it is clear that different types of heterochromatin exist. In fact, we have already considered three of these types in discussing the human centromere (see Figure 4–50).

Each domain of heterochromatin is thought to be formed by the cooperative assembly of a set of non-histone proteins. For example, classical pericentromeric heterochromatin contains more than six such proteins, including heterochromatin protein 1 (HP1), whereas the so-called Polycomb form of heterochromatin contains a similar number of proteins in a non-overlapping set (PcG proteins). There are hundreds of small blocks of heterochromatin spread across the arms of *Drosophila* polytene chromosomes, as identified by their late replication (discussed in Chapter 5). Antibody staining of these regions of heterochromatin suggests that the known forms of heterochromatin can account for no more than half of the heterchromatic polytene bands (**Figure 4–61**). Thus, other types of heterochromatin must exist whose protein composition is not known. It is likely

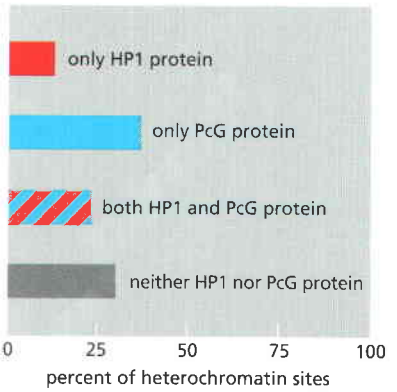


Figure 4–61 Evidence for multiple forms of heterochromatin. In this study, 240 late-replicating sites on the *Drosophila* polytene chromosome arms were examined for the presence of two non-histone proteins. These proteins are known to help compact two different forms of heterochromatin (see text). As indicated, antibody staining suggests that roughly half of the sites are packaged into forms of heterochromatin that are different from either of these two. Experiments such as these demonstrate that we have a great deal more to learn about the packaging of DNA in eucaryotes. (Data from I.F. Zhimulev and E.S. Belyaeva, *BioEssays* 25:1040–1051, 2003. With permission from John Wiley & Sons.)

that each of these types of heterochromatin is differently regulated and has different roles in the cell.

The chromatin structure in each domain ultimately depends on the proteins that bind to specific DNA sequences, and these are known to vary depending on the cell type and its stage of development in a multicellular organism. Thus, both the pattern of chromatin domains and their individual compositions (nucleosome modifications plus non-histone proteins) can vary between tissues. These differences make different genes accessible for genetic readout, helping to explain the cell diversification that accompanies embryonic development (described in Chapter 22). Comparisons of the polytene chromosomes in two different tissues of a fly lend support to this general idea: although the patterns of bands and interbands are largely the same, there are reproducible differences.

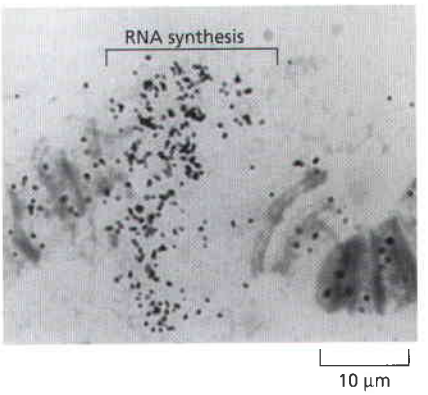


Figure 4–62 **RNA synthesis in polytene chromosome puffs.** An autoradiograph of a single puff in a polytene chromosome from the salivary glands of the freshwater midge *C. tentans*. As outlined in Chapter 1 and described in detail in Chapter 6, the first step in gene expression is the synthesis of an RNA molecule using the DNA as a template. The decondensed portion of the chromosome is undergoing RNA synthesis and has become labeled with ^{3}H-uridine (see p. 603), an RNA precursor molecule that is incorporated into growing RNA chains. (Courtesy of José Bonner.)

Chromatin Loops Decondense When the Genes Within Them Are Expressed

When an insect progresses from one developmental stage to another, distinctive *chromosome puffs* arise and old puffs recede in its polytene chromosomes as new genes become expressed and old ones are turned off (**Figure 4–62**). From inspection of each puff when it is relatively small and the banding pattern is still discernible, it seems that most puffs arise from the decondensation of a single chromosome band.

The individual chromatin fibers that make up a puff can be visualized with an electron microscope. In favorable cases, loops are seen, much like those observed in the amphibian lampbrush chromosomes discussed above. When not expressed, the loop of DNA assumes a thickened structure, possibly a folded 30-nm fiber, but when gene expression is occurring, the loop becomes more extended. In electron micrographs, the chromatin located on either side of the decondensed loop appears considerably more compact, suggesting that a loop constitutes a distinct functional domain of chromatin structure.

Observations made in human cells also suggest that highly folded loops of chromatin expand to occupy an increased volume when a gene within them is expressed. For example, quiescent chromosome regions from 0.4 to 2 million nucleotide pairs in length appear as compact dots in an interphase nucleus when visualized by fluorescence microscopy using FISH or other technologies. However, the same DNA is seen to occupy a larger territory when its genes are expressed, with elongated, punctate structures replacing the original dot.

Chromatin Can Move to Specific Sites Within the Nucleus to Alter Gene Expression

New ways of visualizing individual chromosomes have shown that each of the 46 interphase chromosomes in a human cell tends to occupy its own discrete territory within the nucleus (**Figure 4–63**). However, pictures such as these present only an average view of the DNA in each chromosome. Experiments that specifically localize the heterochromatic regions of a chromosome reveal that they are often closely associated with the nuclear lamina, regardless of the chromosome examined. And DNA probes that preferentially stain gene-rich regions of human

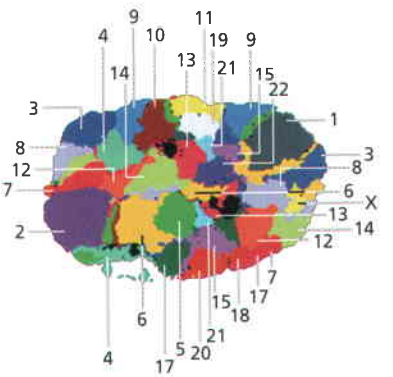


Figure 4–63 **Simultaneous visualization of the chromosome territories for all of the human chromosomes in a single interphase nucleus.** A FISH analysis using a different mixture of fluorochromes for marking the DNA of each chromosome, detected with seven color channels in a fluoresence microscope, allows each chromosome to be distinguished in three-dimensional reconstructions. Below the micrograph, each chromosome is identified in a schematic of the actual image. Note that the two homologous chromosomes (e.g., the two copies of chromosome 9), are not in general co-located. (From M.R. Speicher and N.P. Carter, *Nat. Rev. Genet.* 6:782–792, 2005. With permission from Macmillan Publishers Ltd.)

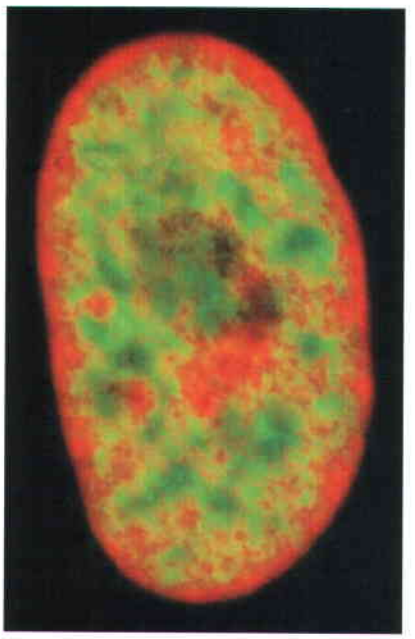


Figure 4–64 The distribution of gene-rich regions of the human genome in an interphase nucleus. Gene-rich regions have been visualized with a fluorescent probe that hybridizes to the Alu interspersed repeat, which is present in more than a million copies in the human genome (see Figure 5–75). For unknown reasons, these sequences cluster in chromosomal regions rich in genes. In this representation, regions enriched for the Alu sequence are *green,* regions depleted for these sequences are *red,* while the average regions are *yellow.* The gene-rich regions are seen to be depleted in the DNA near the nuclear envelope. (From A. Bolzer et al., *PLoS Biol.* 3:826–842, 2005. With permission from Public Library of Science.)

chromosomes produce a striking picture of the interphase nucleus that presumably reflects different average positions for active and inactive genes (**Figure 4–64**).

A variety of different types of experiments have led to the conclusion that the position of a gene in the interior of the nucleus changes when it becomes highly expressed. Thus, a region that becomes very actively transcribed is often found to extend out of its chromosome territory, as if in an extended loop (**Figure 4–65**). We will see in Chapter 6 that the initiation of transcription—the first step in gene expression—requires the assembly of over 100 proteins, and it makes sense that this would occur most rapidly in regions of the nucleus particularly rich in these proteins.

More generally, it is clear that the nucleus is very heterogeneous, with functionally different regions to which portions of chromosomes can move as they are subjected to different biochemical processes—such as when their gene expression changes (**Figure 4–66**). There is evidence that some of these nuclear regions are marked with different inositol phospholipids, reminiscent of the way that the same lipids are used to distinguish different membranes in the cytoplasm (see Figure 13–11). But what these lipids are attached to in the interior of the nucleus is a mystery, as the only known lipid-rich environments are the lipid bilayers of the nuclear envelope.

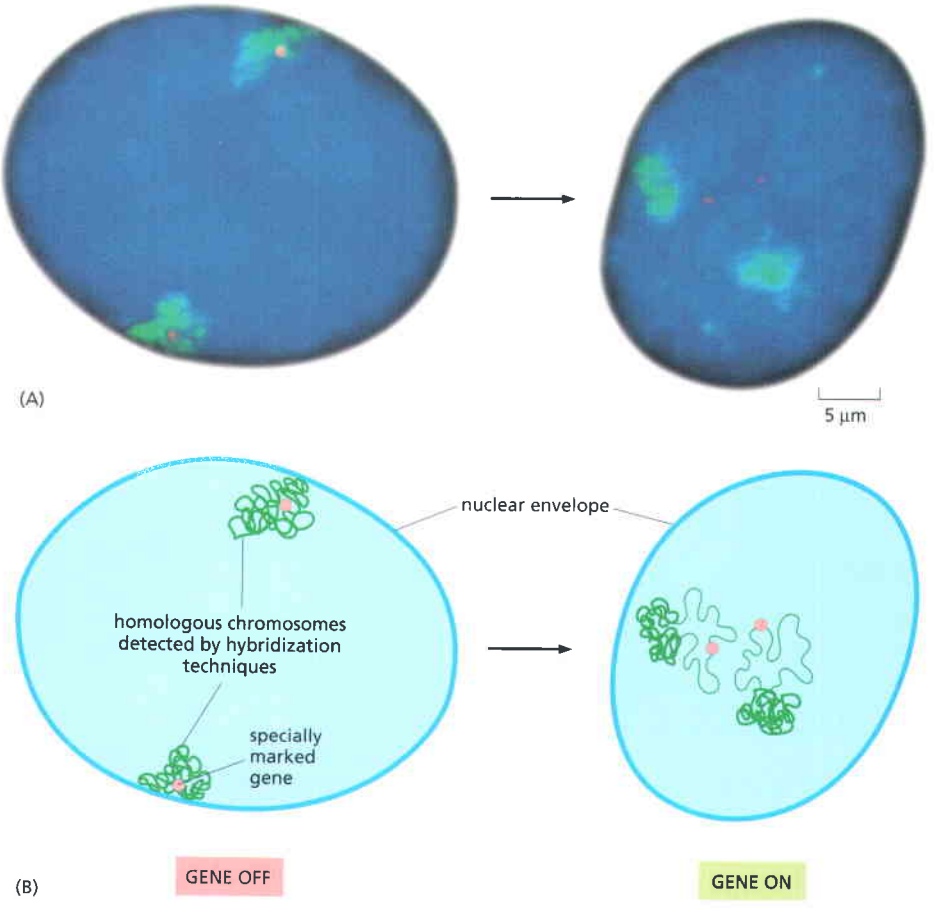


Figure 4–65 An effect of high levels of gene expression on the intranuclear location of chromatin. (A) Fluorescence micrographs of human nuclei showing how the position of a gene changes when it becomes highly transcribed. The region of the chromosome adjacent to the gene *(red)* is seen to leave its chromosomal territory *(green)* only when it is highly active. (B) Schematic representation of a large loop of chromatin that expands when the gene is on, and contracts when the gene is off. Other genes that are less actively expressed can be shown by the same methods to remain inside their chromosomal territory when transcribed. (From J.R. Chubb and W.A. Bickmore, *Cell* 112:403–406, 2003. With permission from Elsevier.)

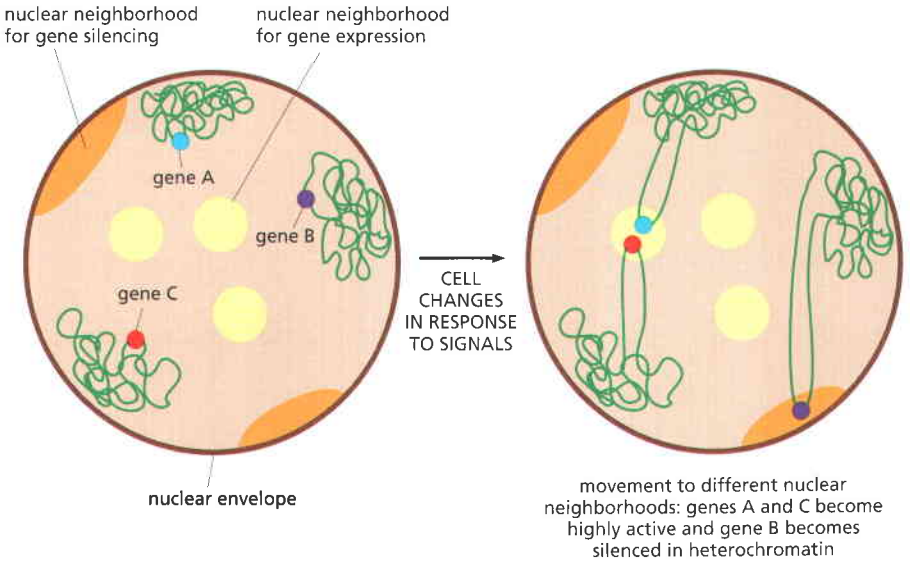


Figure 4–66 The movement of genes to different regions of the nucleus when their expression changes. The interior of the nucleus is very heterogeneous, and different nuclear neighborhoods are known to have distinct effects on gene expression. Movements such as those indicated here presumably reflect changes in the binding affinities that the chromatin and RNA molecules surrounding a gene have for different nuclear neighborhoods. It is thought that the movement is driven by diffusion and does not require a directed movement process, inasmuch as each region of a chromosome can be seen to undergo constant random motion when marked in a way that allows its position to be followed in a living cell.

Networks of Macromolecules Form a Set of Distinct Biochemical Environments inside the Nucleus

In Chapter 6, we describe the function of a variety of subcompartments that are present within the nucleus. The largest and most obvious of these is the nucleolus, a structure well known to microscopists even in the 19th century (see Figure 4–9). Nucleolar regions consist of networks of RNAs and proteins surrounding transcribing ribosomal RNA genes, often existing as multiple nucleoli. The nucleolus is the cell's site of ribosome assembly and maturation, as well as the place where many other specialized reactions occur.

A variety of less obvious organelles are also present inside the nucleus. For example, spherical structures called Cajal bodies and interchromatin granule clusters are present in most plant and animal cells (**Figure 4–67**). Like the nucleolus, these organelles are composed of selected protein and RNA molecules that bind together to create networks that are highly permeable to other protein and RNA molecules in the surrounding nucleoplasm (**Figure 4–68**).

Structures such as these can create distinct biochemical environments by immobilizing select groups of macromolecules, as can other networks of proteins and RNA molecules associated with nuclear pores and with the nuclear envelope. In principle, this allows the molecules that enter these spaces to be processed with great efficiency through complex reaction pathways. Highly permeable, fibrous networks of this sort can thereby impart many of the kinetic advantages of compartmentalization (see p. 186) to reactions that take place in the nucleus (**Figure 4–69**A). However, unlike the membrane-bound compartments in the cytoplasm (discussed in Chapter 12), these nuclear subcompartments—lacking a lipid bilayer membrane—can neither concentrate nor exclude specific small molecules.

The cell has a remarkable ability to construct distinct biochemical environments inside the nucleus. Those thus far mentioned facilitate various aspects of gene expression to be discussed in Chapter 6 (see Figure 6–49). Like the nucleolus, these subcompartments appear to form only as needed, and they create a high local concentration of the many different enzymes and RNA molecules needed for a particular process. In an analogous way, when DNA is damaged by irradiation, the set of enzymes needed to carry out DNA repair are observed to congregate in discrete foci inside the nucleus, creating "repair factories" (see Figure 5–60). And nuclei often contain hundreds of discrete foci representing factories for DNA or RNA synthesis.

It seems likely that all of these entities make use of the type of tethering illustrated in Figure 4–69B, where long flexible lengths of polypeptide chain (or some other polymer) are interspersed with binding sites that concentrate the multiple proteins and/or RNA molecules that are needed to catalyze a particular process. Not surprisingly, tethers are similarly used to help to speed biological processes

Figure 4–67 Electron micrograph showing two very common fibrous nuclear subcompartments. The large sphere here is a Cajal body. The smaller darker sphere is an interchromatin granule cluster, also known as a spreckle (see also Figure 6–49). These "subnuclear organelles" are from the nucleus of a *Xenopus* oocyte. (From K.E. Handwerger and J.G. Gall, *Trends Cell Biol.* 16:19–26, 2006. With permission from Elsevier.)

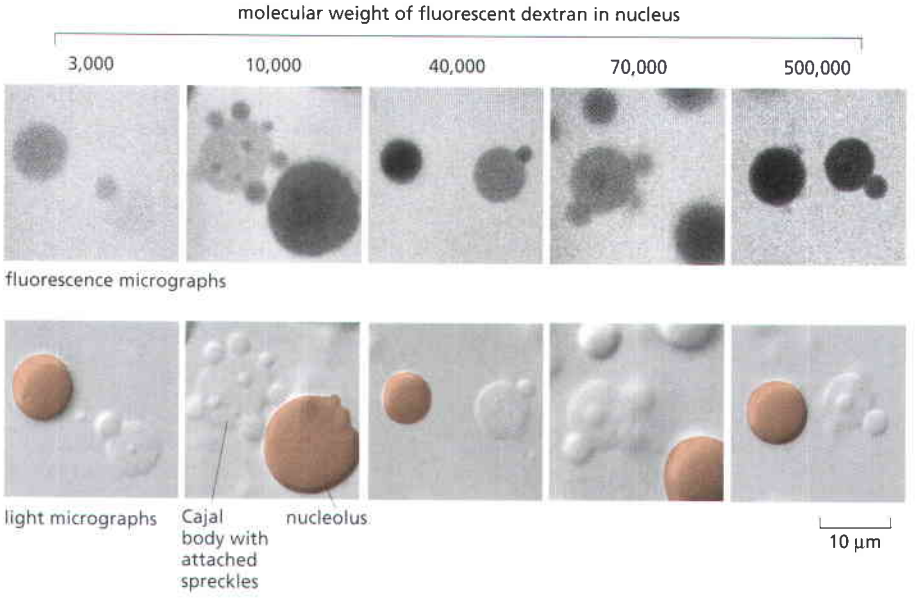


Figure 4–68 An experiment showing that the subnuclear organelles are highly permeable to macromolecules. In these micrographs of a living oocyte nucleus, the top row compares the fluorescence of the interiors of nucleoli, Cajal bodies, and spreckles to the fluorescence of the surrounding nucleoplasm, 12 hours after fluorescent dextrans of the indicated molecular weight had been injected into the nucleoplasm. The brightness of each organelle reflects its permeability, with the most permeable organelle being the brightest. For comparison, the bottom row presents normal light micrographs of the same microscope fields, with the nucleolus in each field of view marked *brown* to distinguish it. Cajal bodies can be seen to be more permeable than nucleoli. However, quantitation shows that a great deal of dextran enters each organelle, even for the largest dextran tested. (From K.E. Handwerger, J.A. Cordero and J.G. Gall, *Mol. Biol. Cell* 16:202–211, 2005. With permission from American Society of Cell Biology.)

in the cytoplasm, increasing specific reaction rates (for example, see Figure 16–38).

Is there also is an intranuclear framework, analogous to the cytoskeleton, on which chromosomes and other components of the nucleus are organized? The *nuclear matrix*, or *scaffold*, has been defined as the insoluble material left in the nucleus after a series of biochemical extraction steps. Many of the proteins and RNA molecules that form this insoluble material are likely to be derived from the fibrous subcompartments of the nucleus just discussed, while others seem to be proteins that help to form the base of chromosomal loops or to attach chromosomes to other structures in the nucleus. Whether or not the nucleus also contains

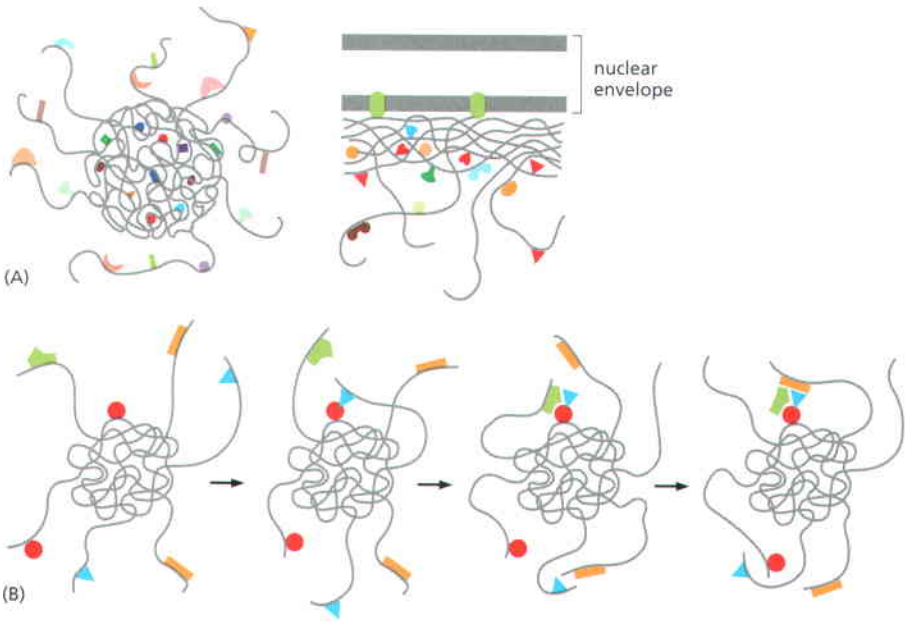


Figure 4–69 Effective compartmentalization without a bilayer membrane. (A) Schematic illustration of the organization of a spherical subnuclear organelle (*left*) and of a postulated similarly organized subcompartment just beneath the nuclear envelope (*right*). In both cases, RNAs and/or proteins (*gray*) associate to form highly porous, gel-like structures that contain binding sites for other specific proteins and RNA molecules (*colored objects*). (B) How the tethering of a selected set of proteins and RNA molecules to long flexible polymer chains, as in A, could create "staging areas" that greatly speed the rates of reactions in subcompartments of the nucleus. The reactions catalyzed will depend on the particular macromolecules that are localized by the tethering. The same type of rate accelerations are of course expected for similar subcompartments established elsewhere in the cell (see also Figure 3–80C).

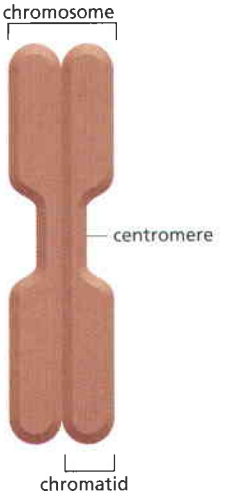


Figure 4–70 A typical mitotic chromosome at metaphase. Each sister chromatid contains one of two identical daughter DNA molecules generated earlier in the cell cycle by DNA replication (see also Figure 17–26).

long filaments that form organized tracks on which nuclear components can move, analogous to some of the filaments in the cytoplasm, is still disputed.

Mitotic Chromosomes Are Formed from Chromatin in Its Most Condensed State

Having discussed the dynamic structure of interphase chromosomes, we now turn to mitotic chromosomes. The chromosomes from nearly all eucaryotic cells become readily visible by light microscopy during mitosis, when they coil up to form highly condensed structures. This condensation reduces the length of a typical interphase chromosome only about tenfold, but it produces a dramatic change in chromosome appearance.

Figure 4–70 depicts a typical **mitotic chromosome** at the metaphase stage of mitosis (for the stages of mitosis, see Figure 17–3). The two daughter DNA molecules produced by DNA replication during interphase of the cell-division cycle are separately folded to produce two sister chromosomes, or *sister chromatids*, held together at their centromeres (see also Figure 4–50). These chromosomes are normally covered with a variety of molecules, including large amounts of RNA–protein complexes. Once this covering has been stripped away, each chromatid can be seen in electron micrographs to be organized into loops of chromatin emanating from a central scaffolding (**Figure 4–71**). Experiments using DNA hybridization to detect specific DNA sequences demonstrate that the order of visible features along a mitotic chromosome at least roughly reflects the order of genes along the DNA molecule. Mitotic chromosome condensation can thus be thought of as the final level in the hierarchy of chromosome packaging (**Figure 4–72**).

The compaction of chromosomes during mitosis is a highly organized and dynamic process that serves at least two important purposes. First, when condensation is complete (in metaphase), sister chromatids have been disentangled from each other and lie side by side. Thus, the sister chromatids can easily separate when the mitotic apparatus begins pulling them apart. Second, the compaction of chromosomes protects the relatively fragile DNA molecules from being broken as they are pulled to separate daughter cells.

The condensation of interphase chromosomes into mitotic chromosomes begins in early M phase, and it is intimately connected with the progression of the cell cycle, as discussed in detail in Chapter 17. During M phase, gene expression shuts down, and specific modifications are made to histones that help to reorganize the chromatin as it compacts. The compaction is aided by a class of proteins called *condensins*, which use the energy of ATP hydrolysis to help coil the two DNA molecules in an interphase chromosome to produce the two chromatids of a mitotic chromosome. Condensins are large protein complexes built from SMC protein dimers: these dimers form when two stiff, elongated protein monomers join at their tails to form a hinge, leaving two globular head domains at the other end that bind DNA and hydrolyze ATP (**Figure 4–73**). When added to purified DNA, condensins can make large right-handed loops in DNA molecules in a reaction that requires ATP. Although it is not yet known how they act on chromatin, the coiling model shown in Figure 4–73C is based on the fact that condensins are a major structural component that end up at the core of metaphase chromosomes, with about one molecule of condensin for every

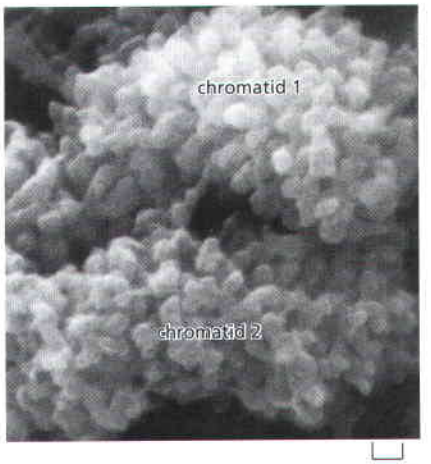


Figure 4–71 A scanning electron micrograph of a region near one end of a typical mitotic chromosome. Each knoblike projection is believed to represent the tip of a separate looped domain. Note that the two identical paired chromatids (drawn in Figure 4–70) can be clearly distinguished. (From M.P. Marsden and U.K. Laemmli, *Cell* 17:849–858, 1979. With permission from Elsevier.)

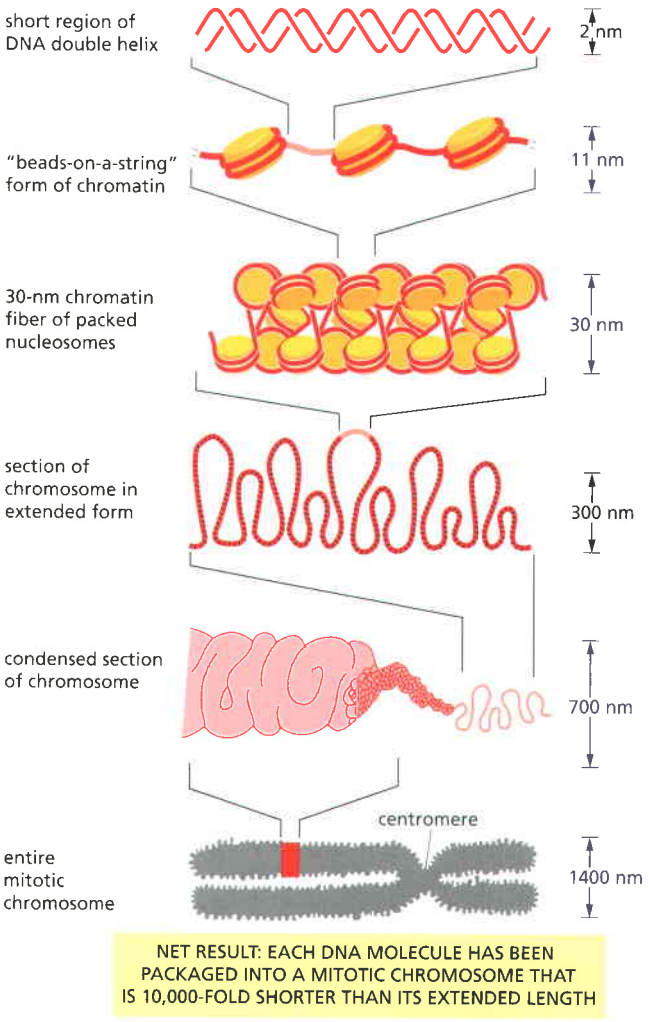
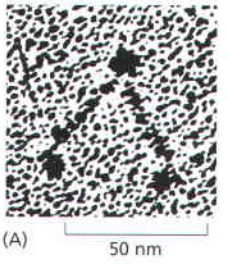


Figure 4–72 Chromatin packing. This model shows some of the many levels of chromatin packing postulated to give rise to the highly condensed mitotic chromosome.



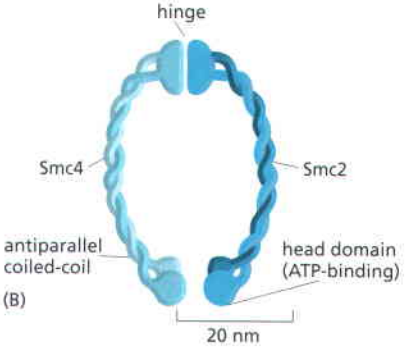
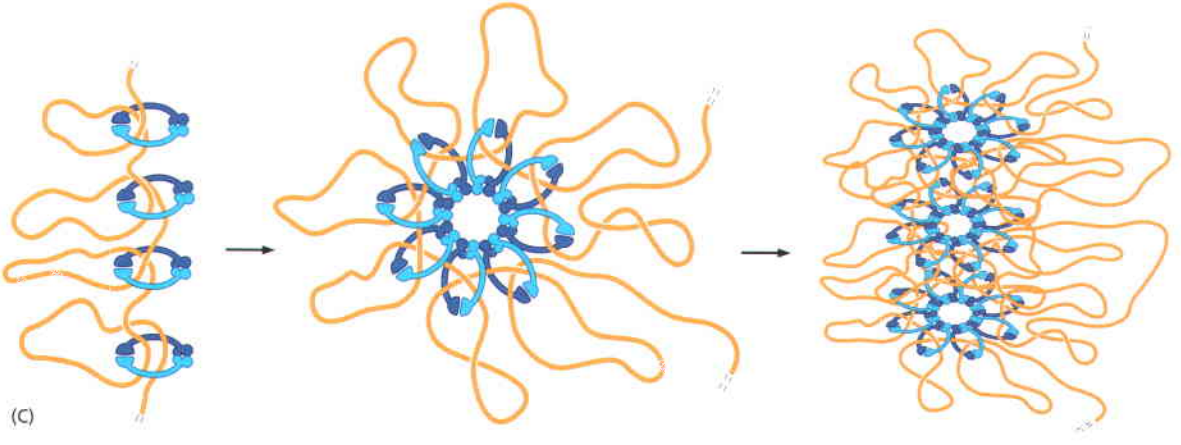


Figure 4–73 The SMC proteins in condensins. (A) Electron micrographs of a purified SMC dimer. (B) The structure of a SMC dimer. The long central region of this protein is an antiparallel coiled-coil (see Figure 3–9) with a flexible hinge in its middle. (C) A model for the way in which the SMC proteins in condensins might compact chromatin. In reality, SMC proteins are components of a much larger condensin complex. It has been proposed that, in the cell, condensins coil long strings of looped chromatin domains (see Figure 4–57). In this way, the condensins could form a structural framework that maintains the DNA in a highly organized state during metaphase of the cell cycle. (A, courtesy of H.P. Erickson; B and C, adapted from T. Hirano, *Nat. Rev. Mol. Cell Biol.* 7:311–322, 2006. With permission from Macmillan Publishers Ltd.)



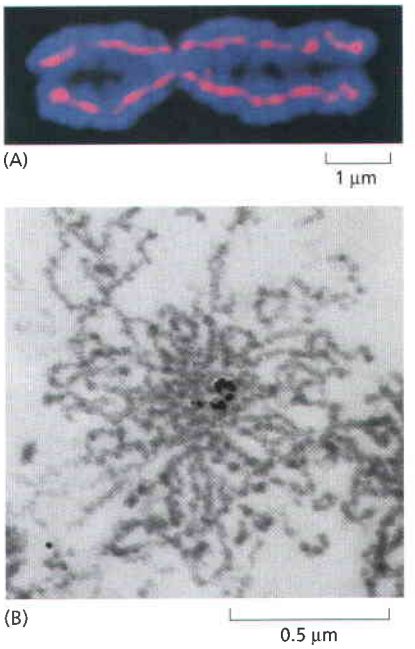


Figure 4–74 The location of condensin in condensed mitotic chromosomes. (A) Fluorescence micrograph of a human chromosome at mitosis, stained with an antibody that localizes condensin. In chromosomes that are this highly condensed, the condensin is seen to be concentrated in punctate structures along the chromosome axis. Similar experiments show a similar location for DNA topoisomerase II, an enzyme that makes reversible double-strand breaks in DNA that allow one DNA double helix to pass through another (see Figure 5–23). (B) Immunogold electron microscopy reveals localization of condensin (*black* dots). Here a chromatid is seen in cross section, with the chromosome axis perpendicular to the plane of the paper. (A, from K. Maeshima and U.K. Laemmli, *Dev. Cell* 4:467–480, 2003. With permission from Elsevier. B, courtesy of U.K. Laemmli, from K. Maeshima, M. Eltsov and U.K. Laemmli, *Chromosoma* 114:365–375, 2005. With permission from Springer.)

10,000 nucleotides of DNA (**Figure 4–74**). When condensins are experimentally depleted from a cell, chromosome condensation still occurs, but the process is abnormal.

Summary

Chromosomes are generally decondensed during interphase, so that the details of their structure are difficult to visualize. Notable exceptions are the specialized lampbrush chromosomes of vertebrate oocytes and the polytene chromosomes in the giant secretory cells of insects. Studies of these two types of interphase chromosomes suggest that each long DNA molecule in a chromosome is divided into a large number of discrete domains organized as loops of chromatin, each loop probably consisting of a 30-nm chromatin fiber that is compacted by further folding. When genes contained in a loop are expressed, the loop unfolds and allows the cell's machinery access to the DNA.

Interphase chromosomes occupy discrete territories in the cell nucleus; that is, they are not extensively intertwined. Euchromatin makes up most of interphase chromosomes and, when not being transcribed, it probably exists as tightly folded 30-nm fibers. However, euchromatin is interrupted by stretches of heterochromatin, in which the 30-nm fibers are subjected to additional packing that usually renders it resistant to gene expression. Heterochromatin exists in several forms, some of which are found in large blocks in and around centromeres and near telomeres. But heterochromatin is also present at many other positions on chromosomes, where it can serve to regulate developmentally important genes.

The interior of the nucleus is highly dynamic, with heterochromatin often positioned near the nuclear envelope and loops of chromatin moving away from their chromosome territory when genes are very highly expressed. This reflects the existence of nuclear subcompartments, where different sets of biochemical reactions are facilitated by an increased concentration of selected proteins and RNAs. The components involved in forming a subcompartment can self-assemble into discrete organelles such as nucleoli or Cajal bodies; they can also be tethered to fixed structures such as the nuclear envelope.

During mitosis, gene expression shuts down and all chromosomes adopt a highly condensed conformation in a process that begins early in M phase to package the two DNA molecules of each replicated chromosome as two separately folded chromatids. This process is accompanied by histone modifications that facilitate chromatin packing. However, satisfactory completion of this orderly process, which reduces the end-to-end distance of each DNA molecule from its interphase length by an additional factor of ten, requires condensin proteins.

HOW GENOMES EVOLVE

In this chapter, we have discussed the structure of genes and the ways that they are packaged and arranged in chromosomes. In this final section, we provide an overview of some of the ways that genes and genomes have evolved over time to produce the vast diversity of modern-day life forms on our planet. Genome

sequencing has revolutionized our view of the process of molecular evolution, uncovering an astonishing wealth of information about specific family relationships among organisms, as well as illuminating evolutionary mechanisms more generally.

It is perhaps not surprising that genes with similar functions can be found in a diverse range of living things. But the great revelation of the past 25 years has been the discovery that the actual nucleotide sequences of many genes are sufficiently well conserved that **homologous** genes—that is, genes that are similar in both their nucleotide sequence and function because of a common ancestry—can often be recognized across vast phylogenetic distances. For example, unmistakable homologs of many human genes are easy to detect in such organisms as nematode worms, fruit flies, yeasts, and even bacteria. In many cases, the resemblance is so close that the protein-coding portion of a yeast gene can be substituted with its human homolog—even though we and yeast are separated by more than a billion years of evolutionary history.

As emphasized in Chapter 3, the recognition of sequence similarity has become a major tool for inferring gene and protein function. Although finding a sequence match does not guarantee similarity in function, it has proven to be an excellent clue. Thus, it is often possible to predict the function of genes in humans for which no biochemical or genetic information is available simply by comparing their nucleotide sequences with the sequences of genes in other organisms.

In general, gene sequences are more tightly conserved than is overall genome structure. As we saw earlier, other features of genome organization such as genome size, number of chromosomes, order of genes along chromosomes, abundance and size of introns, and amount of repetitive DNA are found to differ greatly among organisms, as does the number of genes that an organism contains.

The number of genes is only very roughly correlated with the phenotypic complexity of an organism (see Table 1–1). Much of the increase in gene number observed with increasing biological complexity involves the expansion of families of closely related genes, an observation that establishes gene duplication and divergence as major evolutionary processes. Indeed, it is likely that all present-day genes are descendants—via the processes of duplication, divergence, and reassortment of gene segments—of a few ancestral genes that existed in early life forms.

Genome Alterations are Caused by Failures of the Normal Mechanisms for Copying and Maintaining DNA

Cells in the germline do not have specialized mechanisms for creating changes in the structures of their genomes: evolution depends instead on accidents and mistakes followed by nonrandom survival. Most of the genetic changes that occur result simply from failures in the normal mechanisms by which genomes are copied or repaired when damaged, although the movement of transposable DNA elements also plays an important role. As we will discuss in Chapter 5, the mechanisms that maintain DNA sequences are remarkably precise—but they are not perfect. For example, because of the elaborate DNA-replication and DNA-repair mechanisms that enable DNA sequences to be inherited with extraordinary fidelity, along a given line of descent only about one nucleotide pair in a thousand is randomly changed in the germline every million years. Even so, in a population of 10,000 diploid individuals, every possible nucleotide substitution will have been "tried out" on about 20 occasions in the course of a million years—a short span of time in relation to the evolution of species.

Errors in DNA replication, DNA recombination, or DNA repair can lead either to simple changes in DNA sequence—such as the substitution of one base pair for another—or to large-scale genome rearrangements such as deletions, duplications, inversions, and translocations of DNA from one chromosome to another. In addition to these failures of the genetic machinery, the various mobile DNA elements that will be described in Chapter 5 are an important source of genomic change (see Table 5–3, p. 318). These transposable DNA elements (*transposons*)

are parasitic DNA sequences that colonize genomes and can spread within them. In the process, they often disrupt the function or alter the regulation of existing genes. On occasion, they can even create altogether novel genes through fusions between transposon sequences and segments of existing genes. Over long periods of evolutionary time, transposons have profoundly affected the structure of genomes. In fact, nearly half of the DNA in the human genome has recognizable sequence similarity with known transposon sequences, thereby indicating that these sequences are remnants of past transposition events (see Figure 4–17). Even more of our genome is no doubt derived from transposition events that occurred so long ago (>10^8 years) that the sequences can no longer be traced to transposons.

The Genome Sequences of Two Species Differ in Proportion to the Length of Time That They Have Separately Evolved

The differences between the genomes of species alive today have accumulated over more than 3 billion years. Lacking a direct record of changes over time, we can nevertheless reconstruct the process of genome evolution from detailed comparisons of the genomes of contemporary organisms.

The basic tool of comparative genomics is the phylogenetic tree. A simple example is the tree describing the divergence of humans from the great apes (**Figure 4–75**). The primary support for this tree comes from comparisons of gene or protein sequences. For example, comparisons between the sequences of human genes or proteins and those of the great apes typically reveal the fewest differences between human and chimpanzee and the most between human and orangutan.

For closely related organisms such as humans and chimpanzees, it is relatively easy to reconstruct the gene sequences of the extinct, last common ancestor of the two species (**Figure 4–76**). The close similarity between human and chimpanzee genes is mainly due to the short time that has been available for the accumulation of mutations in the two diverging lineages, rather than to functional constraints that have kept the sequences the same. Evidence for this view comes from the observation that even DNA sequences whose nucleotide order is functionally unconstrained—such as the sequences that code for the fibrinopeptides (see p. 264) or the third position of "synonymous" codons (codons specifying the same amino acid—see Figure 4–76)—are nearly identical in humans and chimpanzees.

For much less closely related organisms, such as humans and chickens (which have evolved separately for about 300 million years), the sequence conservation found in genes is largely due to **purifying selection** (that is, selection that eliminates individuals carrying mutations that interfere with important genetic functions), rather than to an inadequate time for mutations to occur. As a result, protein-coding, RNA-coding, and regulatory sequences in the DNA are often remarkably conserved. In contrast, most DNA sequences in the human and chicken genomes have diverged so far due to multiple mutations that it is often impossible to align them with one another.

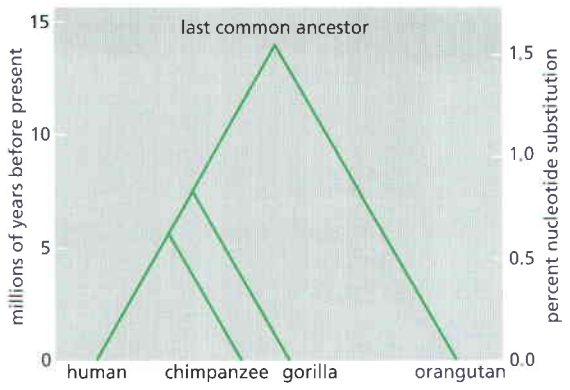


Figure 4–75 A phylogenetic tree showing the relationship between the human and the great apes based on nucleotide sequence data. As indicated, the sequences of the genomes of all four species are estimated to differ from the sequence of the genome of their last common ancestor by a little over 1.5%. Because changes occur independently on both diverging lineages, pairwise comparisons reveal twice the sequence divergence from the last common ancestor. For example, human–orangutan comparisons typically show sequence divergences of a little over 3%, while human–chimpanzee comparisons show divergences of approximately 1.2%. (Modified from F.C. Chen and W.H. Li, *Am. J. Hum. Genet.* 68:444–456, 2001. With permission from University of Chicago Press.)

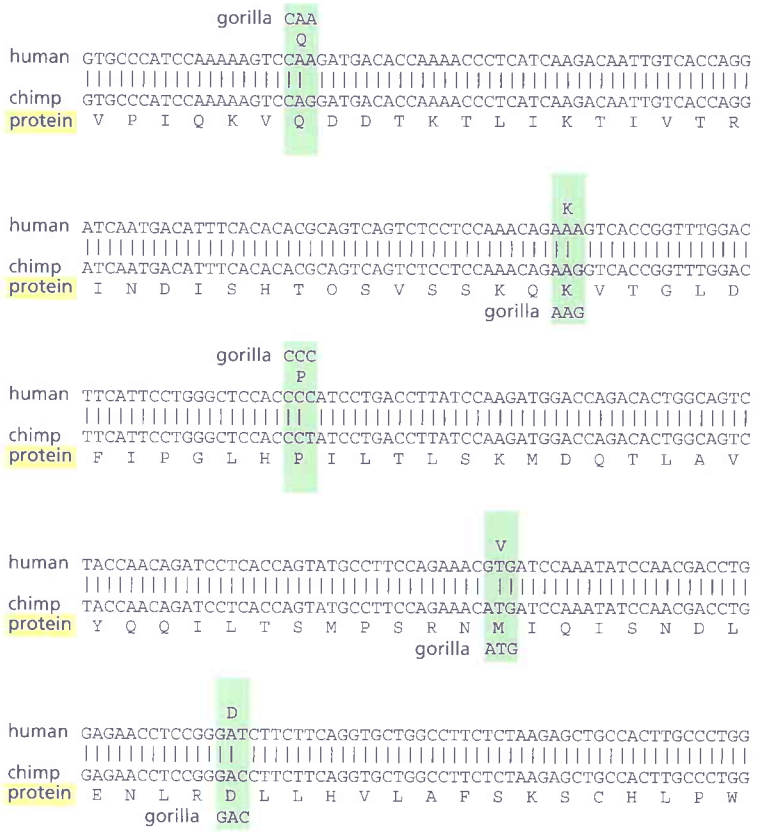


Figure 4–76 Tracing the ancestral sequence from a sequence comparison of the coding regions of human and chimpanzee leptin genes. Leptin is a hormone that regulates food intake and energy utilization in response to the adequacy of fat reserves. As indicated by the codons boxed in green, only 5 nucleotides (of 441 total) differ between these two sequences. Moreover, when the amino acids encoded by both the human and chimpanzee sequences are examined, in only one of the 5 positions does the encoded amino acid differ. For each of the 5 variant nucleotide positions, the corresponding sequence in the gorilla is also indicated. In two cases, the gorilla sequence agrees with the human sequence, while in three cases it agrees with the chimpanzee sequence.

What was the sequence of the leptin gene in the last common ancestor? An evolutionary model that seeks to minimize the number of mutations postulated to have occurred during the evolution of the human and chimpanzee genes would assume that the leptin sequence of the last common ancestor was the same as the human and chimpanzee sequences when they agree; when they disagree, it would use the gorilla sequence as a tie-breaker. For convenience, only the first 300 nucleotides of the leptin coding sequences are given. The remaining 141 are identical between humans and chimpanzees.

Phylogenetic Trees Constructed from a Comparison of DNA Sequences Trace the Relationships of All Organisms

Integration of phylogenetic trees based on molecular sequence comparisons with the fossil record has led to the best available view of the evolution of modern life forms. The fossil record remains important as a source of absolute dates based on the decay of radioisotopes in the rock formations in which fossils are found. However, precise divergence times between species are difficult to establish from the fossil record, even for species that leave good fossils with distinctive morphology.

Such integrated phylogenetic trees suggest that changes in the sequences of particular genes or proteins tend to occur at a nearly constant rate, although rates that differ from the norm by as much as twofold are observed in particular lineages. As discussed above and in Chapter 5, this "molecular clock" runs most rapidly and regularly in sequences that are not subject to purifying selection—such as intergenic regions, portions of introns that lack splicing or regulatory signals, and genes that have been irreversibly inactivated by mutation (the so-called pseudogenes). The clock runs most slowly for sequences that are subject to strong functional constraints—for example, the amino acid sequences of proteins such as actin that engage in specific interactions with large numbers of other proteins and whose structure is therefore highly constrained (see, for example, Figure 16–18).

Occasionally, rapid change is seen in a previously highly conserved sequence. As discussed later in this chapter, such episodes are especially interesting because they are thought to reflect periods of strong positive selection for mutations that conferred a selective advantage in the particular lineage where the rapid change occurred.

Molecular clocks run at rates that are determined both by mutation rates and by the degree of purifying selection on particular sequences. Therefore, a completely different calibration is required for those genes replicated and repaired by different systems within cells. Most notably, in animals, although not in plants, clocks based on functionally unconstrained mitochondrial DNA sequences run

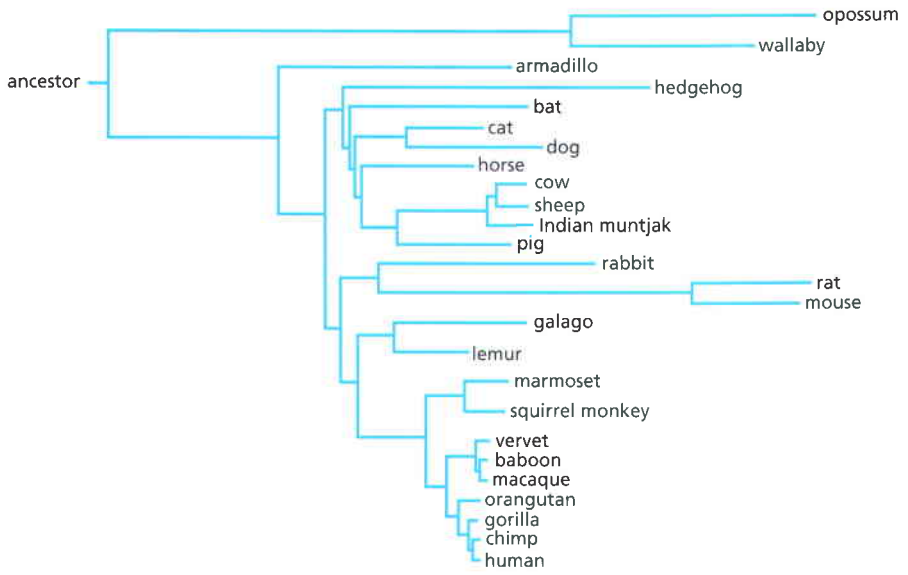


Figure 4–77 A phylogenetic tree highlighting some of the mammals whose genomes are being extensively studied. The length of each line is proportional to the number of "neutral substitutions"—representing the nucleotide changes observed in the absence of purifying selection. (Adapted from G.M. Cooper et al., *Genome Res.* 15:901–913, 2005. With permission from Cold Spring Harbor Laboratory Press.)

much faster than clocks based on functionally unconstrained nuclear sequences, due to an unusually high mutation rate in animal mitochondria.

Molecular clocks have a finer time resolution than the fossil record and are a more reliable guide to the detailed structure of phylogenetic trees than are classical methods of tree construction, which are based on comparisons of the morphology and development of different species. For example, the precise relationship among the great-ape and human lineages was not settled until sufficient molecular-sequence data accumulated in the 1980s to produce the tree that was shown in Figure 4–75. And with huge amounts of DNA sequence now determined from a variety of mammals, much better estimates of our relationship to them are being obtained (**Figure 4–77**).

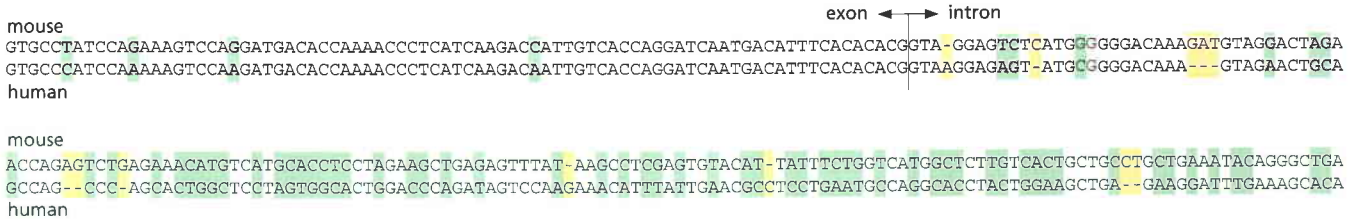
A Comparison of Human and Mouse Chromosomes Shows How The Structures of Genomes Diverge

As would be expected, the human and chimpanzee genomes are much more alike than are the human and mouse genomes. Although the size of the human and mouse genomes are roughly the same and they contain nearly identical sets of genes, there has been a much longer time period over which changes have had a chance to accumulate—approximately 80 million years versus 6 million years. In addition, as indicated in Figure 4–77, rodent lineages (represented by the rat and the mouse) have unusually fast molecular clocks. Hence, these lineages have diverged from the human lineage more rapidly than otherwise expected.

As indicated by the DNA sequence comparison in **Figure 4–78**, mutation has led to extensive sequence divergence between humans and mice at all sites that are not under selection—such as most nucleotide sequences in introns. In contrast, in human–chimpanzee comparisons, nearly all sequence positions are the same simply because not enough time has elapsed since the last common ancestor for large numbers of changes to have occurred.

In contrast to the situation for humans and chimpanzees, local gene order and overall chromosome organization have diverged greatly between humans

Figure 4–78 Comparison of a portion of the mouse and human leptin genes. Positions where the sequences differ by a single nucleotide substitution are boxed in *green*, and positions that differ by the addition or deletion of nucleotides are boxed in *yellow*. Note that the coding sequence of the exon is much more conserved than is the adjacent intron sequence.



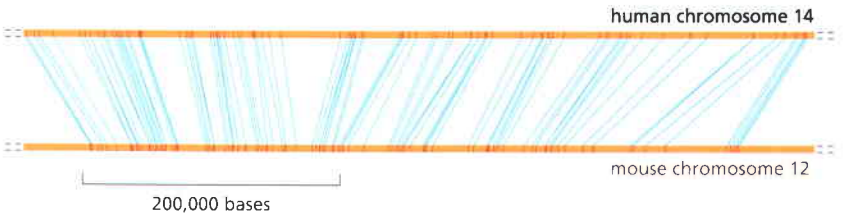


Figure 4–79 **Comparison of a syntenic portion of mouse and human genomes.** About 90 percent of the two genomes can be aligned in this way. Note that while there is an identical order of the matched index sequences (*red marks*), there has been a net loss of DNA in the mouse lineage that is interspersed throughout the entire region. This type of net loss is typical for all such regions, and it accounts for the fact that the mouse genome contains 14 percent less DNA than does the human genome. (Adapted from Mouse Sequencing Consortium, *Nature* 420:520–573, 2002. With permission from Macmillan Publishers Ltd.)

and mice. According to rough estimates, a total of about 180 break-and-rejoin events have occurred in the human and mouse lineages since these two species last shared a common ancestor. In the process, although the number of chromosomes is similar in the two species (23 per haploid genome in the human versus 20 in the mouse), their overall structures differ greatly. Nonetheless, even after the extensive genomic shuffling, there are many large blocks of DNA in which the gene order is the same in the human and the mouse. These stretches of conserved gene order in chromosomes are referred to as regions of *synteny.*

An unexpected conclusion from a detailed comparison of the complete mouse and human genome sequences, confirmed from subsequent comparisons between the genomes of other vertebrates, is that small blocks of sequences are being deleted from and added to genomes at a surprisingly rapid rate. Thus, if we assume that our common ancestor had a genome of human size (about 3 billion nucleotide pairs), mice would have lost a total of about 45 percent of that genome from accumulated deletions during the past 80 million years, while humans would have lost about 25 percent. However, substantial sequence gains from many small chromosome duplications and from the multiplication of transposons have compensated for these deletions. As a result, our genome size is unchanged from that of the last common ancestor for humans and mice, while the mouse genome is smaller by only 0.3 billion nucleotides.

Good evidence for the loss of DNA sequences in small blocks during evolution can be obtained from a detailed comparison of most regions of synteny in the human and mouse genomes. The comparative shrinkage of the mouse genome can be clearly seen from such comparisons, with the net loss of sequences scattered throughout the long stretches of DNA that are otherwise homologous (**Figure 4–79**).

DNA is added to genomes both by the spontaneous duplication of chromosomal segments that contain tens of thousands of nucleotide pairs (as will be discussed shortly), and by active transposition (most transposition events are duplicative, because the original copy of the transposon stays where it was when a copy inserts at the new site; for example, see Figure 5–74). Comparison of the DNA sequences derived from transposons in the human and the mouse therefore readily reveals some of the sequence additions (**Figure 4–80**).

For unknown reasons, all mammals have genome sizes of about 3 billion nucleotide pairs that contain nearly identical sets of genes, even though only on the order of 150 million nucleotide pairs appear to be under sequence-specific functional constraints.

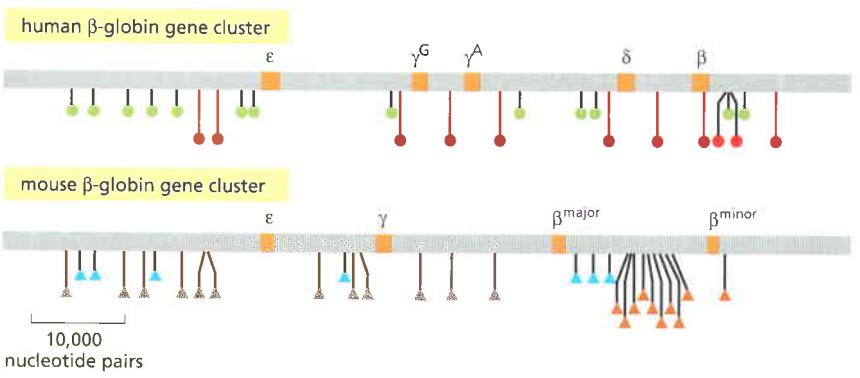


Figure 4–80 **A comparison of the β-globin gene cluster in the human and mouse genomes, showing the location of transposable elements.** This stretch of human genome contains five functional β-globin-like genes *(orange);* the comparable region from the mouse genome has only four. The positions of the human Alu sequence are indicated by green circles, and the human L1 sequences by red circles. The mouse genome contains different but related transposable elements: the positions of B1 elements (which are related to the human Alu sequences) are indicated by *blue triangles,* and the positions of the mouse L1 elements (which are related to the human L1 sequences) are indicated by *orange triangles.* The absence of transposable elements from the globin structural genes can be attributed to purifying selection, which would have eliminated any insertion that compromised gene function. (Courtesy of Ross Hardison and Webb Miller.)

The Size of a Vertebrate Genome Reflects the Relative Rates of DNA Addition and DNA Loss in a Lineage

Now that we know the complete sequence of a number of vertebrate genomes, we see that genome size can vary considerably, apparently without a drastic effect on the organism or its number of genes. Thus, while the mouse and dog genomes are both in the typical mammalian size range, the chicken has a genome that is only about one-third human size (one billion nucleotide pairs). A particularly notable example of an organism with a genome of anomalous size is the puffer fish, *Fugu rubripes* (**Figure 4–81**), which has a tiny genome for a vertebrate (0.4 billion nucleotide pairs compared to 1 billion or more for many other fish). The small size of the *Fugu* genome is largely due to the small size of its introns. Specifically, *Fugu* introns, as well as other noncoding segments of the *Fugu* genome, lack the repetitive DNA that makes up a large portion of the genomes of most well-studied vertebrates. Nevertheless, the positions of *Fugu* introns are nearly perfectly conserved relative to their positions in mammalian genomes (**Figure 4–82**).

While initially a mystery, we now have a simple explanation for such large differences in genome size between similar organisms: because all vertebrates experience a continuous process of DNA loss and DNA addition, the size of a genome merely depends on the balance between these opposing processes acting over millions of years. Suppose, for example, that in the lineage leading to *Fugu*, the rate of DNA addition happened to slow greatly. Over long periods of time, this would result in a major "cleansing" from this fish genome of those DNA sequences whose loss could be tolerated. In retrospect, the process of purifying selection in the *Fugu* lineage has partitioned those vertebrate DNA sequences most likely to be functional into only 400 million nucleotide pairs of DNA, providing a major resource for scientists.

Figure 4–81 The puffer fish, *Fugu rubripes*. (Courtesy of Byrappa Venkatesh.)

We Can Reconstruct the Sequence of Some Ancient Genomes

The genomes of ancestral organisms can be inferred, but never directly observed: there are no ancient organisms alive today. Although a modern organism such as the horseshoe crab looks remarkably similar to fossil ancestors that lived 200 million years ago, there is every reason to believe that the horseshoe-crab genome has been changing during all that time at a rate similar to that occurring in other evolutionary lineages. Selection constraints must have maintained key functional properties of the horseshoe-crab genome to account for the morphological stability of the lineage. However, genome sequences reveal that the fraction of the genome subject to purifying selection is small; hence the genome of the modern horseshoe crab must differ greatly from that of its extinct ancestors, known to us only through the fossil record.

Is there any way around this problem? Can we ever hope to decipher large sections of the genome sequence of the extinct ancestors of organisms that are

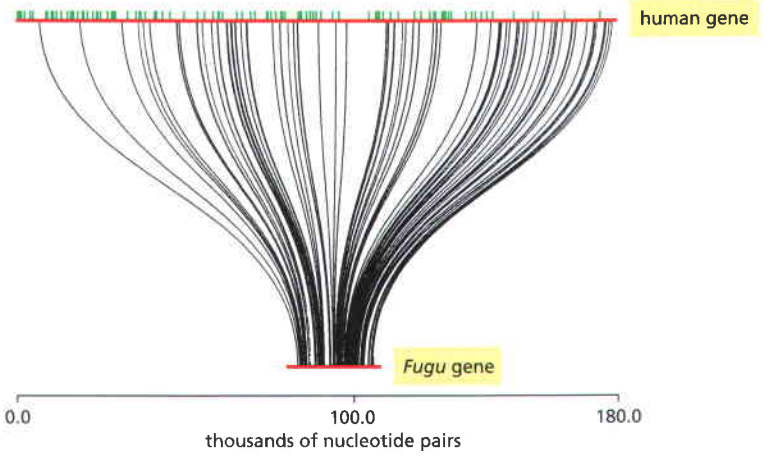


Figure 4–82 Comparison of the genomic sequences of the human and *Fugu* genes encoding the protein huntingtin. Both genes (indicated in *red*) contain 67 short exons that align in 1:1 correspondence to one another; these exons are connected by curved lines. The human gene is 7.5 times larger than the *Fugu* gene (180,000 versus 27,000 nucleotide pairs). The size difference is entirely due to larger introns in the human gene. The larger size of the human introns is due in part to the presence of retrotransposons, whose positions are represented by green vertical lines; the *Fugu* introns lack retrotransposons. In humans, mutation of the huntingtin gene causes Huntington's disease, an inherited neurodegenerative disorder. (Adapted from S. Baxendale et al., *Nat. Genet.* 10:67–76, 1995. With permission from Macmillan Publishers Ltd.)

alive today? For organisms that are as closely related as human and chimp, we saw that this may not be difficult. In that case, reference to the gorilla sequence can be be used to sort out which of the few differences between human and chimp DNA sequences was inherited from our common ancestor some 6 million years ago (see Figure 4–76). For an ancestor that has produced a large number of different organisms alive today, the DNA sequences of many species can be compared simultaneously to unscramble the ancestral sequence, allowing scientists to trace DNA sequences much farther back in time. For example, from the complete genome sequences of 20 modern mammals that will soon be obtained, it should be possible to decipher most of the genome sequence of the 100 million year-old Boreoeutherian mammal that gave rise to species as diverse as dog, mouse, rabbit, armadillo and human (see Figure 4–77).

Multispecies Sequence Comparisons Identify Important DNA Sequences of Unknown Function

The massive quantity of DNA sequence now in databases (more than a hundred billion nucleotide pairs) provides a rich resource that scientists can mine for many purposes. We have already discussed how this information can be used to unscramble the evolutionary pathways that have led to modern organisms. But sequence comparisons also provide many insights into how cells and organisms function. Perhaps the most remarkable discovery in this realm has been the observation that, although only about 1.5% of the human genome codes for proteins, about three times this amount (in total, 5% of the genome—see Table 4–1, p. 206) has been strongly conserved during mammalian evolution. This mass of conserved sequence is most clearly revealed when we align and compare DNA synteny blocks from many different species. In this way, so-called *multispecies conserved sequences* can be readily identified (**Figure 4–83**). Most of the noncoding conserved sequences discovered in this way turn out to be relatively short, containing between 50 and 200 nucleotide pairs. The strict conservation

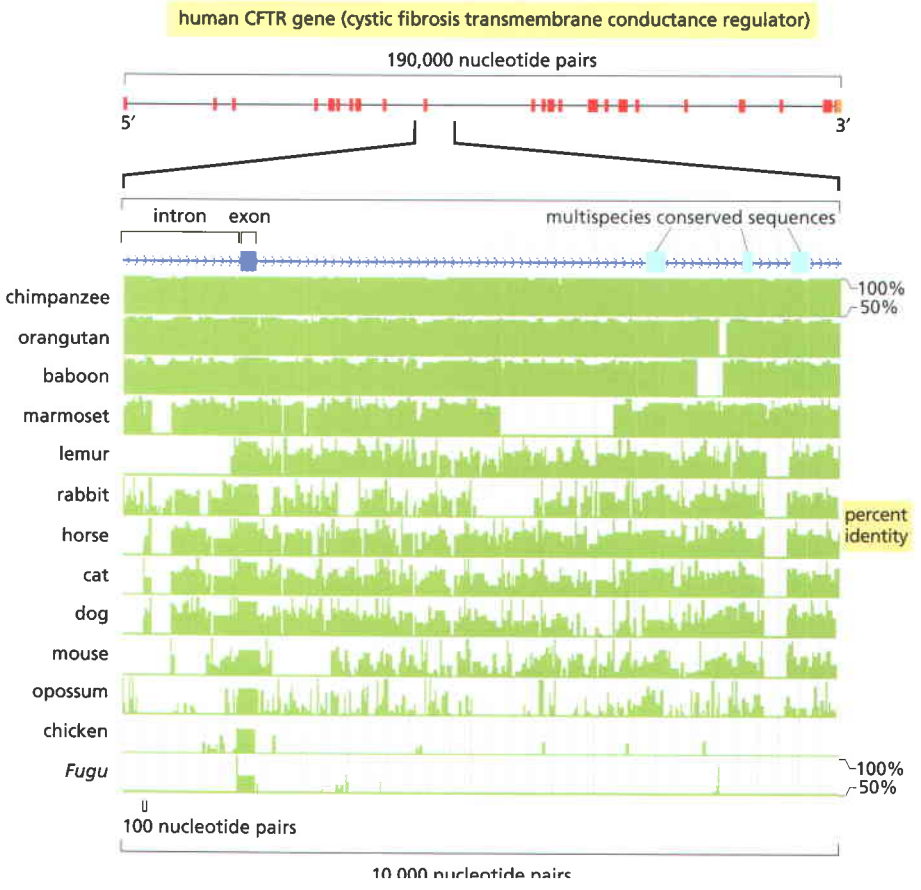


Figure 4–83 The detection of multispecies conserved sequences. In this example, genome sequences for each of the organisms shown have been compared with the indicated region of the human CFTR gene, scanning in 25 nucleotide blocks. For each organism, the percent identity across its syntenic sequences is plotted in *green*. In addition, a computational algorithm has been used to detect the sequences within this region that are most highly conserved when the sequences from all of the organisms are taken into account. Besides the exon, three other blocks of multispecies conserved sequences are shown. The function of most such sequences in the human genome is not known. (Courtesy of Eric D. Green.)

implies that they have important functions that have been maintained by purifying selection. The puzzle is to unravel what those functions are. Some of the conserved sequence that does not code for protein codes for untranslated RNA molecules that are known to have important functions, as we shall see in later chapters. Another fraction of the noncoding conserved DNA is certainly involved in regulating the transcription of adjacent genes, as discussed in Chapter 7. But we do not yet know how much of the conserved DNA can be accounted for in these ways, and the bulk of it is still a deep mystery. The solution to this mystery is bound to have profound consequences for medicine, and it reveals how much more we need to learn about the biology of vertebrate organisms.

How can cell biologists tackle this problem? The first step is to distinguish between the conserved regions that code for protein and those that do not, and then, among the latter, to focus on those that do not already have some other identified function, in coding for structural RNA molecules, for example. The next task is to discover what proteins or RNA molecules bind to these mysterious DNA sequences, how they are packaged into chromatin, and whether they ever serve as templates for RNA synthesis. Most of this task still lies before us, but a start has been made, and some remarkable insights have been obtained. One of the most intriguing concerns the evolutionary changes that have made us humans different from other animals—changes, that is, in sequences that have been conserved among our close relatives but have undergone sudden rapid change in the human sublineage.

Accelerated Changes in Previously Conserved Sequences Can Help Decipher Critical Steps in Human Evolution

As soon as both the human and the chimpanzee genome sequences became available, scientists began searching for DNA sequence changes that might account for the striking differences between us and them. With 3 billion nucleotide pairs to compare in the two species, this might seem an impossible task. But the job was made much easier by confining the search to 35,000 clearly defined multispecies conserved sequences (a total of about 5 million nucleotide pairs), representing parts of the genome that are most likely to be functionally important. Though these sequences are conserved strongly, they are not conserved perfectly, and when the version in one species is compared with that in another they are generally found to have drifted apart by a small amount corresponding simply to the time elapsed since the last common ancestor. In a small proportion of cases, however, one sees signs of a sudden evolutionary spurt. For example, some DNA sequences that have been highly conserved in other mammalian species are found to have changed exceptionally fast during the six million years of human evolution since we diverged from the chimpanzees. Such *human accelerated regions* (HARs) are thought to reflect functions that have been especially important in making us different in some useful way.

About 50 such sites were identified in one study, one-fourth of which were located near genes associated with neural development. The sequence exhibiting the most rapid change (18 changes between human and chimp, compared to only two changes between chimp and chicken) was examined further and found to encode a 118-nucleotide noncoding RNA molecule that is produced in the human cerebral cortex at a critical time during brain development (**Figure 4–84**). Although the function of this HAR1F RNA is not yet known, this exciting finding is stimulating further studies that will hopefully shed light on crucial features of the human brain.

Gene Duplication Provides an Important Source of Genetic Novelty During Evolution

Evolution depends on the creation of new genes, as well as on the modification of those that already exist. How does this occur? When we compare organisms that seem very different—a primate with a rodent, for example, or a mouse with

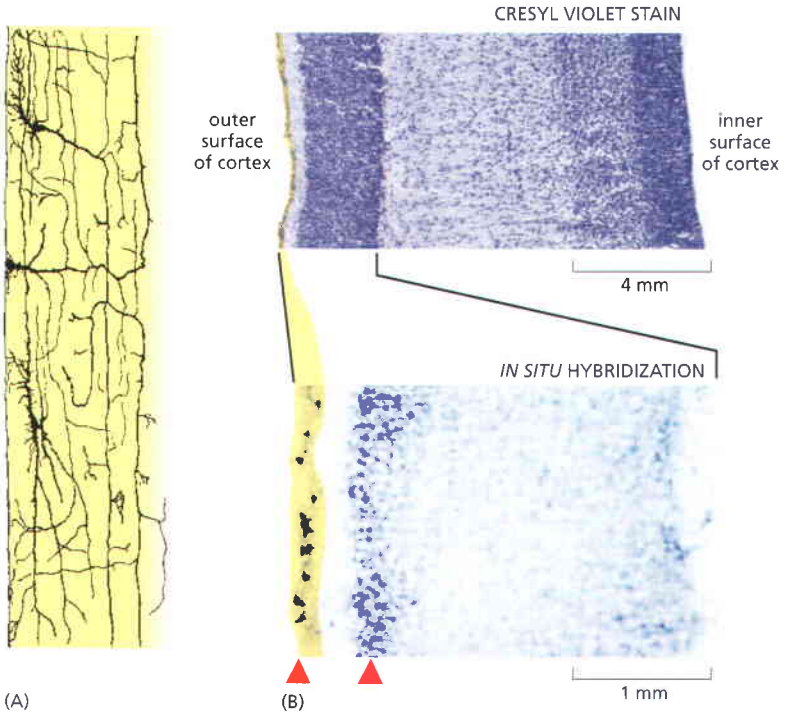


Figure 4–84 Initial characterization of a new gene detected as a previously conserved DNA sequence that evolved rapidly in humans. (A) Drawing by Ramon y Cajal of the outer surface of the human neocortex, highlighting the Cajal–Retzius neurons. (B) Tissue slices from an embryonic human brain showing part of the cortex, with the region containing the Cajal–Retzius neurons highlighted in *yellow*. Upper photograph: cresyl violet stain. Lower photograph: *in situ* hybridization. The *red arrows* indicate the cells that produce HAR1F RNA as detected by *in situ* hybridization *(blue)*. HAR1F is a novel noncoding RNA that has evolved rapidly in the human lineage leading from the great apes. The Cajal–Retzius neurons make this RNA at the time when the neocortex is developing. The results are intriguing, because a large neocortex is special to humans; for the behavior of cells in forming this cortex, see Figure 22–99. (Adapted from K.S. Pollard et al., *Nature* 443:167–172, 2006. With permission from Macmillan Publishers Ltd.)

a fish—we rarely encounter genes in the one species that have no homolog in the other. Genes without homologous counterparts are relatively scarce even when we compare such divergent organisms as a mammal and a worm. On the other hand, we frequently find gene families that have different numbers of members in different species. To create such families, genes have been repeatedly duplicated, and the copies have then diverged to take on new functions that often vary from one species to another.

The genes encoding nuclear hormone receptors in humans, a nematode worm, and a fruit fly illustrate this point (**Figure 4–85**). Many of the subtypes of these nuclear receptors (also called intracellular receptors) have close homologs in all three organisms that are more similar to each other than they are to other family subtypes present in the same species. Therefore, much of the functional divergence of this large gene family must have preceded the divergence of these three evolutionary lineages. Subsequently, one major branch of the gene family underwent an enormous expansion in the worm lineage only. Similar, but smaller, lineage-specific expansions of particular subtypes are evident throughout the gene family tree.

Gene duplication occurs at high rates in all evolutionary lineages, contributing to the vigorous process of DNA addition discussed previously. In a detailed study of spontaneous duplications in yeast, duplications of 50,000 to 250,000 nucleotide pairs were commonly observed, most of which were tandemly repeated. These appeared to result from DNA replication errors that led to the inexact repair of double-strand chromosome breaks. A comparison of the human and chimpanzee genomes reveals that, since the time that these two organisms diverged, segmental duplications have added about 5 million nucleotide pairs to each genome every million years, with an average duplication size being about 50,000 nucleotide pairs (however, there are duplications five times larger, as in yeast). In fact, if one counts nucleotides, duplication events have created more differences between our two species than have single nucleotide substitutions.

Duplicated Genes Diverge

A major question in genome evolution concerns the fate of newly duplicated genes. In most cases, there is presumed to be little or no selection—at least initially—to maintain the duplicated state since either copy can provide an equiv-

alent function. Hence, many duplication events are likely to be followed by loss-of-function mutations in one or the other gene. This cycle would functionally restore the one-gene state that preceded the duplication. Indeed, there are many examples in contemporary genomes where one copy of a duplicated gene can be seen to have become irreversibly inactivated by multiple mutations. Over time, the sequence similarity between such a **pseudogene** and the functional gene whose duplication produced it would be expected to be eroded by the accumulation of many mutations in the pseudogene—the homologous relationship eventually becoming undetectable.

An alternative fate for gene duplications is for both copies to remain functional, while diverging in their sequence and pattern of expression, thus taking on different roles. This process of "duplication and divergence" almost certainly explains the presence of large families of genes with related functions in biologically complex organisms, and it is thought to play a critical role in the evolution of increased biological complexity. An examination of many different eucaryotic genomes suggests that the probability that any particular gene will undergo a duplication event that spreads to most or all individuals in a species is approximately 1% every million years.

Whole-genome duplications offer particularly dramatic examples of the duplication–divergence cycle. A whole-genome duplication can occur quite simply: all that is required is one round of genome replication in a germline cell lineage without a corresponding cell division. Initially, the chromosome number simply doubles. Such abrupt increases in the ploidy of an organism are common, particularly in fungi and plants. After a whole-genome duplication, all genes exist as duplicate copies. However, unless the duplication event occurred so recently that there has been little time for subsequent alterations in genome structure, the results of a series of segmental duplications—occurring at different times—are very hard to distinguish from the end product of a whole-genome duplication. In mammals, for example, the role of whole-genome duplications versus a series of piecemeal duplications of DNA segments is quite uncertain. Nevertheless, it is clear that a great deal of gene duplication has ocurred in the distant past.

Analysis of the genome of the zebrafish, in which either a whole-genome duplication or a series of more local duplications occurred hundreds of millions of years ago, has cast some light on the process of gene duplication and divergence. Although many duplicates of zebrafish genes appear to have been lost by mutation, a significant fraction—perhaps as many as 30–50%—have diverged functionally while both copies have remained active. In many cases, the most obvious functional difference between the duplicated genes is that they are expressed in different tissues or at different stages of development (see Figure 22–46). One attractive theory to explain such an end result imagines that different, mildly deleterious mutations occur quickly in both copies of a duplicated gene set. For example, one copy might lose expression in a particular tissue as a result of a regulatory mutation, while the other copy loses expression in a second tissue. Following such an occurrence, both gene copies would be required to provide the full range of functions that were once supplied by a single gene; hence, both copies would now be protected from loss through inactivating mutations. Over a longer period, each copy could then undergo further changes through which it could acquire new, specialized features.

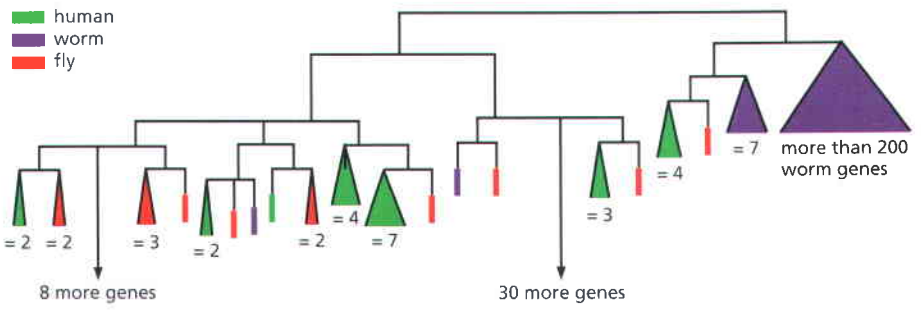


Figure 4–85 A phylogenetic tree based on the inferred protein sequences for all nuclear hormone receptors encoded in the genomes of human (*H. sapiens*), a nematode worm (*C. elegans*), and a fruit fly (*D. melanogaster*). Triangles represent protein subfamilies that have expanded within individual evolutionary lineages; the width of these triangles indicates the number of genes encoding members of these subfamilies. Colored vertical bars represent a single gene. There is no simple pattern to the historical duplications and divergences that have created the gene families encoding nuclear receptors in the three contemporary organisms. The family of nuclear hormone receptors is described in Figure 15–14. These proteins function in cell signaling and gene regulation. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

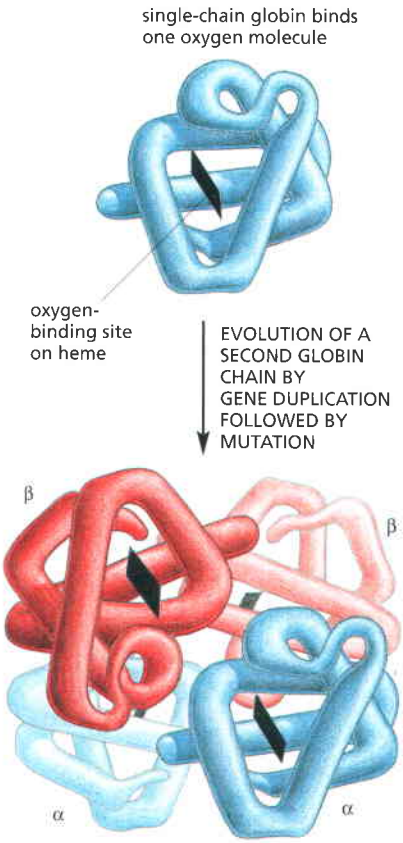


Figure 4–86 A comparison of the structure of one-chain and four-chain globins. The four-chain globin shown is hemoglobin, which is a complex of two α-globin and two β-globin chains. The one-chain globin in some primitive vertebrates forms a dimer that dissociates when it binds oxygen, representing an intermediate in the evolution of the four-chain globin.

The Evolution of the Globin Gene Family Shows How DNA Duplications Contribute to the Evolution of Organisms

The globin gene family provides an especially good example of how DNA duplication generates new proteins, because its evolutionary history has been worked out particularly well. The unmistakable similarities in amino acid sequence and structure among the present-day globins indicate that they all must derive from a common ancestral gene, even though some are now encoded by widely separated genes in the mammalian genome.

We can reconstruct some of the past events that produced the various types of oxygen-carrying hemoglobin molecules by considering the different forms of the protein in organisms at different positions on the phylogenetic tree of life. A molecule like hemoglobin was necessary to allow multicellular animals to grow to a large size, since large animals could no longer rely on the simple diffusion of oxygen through the body surface to oxygenate their tissues adequately. Consequently, hemoglobin-like molecules are found in all vertebrates and in many invertebrates. The most primitive oxygen-carrying molecule in animals is a globin polypeptide chain of about 150 amino acids, which is found in many marine worms, insects, and primitive fish. The hemoglobin molecule in more complex vertebrates, however, is composed of two kinds of globin chains. It appears that about 500 million years ago, during the continuing evolution of fish, a series of gene mutations and duplications occurred. These events established two slightly different globin genes, coding for the α- and β-globin chains, in the genome of each individual. In modern vertebrates, each hemoglobin molecule is a complex of two α chains and two β chains (**Figure 4–86**). The four oxygen-binding sites in the $\alpha_2\beta_2$ molecule interact, allowing a cooperative allosteric change in the molecule as it binds and releases oxygen, which enables hemoglobin to take up and release oxygen more efficiently than the single-chain version.

Still later, during the evolution of mammals, the β-chain gene apparently underwent duplication and mutation to give rise to a second β-like chain that is synthesized specifically in the fetus. The resulting hemoglobin molecule has a higher affinity for oxygen than adult hemoglobin and thus helps in the transfer of oxygen from the mother to the fetus. The gene for the new β-like chain subsequently duplicated and mutated again to produce two new genes, ε and γ, the ε chain being produced earlier in development (to form $\alpha_2\varepsilon_2$) than the fetal γ chain, which forms $\alpha_2\gamma_2$. A duplication of the adult β-chain gene occurred still later, during primate evolution, to give rise to a δ-globin gene and thus to a minor form of hemoglobin ($\alpha_2\delta_2$) that is found only in adult primates (**Figure 4–87**).

Each of these duplicated genes has been modified by point mutations that affect the properties of the final hemoglobin molecule, as well as by changes in regulatory regions that determine the timing and level of expression of the gene. As a result, each globin is made in different amounts at different times of human development (see Figure 7–64B).

The end result of the gene duplication processes that have given rise to the diversity of globin chains is seen clearly in the human genes that arose from

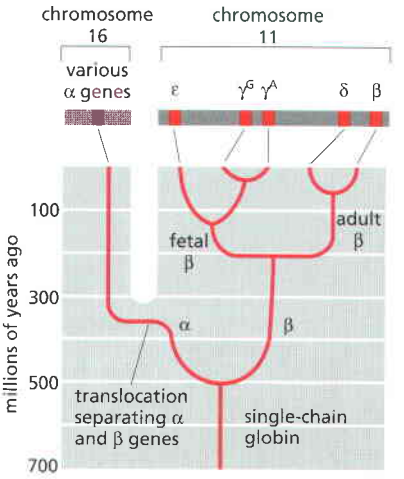


Figure 4–87 An evolutionary scheme for the globin chains that carry oxygen in the blood of animals. The scheme emphasizes the β-like globin gene family. A relatively recent gene duplication of the γ-chain gene produced γ^G and γ^A, which are fetal β-like chains of identical function. The location of the globin genes in the human genome is shown at the top of the figure (see also Figure 7–64).

the original β gene, which are arranged as a series of homologous DNA sequences located within 50,000 nucleotide pairs of one another. A similar cluster of α-globin genes is located on a separate human chromosome. Because the α- and β-globin gene clusters are on separate chromosomes in birds and mammals but are together in the frog *Xenopus*, it is believed that a chromosome translocation event separated the two gene clusters about 300 million years ago (see Figure 4–87).

There are several duplicated globin DNA sequences in the α- and β-globin gene clusters that are not functional genes but pseudogenes. These have a close sequence similarity to the functional genes but have been disabled by mutations that prevent their expression. The existence of such pseudogenes makes it clear that, as expected, not every DNA duplication leads to a new functional gene. We also know that nonfunctional DNA sequences are not rapidly discarded, as indicated by the large excess of noncoding DNA that is found in mammalian genomes.

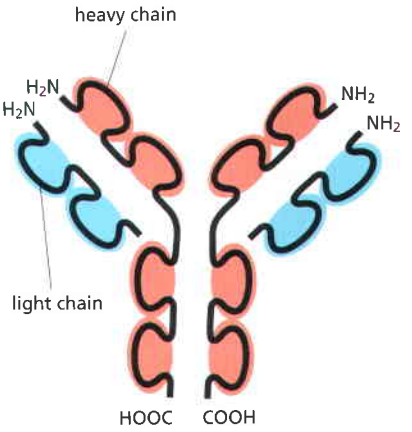


Figure 4–88 Schematic view of an antibody (immunoglobulin) molecule. This molecule is a complex of two identical heavy chains and two identical light chains. Each heavy chain contains four similar, covalently linked domains, while each light chain contains two such domains. Each domain is encoded by a separate exon, and all of the exons are thought to have evolved by the serial duplication of a single ancestral exon.

Genes Encoding New Proteins Can Be Created by the Recombination of Exons

The role of DNA duplication in evolution is not confined to the expansion of gene families. It can also act on a smaller scale to create single genes by stringing together short duplicated segments of DNA. The proteins encoded by genes generated in this way can be recognized by the presence of repeating similar protein domains, which are covalently linked to one another in series. The immunoglobulins (**Figure 4–88**) and albumins, for example, as well as most fibrous proteins (such as collagens) are encoded by genes that have evolved by repeated duplications of a primordial DNA sequence.

In genes that have evolved in this way, as well as in many other genes, each separate exon often encodes an individual protein folding unit, or domain. It is believed that the organization of DNA coding sequences as a series of such exons separated by long introns has greatly facilitated the evolution of new proteins. The duplications necessary to form a single gene coding for a protein with repeating domains, for example, can often occur by breaking and rejoining the DNA anywhere in the long introns on either side of an exon; without introns there would be only a few sites in the original gene at which a recombinational exchange between DNA molecules could duplicate the domain. By enabling the duplication to occur by recombination at many potential sites rather than just a few, introns increase the probability of a favorable duplication event.

More generally, we know from genome sequences that the various parts of genes—both their individual exons and their regulatory elements—have served as modular elements that have been duplicated and moved about the genome to create the great diversity of living things. Thus, for example, many present-day proteins are formed as a patchwork of domains from different origins, reflecting their long evolutionary history (see Figure 3–19).

Neutral Mutations Often Spread to Become Fixed in a Population, with a Probability that Depends on Population Size

In comparisons between two species that have diverged from one another by millions of years, it makes little difference which individuals from each species are compared. For example, typical human and chimpanzee DNA sequences differ from one another by about 1%. In contrast, when the same region of the genome is sampled from two different humans, the differences are typically less than 0.1%. For more distantly related organisms, the inter-species differences overshadow intra-species variation even more dramatically. However, each "fixed difference" between the human and the chimpanzee (in other words, each difference that is now characteristic of all or nearly all individuals of each species) started out as a new mutation in a single individual. If the size of the

interbreeding population in which the mutation occurred is N, the initial allele frequency of a new mutation would be $1/(2N)$ for a diploid organism. How does such a rare mutation become fixed in the population, and hence become a characteristic of the species rather than of a particular individual genome?

The answer to this question depends on the functional consequences of the mutation. If the mutation has a significantly deleterious effect, it will simply be eliminated by purifying selection and will not become fixed. (In the most extreme case, the individual carrying the mutation will die without producing progeny.) Conversely, the rare mutations that confer a major reproductive advantage on individuals who inherit them can spread rapidly in the population. Because humans reproduce sexually and genetic recombination occurs each time a gamete is formed (discussed in Chapter 5), the genome of each individual who has inherited the mutation will be a unique recombinational mosaic of segments inherited from a large number of ancestors. The selected mutation along with a modest amount of neighboring sequence—ultimately inherited from the individual in which the mutation occurred—will simply be one piece of this huge mosaic.

The great majority of mutations that are not harmful are not beneficial either. These selectively neutral mutations can also spread and become fixed in a population, and they make a large contribution to the evolutionary change in genomes. Their spread is not as rapid as the spread of the rare strongly advantageous mutations. The process by which such neutral genetic variation is passed down through an idealized interbreeding population can be described mathematically by equations that are surprisingly simple. The idealized model that has proven most useful for analyzing human genetic variation assumes a constant population size and random mating, as well as selective neutrality for the mutations. While neither of the first two assumptions is a good description of human population history, they nonetheless provide a useful starting point for analyzing intra-species variation.

When a new neutral mutation occurs in a constant population of size N that is undergoing random mating, the probability that it will ultimately become fixed is approximately $1/(2N)$. For those mutations that do become fixed, the average time to fixation is approximately $4N$ generations. A detailed analysis of data on human genetic variation suggests an ancestral population size of approximately 10,000 during the period when the current pattern of genetic variation was largely established. With a population that has reached this size, the probability that a new, selectively neutral mutation would become fixed is small (5×10^{-5}), while the average time to fixation would be on the order of 800,000 years (assuming a 20-year generation time). Thus, while we know that the human population has grown enormously since the development of agriculture approximately 15,000 years ago, most of the present-day set of common human genetic variants reflects the mixture of variants that was already present long before this time, when the human population was still small enough to allow their widespread dissemination.

A Great Deal Can Be Learned from Analyses of the Variation Among Humans

Even though most of the variation among modern humans originates from variation present in a comparatively tiny group of ancestors, the number of variations encountered is very large. One important source of variation, which was missed for many years, is the presence of many duplications and deletions of large blocks of DNA. According to one estimate, when any individual human is compared with the standard reference genome in the database, one should expect to find roughly 100 differences involving long sequence blocks. Some of these "copy number variations" will be very common (**Figure 4–89**), while others will be present in only a minority of humans (**Figure 4–90**). From an initial sampling, nearly half will contain known genes. In retrospect this type of variation is not surprising, given the extensive history of DNA addition and DNA loss in vertebrate genomes (for example, see Figure 4–79).

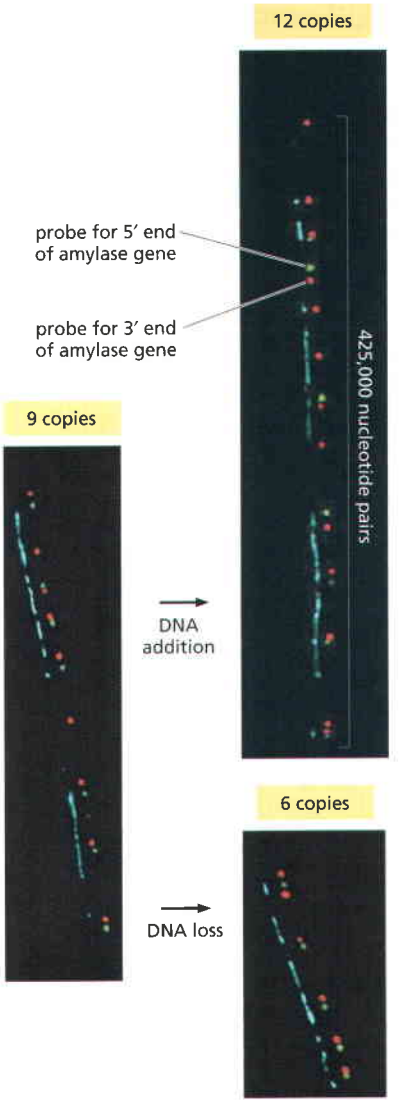


Figure 4–89 Visualization of a frequent type of variation among humans. About half of the humans tested had nine copies of the amylase gene (*left*), which produces an important enzyme that digests starch. In other humans, there has been either DNA loss or DNA addition to produce an altered chromosome, resulting from the deletion (loss) or the duplication (addition) of a part of this region. To obtain these images, stretched chromatin fibers have been hybridized with differently colored probes to the two ends of the amylase gene, as indicated. The *blue* lines mark the general paths of the chromatin. They have been determined by a second stain and displaced to one side for clarity. (Adapted from A.J. Iafrate et al., *Nat. Genet.* 36:949–951, 2004. With permission from Macmillan Publishers Ltd.)

The intra-species variations that have been most extensively characterized are **single-nucleotide polymorphisms** (**SNPs**). These are simply points in the genome sequence where one large fraction of the human population has one nucleotide, while another substantial fraction has another. Two human genomes sampled from the modern world population at random will differ at approximately 2.5×10^6 such sites (1 per 1300 nucleotide pairs). As will be described in the overview of genetics in Chapter 8, mapped sites in the human genome that are **polymorphic**—meaning that there is a reasonable probability (generally more than 1%) that the genomes of two individuals will differ at that site—are extremely useful for genetic analyses, in which one attempts to associate specific traits (phenotypes) with specific DNA sequences for medical or scientific purposes (see p. 560).

Against the background of ordinary SNPs inherited from our prehistoric ancestors, certain sequences with exceptionally high mutation rates stand out. A dramatic example is provided by CA repeats, which are ubiquitous in the human genome and in the genomes of other eucaryotes. Sequences with the motif $(CA)_n$ are replicated with relatively low fidelity because of a slippage that occurs between the template and the newly synthesized strands during DNA replication; hence, the precise value of n can vary over a considerable range from one genome to the next. These repeats make ideal DNA-based genetic markers, since most humans are heterozygous—carrying two values of n at any particular CA repeat, having inherited one repeat length (n) from their mother and a different repeat length from their father. While the value of n changes sufficiently rarely that most parent–child transmissions propagate CA repeats faithfully, the changes are sufficiently frequent to maintain high levels of heterozygosity in the human population. These and some other simple repeats that display exceptionally high variability therefore provide the basis for identifying individuals by DNA analysis in crime investigations, paternity suits, and other forensic applications (see Figure 8–47).

While most of the SNPs and copy number variations in the human genome sequence are thought to have no effect on phenotype, a subset of them must be

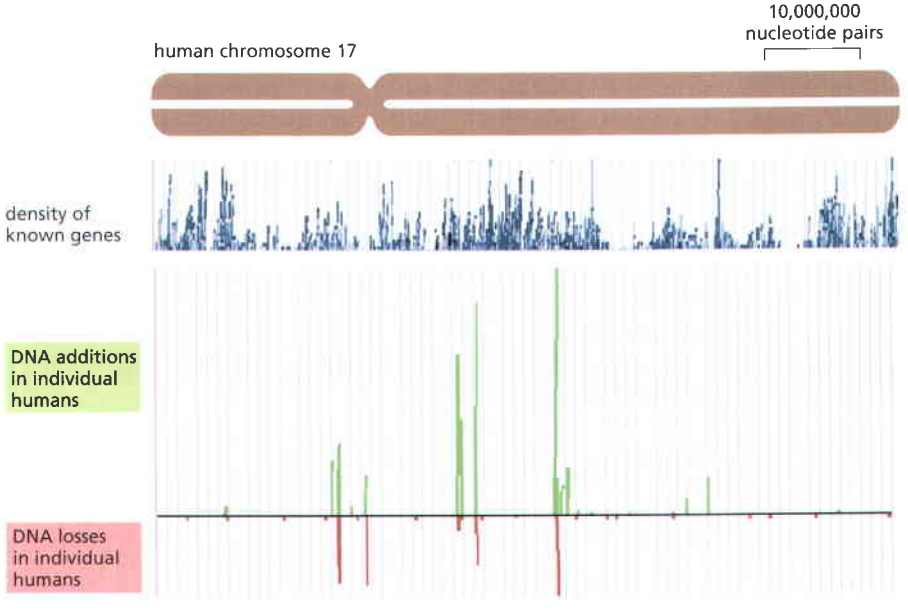


Figure 4–90 Detection of copy number variants on human chromosome 17. When 100 individuals were tested by a DNA microarray analysis that detects the copy number of DNA sequences throughout the entire length of this chromosome, the indicated distributions of DNA additions (*green bars*) and DNA losses (*red bars*) were observed compared with an arbitrary human reference sequence. The shortest *red* and *green bars* represent a single occurrence among the 200 chromosomes examined, whereas the longer bars indicate that the addition or loss was correspondingly more frequent. The results show preferred regions where the variations occur, and these tend to be in or near regions that already contain blocks of segmental duplications. Many of the changes include known genes. (Adapted from J.L. Freeman et al., *Genome Res.* 16:949–961, 2006. With permission from Cold Spring Harbor Laboratory Press.)

responsible for nearly all the heritable aspects of human individuality. We know that even a single nucleotide change that alters one amino acid in a protein can cause a serious disease, as for example in sickle cell anemia, which is caused by such a mutation in hemoglobin. <TTTT> We also know that gene dosage—a doubling or halving of the copy number of some genes—can have a profound effect on human development by altering the level of gene product. There is therefore every reason to suppose that some of the many differences between any two human beings will have substantial effects on human health, physiology, and behavior, whether they be SNPs or copy number variations. The major challenge in human genetics is to learn to recognize those relatively few variations that are functionally important against a large background of neutral variation in the genomes of different humans.

Summary

Comparisons of the nucleotide sequences of present-day genomes have revolutionized our understanding of gene and genome evolution. Because of the extremely high fidelity of DNA replication and DNA repair processes, random errors in maintaining the nucleotide sequences in genomes occur so rarely that only about one nucleotide in 1000 is altered every million years in any particular line of descent. Not surprisingly, therefore, a comparison of human and chimpanzee chromosomes—which are separated by about 6 million years of evolution—reveals very few changes. Not only are our genes essentially the same, but their order on each chromosome is almost identical. Although a substantial number of segmental duplications and segmental deletions have occurred in the past 6 million years, even the positions of the transposable elements that make up a major portion of our noncoding DNA are mostly unchanged.

When one compares the genomes of two more distantly related organisms—such as a human and a mouse, separated by about 80 million years—one finds many more changes. Now the effects of natural selection can be clearly seen: through purifying selection, essential nucleotide sequences—both in regulatory regions and in coding sequences (exon sequences)—have been highly conserved. In contrast, nonessential sequences (for example, much of the DNA in introns) have been altered to such an extent that an accurate alignment according to ancestry is frequently not possible.

Because of purifying selection, the comparison of the genome sequences of multiple related species is an especially powerful way to find DNA sequences with important functions. Although about 5% of the human genome has been conserved as a result of purifying selection, the function of the majority of this DNA (tens of thousands of multispecies conserved sequences) remains mysterious. Future experiments characterizing their functions should teach us a great deal about vertebrate biology.

Other sequence comparisons show that a great deal of the genetic complexity of present-day organisms is due to the expansion of ancient gene families. DNA duplication followed by sequence divergence has clearly been a major source of genetic novelty during evolution. The genomes of any two humans will differ from each other both because of nucleotide substitutions (single nucleotide polymorphisms, or SNPs) and because of inherited DNA gains and DNA losses that cause copy number variants. Understanding these differences will improve both medicine and our understanding of human biology.

PROBLEMS

Which statements are true? Explain why or why not.

4–1 Human females have 23 different chromosomes, whereas human males have 24.

4–2 In a comparison between the DNAs of related organisms such as humans and mice, identifying the conserved DNA sequences facilitates the search for functionally important regions.

4–3 The four core histones are relatively small proteins with a very high proportion of positively charged amino acids; the positive charge helps the histones bind tightly to DNA, regardless of its nucleotide sequence.

4–4 Nucleosomes bind DNA so tightly that they cannot move from the positions where they are first assembled.

4–5 Gene duplication and divergence is thought to have played a critical role in the evolution of increased biological complexity.

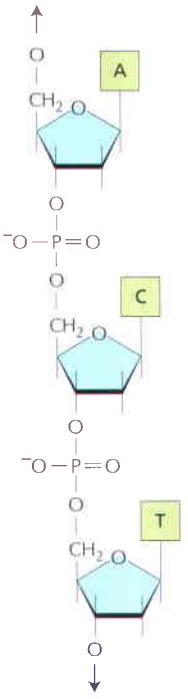


Figure Q4–1 Three nucleotides from the interior of a single strand of DNA (Problem 4–7). *Arrows* at the ends of the DNA strand indicate that the structure continues in both directions.

Discuss the following problems.

4–6 DNA isolated from the bacterial virus M13 contains 25% A, 33% T, 22% C, and 20% G. Do these results strike you as peculiar? Why or why not? How might you explain these values?

4–7 A segment of DNA from the interior of a single strand is shown in **Figure Q4–1**. What is the polarity of this DNA from top to bottom?

4–8 Human DNA contains 20% C on a molar basis. What are the mole percents of A, G, and T?

4–9 Chromosome 3 in orangutans differs from chromosome 3 in humans by two inversion events (**Figure Q4–2**). Draw the intermediate chromosome that resulted from the first inversion and explicitly indicate the segments included in each inversion.

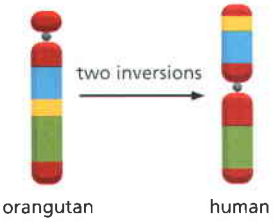


Figure Q4–2 Chromosome 3 in orangutans and humans (Problem 4–9). Differently *colored blocks* indicate segments of the chromosomes that were derived by previous fusions.

4–10 Assuming that the 30-nm chromatin fiber contains about 20 nucleosomes (200 bp/nucleosome) per 50 nm of length, calculate the degree of compaction of DNA associated with this level of chromatin structure. What fraction of the 10,000-fold condensation that occurs at mitosis does this level of DNA packing represent?

4–11 In contrast to histone acetylation, which always correlates with gene activation, histone methylation can lead to either transcriptional activation or repression. How do you suppose that the same modification—methylation—can mediate different biological outcomes?

4–12 Why is a chromosome with two centromeres (a dicentric chromosome) unstable? Would a back-up centromere not be a good thing for a chromosome, giving it two chances to form a kinetochore and attach to microtubules during mitosis? Would that not help to ensure that the chromosome did not get left behind at mitosis?

4–13 HP1 proteins, a family of proteins found in heterochromatin, are implicated in gene silencing and chromatin structure. The three proteins in humans—HP1α, HP1β, and HP1γ—share a highly conserved chromodomain, which is thought to direct chromatin localization. To determine whether these proteins could bind to the histone H3 N-terminus, you have covalently attached to separate beads various versions of the H3 N-terminal peptide—unmodified,

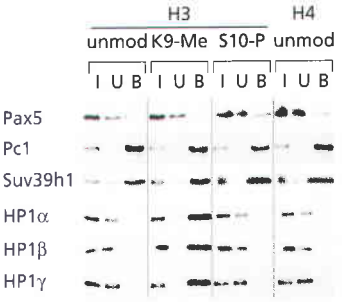


Figure Q4–3 Pull-down assays to determine binding specificity of HP1 proteins (Problem 4–13). Each protein at the left was detected by immunoblotting using a specific antibody after separation by SDS-polyacrylamide gel electrophoresis. For each histone N-terminal peptide the total input protein (I), the unbound protein (U), and the bound protein (B) are indicated. (Adapted from M. Lachner et al., *Nature* 410:116–120, 2001. With permission from Macmillan Publishers Ltd.)

Lys-9-dimethylated (K9-Me), and Ser-10-phosphorylated (S10-P)—along with an unmodified tail from histone H4. This arrangement allows you to incubate the beads with various proteins, wash away unbound proteins, and then elute bound proteins for assay by Western blotting. The results of your 'pull-down' assay for the HP1 proteins are shown in **Figure Q4–3**, along with the results from several control proteins, including Pax5, a gene regulatory protein, polycomb protein Pc1, which is known to bind to histones, and Suv39h1, a histone methyltransferase.

Based on these results, which of the proteins tested bind to the unmodified tails of histones? Do any of the HP1 proteins and control proteins selectively bind to the modified histone N-terminal peptides? What histone modification would you predict would be found in heterochromatin?

4–14 Mobile pieces of DNA—transposable elements—that insert themselves into chromosomes and accumulate during evolution make up more than 40% of the human genome. Transposable elements of four types—long interspersed elements (LINEs), short interspersed elements (SINEs), LTR retrotransposons, and DNA transposons—are inserted more or less randomly throughout the human genome. These elements are conspicuously rare at the four homeobox gene clusters, *HoxA*, *HoxB*, *HoxC*, and *HoxD*, as illustrated for *HoxD* in **Figure Q4–4**, along with an equivalent region of chromosome 22, which lacks a *Hox* cluster. Each *Hox* cluster is about 100 kb in length and contains 9 to 11 genes, whose differential expression along the anteroposterior axis of the developing embryo establishes the basic body plan for humans (and for other animals). Why do you suppose that transposable elements are so rare in the *Hox* clusters?

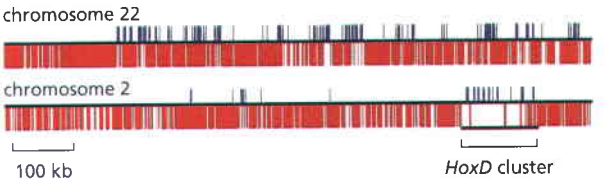


Figure Q4–4 Transposable elements and genes in 1 Mb regions of chromosomes 2 and 22 (Problem 4–14). Lines that project *upward* indicate exons of known genes. Lines that project *downward* indicate transposable elements; they are so numerous (constituting more than 40% of the human genome) that they merge into nearly a solid block outside the *Hox* clusters. (Adapted from E. Lander et al., *Nature* 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

REFERENCES

General

Hartwell L, Hood L, Goldberg ML et al (2006) Genetics: from Genes to Genomes, 3rd ed. Boston: McGraw Hill.

Olson MV (2002) The Human Genome Project: a player's perspective. *J Mol Biol* 319:931–942.

Strachan T & Read AP (2004) Human Molecular Genetics. New York: Garland Science.

Wolffe A (1999) Chromatin: Structure and Function, 3rd ed. New York: Academic Press.

The Structure and Function of DNA

Avery OT, MacLeod CM & McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of *pneumococcal* types. *J Exp Med* 79:137–158.

Meselson M & Stahl FW (1958) The replication of DNA in *E. coli*. *Proc Natl Acad Sci USA* 44:671–682.

Watson JD & Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acids. *Nature* 171:737–738.

Chromosomal DNA and Its Packaging in the Chromatin Fiber

Jin J, Cai Y, Li B et al (2005) In and out: histone variant exchange in chromatin. *Trends Biochem Sci* 30:680–687.

Kornberg RD & Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98:285–294.

Li G, Levitus M, Bustamante C & Widom J (2005) Rapid spontaneous accessibility of nucleosomal DNA. *Nature Struct Mol Biol* 12:46–53.

Lorch Y, Maier-Davis B & Kornberg RD (2006) Chromatin remodeling by nucleosome disassembly *in vitro*. *Proc Natl Acad Sci USA* 103:3090–3093.

Luger K, Mader AW, Richmond RK et al (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389:251–260.

Luger K & Richmond TJ (1998) The histone tails of the nucleosome. *Curr Opin Genet Dev* 8:140–146.

Malik H S & Henikoff S (2003) Phylogenomics of the nucleosome *Nature Struct Biol* 10:882-891.

Ried T, Schrock E, Ning Y & Wienberg J (1998) Chromosome painting: a useful art. *Hum Mol Genet* 7:1619–1626.

Robinson PJ & Rhodes R (2006) Structure of the 30 nm chromatin fibre: A key role for the linker histone. *Curr Opin Struct Biol* 16:1–8.

Saha A, Wittmeyer J & Cairns BR (2006) Chromatin remodeling: the industrial revolution of DNA around histones. *Nature Rev Mol Cell Biol* 7:437–446.

Woodcock CL (2006) Chromatin architecture. *Curr Opin Struct Biol* 16:213–220.

The Regulation of Chromatin Structure

Egger G, Liang, G, Aparicio A & Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429:457–463.

Henikoff S (1990) Position-effect variegation after 60 years. *Trends Genet* 6:422–426.

Henikoff S & Ahmad K (2005) Assembly of variant histones into chromatiin. *Annu Rev Cell Dev Biol* 21:133–153.

Gaszner M & Felsenfeld G (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Rev Genet* 7:703–713.

Hake SB & Allis CD (2006) Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis." *Proc Natl Acad Sci USA* 103:6428–6435.

Jenuwein T (2006) The epigenetic magic of histone lysine methylation. *FEBS J* 273:3121–3135.

Martin C & Zhang Y (2005) The diverse functions of histone lysine methylation. *Nature Rev Mol Cell Biol* 6:838–849.

Mellone B, Erhardt S & Karpen GH (2006) The ABCs of centromeres. *Nature Cell Biol* 8:427–429.

Peterson CL & Laniel MA (2004) Histones and histone modifications. *Curr Biol* 14:R546–R551.

Ruthenburg AJ, Allis CD & Wysocka J (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol Cell* 25:15–30.

Shahbazian MD & Grunstein M (2007) Functions of site-specific histone acetylation and deacetylation. *Annu Rev Biochem* 76:75–100.

The Global Structure of Chromosomes

Akhtar A & Gasser SM (2007) The nuclear envelope and transcriptional control. *Nature Rev Genet* 8:507–517.

Callan HG (1982) Lampbrush chromosomes. *Proc Roy Soc Lond Ser B* 21:417–448.

Chakalova L, Debrand E, Mitchel JA et al (2005) Replication and transcription: shaping the landscape of the genome. *Nature Rev Genet* 6:669–678.

Cremer T, Cremer M, Dietzel S et al (2006) Chromosome territories—a functional nuclear landscape. *Curr Opin Cell Biol* 18:307–316.

Ebert A, Lein S, Schotta G & Reuter G (2006) Histone modification and the control of heterochromatic gene silencing in *Drosophila*. *Chromosome Res* 14:377–392.

Fraser P & Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. *Nature* 447:413–417.

Handwerger KE & Gall JG (2006) Subnuclear organelles: new insights into form and function. *Trends Cell Biol* 16:19–26.

Hirano T (2006) At the heart of the chromosome: SMC proteins in action. *Nature Rev Mol Cell Biol* 7:311–322.

Lamond AI & Spector DL (2003) Nuclear speckles: a model for nuclear organelles. *Nature Rev Mol Cell Biol* 4:605–612.

Maeshima K & Laemmli UK (2003) A two-step scaffolding model for mitotic chromosome assembly. *Dev Cell* 4:467–480.

Sims JK, Houston SI, Magazinnik T & Rice JC (2006) A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. *J Biol Chem* 281:12760–12766.

Speicher MR & Carter NP (2005) The new cytogenetics: blurring the boundaries with molecular biology. *Nature Rev Genet* 6:782–792.

Zhimulev IF (1998) Morphology and structure of polytene chromosomes. *Adv Genet* 37:1–566.

How Genomes Evolve

Batzer MA & Deininger PL (2002) ALU repeats and human genomic diversity. *Nature Rev Genet* 3:370–379.

Blanchette M, Green ED, Miller W, & Haussler D (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res* 14:2412–2423.

Cheng Z, Ventura M, She X et al (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437:88–93.

Feuk L, Carson AR & Scherer S (2006) Structural variation in the human genome. *Nature Rev Genet* 7:85–97.

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

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

Koszul R, Caburet S, Dujon B & Fischer G (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* 23:234–243.

Margulies EH, NISC Comparative Sequencing Program & Green ED (2003) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harbor Symp Quant Biol* 68:255–263.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

Pollard KS, Salama SR, Lambert N et al (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172.

Sharp AJ, Cheng Z & Eichler EE (2007) Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 7:407–442.

Siepel A, Bejerano G, Pedersen JS et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.

The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.