CHAPTER **NINE** 9

How Genes and Genomes Evolve

For a given individual, the nucleotide sequence of the genome in every one of its cells is virtually the same. But compare the DNA of two individuals—even parent and child—and that is no longer the case: the genomes of individuals within a species contain slightly different information. And between members of different species, the deviations are even more extensive.

Such differences in DNA sequence are responsible for the diversity of life on Earth, from the subtle variations in hair color, eye color, and skin color that characterize members of our own species (**Figure 9–1**) to the dramatic differences in phenotype that distinguish a fish from a fungus or a robin from a rose. But if all life emerged from a common ancestor—a single-celled organism that existed some 3.5 billion years ago—where did these genetic improvisations come from? How did they arise, why were they preserved, and how do they contribute to the breathtaking biological diversity that surrounds us?

Improvements in the methods used to sequence and analyze whole genomes—from pufferfish to people—are now allowing us to address some of these questions. In Chapter 10, we describe these revolutionary technologies, which continue to transform the modern era of genomics. In this chapter, we present some of the fruits of these technological innovations. We discuss how genes and genomes have been sculpted over billions of years to give rise to the spectacular menagerie of life-forms that crowd every corner of the planet. We examine the molecular mechanisms that generate genetic diversity, and we consider how the information in present-day genomes can be deciphered to yield a historical record of the evolutionary processes that have shaped these DNA

Figure 9–1 Small differences in DNA sequence account for differences in appearance between one individual and the next. A group of schoolchildren displays a sampling of the characteristics that define the unity and diversity of our own species. (joSon/Getty Images.)

sequences. We also take a brief look at mobile genetic elements and consider how these elements, along with modern-day viruses, can carry genetic information from place to place and from organism to organism. Finally, we end the chapter by taking a closer look at the human genome to see what the DNA sequences from individuals all around the world tell us about who we are and where we come from.

GENERATING GENETIC VARIATION

There is no natural mechanism for making long stretches of entirely novel nucleotide sequences. Thus evolution is more a tinkerer than an inventor: it uses as its raw materials the DNA sequences that each organism inherits from its ancestors. In this sense, no gene or genome is ever entirely new. Instead, the astonishing diversity in form and function in the living world is all the result of variations on preexisting themes. As genetic variations pile up over millions of generations, they can produce radical change.

Several basic types of genetic change are especially crucial in evolution (Figure 9–2):

- *Mutation within a gene:* An existing gene can be modified by a mutation that changes a single nucleotide or deletes or duplicates one or more nucleotides. These mutations can alter the splicing of a gene's RNA transcript or change the stability, activity, location, or interactions of its encoded protein or RNA product.
- *Mutation within regulatory DNA sequences:* When and where a gene is expressed can be affected by a mutation in the stretches of DNA sequence that regulate the gene's activity (described in Chapter 8). For example, humans and fish have a surprisingly large number of genes in common, but changes in the regulation of those shared genes underlie many of the most dramatic differences between those species.
- *Gene duplication and divergence:* An existing gene, or even a whole genome, can be duplicated. As the cell containing this duplication, and its progeny, continue to divide, the original DNA sequence and the duplicate sequence can acquire different mutations and thereby assume new functions and patterns of expression.
- *Exon shuffling:* Two or more existing genes can be broken and rejoined to make a hybrid gene containing DNA segments that originally belonged to separate genes. In eukaryotes, such breaking and rejoining often occurs within the long intron sequences, which do not encode protein. Because these intron sequences are removed by RNA splicing, the breaking and joining do not have to be precise to produce a functional gene.
- *Transposition of mobile genetic elements:* Specialized DNA sequences that can move from one chromosomal location to another can alter the activity or regulation of a gene; they can also promote gene duplication, exon shuffling, and other genome rearrangements.
- *Horizontal gene transfer:* A piece of DNA can be passed from the genome of one cell to that of another—even to that of another species. This process, which is rare among eukaryotes but common among bacteria, differs from the usual "vertical" transfer of genetic information from parent to progeny.

Each of these forms of genetic variation has played an important part in the evolution of modern organisms. And they still play that part today, as organisms continue to evolve. In this section, we discuss these basic mechanisms of genetic change, and we consider their consequences for

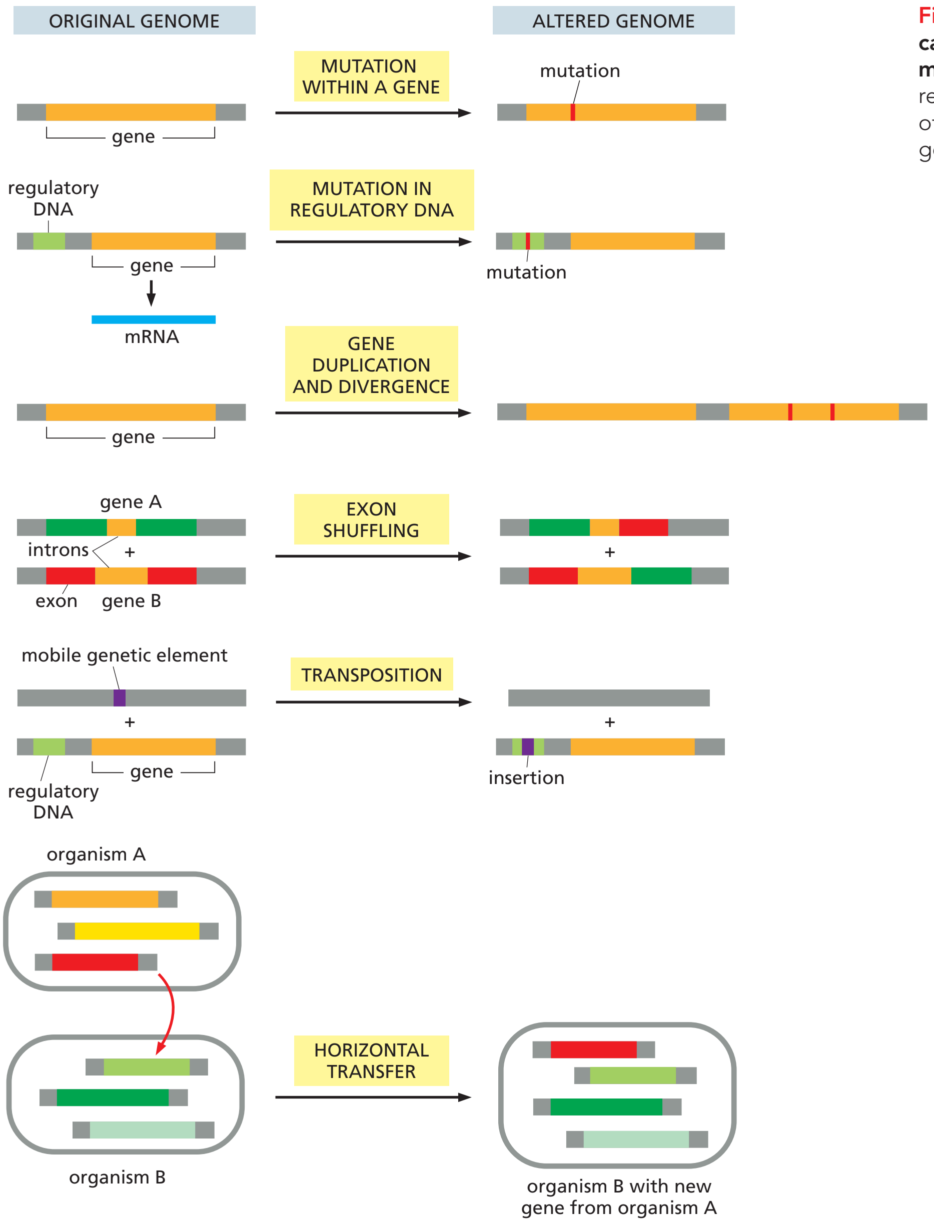


Figure 9–2 Genes and genomes can be altered by several different mechanisms. Small mutations, duplications, rearrangements, and even the infusion of fresh genetic material all contribute to genome evolution.

genome evolution. But first, we pause to consider the contribution of sex—the mechanism that many organisms use to pass genetic information on to future generations.

In Sexually Reproducing Organisms, Only Changes to the Germ Line Are Passed On to Progeny

For bacteria and unicellular organisms that reproduce asexually, the inheritance of genetic information is fairly straightforward. Each individual duplicates its genome and donates one copy to each daughter cell when the individual divides in two. The family tree of such unicellular organisms is simply a branching diagram of cell divisions that directly links each individual to its progeny and to its ancestors.

For a multicellular organism that reproduces sexually, however, the family connections are considerably more complex. Although individual cells within that organism divide, only the specialized reproductive cells—the **gametes**—carry a copy of its genome to the next generation of organisms (discussed in Chapter 19). All the other cells of the body—the **somatic cells**—are doomed to die without leaving evolutionary descendants of

QUESTION 9–1

In this chapter, we argue that genetic variability is beneficial for a species because it enhances that species' ability to adapt to changing conditions. Why, then, do you think that cells go to such great lengths to ensure the fidelity of DNA replication?

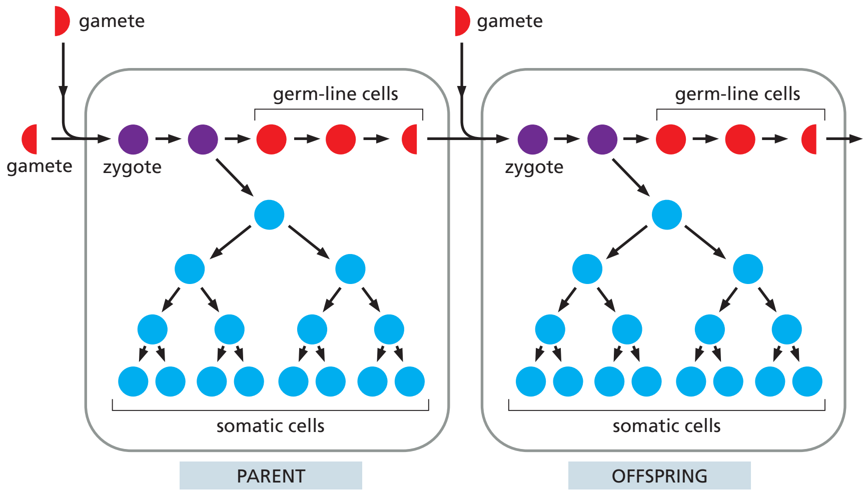


Figure 9–3 Germ-line cells and somatic cells have fundamentally different functions. In sexually reproducing organisms, genetic information is propagated into the next generation exclusively by germ-line cells (*red*). This cell lineage includes the specialized reproductive cells—the gametes (eggs and sperm, half circles)—which contain only half the number of chromosomes than do the other cells in the body (full circles). When two gametes come together during fertilization, they form a fertilized egg or zygote (*purple*), which once again contains a full set of chromosomes (discussed in Chapter 19). The zygote gives rise to both germ-line cells and to somatic cells (*blue*). Somatic cells form the body of the organism but do not contribute their DNA to the next generation.

their own (**Figure 9–3**). In a sense, somatic cells exist only to support the **germ-line** cell lineage that gives rise to the gametes.

A mutation that occurs in a somatic cell—although it might have unfortunate consequences for the individual in which it occurs (causing cancer, for example)—will not be transmitted to the organism's offspring. For a mutation to be passed on to the next generation, it must alter the germ line (**Figure 9–4**). Thus, when we track the genetic changes that accumulate during the evolution of sexually reproducing organisms, we are looking at events that took place in a germ-line cell. It is through a series of germ-line cell divisions that sexually reproducing organisms trace their descent back to their ancestors and, ultimately, back to the ancestors of us all—the first cells that existed, at the origin of life more than 3.5 billion years ago.

In addition to perpetuating a species, sex also introduces its own form of genetic change: when gametes from a male and female unite during fertilization, they generate offspring that are genetically distinct from either parent. We discuss this form of genetic diversification, which occurs only in sexually reproducing species, in detail in Chapter 19. The mechanisms for generating genetic change we discuss in this chapter, on the other hand, apply to all living things—and we return to them now.

Point Mutations Are Caused by Failures of the Normal Mechanisms for Copying and Repairing DNA

Despite the elaborate mechanisms that exist to faithfully copy and repair DNA sequences, every nucleotide pair in an organism's genome runs a

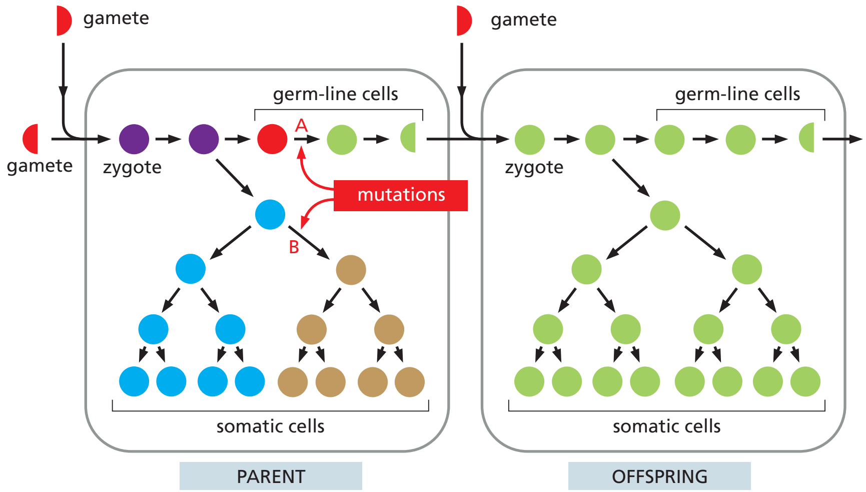


Figure 9–4 Mutations in germ-line cells and somatic cells have different consequences. A mutation that occurs in a germ-line cell (A) can be passed on to the next generation (*green*). By contrast, a mutation that arises in a somatic cell (B) affects only the progeny of that cell (*orange*) and will not be passed on to the organism's offspring. As we discuss in Chapter 20, somatic mutations are responsible for most human cancers (see pp. 720–721).

small risk of changing each time a cell divides. Changes that affect a single nucleotide pair are called **point mutations**. These typically arise from rare errors in DNA replication or repair (discussed in Chapter 6).

The point mutation rate has been determined directly in experiments with bacteria such as *E. coli.* Under laboratory conditions, *E. coli* divides about once every 20–25 minutes; in less than a day, a single *E. coli* can produce more descendants than there are humans on Earth—enough to provide a good chance for almost any conceivable point mutation to occur. A culture containing 10^9 *E. coli* cells thus harbors millions of mutant cells whose genomes differ subtly from a single ancestor cell. A few of these mutations may confer a selective advantage on individual cells: resistance to a poison, for example, or the ability to survive when deprived of a standard nutrient. By exposing the culture to a selective condition—adding an antibiotic or removing an essential nutrient, for example—one can find these needles in the haystack; that is, the cells that have undergone a specific mutation enabling them to survive in conditions where the original cells cannot (**Figure 9–5**). Such experiments have revealed that the overall point mutation frequency in *E. coli* is about 3 changes for each 10^{10} nucleotide pairs replicated. With a genome size of 4.6 million nucleotide pairs, this mutation rate means that approximately 99.99% of the time, the two daughter cells produced in a round of cell division will inherit exactly the same genome sequence of the parent *E. coli* cell; mutant cells are therefore produced only rarely.

The overall mutation rate in humans, as determined by comparing the DNA sequences of children and their parents (and estimating how many times the parental germ cells divided before producing gametes), is about one-third that of *E. coli*—which suggests that the mechanisms that

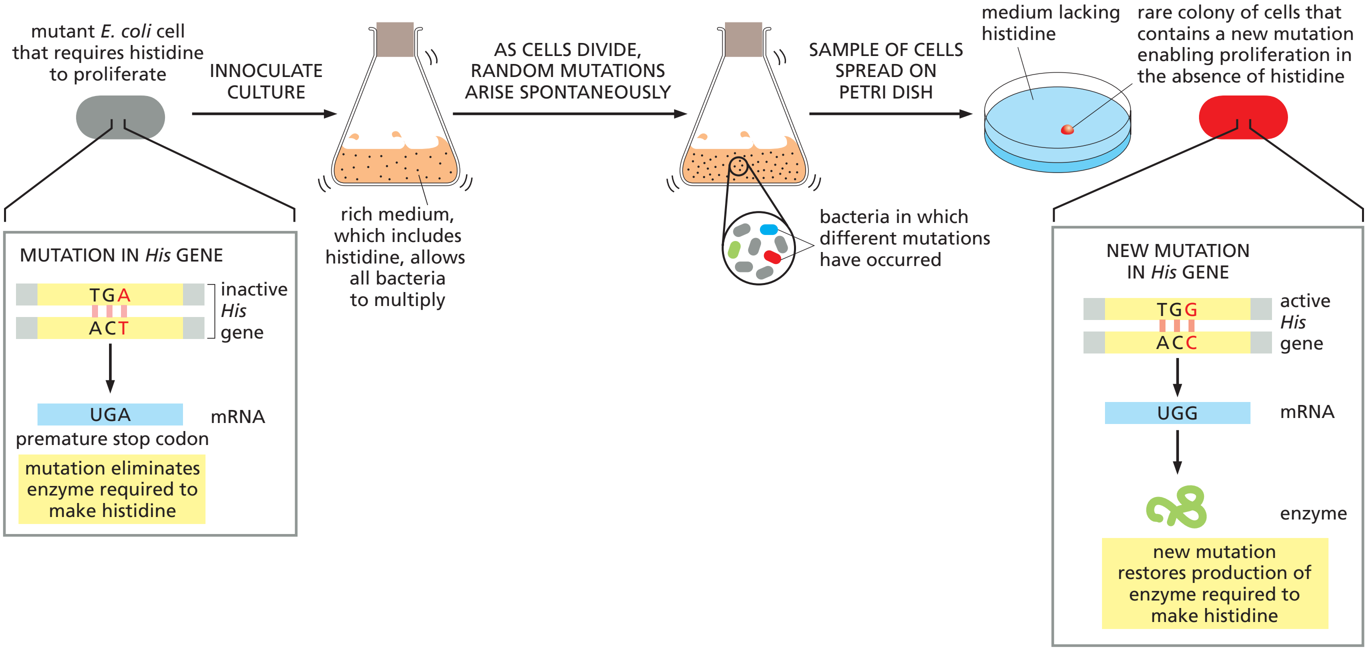

mutant E. coli cell that requires histidine to proliferate
INNOCULATE CULTURE
rich medium, which includes histidine, allows all bacteria to multiply
AS CELLS DIVIDE, RANDOM MUTATIONS ARISE SPONTANEOUSLY
bacteria in which different mutations have occurred
SAMPLE OF CELLS SPREAD ON PETRI DISH
medium lacking histidine
rare colony of cells that contains a new mutation enabling proliferation in the absence of histidine
MUTATION IN His GENE
TGA
ACT
inactive His gene
UGA
mRNA
premature stop codon
mutation eliminates enzyme required to make histidine
NEW MUTATION IN His GENE
TGG
ACC
active His gene
UGG
mRNA
enzyme
new mutation restores production of enzyme required to make histidine

Figure 9–5 Mutation rates can be measured in the laboratory. In this experiment, an *E. coli* strain that carries a deleterious point mutation in the *His* gene—which is needed to manufacture the amino acid histidine—is used. The mutation has converted a G-C nucleotide pair to an A-T, resulting in a premature stop signal in the mRNA produced from the mutant gene (*left* box). As long as histidine is supplied in the growth medium, this strain can grow and divide normally. If a large number of mutant cells (say 10^{10}) is spread on an agar plate that lacks histidine, the great majority will die. The rare survivors will contain a new mutation in which the A-T is changed back to a G-C. This "reversion" corrects the original defect and allows the bacterium to make the enzyme it needs to survive in the absence of histidine. Such mutations happen by chance and only rarely, but the ability to work with very large numbers of *E. coli* cells makes it possible to detect this change and to accurately measure its frequency.

evolved to maintain genome integrity operate with an efficiency that does not greatly differ between even distantly related species.

Point mutations can destroy a gene's activity or—very rarely—improve it (as shown in Figure 9–5). More often, however, they do neither of these things. At many sites in the genome, a point mutation has absolutely no effect on the organism's appearance, viability, or ability to reproduce. Such *neutral mutations* often fall in regions of the gene where the DNA sequence is unimportant, including most of an intron's sequence. In cases where they occur within an exon, neutral mutations can change the third position of a codon such that the amino acid it specifies is unchanged—or is so similar that the protein's function is unaffected.

Mutations Can Also Change the Regulation of a Gene

Point mutations that lie outside the coding sequences of genes can sometimes affect regulatory DNA sequences—elements that control the timing, location, and level of gene expression. Such mutations in regulatory DNA sequences can have a profound effect on the protein's production and thereby on the organism. For example, a small number of people are resistant to malaria because of a point mutation that affects the expression of a cell-surface receptor to which the malaria parasite *Plasmodium vivax* binds. The mutation prevents the receptor from being produced in red blood cells, rendering the individuals who carry this mutation immune to malarial infection.

Point mutations in regulatory DNA sequences also have a role in our ability to digest lactose, the main sugar in milk. Our earliest ancestors were lactose intolerant, because the enzyme that breaks down lactose—called lactase—was made only during infancy. Adults, who were no longer exposed to breast milk, did not need the enzyme. When humans began to get milk from domesticated cattle some 10,000 years ago, variant genes—the product of random mutation—enabled those who carried the variation to continue to express lactase as adults, and thus take advantage of nutrition provided by cow's milk. We now know that people who retain the ability to digest milk as adults contain a point mutation in the regulatory DNA sequence of the lactase gene, allowing it to be efficiently transcribed throughout life. In a sense, these milk-drinking adults are "mutants" with respect to their ancestors. It is remarkable how quickly this adaptation spread through the human population, especially in societies that depended heavily on milk for nutrition (**Figure 9–6**).

These evolutionary changes in the regulatory DNA sequence of the lactase gene occurred relatively recently (10,000 years ago), well after humans became a distinct species. However, much more ancient changes in regulatory DNA sequences have occurred in other genes, and some of these are thought to underlie many of the profound differences among species (**Figure 9–7**).

DNA Duplications Give Rise to Families of Related Genes

Point mutations can influence the activity of an existing gene, but how do new genes with new functions come into being? Gene duplication is perhaps the most important mechanism for generating new genes from old ones. Once a gene has been duplicated, each of the two copies is free to accumulate mutations—as long as whatever activities the original gene may have had are not lost. Over time, as mutations continue to accumulate in the descendants of the original cell in which gene duplication occurred, some of these genetic changes allow one of the gene copies to perform a different function.

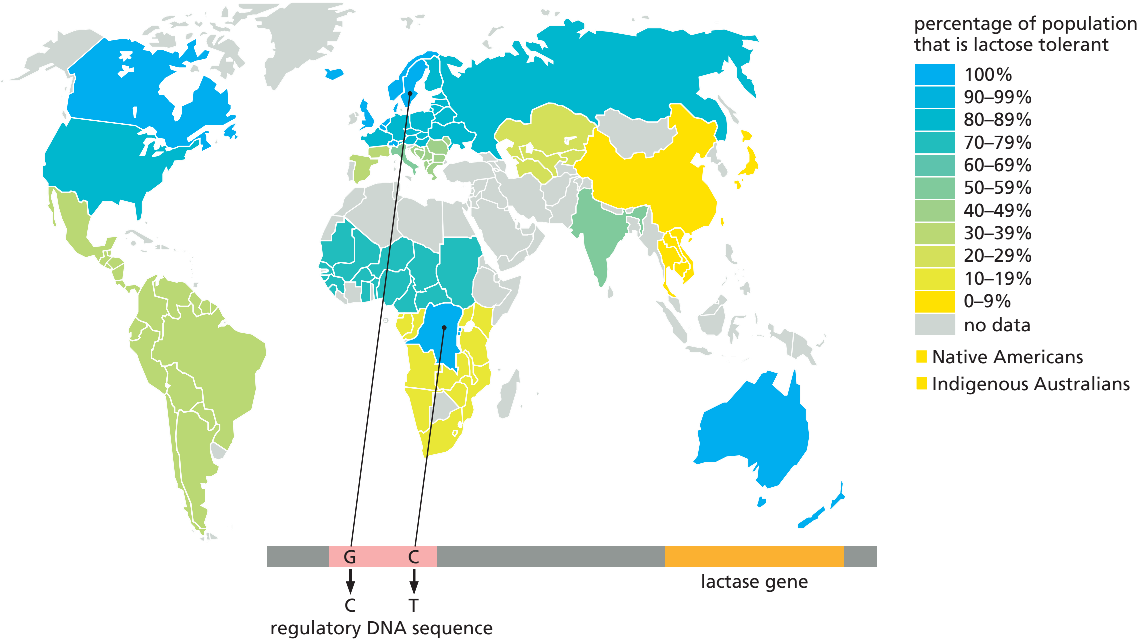


Figure 9–6 The widespread ability of adult humans to digest milk followed the domestication of cattle. Approximately 10,000 years ago, humans in northern Europe and central Africa began to raise cattle. The subsequent availability of cow's milk—particularly during periods of starvation—gave a selective advantage to those humans able to digest lactose as adults. Two independent point mutations that allow the expression of lactase in adults arose in human populations—one in northern Europe and another in central Africa. These mutations have since spread through different regions of the world.

By repeated rounds of this process of **gene duplication and divergence** over many millions of years, one gene can give rise to a whole family of genes, each with a specialized function, within a single genome. Analysis of genome sequences reveals many examples of such **gene families**: in *Bacillus subtilis*, for example, nearly half of the genes have one or more obvious relatives elsewhere in the genome. And in vertebrates, the globin family of genes, which encode oxygen-carrying proteins, clearly arose from a single primordial gene, as we see shortly. But how does gene duplication occur in the first place?

Many gene duplications are believed to be generated by *homologous recombination*. As discussed in Chapter 6, homologous recombination provides an important mechanism for mending a broken double helix; it allows an intact chromosome to be used as a template to repair a damaged sequence on its homolog. But as we discuss in Chapter 19, homologous recombination can also catalyze *crossovers* in which two

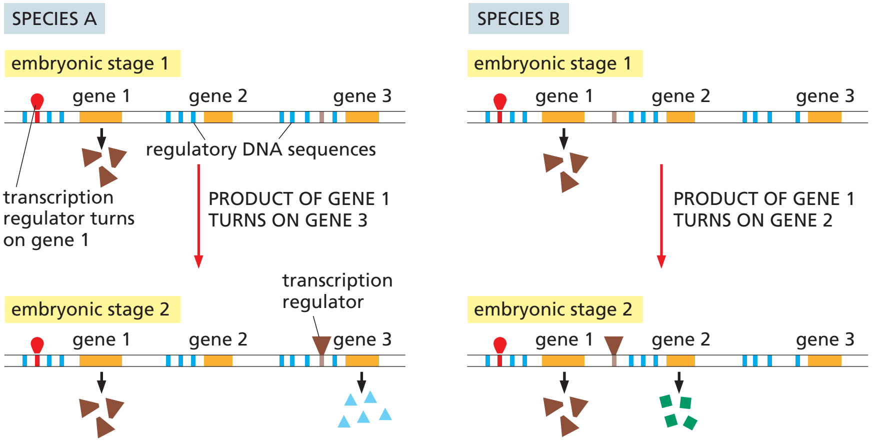


Figure 9–7 Changes in regulatory DNA sequences can have dramatic consequences for the development of an organism. In this hypothetical example, the genomes of two closely related species A and B contain the same three genes (1, 2, and 3) and encode the same two transcription regulators (*red* oval, *brown* triangle). However, the regulatory DNA sequences controlling expression of genes 2 and 3 are different in the two species. Although both express gene 1 during embryonic stage 1, the differences in their regulatory DNA sequences cause them to express different genes in stage 2. In principle, a collection of such regulatory changes can have profound effects on an organism's developmental program—and, ultimately, on the appearance of the adult.

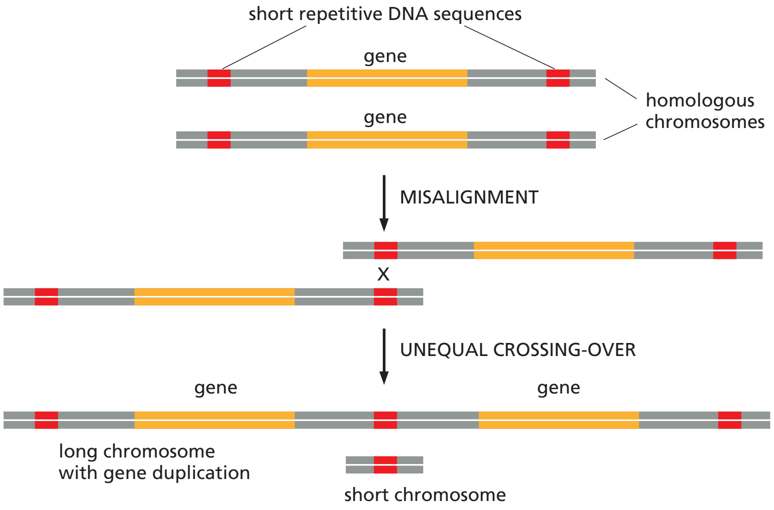


Figure 9–8 Gene duplication can be caused by crossovers between short, repeated DNA sequences in adjacent homologous chromosomes. The two chromosomes shown here undergo homologous recombination at short repeated sequences (*red*), that bracket a gene (*orange*). For simplicity, only one gene is shown on each homolog. The repeated sequences can be remnants of mobile genetic elements, which are present in many copies in the human genome, as we discuss shortly. When crossing-over occurs unequally, as shown, one chromosome will get two copies of the gene, while the other will get none. The type of homologous recombination that produces gene duplications is called *unequal crossing-over* because the resulting products are unequal in size. If this process occurs in the germ line, some progeny will inherit the long chromosome, while others will inherit the short one.

chromosomes are broken and joined up to produce hybrid chromosomes. Crossovers take place only between regions of chromosomes that have nearly identical DNA sequences; for this reason, they usually occur between homologous chromosomes and generate hybrid chromosomes in which the order of genes is exactly the same as on the original chromosomes. This process occurs extensively during meiosis, as we see in Chapter 19.

On rare occasions, however, a crossover can occur between a pair of short DNA sequences—identical or very similar—that fall on either side of a gene. If these short sequences are not aligned properly during recombination, a lopsided exchange of genetic information can occur. Such unequal crossovers can generate one chromosome that has an extra copy of the gene and another with no copy (**Figure 9–8**); this shorter chromosome will eventually be lost.

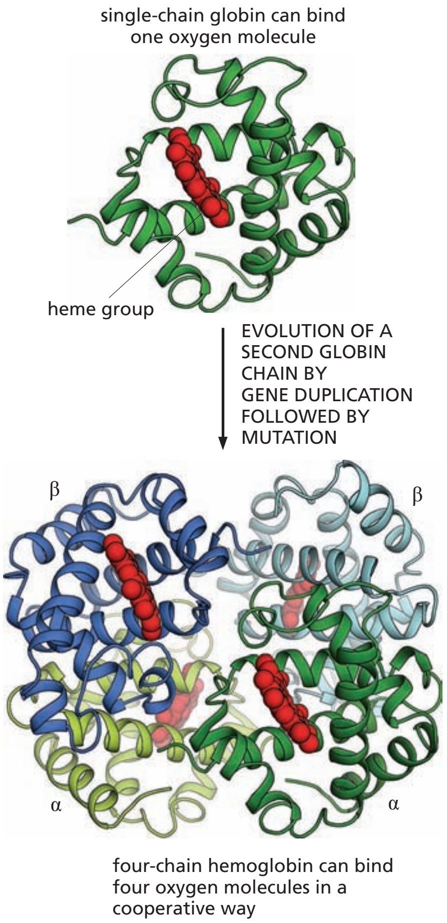
Once a gene has been duplicated in this way, extra copies of the gene can be added by the same mechanism. As a result, entire sets of closely related genes, arranged in series, are commonly found in genomes.

Duplication and Divergence Produced the Globin Gene Family

The evolutionary history of the globin gene family provides a striking example of how gene duplication and divergence has generated new proteins. The unmistakable similarities in amino acid sequence and structure among present-day globin proteins indicate that all the globin genes must derive from a single ancestral gene.

The simplest globin protein has a single polypeptide chain of about 150 amino acids, and is found in many marine worms, insects, and primitive fish. Like our hemoglobin, this protein transports oxygen molecules throughout the animal's body. The oxygen-carrying protein in the blood of adult mammals and most other vertebrates, however, is more complex; it is composed of four globin chains of two distinct types—α globin and β globin (**Figure 9–9**). The four oxygen-binding sites in the $\alpha_2\beta_2$ molecule interact, allowing an allosteric change in the molecule as it binds and releases oxygen. This structural shift enables the four-chain hemoglobin molecule to efficiently take up and release four oxygen molecules in an all-or-none fashion, a feat not possible for the single-chain version. Such efficiency is particularly important for large multicellular animals, which cannot rely on the simple diffusion of oxygen through the body to oxygenate their tissues adequately.

Figure 9–9 An ancestral globin gene encoding a single-chain globin molecule gave rise to the pair of genes that produce four-chain hemoglobin proteins of modern humans and other mammals. The mammalian hemoglobin molecule is a complex of two α-globin (*green*) and two β-globin (*blue*) chains. Each chain contains a tightly bound heme group (*red*) that is responsible for binding oxygen.


single-chain globin can bind one oxygen molecule
heme group
EVOLUTION OF A SECOND GLOBIN CHAIN BY GENE DUPLICATION FOLLOWED BY MUTATION
β
β
α
α
four-chain hemoglobin can bind four oxygen molecules in a cooperative way

The α- and β-globin genes are the result of a gene duplication that occurred early in vertebrate evolution. Genome analyses suggest that one of our distant ancestors had a single globin gene. But about 500 million years ago, a gene duplication followed by an accumulation of different mutations in each gene copy is thought to have given rise to two slightly different globin genes, one encoding α globin, the other encoding β globin. Still later, as the different mammals began diverging from their common ancestor, the β-globin gene underwent its own duplication and divergence to give rise to a second β-like globin gene that is expressed specifically in the fetus (**Figure 9–10**). The resulting fetal hemoglobin molecule has a higher affinity for oxygen compared with adult hemoglobin, a property that helps transfer oxygen from mother to fetus.

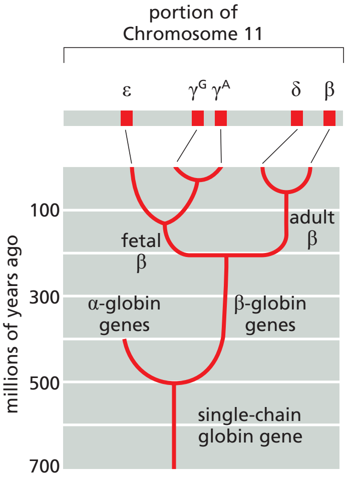
Subsequent rounds of duplication and divergence in both the α- and β-globin genes gave rise to additional members of these families. Each of these duplicated genes has been modified by point mutations that affect the properties of the final hemoglobin molecule, and by changes in regulatory DNA sequences that determine when—and how strongly—each gene is expressed. As a result, each globin differs slightly in its ability to bind and release oxygen and in the stage of development during which it is expressed.

In addition to these specialized globin genes, there are several duplicated DNA sequences in the α- and β-globin gene clusters that are not functional genes. They are similar in DNA sequence to the functional globin genes, but they have been disabled by the accumulation of many inactivating mutations. The existence of such *pseudogenes* makes it clear that not every DNA duplication leads to a new functional gene. In fact, most gene duplication events are unsuccessful in that one copy is gradually inactivated by mutation. Although we have focused here on the evolution of the globin genes, similar rounds of gene duplication and divergence have clearly taken place in many other gene families present in the human genome.

Figure 9–10 Repeated rounds of duplication and mutation generated the globin gene family in humans. About 500 million years ago, an ancestral globin gene duplicated and gave rise to both the β-globin gene family (including the five genes shown) and the α-globin gene family. In most vertebrates, a molecule of hemoglobin (see Figure 9–9) is formed from two chains of α globin and two chains of β globin—which can be any one of the five subtypes of the β family listed here.

The evolutionary scheme shown was worked out by comparing globin genes from many different organisms. The nucleotide sequences of the γ^G and γ^A genes—which produce the β-globin-like chains that form fetal hemoglobin—are much more similar to each other than either of them is to the adult β gene. The δ-globin gene encodes a minor form of adult β-globin. In humans, the β-globin genes are located in a cluster on Chromosome 11.

A subsequent chromosome breakage event, which occurred about 300 million years ago, is believed to have separated the α- and β-globin genes; the α-globin genes now reside on human Chromosome 16 (not shown).



Whole-Genome Duplications Have Shaped the Evolutionary History of Many Species

Almost every gene in the genomes of vertebrates exists in multiple versions, suggesting that, rather than single genes being duplicated in a piecemeal fashion, the whole vertebrate genome was long ago duplicated in one fell swoop. Early in vertebrate evolution, it appears that the entire genome actually underwent duplication twice in succession, giving rise to four copies of every gene. In some groups of vertebrates, such as the salmon and carp families (including the zebrafish; see Figure 1–38), there may have been yet another duplication, creating an eightfold multiplicity of genes.

The precise history of whole-genome duplications in vertebrate evolution is difficult to chart because many other changes, including the loss of genes, have occurred since these ancient evolutionary events. In some organisms, however, full genome duplications are especially obvious, as they have occurred relatively recently, evolutionarily speaking. The frog genus *Xenopus*, for example, includes closely related species that differ dramatically in DNA content: some are diploid—containing two complete sets of chromosomes—whereas others are tetraploid or octoploid. Such large-scale duplications can happen if cell division fails to occur following a round of genome replication in the germ line of a particular individual. Once an accidental doubling of the genome occurs in a germ-line cell, it will be faithfully passed on to germ-line progeny cells in that individual and, ultimately, to any offspring these cells might produce.

Whole-genome duplications are also common in plants, including many of those that we eat. These genome duplications generally make the plant easier to cultivate and its fruit more palatable. In some cases, genome duplication renders the plant sterile so that it cannot produce seeds; such is the case with seedless grapes. Apples, leeks, and potatoes are all tetraploid, whereas strawberries and sugarcane are octoploid (**Figure 9–11**).

Novel Genes Can Be Created by Exon Shuffling

As we discussed in Chapter 4, many proteins are composed of smaller functional *domains*. In eukaryotes, each of these protein domains is usually encoded by a separate exon, which is surrounded by long stretches of noncoding introns (see Figures 7–18 and 7–19). This organization of eukaryotic genes can facilitate the evolution of new proteins by allowing exons from one gene to be added to another—a process called **exon shuffling**.

Such duplication and movement of exons is promoted by the same type of recombination that gives rise to gene duplications (see Figure 9–8). In this case, recombination occurs within the introns that surround the exons.

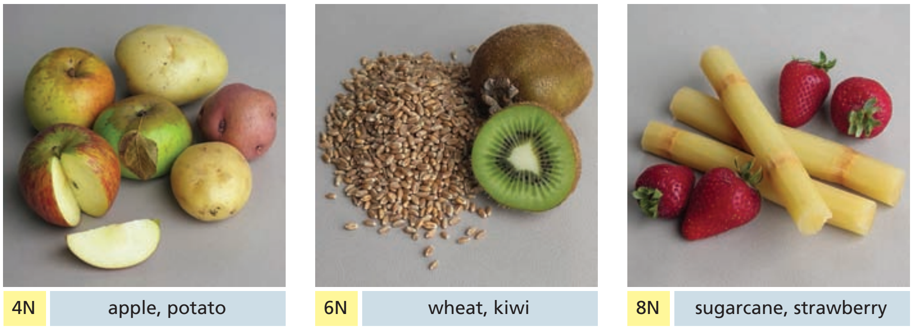


Figure 9–11 Many crop plants have undergone whole-genome duplication. Many of these duplications, which arose spontaneously, were propagated by plant breeders because they rendered the plants easier to cultivate or made their fruits larger, more flavorful, or devoid of indigestible seeds. N indicates the ploidy of each type of plant: for example, wheat and kiwi are hexaploid—possessing six complete sets of chromosomes (6N).

If the introns in question are from two different genes, this recombination can generate a hybrid gene that includes complete exons from both. The results of such exon shuffling are seen in many present-day proteins, which contain a patchwork of many different protein domains (**Figure 9–12**).

It has been proposed that nearly all the proteins encoded by the human genome (approximately 19,000) arose from the duplication and shuffling of a few thousand distinct exons, each encoding a protein domain of approximately 30–50 amino acids. This remarkable idea suggests that the great diversity of protein structures is generated from a fairly small universal "parts list," pieced together in different combinations.

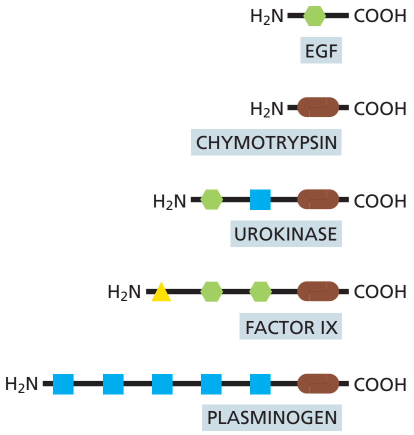


Figure 9–12 Exon shuffling during evolution can generate proteins with new combinations of protein domains. Each type of colored symbol represents a different protein domain. These different domains were joined together by exon shuffling during evolution to create the modern-day human proteins shown here. EGF, epidermal growth factor.

The Evolution of Genomes Has Been Profoundly Influenced by Mobile Genetic Elements

Mobile genetic elements—DNA sequences that can move from one chromosomal location to another—are an important source of genomic change and have profoundly affected the structure of modern genomes. These parasitic DNA sequences can colonize a genome and then spread within it. In the process, they often disrupt the function or alter the regulation of existing genes; sometimes they even create novel genes through fusions between mobile sequences and segments of existing genes.

The insertion of a mobile genetic element into the coding sequence of a gene or into its regulatory DNA sequence can cause the "spontaneous" mutations that are observed in many of today's organisms. Mobile genetic elements can severely disrupt a gene's activity if they land directly within its coding sequence. Such an insertion mutation destroys the gene's capacity to encode a useful protein—as is the case for a number of mutations that cause hemophilia in humans, for example.

The activity of mobile genetic elements can also change the way existing genes are regulated. An insertion of an element into a regulatory DNA sequence, for instance, will often have a striking effect on where or when genes are expressed (**Figure 9–13**). Many mobile genetic elements carry DNA sequences that are recognized by specific transcription regulators; if these elements insert themselves near a gene, that gene can be brought under the control of these transcription regulators, thereby changing the gene's expression pattern. Thus, mobile genetic elements can be a major source of developmental changes: they have been particularly important in the evolution of domesticated plants. For example, the development of modern corn from a wild, grassy plant called teosinte required only a small number of genetic alterations. One of these changes was the insertion of a mobile genetic element upstream of a gene active in seed development, which transformed the small, hard seeds of teosinte into the plentiful soft kernels of modern corn (**Figure 9–14**).

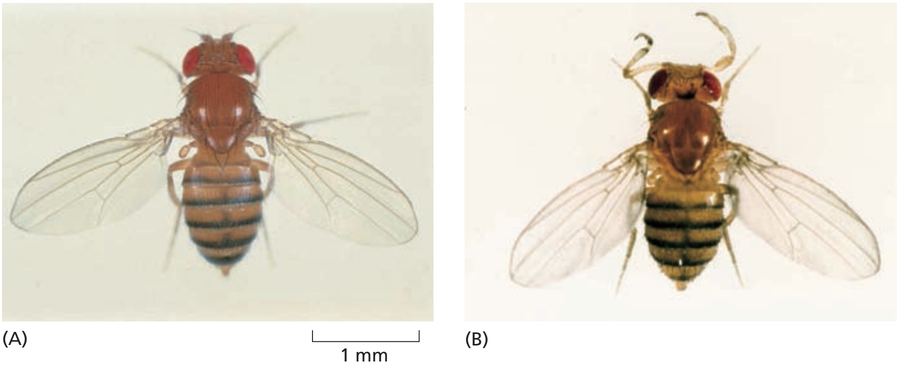


Figure 9–13 Mutation due to a mobile genetic element can induce dramatic alterations in the body plan of an organism. (A) A normal fruit fly (*Drosophila melanogaster*). (B) A mutant fly in which the antennae have been replaced by legs because of a mutation in a regulatory DNA sequence that causes genes for leg formation to be activated in the positions normally reserved for antennae. Although this particular change is not advantageous to the fly, it illustrates how the movement of a transposable element can produce a major change in the appearance of an organism. (A, Edward B. Lewis. Courtesy of the Archives, California Institute of Technology; B, courtesy of Matthew Scott.)

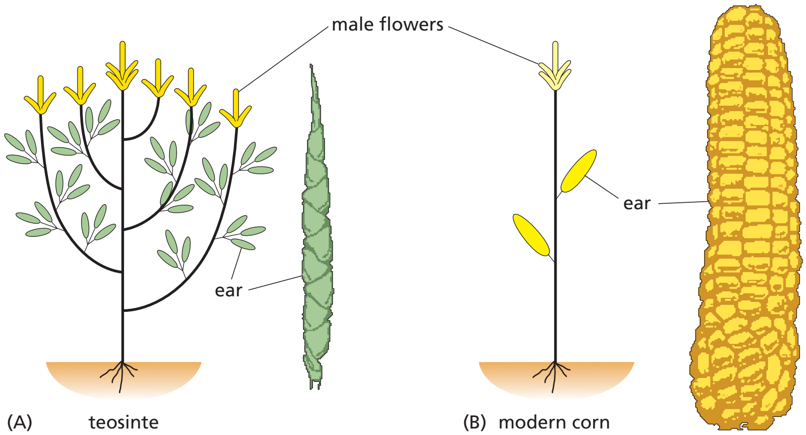


Figure 9–14 The insertion of a mobile genetic element helped produce modern corn. Today's corn plants were originally bred from a wild plant called teosinte (A). This wild ancestor produced numerous ears that contained small, hard seeds. (B) Modern corn, by contrast, produces fewer cobs—but they contain numerous plump, sweet kernels. The insertion of a mobile genetic element near a gene involved in seed development helped drive the change. Here, the two plants are drawn to the same scale; for simplicity, the leaves are not shown.

Finally, mobile genetic elements provide opportunities for genome rearrangements by serving as targets of homologous recombination (see Figure 9–8). For example, the duplications that gave rise to the β-globin gene cluster are thought to have occurred by crossovers between the abundant mobile genetic elements sprinkled throughout the human genome. Later in the chapter, we describe these elements in more detail and discuss the mechanisms that have allowed them to establish a stronghold within our genome.

QUESTION 9–2

Why do you suppose that horizontal gene transfer is more prevalent in single-celled organisms than in multicellular organisms?

Genes Can Be Exchanged Between Organisms by Horizontal Gene Transfer

So far we have considered genetic changes that take place within the genome of an individual organism. However, genes and other portions of genomes can also be exchanged between individuals of different species. This mechanism of **horizontal gene transfer** is rare among eukaryotes but common among bacteria, which can exchange DNA by the process of conjugation (**Figure 9–15** and **Movie 9.1**).

E. coli, for example, has acquired about one-fifth of its genome from other bacterial species within the past 100 million years. And such genetic exchanges are currently responsible for the rise of new and potentially dangerous strains of drug-resistant bacteria. Genes that confer resistance to antibiotics are readily transferred from species to species, providing the recipient bacterium with an enormous selective advantage in evading the antimicrobial compounds that constitute modern medicine's frontline attack against bacterial infection. As a result, many antibiotics are no longer effective against the common bacterial infections for which they were originally used; as an example, most strains of *Neisseria gonorrhoeae*, the bacterium that causes gonorrhea, are now resistant to penicillin, which is therefore no longer the primary drug used to treat this disease.

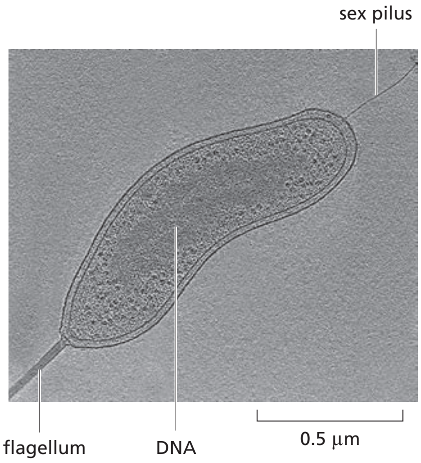


Figure 9–15 Bacterial cells can exchange DNA through conjugation. Conjugation begins when a donor cell captures a recipient cell using a fine appendage called a sex pilus. Following capture, DNA moves from the donor cell, through the pilus, into the recipient cell. In this cryoelectron micrograph, the sex pilus is clearly distinguished from the flagellum. Conjugation is one of several ways in which bacteria carry out horizontal gene transfer. (From C.M. Oikonomou and G.J. Jensen, *Nat. Rev. Microbiol.* 14:205–220, 2016. With permission from Macmillan Publishers Ltd.)

RECONSTRUCTING LIFE'S FAMILY TREE

The nucleotide sequences of present-day genomes provide a record of those genetic changes that have survived the test of time. By comparing the genomes of a variety of living organisms, we can thus begin to decipher our evolutionary history and see how our ancestors veered off in adventurous new directions that led us to where we are today.

The most astonishing revelation of such genome comparisons has been that **homologous genes**—those that are similar in nucleotide sequence because of their common ancestry—can be recognized across vast evolutionary distances. Unmistakable homologs of many human genes are easy to detect in organisms such as worms, fruit flies, yeasts, and even bacteria. Although the lineage that led to the evolution of vertebrates is thought to have diverged from the one that led to nematode worms and insects more than 600 million years ago, when we compare the genomes of the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* with that of *Homo sapiens*, we find that about 50% of the genes in each of these species have clear homologs in one or both of the other two species. In other words, clearly recognizable versions of at least half of all human genes must have already been present in the common ancestor of worms, flies, and humans.

By tracing such relationships among genes, we can begin to define the evolutionary relationships among different species, placing each bacterium, animal, plant, or fungus in a single vast family tree of life. In this section, we discuss how these relationships are determined and what they tell us about our genetic heritage.

Genetic Changes That Provide a Selective Advantage Are Likely to Be Preserved

Evolution is commonly thought of as progressive, but at the molecular level the process is random. Consider the fate of a point mutation that occurs in a germ-line cell. On rare occasions, the mutation might cause a change for the better. But most often it will either have no consequence or cause serious damage. Mutations of the first type will tend to be perpetuated, because the organism that inherits them will have an increased likelihood of reproducing itself. Mutations that are deleterious will usually be lost. And mutations that are *selectively neutral* may or may not persist, depending on factors such as the size of the population, or whether the individual carrying the neutral mutation also harbors a favorable mutation located nearby. Through endless repetition of such cycles of mutation and natural selection—a molecular form of trial and error—organisms gradually evolve. Their genomes change and they develop new ways to exploit the environment—to outcompete others and to reproduce successfully.

Clearly, some parts of the genome can accumulate mutations more easily than others in the course of evolution. A segment of DNA that does not code for protein or RNA and has no significant regulatory role is free to change at a rate limited only by the frequency of random mutation. In contrast, deleterious alterations in a gene that codes for an essential protein or RNA molecule cannot be accommodated so easily: when mutations occur, the faulty organism will almost always be eliminated or fail to reproduce. Genes of this latter sort are therefore *highly conserved*; that is, the products they encode, whether RNA or protein, are very similar from organism to organism. Throughout the 3.5 billion years or more of evolutionary history, the most highly conserved genes remain perfectly recognizable in all living species. They encode crucial proteins such as DNA and RNA polymerases, and they are the ones we turn to

QUESTION 9–3

Highly conserved genes such as those for ribosomal RNA are present as clearly recognizable relatives in all organisms on Earth; thus, they have evolved very slowly over time. Were such genes "born" perfect?

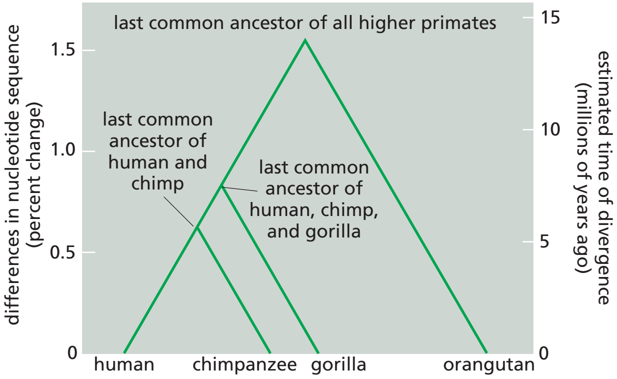


Figure 9–16 Phylogenetic trees display the relationships among modern life-forms. In this family tree of higher primates, humans fall closer to chimpanzees than to gorillas or orangutans, as there are fewer differences between human and chimp DNA sequences than there are between those of humans and gorillas, or of humans and orangutans. As indicated, the genome sequences of each of these four species are estimated to differ from the sequence of the last common ancestor of higher primates by about 1.5%. Because changes occur independently in each lineage after two species diverge from a common ancestor, the genetic differences between any two species will be twice as much as the amount of change between each of the species and the common ancestor. For example, although humans and orangutans each differ from their common ancestor by about 1.5% in terms of nucleotide sequence, they typically differ from one another by slightly more than 3%; human and chimp genomes differ by about 1.2%. This phylogenetic tree is based solely on nucleotide sequences of species alive today, as indicated on the *left* side of the graph; the estimated dates of divergence, shown on the *right* side of the graph, are derived from analysis of the fossil record. (Modified from F.C. Chen and W.H. Li, *Am. J. Hum. Genet.* 68:444–456, 2001.)

when we wish to trace family relationships among the most distantly related organisms in the tree of life.

Closely Related Organisms Have Genomes That Are Similar in Organization as Well as Sequence

For species that are closely related, it is often most informative to focus on selectively neutral mutations. Because they accumulate steadily at a rate that is unconstrained by selection pressures, these mutations provide a metric for gauging how much modern species have diverged from their common ancestor. Such sequence comparisons allow the construction of a **phylogenetic tree**, a diagram that depicts the evolutionary relationships among a group of organisms. As an example, **Figure 9–16** presents a phylogenetic tree that lays out the relationships among higher primates.

As indicated in this figure, chimpanzees are our closest living relative among the higher primates. Not only do chimpanzees seem to have essentially the same set of genes as we do, but their genes are arranged in nearly the same way on their chromosomes. The only substantial exception is human Chromosome 2, which arose from a fusion of two chromosomes that remain separate in the chimpanzee, gorilla, and orangutan. Humans and chimpanzees are so closely related that it is possible to use DNA sequence comparisons to reconstruct the amino acid sequences of proteins that must have been present in the now-extinct, common ancestor of the two species (**Figure 9–17**).

Even the rearrangement of genomes by crossing over, which we described earlier, has produced only minor differences between the human and chimp genomes. For example, both the chimp and human genomes contain a million copies of a type of mobile genetic element called an *Alu* sequence. More than 99% of these elements are in corresponding positions in both genomes, indicating that most of the *Alu* sequences in our genome were in place before humans and chimpanzees diverged.

Functionally Important Genome Regions Show Up as Islands of Conserved DNA Sequence

As we delve back further into our evolutionary history and compare our genomes with those of more distant relatives, the picture begins to change. The lineages of humans and mice, for example, diverged about 75 million years ago. These genomes are about the same size, contain practically the same genes, and are both riddled with mobile genetic elements. However, the mobile genetic elements found in mouse and human DNA, although similar in nucleotide sequence, are distributed

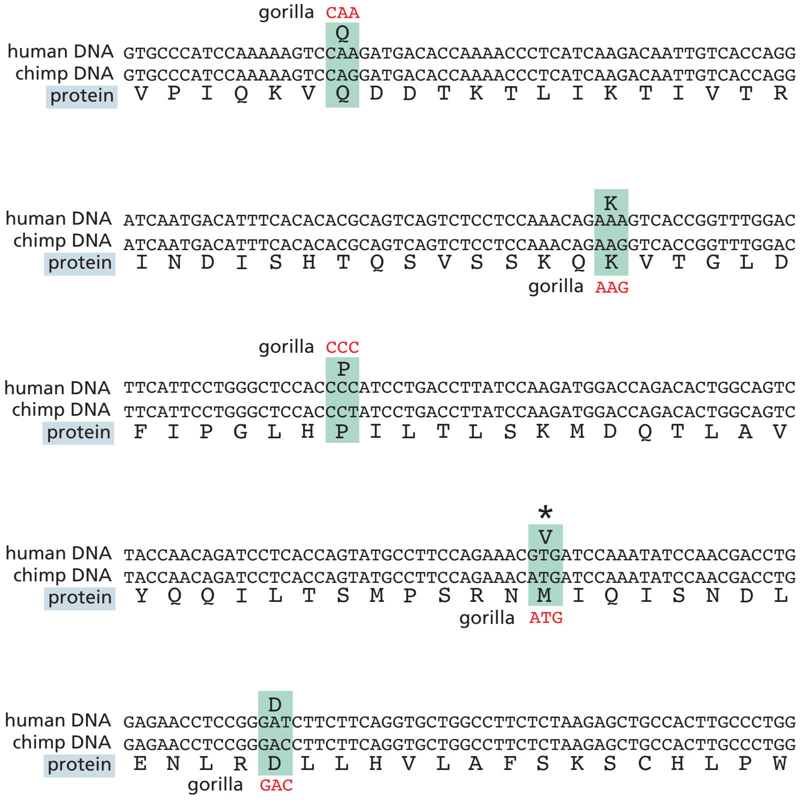

gorilla CAA
Q
human DNA GTGCCCATCCAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGG
chimp DNA GTGCCCATCCAAAAAGTCCAGGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGG
protein V P I Q K V Q D D T K T L I K T I V T R
K
human DNA ATCAATGACATTTCACACACGCAGTCAGTCTCCTCCAAACAGAAAGTCACCGGTTTGGAC
chimp DNA ATCAATGACATTTCACACACGCAGTCAGTCTCCTCCAAACAGAAGGTCACCGGTTTGGAC
protein I N D I S H T Q S V S S K Q K V T G L D
gorilla AAG
gorilla CCC
P
human DNA TTCATTCCTGGGCTCCACCCCATCCTGACCTTATCCAAGATGGACCAGACACTGGCAGTC
chimp DNA TTCATTCCTGGGCTCCACCCTATCCTGACCTTATCCAAGATGGACCAGACACTGGCAGTC
protein F I P G L H P I L T L S K M D Q T L A V
*
V
human DNA TACCAACAGATCCTCACCAGTATGCCTTCCAGAAACGTGATCCAAATATCCAACGACCTG
chimp DNA TACCAACAGATCCTCACCAGTATGCCTTCCAGAAACATGATCCAAATATCCAACGACCTG
protein Y Q Q I L T S M P S R N M I Q I S N D L
gorilla ATG
D
human DNA GAGAACCTCCGGGATCTTCTTCAGGTGCTGGCCTTCTCTAAGAGCTGCCACTTGCCCTGG
chimp DNA GAGAACCTCCGGGACCTTCTTCAGGTGCTGGCCTTCTCTAAGAGCTGCCACTTGCCCTGG
protein E N L R D L L H V L A F S K S C H L P W
gorilla GAC

Figure 9–17 Ancestral gene sequences can be reconstructed by comparing closely related present-day species. Shown here, in five contiguous segments of DNA, are the nucleotide sequences that encode the mature leptin protein from humans and chimpanzees. Leptin is a hormone that regulates food intake and energy utilization. As indicated by the codons boxed in *green*, only five nucleotides differ between the chimp and human sequences. Only one of these changes (marked with an asterisk) results in a change in the amino acid sequence.

The nucleotide sequence of the last common ancestor was probably the same as the human and chimp sequences where they agree; in the few places where they disagree, the gorilla sequence (*red*) can be used as a “tiebreaker,” as the gorilla sequence is evolutionarily more distant than those of chimp and human (see Figure 9–16). Thus, the amino acid indicated by the asterisk was a methionine in the common ancestor of humans and chimpanzees and is changed to a valine in the human lineage. For convenience, only the first 300 nucleotides of the coding sequences for the mature leptin protein are shown; the last 141 nucleotides of that sequence are identical between humans and chimpanzees.

differently, as they have had more time to proliferate and move around the two genomes after these species diverged (**Figure 9–18**).

In addition to the movement of mobile genetic elements, the large-scale organization of the human and mouse genomes has been scrambled by many episodes of chromosome breakage and recombination over the past 75 million years: it is estimated that about 180 such “break-and-join” events have dramatically altered chromosome organization. For example, in humans most centromeres lie near the middle of the chromosome, whereas those of mouse are located at the chromosome ends.

Regardless of this significant degree of genetic shuffling, one can nevertheless still recognize many blocks of **conserved synteny**, regions in which corresponding genes are strung together in the same order in both species. These genes were neighbors in the ancestral species and, despite all the chromosomal upheavals, they remain neighbors in the two present-day species. More than 90% of the mouse and human genomes can be partitioned into such corresponding regions of conserved synteny. Within these regions, we can align the DNA of mouse with that of humans so that we can compare the nucleotide sequences in detail. Such genome-wide sequence comparisons reveal that, in the roughly

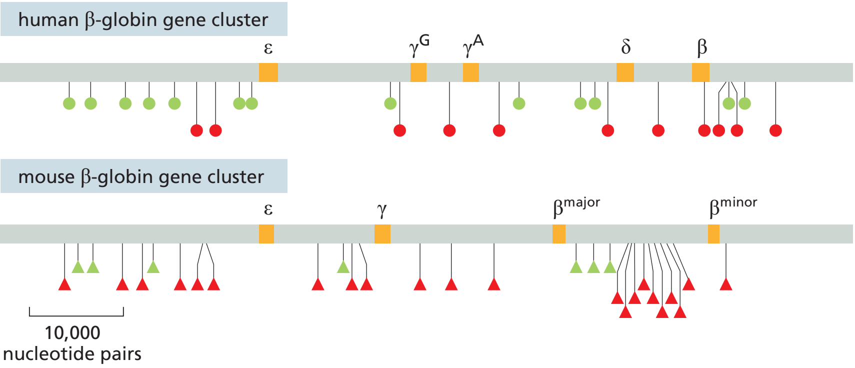


Figure 9–18 Differences in the positions of mobile genetic elements in the human and mouse genomes reflect the long evolutionary time separating the two species. This stretch of human Chromosome 11 (seen also in Figure 9–10) contains five functional β-globin-like genes (*orange*); the comparable region from the mouse genome contains only four. The positions of two types of mobile genetic element—*Alu* sequences (*green*) and *L1* sequences (*red*)—are shown in each genome. Although the mobile genetic elements in human (*circles*) and mouse (*triangles*) are not identical, they are closely related. The absence of these elements within the globin genes can be attributed to *purifying selection*, which would have eliminated any insertion that compromised gene function. (The mobile genetic element that falls inside the human β-globin gene (*far right*) is located within an intron, not in a coding sequence.) (Courtesy of Ross Hardison and Webb Miller.)

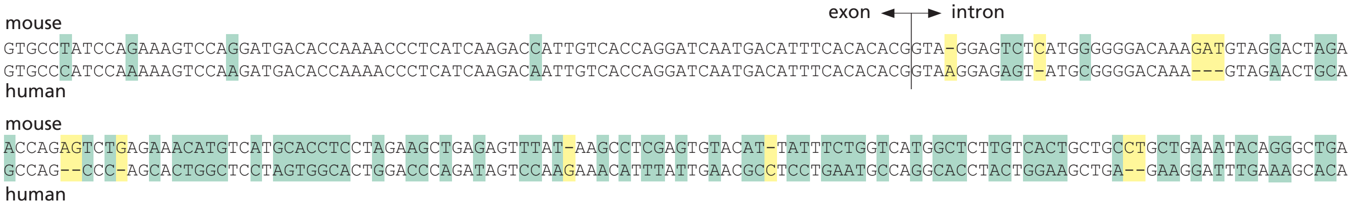

exon
intron
mouse
GTGCCTATCCAGAAAGTCCAGGATGACACCAAAACCCTCATCAAGACCATTGTCACCAGGATCAATGACATTTCACACACGGTA-GGAGTCTCATGGGGGGACAAAGATGTAGGACTAGA
GTGCCCATCCAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGGATCAATGACATTTCACACACGGTAAGGAGAGT-ATGCGGGGACAAA---GTAGAACTGCA
human
mouse
ACCAGAGTCTGAGAAACATGTCATGCACCTCCTAGAAGCTGAGAGTTTAT-AAGCCTCGAGTGTACAT-TATTTCTGGTCATGGCTCTTGTCACTGCTGCCTGCTGAAATACAGGGCTGA
GCCAG--CCC-AGCACTGGCTCCTAGTGGCACTGGACCCAGATAGTCCAAGAAACATTTATTGAACGCCTCCTGAATGCCAGGCACCTACTGGAAGCTGA--GAAGGATTTGAAAGCACA
human

Figure 9–19 Accumulated mutations have resulted in considerable divergence in the nucleotide sequences of the human and the mouse genomes. Shown here in two contiguous segments of DNA are portions of the human and mouse leptin gene sequences. Positions where the sequences differ by a single nucleotide substitution are boxed in *green*, and positions where they 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 the adjacent intron sequence.

75 million years since humans and mice diverged from their common ancestor, about 50% of the nucleotides have changed. However, these differences are not dispersed evenly across the genome. By observing where the human and mouse sequences have remained nearly the same, one can thus see very clearly the regions where genetic changes are not tolerated (**Figure 9–19**). These sequences have been conserved by **purifying selection**—that is, by the elimination of individuals carrying mutations that interfere with important functions.

The power of *comparative genomics* can be further increased by stacking our genome up against the genomes of additional animals, including the rat, chicken, and dog. Such comparisons take advantage of the results of the "natural experiment" that has lasted for hundreds of millions of years, and they highlight some of the most important regions of these genomes. These comparisons reveal that roughly 4.5% of the human genome consists of DNA sequences that are highly conserved in many other mammals (**Figure 9–20**). Surprisingly, only about one-third

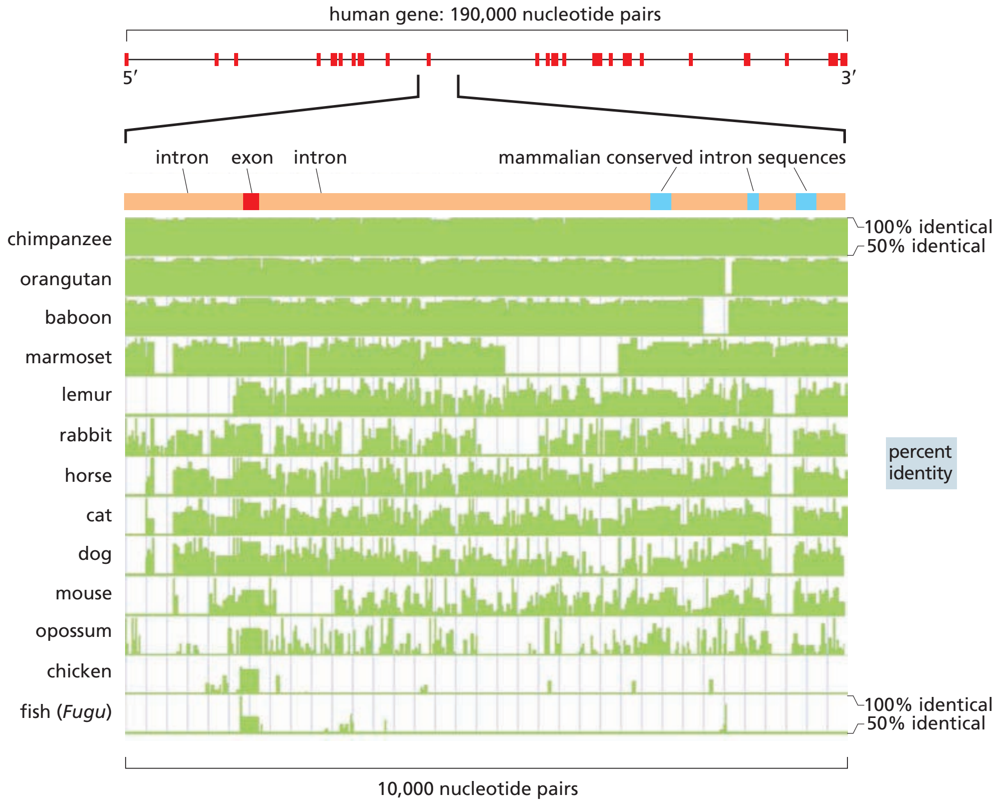


Figure 9–20 Comparison of nucleotide sequences from many different vertebrates reveals regions of high conservation. The nucleotide sequence examined in this diagram is a small segment of the human gene for a plasma membrane transporter protein. The upper part of the diagram shows the location of the exons (*red*) in both the complete gene (*top*) and in the expanded region of the gene. Three blocks of intron sequence that are conserved in mammals are shown in *blue*. In the lower part of the figure, the DNA sequence of the expanded segment of 10,000 nucleotide pairs is aligned with the corresponding sequences of different vertebrates; the percent identity with the human sequences for successive stretches of 100 nucleotide pairs is plotted in *green*, with only identities above 50% shown. Note that the sequence of the exon is highly conserved in all the species, including chicken and fish, but the three intron sequences that are conserved in mammals are not conserved in chickens or fish. The functions of most conserved intron sequences in the human genome (including these three) are not known. (Courtesy of Eric D. Green.)

of these sequences code for proteins. Some of the conserved noncoding sequences correspond to regulatory DNA, whereas others are transcribed to produce RNA molecules that are not translated into protein but serve a variety of functions (see Chapter 8). The functions of many of these conserved noncoding sequences, however, remain unknown. The unexpected discovery of these mysterious conserved DNA sequences suggests that we understand much less about the cell biology of mammals than we had previously imagined. With the plummeting cost and accelerating speed of whole-genome sequencing, we can expect many more surprises that will lead to an increased understanding in the years ahead.

Genome Comparisons Show That Vertebrate Genomes Gain and Lose DNA Rapidly

Going back even further in evolution, we can compare our genome with those of more distantly related vertebrates. The lineages of fish and mammals diverged about 400 million years ago. This stretch of time is long enough for random sequence changes and differing selection pressures to have obliterated almost every trace of similarity in nucleotide sequence—except where purifying selection has operated to prevent change. Regions of the genome conserved between humans and fishes thus stand out even more strikingly than those conserved between different mammals. In fishes, one can still recognize most of the same genes as in humans and even many of the same regulatory DNA sequences. On the other hand, the extent of duplication of any given gene is often different, resulting in different numbers of members of gene families in the two species.

Even more striking is the finding that although all vertebrate genomes contain roughly the same number of genes, their overall size varies considerably. Whereas human, dog, and mouse are all in the same size range (around 3×10^9 nucleotide pairs), the chicken genome is only one-third this size. An extreme example of genome compression is the pufferfish *Fugu rubripes* (**Figure 9–21**). The fish's tiny genome is about one-eighth the size of mammalian genomes, largely because of the small size of its intergenic regions, which are missing nearly all of the repetitive DNA that makes up a large portion of most mammalian genomes. The *Fugu* introns are also short in comparison to human introns. Nonetheless, the positions of most *Fugu* introns are perfectly conserved when compared with their positions in the genomes of mammals. Clearly, the intron structure of most vertebrate genes was already in place in the common ancestor of fish and mammals.

What factors could be responsible for the size differences among modern vertebrate genomes? Detailed comparisons of many genomes have led to the unexpected finding that small blocks of sequence are being lost from and added to genomes at a surprisingly rapid rate. It seems likely, for example, that the *Fugu* genome is so tiny because it lost DNA sequences faster than it gained them. Over long periods, this imbalance apparently cleared out those DNA sequences whose loss could be tolerated. This "cleansing" process has been enormously helpful to biologists: by "trimming the fat" from the *Fugu* genome, evolution has provided a conveniently slimmed-down version of a vertebrate genome in which the only DNA sequences that remain are those that are very likely to have important functions.

Figure 9–21 The pufferfish, *Fugu rubripes*, has a remarkably compact genome. At 400 million nucleotide pairs, the *Fugu* genome is only one-quarter the size of the zebrafish genome, even though the two species have nearly the same genes. (From a woodcut by Hiroshige, courtesy of Arts and Designs of Japan.)

Sequence Conservation Allows Us to Trace Even the Most Distant Evolutionary Relationships

As we go back further still to the genomes of our even more distant relatives—beyond apes, mice, fish, flies, worms, plants, and yeasts, all

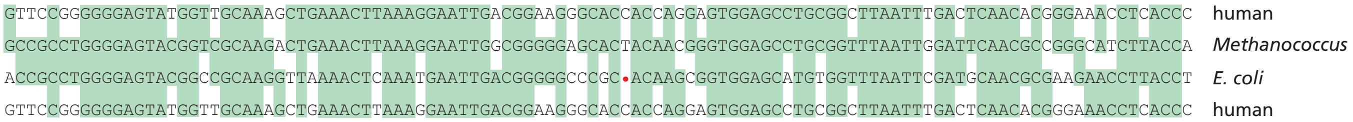

GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCC human
GCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTTAAAGGAATTGGCGGGGGAGCACTACAACGGGTGGAGCCTGCGGTTTAATTGGATTCAACGCCGGGCATCTTACCA Methanococcus
ACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGC•ACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCT E. coli
GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCC human

Figure 9–22 Some genetic information has been conserved since the beginnings of life. A part of the gene for the small ribosomal subunit rRNA (see Figure 7–35) is shown. Corresponding segments of nucleotide sequence from this gene in three distantly related species (*Methanococcus jannaschii* and *Escherichia coli*, both prokaryotes, and *Homo sapiens*, a eukaryote) are aligned in parallel. Sites where the nucleotides are identical between any two species are indicated by *green* shading; the human sequence is repeated at the bottom of the alignment so that all three two-way comparisons can be seen. The *red* dot halfway along the *E. coli* sequence denotes a site where a nucleotide has been either deleted from the bacterial lineage in the course of evolution or inserted in the other two lineages. Note that the three sequences have all diverged from one another to a roughly similar extent, while still retaining unmistakable similarities.

the way to bacteria—we find fewer and fewer resemblances to our own genome. Yet even across this enormous evolutionary divide, purifying selection has maintained a few hundred fundamentally important genes. By comparing the sequences of these genes in different organisms and seeing how far they have diverged, we can attempt to construct a phylogenetic tree that goes all the way back to the ultimate ancestors—the cells at the very origins of life, from which we all derive.

To construct such a tree, biologists have focused on one particular gene that is conserved in all living species: the gene that codes for the ribosomal RNA (rRNA) of the small ribosomal subunit (shown schematically in Figure 7–35). Because the process of translation is fundamental to all living cells, this component of the ribosome has been highly conserved since early in the history of life on Earth (**Figure 9–22**).

By applying the same principles used to construct the primate family tree (see Figure 9–16), the small-subunit rRNA nucleotide sequences have been used to create a single, all-encompassing tree of life. Although many aspects of this phylogenetic tree were anticipated by classical taxonomy (which is based on the outward appearance of organisms), there were also many surprises. Perhaps the most important was the realization that some of the organisms that were traditionally classed as "bacteria" are as widely divergent in their evolutionary origins as is any prokaryote from any eukaryote. As discussed in Chapter 1, it is now apparent that the prokaryotes comprise two distinct groups—the *bacteria* and the *archaea*—that diverged early in the history of life on Earth. The living world therefore has three major divisions or *domains*: bacteria, archaea, and eukaryotes (**Figure 9–23**).

Although we humans have been classifying the visible world since antiquity, we now realize that most of life's genetic diversity lies in the world of microscopic organisms. These microbes have tended to go unnoticed, unless they cause disease or rot the timbers of our houses. Yet they make up most of the total mass of living matter on our planet. Many of these

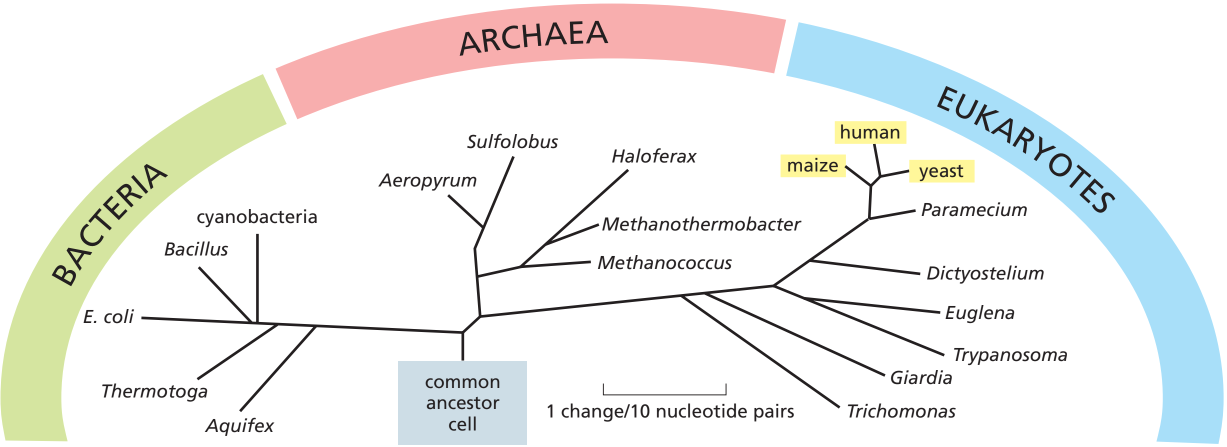


Figure 9–23 The tree of life has three major divisions. Each branch on the tree is labeled with the name of a representative member of that group, and the length of each branch corresponds to the degree of difference in the DNA sequences that encode their small-subunit rRNAs (see Figure 9–22). Note that all the organisms we can see with the unaided eye—animals, plants, and some fungi (highlighted in *yellow*)—represent only a small subset of the diversity of life.

organisms cannot be grown under laboratory conditions. Thus it is only through the analysis of DNA sequences, obtained from around the globe, that we are beginning to obtain a more detailed understanding of all life on Earth—knowledge that is less distorted by our biased perspective as large animals living on dry land.

MOBILE GENETIC ELEMENTS AND VIRUSES

The tree of life depicted in Figure 9–23 includes representatives from life's most distant branches, from the cyanobacteria that release oxygen into Earth's atmosphere to the animals, like us, that use that oxygen to boost their metabolism. What the diagram does not encompass, however, are the parasitic genetic elements that operate on the outskirts of life. Although these elements are built from the same nucleic acids contained in all life-forms and can multiply and move from place to place, they do not cross the threshold of actually being alive. Yet because of their prevalence and their penchant for propagating themselves, these diminutive genetic parasites have major implications for the evolution of species and for human health.

We briefly discussed these **mobile genetic elements**, earlier in the chapter, and here we consider them in greater detail. Known informally as jumping genes, mobile genetic elements are found in virtually all cells. Their DNA sequences make up almost half of the human genome. Although they can insert themselves into virtually any region of the genome, most mobile genetic elements lack the ability to leave the cell in which they reside. This is not the case for their relatives, the *viruses*. Not much more than strings of genes wrapped in a protective coat, viruses can escape from one cell and infect another.

In this section, we discuss mobile genetic elements and viruses. We review their structure and outline how they operate—and we consider the effects they have on gene expression, genome evolution, and the transmission of disease.

Mobile Genetic Elements Encode the Components They Need for Movement

Mobile genetic elements, also called **transposons,** are typically classified according to the mechanism by which they move or *transpose*. In bacteria, the most common mobile genetic elements are the *DNA-only transposons*. The name is derived from the fact that the element moves from one place to another as a piece of DNA, as opposed to being converted into an RNA intermediate—which is the case for another type of mobile element we discuss shortly. Bacteria contain many different DNA-only transposons. Some move to the target site using a simple cut-and-paste mechanism, whereby the element is simply excised from the genome and inserted into a different site. Other DNA-only transposons replicate before transposing; in this case, the new copy of the transposon inserts into a second chromosomal site, while the original copy remains intact at its previous location (**Figure 9–24**).

Each mobile genetic element typically encodes a specialized enzyme, called a *transposase*, that mediates its movement. These enzymes recognize and act on unique DNA sequences that are present on the mobile genetic elements that code for the transposase. Many mobile genetic elements also harbor additional genes: some mobile genetic elements, for example, carry antibiotic-resistance genes, which have contributed greatly to the widespread dissemination of antibiotic resistance in bacterial populations (**Figure 9–25**).

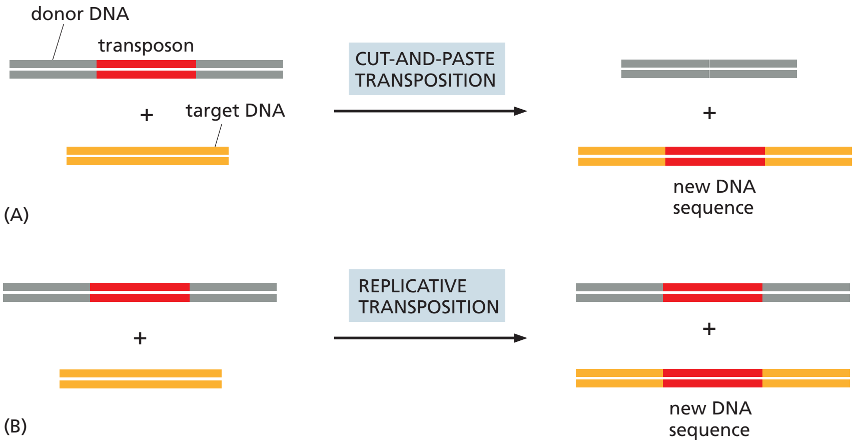


Figure 9–24 The most common mobile genetic elements in bacteria, DNA-only transposons, move by two types of mechanism. (A) In cut-and-paste transposition, the element is cut out of the donor DNA and inserted into the target DNA, leaving behind a broken donor DNA molecule, which is subsequently repaired. (B) In replicative transposition, the mobile genetic element is copied by DNA replication. The donor molecule remains unchanged, and the target molecule receives a copy of the mobile genetic element. In general, a particular type of transposon moves by only one of these mechanisms. However, the two mechanisms have many enzymatic similarities, and a few transposons can move by either mechanism. The donor and target DNAs can be part of the same DNA molecule or reside on different DNA molecules.

In addition to relocating themselves, mobile genetic elements occasionally rearrange the DNA sequences of the genome in which they are embedded. For example, if two mobile genetic elements that are recognized by the same transposase integrate into neighboring regions of the same chromosome, the DNA between them can be accidentally excised and inserted into a different gene or chromosome (**Figure 9–26**). In eukaryotic genomes, such accidental transposition provides a pathway for generating novel genes, both by altering gene expression and by duplicating existing genes.

QUESTION 9–4

Many transposons move within a genome by replicative mechanisms (such as those shown in Figure 9–24B). They therefore increase in copy number each time they transpose. Although individual transposition events are rare, many transposons are found in multiple copies in genomes. What do you suppose keeps the transposons from completely overrunning their hosts' genomes?

The Human Genome Contains Two Major Families of Transposable Sequences

The sequencing of human genomes has revealed many surprises, as we describe in detail in the next section. But one of the most stunning was the finding that a large part of our DNA is not entirely our own. Nearly half of the human genome is made up of mobile genetic elements, which number in the millions. Some of these elements have moved from place to place within the human genome using the cut-and-paste mechanism discussed earlier (see Figure 9–24A). However, most have moved not as DNA, but via an RNA intermediate. These **retrotransposons** appear to be unique to eukaryotes.

One abundant human retrotransposon, the ***L1* element** (sometimes referred to as *LINE-1, a long interspersed nuclear element*), is transcribed into RNA by a host cell's RNA polymerase. A double-stranded DNA copy of this RNA is then made using an enzyme called **reverse transcriptase**, an unusual DNA polymerase that can use RNA as a template. The reverse transcriptase is encoded by the *L1* element itself. The DNA copy of the element is then free to reintegrate into another site in the genome (**Figure 9–27**).

L1 elements constitute about 15% of the human genome. Although most copies have been immobilized by the accumulation of deleterious

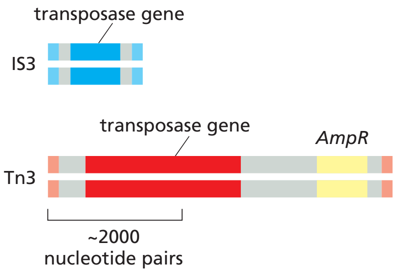


Figure 9–25 Transposons contain the components they need for transposition. Shown here are two types of bacterial DNA-only transposons. Each carries a gene that encodes a transposase (*blue* and *red*)—the enzyme that catalyzes the element's movement—as well as DNA sequences (*red*) that are recognized by each transposase. Some transposons carry additional genes (*yellow*) that encode enzymes that inactivate antibiotics such as ampicillin (*AmpR*). The spread of these transposons is a serious problem in medicine, as it has allowed many disease-causing bacteria to become resistant to antibiotics developed during the twentieth century.

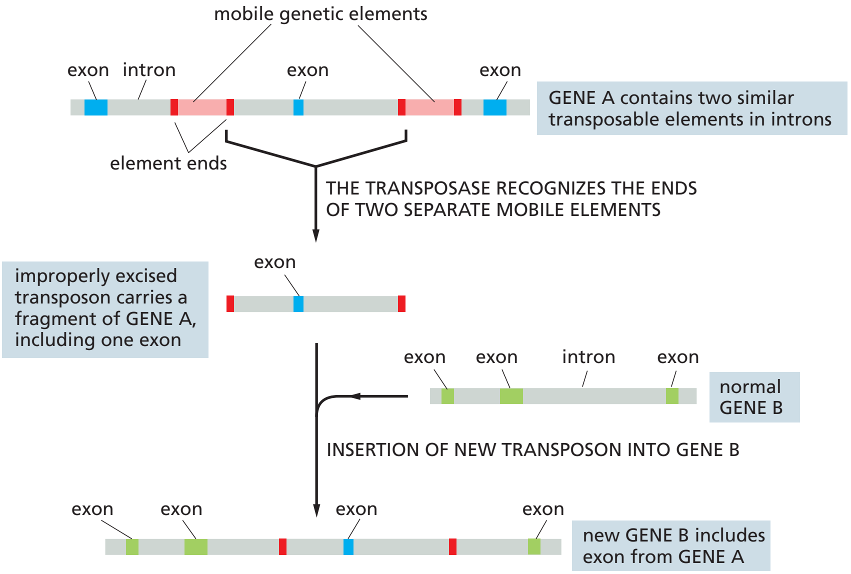

mobile genetic elements
exon
intron
exon
exon
GENE A contains two similar transposable elements in introns
element ends
THE TRANSPOSASE RECOGNIZES THE ENDS OF TWO SEPARATE MOBILE ELEMENTS
improperly excised transposon carries a fragment of GENE A, including one exon
exon
exon
exon
intron
exon
normal GENE B
INSERTION OF NEW TRANSPOSON INTO GENE B
exon
exon
exon
exon
new GENE B includes exon from GENE A

Figure 9–26 Mobile genetic elements can move exons from one gene to another. When two mobile genetic elements of the same type (*red*) happen to insert near each other in a chromosome, the transposition mechanism occasionally recognizes the ends of two different elements (instead of the two ends of the same element). As a result, the chromosomal DNA that lies between the mobile genetic elements gets excised and moved to a new site. Such inadvertent transposition of chromosomal DNA can either generate novel genes, as shown, or alter gene regulation (not shown).

mutations, a few still retain the ability to transpose. Their movement can sometimes precipitate disease: for example, movement in the germline of an *L1* element into the gene that encodes Factor VIII—a protein essential for proper blood clotting—caused hemophilia in a child with no family history of the disease.

Another type of retrotransposon, the ***Alu* sequence**, is present in about 1 million copies, making up about 10% of our genome. *Alu* elements do not encode their own reverse transcriptase and thus depend on enzymes already present in the cell to help them move.

Comparisons of the sequence and locations of the *L1* and *Alu* elements in different mammals suggest that these sequences have proliferated in primates relatively recently in evolutionary history (see Figure 9–18). Given that the placement of mobile genetic elements can have profound effects on gene expression, it is humbling to contemplate how many of our uniquely human qualities we might owe to these prolific genetic parasites.

Viruses Can Move Between Cells and Organisms

Viruses are also mobile, but unlike the transposons we have discussed so far, they can actually escape from cells and move to other cells and organisms. Viruses were first categorized as disease-causing agents that, by virtue of their tiny size, passed through ultrafine filters that can hold back even the smallest bacterial cell. We now know that viruses are essentially small genomes enclosed by a protective protein coat, and that they must enter a cell and coopt its molecular machinery to express their genes, make their proteins, and reproduce. Although the first viruses that were discovered attack mammalian cells, it is now recognized that many types of viruses exist, and virtually all organisms—including plants, animals, and bacteria—can serve as viral hosts.

Viral reproduction is often lethal to the host cells; in many cases, the infected cell breaks open (lyses), releasing progeny viruses, which can then infect neighboring cells. Many of the symptoms of viral infections reflect this lytic effect of the virus. The cold sores formed by herpes simplex virus and the blisters caused by the chickenpox virus, for example, reflect the localized killing of human skin cells.

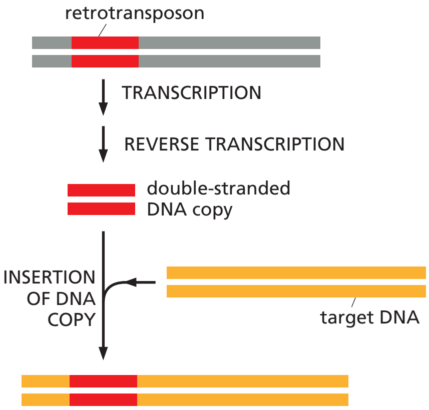


Figure 9–27 Retrotransposons move via an RNA intermediate. These transposable elements are first transcribed into an RNA intermediate (not shown). Next, a double-stranded DNA copy of this RNA is synthesized by the enzyme reverse transcriptase. This DNA copy is then inserted into the target location, which can be on either the same or a different DNA molecule. The donor retrotransposon remains at its original location, so each time it transposes, it duplicates itself. These mobile genetic elements are called retrotransposons because at one stage in their transposition their genetic information flows backward, from RNA to DNA.

TABLE 9–1 VIRUSES THAT CAUSE HUMAN DISEASE

Virus	Genome Type	Disease
Herpes simplex virus	double-stranded DNA	recurrent cold sores
Epstein–Barr virus (EBV)	double-stranded DNA	infectious mononucleosis
Varicella-zoster virus	double-stranded DNA	chickenpox and shingles
Smallpox virus	double-stranded DNA	smallpox
Hepatitis B virus	part single-, part double-stranded DNA	serum hepatitis
Human immunodeficiency virus (HIV)	single-stranded RNA	acquired immune deficiency syndrome (AIDS)
Influenza virus type A	single-stranded RNA	respiratory disease (flu)
Poliovirus	single-stranded RNA	poliomyelitis
Rhinovirus	single-stranded RNA	common cold
Hepatitis A virus	single-stranded RNA	infectious hepatitis
Hepatitis C virus	single-stranded RNA	non-A, non-B type hepatitis
Yellow fever virus	single-stranded RNA	yellow fever
Rabies virus	single-stranded RNA	rabies encephalitis
Mumps virus	single-stranded RNA	mumps
Measles virus	single-stranded RNA	measles

Most viruses that cause human disease have genomes made of either double-stranded DNA or single-stranded RNA (Table 9–1). However, viral genomes composed of single-stranded DNA and of double-stranded RNA are also known. The simplest viruses found in nature have a small genome, composed of as few as three genes, enclosed by a protein coat built from many copies of a single polypeptide chain. More complex viruses have larger genomes of up to several hundred genes, surrounded by an elaborate shell composed of many different proteins (Figure 9–28). The amount of genetic material that can be packaged inside a viral protein shell is limited. Because these shells are too small to encase the genes needed to encode the many enzymes and other proteins that are required to replicate even the simplest virus, viruses must hijack their host's biochemical machinery to reproduce themselves (Figure 9–29). A viral genome will typically encode both viral coat proteins and proteins that help the virus to commandeer the host enzymes needed to replicate its genetic material.

QUESTION 9–5

Discuss the following statement: "Viruses exist in the twilight zone of life: outside cells they are simply dead assemblies of molecules; inside cells, however, they are alive."

Retroviruses Reverse the Normal Flow of Genetic Information

Although there are many similarities between bacterial and eukaryotic viruses, one important class of viruses—the **retroviruses**—is found only in eukaryotic cells. In many respects, retroviruses resemble the retrotransposons we just discussed. A key feature of the replication cycle of both is a step in which DNA is synthesized using RNA as a template—hence the prefix *retro,* which refers to the reversal of the usual flow of information from DNA to RNA. Retroviruses are thought to have derived from a retrotransposon that long ago acquired additional genes encoding

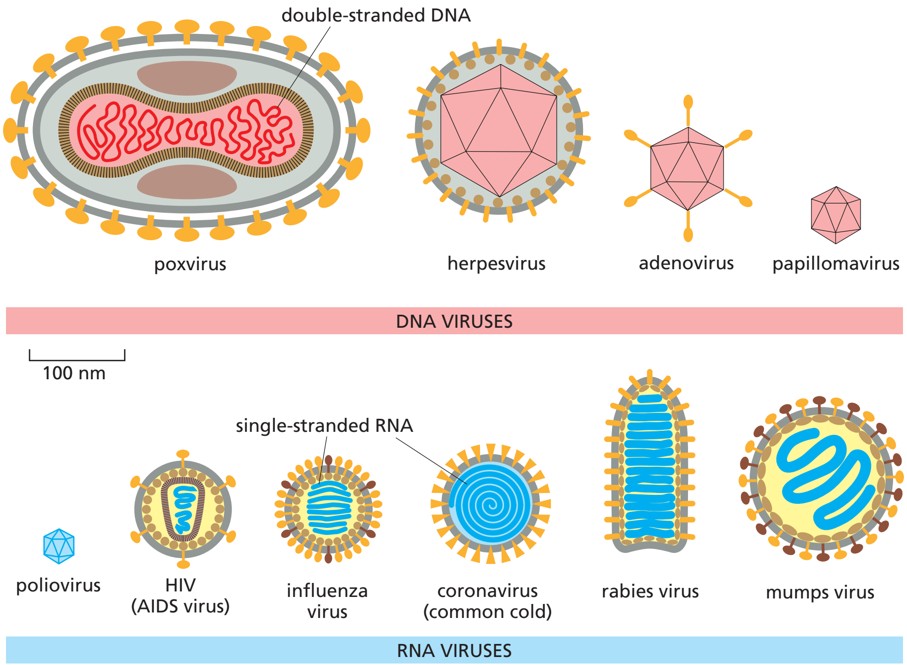


Figure 9–28 Viruses come in different shapes and sizes. Some of the viruses are shown in cross section (such as poxvirus and HIV). For others, the outer structure is emphasized. Some viruses (such as papilloma and polio) contain an outer surface that is composed solely of viral-encoded proteins. Others (such as poxvirus and HIV) bear a lipid-bilayer envelope (*gray*) in which viral-encoded proteins are embedded.

the coat proteins and other proteins required to make a virus particle. The RNA stage of its replicative cycle could then be packaged into a viral particle that could leave the cell.

Like retrotransposons, retroviruses use the enzyme reverse transcriptase to convert RNA into DNA. The enzyme is encoded by the retroviral genome, and a few molecules of the enzyme are packaged along with the RNA genome in each virus particle. When the single-stranded RNA genome of the retrovirus enters a cell, the reverse transcriptase brought in with it makes a complementary DNA strand to form a DNA/RNA hybrid double helix. The RNA strand is removed, and the reverse transcriptase (which can use either DNA or RNA as a template) now synthesizes a complementary DNA strand to produce a DNA double helix. This DNA is then inserted, or integrated, into a randomly selected site in the host genome by a virally encoded *integrase* enzyme. In this integrated state, the virus is *latent*: each time the host cell divides, it passes on a copy of the integrated viral genome, which is known as a *provirus*, to its progeny cells.

The next step in the replication of a retrovirus—which can take place long after its integration into the host genome—is the copying of the integrated viral DNA into RNA by a host-cell RNA polymerase, which produces large numbers of single-stranded RNAs identical to the original infecting genome. These viral RNAs are then translated by the host-cell ribosomes to produce the viral shell proteins, the envelope proteins, and reverse transcriptase—all of which are assembled with the RNA genome into new virus particles. The steps involved in the integration and replication of a retrovirus are shown in Figure 9–30.

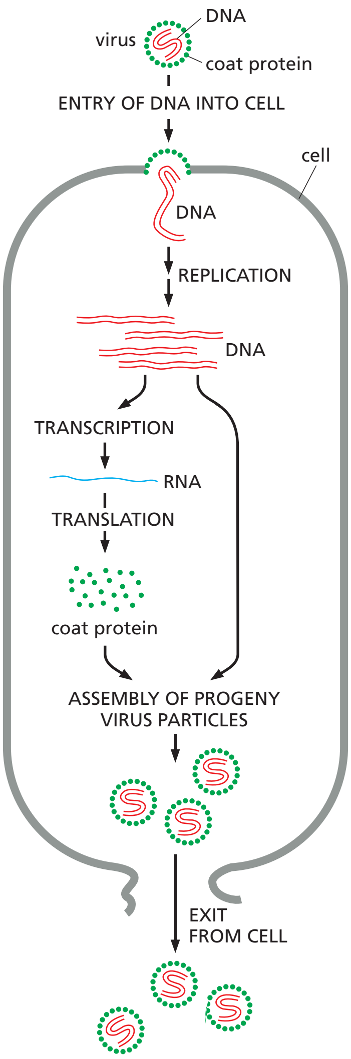


Figure 9–29 Viruses commandeer the host cell's molecular machinery to reproduce. The hypothetical virus illustrated here consists of a small, double-stranded DNA molecule that encodes just a single type of viral coat protein. To reproduce, the viral genome must first enter a host cell, where it is replicated to produce multiple copies, which are transcribed and translated to produce the viral coat protein. The viral genomes can then assemble spontaneously with the coat protein to form new virus particles, which escape from the cell by lysing it.

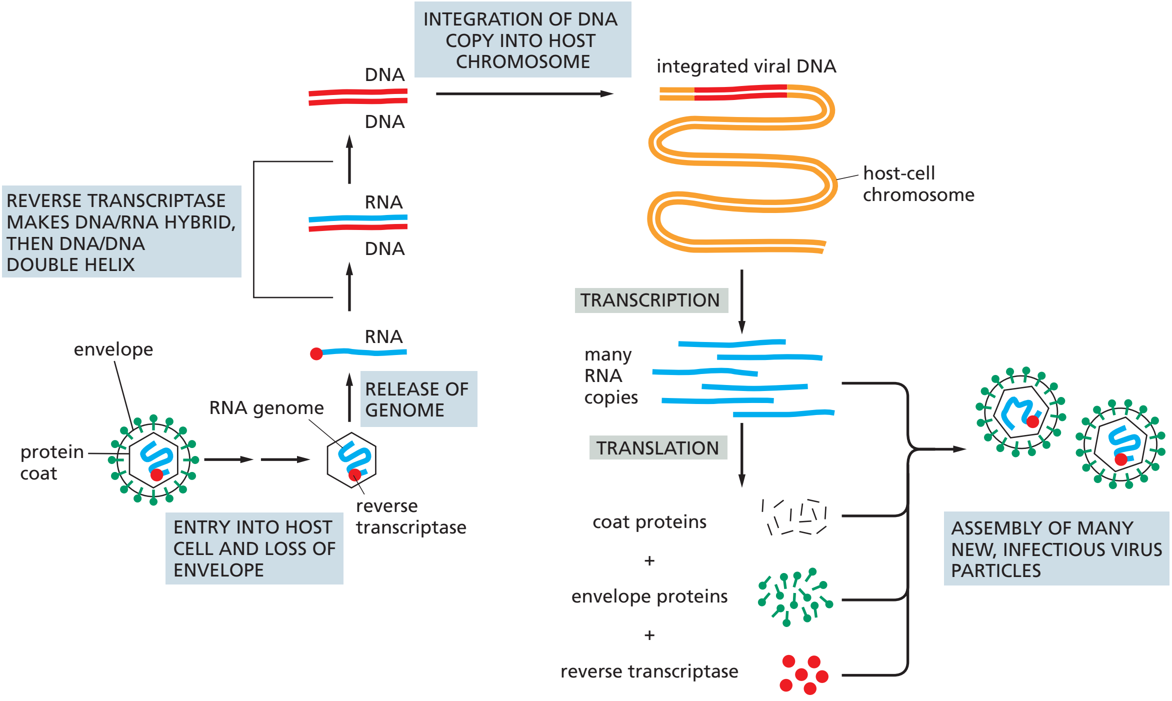


Figure 9–30 Infection by a retrovirus includes reverse transcription and integration of the viral genome into the host cell's DNA. The retrovirus genome consists of an RNA molecule (*blue*) that is typically between 7000 and 12,000 nucleotides in size. It is packaged inside a protein coat, which is surrounded by a lipid-bilayer envelope that contains virus-encoded envelope proteins (*green*). The enzyme reverse transcriptase (*red* circle), encoded by the viral genome and packaged with its RNA, first makes a single-stranded DNA copy of the viral RNA molecule and then a second DNA strand, generating a double-stranded DNA copy of the RNA genome. This DNA double helix is then integrated into a host chromosome, a step required for the synthesis of new viral RNA molecules by a host-cell RNA polymerase.

The human immunodeficiency virus (HIV), which is the cause of AIDS, is a retrovirus. As with other retroviruses, the HIV genome can persist in a latent state as a provirus embedded in the chromosomes of an infected cell. This ability to hide in host cells complicates attempts to treat the infection with antiviral drugs. But because the HIV reverse transcriptase is not used by cells for any purpose of their own, it is one of the prime targets of drugs currently used to treat AIDS.

EXAMINING THE HUMAN GENOME

The human genome contains an enormous amount of information about who we are and where we came from (**Figure 9–31**). Its 3.2×10^9 nucleotide pairs, spread out over 23 sets of chromosomes—22 autosomes and a pair of sex chromosomes (X and Y)—provide the instructions needed to build a human being. Yet, 25 years ago, biologists actively debated the value of determining the *human genome sequence*—the complete list of nucleotides contained in our chromosomes.

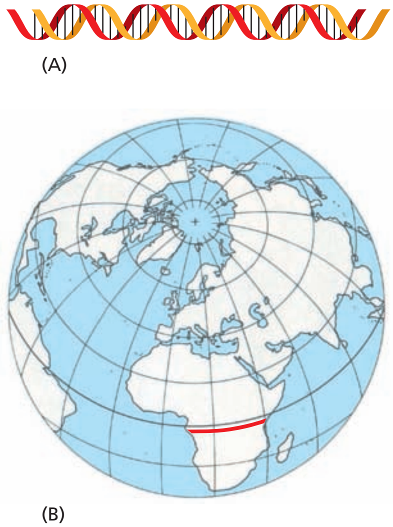


Figure 9–31 The 3 billion nucleotide pairs of the human genome contain a vast amount of information, including clues about our origins. If each nucleotide pair is drawn to span 1 mm, as shown in (A), the human genome would extend 3200 km (approximately 2000 miles)—far enough to stretch across central Africa, where humans first arose (*red* line in B). At this scale, there would be, on average, a protein-coding gene every 150 m. An average gene would extend for about 30 m, but the coding sequences (exons) in this gene would add up to only just over a meter; the rest would be introns.

The task was not simple. An international consortium of investigators labored tirelessly for the better part of a decade—and spent nearly $3 billion—to give us our first glimpse of this genetic blueprint. But the effort turned out to be well worth the cost, as the data continue to shape our thinking about how our genome functions and how it has evolved.

The first human genome sequence was just the beginning. The spectacular improvements in sequencing technologies (which we discuss in Chapter 10), coupled with powerful new tools for handling massive amounts of data, are taking genomics to a whole new level. The cost of DNA sequencing has dropped enormously since the Human Genome Project was launched in 1990, such that a whole human genome can now be sequenced in a few days for about $1000. Investigators around the world are collaborating to collect and compare the nucleotide sequences of thousands of human genomes. This resulting deluge of data offers tantalizing clues as to what makes us human, and what makes each of us unique.

Although it will take many years to analyze the rapidly accumulating genome data, the recent findings have already influenced the content of every chapter in this book. In this section, we describe some of the most striking features of the human genome—many of which were entirely unexpected. We review what genome comparisons can tell us about how we evolved, and we discuss some of the mysteries that still remain.

The Nucleotide Sequences of Human Genomes Show How Our Genes Are Arranged

When the DNA sequence of human Chromosome 22, one of the smallest human chromosomes, was completed in 1999, it became possible for the first time to see exactly how genes are arranged along an entire vertebrate chromosome (Figure 9–32). The subsequent publication of the

QUESTION 9–6

Mobile genetic elements, such as the *Alu* sequences, are found in many copies in human DNA. In what ways could the presence of an *Alu* sequence affect a nearby gene?

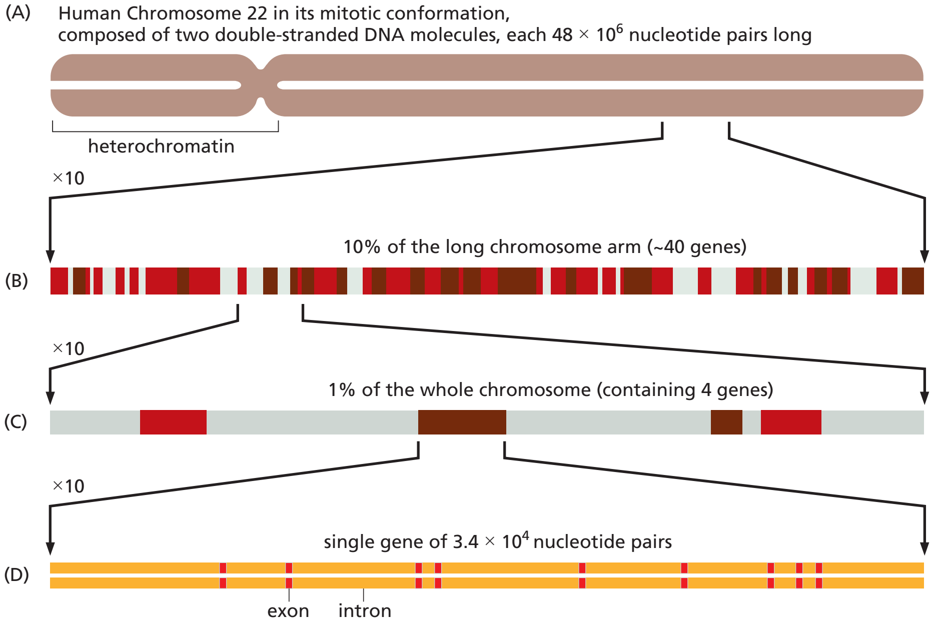

(A)
Human Chromosome 22 in its mitotic conformation, composed of two double-stranded DNA molecules, each 48 × 10⁶ nucleotide pairs long
heterochromatin
×10
(B)
10% of the long chromosome arm (~40 genes)
×10
(C)
1% of the whole chromosome (containing 4 genes)
×10
(D)
single gene of 3.4 × 10⁴ nucleotide pairs
exon
intron

Figure 9–32 The sequence of Chromosome 22 shows how human chromosomes are organized. (A) Chromosome 22, one of the smallest human chromosomes, contains 48×10^6 nucleotide pairs and makes up approximately 1.5% of the human genome. Most of the short arm of Chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin), as discussed in Chapter 5. (B) A tenfold expansion of a portion of Chromosome 22 shows about 40 genes. 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 (*yellow*) is relatively unimportant. (Adapted from The International Human Genome Sequencing Consortium, *Nature* 409:860–921, 2001.)

TABLE 9–2 SOME VITAL STATISTICS FOR THE HUMAN GENOME

DNA Length	3.2×10^9 Nucleotide Pairs*
Number of protein-coding genes	approximately 19,000
Number of non-protein-coding genes**	approximately 5000
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***	approximately 11,000
Percentage of DNA sequence in exons (protein-coding sequences)	1.5%
Percentage of DNA conserved with other mammals that does not encode protein****	3.0%
Percentage of DNA in high-copy repetitive elements	approximately 50%

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

**These include genes that encode structural, catalytic, and regulatory RNAs.

***A pseudogene is a DNA sequence that closely resembles that of a functional gene but contains 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.

****This includes DNA encoding 5′ and 3′ UTRs (untranslated regions of mRNAs), regulatory DNA sequences, and conserved regions of unknown function.

whole human genome sequence—a first draft in 2001 and a finished draft in 2004—provided a more panoramic view of the complete genetic landscape, including how many genes we have, what those genes look like, and how they are distributed across the genome (**Table 9–2**).

The first striking feature of the human genome is how little of it—less than 2%—codes for proteins (**Figure 9–33**). In addition, almost half of our DNA is made up of mobile genetic elements that have colonized our genome over evolutionary time. Because these elements have accumulated

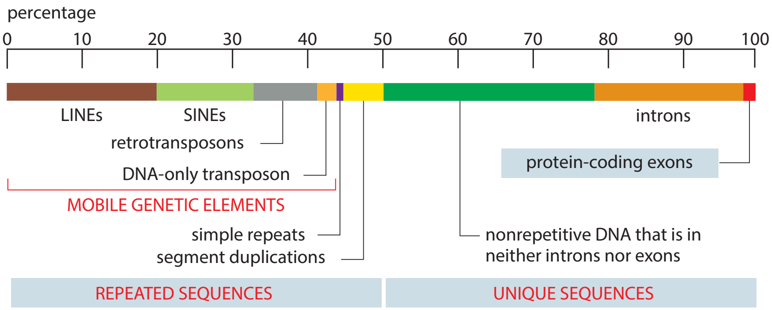


Figure 9–33 The bulk of the human genome is made of repetitive nucleotide sequences and other noncoding DNA. About half of our genome consists of repeated sequences. These include the LINEs (long interspersed nuclear elements, such as *L1*), SINEs (short interspersed nuclear elements, such as *Alu*), other retrotransposons, and DNA-only transposons—mobile genetic elements that have multiplied in our genome by replicating themselves and inserting the new copies in different positions. Most of these mobile genetic elements are fossils—remnants that are no longer capable of transposition. Simple repeats are short nucleotide sequences (less than 14 nucleotide pairs) that are repeated again and again for long stretches. Segment duplications are large blocks of the genome (1000–200,000 nucleotide pairs) that are present at two or more locations in the genome. These, too, represent repeated DNA sequences. The most highly repeated blocks of DNA in heterochromatin have not yet been completely sequenced; these comprise about 10% of human DNA sequences and are not represented in this diagram.

The unique sequences that are not part of any introns or exons (*dark green*) include regulatory DNA sequences, sequences that code for functional RNA, and sequences whose functions are not known. (Data courtesy of E.H. Margulies.)

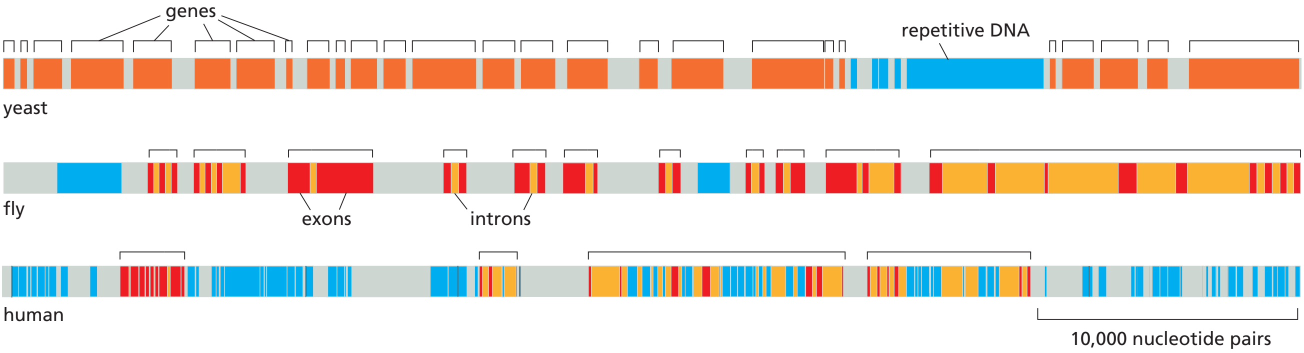


Figure 9–34 Genes are sparsely distributed in the human genome. Compared to some other eukaryotic genomes, the human genome is less gene-dense. Shown here are DNA segments about 50,000 nucleotide pairs in length from bakers yeast, *Drosophila*, and human. The human segment contains only 4 genes, compared to 26 in the yeast and 11 in the fly. Exons are shown in *orange*, introns in *yellow*, repetitive elements in *blue*, and intergenic DNA in *gray*. The genes of yeast and flies are generally more compact, with fewer introns, than the genes of humans.

mutations, most can no longer move; rather, they are relics from an earlier evolutionary era when mobile genetic elements ran rampant through our genome.

It was a surprise to discover how few protein-coding genes our genome actually contains. Earlier estimates had been in the neighborhood of 100,000 (as discussed in **How We Know**, pp. 324–325). Although the exact count is still being refined, current estimates place the number of human protein-coding genes at about 19,000, with perhaps another 5000 genes encoding functional RNAs that are not translated into proteins. This estimate brings us much closer to the gene numbers for simpler multicellular animals—for example, 14,000 protein-coding genes for *Drosophila*, 22,000 for *C. elegans*, and 28,000 for the small weed *Arabidopsis* (see Table 1–2).

The number of protein-coding genes we have may be unexpectedly small, but their relative size is unusually large. Only about 1300 nucleotide pairs are needed to encode an average-sized human protein of about 430 amino acids. Yet the average length of a human gene is 27,000 nucleotide pairs. Most of this DNA is in noncoding introns. In addition to the voluminous introns (see Figure 9–32D), each gene is associated with regulatory DNA sequences that ensure that the gene is expressed at the proper level, time, and place. In humans, these regulatory DNA sequences are typically interspersed along tens of thousands of nucleotide pairs, much of which seems to be “spacer” DNA. Indeed, compared to many other eukaryotic genomes, the human genome is much less densely packed (**Figure 9–34**).

Although exons and their associated regulatory DNA sequences comprise less than 2% of the human genome, comparative studies indicate that about 4.5% of the human genome is highly conserved when compared with other mammalian genomes (see Figure 9–20). An additional 5% of the genome shows reduced variation in the human population, as determined by comparing the DNA sequence of thousands of individuals. This reduced variation reflects the relative importance of these sequences compared with the majority of the genome. Taken together, such analyses suggest that only about 10% of the human genome contains sequences that truly matter—but we do not yet know the function of much of this DNA.

Differences in Gene Regulation May Help Explain How Animals with Similar Genomes Can Be So Different

We now have the complete genome sequences for many different mammals, including humans, chimpanzees, gorillas, orangutans, dogs, cats, and mice. All of these species contain essentially the same protein-coding genes, which raises a fundamental question: What makes these creatures so different from one another? And what makes humans different from other animals?

COUNTING GENES

How many genes does it take to make a human? It seems a natural thing to wonder. If about 6000 genes can produce a yeast and 14,000 a fly, how many are needed to make a human being—a creature curious and clever enough to study its own genome? Until researchers completed the first draft of the human genome sequence, the most frequently cited estimate was 100,000. But where did that figure come from? And how was the revised estimate of only 19,000 protein-coding genes derived?

Walter Gilbert, a physicist-turned-biologist who won a Nobel Prize for developing techniques for sequencing DNA, was one of the first to throw out a ballpark estimate of the number of human genes. In the mid-1980s, Gilbert suggested that humans could have 100,000 genes, an estimate based on the average size of the few human genes known at the time (about 3×10^4 nucleotide pairs) and the size of our genome (about 3×10^9 nucleotide pairs). This back-of-the-envelope calculation yielded a number with such a pleasing roundness that it wound up being quoted widely in articles and textbooks.

The calculation provides an estimate of the number of genes a human could have in principle, but it does not address the question of how many genes we actually have. As it turns out, that question is not so easy to answer, even with the complete human genome sequence in hand. The problem is, how does one identify a gene? Consider protein-coding genes, which comprise only 1.5% of the human genome. Looking at a given piece of raw DNA sequence—an apparently random string of As, Ts, Gs, and Cs—how can one tell which parts represent protein-coding segments? Being able to accurately and reliably distinguish the rare coding sequences from the more plentiful noncoding sequences in a genome is necessary before one can hope to locate and count its genes.

Signals and chunks

As always, the situation is simplest in bacteria and simple eukaryotes such as yeasts. In these genomes, genes that encode proteins are identified by searching through the entire DNA sequence looking for **open reading frames** (**ORFs**). These are long sequences—say, 100 codons or more—that lack stop codons. A random sequence of nucleotides will by chance encode a stop codon about once every 20 codons (as there are three stop codons in the set of 64 possible codons—see Figure 7–27). So finding an ORF—a continuous nucleotide sequence that encodes more than 100 amino acids—is the first step in identifying a good candidate for a protein-coding gene. Today, computer programs are used to search for such ORFs, which begin with an initiation codon, usually ATG, and end with a termination codon, TAA, TAG, or TGA (**Figure 9–35**).

In animals and plants, the process of identifying ORFs is complicated by the presence of large intron sequences, which interrupt the protein-coding portions of genes. As we have seen, these introns are generally much larger than the exons, which might represent only a few percent of the gene. In human DNA, exons sometimes contain as few as 50 codons (150 nucleotide pairs), while introns may exceed 10,000 nucleotide pairs in length. Fifty codons is too short to generate a statistically significant

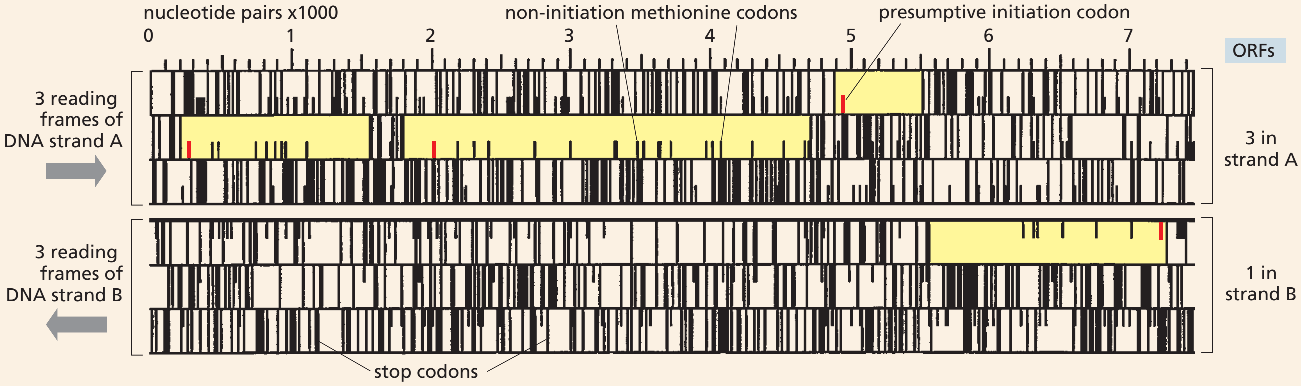

nucleotide pairs x1000
0
1
2
3
4
5
6
7
non-initiation methionine codons
presumptive initiation codon
ORFs
3 reading frames of DNA strand A
3 in strand A
3 reading frames of DNA strand B
1 in strand B
stop codons

Figure 9–35 Computer programs are used to identify protein-coding genes. In this example, a DNA sequence of 7500 nucleotide pairs from the pathogenic yeast *Candida albicans* was fed into a computer, which then calculated the proteins that could, in theory, be produced from each of its six possible reading frames—three on each of the two strands (see Figure 7–28). The output shows the location of start and stop codons for each reading frame. The reading frames are laid out in horizontal columns. Stop, or termination, codons (TGA, TAA, and TAG) are represented by tall, vertical black lines, and methionine codons (ATG) are represented by shorter black lines. Four open reading frames, or ORFs (shaded *yellow*), can be clearly identified by the statistically significant absence of stop codons. For each ORF, the presumptive initiation codon (ATG) is indicated in *red*. The additional ATG codons (*black*) in the ORFs code for methionine in the protein.

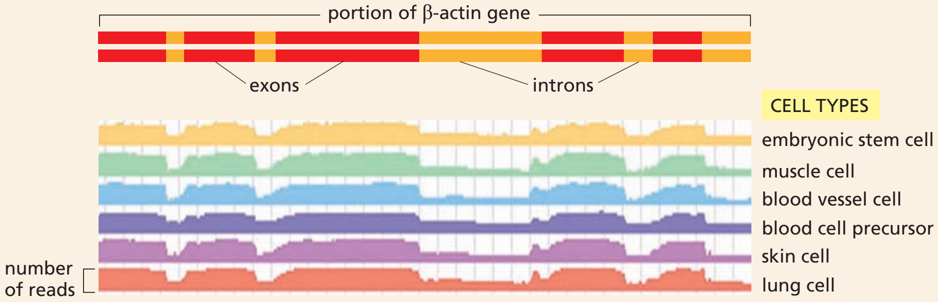


Figure 9–36 RNA sequencing can be used to characterize protein-coding genes. Presented here is a set of data corresponding to RNAs produced from a segment of the gene for β-actin, which is depicted schematically at the top. Millions of RNA "sequence reads," each approximately 200 nucleotides long, were collected from a variety of cell types (*right*) and matched to DNA sequences within the β-actin gene. The height of each trace is proportional to how often each sequence appears in a read. Exon sequences are present at high levels, reflecting their presence in mature β-actin mRNAs. Intron sequences are present at low levels, most likely reflecting their presence in pre-mRNA molecules that have not yet been spliced or spliced introns that have not yet been degraded.

"ORF signal," as it is not all that unusual for 50 random codons to lack a stop signal. Moreover, introns are so long that they are likely to contain by chance quite a bit of "ORF noise," numerous stretches of sequence lacking stop signals. Finding the true ORFs in this sea of information in which the noise often outweighs the signal can be difficult. To make the task more manageable, computers are used to search for other distinctive features that mark the presence of a protein-coding gene. These include the splicing sequences that signal an intron–exon boundary (see Figure 7–20), regulatory DNA sequences, or conservation with coding sequences from other organisms.

In 1992, researchers used a computer program to predict protein-coding regions in a preliminary human sequence. They found two genes in a 58,000-nucleotide-pair segment of Chromosome 4, and five genes in a 106,000-nucleotide-pair segment of Chromosome 19. That works out to an average of 1 gene every 23,000 nucleotide pairs. Extrapolating from that density to the whole genome would give humans nearly 130,000 genes. It turned out, however, that the chromosomes the researchers analyzed had been chosen for sequencing precisely because they appeared to be gene-rich. When the estimate was adjusted to take into account the gene-poor regions of the human genome—guessing that half of the human genome had maybe one-tenth of that gene-rich density—the estimated number dropped to 71,000.

Matching RNAs

Of course, these estimates are based on what we think genes look like; to get around this bias, we must employ more direct, experiment-based methods for locating genes. Because genes are transcribed into RNA, the preferred strategy for finding genes involves isolating all of the RNAs produced by a particular cell type and determining their nucleotide sequence—a technique called RNA-Seq. These sequences are then mapped back to the genome to locate their genes. For protein-coding genes, exon segments are more highly represented among the sequenced transcripts, as intron sequences tend to be spliced out and destroyed. Because different cell types express different genes, and splice their RNA transcripts differently, a variety of cell types are used in the analysis (**Figure 9–36**).

Thanks to RNA-Seq, the number of predicted protein-coding genes has dropped even further, because the technique detects only those genes that are actively transcribed. At the same time, the approach also allowed the detection of genes that do not code for proteins, but instead encode functional or regulatory RNAs. Many noncoding RNAs were first identified through RNA-Seq.

Human gene countdown

Based on a combination of all of these computational and experimental techniques, current estimates of the total number of human genes are now converging around 24,000, of which approximately 19,000 are protein-coding. It could be many years, however, before we have the final answer to how many genes it takes to make a human. In the end, having an exact count will not be nearly as important as understanding the functions of each gene and how they interact to build the living organism.

The instructions needed to produce a multicellular animal from a fertilized egg are provided, in large part, by the regulatory DNA sequences associated with each gene. These noncoding DNA sequences contain, scattered within them, dozens of separate regulatory elements, including short DNA segments that serve as binding sites for specific transcription regulators (discussed in Chapter 8). Regulatory DNA sequences ultimately dictate each organism's developmental program—the rules its cells follow as they proliferate, assess their positions in the embryo, and specialize by switching on and off specific genes at the right time and place. The evolution of species is likely to have more to do with innovations in regulatory DNA sequences than in the proteins or functional RNAs the genes encode.

Given the importance of regulatory DNA sequences in defining the characteristics of a species, one place to begin searching for clues to identity is in the regulatory DNA sequences that are highly conserved across mammalian species, but are altered or absent in our own genome. One study identified more than 500 such sequences, providing some intriguing clues as to what makes us human. One of these regulatory DNA sequences, missing in humans, seems to suppress the proliferation of neurons in the brain. Although further investigation is required, it is possible that the loss of this sequence—or changes in other neural-specific regulatory DNA sequences—played an instrumental role in the evolution of the human brain.

Another regulatory DNA sequence lost in the human lineage directs the formation of penile spines—structures present in a wide variety of mammals including chimpanzees, bonobos, gorillas, orangutans, gibbons, rhesus monkeys, and bushbabies. Whether the loss of these structures provides some advantage to humans is not known; it could be that the change is neutral—neither advantageous nor harmful. Regardless, it is a characteristic that makes us unique.

Thanks to such genetic comparisons, we are beginning to unravel the secrets of how our genome evolved to produce the qualities that define us as a species. But these analyses can only provide information about our distant evolutionary past. To learn about the more recent events in the history of modern *Homo sapiens*, we are turning to the genomes of our closest extinct relations, as we see next.

The Genome of Extinct Neanderthals Reveals Much about What Makes Us Human

In 2010, investigators completed their analysis of the first Neanderthal genome. One of our closest evolutionary relatives, Neanderthals lived side-by-side with the ancestors of modern humans in Europe and Western Asia. By comparing the Neanderthal genome sequence—obtained from DNA that was extracted from a fossilized bone fragment found in a cave in Croatia—with those of people from different parts of the world, researchers identified a handful of genomic regions that have undergone a sudden spurt of changes in modern humans. These regions include genes involved in metabolism, brain development, the voice box, and the shape of the skeleton, particularly the rib cage and brow—all features thought to differ between modern humans and our extinct cousins.

Remarkably, these studies also revealed that many modern humans—particularly those that hail from Europe and Asia—share about 2% of their genomes with Neanderthals. This genetic overlap indicates that our ancestors mated with Neanderthals—before outcompeting or actively exterminating them—on the way out of Africa (**Figure 9–37**). This ancient relationship left a permanent mark in the human genome.

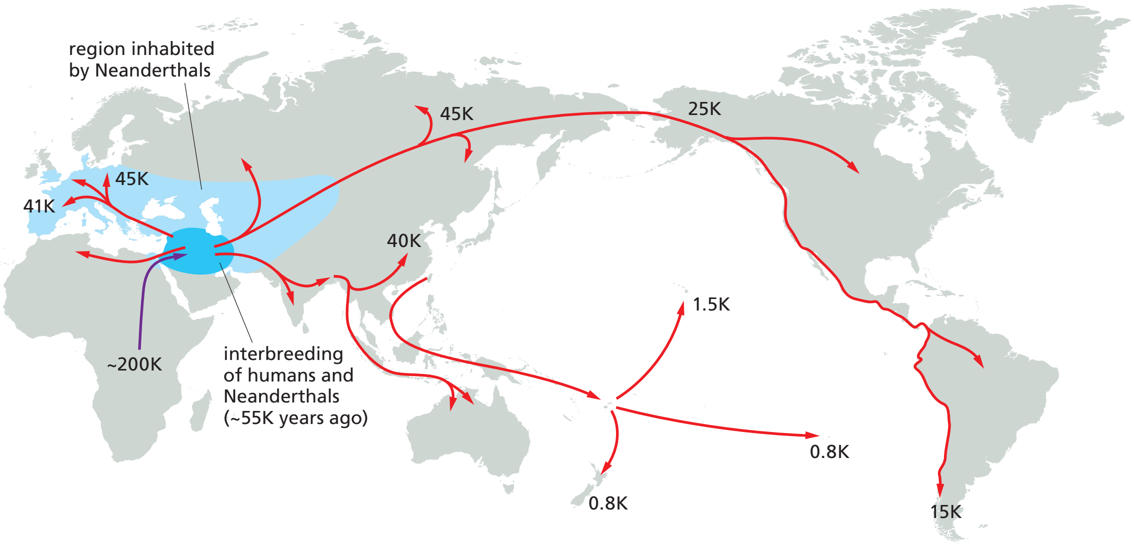

region inhabited by Neanderthals
45K
41K
45K
25K
40K
1.5K
~200K
interbreeding of humans and Neanderthals (~55K years ago)
0.8K
0.8K
15K

Figure 9–37 Ancestral humans encountered Neanderthals on their way out of Africa. Modern humans descended from a relatively small population—perhaps as few as 10,000 individuals—that existed in Africa approximately 200,000 (200 K) years ago. Among that small group of ancestors, some migrated northward, and their descendants spread across the globe. As ancestral humans left Africa, around 130,000 years ago (*purple* arrows), they encountered Neanderthals who inhabited the region indicated in *light blue*. As a result of interbreeding (in the region shown in *dark blue*), the humans that subsequently spread throughout Europe and Asia (*red* arrows) carried with them traces of Neanderthal DNA. Ultimately, ancestral humans continued their global spread to the New World, reaching North America approximately 25,000 years ago and the southern regions of South America 15,000 years later. This scenario is based on many types of data, including fossil records, anthropological studies, and the genome sequences of Neanderthals and of humans from around the world. (Adapted from M.A. Jobling et al., *Human Evolutionary Genetics*, 2nd ed. New York: Garland Science, 2014.)

Genome Variation Contributes to Our Individuality—But How?

With the possible exception of some identical twins, no two people have exactly the same genome sequence. When the same region of the genome from two different humans is compared, the nucleotide sequences typically differ by about 0.1%. This degree of variation represents about 1 difference in every 1000 nucleotide pairs—or some 3 million genetic differences between the genome of one person and the next. Detailed analyses of human genetic variation suggest that the bulk of this variation was already present early in our evolution, perhaps 200,000 years ago, when the human population was still small. Yet much of this variation has been reshuffled as more and more generations of humans have arisen. Thus, although a great deal of the genetic diversity in present-day humans was inherited from our early human ancestors, each individual inherits a unique combination of this ancient genetic variation.

Sprinkled on top of this "tossed salad" of ancient variation are mutations that are much more recent. At birth, each human's genome contains approximately 70 new mutations that were not present in the genomes of either parent. Combined with the jumbled collection of ancient variation we acquired from our ancestors, these recent mutations further distinguish one individual from another. Most of the variation in the human genome takes the form of single base-pair changes. Although some of these base-pair changes are unique to individual humans, many more are preserved from our distant ancestors and are therefore widespread in the human population. Those single-base changes that are present in at least 1% of the population are called **single-nucleotide polymorphisms** (**SNPs**, pronounced "snips"). These polymorphisms are simply points in the genome that differ in nucleotide sequence between one portion of the population and another—positions where, for example, more than 1%

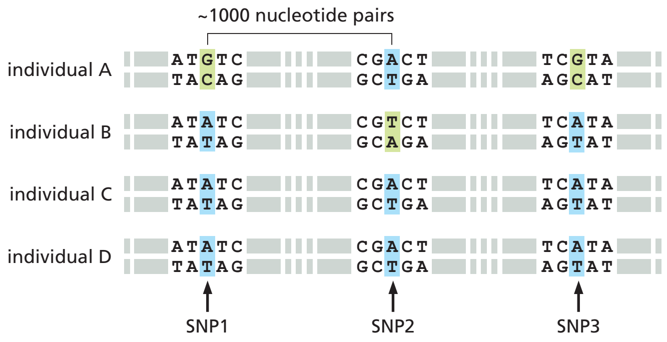


Figure 9–38 Single-nucleotide polymorphisms (SNPs) are points in the genome that differ by a single nucleotide pair between one portion of the population and another. Here, the differences are highlighted in *green* and *blue*. By convention, to count as a polymorphism, a genetic difference must be present in at least 1% of the total population of the species. Most, but not all, SNPs in the human genome occur in regions where they do not affect the function of a gene. As indicated by the bracket, when comparing any two humans one finds, on average, about one SNP per every 1000 nucleotide pairs.

of the population has a G-C nucleotide pair, while the rest have an A-T (**Figure 9–38**). Two human genomes chosen at random from the world's population will differ by approximately 2.5 × 10^6 SNPs that are scattered throughout the genome.

Most of these SNPs are genetically silent, as they fall within noncritical regions of the genome. Such variations have no effect on how we look or how our cells function. This means that only a small subset of the variation we observe in our DNA is responsible for the heritable differences from one human to the next. We discussed one such difference—that responsible for the ability of some adults to digest milk—earlier in the chapter. However, it remains a major challenge to identify the thousands of other genetic variations that are functionally important—a problem we return to in Chapter 19.

Genome sequences hold the secrets to why humans look, think, and act the way we do—and why one human differs from another. Our genome contains the instructions that guide the countless decisions made by all of our cells as they interact with one another to build our tissues and organs. But we are only just beginning to learn the grammar and rules by which this genetic information orchestrates our biology and our behavior. Deciphering this code—which has been shaped by evolution and refined by individual variation—is one of the great challenges facing the next generation of cell biologists.

ESSENTIAL CONCEPTS

- **By comparing the DNA and protein sequences of contemporary organisms, we are beginning to reconstruct how genomes have evolved in the billions of years that have elapsed since the appearance of the first cells.**
- **Genetic variation—the raw material for evolutionary change—arises through a variety of mechanisms that alter the nucleotide sequence of genomes. These changes in sequence range from simple point mutations to larger-scale deletions, duplications, and rearrangements.**
- **Genetic changes that give an organism a selective advantage are likely to be perpetuated. Changes that compromise an organism's fitness or ability to reproduce are eliminated through natural selection.**
- **Gene duplication is one of the most important sources of genetic diversity. Once duplicated, the two genes can accumulate different mutations and thereby diversify to perform different roles.**
- **Repeated rounds of gene duplication and divergence during evolution have produced many large gene families.**
- **The evolution of new proteins is thought to have been greatly facilitated by the swapping of exons between genes to create hybrid proteins with new functions.**

- The human genome contains 3.2 × 10^9 nucleotide pairs distributed among 23 pairs of chromosomes—22 autosomes and a pair of sex chromosomes. Less than a tenth of this DNA is transcribed to produce protein-coding or otherwise functional RNAs.
- Individual humans differ from one another by an average of 1 nucleotide pair in every 1000; this and other genetic variation underlies most of our individuality and provides the basis for identifying individuals by DNA analysis.
- Nearly half of the human genome consists of mobile genetic elements that can move from one site to another within a genome. Two classes of these elements have multiplied to especially high copy numbers.
- Viruses are genes packaged in protective coats that can move from cell to cell and organism to organism, but they require host cells to reproduce.
- Some viruses have RNA instead of DNA as their genetic material. To reproduce, retroviruses copy their RNA genomes into DNA, and integrate into the host-cell genome.
- Comparing genome sequences of different species provides a powerful way to identify conserved, functionally important DNA sequences.
- Related species, such as human and mouse, have many genes in common; evolutionary changes in the regulatory DNA sequences that affect how these genes are expressed are especially important in determining the differences between species.
- A comparison of genome sequences from people around the world has helped reveal how humans have evolved and spread across the globe.

KEY TERMS

Alu sequence
conserved synteny
exon shuffling
gamete
gene duplication and divergence
gene family
germ line
homologous gene
horizontal gene transfer
L1 element
mobile genetic element
open reading frame (ORF)
phylogenetic tree
point mutation
purifying selection
retrotransposon
retrovirus
reverse transcriptase
single-nucleotide polymorphism (SNP)
somatic cell
transposon
virus

QUESTIONS

QUESTION 9–7

Discuss the following statement: “Mobile genetic elements are parasites. They are always harmful to the host organism.”

QUESTION 9–8

Human Chromosome 22 (48 × 10^6 nucleotide pairs in length) has about 700 protein-coding genes, which average 19,000 nucleotide pairs in length and contain an average of 5.4 exons, each of which averages 266 nucleotide pairs. What fraction of the average protein-coding gene is converted into mRNA? What fraction of the chromosome do these genes occupy?

QUESTION 9–9

(True or False?) The DNA sequence of most of the human genome is unimportant. Explain your answer.

QUESTION 9–10

Mobile genetic elements make up nearly half of the human genome and are inserted more or less randomly throughout it. However, in some spots these elements are rare, as illustrated for a cluster of genes called HoxD, which lies on Chromosome 2 (Figure Q9–10). This cluster is about 100 kb in length and contains nine genes whose differential expression along the length of the developing

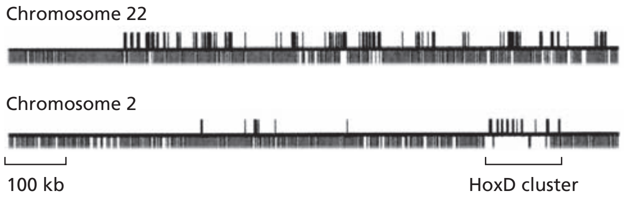


Figure Q9–10

embryo helps establish the basic body plan for humans (and other animals). In Figure Q9–10, lines that project *upward* indicate exons of known genes. Lines that project *downward* indicate mobile genetic elements; they are so numerous they merge into nearly a solid block outside the HoxD cluster. For comparison, an equivalent region of Chromosome 22 is shown. Why do you suppose that mobile genetic elements are so rare in the HoxD cluster?

QUESTION 9–11

An early graphical method for comparing nucleotide sequences—the so-called diagon plot—still yields one of the best visual comparisons of sequence relatedness. An example is illustrated in Figure Q9–11, in which the human β-globin gene is compared with the human cDNA for β globin (which contains only the coding portion of the gene; Figure Q9–11A) and with the mouse β-globin gene (Figure Q9–11B). Diagon plots are generated by comparing blocks of sequence, in this case blocks of 11 nucleotides at a time. If 9 or more of the nucleotides match, a dot is placed on the diagram at the coordinates corresponding to the blocks being compared. A comparison of all possible blocks generates diagrams such as the ones shown in Figure Q9–11, in which sequence similarities show up as diagonal lines.

A. From the comparison of the human β-globin gene with the human β-globin cDNA (Figure Q9–11A), can you deduce the positions of exons and introns in the β-globin gene?

B. Are the exons of the human β-globin gene (indicated by shading in Figure Q9–11B) similar to those of the mouse β-globin gene? Identify and explain any key differences.

C. Is there any sequence similarity between the human and mouse β-globin genes that lies outside the exons? If so, identify its location and offer an explanation for its preservation during evolution.

D. Did the mouse or human gene undergo a change of intron length during their evolutionary divergence? How can you tell?

QUESTION 9–12

Your advisor suggests that you write a computer program that will identify the exons of protein-coding genes directly from the sequence of the human genome. In preparation for that task, you decide to write down a list of the features that might distinguish protein-coding sequences from intronic DNA and from other sequences in the genome. What features would you list? (You may wish to review basic aspects of gene expression in Chapter 7.)

QUESTION 9–13

You are interested in finding out the function of a particular gene in the mouse genome. You have determined the nucleotide sequence of the gene, defined the portion that codes for its protein product, and searched the relevant database for similar sequences; however, neither the gene nor the encoded protein resembles anything previously described. What types of additional information about the gene and the encoded protein would you like to know in order to narrow down its function, and why? Focus on the information you would want, rather than on the techniques you might use to get that information.

QUESTION 9–14

Why do you expect to encounter a stop codon about every 20 codons or so in a random sequence of DNA?

QUESTION 9–15

Which of the processes listed below contribute significantly to the evolution of new protein-coding genes?

A. Duplication of genes to create extra copies that can acquire new functions.

B. Formation of new genes *de novo* from noncoding DNA in the genome.

C. Horizontal transfer of DNA between cells of different species.

D. Mutation of existing genes to create new functions.

E. Shuffling of protein domains by gene rearrangement.

QUESTION 9–16

Some protein sequences evolve more rapidly than others. But how can this be demonstrated? One approach is to compare several genes from the same two species, as shown for rat and human in the table. Two measures of rates of nucleotide substitution are indicated in the table. Nonsynonymous changes refer to single-nucleotide changes in the DNA sequence that alter the encoded amino acid (for example, ATC → TTC, which gives isoleucine → phenylalanine). Synonymous changes refer

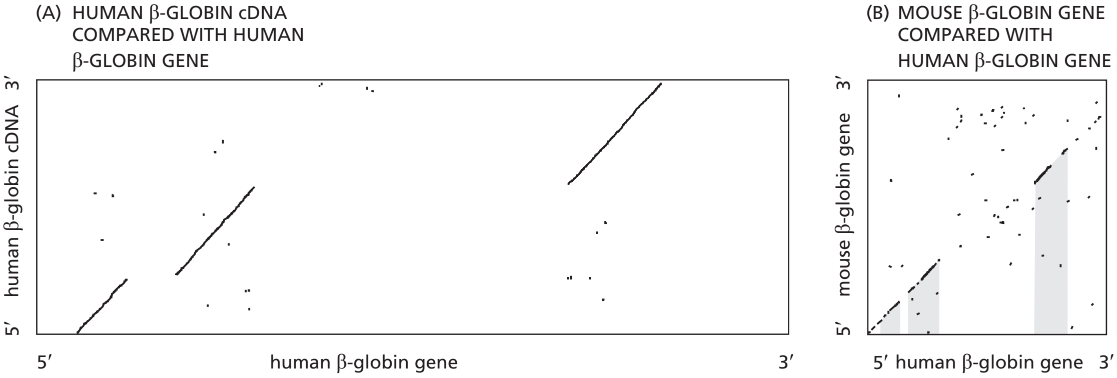


Figure Q9–11

Gene	Amino Acids	Rates of Change	
		Nonsynonymous	Synonymous
Histone H3	135	0.0	4.5
Hemoglobin α	141	0.6	4.4
Interferon γ	136	3.1	5.5

Rates were determined by comparing rat and human sequences and are expressed as nucleotide changes per site per 10^9 years. The average rate of nonsynonymous changes for several dozen rat and human genes is about 0.8.

to those that do not alter the encoded amino acid (ATC → ATT, which gives isoleucine → isoleucine, for example). (As is apparent in the genetic code, Figure 7–27, there are many cases where several codons correspond to the same amino acid.)

A. Why are there such large differences between the synonymous and nonsynonymous rates of nucleotide substitution?

B. Considering that the rates of synonymous changes are about the same for all three genes, how is it possible for the histone H3 gene to resist so effectively those nucleotide changes that alter its amino acid sequence?

C. In principle, a protein might be highly conserved because its gene exists in a "privileged" site in the genome that is subject to very low mutation rates. What feature of the data in the table argues against this possibility for the histone H3 protein?

QUESTION 9–17

Hemoglobin-like proteins were discovered in legumes, where they function in root nodules to lower the oxygen concentration, allowing the resident bacteria to fix nitrogen. These plant "hemoglobins" impart a characteristic pink color to the root nodules. The discovery of hemoglobin in plants was initially surprising because scientists regarded hemoglobin as a distinctive feature of animal blood. It was hypothesized that the plant hemoglobin gene was acquired by horizontal transfer from an animal. Many more hemoglobin-like genes have now been discovered and sequenced from a variety of organisms, and a phylogenetic tree of hemoglobins is shown in Figure Q9–17.

A. Does the evidence in the tree support or refute the hypothesis that the plant hemoglobins arose by horizontal gene transfer from animals?

B. Supposing that the plant hemoglobin genes were originally derived by horizontal transfer (from a parasitic nematode, for example), what would you expect the phylogenetic tree to look like?

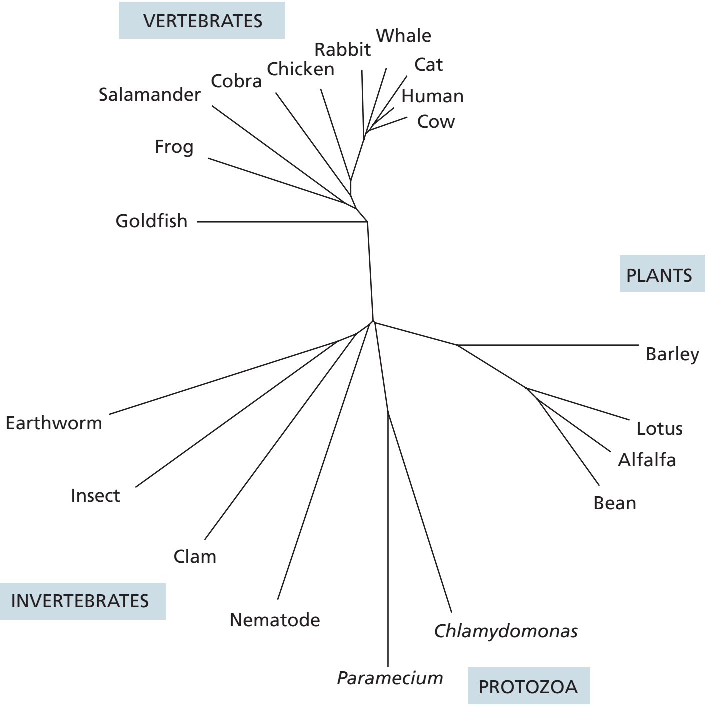


Figure Q9–17

QUESTION 9–18

The accuracy of DNA replication in the human germ-cell line is such that on average only about 0.6 out of the 6 billion nucleotides is altered at each cell division. Because most of our DNA is not subject to any precise constraint on its sequence, most of these changes are selectively neutral. Any two modern humans chosen at random will differ by about 1 nucleotide pair per 1000. Suppose we are all descended from a single pair of ancestors (an "Adam and Eve") who were genetically identical and homozygous (each chromosome was identical to its homolog). Assuming that all germ-line mutations that arise are preserved in descendants, how many cell generations must have elapsed since the days of our original ancestor parents for 1 difference per 1000 nucleotides to have accumulated in modern humans? Assuming that each human generation corresponds on average to 200 cell-division cycles in the germ-cell lineage and allowing 30 years per human generation, how many years ago would this ancestral couple have lived?

QUESTION 9–19

Reverse transcriptases do not proofread as they synthesize DNA using an RNA template. What do you think the consequences of this are for the treatment of AIDS?