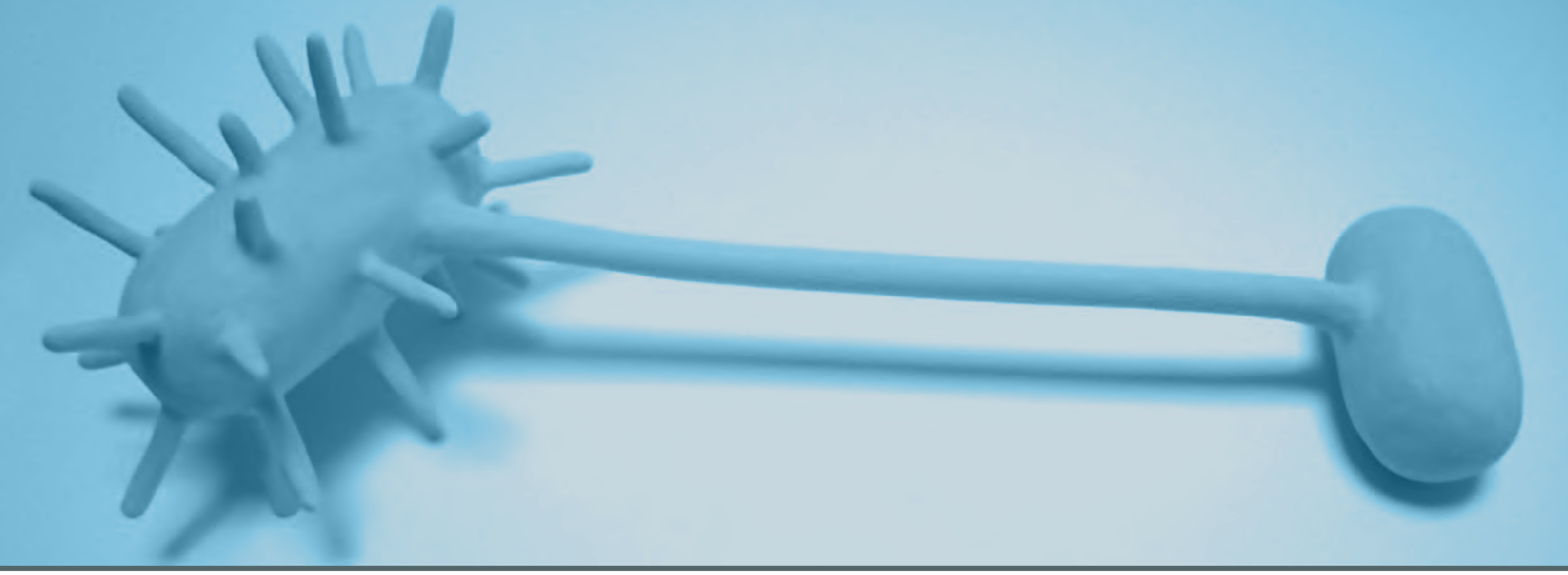


CHAPTER **NINE** 9

How Genes and Genomes Evolve

For a given individual, the nucleotide sequence of the genome in virtually every one of its cells is 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 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 and the plague bacterium to people from around the world—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. Our ability to compare the genomes of a wide-ranging collection of organisms has provided striking confirmation of Darwin's explanations for the diversity of life on Earth—revealing how processes of mutation and natural selection have been sculpting DNA sequences for billions of years, giving rise to the spectacular menagerie of present-day life-forms that crowd every corner of the planet.

Figure 9–1 **Small differences in DNA sequence account for differences in appearance between one individual and the next.** A group of English schoolchildren displays a sampling of the characteristics that define the unity and diversity of our own species. (Courtesy of Fiona Pragoff, Wellcome Images.)

In this chapter, we discuss how genes and genomes change over time. 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 sequences. We 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 our own DNA sequences tell us about who we are and where we come from.

GENERATING GENETIC VARIATION

Evolution is more a tinkerer than an inventor: it uses as its raw materials the DNA sequences that each organism inherits from its ancestors. There is no natural mechanism for making long stretches of entirely novel nucleotide sequences. 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 transcript or change the stability, activity, location, or interactions of its encoded protein or RNA product.
- *Mutation within regulatory DNA*: 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*: An existing gene, a larger segment of DNA, or even a whole genome can be duplicated, creating a set of closely related genes within a single cell. As this cell and its progeny divide, the original DNA sequence and its duplicate can acquire additional 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, the breaking and rejoining often occurs within the long intron sequences, which do not encode protein. Because intron sequences are removed by RNA splicing, the breaking and joining do not have to be precise to result in a functional gene.
- *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 transferred 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.

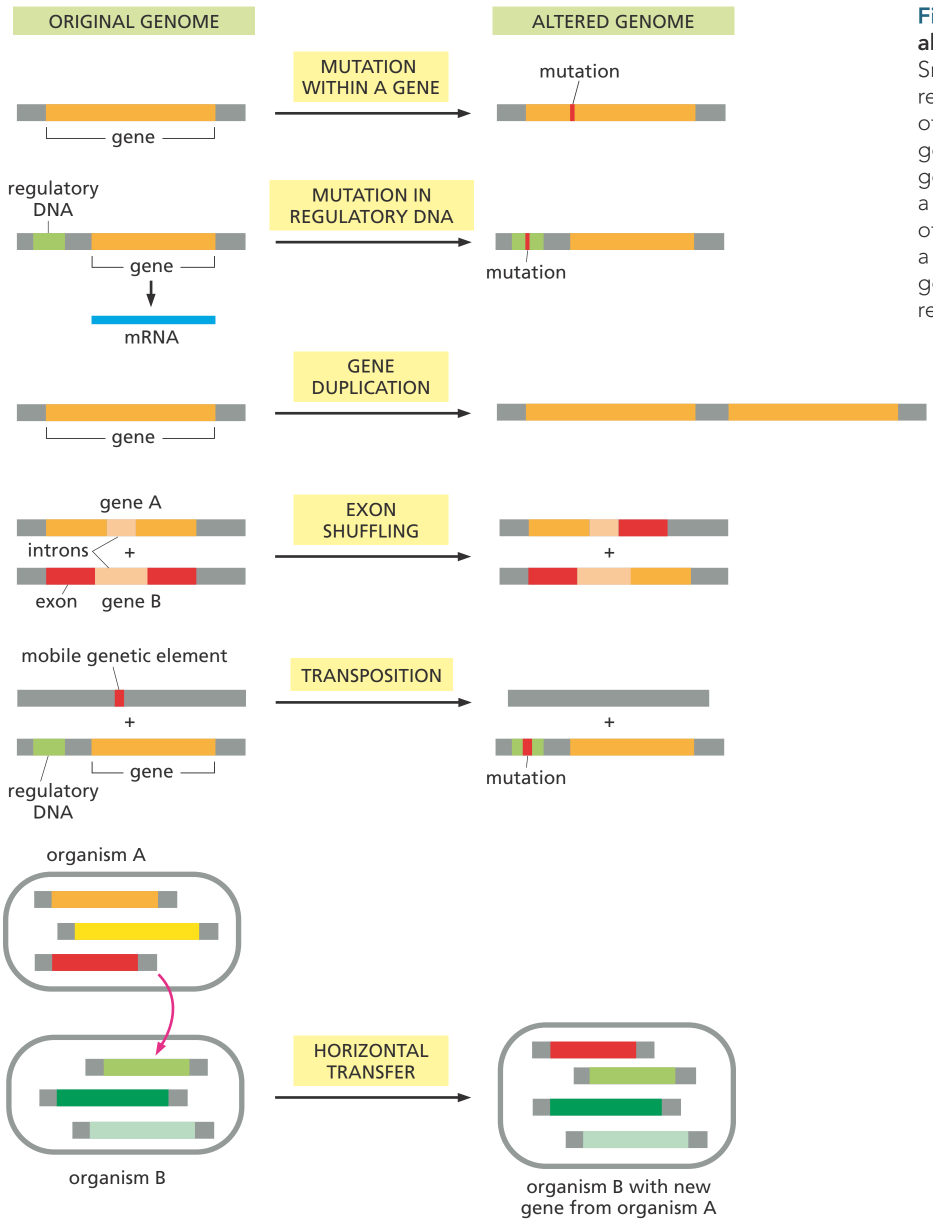


Figure 9–2 Genes and genomes can be altered by several different mechanisms. Small mutations, duplications, deletions, rearrangements, and even the infusion of fresh genetic material all contribute to genome evolution. Although the mobile genetic element here is shown interrupting a gene regulatory sequence, the movement of these parasitic elements can promote a variety of genetic variations, including gene duplication, exon shuffling, and other regulatory and structural alterations.

Each of these forms of genetic variation—from the simple mutations that occur within a gene to the more extensive duplications, deletions, rearrangements, and additions that occur within a genome—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 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 mainly 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.

QUESTION 9–1

In this chapter, it is argued 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?

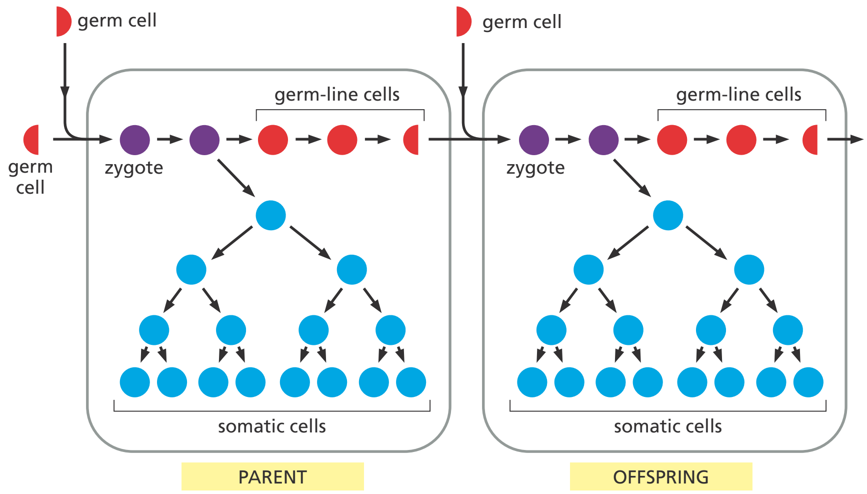


Figure 9–3 Germ-line cells and somatic cells have fundamentally different functions. In sexually reproducing organisms, genetic information is propagated into the next generation exclusively by germ-line cells (*red*). This cell lineage includes the specialized reproductive cells—the germ cells (eggs and sperm, *red* half circles)—which contain only half the number of chromosomes than do the other cells in the body (full circles). When two germ cells 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.

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 **germ cells**—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 their own (Figure 9–3). In a sense, somatic cells exist only to help the germ cells survive and propagate.

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**—the cell lineage that gives rise to the germ cells (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 germ cells 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 in detail in Chapter 19. In the meantime, aside from this mating-based genome

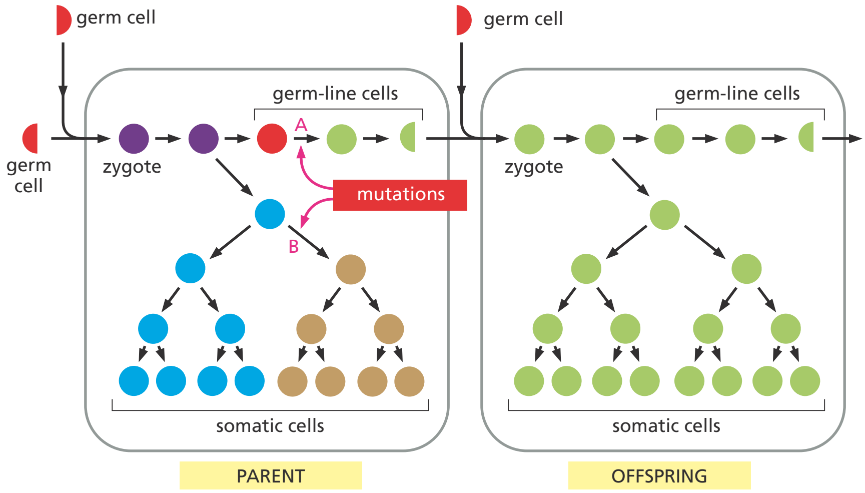


Figure 9–4 Mutations in germ-line cells and somatic cells have different consequences. A mutation that occurs in a germ-line cell (A) can be passed on to the cell's progeny and, ultimately, to the progeny of the organism (*green*). By contrast, a mutation that arises in a somatic cell (B) affects only the progeny of that cell (*orange*) and will not be passed on to the organism's progeny. As we discuss in Chapter 20, somatic mutations are responsible for most human cancers (see pp. 714–717).

reshuffling, which influences how mutations are inherited in organisms that reproduce sexually, most of the mechanisms that generate genetic change are the same for all living things, as we now discuss.

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, each nucleotide pair in an organism's genome runs a 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 the ancestor cell. Some 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 per 10^{10} nucleotide pairs each cell generation. The 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), is

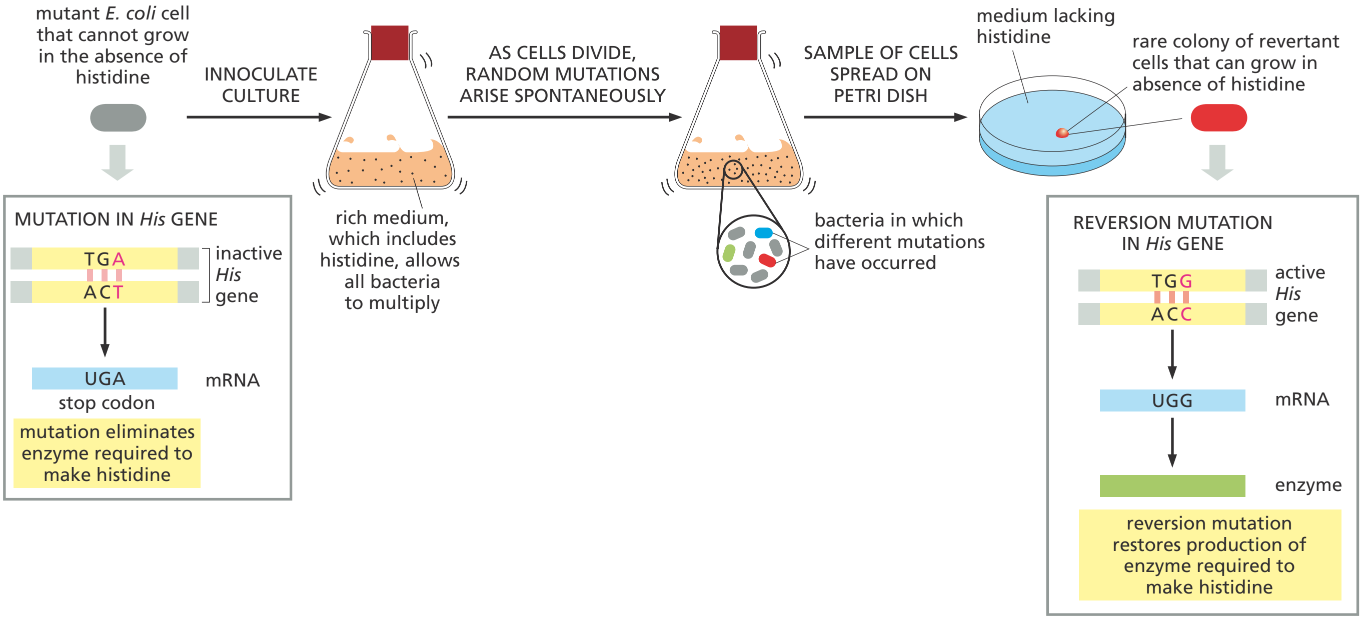

mutant E. coli cell that cannot grow in the absence of histidine
INNOCULATE CULTURE
AS CELLS DIVIDE, RANDOM MUTATIONS ARISE SPONTANEOUSLY
SAMPLE OF CELLS SPREAD ON PETRI DISH
medium lacking histidine
rare colony of revertant cells that can grow in absence of histidine
rich medium, which includes histidine, allows all bacteria to multiply
bacteria in which different mutations have occurred
MUTATION IN His GENE
TGA
ACT
inactive His gene
UGA
mRNA
stop codon
mutation eliminates enzyme required to make histidine
REVERSION MUTATION IN His GENE
TGG
ACC
active His gene
UGG
mRNA
enzyme
reversion 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 converts 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 "reversion" mutation (in which the A-T is changed back to a G-C). This reversion corrects the original defect and now 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.

about one-third that of *E. coli*—which suggests that the mechanisms that evolved to maintain genome integrity operate with an efficiency that does not differ significantly in 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.

Point Mutations Can Change the Regulation of a Gene

Mutations in the coding sequences of genes are fairly easy to spot because they change the amino acid sequence of the encoded protein in predictable ways. But mutations in regulatory DNA are more difficult to recognize, because they don't affect protein sequence and can be located some distance from the coding sequence of the gene.

Despite these difficulties, many examples have been discovered where point mutations in regulatory DNA 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 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 domestic animals some 10,000 years ago, variant genes—produced by random mutation—enabled those who carried the variation to continue to express lactase as adults. We now know that people who retain the ability to digest milk as adults contain a point mutation in the regulatory DNA 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 ability to digest lactose. It is remarkable how quickly this trait spread through the human population, especially in societies that depended heavily on milk for nutrition (**Figure 9–6**).

These evolutionary changes in the regulatory 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 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 that might allow it to perform a slightly different function—as long as the original activity of the gene is not lost. This specialization of duplicated genes occurs gradually, as mutations

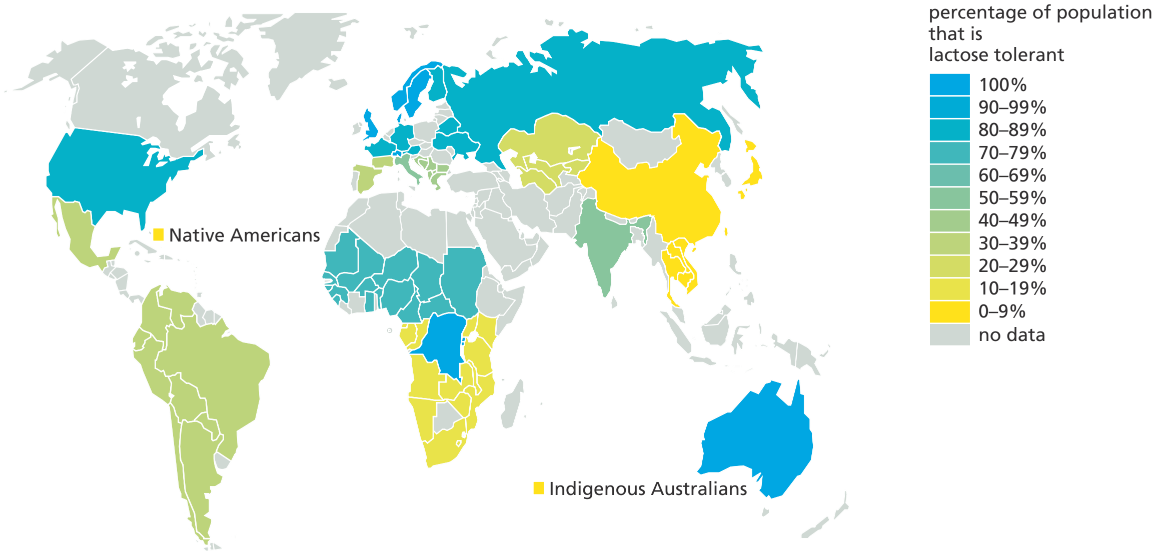


Figure 9–6 The ability of adult humans to digest milk followed the domestication of cattle. Approximately 10,000 years ago, humans in northern Europe and central Africa began to raise cattle. The subsequent availability of cow's milk—particularly during periods of starvation—gave a selective advantage 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. For example, the migration of Northern Europeans to North America and Australia explains why most people living on these continents can digest lactose as adults; the native populations of North America and Australia, however, remain lactose intolerant.

accumulate in the descendants of the original cell in which gene duplication occurred. 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?

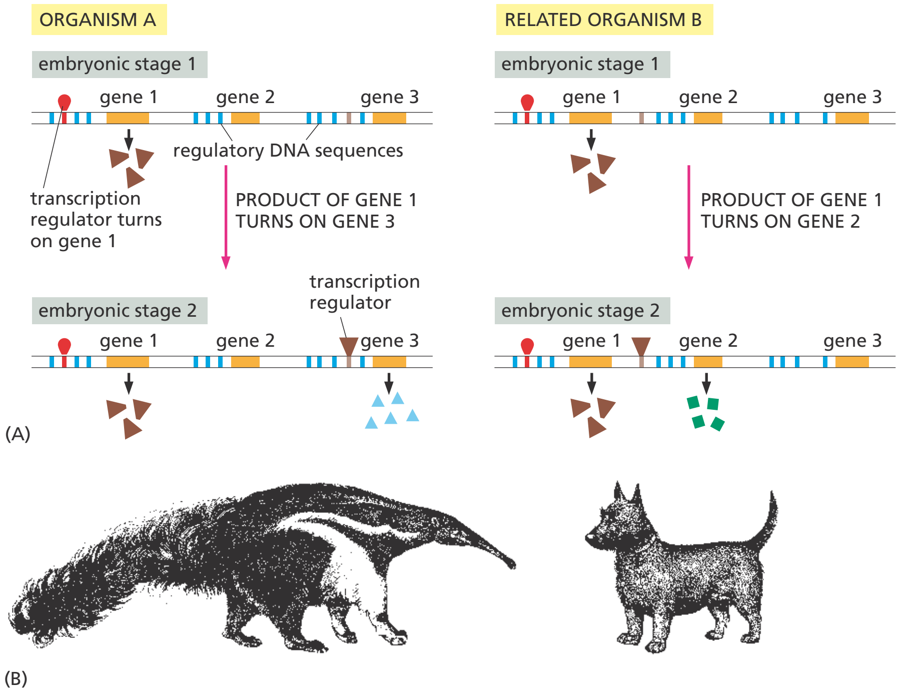


Figure 9–7 Changes in regulatory DNA sequences can have dramatic consequences for the development of an organism. (A) In this hypothetical example, the genomes of organisms A and B contain the same three genes (1, 2, and 3) and encode the same two transcription regulators (*red oval, brown triangle*). However, the regulatory DNA controlling expression of genes 2 and 3 is different in the two organisms. Although both express the same gene—gene 1—during embryonic stage 1, the differences in their regulatory DNA cause them to express different genes in stage 2. (B) 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.

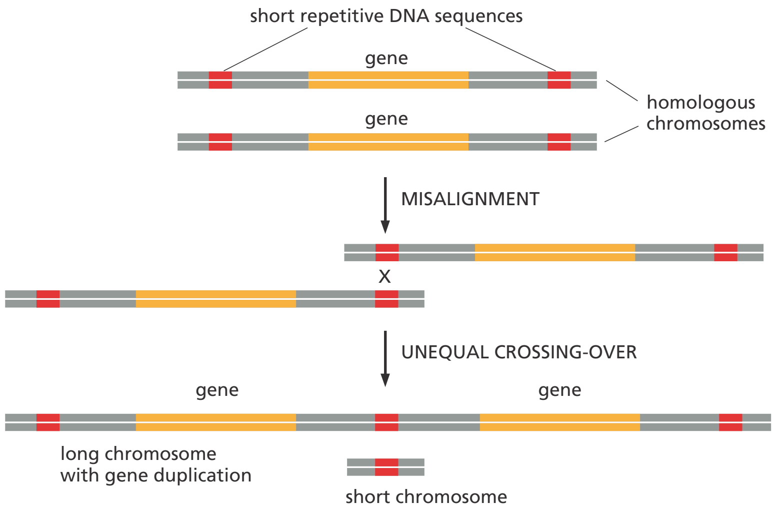


Figure 9–8 Gene duplication can be caused by crossovers between short, repeated DNA sequences in adjacent homologous chromosomes. The two chromosomes shown here undergo homologous recombination at short repeated sequences (*red*), that bracket a gene (*orange*). These repeated sequences can be remnants 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.

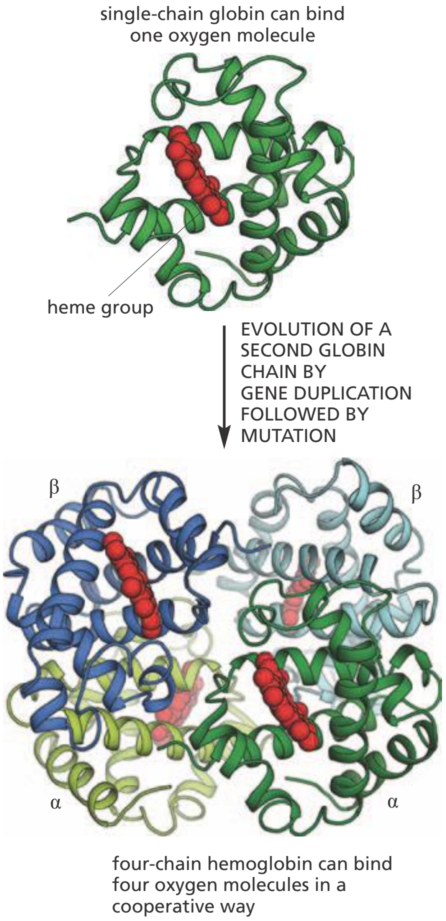
Many gene duplications are believed to be generated by *homologous recombination*. As discussed in Chapter 6, homologous recombination provides an important mechanism for mending a broken double helix; it allows an intact chromosome to be used as a template to repair a damaged sequence on its homolog. Homologous recombination normally takes place only after two long stretches of nearly identical DNA become paired, so that the information in the intact piece of DNA can be used to "restore" the sequence in the broken DNA. On rare occasions, however, a recombination event can occur between a pair of shorter 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**). Once a gene has been duplicated in this way, subsequent unequal crossovers can readily add extra copies to the duplicated set by the same mechanism. As a result, entire sets of closely related genes, arranged in series, are commonly found in genomes.

The Evolution of the Globin Gene Family Shows How Gene Duplication and Divergence Can Produce New Proteins

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 the present-day globin proteins indicate that all the globin genes must derive from a single ancestral gene.

The simplest globin protein has a polypeptide chain of about 150 amino acids, which 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. This efficiency is particularly important for large multicellular animals, which

Figure 9–9 An ancestral globin gene encoding a single-chain globin molecule is thought to have given 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- and two β-globin chains. Each chain has a 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

cannot rely on the simple diffusion of oxygen through the body to oxygenate their tissues adequately.

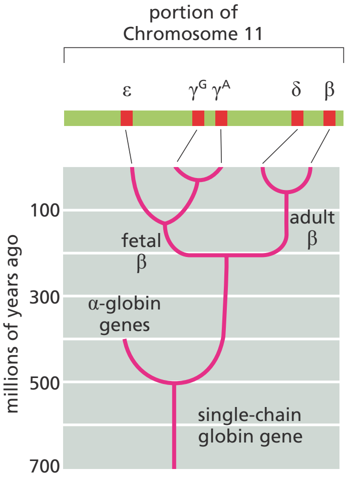
The α- and β-globin genes are the result of gene duplications that occurred early in vertebrate evolution. Genome analyses suggest that one of our ancient ancestors had a single globin gene. But about 500 million years ago, gene duplications followed by mutation are thought to have given rise to two 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 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 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 mutations that inactivate them. The existence of such *pseudogenes* makes it clear that, as might be expected, not every DNA duplication leads to a new functional gene. 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 are thought to have generated the globin gene family in humans. About 500 million years ago, an ancestral globin gene duplicated and gave rise to the β-globin gene family (including the five genes shown) and the related 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. And the δ-globin gene that arose during primate evolution encodes a minor β-globin form that's only made in adult primates. 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).



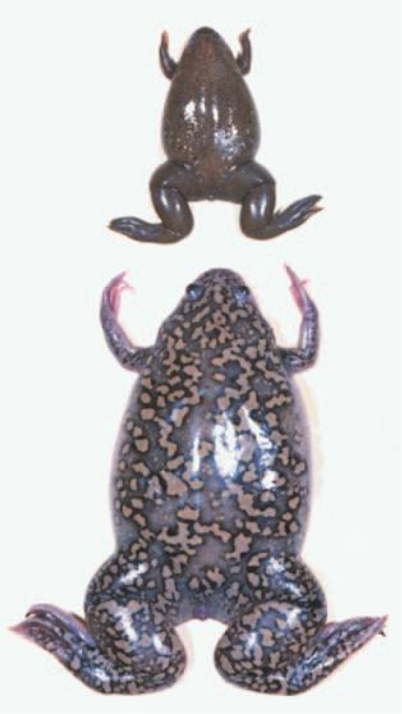


Figure 9–11 Different species of the frog *Xenopus* have different DNA contents. *X. tropicalis* (*above*) has an ordinary diploid genome with two sets of chromosomes in every somatic cell; the tetraploid *X. laevis* (*below*) has a duplicated genome containing twice as much DNA per cell. (Courtesy of Enrique Amaya.)

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 1–37), 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 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, comprises a set of closely similar species related to one another by repeated duplications or triplications of the whole genome (**Figure 9–11**). 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.

Novel Genes Can Be Created by Exon Shuffling

As we discussed in Chapter 4, many proteins are composed of a set 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 7–17 and 7–18). 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**.

This 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. If the introns in question are from two different genes, this recombination can generate a hybrid gene that includes complete exons from both. The presumed 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 all the proteins encoded by the human genome (approximately 21,000) 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

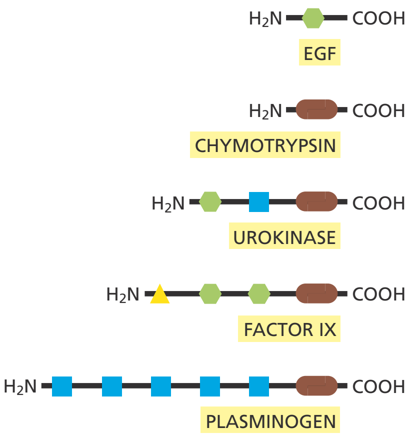


Figure 9–12 Exon shuffling during evolution can generate proteins with new combinations of protein domains. Each type of colored symbol represents a different protein domain. These different domains are thought to have been joined together by exon shuffling during evolution to create the modern-day human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

diversity of protein structures is generated from a quite small universal "list of parts," pieced together in different combinations.

The Evolution of Genomes Has Been Profoundly Influenced by the Movement of 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 region 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 region, for instance, will often have a striking effect on where and 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 are thought to have been particularly important in the evolution of the body plans of multicellular plants and animals.

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.

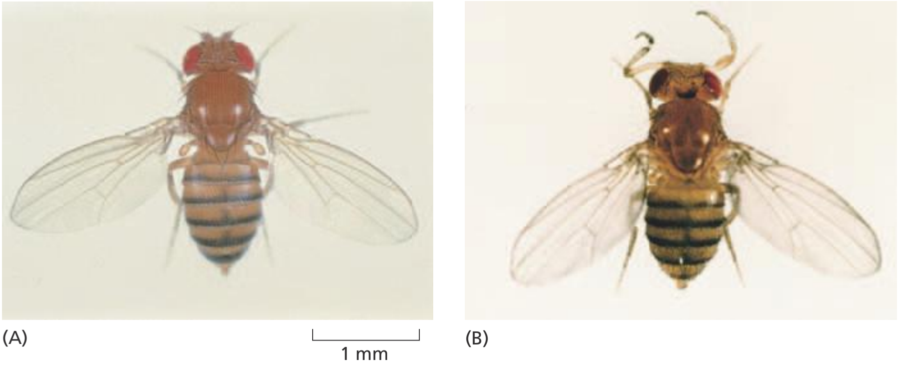


Figure 9–13 Mutation due to a mobile genetic element can induce dramatic alterations in the body plan of an organism. (A) A normal fruit fly (*Drosophila melanogaster*). (B) A mutant fly in which the antennae have been transformed into legs because of a mutation in a regulatory DNA sequence that causes 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, courtesy of E.B. Lewis; B, courtesy of Matthew Scott.)

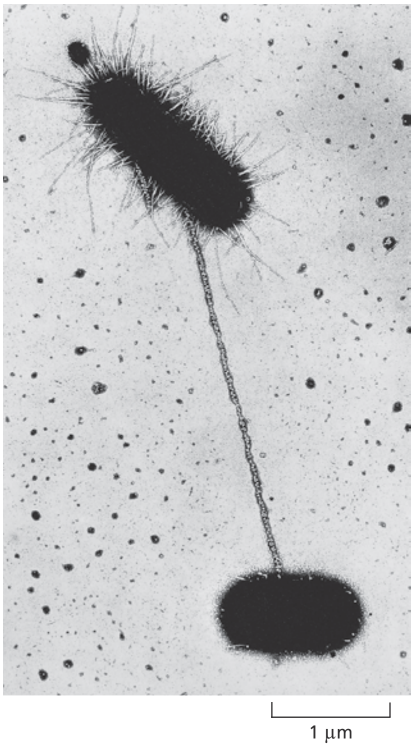


Figure 9–14 Bacterial cells can exchange DNA through conjugation. Conjugation begins when a donor cell (*top*) attaches to a recipient cell (*bottom*) by a fine appendage, called a sex pilus. DNA from the donor cell then moves through the pilus into the recipient cell. In this electron micrograph, the sex pilus has been labeled along its length by viruses that specifically bind to it and make the structure more visible. Conjugation is one of several ways in which bacteria carry out horizontal gene transfer. (Courtesy of Charles C. Brinton Jr. and Judith Carnahan.)

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 9–14** 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.

QUESTION 9–2

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

RECONSTRUCTING LIFE'S FAMILY TREE

We have seen how genomes can change over evolutionary time. The nucleotide sequences of present-day genomes provide a record of those changes that conferred biological success. By comparing the genomes of a variety of living organisms, we can thus begin to decipher our evolutionary history, seeing 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*, the fruit fly *Drosophila melanogaster*, and *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 were already 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 *selectively neutral* may or may not be passed on. And mutations that are deleterious will be lost. Through endless repetition of such cycles of error and trial—of mutation and natural selection—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 proteins they encode 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 when we wish to trace family relationships among the most distantly related organisms in the tree of life.

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?

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 comparisons of nucleotide changes allow the construction of a **phylogenetic tree**, a diagram that depicts the evolutionary relationships among a group of organisms. Figure 9–15 presents a phylogenetic tree that lays out the relationships among higher primates.

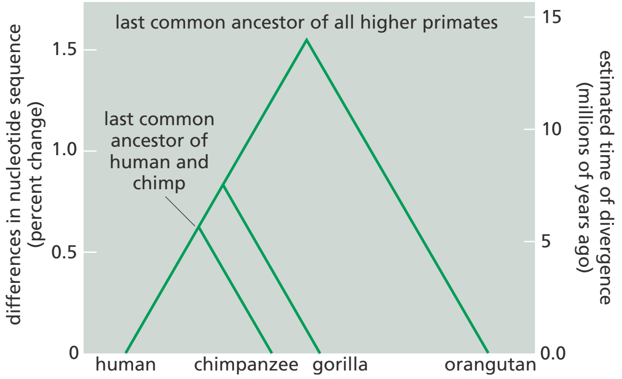


Figure 9–15 Phylogenetic trees display the relationships among modern life-forms. In this family tree of higher primates, humans fall closer to chimpanzees than to gorillas or orangutans, as there are fewer differences between human and chimp DNA sequences than there are between those of humans and gorillas, 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, the divergence between any two species will be twice as much as the amount of change that takes place between each of the species and their last common ancestor. For example, although humans and orangutans 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%. Although this phylogenetic tree is based solely on nucleotide sequences, the estimated dates of divergence, shown on the *right* side of the graph, derive from data obtained from the fossil record. (Modified from F.C. Chen and W.H. Li, *Am. J. Hum. Genet.* 68:444–456, 2001. With permission from Elsevier.)

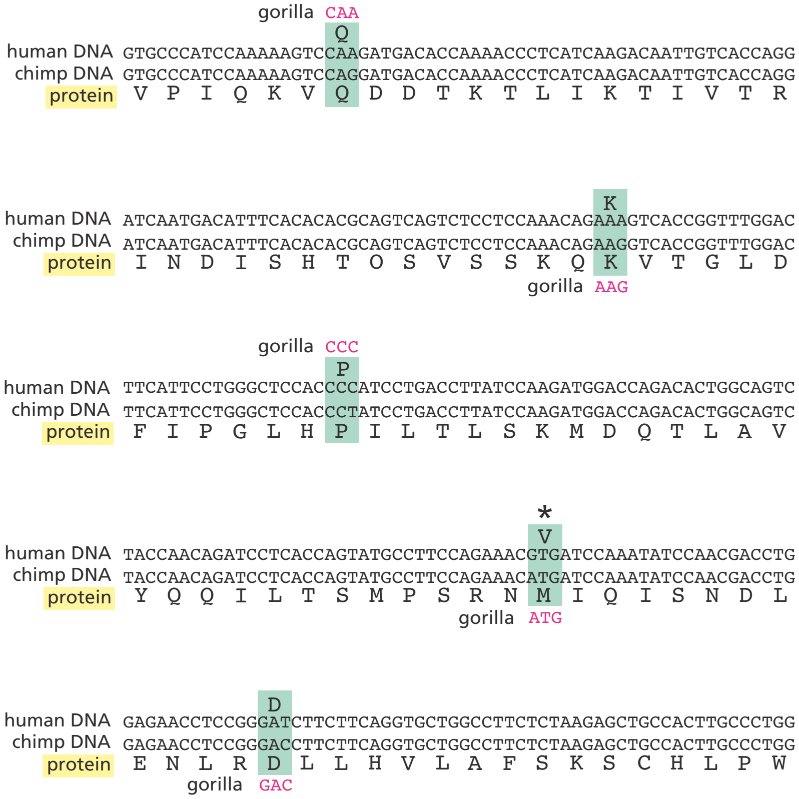

gorilla CAA
Q
human DNA GTGCCCATCCAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGG
chimp DNA GTGCCCATCCAAAAAGTCCAGGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGG
protein V P I Q K V Q D D T K T L I K T I V T R
K
human DNA ATCAATGACATTTCACACACGCAGTCAGTCTCCTCCAAACAGAAAGTCACCGGTTTGGAC
chimp DNA ATCAATGACATTTCACACACGCAGTCAGTCTCCTCCAAACAGAAGGTCACCGGTTTGGAC
protein I N D I S H T O 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 9–16 Ancestral gene sequences can be reconstructed by comparing closely related present-day species. Shown here, in five contiguous segments of DNA, are nucleotide sequences from the protein-coding region of the leptin gene from humans and chimpanzees. Leptin is a hormone that regulates food intake and energy utilization. As indicated by the codons boxed in *green*, only 5 out of a total 441 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." This strategy is based on the relationship shown in Figure 9–15: differences between humans and chimpanzees reflect relatively recent events in evolutionary history, and the gorilla sequence reveals the most likely precursor sequence. For convenience, only the first 300 nucleotides of the leptin-coding sequences are shown. The last 141 nucleotides are identical between humans and chimpanzees.

It is clear from this figure that 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. 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 sequence of genes that must have been present in the now-extinct, common ancestor of the two species (**Figure 9–16**).

Even the rearrangement of genomes by recombination, 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 sequence, are distributed differently, as they have had more time to proliferate and move around the two genomes since these species diverged (**Figure 9–17**).

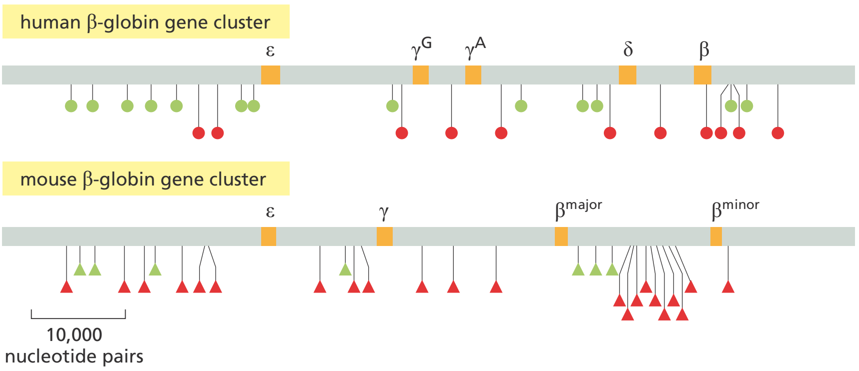
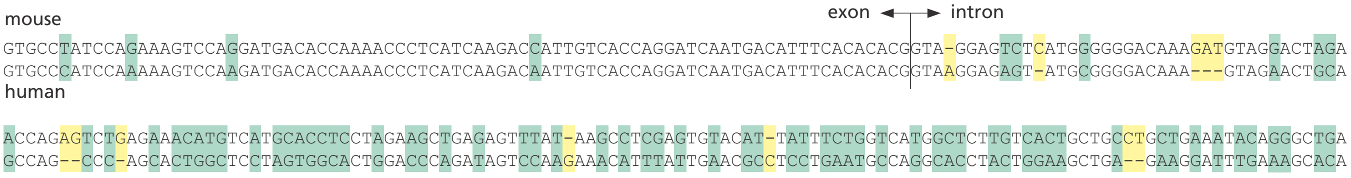


Figure 9–17 The positions of mobile genetic elements in the human and mouse genomes reflect the long evolutionary time separating the two species. This stretch of human Chromosome 11 (introduced in Figure 9–10) contains five functional β-globin-like genes (*orange*); the comparable region from the mouse 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 natural selection, which most likely eliminated any insertion that compromised gene function. (The mobile genetic element that falls inside the human β-globin gene (*far right*) is actually located within an intron.) (Courtesy of Ross Hardison and Webb Miller.)

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 in the past 75 million years: it is estimated that about 180 such "break-and-join" events have dramatically altered chromosome structure. For example, in humans most centromeres lie near the middle of the chromosome, whereas those of mouse are located at the chromosome ends.

In spite of this significant degree of genetic shuffling, one can nevertheless still recognize many blocks of **conserved synteny**, regions where 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 75 million years since humans and mice diverged from their common ancestor, about 50% of the nucleotides have changed. Against this background of dissimilarity, however, one can now begin to see very clearly the regions where changes are not tolerated, so that the human and mouse sequences have remained nearly the same (**Figure 9–18**). Here, the 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 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 9–19**). Surprisingly, only about one-third of these sequences code for proteins. Some of the conserved noncoding sequences correspond



```
GTGCCTATCCAGAAAGTCCAGGATGACACCAAAACCCTCATCAAGACCATTGTCACCAGGATCAATGACATTTCACACACGGTA-GGAGTCTCATGGGGGACAAAGATGTAGGACTAGA
GTGCCCATCCAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAGGATCAATGACATTTCACACACGGTAAGGAGAGT-ATGCGGGGACAAA---GTAGAACTGCA

ACCAGAGTCTGAGAAACATGTCATGCACCTCCTAGAAGCTGAGAGTTTAT-AAGCCTCGAGTGTACAT-TATTTCTGGTCATGGCTCTTGTCACTGCTGCCTGCTGAAATACAGGGCTGA
GCCAG--CCC-AGCACTGGCTCCTAGTGGCACTGGACCCAGATAGTCCAAGAAACATTTATTGAACGCCTCCTGAATGCCAGGCACCTACTGGAAGCTGA--GAAGGATTTGAAAGCACA
```

Figure 9–18 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.

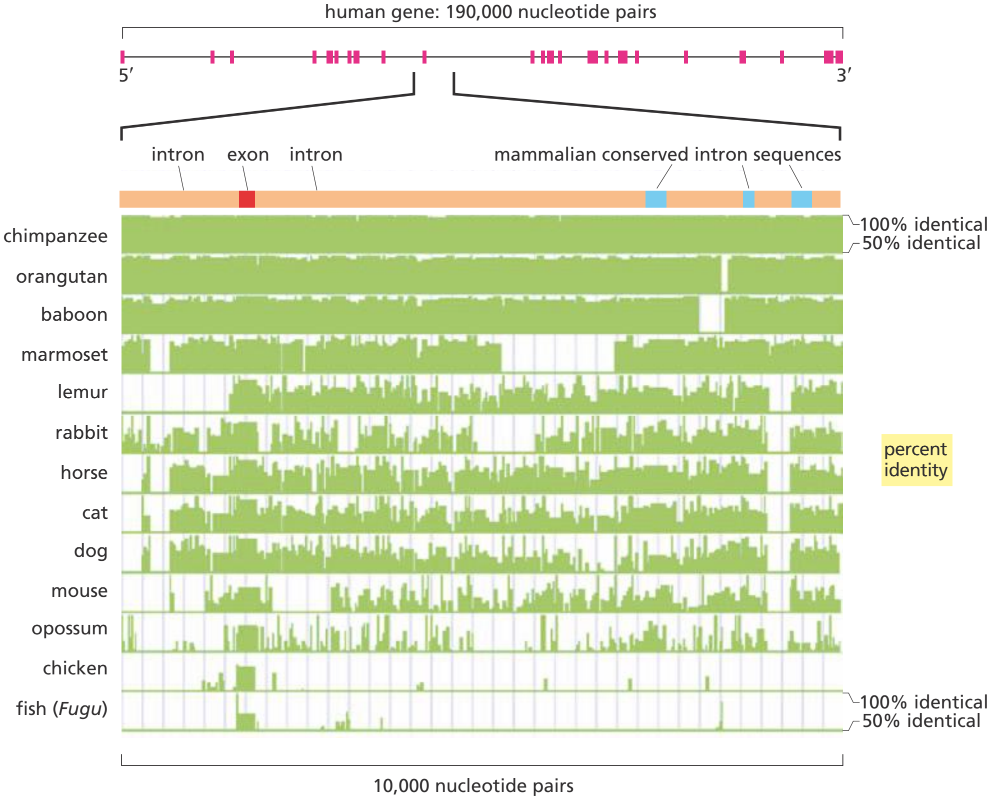


Figure 9–19 Comparison of nucleotide sequences from many different vertebrates reveals regions of high conservation. The nucleotide sequence examined in this diagram is a small segment of the human gene for a plasma membrane transporter protein. Exons in the complete gene (*top*) and in the expanded region of the gene are indicated in *red*. Three blocks of intron sequence that are conserved in mammals are shown in *blue*. In the lower part of the figure, the expanded human DNA sequence 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.)

to regulatory DNA, whereas others are transcribed to produce RNA molecules that are not translated into protein but serve regulatory functions (discussed in Chapter 8). The functions of the majority 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 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 segments of regulatory DNA. 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.

But 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 9–20**), whose tiny genome is one-tenth the size of mammalian genomes, largely because of the small size of

Figure 9–20 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.)

its introns. *Fugu* introns, as well as other noncoding segments in the animal's genome, lack the repetitive DNA that makes up a large portion of most mammalian genomes. Nonetheless, the positions of most *Fugu* introns are perfectly conserved when compared with their positions in mammalian genomes (**Figure 9–21**). 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.

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 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 (see Figure 7–32). Because the process of translation is fundamental to all living cells, this

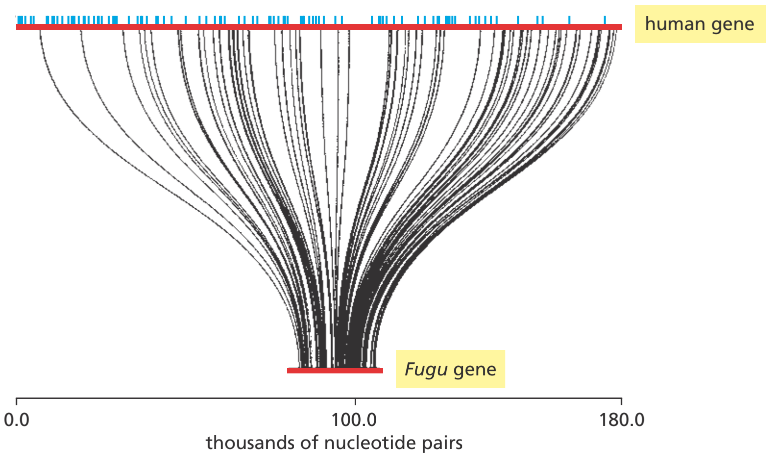


Figure 9–21 The positions of introns and exons are conserved between *Fugu* and humans. Comparison of the nucleotide sequences of the genes that encode the huntingtin protein in human and in *Fugu*. Both genes (*red*) contain 67 short exons, which align in 1:1 correspondence with one another; the corresponding exons are connected by the curved black lines. The human gene is 7.5 times larger than the *Fugu* gene (180,000 versus 24,000 nucleotide pairs), due entirely to the larger introns in the human sequence. The larger size of the human introns is due in part to mobile genetic elements, whose positions are represented by the *blue* vertical lines. These elements are absent in *Fugu*. In humans, mutation of this gene causes Huntington's disease, an inherited neurodegenerative disorder of the brain. (Adapted from S. Baxendale et al., *Nat. Genet.* 10:67–76, 1995. With permission from Macmillan Publishers Ltd.)

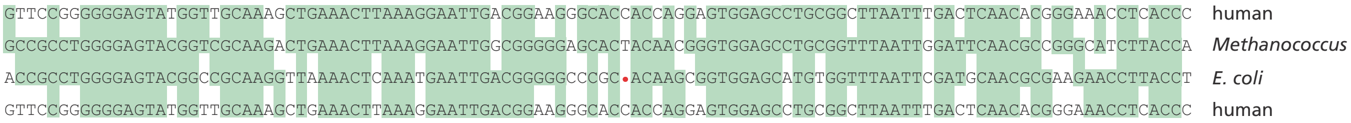

GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAACCTCACCC human
GCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTTAAAGGAATTGGCGGGGGAGCACTACAACGGGTGGAGCCTGCGGTTTAATTGGATTCAACGCCGGGCATCTTACCA Methanococcus
ACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGC•ACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCT 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 subunit rRNA (see Figure 7–32) is shown. Corresponding segments of nucleotide sequence from this gene in three distantly related species (*Methanococcus jannaschii*, an archaeon; *Escherichia coli*, a bacterium; and *Homo sapiens*, a eukaryote) are aligned in parallel. Sites where the nucleotides are identical between 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.

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 9–15), 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 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.

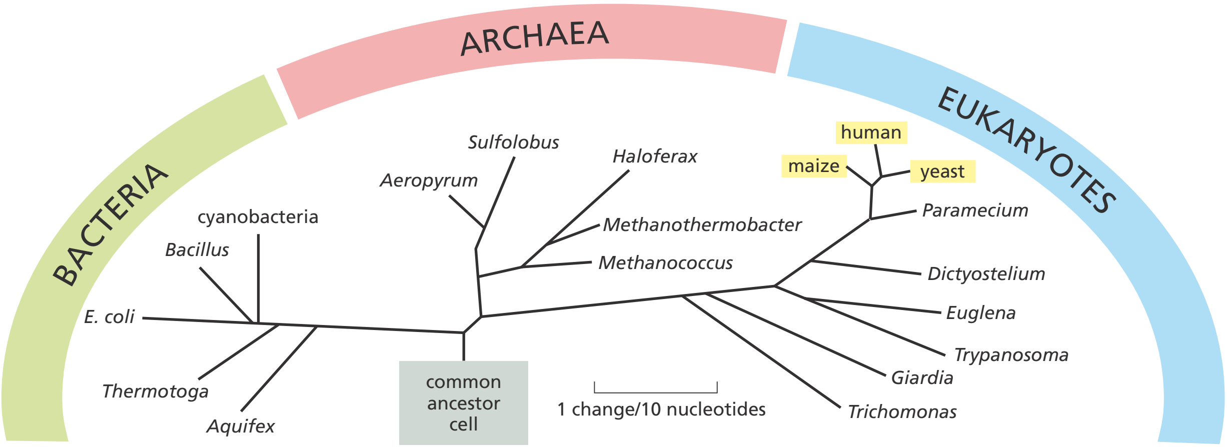


Figure 9–23 The tree of life has three major divisions. Each branch on the tree is labeled with the name of a representative member of that group, and the length of each branch corresponds to the degree of difference in the DNA sequences that encode their small subunit rRNAs (see Figure 9–22). Note that 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.

TRANSPOSONS 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 the 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 behavior, these diminutive genetic parasites have major implications for the evolution of species and for human health.

Mobile genetic elements, known informally as jumping genes, 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 DNA sequence, 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 briefly discuss mobile genetic elements as well as 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 below. 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 their DNA before inserting into the new chromosomal site, leaving the original copy 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 each mobile genetic element. Many mobile genetic elements also carry additional genes: some

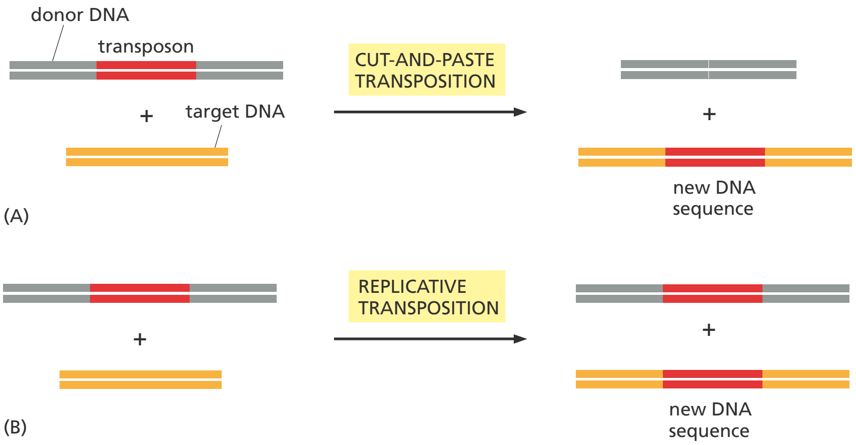


Figure 9–24 The most common mobile genetic elements in bacteria, DNA-only transposons, move by two types of mechanism. (A) In cut-and-paste transposition, the element is cut out of the donor DNA and inserted into the target DNA, leaving behind a broken donor DNA molecule, which is subsequently repaired. (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.

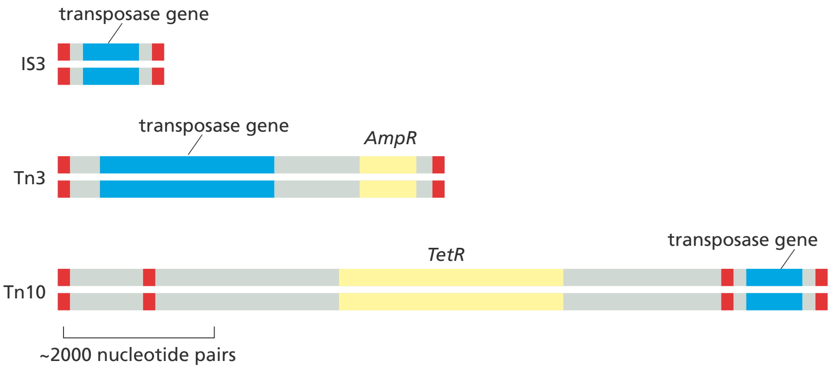


Figure 9–25 Transposons contain the components they need for transposition. Shown here are three types of bacterial DNA-only transposons. Each carries a gene that encodes a transposase (*blue*)—the enzyme that catalyzes the element's movement—as well as DNA sequences (*red*) that are recognized by that transposase. Some transposons carry additional genes (*yellow*) that encode enzymes that inactivate antibiotics such as ampicillin (*AmpR*) and tetracycline (*TetR*). 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, for example, carry antibiotic-resistance genes, which have contributed greatly to the widespread dissemination of antibiotic resistance in bacterial populations (**Figure 9–25**).

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.

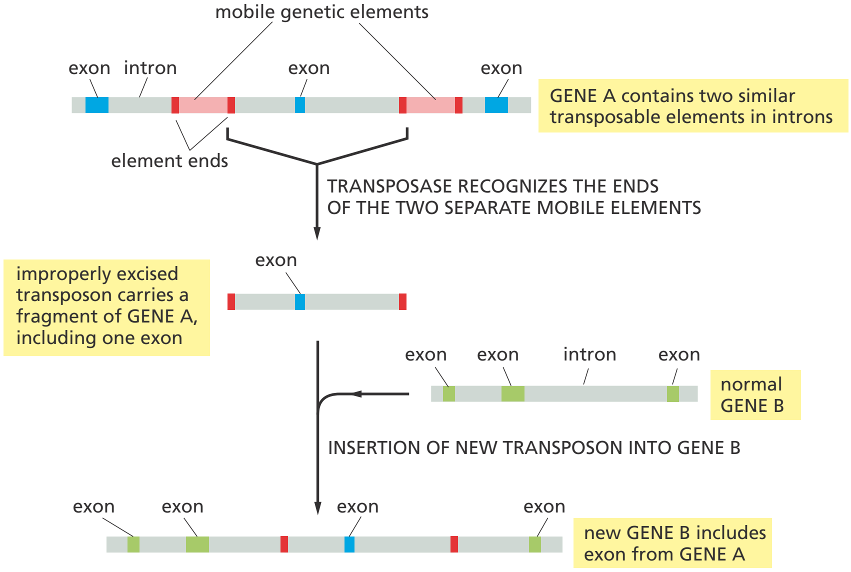

mobile genetic elements
exon
intron
exon
exon
GENE A contains two similar transposable elements in introns
element ends
TRANSPOSASE RECOGNIZES THE ENDS OF THE TWO SEPARATE MOBILE ELEMENTS
improperly excised transposon carries a fragment of GENE A, including one exon
exon
exon
exon
intron
exon
normal 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).

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**).

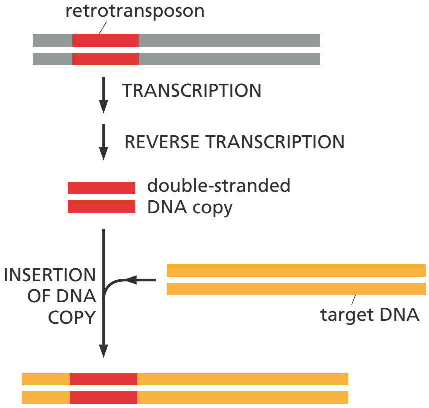


Figure 9–27 Retrotransposons move via an RNA intermediate. These transposable elements are first transcribed into an RNA intermediate. Next, a double-stranded DNA copy of this RNA is synthesized by the enzyme reverse transcriptase. This DNA copy is then inserted into the target location, which can be 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.

L1 elements constitute about 15% of the human genome. Although most copies have been immobilized by the accumulation of deleterious mutations, a few still retain the ability to transpose. Their movement can sometimes precipitate disease: for example, about 40 years ago, movement of an *L1* element into the gene that encodes Factor VIII—a protein essential for proper blood clotting—caused hemophilia in an individual 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 9–17). 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 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.

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 encode the

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."

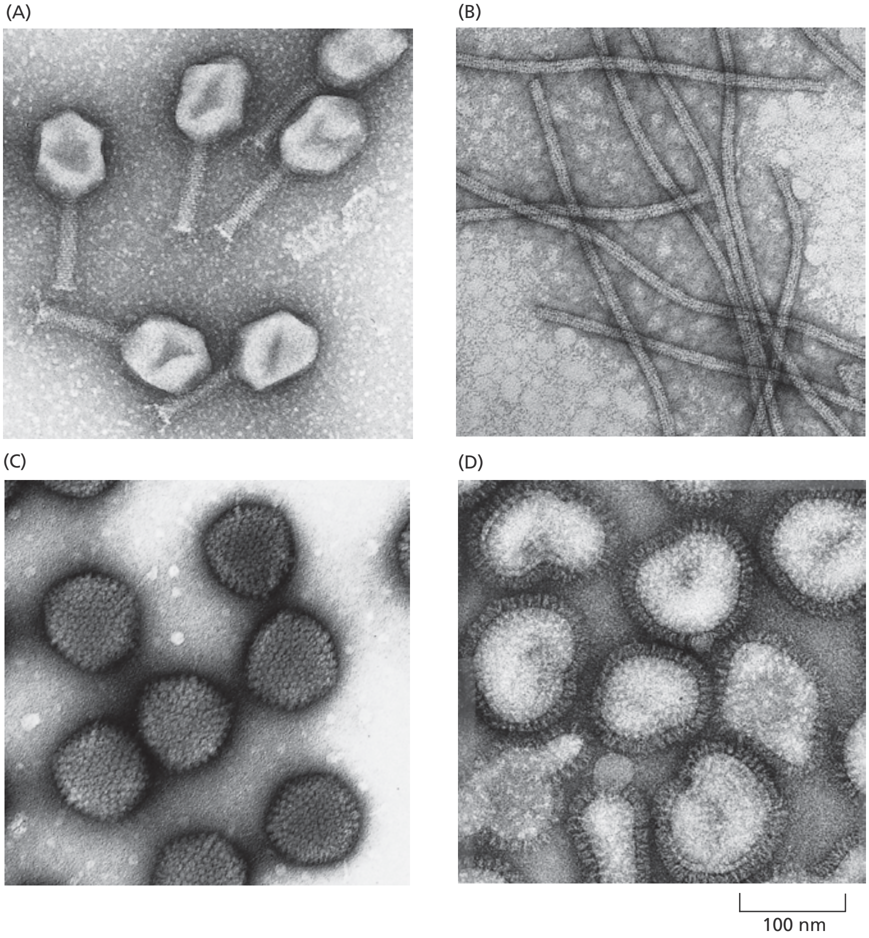


Figure 9–28 Viruses come in different shapes and sizes. These electron micrographs of virus particles are all shown at the same scale. (A) T4 bacteriophage, a large DNA-containing virus that infects *E. coli* cells. The DNA is stored in the viral head and is injected into the bacterium through the cylindrical tail. (B) Potato virus X, a tubelike plant virus that contains an RNA genome. (C) Adenovirus, a DNA-containing animal virus that can infect human cells. (D) Influenza virus, a large RNA-containing animal virus whose protein coat is further enclosed in a lipid-bilayer-based envelope. The spikes protruding from the envelope are viral coat proteins embedded in the lipid bilayer. (A, courtesy of James R. Paulson; B, courtesy of Graham Hills; C, courtesy of Mei Lie Wong; D, courtesy of R.C. Williams and H.W. Fisher.)

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**). The viral genome will typically encode both viral coat proteins and proteins that help them to coopt the host enzymes needed to replicate their genetic material.

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 life 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 DNA information to RNA. Retroviruses are thought to have derived from a retrotransposon that long ago acquired additional genes encoding 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. The complete life cycle of a retrovirus is shown in **Figure 9–30**.

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

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

(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 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.

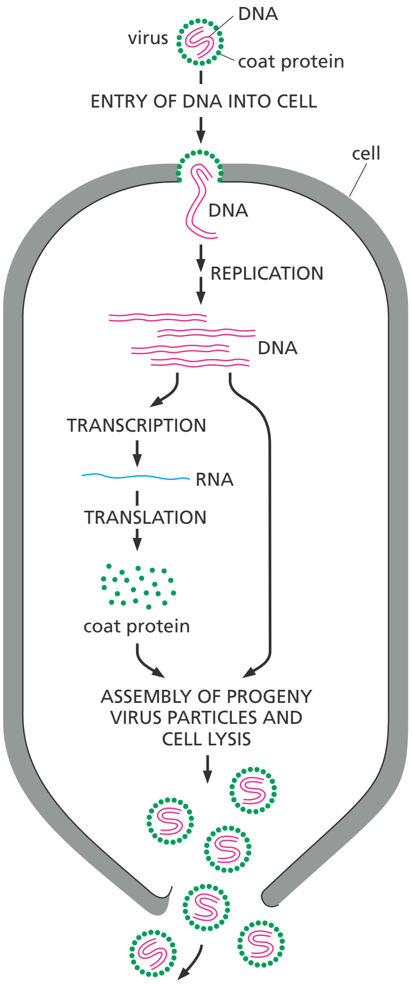


Figure 9–29 Viruses commandeer the host cell's molecular machinery to reproduce. The hypothetical simple virus illustrated here consists of a small double-stranded DNA molecule that encodes just a single type of viral coat protein. To reproduce, the viral genome must first enter a host cell, where it 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.

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

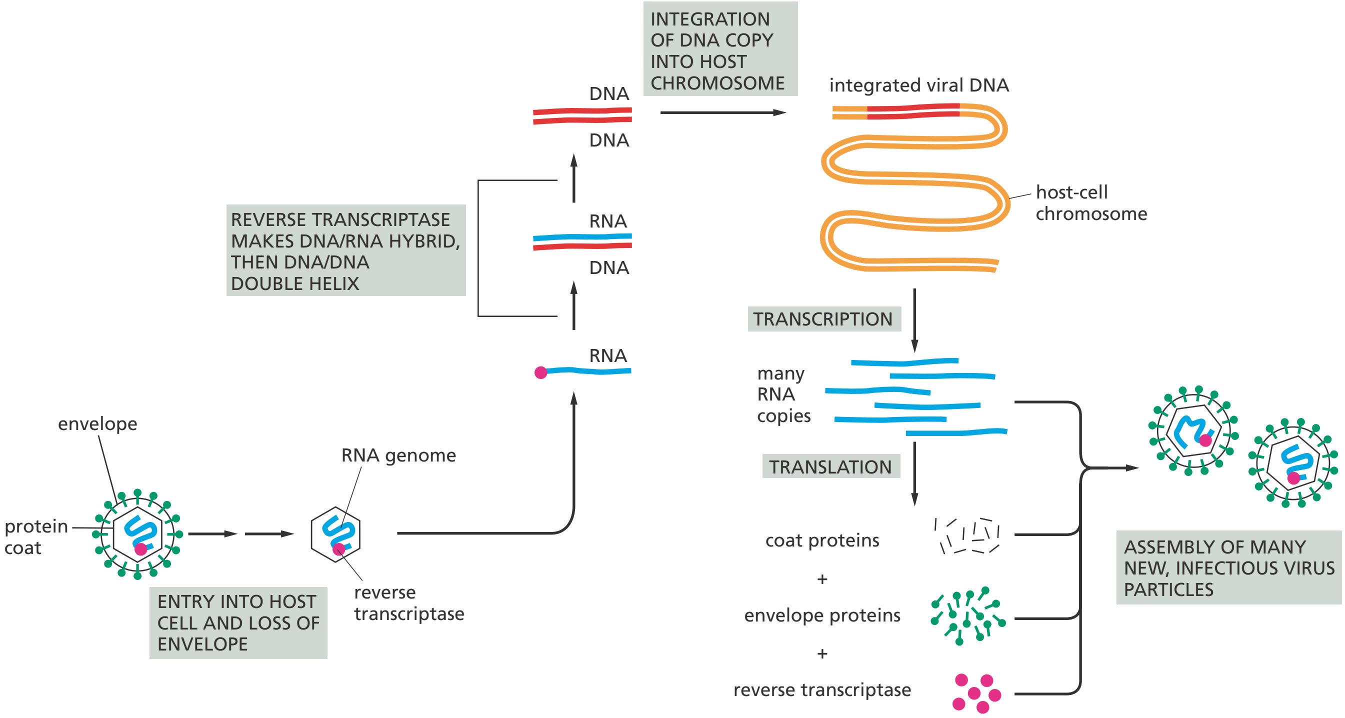

INTEGRATION OF DNA COPY INTO HOST CHROMOSOME
DNA
DNA
integrated viral DNA
host-cell chromosome
REVERSE TRANSCRIPTASE MAKES DNA/RNA HYBRID, THEN DNA/DNA DOUBLE HELIX
RNA
DNA
RNA
TRANSCRIPTION
many RNA copies
TRANSLATION
envelope
RNA genome
protein coat
ENTRY INTO HOST CELL AND LOSS OF ENVELOPE
reverse transcriptase
coat proteins
+
envelope proteins
+
reverse transcriptase
ASSEMBLY OF MANY NEW, INFECTIOUS VIRUS PARTICLES

Figure 9–30 The life cycle of 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-based 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.

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 DNA.

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. Spectacular improvements in sequencing technologies, 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 about 100,000-fold 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 promises to tell us what makes us human, and what makes each of us unique.

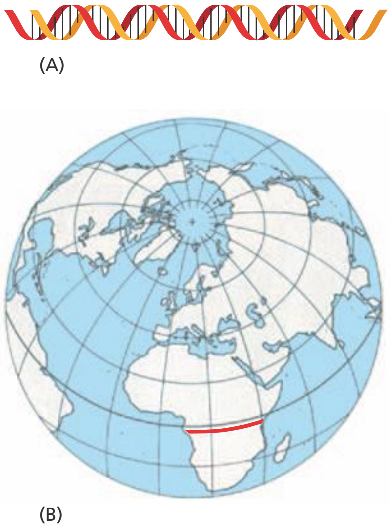


Figure 9–31 The 3 billion nucleotide pairs of the human genome contain a vast amount of information, including clues about our origins. If each nucleotide pair is drawn to span 1 mm, as shown in (A), the human genome would extend 3200 km (approximately 2000 miles)—far enough to stretch across central 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 30 m, but the coding sequences (exons) in this gene would add up to only just over a meter; the rest would be introns.

Although it will take decades 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 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 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 (see **How We Know**, pp. 316–317). Although the exact count is still being refined, current estimates place the number of human

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?

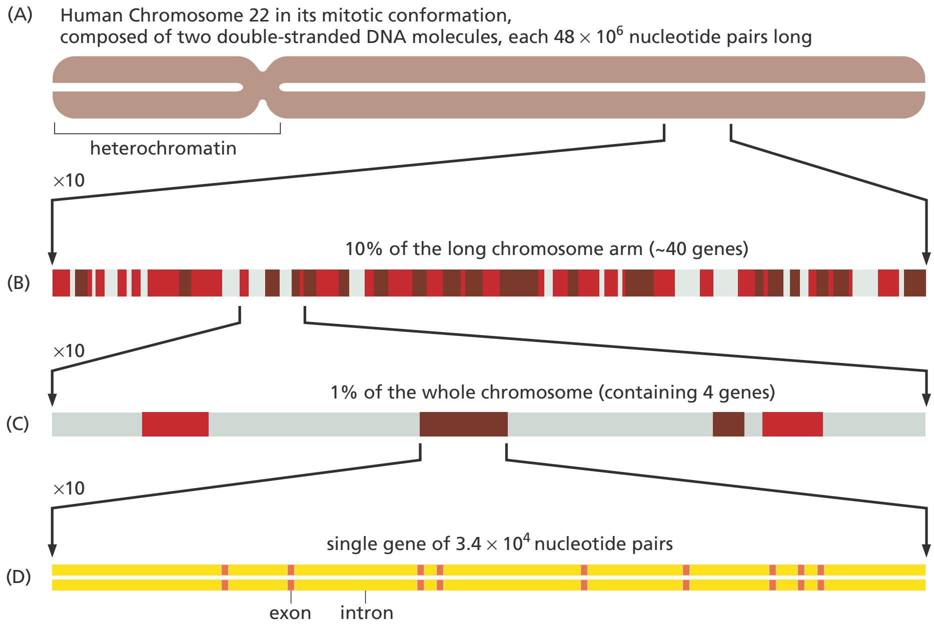

(A)
Human Chromosome 22 in its mitotic conformation, composed of two double-stranded DNA molecules, each 48 × 10^6 nucleotide pairs long
heterochromatin
×10
10% of the long chromosome arm (~40 genes)
(B)
×10
1% of the whole chromosome (containing 4 genes)
(C)
×10
single gene of 3.4 × 10^4 nucleotide pairs
(D)
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 entire human genome. Most of the left arm of Chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin), 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 (*orange*) 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. With permission from Macmillan Publishers Ltd.)

TABLE 9–2 SOME VITAL STATISTICS FOR THE HUMAN GENOME

DNA length	3.2×10^9 nucleotide pairs*
Number of protein-coding genes	approximately 21,000
Number of non-protein-coding genes**	approximately 9000
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****	3.5%
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, and conserved regions of unknown function.

protein-coding genes at about 21,000. Perhaps another 9000 genes encode functional RNAs that are not translated into proteins. The estimate of 30,000 total genes brings us much closer to the gene numbers for simpler multicellular animals—for example, 13,000 for *Drosophila*, 21,000 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

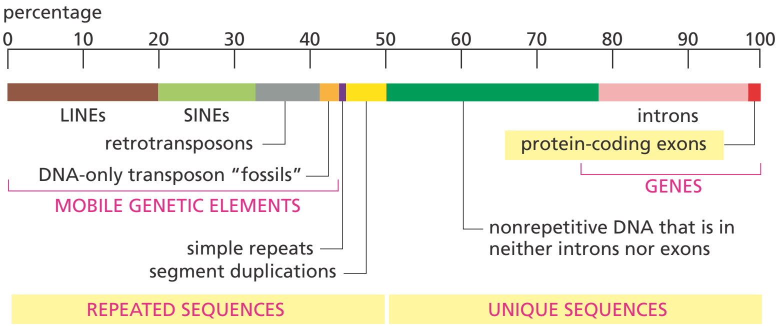
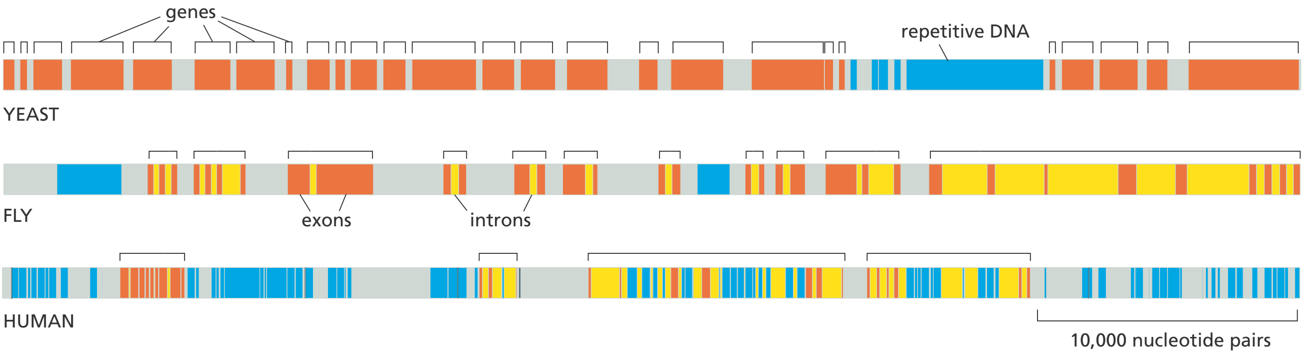


Figure 9–33 The bulk of the human genome is made of repetitive nucleotide sequences and other noncoding DNA. The LINEs (which include *L1*), SINEs (short interspersed nuclear element, which include *Alu*), retrotransposons, and DNA-only transposons are mobile genetic elements that have multiplied in our genome by replicating themselves and inserting the new copies in different positions. 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. The unique sequences that are not part of any introns or exons (*dark green*) include gene regulatory sequences, sequences that code for functional RNA, and sequences whose functions are not known. The most highly repeated blocks of DNA in heterochromatin have not yet been completely sequenced; therefore about 10% of human DNA sequences are not represented in this diagram. (Data courtesy of E.H. Margulies.)



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**).

Figure 9–34 Genes are sparsely distributed in the human genome. Compared to these other eukaryotic genomes, the human genome is less gene-dense. Shown here are DNA segments about 50,000 nucleotide pairs in length from 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 "spacer" DNA in *gray*. The genes of yeast and flies are generally more compact, with fewer introns, than the genes of humans.

Although exons and their associated gene regulatory sequences comprise less than 2% of the human genome, comparative studies indicate that about 5% of the human genome is highly conserved when compared with other mammalian genomes (see Figure 9–19). An additional 4% of the genome shows reduced variation in the human population, as determined by comparing the DNA sequence of thousands of individuals. Taken together, this conservation suggests that about 9% of the human genome contains sequences that are likely to be functionally important—but we do not yet know the function of much of this DNA.

Accelerated Changes in Conserved Genome Sequences Help Reveal What Makes Us Human

When the chimpanzee genome sequence became available in 2005, scientists began searching for DNA sequence changes that might account for the striking differences between us and them (**Figure 9–35**). With about 3 billion nucleotide pairs to compare between the two species, the task is daunting. But the search is made much easier by confining the comparison to those sequences that are highly conserved across multiple mammalian species (see Figure 9–19). These conserved sequences represent parts of the genome that are most likely to be functionally important—and are thus areas of particular interest when we search for genetic changes that make humans different from our mammalian cousins.

Although these sequences are conserved, they are not identical: when the version from one mammal is compared with that of another, they are typically found to have drifted apart by a small amount, which corresponds to the time elapsed since the species diverged during evolution. In a small proportion of cases, however, the sequences show signs of a sudden evolutionary spurt. For example, some DNA sequences that have been highly conserved in most mammalian species are found to have changed exceptionally fast during the last six million years of human evolution. Such *human accelerated regions* are thought to reflect functions that have been especially important in making us the unique animal that we are.

One study identified about 50 such sites—one-quarter of which were located near genes associated with brain development. The sequence

Figure 9–35 DNA sequences that have changed rapidly in the past six million years may account for the differences between chimps and humans. Many of these changes may have affected the way human brains develop. Shown here is anthropologist Jane Goodall with one of her chimpanzee subjects. (Courtesy of the Jane Goodall Institute of Canada.)

COUNTING GENES

How many genes does it take to make a human? It seems a natural thing to wonder. If 6000 genes can produce a yeast and 13,000 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 21,000 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 7–25). 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 9–36**).

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

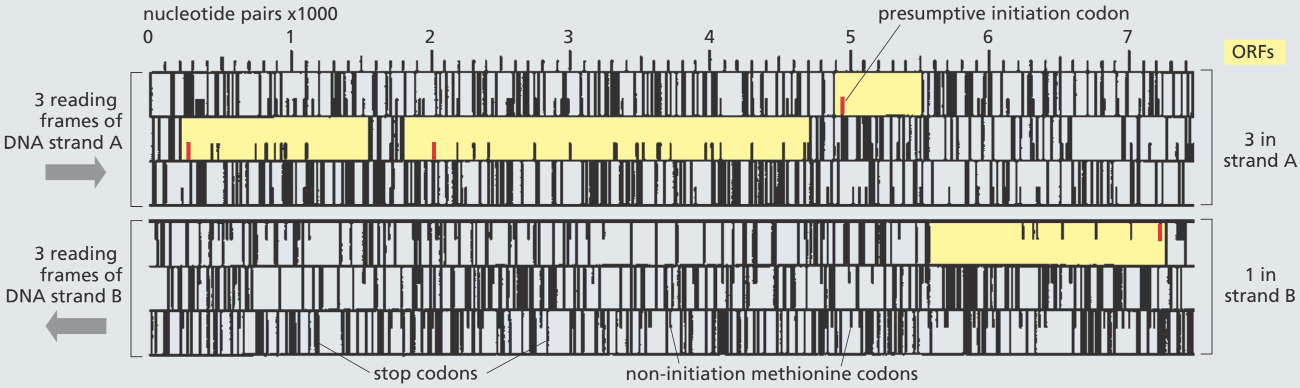

nucleotide pairs x1000
0
1
2
3
4
5
6
7
presumptive initiation codon
ORFs
3 reading frames of DNA strand A
3 in strand A
3 reading frames of DNA strand B
1 in strand B
stop codons
non-initiation methionine codons

Figure 9–36 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 7–26). 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 in the ORFs code for methionine in the protein.

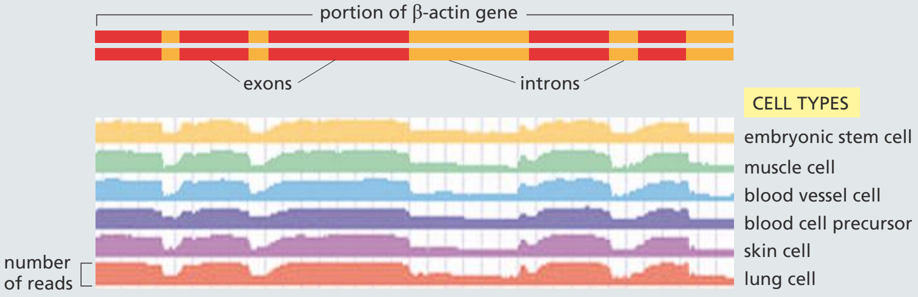


Figure 9–37 RNA sequencing can be used to identify protein-coding genes. Presented here is a set of data corresponding to RNAs produced from a segment of the gene for β-actin, which is depicted schematically at the top. Millions of RNA "sequence reads," each approximately 200 nucleotides long, were collected 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 7–19), gene regulatory 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 9–37**).

RNA Seq also offers a few additional benefits. First, the relative abundance of each sequence can be used to assess how highly its gene is expressed. Furthermore, the approach also locates 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 number of human genes are now converging around 30,000. 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.

exhibiting the most rapid change (18 changes between human and chimp, compared with only two changes between chimp and chicken) was examined further and found to encode a short, non-protein-coding RNA that is produced in the human cerebral cortex at a critical time during brain development. Although the function of this RNA is not yet known, this exciting finding is stimulating further studies that might help shed light on features of the human brain that distinguish us from chimps.

Similar studies have identified genes that may have played a role in even more recent human evolution. In 2010, investigators completed their analysis of the first Neanderthal genome. Our closest evolutionary relative, 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 five people from different parts of the world, the 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, and the shape of the skeleton, particularly the rib cage and head—all features thought to differ between modern humans and our extinct cousins.

Remarkably, these studies also revealed that some modern humans—those that hail from Europe and Asia—share from 1 to 4 percent of their genomes with Neanderthals. This genetic overlap suggests that our ancestors may have mated with Neanderthals—before outcompeting or actively exterminating them—on the way out of Africa, a relationship that left a permanent mark in the human genome.

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%. That might seem an insignificant degree of variation, but considering the size of the human genome, it amounts to some 3 million genetic differences per genome between 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 100,000 years ago, when the human population was still small. This means that a great deal of the genetic variation in present-day humans was inherited from our early human ancestors.

Most of the genetic variation in the human genome takes the form of single base changes 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 more than 1% of the population has a G-C nucleotide pair, for example, while another has 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.

Another important source of variation inherited from our ancestors involves the duplication and deletion of large segments of DNA. When the genome of any person is compared with a standard reference genome, one observes roughly 100 instances in which a relatively long stretch of DNA has been gained or lost. Some of these **copy-number variations** (**CNVs**) are very common, whereas others are present in only a small minority of people. From an initial sampling, nearly half of

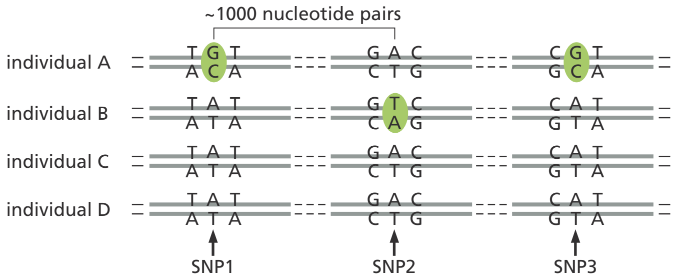


Figure 9–38 **Single-nucleotide polymorphisms (SNPs) are points in the genome that differ by a single nucleotide pair between one portion of the population and another.** By convention, to count as a polymorphism, a genetic difference must be present in at least 1% of the total population of the species.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

these segments contain known genes and can affect one's susceptibility to certain diseases. In retrospect, this type of structural variation is not surprising, given the extensive history of DNA addition and DNA loss in vertebrate genomes discussed earlier. Exactly how it contributes to our individuality, however, remains to be determined.

In addition to the SNPs and the CNVs that we inherited from our ancestors, humans also possess repetitive nucleotide sequences that are particularly prone to new mutations. CA repeats, for example, are ubiquitous in the human genome. Nucleotide sequences containing large numbers of CA repeats are often replicated inaccurately (imagine trying to copy a word that is nothing more than a string of CACACACAC...); hence, the precise length of such repeats can vary widely between individuals and can increase from one generation to the next. Because they show such exceptional variability, and because this variability has arisen so recently in human history, CA repeats, and others like them, make ideal markers for distinguishing the DNA of individual humans. For this reason, differences in the numbers of *short tandem repeats* at different positions in the genome are used to identify individuals by *DNA fingerprinting* in crime investigations, paternity suits, and other forensic applications (see Figure 10–18).

Most of the variations in the human genome sequence 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. It remains a major challenge to identify those genetic variations that are functionally important—a problem we return to in Chapter 19.

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

The finding that humans, chimps, and mice contain essentially the same protein-coding genes has raised a fundamental question: What makes these creatures so different from one another?

To a large extent, the instructions needed to produce a multicellular animal from a fertilized egg are provided by the regulatory DNA 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 ultimately dictates 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 gene regulatory sequences than in the proteins or functional RNAs those genes encode.

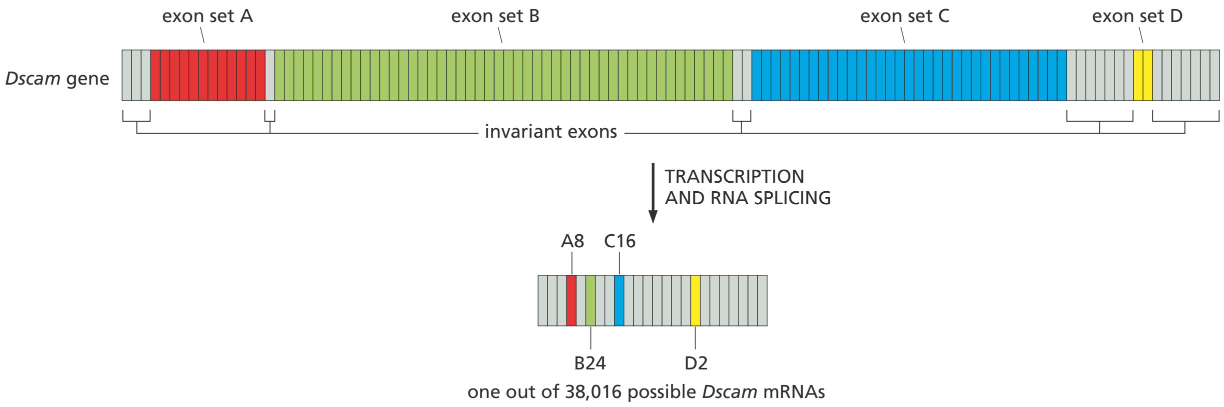


Figure 9–39 Alternative splicing of RNA transcripts can produce many distinct proteins. The *Drosophila* Dscam proteins are receptors that help nerve cells make their appropriate connections. The final mRNA transcript contains 24 exons, four of which (denoted A, B, C, and D) are present in the *Dscam* gene as arrays of alternative exons. Each mature mRNA contains 1 of 12 alternatives for exon A (*red*), 1 of 48 alternatives for exon B (*green*), 1 of 33 alternatives for exon C (*blue*), 1 of 2 alternatives for exon D (*yellow*), and all of the 19 invariant exons (*gray*). If all possible splicing combinations were used, 38,016 different proteins could in principle be produced from the *Dscam* gene. Only one of the many possible splicing patterns and the mature mRNA it produces is shown. (Adapted from D.L. Black, *Cell* 103:367–370, 2000. With permission from Elsevier.)

Although we have made great strides in recognizing many of these regulatory sequences amidst the excess of noncritical "spacer" DNA, we still do not know how to "read" these sequences so that we can predict exactly how they operate in cells to control development. For example, the same short stretch of regulatory DNA may be recognized by several different transcription regulators, so simply knowing its nucleotide sequence will not reveal which transcription regulator—or regulators—might bind to the sequence in a particular cell at a particular time or place. In addition, gene expression is controlled by complex combinations of proteins (see Figure 8–12), which further complicates our attempts to decipher when in development and in which type of cell any given gene will be expressed.

Even if we could predict when a particular protein-coding gene would be expressed, we would not necessarily be able to predict what protein that gene would produce. Recent studies suggest that more than 90% of human genes undergo alternative RNA splicing, which allows cells to produce a range of related but distinct proteins from a single gene (see Figure 7–22). RNA splicing is often regulated, so that one form of a protein is produced in one type of cell, while other forms are produced preferentially in other cell types. In one extreme example, from *Drosophila*, a single gene can produce thousands of different protein variants through alternative RNA splicing (**Figure 9–39**). Thus an organism can produce far more proteins than it has genes. We do not yet know enough about alternative splicing to predict exactly which human genes are subject to this process—and when, where, and how during development such regulation occurs. Nonetheless, it seems likely that these differences in alternative RNA splicing could help explain how animals with very similar protein-coding genes develop so differently.

Another part of the explanation may involve regulatory RNAs, such as the microRNAs and long noncoding RNAs discussed in Chapter 8. Thus for example, microRNAs have diverse roles in controlling gene expression, especially during development. They regulate as many as one-third of all human genes, for example, yet few of them have been studied in any detail—and new ones are still being found. And even less is known about the long noncoding RNAs.

The information that guides the countless decisions made by developing cells as they divide and specialize is all contained within the genome sequence of an organism. But we are only just beginning to learn the grammar and rules by which this genetic information orchestrates development. 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 the most 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 themselves.
- Some viruses have RNA instead of DNA as their genetic material. Retroviruses copy their RNA genomes into DNA before integrating 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.

KEY TERMS

Alu sequence
conserved synteny
copy-number variation
divergence
exon shuffling
gene duplication and divergence
gene family
germ cell
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 harmful to the host organism and therefore place it at an evolutionary disadvantage."

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/False) The majority of human DNA is unimportant junk. 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 embryo helps establish the basic body plan for humans (and for other animals). Why do you suppose that mobile genetic elements are so rare in this cluster? 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.

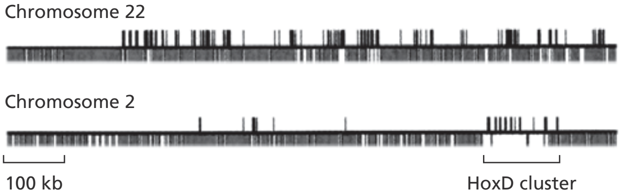


Figure Q9–10

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 to 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, a brilliant bioinformatician, has high regard for your intellect and industry. She 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

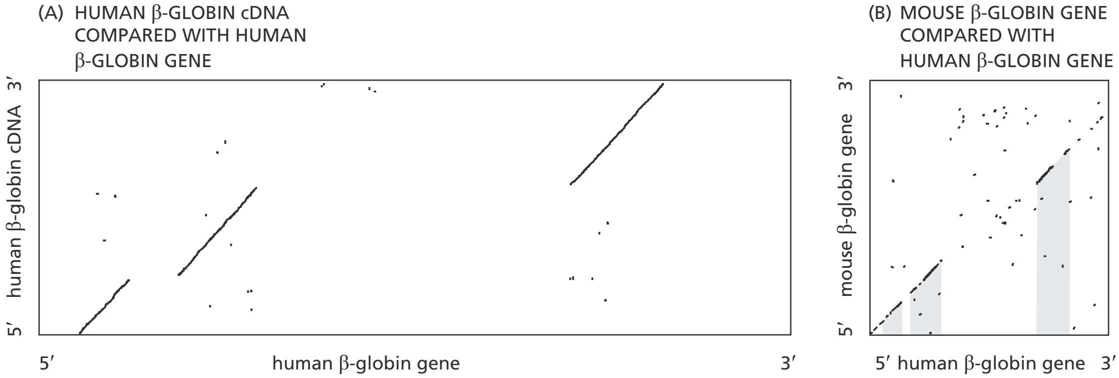


Figure Q9–11

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

The genetic code (see Figure 7–25) relates the nucleotide sequence of mRNA to the amino acid sequence of encoded proteins. Ever since the code was deciphered, some have claimed it must be a frozen accident—that is, the system randomly fell into place in some ancestral organism and was then perpetuated unchanged throughout evolution; others have argued that the code has been shaped by natural selection.

A striking feature of the genetic code is its inherent resistance to the effects of mutation. For example, a change in the third position of a codon often specifies the same amino acid or one with similar chemical properties. But is the natural code more resistant to mutation than other possible versions? The answer is an emphatic "Yes," as illustrated in **Figure Q9–15**. Only one in a million computer-generated "random" codes is more error-resistant than the natural genetic code.

Does the resistance to mutation of the actual genetic code argue in favor of its origin as a frozen accident or as a result of natural selection? Explain your reasoning.

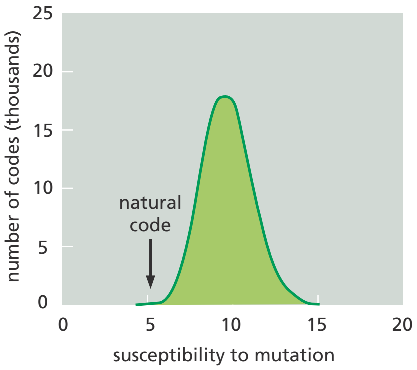


Figure Q9–15

QUESTION 9–16

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 9–17

Some genes 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 above. 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 (ATC → TTC, which gives isoleucine → phenylalanine, for example). Synonymous changes refer 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 7–25, there are many cases where several codons correspond to the same amino acid.)

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.

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 9–18

Plant hemoglobins were found initially in legumes, where they function in root nodules to lower the oxygen concentration, allowing the resident bacteria to fix nitrogen. These 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 genes have now been sequenced from a variety of organisms, and a phylogenetic tree of hemoglobins is shown in **Figure Q9–18**.

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

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?

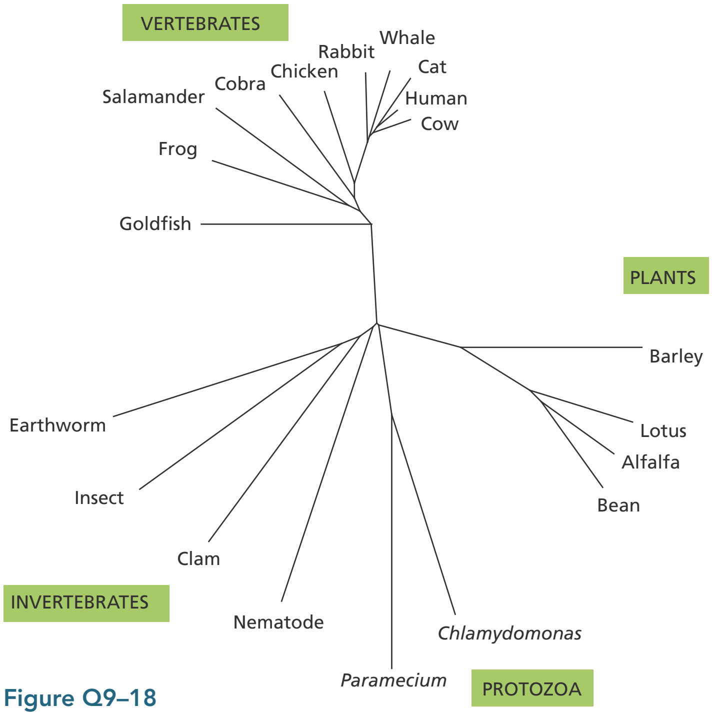


Figure Q9–18

QUESTION 9–19

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 picked at random will show about 1 difference of nucleotide sequence in every 1000 nucleotides. Suppose we are all descended from a single pair of ancestors—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 Adam and Eve 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 9–20

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?

CHAPTER **TEN** 10

Modern Recombinant DNA Technology

Since the turn of the century, biologists have amassed an unprecedented wealth of information on the genes that direct the development and behavior of living things. Thanks to advances in our ability to rapidly determine the nucleotide sequence of entire genomes, we now have access to the complete molecular blueprints for thousands of different organisms, from the platypus to the plague bacterium, and for thousands of different people from all over the world.

This information explosion would not have been possible without the technological revolution that enabled us to manipulate DNA molecules. In the early 1970s, it became possible, for the first time, to isolate a selected piece of DNA from the many millions of nucleotide pairs in a typical chromosome—and to replicate, sequence, and modify this DNA. These modified DNA molecules can then be introduced into another organism's genome, where they become a functional and heritable part of that organism's genetic instructions.

These technical breakthroughs—dubbed **recombinant DNA technology**, or *genetic engineering*—have had a dramatic impact on all aspects of cell biology. They have advanced our understanding of the organization and evolutionary history of complex eukaryotic genomes (as discussed in Chapter 9) and have led to the discovery of whole new classes of genes, RNAs, and proteins. They continue to generate new ways of determining the functions of genes and proteins in living organisms, and they provide an important set of tools for unraveling the mechanisms—still poorly understood—by which a complex organism can develop from a single fertilized egg.

Recombinant DNA technology has also had a profound influence on our understanding and treatment of disease: it is used, for example, to detect the mutations in human genes that are responsible for inherited disorders or that predispose us to a variety of common diseases, including cancer; it is used to produce an increasing number of pharmaceuticals, such as insulin for diabetics and blood-clotting proteins for hemophiliacs. But recombinant DNA technology also has applications outside the clinic. It allows, for example, forensic science to identify or acquit suspects in a crime. Even our laundry detergents contain heat-stable, stain-removing proteases, courtesy of DNA technology. Of all the discoveries described in this book, those that led to the development of recombinant DNA technology have the greatest impact on our everyday lives.

In this chapter, we present a brief overview of how we learned to manipulate DNA, identify genes, and produce many copies of any given nucleotide sequence in the laboratory. We discuss several approaches to exploring gene function, including new ways to monitor gene expression and to inactivate or modify genes in cells, animals, and plants. These methods—which are continuously being improved and made ever-more powerful—are not only revolutionizing the way we do science, they are transforming our understanding of cell biology and human disease. Indeed, they are responsible for a substantial portion of the information we present in this book.

QUESTION 10–1

DNA sequencing of your own two β-globin genes (one from each of your two Chromosome 11s) reveals a mutation in one of the genes. Given this information alone, should you worry about being a carrier of an inherited disease that could be passed on to your children? What other information would you like to have to assess your risk?

MANIPULATING AND ANALYZING DNA MOLECULES

Humans have been experimenting with DNA, albeit without realizing it, for millennia. The roses in our gardens, the corn on our plate, and the dogs in our yards are all the product of selective breeding that has taken place over many, many generations (**Figure 10–1**). But it wasn't until the development of recombinant DNA techniques in the 1970s that we could begin to engineer organisms with desired properties by directly tinkering with their genes.

Isolating and manipulating individual genes is not a trivial matter. Unlike a protein, a gene does not exist as a discrete entity in cells; it is a small part of a much larger DNA molecule. Even bacterial genomes, which are much less complex than the chromosomes of eukaryotes, are enormously long. The *E. coli* genome, for example, contains 4.6 million nucleotide pairs.

How, then, can a single gene be separated from a eukaryotic genome—which is considerably larger—so that it can be handled in the laboratory? The solution to this problem emerged, in large part, with the discovery of a class of bacterial enzymes known as *restriction nucleases*. These

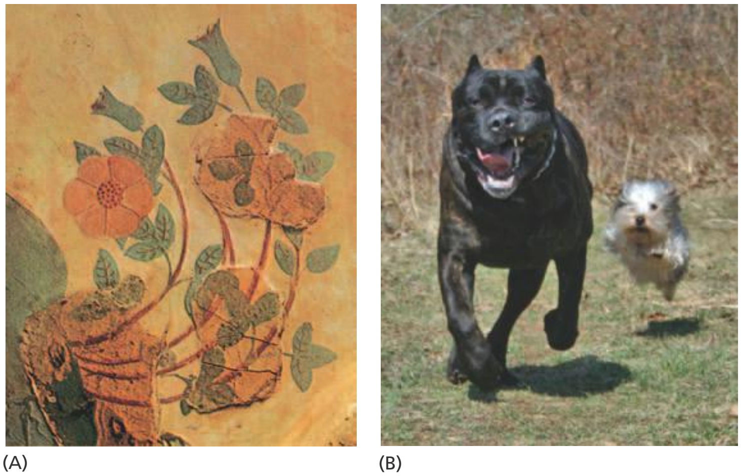


Figure 10–1 By breeding plants and animals, humans have been unwittingly experimenting with DNA for millennia. (A) The oldest known depiction of a rose in Western art, from the palace of Knossos in Crete, around 2000 BC. Modern roses are the result of centuries of breeding between such wild roses. (B) Dogs have been bred to exhibit a wide variety of characteristics, including different head shapes, coat colors, and of course size. All dogs, regardless of breed, belong to a single species that was domesticated from the gray wolf some 10,000 to 15,000 years ago. (B, from A.L. Shearin & E.A. Ostrander, *PLoS Biol.* 8:e1000310, 2010.)

enzymes cut double-stranded DNA at particular sequences. They can therefore be used to produce a reproducible set of specific DNA fragments from any genome. In this section, we describe how these enzymes work and how the DNA fragments they produce can be separated and visualized. We then discuss how these fragments can be probed to identify the ones that contain the DNA sequence of interest.

Restriction Nucleases Cut DNA Molecules at Specific Sites

Like many of the tools of recombinant DNA technology, restriction nucleases were discovered by researchers trying to understand an intriguing biological phenomenon. It had been observed that certain bacteria always degraded "foreign" DNA that was introduced into them experimentally. A search for the mechanism responsible revealed a novel class of bacterial nucleases that cleave DNA at specific nucleotide sequences. The bacteria's own DNA is protected from cleavage by chemical modification of these specific sequences. Because these enzymes function to restrict the transfer of DNA between strains of bacteria, they were called **restriction nucleases**. The pursuit of this seemingly arcane biological puzzle set off the development of technologies that have forever changed the way cell and molecular biologists study living things.

Different bacterial species produce different restriction nucleases, each cutting at a different, specific nucleotide sequence (**Figure 10–2**). Because these target sequences are short—generally four to eight nucleotide pairs—many sites of cleavage will occur, purely by chance, in any long DNA molecule. The reason restriction nucleases are so useful in the laboratory is that each enzyme will cut a particular DNA molecule, at the same sites. Thus for a given sample of DNA, a particular restriction nuclease will reliably generate the same set of DNA fragments.

The size of the resulting fragments depends on the target sequences of the restriction nucleases. As shown in Figure 10–2, the enzyme HaeIII cuts at a sequence of four nucleotide pairs; a sequence this long would be expected to occur purely by chance approximately once every 256 nucleotide pairs (1 in 4^4). In comparison, a restriction nuclease with a target sequence that is eight nucleotides long would be expected to cleave DNA on average once every 65,536 nucleotide pairs (1 in 4^8). This difference in sequence selectivity makes it possible to cleave a long DNA molecule into the fragment sizes that are most suitable for a given application.

Gel Electrophoresis Separates DNA Fragments of Different Sizes

After a large DNA molecule is cleaved into smaller pieces with a restriction nuclease, the DNA fragments can be separated from one another on

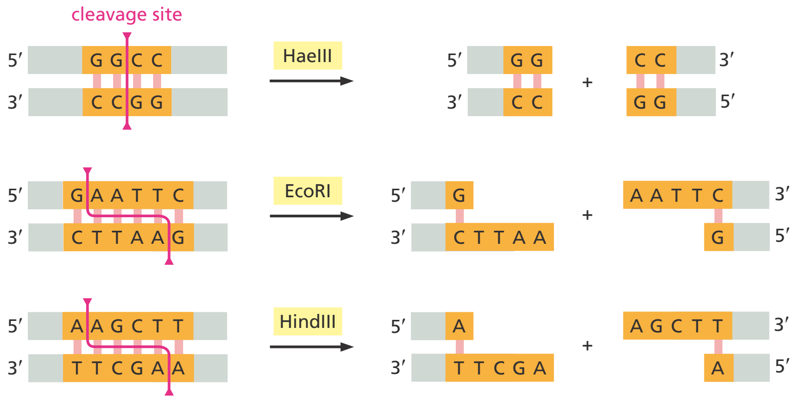


Figure 10–2 Restriction nucleases cleave DNA at specific nucleotide sequences. Target sequences are often palindromic (that is, the nucleotide sequence is symmetrical around a central point). Here, both strands of the DNA double helix are cut at specific points within the target sequence (*orange*). Some enzymes, such as HaeIII, cut straight across the double helix and leave two blunt-ended DNA molecules; with others, such as EcoRI and HindIII, the cuts on each strand are staggered. These staggered cuts generate "sticky ends"—short, single-stranded overhangs that help the cut DNA molecules join back together through complementary base-pairing. This rejoining of DNA molecules becomes important for DNA cloning, as we discuss later. Restriction nucleases are usually obtained from bacteria, and their names reflect their origins: for example, the enzyme EcoRI comes from *Escherichia coli*.

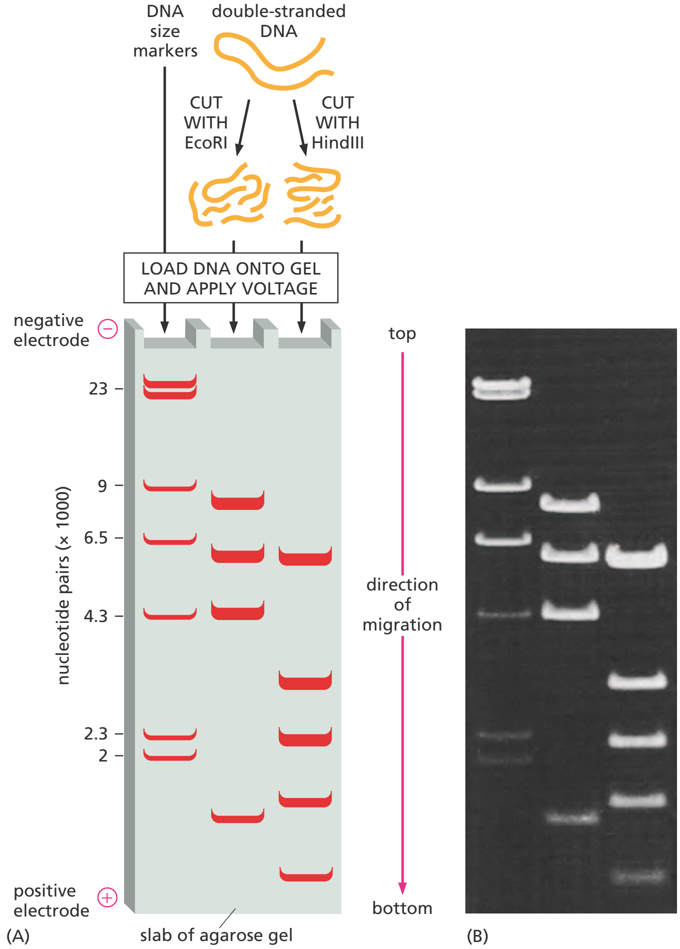


Figure 10–3 DNA molecules can be separated by size using gel electrophoresis. (A) Schematic illustration compares the results of cutting the same DNA molecule (in this case, the genome of a virus that infects parasitic wasps) with two different restriction nucleases, EcoRI (*middle*) and HindIII (*right*). The fragments are then separated by gel electrophoresis. Because larger fragments migrate more slowly than smaller ones, the lowermost bands on the gel contain the smallest DNA fragments. The sizes of the fragments can be estimated by comparing them to a set of DNA fragments of known sizes (*left*). (B) Photograph of an actual gel shows the positions of DNA bands that have been labeled with a fluorescent dye. (B, from U. Albrecht et al., *J. Gen. Virol.* 75:3353–3363, 1994.)

the basis of their length by gel electrophoresis—the same method used to separate mixtures of proteins (see Panel 4–5, p. 167). A mixture of DNA fragments is loaded at one end of a slab of agarose or polyacrylamide gel, which contains a microscopic network of pores. When a voltage is applied across the gel, the negatively charged DNA fragments migrate toward the positive electrode; larger fragments will migrate more slowly because their progress is impeded to a greater extent by the gel matrix. Over several hours, the DNA fragments become spread out across the gel according to size, forming a ladder of discrete bands, each composed of a collection of DNA molecules of identical length (**Figure 10–3**). To isolate a desired DNA fragment, the small section of the gel that contains the band is excised with a scalpel or a razor blade, and the DNA is then extracted.

QUESTION 10–2

Which products result when the double-stranded DNA molecule *below* is digested with (A) EcoRI, (B) HaeIII, (C) HindIII, or (D) all three of these enzymes together? (See Figure 10–2 for the target sequences of these enzymes.)

5′-AAGAATTGCGGAATTCGGGCCTTAAGCGCCGCGTCGAGGCCTTAAA-3′
3′-TTCTTAACGCCTTAAGCCCGGAATTCGCGGCGCAGCTCCGGAATTT-5′

Bands of DNA in a Gel Can Be Visualized Using Fluorescent Dyes or Radioisotopes

The separated DNA bands on an agarose or polyacrylamide gel are not, by themselves, visible. To see these bands, the DNA must be labeled or stained in some way. One sensitive method involves exposing the gel to a dye that fluoresces under ultraviolet (UV) light when it is bound to DNA. When the gel is placed on a UV light box, the individual bands glow bright orange—or bright white when the gel is photographed in black and white (see Figure 10–3B).

An even more sensitive detection method involves incorporating a radioisotope into the DNA molecules before they are separated by electrophoresis; ^{32}P is often used, as it can be incorporated into the phosphates of DNA. Because the β particles emitted from ^{32}P can activate the radiation-sensitive particles in photographic film, a sheet of film placed flat on top of the agarose gel will, when developed, show the position of all the DNA bands.

Exposing a gel to a fluorescent dye that binds to DNA—or starting with DNA that has been pre-labeled with ^{32}P—will allow every band on the gel to be seen. But it does not reveal which of those bands contains a DNA sequence of interest. To do that, a probe is designed to bind specifically to the desired nucleotide sequence by complementary base-pairing, as we see next.

Hybridization Provides a Sensitive Way to Detect Specific Nucleotide Sequences

Under normal conditions, the two strands of a DNA double helix are held together by hydrogen bonds between the complementary base pairs (see Figure 5–6). But these relatively weak, noncovalent bonds can be fairly easily broken. Such *DNA denaturation* will release the two strands from each other, but does not break the covalent bonds that link together the nucleotides within each strand. Perhaps the simplest way to achieve this separation involves heating the DNA to around 90°C. When the conditions are reversed—by slowly lowering the temperature—the complementary strands will readily come back together to re-form a double helix. This **hybridization**, or *DNA renaturation*, is driven by the re-formation of the hydrogen bonds between complementary base pairs (**Figure 10–4**).

This fundamental capacity of a single-stranded nucleic acid molecule, either DNA or RNA, to form a double helix with a single-stranded molecule of a complementary sequence provides a very powerful and sensitive technique for detecting specific nucleotide sequences in both DNA and RNA. Today, one simply designs a short, single-stranded *DNA probe* that is complementary to the nucleotide sequence of interest. Because the nucleotide sequences of so many genomes are known—and are stored in publicly accessible databases—designing such a probe is straightforward. The desired probe can then be synthesized in the laboratory—usually by a

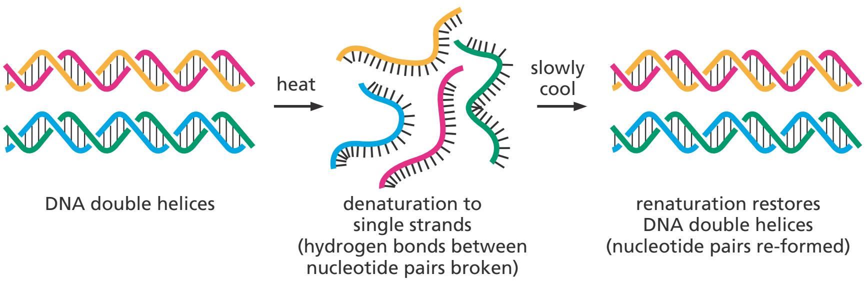

heat
slowly cool
DNA double helices
denaturation to single strands (hydrogen bonds between nucleotide pairs broken)
renaturation restores DNA double helices (nucleotide pairs re-formed)

Figure 10–4 A molecule of DNA can undergo denaturation and renaturation (hybridization). For two single-stranded molecules to hybridize, they must have complementary nucleotide sequences that allow base-pairing. In this example, the *red* and *orange* strands are complementary to each other, and the *blue* and *green* strands are complementary to each other. Although denaturation by heating is shown, DNA can also be renatured after being denatured by alkali treatment. The 1961 discovery that single strands of DNA could readily re-form a double helix in this way was a big surprise to scientists.

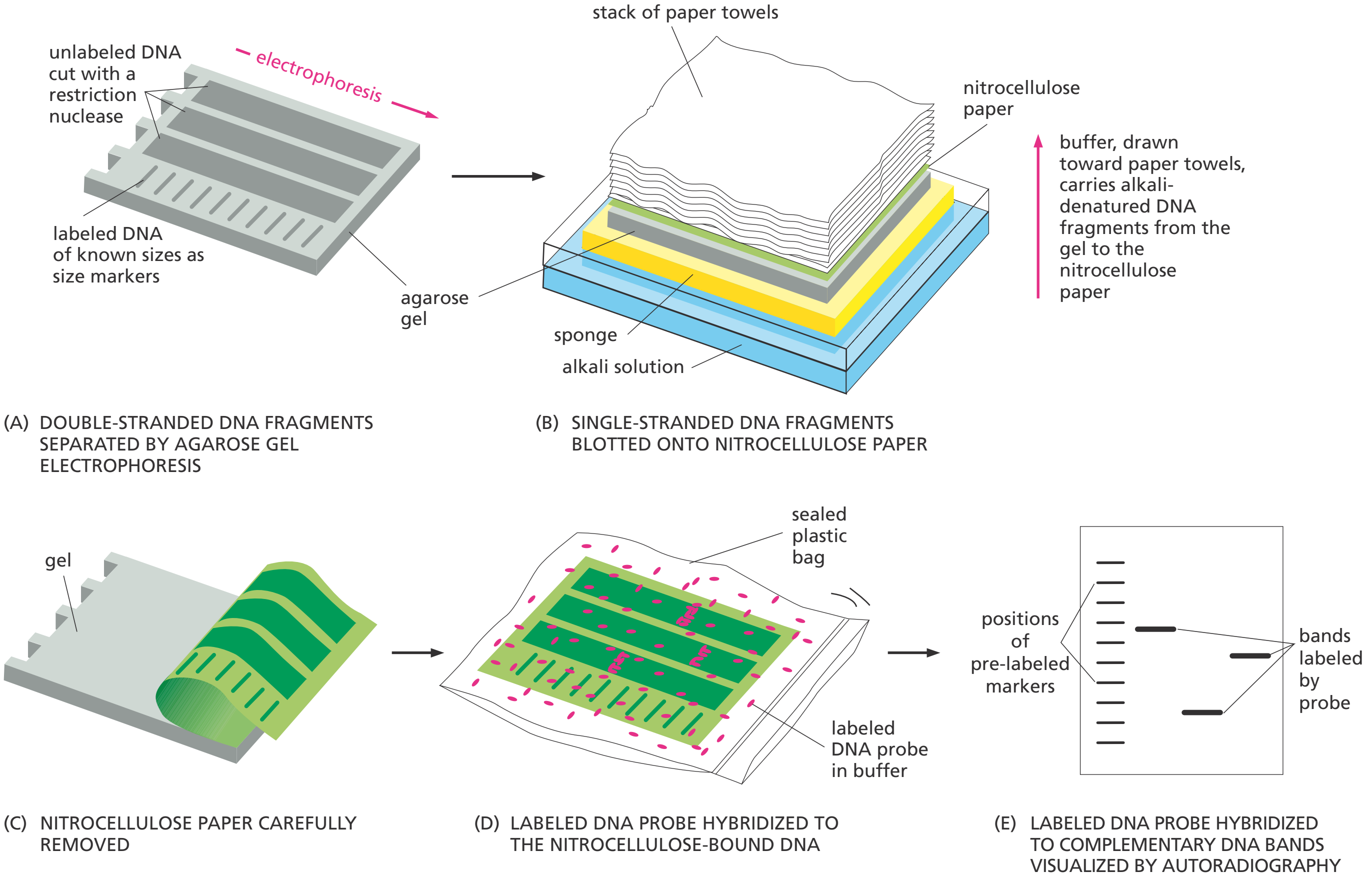

unlabeled DNA cut with a restriction nuclease
electrophoresis
labeled DNA of known sizes as size markers
agarose gel
stack of paper towels
nitrocellulose paper
buffer, drawn toward paper towels, carries alkali-denatured DNA fragments from the gel to the nitrocellulose paper
sponge
alkali solution
(A) DOUBLE-STRANDED DNA FRAGMENTS SEPARATED BY AGAROSE GEL ELECTROPHORESIS
(B) SINGLE-STRANDED DNA FRAGMENTS BLOTTED ONTO NITROCELLULOSE PAPER
gel
sealed plastic bag
labeled DNA probe in buffer
positions of pre-labeled markers
bands labeled by probe
(C) NITROCELLULOSE PAPER CAREFULLY REMOVED
(D) LABELED DNA PROBE HYBRIDIZED TO THE NITROCELLULOSE-BOUND DNA
(E) LABELED DNA PROBE HYBRIDIZED TO COMPLEMENTARY DNA BANDS VISUALIZED BY AUTORADIOGRAPHY

Figure 10–5 Gel-transfer hybridization, or Southern blotting, is used to detect specific DNA fragments. (A) The mixture of double-stranded DNA fragments generated by restriction nuclease treatment of DNA is separated according to length by gel electrophoresis. (B) A sheet of nitrocellulose paper is laid over the gel, and the separated DNA fragments are denatured with alkali and transferred to the sheet by blotting. In this process, a stack of absorbent paper towels is used to suck buffer up through the gel, transferring the single-stranded DNA fragments from the gel to the nitrocellulose paper. (C) The nitrocellulose sheet is carefully peeled off the gel. (D) The sheet containing the bound single-stranded DNA fragments is exposed to a radioactive, single-stranded DNA probe specific for the DNA sequence of interest under conditions that favor hybridization. (E) The sheet is washed thoroughly, so that only probe molecules that have hybridized to the DNA on the paper remain attached. After autoradiography, the DNA that has hybridized to the labeled probe will show up as a band on the autoradiograph. An adaptation of this technique, used to detect specific RNA sequences, is called *Northern blotting.* In this case, RNA molecules are electrophoresed through the gel, and the probe is usually a single-stranded DNA molecule. The same procedures can be carried out with non-radioactive probes using an appropriate method of detection.

commercial organization or a centralized academic facility. Such probes carry a fluorescent or radioactive label to facilitate detection of the nucleotide sequence to which they bind.

Once a suitable probe has been obtained, it can be used in a variety of situations to search for nucleic acids with a complementary sequence—for example, finding a sequence of interest among DNA fragments that have been separated on an agarose gel. In this case, the fragments are first transferred to a special sheet of paper, which is then exposed to the labeled probe. This common technique, called *Southern blotting*, was named after the scientist who invented it (**Figure 10–5**).

DNA probes are widely used in cell biology. Later in the chapter, we discuss how they can be used to determine in which tissues and at what stages of development a gene is transcribed. But first, we consider how hybridization facilitates the process of DNA cloning.

DNA CLONING IN BACTERIA

The term **DNA cloning** refers to the production of many identical copies of a DNA sequence. It is this amplification that makes it possible to separate a defined segment of DNA—often a gene of interest—from the rest of a cell's genome. DNA cloning is one of the most important feats of recombinant DNA technology, as it is the starting point for understanding the function of any stretch of DNA within the genome.

In this section, we describe the classical approach to DNA cloning, in which one copies all of the DNA from a cell or tissue and then finds and isolates the specific DNA of interest. Later, we discuss how the development of the *polymerase chain reaction* (*PCR*) has facilitated a more direct approach to cloning, allowing one to copy, in a test tube, only the DNA fragment of interest.

DNA Cloning Begins with Genome Fragmentation and Production of Recombinant DNAs

Whole genomes, even small ones, are too large and unwieldy to be handled easily in the laboratory. Thus the first step in cloning any gene is to break the genome into smaller, more manageable pieces. These fragments can then be joined together, or recombined, to produce the DNA molecules that will be amplified. Our ability to generate such **recombinant DNA molecules** is made possible by the use of molecular tools that are provided by cells themselves.

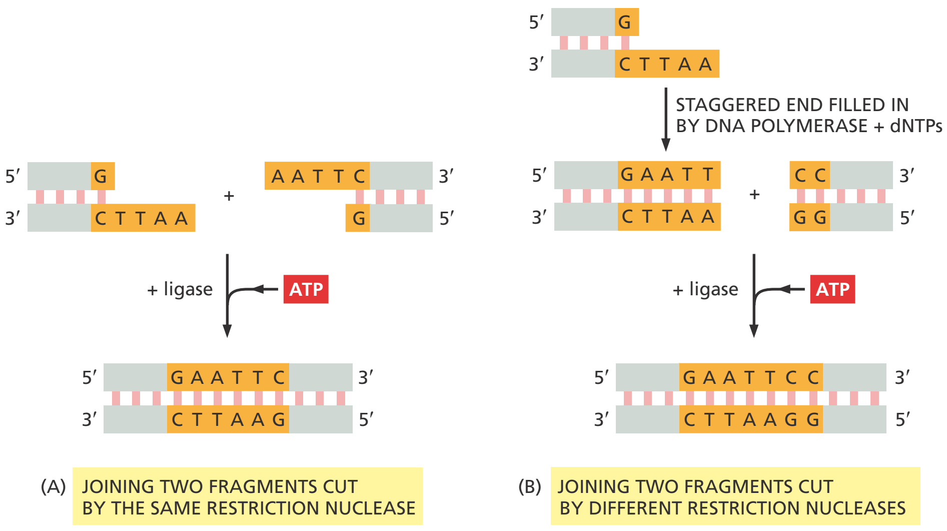
As we discussed earlier, bacterial restriction nucleases can be used to cut long DNA molecules into conveniently sized fragments (see Figure 10–2). These fragments can then be joined to one another—or to any piece of DNA—using **DNA ligase**, an enzyme that reseals the nicks that arise in the DNA backbone during DNA replication and DNA repair in cells (see Figure 6–18). DNA ligase allows investigators to join together any two pieces of DNA in a test tube, producing recombinant DNA molecules that are not found in nature (**Figure 10–6**).

The production of recombinant DNA molecules in this way is a key step in the classical approach to DNA cloning. It allows the DNA fragments generated by treatment with a restriction nuclease to be inserted into another, special DNA molecule that serves as a carrier, or *vector*, which can be copied—and thereby amplified—inside a cell, as we discuss next.

Recombinant DNA Can Be Inserted Into Plasmid Vectors

The vectors typically used for gene cloning are relatively small, circular DNA molecules called **plasmids**. (**Figure 10–7**). Each plasmid contains a replication origin, which enables it to replicate in a bacterial cell independently of the bacterial chromosome. It also has cleavage sites for common restriction nucleases, so that the plasmid can be conveniently opened and a foreign DNA fragment inserted.

The plasmids used for cloning are basically streamlined versions of plasmids that occur naturally in many bacteria. Bacterial plasmids were first recognized by physicians and scientists because they often carry



(A) JOINING TWO FRAGMENTS CUT BY THE SAME RESTRICTION NUCLEASE

(B) JOINING TWO FRAGMENTS CUT BY DIFFERENT RESTRICTION NUCLEASES

Figure 10–6 DNA ligase can join together any two DNA fragments *in vitro* to produce recombinant DNA molecules. ATP provides the energy necessary for the ligase to reseal the sugar–phosphate backbone of DNA. (A) DNA ligase can readily join two DNA fragments produced by the same restriction nuclease, in this case EcoRI. Note that the staggered ends produced by this enzyme enable the ends of the two fragments to base-pair correctly with each other, greatly facilitating their rejoining. (B) DNA ligase can also be used to join DNA fragments produced by different restriction nucleases—for example, EcoRI and HaeIII. In this case, before the fragments undergo ligation, DNA polymerase plus a mixture of deoxyribonucleoside triphosphates (dNTPs) are used to fill in the staggered cut produced by EcoRI. Each DNA fragment shown in the figure is oriented so that its 5′ ends are the left end of the upper strand and the right end of the lower strand, as indicated.

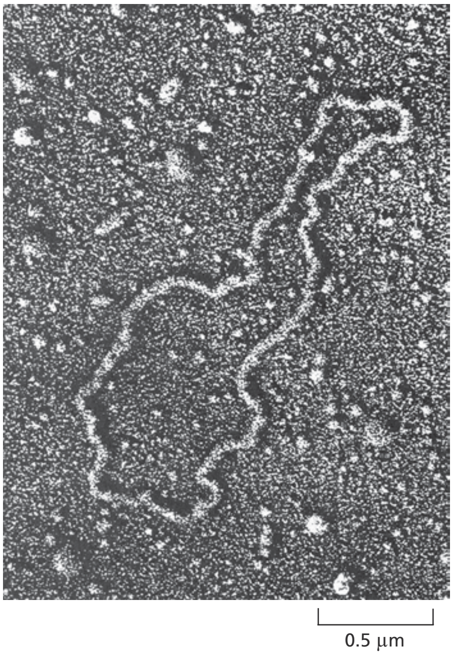


Figure 10–7 Bacterial plasmids are commonly used as cloning vectors. This circular, double-stranded DNA molecule was the first plasmid for DNA cloning; it contains about nine thousand nucleotide pairs. The staining procedure used to make the DNA visible in this electron micrograph causes the DNA to appear much thicker than it actually is. (Courtesy of Stanley N. Cohen, Stanford University.)

genes that render their microbial host resistant to one or more antibiotics. Indeed, historically potent antibiotics—penicillin, for example—are no longer effective against many of today's bacterial infections because plasmids that confer resistance to the antibiotic have spread among bacterial species by horizontal gene transfer (see Figure 9–14).

To insert a piece of DNA into a plasmid vector, the purified plasmid DNA is opened up by a restriction nuclease that cleaves it at a single site, and the DNA fragment to be cloned is then spliced into that site using DNA ligase (**Figure 10–8**). This recombinant DNA molecule is now ready to be introduced into a bacterium, where it will be copied and amplified, as we see next.

Recombinant DNA Can Be Copied Inside Bacterial Cells

To introduce recombinant DNA into a bacterial cell, investigators take advantage of the fact that some bacteria naturally take up DNA molecules present in their surroundings. The mechanism that controls this uptake is called **transformation**, because early observations suggested it could "transform" one bacterial strain into another. Indeed, the first proof that genes are made of DNA came from an experiment in which DNA purified from a pathogenic strain of pneumococcus was used to transform a harmless bacterium into a deadly one (see How We Know, pp. 174–176).

In a natural bacterial population, a source of DNA for transformation is provided by bacteria that have died and released their contents, including DNA, into the environment. In a test tube, however, bacteria such as *E. coli* can be coaxed to take up recombinant DNA that has been created in the laboratory. These bacteria are then suspended in a nutrient-rich broth and allowed to proliferate.

Each time the bacterial population doubles—every 30 minutes or so—the number of copies of the recombinant DNA molecule also doubles. Thus, in 24 hours, the engineered cells will produce hundreds of millions of copies of the plasmid, along with the DNA fragment it contains. The bacteria can then be split open (lysed) and the plasmid DNA purified from

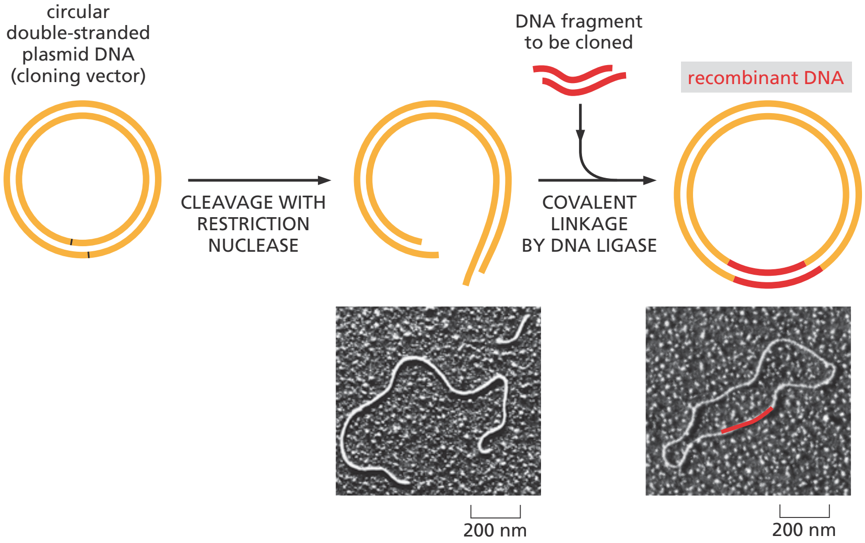

circular double-stranded plasmid DNA (cloning vector)
CLEAVAGE WITH RESTRICTION NUCLEASE
DNA fragment to be cloned
COVALENT LINKAGE BY DNA LIGASE
recombinant DNA
200 nm
200 nm

Figure 10–8 A DNA fragment is inserted into a bacterial plasmid by using the enzyme DNA ligase. The plasmid is first cut open at a single site with a restriction nuclease (in this case, one that produces staggered ends). It is then mixed with the DNA fragment to be cloned, which has been cut with the same restriction nuclease. DNA ligase and ATP are also added to the mix. The staggered ends base-pair, and the nicks in the DNA backbone are sealed by the DNA ligase to produce a complete recombinant DNA molecule. In the accompanying micrographs, we have colored the DNA fragment *red* to make it easier to see. (Micrographs courtesy of Huntington Potter and David Dressler.)

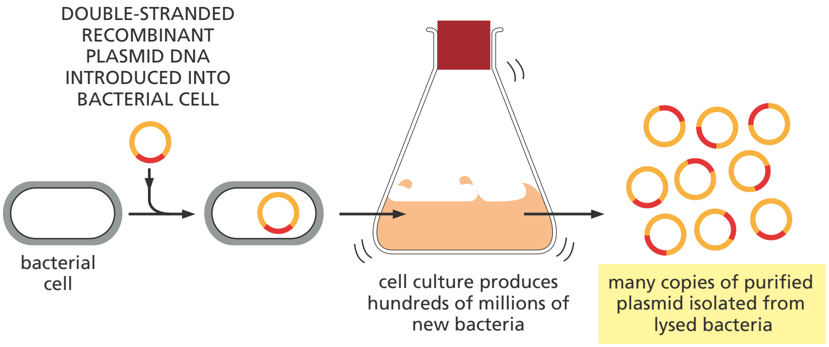

DOUBLE-STRANDED RECOMBINANT PLASMID DNA INTRODUCED INTO BACTERIAL CELL
bacterial cell
cell culture produces hundreds of millions of new bacteria
many copies of purified plasmid isolated from lysed bacteria

Figure 10–9 A DNA fragment can be replicated inside a bacterial cell. To clone a particular fragment of DNA, it is first inserted into a plasmid vector, as shown in Figure 10–8. The resulting recombinant plasmid DNA is then introduced into a bacterium, where it is replicated many millions of times as the bacterium multiplies. For simplicity, the genome of the bacterial cell is not shown.

the rest of the cell contents, including the large bacterial chromosome (**Figure 10–9**).

The DNA fragment can be readily recovered by cutting it out of the plasmid DNA with the same restriction nuclease that was used to insert it, and then separating it from the plasmid DNA by gel electrophoresis (see Figure 10–3). Together, these steps allow the amplification and purification of any segment of DNA from the genome of any organism.

Genes Can Be Isolated from a DNA Library

Thus far, we have described the amplification of a single DNA fragment. In reality, when a genome is cut by a restriction nuclease, millions of different DNA fragments are generated. How can the single fragment that contains the DNA of interest be isolated from this collection? The solution involves introducing all of the fragments into bacteria and then selecting those bacterial cells that have amplified the desired DNA molecule.

The entire collection of DNA fragments can be ligated into plasmid vectors, using conditions that favor the insertion of a single DNA fragment into each plasmid molecule. These recombinant plasmids are then introduced into *E. coli* at a concentration that ensures that no more than one plasmid molecule is taken up by each bacterium. The collection of cloned DNA fragments in this bacterial culture is known as a **DNA library**. Because the DNA fragments were derived directly from the chromosomal DNA of the organism of interest, the resulting collection—called a **genomic library**—should represent the entire genome of that organism (**Figure 10–10**).

To find a particular gene within this library, one can use a labeled DNA probe designed to bind specifically to part of the gene's DNA sequence. Using such a probe, the rare bacterial clones in the DNA library that contain the gene—or a portion of it—can be identified by hybridization (**Figure 10–11**).

But before a gene has been cloned, how can one design a probe to detect it? In the early days of cloning, investigators wishing to study a protein-coding gene would first determine at least part of the protein's amino acid sequence. By applying the genetic code in reverse, they could use this amino acid sequence to deduce the corresponding gene sequence, which allowed them to generate an appropriate DNA probe.

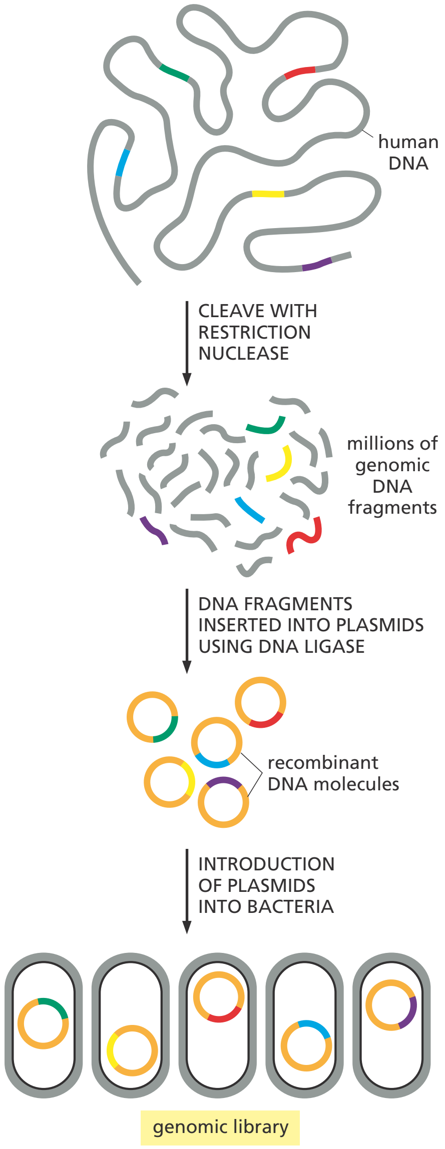

human DNA
CLEAVE WITH RESTRICTION NUCLEASE
millions of genomic DNA fragments
DNA FRAGMENTS INSERTED INTO PLASMIDS USING DNA LIGASE
recombinant DNA molecules
INTRODUCTION OF PLASMIDS INTO BACTERIA
genomic library

Figure 10–10 Human genomic libraries containing DNA fragments representing the whole human genome can be constructed using restriction nucleases and DNA ligase. Such a genomic library consists of a set of bacteria, each carrying a different small fragment of human DNA. For simplicity, only the *colored* DNA fragments are shown in the library; in reality, all of the different *gray* fragments will also be represented.

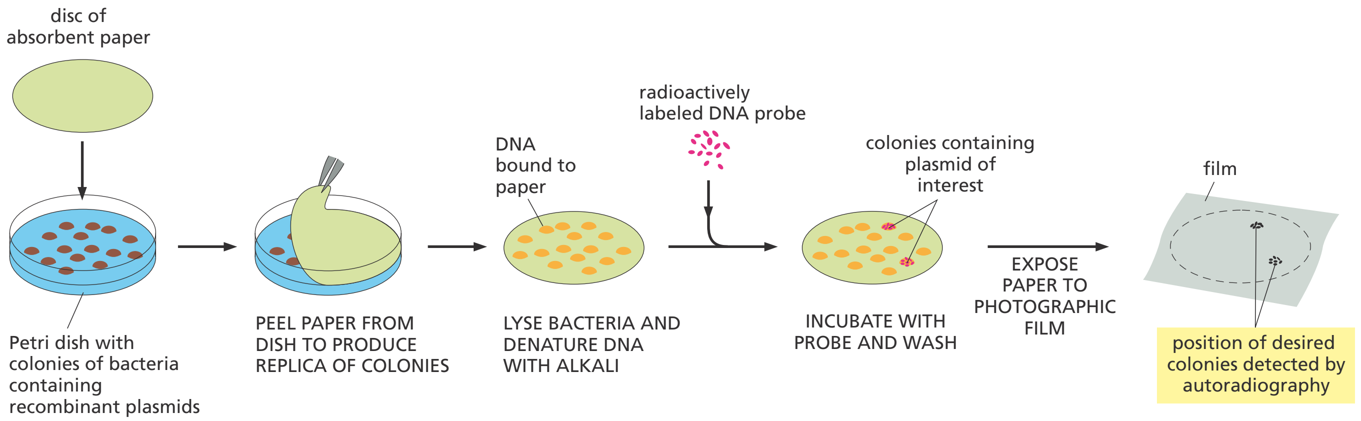

disc of absorbent paper
Petri dish with colonies of bacteria containing recombinant plasmids
PEEL PAPER FROM DISH TO PRODUCE REPLICA OF COLONIES
DNA bound to paper
LYSE BACTERIA AND DENATURE DNA WITH ALKALI
radioactively labeled DNA probe
colonies containing plasmid of interest
INCUBATE WITH PROBE AND WASH
EXPOSE PAPER TO PHOTOGRAPHIC FILM
film
position of desired colonies detected by autoradiography

Figure 10–11 A bacterial colony carrying a particular DNA clone can be identified by hybridization. A replica of the arrangement of the bacterial colonies (clones) on the Petri dish is made by pressing a piece of absorbent paper against the surface of the dish. This replica is treated with alkali (to lyse the cells and dissociate the plasmid DNA into single strands), and the paper is then hybridized to a highly radioactive DNA probe. Those bacterial colonies that have bound the probe are identified by autoradiography. Living bacterial cells containing the plasmid can then be isolated from the original Petri dish.

Many genes were originally identified and cloned using variations on this basic approach. Now that the complete genome sequences of many organisms, including humans, are known, however, cloning genes is very much easier, faster, and cheaper. The sequence of any gene in an organism can be looked up in an electronic database, making it a simple matter to design a probe that can be synthesized to order. As we discuss shortly, gene cloning today is typically done directly on the original DNA sample, bypassing the use of a DNA library entirely.

cDNA Libraries Represent the mRNAs Produced by Particular Cells

For many applications—for example, when attempting to clone a protein-coding gene, it is advantageous to obtain the gene in a form that contains only the coding sequence; that is, a form that lacks the intron DNA. For some genes, the complete genomic clone—including introns and exons—is too large and unwieldy to handle conveniently in the laboratory (see, for example, Figure 7–18B). What's more, the bacterial or yeast cells typically used to amplify cloned DNA are unable to remove introns from mammalian RNA transcripts. So if the goal is to use a cloned mammalian gene to produce a large amount of the protein it encodes, it is essential to use only the coding sequence of the gene. Fortunately, it is relatively simple to isolate a gene free of all its introns, by using a different type of DNA library, called a **cDNA library**.

A cDNA library is similar to a genomic library in that it also contains numerous clones containing many different DNA sequences. But it differs in one important respect. The DNA that goes into a cDNA library is not genomic DNA; it is DNA copied from the mRNAs present in a particular type of cell. To prepare a **cDNA** library, all of the mRNAs are extracted, and double-stranded DNA copies of these mRNAs are produced by the enzymes *reverse transcriptase* and DNA polymerase (**Figure 10–12**). These **complementary DNA**—or **cDNA**—molecules are then introduced into bacteria and amplified, as described for genomic DNA fragments (see Figure 10–10). The gene of interest—in this case, without its introns—can then be isolated by using a probe that hybridizes to the DNA sequence (see Figure 10–11). We discuss later how such cDNAs can be used to produce purified proteins on a commercial scale.

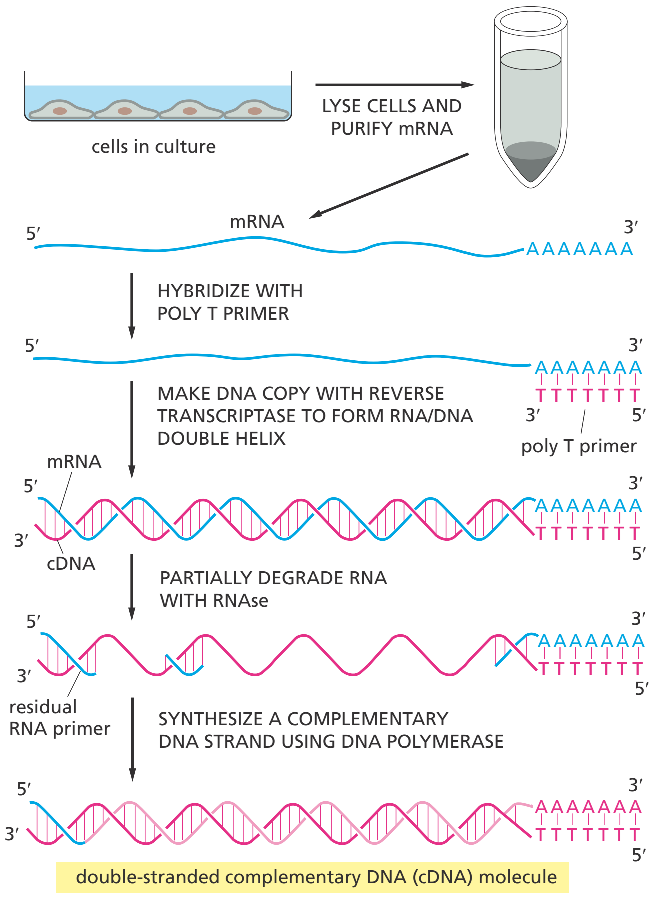

cells in culture
LYSE CELLS AND PURIFY mRNA
mRNA
5′
3′
AAAAAAA
HYBRIDIZE WITH POLY T PRIMER
5′
3′
AAAAAAA
TTTTTTT
3′
5′
poly T primer
MAKE DNA COPY WITH REVERSE TRANSCRIPTASE TO FORM RNA/DNA DOUBLE HELIX
mRNA
5′
3′
AAAAAAA
TTTTTTT
3′
5′
cDNA
PARTIALLY DEGRADE RNA WITH RNAse
5′
3′
AAAAAAA
TTTTTTT
3′
5′
residual RNA primer
SYNTHESIZE A COMPLEMENTARY DNA STRAND USING DNA POLYMERASE
5′
3′
AAAAAAA
TTTTTTT
3′
5′
double-stranded complementary DNA (cDNA) molecule

Figure 10–12 Complementary DNA (cDNA) is prepared from mRNA. Total mRNA is extracted from a selected type of cell, and double-stranded, complementary DNA (cDNA) is produced using reverse transcriptase (see Figure 9–30) and DNA polymerase. For simplicity, the copying of just one of these mRNAs into cDNA is illustrated here. Note that an RNA fragment that remains hybridized to the first cDNA strand after partial RNAse digestion serves as the primer needed for DNA polymerase to begin synthesis of the complementary DNA strand.

There are several important differences between genomic DNA clones and cDNA clones, as illustrated in **Figure 10–13**. Genomic clones represent a random sample of all of the DNA sequences found in an organism's genome and, with very rare exceptions, will contain the same sequences regardless of the cell type from which the DNA came. Also, genomic clones from eukaryotes contain large amounts of noncoding DNA, repetitive DNA sequences, introns, regulatory DNA, and spacer DNA; sequences that code for proteins will make up only a few percent of the library (see Figure 9–33). By contrast, cDNA clones contain predominantly protein-coding sequences, and only those for genes that have been transcribed into mRNA in the cells from which the cDNA was made. As different types of cells produce distinct sets of mRNA molecules, each yields a different cDNA library. Furthermore, patterns of gene expression change during development, so cells at different stages in their development will also yield different cDNA libraries.

As we discuss later, cDNAs are used to assess which genes are expressed in specific cells, at particular times in development, or under a particular set of conditions. In contrast, genomic clones—which include introns and exons, as well as regulatory DNA sequences—provide the starting material for determining the complete nucleotide sequence of an organism's genome.

QUESTION 10–3

Discuss the following statement: "From the nucleotide sequence of a cDNA clone, the complete amino acid sequence of a protein can be deduced by applying the genetic code. Thus, protein biochemistry has become superfluous because there is nothing more that can be learned by studying the protein."

DNA CLONING BY PCR

Genomic and cDNA libraries were once the only route to gene cloning, and they are still used for cloning very large genes and for sequencing whole genomes. However, a powerful and versatile method for amplifying DNA, known as the **polymerase chain reaction** (**PCR**), provides a

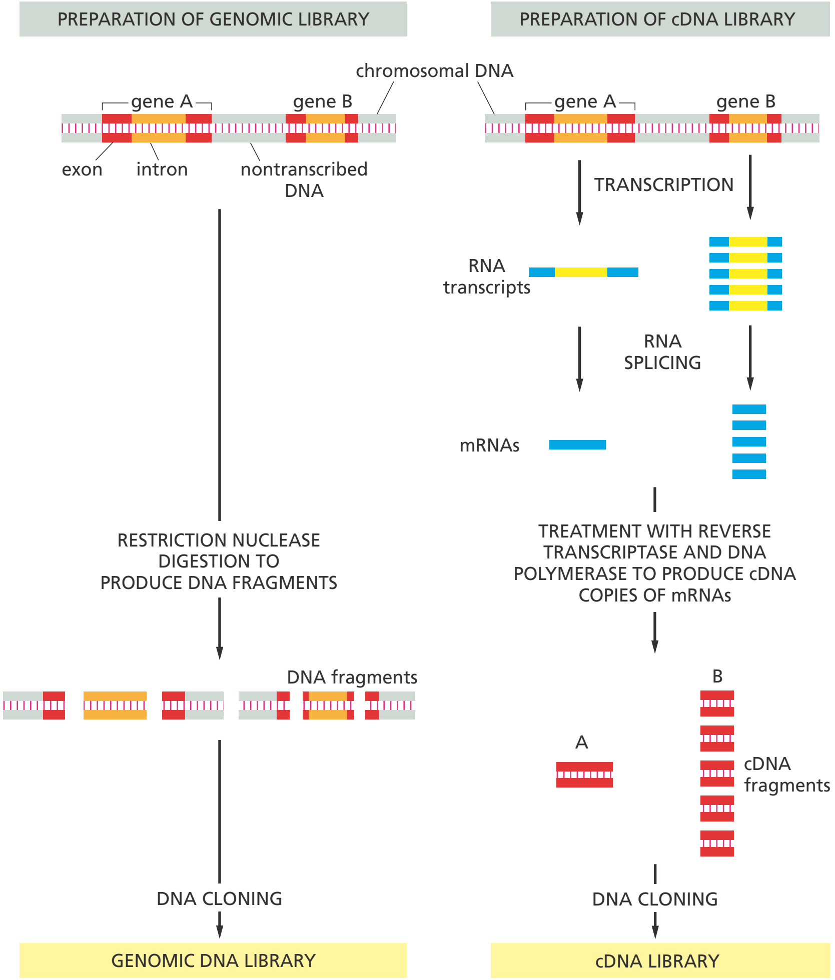

PREPARATION OF GENOMIC LIBRARY
PREPARATION OF cDNA LIBRARY
chromosomal DNA
gene A
gene B
gene A
gene B
exon
intron
nontranscribed DNA
TRANSCRIPTION
RNA transcripts
RNA SPLICING
mRNAs
RESTRICTION NUCLEASE DIGESTION TO PRODUCE DNA FRAGMENTS
TREATMENT WITH REVERSE TRANSCRIPTASE AND DNA POLYMERASE TO PRODUCE cDNA COPIES OF mRNAs
DNA fragments
B
A
cDNA fragments
DNA CLONING
DNA CLONING
GENOMIC DNA LIBRARY
cDNA LIBRARY

Figure 10–13 Genomic DNA clones and cDNA clones derived from the same region of the genome are different. In this example, gene A is infrequently transcribed, whereas gene B is frequently transcribed, and both genes contain introns (*orange*). In the genomic DNA library, both introns and nontranscribed DNA (*gray*) are included in the clones, and most clones will contain either no coding sequence or only part of the coding sequence of a gene (*red*); the DNA sequences that regulate the expression of each gene are also included (not indicated). In the cDNA clones, the intron sequences have been removed by RNA splicing during the formation of the mRNA (*blue*), and a continuous coding sequence is therefore present in each clone. Because gene B is transcribed more frequently than gene A in the cells from which the cDNA library was made, it will be represented much more often than A in the cDNA library. In contrast, genes A and B should be represented equally in the genomic library.

more rapid and straightforward approach to DNA cloning, particularly in organisms whose complete genome sequence is known. Today, most genes are cloned via PCR.

Invented in the 1980s, PCR revolutionized the way that DNA and RNA are analyzed. The technique can amplify any nucleotide sequence rapidly and selectively. Unlike the traditional approach of cloning using vectors—which relies on bacteria to make copies of the desired DNA sequences—PCR is performed entirely in a test tube. Eliminating the need for bacteria makes PCR convenient and incredibly quick—billions of copies of a nucleotide sequence can be generated in a matter of hours. At the same time, PCR is remarkably sensitive: the method can be used to detect the trace amounts of DNA in a drop of blood left at a crime scene or in a few copies of a viral genome in a patient's blood sample. Because of its sensitivity, speed, and ease of use, PCR has many applications in addition to DNA cloning, including forensics and diagnostics.

In this section, we provide a brief overview of how PCR works and how it is used for a range of purposes that require the amplification of specific DNA sequences.

PCR Uses a DNA Polymerase to Amplify Selected DNA Sequences in a Test Tube

The success of PCR depends on the exquisite selectivity of DNA hybridization, along with the ability of DNA polymerase to copy a DNA template reliably, through repeated rounds of replication *in vitro*. The enzyme works by adding nucleotides to the 3′ end of a growing strand of DNA (see Figure 6–11). To initiate the reaction, the polymerase requires a primer—a short nucleotide sequence that provides a 3′ end from which synthesis can begin. The beauty of PCR is that the primers that are added to the reaction mixture not only serve as starting points, they also direct the polymerase to the specific DNA sequence to be amplified. These primers, like the DNA probes used to identify specific nucleotide sequences as discussed earlier, are designed by the experimenter based on the DNA sequence of interest and then synthesized chemically. Thus, PCR can only be used to clone a DNA segment for which the sequence is known in advance. With the large and growing number of genome sequences available in public databases, this requirement is rarely a drawback.

Multiple Cycles of Amplification *In Vitro* Generate Billions of Copies of the Desired Nucleotide Sequence

PCR is an iterative process in which the cycle of amplification is repeated dozens of times. At the start of each cycle, the two strands of the double-stranded DNA template are separated and a unique primer is annealed to each. DNA polymerase is then allowed to replicate each strand independently (**Figure 10–14**). In subsequent cycles, all the newly synthesized DNA molecules produced by the polymerase serve as templates for the next round of replication (**Figure 10–15**). Through this iterative amplification process, many copies of the original sequence can be made—billions after about 20 to 30 cycles.

PCR is now the method of choice for cloning relatively short DNA fragments (say, under 10,000 nucleotide pairs). Each cycle takes only about five minutes, and automation of the whole procedure enables cell-free cloning of a DNA fragment in a few hours, compared with the several days required for cloning in bacteria. The original template for PCR can be either DNA or RNA, so this method can be used to obtain either a full genomic clone (complete with introns and exons) or a cDNA copy of an mRNA (**Figure 10–16**). A major benefit of PCR is that genes can be cloned directly from any piece of DNA or RNA without the time and effort needed to first construct a DNA library.

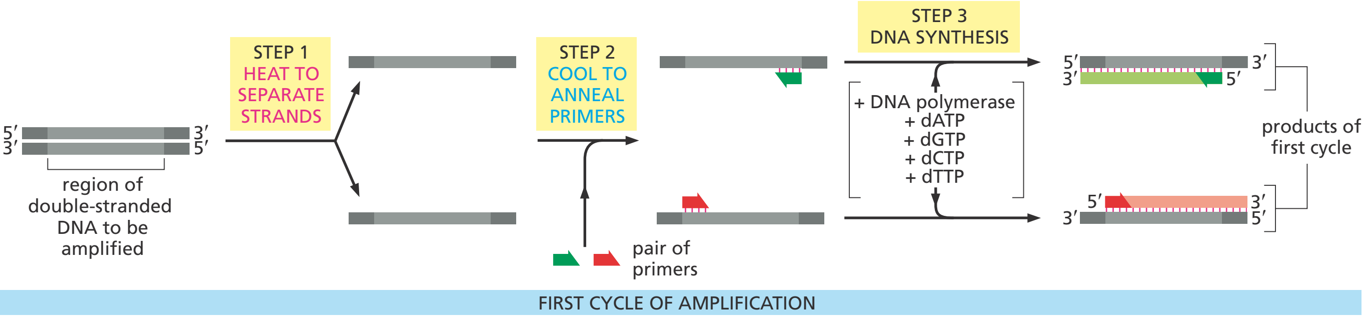


Figure 10–14 A pair of PCR primers directs the amplification of a desired segment of DNA in a test tube. Each cycle of PCR includes three steps: (1) The double-stranded DNA is heated briefly to separate the two strands. (2) The DNA is exposed to a large excess of a pair of specific primers—designed to bracket the region of DNA to be amplified—and the sample is cooled to allow the primers to hybridize to complementary sequences in the two DNA strands. (3) This mixture is incubated with DNA polymerase and the four deoxyribonucleoside triphosphates so that DNA can be synthesized, starting from the two primers. The cycle can then be repeated by reheating the sample to separate the newly synthesized DNA strands (see Figure 10–15).

The technique depends on the use of a special DNA polymerase isolated from a thermophilic bacterium; this polymerase is stable at much higher temperatures than eukaryotic DNA polymerases, so it is not denatured by the heat treatment shown in step 1. The enzyme therefore does not have to be added again after each cycle.

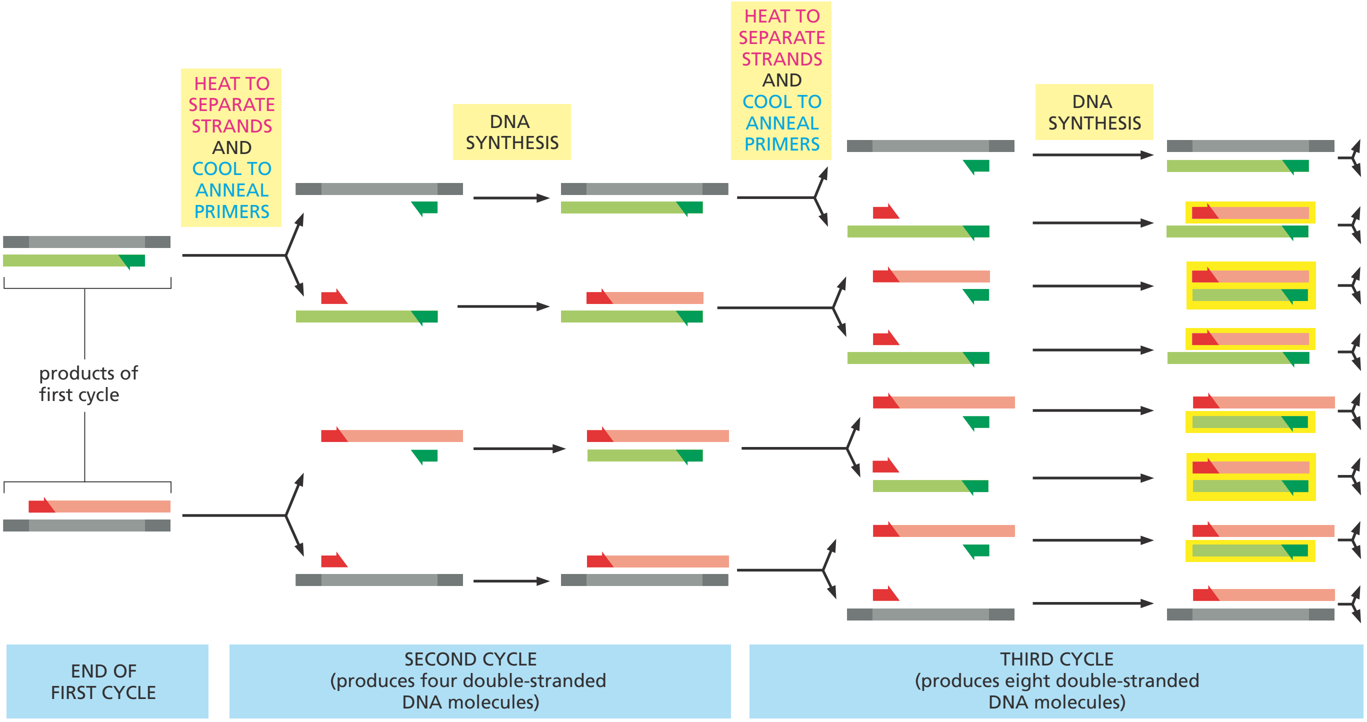

HEAT TO SEPARATE STRANDS AND COOL TO ANNEAL PRIMERS
DNA SYNTHESIS
HEAT TO SEPARATE STRANDS AND COOL TO ANNEAL PRIMERS
DNA SYNTHESIS
products of first cycle
END OF FIRST CYCLE
SECOND CYCLE (produces four double-stranded DNA molecules)
THIRD CYCLE (produces eight double-stranded DNA molecules)

Figure 10–15 PCR uses repeated rounds of strand separation, hybridization, and synthesis to amplify DNA. As the procedure outlined in Figure 10–14 is repeated, all the newly synthesized fragments serve as templates in their turn. Because the polymerase and the primers remain in the sample after the first cycle, PCR involves simply heating and then cooling the same sample, in the same test tube, again and again. Each cycle doubles the amount of DNA synthesized in the previous cycle, so that within a few cycles, the predominant DNA is identical to the sequence bracketed by and including the two primers in the original template. In the example illustrated here, three cycles of reaction produce 16 DNA chains, 8 of which (boxed in *yellow*) correspond exactly to one or the other strand of the original bracketed sequence. After four more cycles, 240 of the 256 DNA chains will correspond exactly to the original sequence, and after several more cycles, essentially all of the DNA strands will be this length. The whole procedure is shown in Movie 10.1.

PCR is Also Used for Diagnostic and Forensic Applications

In addition to its use in gene cloning, PCR is frequently employed to amplify DNA for other, more practical purposes. Because of its extraordinary sensitivity, PCR can be used to detect invading microorganisms at very early stages of infection. In this case, short sequences complementary to a segment of the infectious agent's genome are used as primers, and following many cycles of amplification, even a few copies of an invading bacterial or viral genome in a patient sample can be detected (**Figure 10–17**). For many infections, PCR has replaced the use of antibodies against microbial molecules to detect the presence of pathogens. PCR can also be used to track epidemics, detect bioterrorist attacks, and test food products for the presence of potentially harmful microbes. It is also used to verify the authenticity of a food source—for example, whether a sample of beef actually came from a cow.

Finally, PCR is now widely used in forensic medicine. The method's extreme sensitivity allows forensic investigators to isolate DNA from minute traces of human blood or other tissue to obtain a *DNA fingerprint* of the person who left the sample behind. With the possible exception of identical twins, the genome of each human differs in DNA sequence from that of every other person on Earth. Using primer pairs targeted at genome sequences that are known to be highly variable in the human population, PCR makes it possible to generate a distinctive DNA fingerprint for any individual (**Figure 10–18**). Such forensic analyses can be used not only to point the finger at those who have done wrong, but—equally important—to help exonerate those who have been wrongfully convicted.

QUESTION 10–4

A. If the PCR shown in Figure 10–15 is carried through an additional two rounds of amplification, how many of the DNA fragments labeled in gray, green, or red or outlined in yellow are produced? If many additional cycles are carried out, which fragments will predominate?

B. Assume you start with one double-stranded DNA molecule and amplify a 500-nucleotide-pair sequence contained within it. Approximately how many cycles of PCR amplification will you need to produce 100 ng of this DNA? 100 ng is an amount that can be easily detected after staining with a fluorescent dye. (Hint: for this calculation, you need to know that each nucleotide has an average molecular mass of 330 g/mole.)

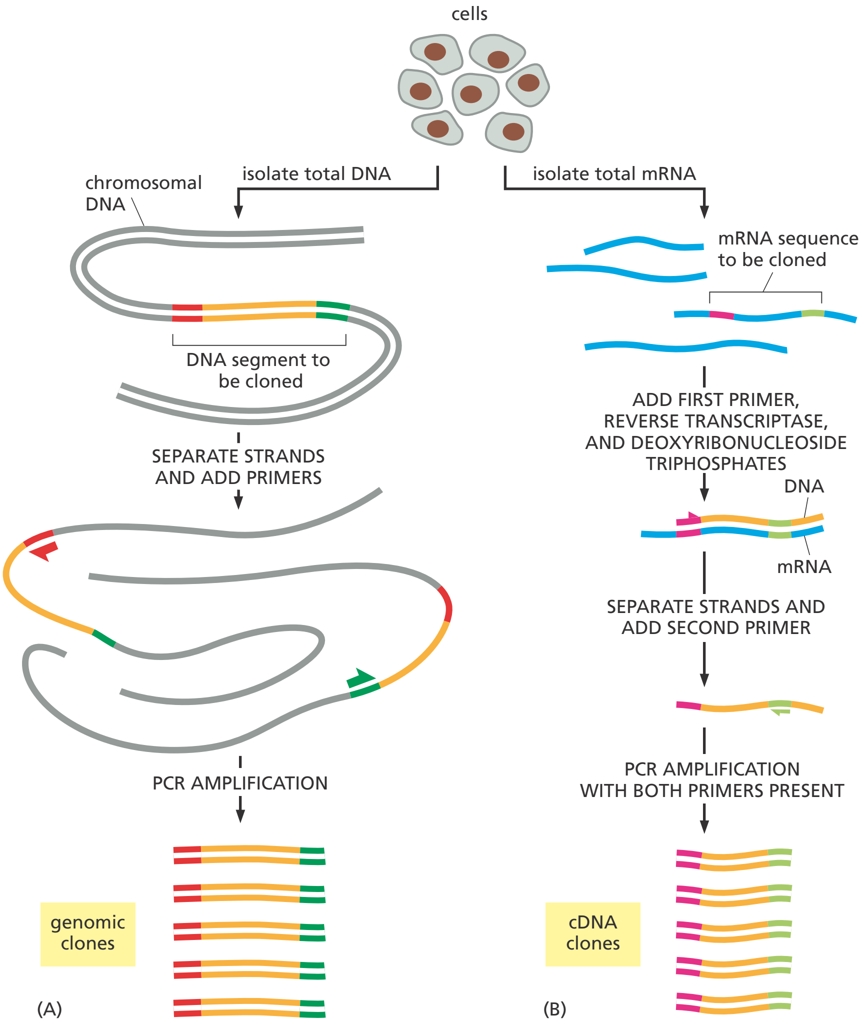

cells
isolate total DNA
isolate total mRNA
chromosomal DNA
DNA segment to be cloned
SEPARATE STRANDS AND ADD PRIMERS
PCR AMPLIFICATION
genomic clones
(A)
mRNA sequence to be cloned
ADD FIRST PRIMER, REVERSE TRANSCRIPTASE, AND DEOXYRIBONUCLEOSIDE TRIPHOSPHATES
DNA
mRNA
SEPARATE STRANDS AND ADD SECOND PRIMER
PCR AMPLIFICATION WITH BOTH PRIMERS PRESENT
cDNA clones
(B)

Figure 10–16 PCR can be used to obtain either genomic or cDNA clones. (A) To use PCR to clone a segment of chromosomal DNA, total DNA is first purified from cells. PCR primers that flank the stretch of DNA to be cloned are added, and many cycles of PCR are completed (see Figure 10–15). Because only the DNA between (and including) the primers is amplified, PCR provides a way to obtain selectively any short stretch of chromosomal DNA in an effectively pure form. (B) To use PCR to obtain a cDNA clone of a gene, total mRNA is first purified from cells. The first primer is added to the population of mRNAs, and reverse transcriptase is used to make a DNA strand complementary to the specific RNA sequence of interest. The second primer is then added, and the DNA molecule is amplified through many cycles of PCR.

EXPLORING AND EXPLOITING GENE FUNCTION

The procedures we have described thus far enable biologists to obtain large amounts of DNA in a form that is easy to work with in the laboratory. Whether present as fragments stored in a DNA library in bacteria or as a collection of PCR products nestled in the bottom of a test tube, this DNA also provides the raw material for experiments designed to unravel how individual genes—and the RNA molecules and proteins they encode—function in cells and organisms.

This is where creativity comes in. There are as many ways to study gene function as there are scientists interested in studying it. The techniques

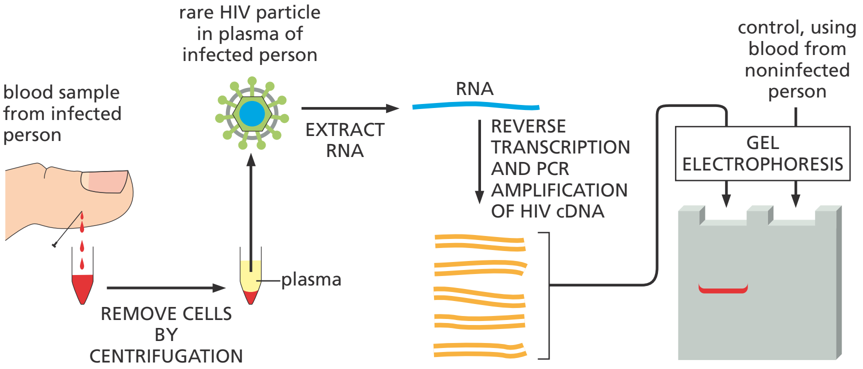

rare HIV particle in plasma of infected person
blood sample from infected person
EXTRACT RNA
RNA
REVERSE TRANSCRIPTION AND PCR AMPLIFICATION OF HIV cDNA
control, using blood from noninfected person
GEL ELECTROPHORESIS
plasma
REMOVE CELLS BY CENTRIFUGATION

Figure 10–17 PCR can be used to detect the presence of a viral genome in a sample of blood. Because of its ability to amplify enormously the signal from every single molecule of nucleic acid, PCR is an extraordinarily sensitive method for detecting trace amounts of virus in a sample of blood or tissue without the need to purify the virus. For HIV, the virus that causes AIDS, the genome is a single-stranded molecule of RNA, as illustrated here. In addition to HIV, many other viruses that infect humans are now detected in this way.

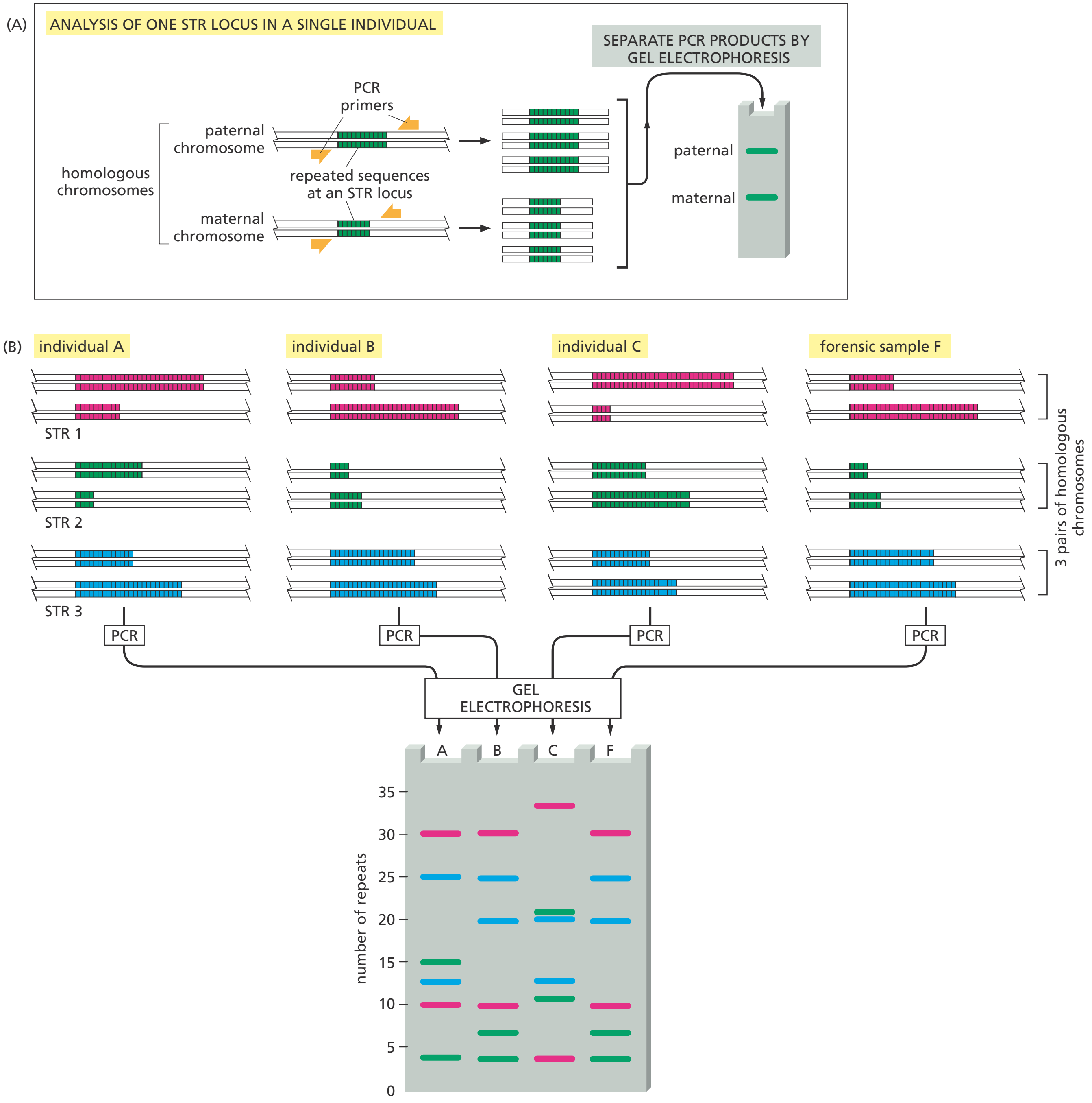


Figure 10–18 PCR is used in forensic science to distinguish one individual from another. The DNA sequences analyzed are short tandem repeats (STRs) composed of sequences such as CACACA… or GTGTGT…. STRs are found in various positions (loci) in the human genome. The number of repeats in each STR locus is highly variable in the population, ranging from 4 to 40 in different individuals. Because of the variability in these sequences, individuals will usually inherit a different number of repeats at each STR locus from their mother and from their father; two unrelated individuals, therefore, rarely contain the same pair of sequences at a given STR locus. (A) PCR using primers that recognize unique sequences on either side of one particular STR locus produces a pair of bands of amplified DNA from each individual, one band representing the maternal STR variant and the other representing the paternal STR variant. The length of the amplified DNA, and thus its position after gel electrophoresis, will depend on the exact number of repeats at the locus. (B) In the schematic example shown here, the same three STR loci are analyzed in samples from three suspects (individuals A, B, and C), producing six bands for each individual. Although different people can have several bands in common, the overall pattern is quite distinctive for each person. The band pattern can therefore serve as a *DNA fingerprint* to identify an individual nearly uniquely. The fourth lane (F) contains the products of the same PCR amplifications carried out on a hypothetical forensic DNA sample, which could have been obtained from a single hair or a tiny spot of blood left at a crime scene.

The more loci that are examined, the more confident one can be about the results. When examining the variability at 5–10 different STR loci, the odds that two random individuals would share the same fingerprint by chance are approximately one in 10 billion. In the case shown here, individuals A and C can be eliminated from inquiries, while B is a clear suspect. A similar approach is now used routinely in paternity testing.

an investigator chooses often depend on his or her background and training: a geneticist might, for example, engineer mutant organisms in which the activity of the gene has been disrupted, whereas a biochemist might take the same gene and produce large amounts of its protein to determine its three-dimensional structure.

In this section, we present a few of the methods that investigators currently use to study the function of a gene—all of which depend on recombinant DNA technology. Because a gene's activity is specified by its nucleotide sequence, we begin by outlining the techniques used to determine—and begin to interpret—the nucleotide sequence of a stretch of DNA. We then explore a variety of approaches for investigating when and where a gene is expressed. We describe how disrupting the activity of a gene in a cell, tissue, or whole plant or animal can provide insights into what that gene normally does. Finally, we explain how recombinant DNA technology can be harnessed to produce large amounts of any protein. Together, the methods we discuss have revolutionized all aspects of cell biology.

Whole Genomes Can Be Sequenced Rapidly

In the late 1970s, researchers developed several schemes for determining, simply and quickly, the nucleotide sequence of any purified DNA fragment. The one that became the most widely used is called **dideoxy sequencing** or **Sanger sequencing** (after the scientist who invented it). The technique uses DNA polymerase, along with special chain-terminating nucleotides called dideoxyribonucleoside triphosphates (**Figure 10–19**), to make partial copies of the DNA fragment to be sequenced. It ultimately produces a collection of different DNA copies that terminate at every position in the original DNA sequence.

Until recently, these DNA copies, which differ in length by a single nucleotide, would then be separated by gel electrophoresis, and the nucleotide sequence of the original DNA would be determined manually from the order of labeled DNA fragments in the gel (**Figure 10–20**). These days, however, Sanger sequencing is fully automated: robotic devices mix the reagents—including the four different chain-terminating dideoxynucleotides, each tagged with a different-colored fluorescent dye—and load the reaction samples onto long, thin capillary gels, which have replaced the flat gel slabs used since the 1970s. A detector then records the color of each band in the gel, and a computer translates the information into a nucleotide sequence (**Figure 10–21**). How such sequence information is then analyzed to assemble a complete genome sequence—for example, the first draft of the human genome—is described in **How We Know**, pp. 344–345.

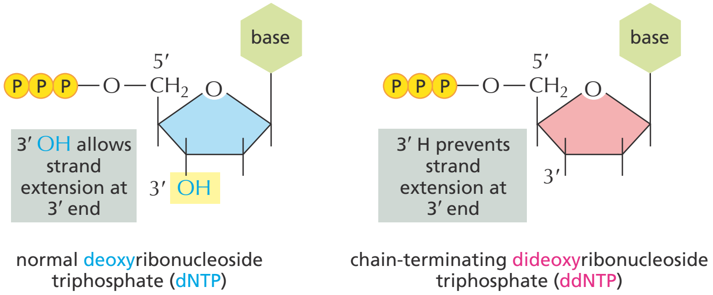

base
5′
P P P
O
CH_2
O
3′ OH allows strand extension at 3′ end
3′ OH
normal deoxyribonucleoside triphosphate (dNTP)
base
5′
P P P
O
CH_2
O
3′ H prevents strand extension at 3′ end
3′
chain-terminating dideoxyribonucleoside triphosphate (ddNTP)

Figure 10–19 The dideoxy, or Sanger, method of sequencing DNA relies on chain-terminating dideoxynucleoside triphosphates (ddNTPs). These ddNTPs are derivatives of the normal deoxyribonucleoside triphosphates that lack the 3′ hydroxyl group. When incorporated into a growing DNA strand, they block further elongation of that strand.

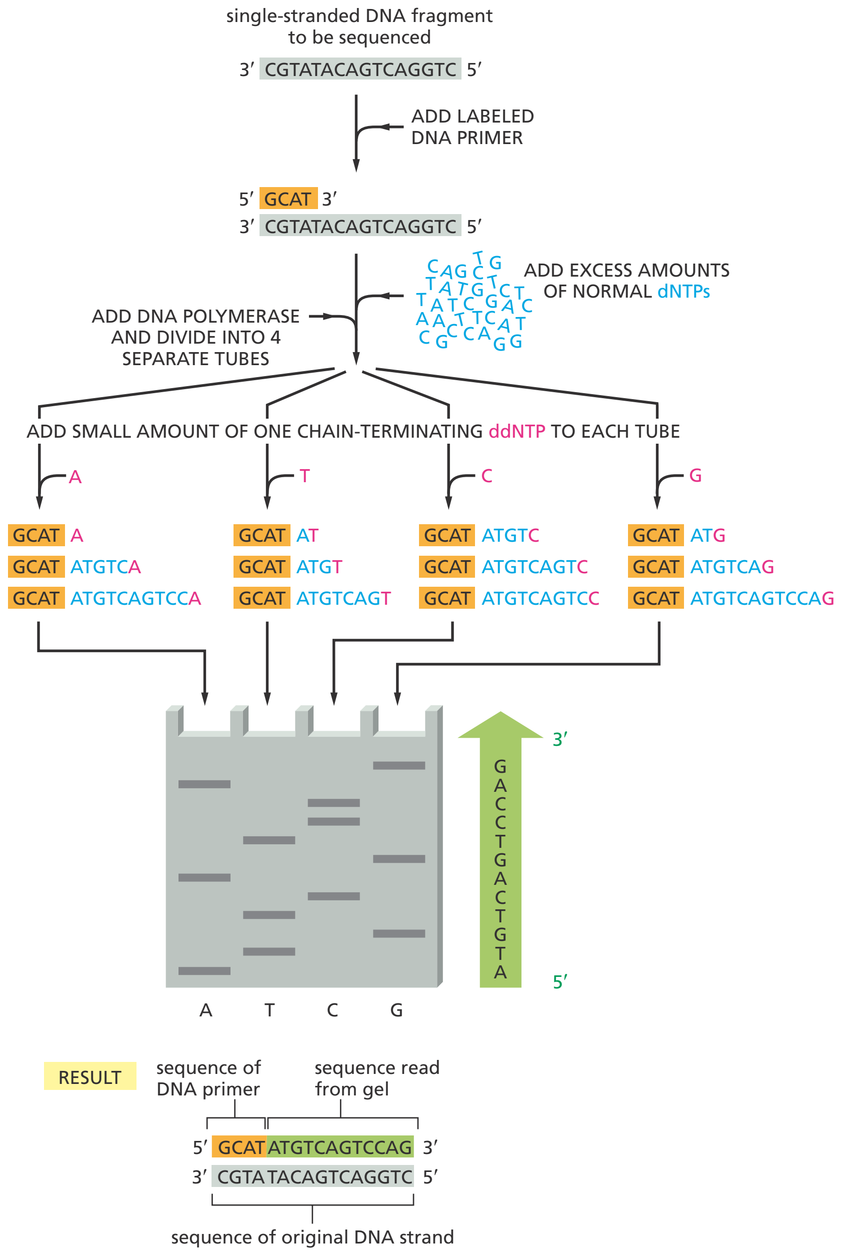


Figure 10–20 The Sanger method produces four sets of labeled DNA molecules. To determine the complete sequence of a single-stranded fragment of DNA (*gray*), the DNA is first hybridized with a short DNA primer (*orange*) that is labeled with a fluorescent dye or radioisotope. DNA polymerase and an excess of all four normal deoxyribonucleoside triphosphates (*blue* A, C, G, and T) are added to the primed DNA, which is then divided into four reaction tubes. Each of these tubes receives a small amount of a single chain-terminating dideoxyribonucleoside triphosphate (*red* A, C, G, or T). Because the chain-terminating ddNTPs will be incorporated only occasionally, each reaction produces a set of DNA copies that terminate at different points in the sequence. The products of these four reactions are separated by electrophoresis in four parallel lanes of a polyacrylamide gel (labeled here A, T, C, and G). In each lane, the bands represent fragments that have terminated at a given nucleotide (e.g., A in the leftmost lane) but at different positions in the DNA. By reading off the bands in order, starting at the bottom of the gel and reading across all lanes, the DNA sequence of the newly synthesized strand can be determined. The sequence, which is given in the *green arrow* to the right of the gel, is complementary to the sequence of the original *gray* single-stranded DNA, as shown on the bottom.

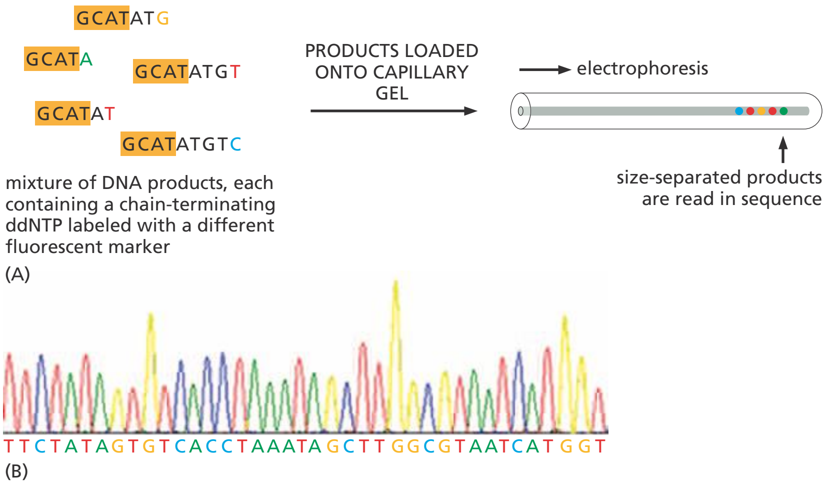

GCATATG
GCATA
GCATATGT
GCATAT
GCATATGTC
mixture of DNA products, each containing a chain-terminating ddNTP labeled with a different fluorescent marker
(A)
PRODUCTS LOADED ONTO CAPILLARY GEL
electrophoresis
0
size-separated products are read in sequence
TTCTATAGTGTCACCTAAATAGCTTGGCGTAATCATGGT
(B)

Figure 10–21 Fully automated machines can set up and run Sanger sequencing reactions. (A) The automated method uses an excess amount of normal dNTPs plus a mixture of four different chain-terminating ddNTPs, each of which is labeled with a fluorescent tag of a different color. The reaction products are loaded onto a long, thin capillary gel and separated by electrophoresis. A camera reads the color of each band on the gel and feeds the data to a computer that assembles the sequence (not shown). (B) A tiny part of the data from such an automated sequencing run. Each colored peak represents a nucleotide in the DNA sequence.

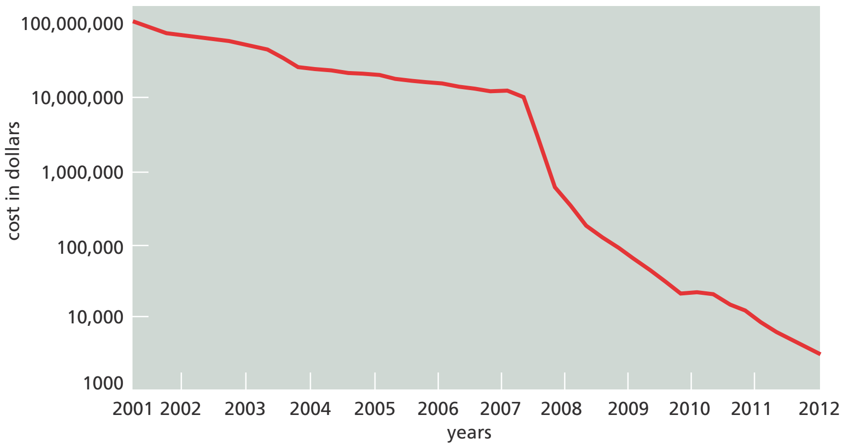


Figure 10–22 The cost of DNA sequencing has dropped precipitously since the advent of next-generation sequencing technologies. Shown here are the costs of sequencing a human genome which was $100 million in 2001 and not much more than a thousand dollars by the end of 2012. (Data from the National Human Genome Research Initiative.)

Next-Generation Sequencing Techniques Make Genome Sequencing Faster and Cheaper

The Sanger method has made it possible to sequence the genomes of humans and of many other organisms including most of those discussed in this book. But newer methods, developed since 2005, have made genome sequencing even more rapid—and very much cheaper. With these so-called *second-generation sequencing methods*, the cost of sequencing DNA has plummeted (**Figure 10–22**). At the same time, the number of genomes that have been sequenced has skyrocketed. These rapid methods allow multiple genomes to be sequenced in parallel in a matter of weeks, enabling investigators to examine thousands of human genomes, catalog the variation in nucleotide sequences from people around the world, and uncover the mutations that increase the risk of various diseases—from cancer to autism—as we discuss in Chapter 19.

Although each method differs in detail, most rely on PCR amplification of a random collection of DNA fragments attached to a solid support, such as a glass slide or a microwell plate. For each fragment, the amplification generates a “cluster” that contains about 1000 copies of an individual DNA fragment. These clusters—tens of millions of which can fit on a single slide or plate—are then sequenced at the same time (**Figure 10–23**).

Even more remarkable are the newest, *third-generation sequencing methods*, which permit the sequencing of just a single molecule of DNA. In one of these techniques, for example, each DNA molecule is slowly pulled through a very tiny channel, like thread through the eye of a needle. Because each of the four nucleotides has a different, characteristic shape, the way a nucleotide obstructs the pore as it passes through reveals

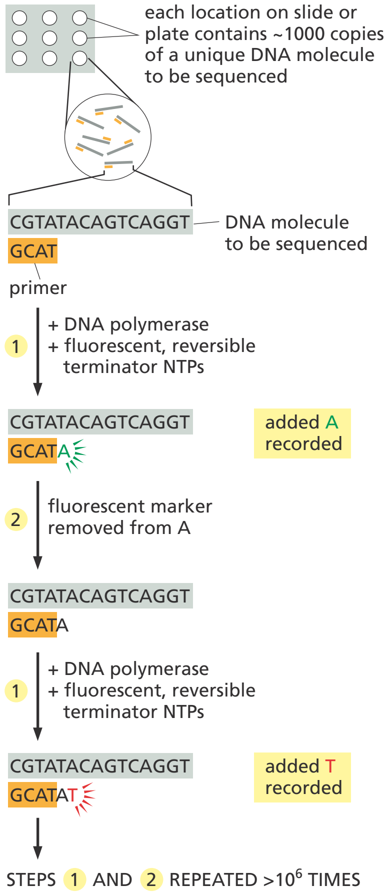

each location on slide or plate contains ~1000 copies of a unique DNA molecule to be sequenced
CGTATACAGTCAGGT
GCAT
DNA molecule to be sequenced
primer
1
+ DNA polymerase
+ fluorescent, reversible terminator NTPs
CGTATACAGTCAGGT
GCATA
added A recorded
2
fluorescent marker removed from A
CGTATACAGTCAGGT
GCATA
1
+ DNA polymerase
+ fluorescent, reversible terminator NTPs
CGTATACAGTCAGGT
GCATAT
added T recorded
STEPS 1 AND 2 REPEATED >10^6 TIMES

Figure 10–23 Second-generation sequencing methods rely on massively parallel sequencing reactions carried out on clusters of PCR-amplified DNA. Each spot on a slide or plate contains about a thousand copies of a single DNA fragment. In the first step, the plate is incubated with DNA polymerase and a special set of four nucleoside triphosphates (NTPs) that terminate DNA synthesis in a reversible manner, each of which carries a fluorescent marker of a different color; no normal dNTPs are present. A camera then images and records the fluorescence at each position on the plate. In the second step, the DNA is chemically treated to remove the fluorescent markers and chemical blockers from each nucleoside; strand synthesis then continues after a new batch of fluorescent NTPs is added. These steps are repeated until the sequence is complete. The snapshots of each round of synthesis are compiled by computer to yield the sequence of the cluster of fragments located at each of the potentially millions of positions on the plate.

SEQUENCING THE HUMAN GENOME

When DNA sequencing techniques became fully automated, determining the order of the nucleotides in a piece of DNA went from being an elaborate Ph.D. thesis project to a routine laboratory chore. Feed DNA into the sequencing machine, add the necessary reagents, and out comes the sought-after result: the order of As, Ts, Gs, and Cs. Nothing could be simpler.

So why was sequencing the human genome such a formidable task? Largely because of its size. The DNA sequencing methods employed at the time were limited by the physical size of the gel used to separate the labeled fragments (see Figure 10–20). At most, only a few hundred nucleotides could be read from a single gel. How, then, do you handle a genome that contains billions of nucleotide pairs?

The solution is to break the genome into fragments and sequence these smaller pieces. The main challenge then comes in piecing the short fragments together in the correct order to yield a comprehensive sequence of a whole chromosome, and ultimately a whole genome. There are two main strategies for accomplishing this genomic breakage and reassembly: the shotgun method and the clone-by-clone approach.

Shotgun sequencing

The most straightforward approach to sequencing a genome is to break it into random fragments, separate and sequence each of the single-stranded fragments, and then use a powerful computer to order these pieces using sequence overlaps to guide the assembly (**Figure 10–24**). This approach is called the shotgun sequencing strategy. As an analogy, imagine shredding several copies of *Essential Cell Biology* (*ECB*), mixing up the pieces, and then trying to put one whole copy of the book back together again by matching up the words or phrases or sentences that appear on each piece. (Several copies would be needed to generate enough overlap for reassembly.) It could be done, but it would be much easier if the book were, say, only two pages long.

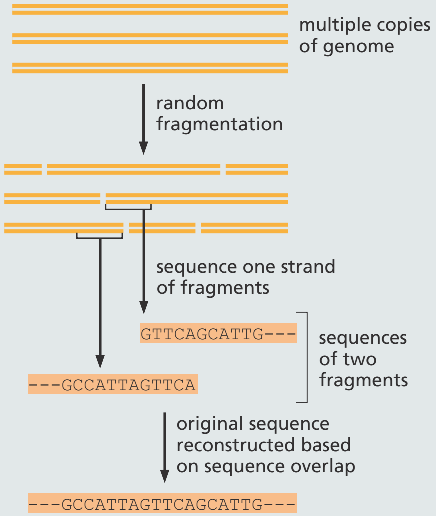

multiple copies of genome
random fragmentation
sequence one strand of fragments
GTTCAGCATTG---
---GCCATTAGTTCA
sequences of two fragments
original sequence reconstructed based on sequence overlap
---GCCATTAGTTCAGCATTG---

Figure 10–24 Shotgun sequencing is the method of choice for small genomes. The genome is first broken into much smaller, overlapping fragments. Each fragment is then sequenced, and the genome is assembled based on overlapping sequences.

For this reason, a straight-out shotgun approach is the strategy of choice only for sequencing small genomes. The method proved its worth in 1995, when it was used to sequence the genome of the infectious bacterium *Haemophilus influenzae*, the first organism to have its complete genome sequence determined. The trouble with shotgun sequencing is that the reassembly process can be derailed by repetitive nucleotide sequences. Although rare in bacteria, these sequences make up a large fraction of vertebrate genomes (see Figure 9–33). Highly repetitive DNA segments make it difficult to piece DNA sequences back together accurately (**Figure 10–25**). Returning to the *ECB* analogy, this chapter alone contains more than a few instances of the phrase "the human genome." Imagine that one slip of paper from the shredded *ECB*s contains the information: "So why was sequencing the human genome" (which appears at the start of this section); another contains the information: "the human genome sequence consortium combined shotgun sequencing with a clone-by-clone approach" (which appears below). You might be tempted to join these two segments together based on the overlapping phrase "the human genome." But you would wind up with the nonsensical statement: "So why was sequencing the human genome sequence consortium combined shotgun sequencing with a clone-by-clone approach." You would also lose the several paragraphs of important text that originally appeared between these two instances of "the human genome."

And that's just in this section. The phrase "the human genome" appears in many chapters of this book. Such repetition compounds the problem of placing each fragment in its correct context. To circumvent these assembly problems, researchers in the human genome sequence consortium combined shotgun sequencing with a clone-by-clone approach.

Clone-by-clone

In this approach, researchers started by preparing a genomic DNA library. They broke the human genome into overlapping fragments, 100–200 kilobase pairs in size. They then plugged these segments into bacterial artificial chromosomes (BACs) and inserted them into *E. coli.* (BACs are similar to the bacterial plasmids discussed earlier, except they can carry much larger pieces of DNA.) As the bacteria divided, they copied the BACs,

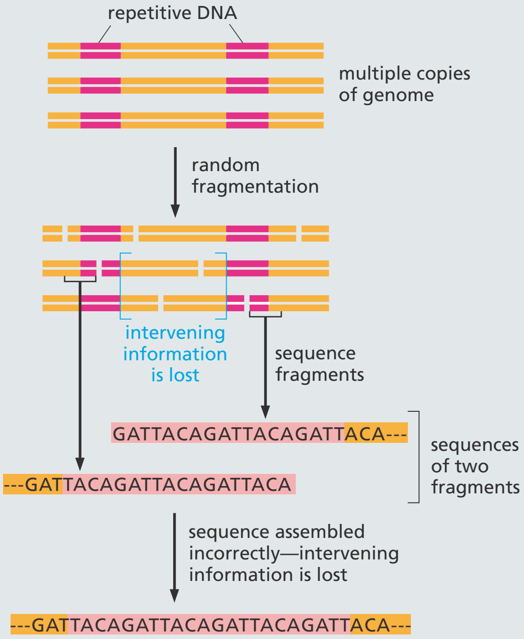

repetitive DNA
multiple copies of genome
random fragmentation
intervening information is lost
sequence fragments
GATTACAGATTACAGATTACA---
---GATTACAGATTACAGATTACA
sequences of two fragments
sequence assembled incorrectly—intervening information is lost
---GATTACAGATTACAGATTACAGATTACA---

Figure 10–25 Repetitive DNA sequences in a genome make it difficult to accurately assemble its fragments. In this example, the DNA contains two segments of repetitive DNA, each made of many copies of the sequence GATTACA. When the resulting sequences are examined, two fragments from different parts of the DNA appear to overlap. Assembling these sequences incorrectly would result in a loss of the information (in brackets) that lies between the original repeats.

thus producing a collection of overlapping cloned fragments (see Figure 10–10).

The researchers then determined where each of these DNA fragments fit into the existing map of the human genome. To do this, different restriction nucleases were used to cut each clone to generate a unique restriction-site "signature." The locations of the restriction sites in each fragment allowed researchers to map each BAC clone onto a restriction map of a whole human genome that had been generated previously using the same set of restriction nucleases (**Figure 10–26**).

Knowing the relative positions of the cloned fragments, the researchers then selected some 30,000 BACs, sheared each into smaller fragments, and determined the nucleotide sequence of each BAC separately using the shotgun method. They could then assemble the whole genome sequence by stitching together the sequences of thousands of individual BACs that span the length of the genome.

The beauty of this approach was that it was relatively easy to accurately determine where the BAC fragments belong in the genome. This mapping step reduces the likelihood that regions containing repetitive sequences will be assembled incorrectly, and it virtually eliminates the possibility that sequences from different chromosomes will be mistakenly joined together. Returning to the textbook analogy, the BAC-based approach is akin to first separating your copies of *ECB* into individual pages and then shredding each page into its own separate pile. It should be much easier to put the book back together when one pile of fragments contains words from page 1, a second pile from page 2, and so on. And there's virtually no chance of mistakenly sticking a sentence from page 40 into the middle of a paragraph on page 412.

All together now

The clone-by-clone approach produced the first draft of the human genome sequence in 2000 and the completed sequence in 2004. As the set of instructions that specify all of the RNA and protein molecules needed to build a human being, this string of genetic bits holds the secrets to human development and physiology. But the sequence was also of great value to researchers interested in comparative genomics or in the physiology of other organisms: it eased the assembly of nucleotide sequences from other mammalian genomes—mice, rats, dogs, and other primates. It also made it much easier to determine the nucleotide sequences of the genomes of individual humans by providing a framework on which the new sequences could be simply superimposed.

The first human sequence was the only mammalian genome completed in this methodical way. But the human genome project was an unqualified success in that it provided the techniques, confidence, and momentum that drove the development of the next generation of DNA sequencing methods, which are now rapidly transforming all areas of biology.

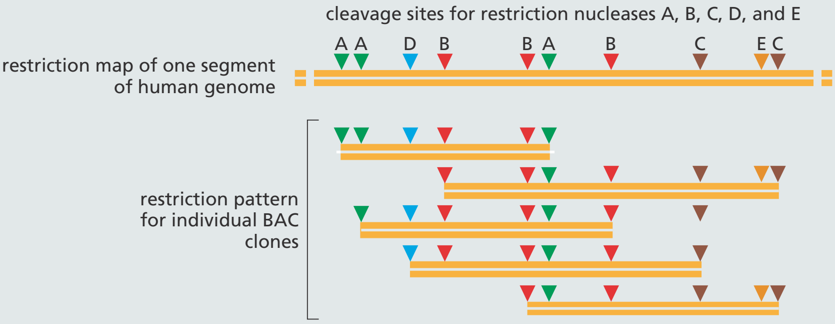

cleavage sites for restriction nucleases A, B, C, D, and E
A A D B B A B C E C
restriction map of one segment of human genome
restriction pattern for individual BAC clones

Figure 10–26 Individual BAC clones are positioned on the physical map of the human genome sequence on the basis of their restriction site "signatures." Clones are digested with five different restriction nucleases, and the sites at which the different enzymes cut each clone are recorded. The distinctive pattern of restriction sites allows investigators to order the fragments and place them on a restriction map of a human genome that had been previously generated using the same nucleases.

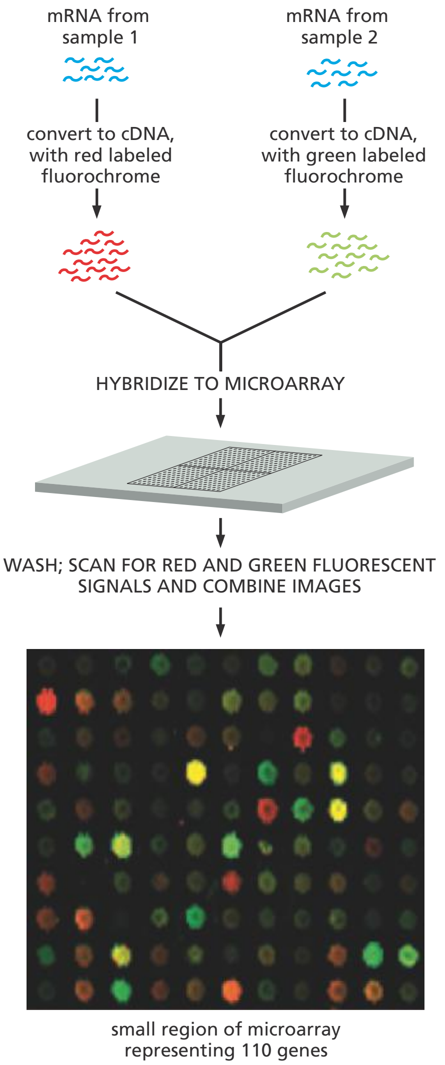

mRNA from sample 1
mRNA from sample 2
convert to cDNA, with red labeled fluorochrome
convert to cDNA, with green labeled fluorochrome
HYBRIDIZE TO MICROARRAY
WASH; SCAN FOR RED AND GREEN FLUORESCENT SIGNALS AND COMBINE IMAGES
small region of microarray representing 110 genes

Figure 10–27 DNA microarrays are used to analyze the production of thousands of different mRNAs in a single experiment. In this example, mRNA is collected from two different cell samples—for example, cells treated with a hormone and untreated cells of the same type—to allow for a direct comparison of the specific genes expressed under both conditions. The mRNAs are converted to cDNAs that are labeled with a red fluorescent dye for one sample, and a green fluorescent dye for the other. The labeled samples are mixed and then allowed to hybridize to the microarray. After incubation, the array is washed and the fluorescence scanned. Only a small proportion of the microarray, representing 110 genes, is shown. *Red* spots indicate that the gene in sample 1 is expressed at a higher level than the corresponding gene in sample 2, and *green* spots indicate the opposite. *Yellow* spots reveal genes that are expressed at about equal levels in both cell samples. The intensity of the fluorescence provides an estimate of how much RNA is present from a gene. *Dark* spots indicate little or no expression of the gene whose fragment is located at that position in the array.

its identity—information that is then used to compile the sequence of the DNA molecule. Such methods require no amplification or chemical labeling, and thereby reduce the cost and time of sequencing even further, making it possible to obtain a complete human genome sequence for under $1000 in hours.

Comparative Genome Analyses Can Identify Genes and Predict Their Function

Strings of nucleotides, at first glance, reveal nothing about how that genetic information directs the development of a living organism—or even what type of organism it might encode. One way to learn something about the function of a particular nucleotide sequence is to compare it with the multitude of sequences available in public databases. Using a computer program to search for sequence similarity, one can determine whether a nucleotide sequence contains a gene and what that gene is likely to do—based on the gene's known activity in other organisms.

Comparative analyses have revealed that the coding regions of genes from a wide variety of organisms show a large degree of sequence conservation (see Figure 9–19). The sequences of noncoding regions, however, tend to diverge over evolutionary time (see Figure 9–18). Thus, a search for sequence similarity can often indicate from which organism a particular piece of DNA was derived, and which species are most closely related. Such information is particularly useful when the origin of a DNA sample is unknown—because it was extracted, for example, from a sample of soil or seawater or the blood of a patient with an undiagnosed infection.

But knowing where a nucleotide sequence comes from—or even what activity it might have—is only the first step toward determining what role it has in the development or physiology of the organism. The knowledge that a particular DNA sequence encodes a transcription regulator, for example, does not reveal when and where that protein is produced, or which genes it might regulate. To learn that, investigators must head back to the laboratory.

Analysis of mRNAs By Microarray or RNA-Seq Provides a Snapshot of Gene Expression

As we discussed in Chapter 8, a cell expresses only a subset of the thousands of genes available in its genome. This subset differs from one cell type to another. One way to determine which genes are being expressed in a population of cells or in a tissue is to analyze which mRNAs are being produced.

The first tool that allowed investigators to analyze simultaneously the thousands of different RNAs produced by cells or tissues was the **DNA microarray**. Developed in the 1990s, DNA microarrays are glass microscope slides that contain hundreds of thousands of DNA fragments, each of which serves as a probe for the mRNA produced by a specific gene. Such microarrays allow investigators to monitor the expression of every gene in an entire genome in a single experiment. To do the analysis, mRNAs are extracted from cells or tissues and converted to cDNAs (see Figure 10–12). The cDNAs are fluorescently labeled and allowed to hybridize to the fragments on the microarray. An automated fluorescence microscope then determines which mRNAs were present in the original sample based on the array positions to which the cDNAs are bound (**Figure 10–27**).

Although microarrays are relatively inexpensive and easy to use, they suffer from one obvious drawback: the sequences of the mRNA samples to be analyzed must be known in advance and represented by a corresponding probe on the array. With the development of next-generation sequencing technologies, investigators increasingly use a more direct approach for cataloging the RNAs produced by a cell. The RNAs are converted to cDNAs, which are then sequenced using second-generation sequencing methods. The approach, called **RNA-Seq**, provides a more quantitative analysis of the *transcriptome*—the complete collection of RNAs produced by a cell under a certain set of conditions. It also determines the number of times a particular sequence appears in a sample and detects rare mRNAs, RNA transcripts that are alternatively spliced, mRNAs that harbor sequence variations, and noncoding RNAs. For these reasons, RNA-Seq is replacing microarrays as the method of choice for analyzing the transcriptome.

In Situ Hybridization Can Reveal When and Where a Gene Is Expressed

Although microarrays and RNA-Seq provide a list of genes that are being expressed by a cell or tissue, they do not reveal exactly where in the cell or tissue those mRNAs are produced. To see where a particular RNA is made, investigators use a technique called ***in situ* hybridization** (from the Latin *in situ*, "in place"), which allows a specific nucleic acid sequence—either DNA or RNA—to be visualized in its normal location.

In situ hybridization uses single-stranded DNA or RNA probes, labeled with either fluorescent dyes or radioactive isotopes, to detect complementary nucleic acid sequences within a tissue, a cell (**Figure 10–28**), or even an isolated chromosome (**Figure 10–29**). The latter application is used in the clinic to determine, for example, whether fetuses carry abnormal chromosomes.

In situ hybridization is frequently used to study the expression patterns of a particular gene or collection of genes in an adult or developing tissue. In one particularly ambitious project, neuroscientists are using the method to assemble a three-dimensional map of all the genes expressed in both the mouse and human brain (**Figure 10–30**). Knowing where and when a gene is expressed can provide important clues about its function.

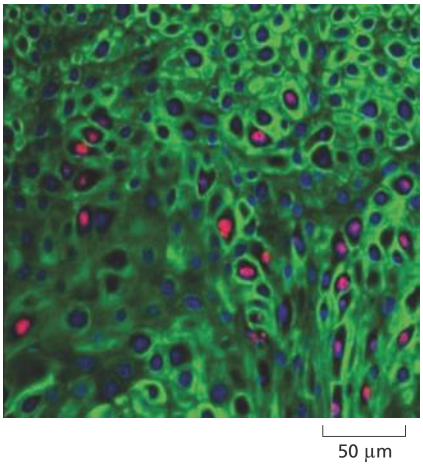


Figure 10–28 *In situ* hybridization can be used to detect the presence of a virus in cells. In this micrograph, the nuclei of cultured epithelial cells infected with the human papillomavirus (HPV) are stained *pink* by a fluorescent probe that recognizes a viral DNA sequence. The cytoplasm of all cells is stained *green*. (Courtesy of Hogne Røed Nilsen.)

Reporter Genes Allow Specific Proteins to be Tracked in Living Cells

For a gene that encodes a protein, the location of the protein within the cell, tissue, or organism yields clues to the gene's function. Traditionally, the most effective way to visualize a protein within a cell or tissue involved using a labeled antibody. That approach requires the generation of an antibody that specifically recognizes the protein of interest—a process that can be time-consuming and has no guarantee of success.

An alternative approach is to use the regulatory DNA sequences of the protein-coding gene to drive the expression of some type of

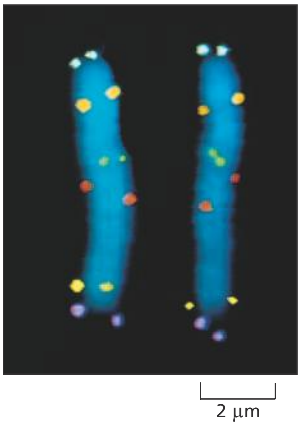


Figure 10–29 *In situ* hybridization can be used to locate genes on isolated chromosomes. Here, six different DNA probes have been used to mark the locations of their respective nucleotide sequences on human Chromosome 5 isolated from a mitotic cell in metaphase (see Figure 5–16 and Panel 18–1, pp. 622–623). The DNA probes have been labeled with different chemical groups and are detected using fluorescent antibodies specific for those groups. Both the maternal and paternal copies of Chromosome 5 are shown, aligned side by side. Each probe produces two dots on each chromosome because chromosomes undergoing mitosis have already replicated their DNA; therefore, each chromosome contains two identical DNA helices. The technique employed here is nicknamed FISH, for *fluorescence* in situ *hybridization*. (Courtesy of David C. Ward.)

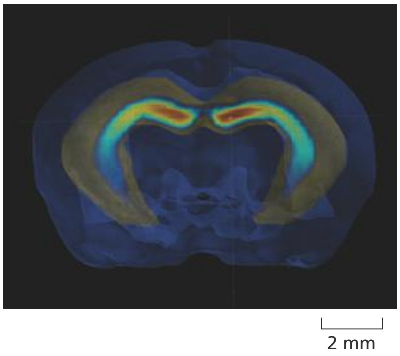


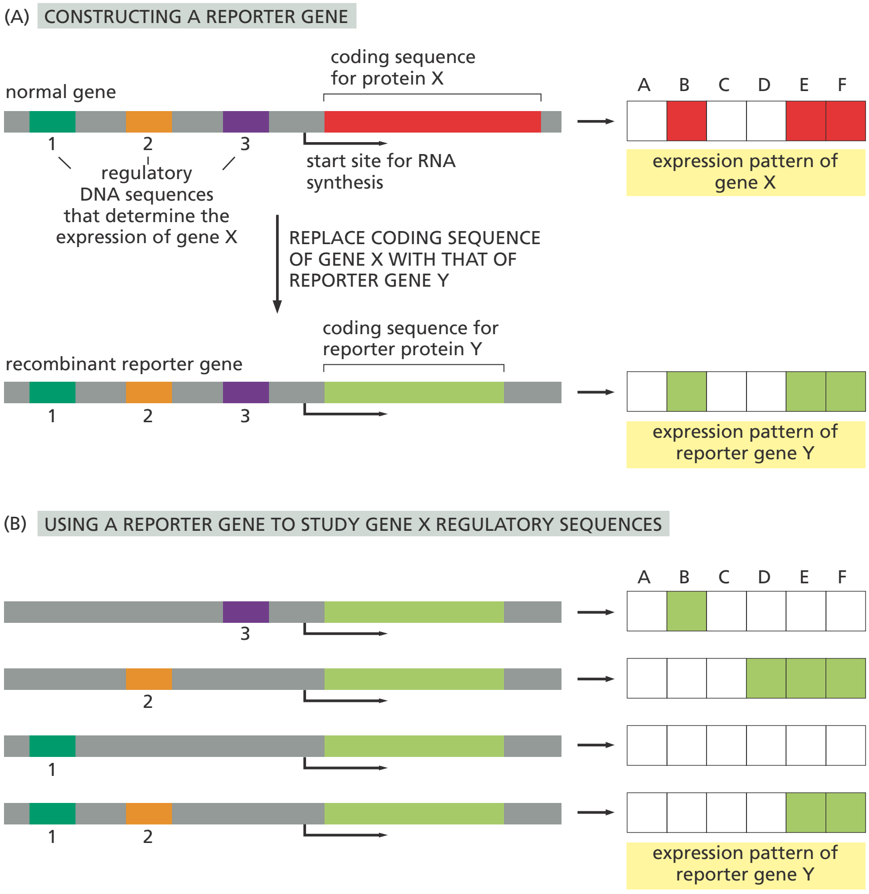
Figure 10–30 *In situ* hybridization has been used to generate an atlas of gene expression in the mouse brain. This computer-generated image shows the expression of genes specific to an area of the brain associated with learning and memory. Similar maps of expression patterns of all known genes in the mouse brain are compiled in the brain atlas project, which is available for free online. (From M. Hawrylycz et al., *PLoS Comput. Biol.* 7:e1001065, 2011.)

reporter gene, one that encodes a protein that can be easily monitored by its fluorescence or enzymatic activity. A recombinant gene of this type usually mimics the expression of the gene of interest, producing the reporter protein when, where, and in the same amounts as the normal protein would be made (**Figure 10–31A**). The same approach can be used to study the regulatory DNA sequences that control the gene's expression (**Figure 10–31B**).

One of the most popular reporter proteins used today is **green fluorescent protein** (**GFP**), the molecule that gives luminescent jellyfish their greenish glow. In many cases, the gene that encodes GFP is simply attached to one end of the gene of interest. The resulting *GFP fusion protein* often behaves in the same way as the normal protein produced by the gene of interest, and its location can be monitored by fluorescence microscopy (**Figure 10–32**). GFP fusion has become a standard strategy for tracking not only the location but also the movement of specific proteins in living cells. In addition, the use of multiple GFP variants that fluoresce at different wavelengths can provide insights into how different cells interact in a living tissue (**Figure 10–33**).

The Study of Mutants Can Help Reveal the Function of a Gene

Although it may seem counterintuitive, one of the best ways to determine a gene's function is to see what happens to an organism when the gene is inactivated by a mutation. Before the advent of gene cloning, geneticists


(A) CONSTRUCTING A REPORTER GENE
normal gene
coding sequence for protein X
1
2
3
regulatory DNA sequences that determine the expression of gene X
start site for RNA synthesis
A B C D E F
expression pattern of gene X
REPLACE CODING SEQUENCE OF GENE X WITH THAT OF REPORTER GENE Y
recombinant reporter gene
coding sequence for reporter protein Y
1
2
3
expression pattern of reporter gene Y
(B) USING A REPORTER GENE TO STUDY GENE X REGULATORY SEQUENCES
A B C D E F
3
2
1
1
2
expression pattern of reporter gene Y

CONCLUSIONS —regulatory sequence 3 turns on gene X in cell B
—regulatory sequence 2 turns on gene X in cells D, E, and F
—regulatory sequence 1 turns off gene X in cell D

Figure 10–31 Reporter genes can be used to determine the pattern of a gene's expression. (A) Suppose the goal is to find out which cell types (A–F) express protein X, but it is difficult to detect the protein directly—with antibodies, for example. Using recombinant DNA techniques, the coding sequence for protein X can be replaced with the coding sequence for reporter protein Y, which can be easily monitored visually; two commonly used reporter proteins are the enzyme β-galactosidase (see Figure 8–13C) and green fluorescent protein (GFP, see Figure 10–32). The expression of the reporter protein Y will now be controlled by the regulatory sequences (here labeled 1, 2, and 3) that control the expression of the normal protein X. (B) To determine which regulatory sequences normally control expression of gene X in particular cell types, reporters with various combinations of the regulatory regions associated with gene X can be constructed. These recombinant DNA molecules are then tested for expression after their introduction into the different cell types.

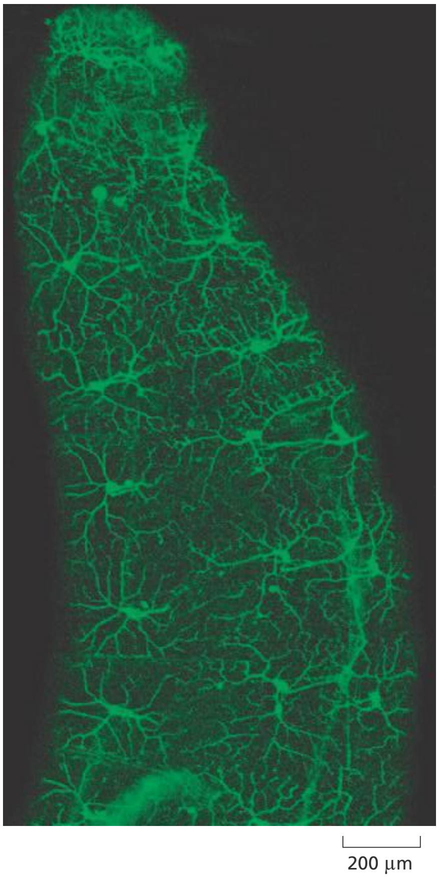


Figure 10–32 Green fluorescent protein (GFP) can be used to identify specific cells in a living animal. For this experiment, carried out in the fruit fly, recombinant DNA techniques were used to join the gene encoding GFP to the regulatory DNA sequences that direct the production of a particular *Drosophila* protein. Both the GFP and the normal fly protein are made only in a specialized set of neurons. This image of a live fly embryo was captured by a fluorescence microscope and shows approximately 20 neurons, each with long projections (axons and dendrites) that communicate with other (nonfluorescent) cells. These neurons, located just under the embryo's surface, allow the organism to sense its immediate environment. (From W.B. Grueber et al., *Curr. Biol.* 13:618–626, 2003. With permission from Elsevier.)

studied the mutant organisms that arise spontaneously in a population. The mutants of most interest were often selected because of their unusual *phenotype*—fruit flies with white eyes or curly wings, for example. The gene responsible for the mutant phenotype could then be studied by breeding experiments, as Gregor Mendel did with peas in the nineteenth century (discussed in Chapter 19).

Although mutant organisms can arise spontaneously, they do so infrequently. The process can be accelerated by treating organisms with either radiation or chemical mutagens, which randomly disrupt gene activity. Such random mutagenesis generates large numbers of mutant organisms, each of which can then be studied individually. This "classical genetic approach," which we discuss in detail in Chapter 19, is most applicable to organisms that reproduce rapidly and can be analyzed genetically in the laboratory—such as bacteria, yeasts, nematode worms, and fruit flies—although it has also been used in zebrafish and mice.

RNA Interference (RNAi) Inhibits the Activity of Specific Genes

Recombinant DNA technology has made possible a more targeted genetic approach to studying gene function. Instead of beginning with a randomly generated mutant and then identifying the responsible gene, a gene of known sequence can be inactivated deliberately and the effects on the cell or organism's phenotype can be observed. Because this strategy is essentially the reverse of that used in classical genetics—which goes from mutants to genes—it is often referred to as *reverse genetics*.

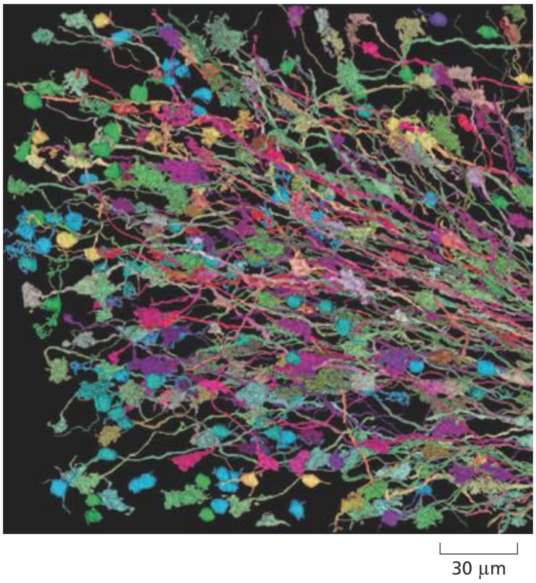


Figure 10–33 GFPs that fluoresce at different wavelengths help reveal the connections that individual neurons make within the brain. This image shows differently colored neurons in one region of a mouse brain. The neurons randomly express different combinations of differently colored GFPs, making it possible to distinguish and trace many individual neurons within a population. The stunning appearance of these labeled neurons have earned these animals the colorful nickname "brainbow mice." (From J. Livet et al., *Nature* 450:56–62, 2007. With permission from Macmillan Publishers Ltd.)

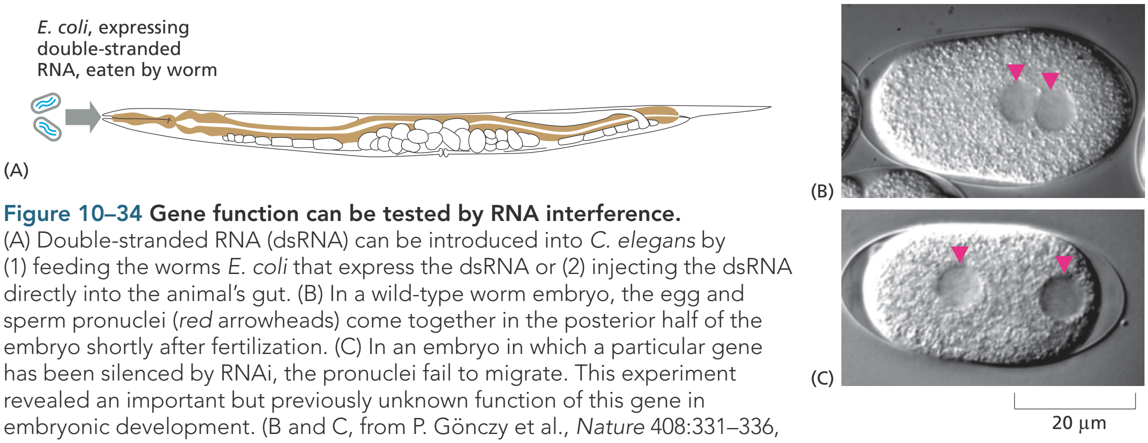


Figure 10–34 Gene function can be tested by RNA interference. (A) Double-stranded RNA (dsRNA) can be introduced into *C. elegans* by (1) feeding the worms *E. coli* that express the dsRNA or (2) injecting the dsRNA directly into the animal's gut. (B) In a wild-type worm embryo, the egg and sperm pronuclei (*red* arrowheads) come together in the posterior half of the embryo shortly after fertilization. (C) In an embryo in which a particular gene has been silenced by RNAi, the pronuclei fail to migrate. This experiment revealed an important but previously unknown function of this gene in embryonic development. (B and C, from P. Gönczy et al., *Nature* 408:331–336, 2000. With permission from Macmillan Publishers Ltd.)

One of the fastest and easiest ways to silence genes in cells and organisms is via **RNA interference** (**RNAi**). Discovered in 1998, RNAi exploits a natural mechanism used in a wide variety of plants and animals to protect themselves against certain viruses and the proliferation of mobile genetic elements (discussed in Chapter 9). The technique involves introducing into a cell or organism double-stranded RNA molecules with a nucleotide sequence that matches the gene to be inactivated. The double-stranded RNA is cleaved and processed by special RNAi machinery to produce shorter, double-stranded fragments called small interfering RNAs (siRNAs). These siRNAs are unwound to form single-stranded RNA fragments that hybridize with the target gene's mRNAs and direct their degradation (see Figure 8–26). In some organisms, the same fragments can direct the production of more siRNAs allowing continued inactivation of the target mRNAs.

RNAi is frequently used to inactivate genes in cultured mammalian cell lines, *Drosophila*, and the nematode *C. elegans*. Introducing double-stranded RNAs into *C. elegans* is particularly easy: the worm can be fed with *E. coli* that have been genetically engineered to produce the double-stranded RNAs that trigger RNAi (**Figure 10–34**). These RNAs get converted into siRNAs, which get distributed throughout the animal's body to inhibit expression of the target gene in various tissues. For the many organisms whose genomes have been completely sequenced, RNAi can, in principle, be used to explore the function of any gene, and large collections of DNA vectors that produce these double-stranded RNAs are available for several species.

A Known Gene Can Be Deleted or Replaced With an Altered Version

Despite its usefulness, RNAi has some limitations. Non-target genes are sometimes inhibited along with the gene of interest, and certain cell types are resistant to RNAi entirely. Even for cell types in which the mechanism functions effectively, gene inactivation by RNAi is often temporary, earning the description "gene knockdown."

Fortunately, there are other, more specific and effective means of eliminating gene activity in cells and organisms. Using recombinant DNA techniques, the coding sequence of a cloned gene can be mutated *in vitro* to change the functional properties of its protein product. Alternatively, the coding region can be left intact and the regulatory region of the gene changed, so that the amount of protein made will be altered or the gene will be expressed in a different type of cell or at a different time during development. By re-introducing this altered gene back into the organism from which it originally came, one can produce a mutant organism that can be studied to determine the gene's function. Often the altered gene is inserted into the genome of reproductive cells so that it can be stably inherited by subsequent generations. Organisms whose genomes have been altered in this way are known as **transgenic organisms**, or *genetically modified organisms* (*GMOs*); the introduced gene is called a *transgene*.

To study the function of a gene that has been altered *in vitro,* ideally one would like to generate an organism in which the normal gene has been replaced by the altered one. In this way, the function of the mutant protein can be analyzed in the absence of the normal protein. A common way of doing this in mice makes use of cultured mouse embryonic stem (ES) cells (discussed in Chapter 20). These cells are first subjected to targeted gene replacement before being transplanted into a developing embryo to produce a mutant mouse, as illustrated in **Figure 10–35**.

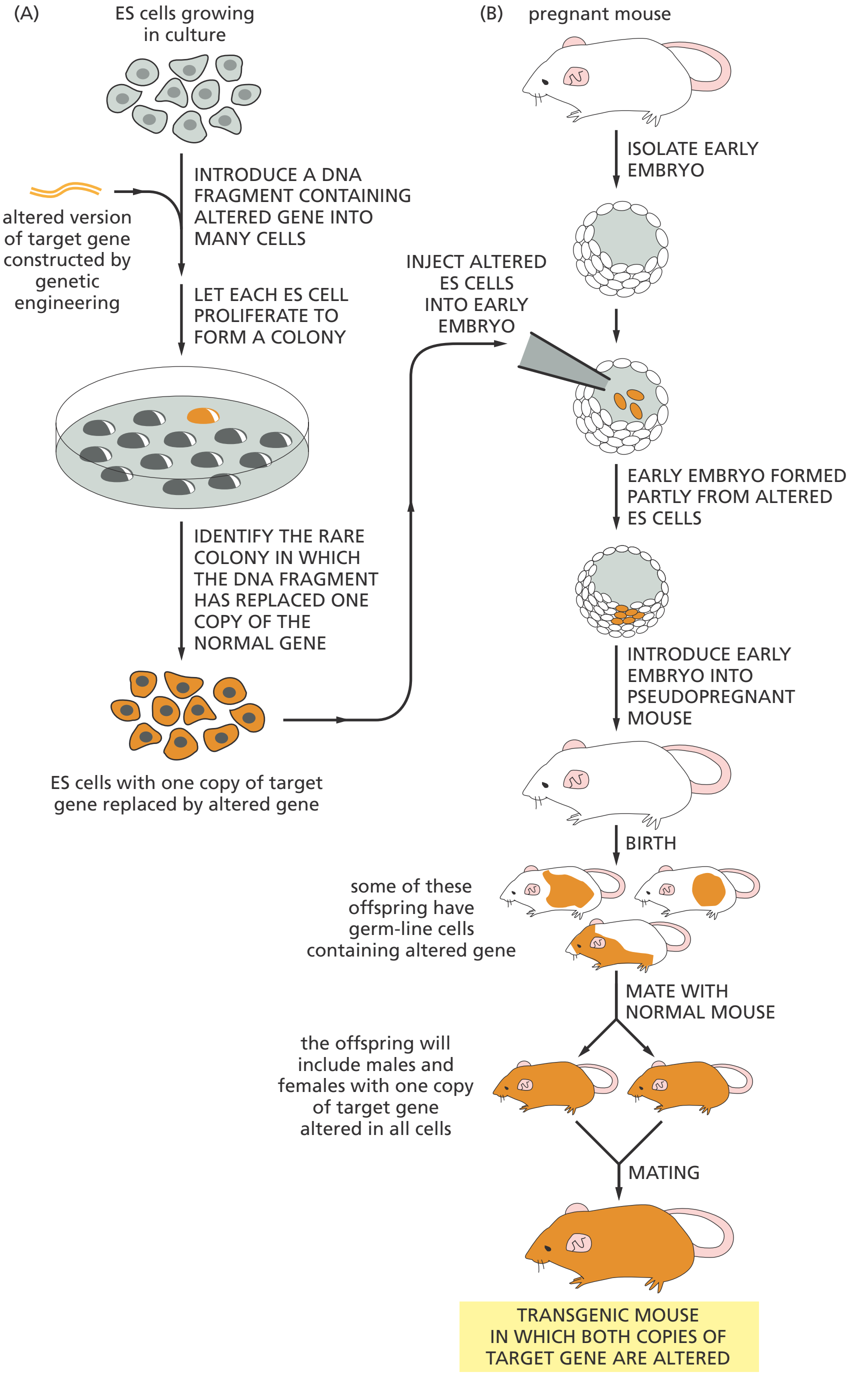

(A)
ES cells growing in culture
altered version of target gene constructed by genetic engineering
INTRODUCE A DNA FRAGMENT CONTAINING ALTERED GENE INTO MANY CELLS
LET EACH ES CELL PROLIFERATE TO FORM A COLONY
IDENTIFY THE RARE COLONY IN WHICH THE DNA FRAGMENT HAS REPLACED ONE COPY OF THE NORMAL GENE
ES cells with one copy of target gene replaced by altered gene
(B)
pregnant mouse
ISOLATE EARLY EMBRYO
INJECT ALTERED ES CELLS INTO EARLY EMBRYO
EARLY EMBRYO FORMED PARTLY FROM ALTERED ES CELLS
INTRODUCE EARLY EMBRYO INTO PSEUDOPREGNANT MOUSE
BIRTH
some of these offspring have germ-line cells containing altered gene
MATE WITH NORMAL MOUSE
the offspring will include males and females with one copy of target gene altered in all cells
MATING
TRANSGENIC MOUSE IN WHICH BOTH COPIES OF TARGET GENE ARE ALTERED

Figure 10–35 Targeted gene replacement in mice utilizes embryonic stem (ES) cells. (A) First, an altered version of the gene is introduced into cultured ES cells. In a few rare ES cells, the altered gene will replace the corresponding normal gene through homologous recombination. Although the procedure is often laborious, these rare cells can be identified and cultured to produce many descendants, each of which carries an altered gene in place of one of its two normal corresponding genes. (B) Next, the altered ES cells are injected into a very early mouse embryo; the cells are incorporated into the growing embryo, which then develops into a mouse that contains some somatic cells (colored *orange*) that carry the altered gene. Some of these mice may also have germ-line cells that contain the altered gene; when bred with a normal mouse, some of the progeny of these mice will contain the altered gene in all of their cells. Such a mouse is called a "knock-in" mouse. If two such mice are bred, one can obtain progeny that contain two copies of the altered gene—one on each chromosome—in all of their cells.

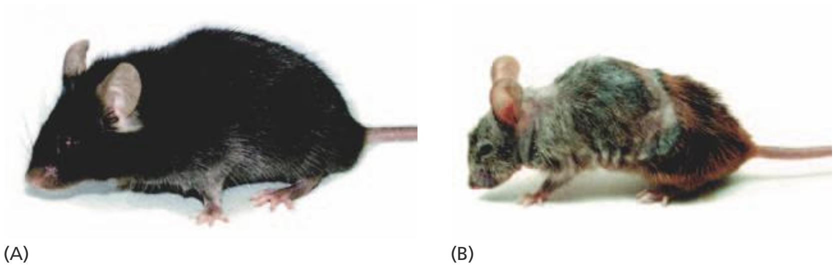


Figure 10–36 Transgenic mice with a mutant DNA helicase show premature aging. The helicase, encoded by the *Xpd* gene, is involved in both transcription and DNA repair. Compared with a wild-type mouse (A), a transgenic mouse that expresses a defective version of *Xpd* (B) exhibits many of the symptoms of premature aging, including osteoporosis, emaciation, early graying, infertility, and reduced life-span. The mutation in *Xpd* used here impairs the activity of the helicase and mimics a human mutation that causes trichothiodystrophy, a disorder characterized by brittle hair, skeletal abnormalities, and a greatly reduced life expectancy. These results support the hypothesis that an accumulation of DNA damage contributes to the aging process in both humans and mice. (From J. de Boer et al., *Science* 296:1276–1279, 2002. With permission from the AAAS.)

Using a similar strategy, the activity of both copies of a gene can also be eliminated entirely, creating a "**gene knockout**." To do this, one can either introduce an inactive, mutant version of the gene into cultured ES cells or delete the gene altogether. The ability to use ES cells to produce such "knockout mice" revolutionized the study of gene function, and the technique is now being used to systematically determine the function of every mouse gene (**Figure 10–36**). A variation of this technique is used to produce *conditional knockout mice*, in which a known gene can be disrupted more selectively—only in a particular cell type or at a certain time in development. Such conditional knockouts are useful for studying genes with a critical function during development, because mice missing these crucial genes often die before birth.

Mutant Organisms Provide Useful Models of Human Disease

Technically speaking, transgenic approaches could be used to alter genes in the human germ line. For ethical reasons, such manipulations are unlawful. But transgenic technologies are widely used to generate animal models of human diseases in which mutant genes play a major part.

With the explosion of DNA sequencing technologies, investigators can rapidly search the genomes of patients for mutations that cause or greatly increase the risk of their disease (discussed in Chapter 19). These mutations can then be introduced into animals, such as mice, that can be studied in the laboratory. The resulting transgenic animals, which often mimic some of the phenotypic abnormalities associated with the condition in patients, can be used to explore the cellular and molecular basis of the disease and to screen for drugs that could potentially be used therapeutically in humans.

An encouraging example is provided by *fragile X syndrome*, a neuropsychiatric disorder associated with intellectual impairment, neurological abnormalities, and often autism. The disease is caused by a mutation in the *fragile X mental retardation gene* (*FMR1*), which encodes a protein that inhibits the translation of mRNAs into proteins at synapses—the junctions where nerve cells communicate with one another (see Figure 12–38). Transgenic mice in which the *FMR1* gene has been disabled show many of the same neurological and behavioral abnormalities seen in patients with the disorder, and drugs that return synaptic protein synthesis to near-normal levels also reverse many of the problems seen in these mutant mice. Preliminary studies suggest that at least one of these drugs may benefit patients with the disease.

Transgenic Plants Are Important for Both Cell Biology and Agriculture

Although we tend to think of recombinant DNA research in terms of animal biology, these techniques have also had a profound impact on the study of plants. In fact, certain features of plants make them especially amenable to recombinant DNA methods.

When a piece of plant tissue is cultured in a sterile medium containing nutrients and appropriate growth regulators, some of the cells are stimulated to proliferate indefinitely in a disorganized manner, producing a mass of relatively undifferentiated cells called a *callus*. If the nutrients and growth regulators are carefully manipulated, one can induce the formation of a shoot within the callus, and in many species a whole new plant can be regenerated from such shoots. In a number of plants—including tobacco, petunia, carrot, potato, and *Arabidopsis*—a single cell from such a callus can be grown into a small clump of cells from which a whole plant can be regenerated (see Figure 8–2B). Just as mutant mice can be derived by the genetic manipulation of embryonic stem cells in culture, so transgenic plants can be created from plant cells transfected with DNA in culture (**Figure 10–37**).

The ability to produce transgenic plants has greatly accelerated progress in many areas of plant cell biology. It has played an important part, for example, in isolating receptors for growth regulators and in analyzing the mechanisms of morphogenesis and of gene expression in plants. These techniques have also opened up many new possibilities in agriculture that could benefit both the farmer and the consumer. They have made it possible, for example, to modify the ratio of lipid, starch, and protein in seeds, to impart pest and virus resistance to plants, and to create modified plants that tolerate extreme habitats such as salt marshes or water-stressed soil. One variety of rice has been genetically engineered to produce β-carotene, the precursor of vitamin A. If it replaced conventional rice, this "golden rice"—so called because of its faint yellow color—could help to alleviate severe vitamin A deficiency, which causes blindness in hundreds of thousands of children in the developing world each year.

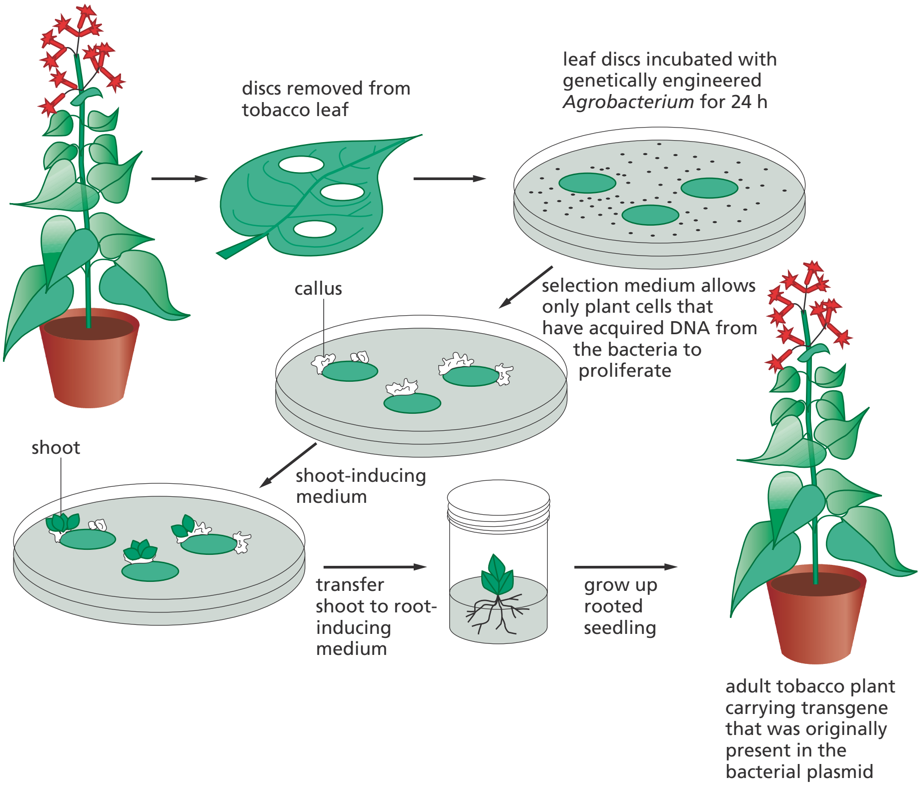

discs removed from tobacco leaf
leaf discs incubated with genetically engineered *Agrobacterium* for 24 h
callus
selection medium allows only plant cells that have acquired DNA from the bacteria to proliferate
shoot
shoot-inducing medium
transfer shoot to root-inducing medium
grow up rooted seedling
adult tobacco plant carrying transgene that was originally present in the bacterial plasmid

Figure 10–37 Transgenic plants can be made using recombinant DNA techniques optimized for plants. A disc is cut out of a leaf and incubated in a culture of *Agrobacterium* that carries a recombinant plasmid with both a selectable marker and a desired genetically engineered gene. The wounded plant cells at the edge of the disc release substances that attract the bacteria, which inject their DNA into the plant cells. Only those plant cells that take up the appropriate DNA and express the selectable marker gene survive and proliferate and form a callus. The manipulation of growth factors supplied to the callus induces it to form shoots, which subsequently root and grow into adult plants carrying the engineered gene.

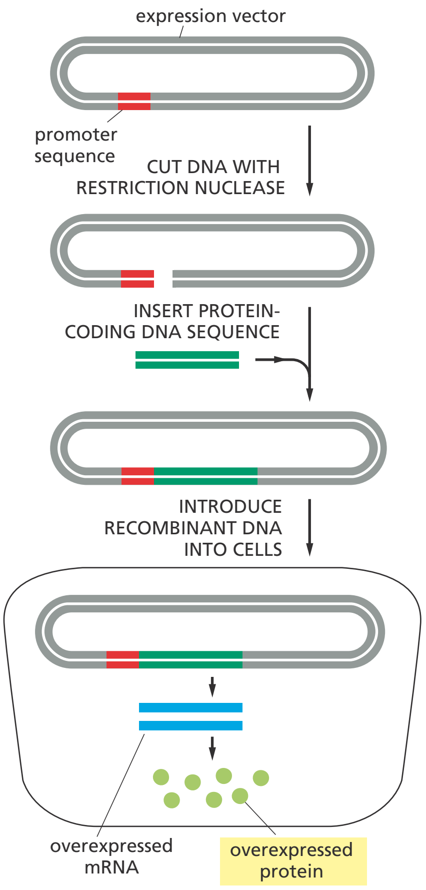

expression vector
promoter sequence
CUT DNA WITH RESTRICTION NUCLEASE
INSERT PROTEIN-CODING DNA SEQUENCE
INTRODUCE RECOMBINANT DNA INTO CELLS
overexpressed mRNA
overexpressed protein

Figure 10–38 Large amounts of a protein can be produced from a protein-coding DNA sequence inserted into an expression vector and introduced into cells. Here, a plasmid vector has been engineered to contain a highly active promoter, which causes unusually large amounts of mRNA to be produced from the inserted protein-coding gene. Depending on the characteristics of the cloning vector, the plasmid is introduced into bacterial, yeast, insect, or mammalian cells, where the inserted gene is efficiently transcribed and translated into protein.

Even Rare Proteins Can Be Made in Large Amounts Using Cloned DNA

One of the most important contributions of DNA cloning and genetic engineering to cell biology is that they make it possible to produce any protein, including the rare ones, in nearly unlimited amounts. Such high-level production is usually accomplished by using specially designed vectors known as *expression vectors*. These vectors include transcription and translation signals that direct an inserted gene to be expressed at very high levels. Different expression vectors are designed for use in bacterial, yeast, insect, or mammalian cells, each containing the appropriate regulatory sequences for transcription and translation in these cells (**Figure 10–38**). The expression vector is replicated at each round of cell division, so that the transfected cells in the culture are able to synthesize very large amounts of the protein of interest—often comprising 1–10% of the total cell protein. It is usually a simple matter to purify this protein away from the other proteins made by the host cell.

This technology is now used to make large amounts of many medically useful proteins, including hormones (such as insulin), growth factors, and viral coat proteins for use in vaccines. Expression vectors also allow scientists to produce many proteins of biological interest in large enough amounts for detailed structural and functional studies that were once impossible—especially for proteins that are normally present in very small amounts, such as some receptors and transcription regulators. Recombinant DNA techniques thus allow scientists to move with ease from protein to gene, and vice versa, so that the functions of both can be explored on multiple fronts (**Figure 10–39**).

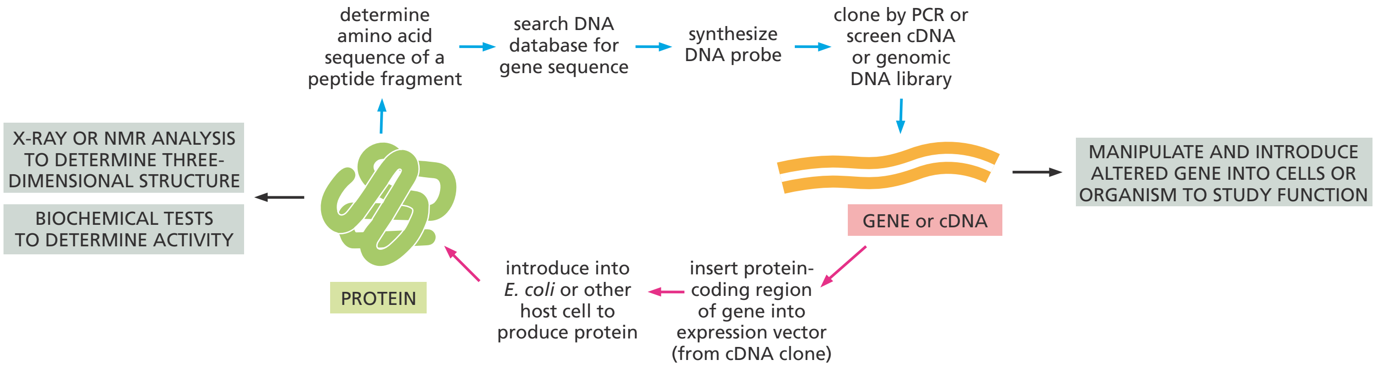

determine amino acid sequence of a peptide fragment
search DNA database for gene sequence
synthesize DNA probe
clone by PCR or screen cDNA or genomic DNA library
X-RAY OR NMR ANALYSIS TO DETERMINE THREE-DIMENSIONAL STRUCTURE
BIOCHEMICAL TESTS TO DETERMINE ACTIVITY
MANIPULATE AND INTRODUCE ALTERED GENE INTO CELLS OR ORGANISM TO STUDY FUNCTION
GENE or cDNA
PROTEIN
introduce into *E. coli* or other host cell to produce protein
insert protein-coding region of gene into expression vector (from cDNA clone)

Figure 10–39 Recombinant DNA techniques make it possible to move experimentally from gene to protein and from protein to gene. A small quantity of a purified protein or peptide fragment is used to obtain a partial amino acid sequence, which is used to search a DNA database for the corresponding nucleotide sequence. This sequence is used to synthesize a DNA probe, which can be used either to pick out the corresponding gene from a DNA library by DNA hybridization (see Figure 10–11) or to clone the gene by PCR from a sequenced genome (see Figure 10–16). Once the gene has been isolated and sequenced, its protein-coding sequence can be inserted into an expression vector to produce large quantities of the protein (see Figure 10–38), which can then be studied biochemically or structurally. In addition to producing protein, the gene or DNA can also be manipulated and introduced into cells or organisms to study its function. (NMR stands for nuclear magnetic resonance; see How We Know, pp. 162–163.)

ESSENTIAL CONCEPTS

- Recombinant DNA technology has revolutionized the study of cells, making it possible to pick out any gene at will from the thousands of genes in a cell and to determine its nucleotide sequence.
- A crucial element in this technology is the ability to cut a large DNA molecule into a specific and reproducible set of DNA fragments using restriction nucleases, each of which cuts the DNA double helix only at a particular nucleotide sequence.
- DNA fragments can be separated from one another on the basis of size by gel electrophoresis.
- Nucleic acid hybridization can detect any given DNA or RNA sequence in a mixture of nucleic acid fragments. This technique depends on highly specific base-pairing between a labeled, single-stranded DNA or RNA probe and another nucleic acid with a complementary sequence.
- DNA cloning techniques enable any DNA sequence to be selected from millions of other sequences and produced in unlimited amounts in pure form.
- DNA fragments can be joined together *in vitro* by using DNA ligase to form recombinant DNA molecules that are not found in nature.
- DNA fragments can be maintained and amplified by inserting them into a larger DNA molecule capable of replication, such as a plasmid. This recombinant DNA molecule is then introduced into a rapidly dividing host cell, usually a bacterium, so that the DNA is replicated at each cell division.
- A collection of cloned fragments of chromosomal DNA representing the complete genome of an organism is known as a genomic library. The library is often maintained as millions of clones of bacteria, each different clone carrying a different fragment of the organism's genome.
- cDNA libraries contain cloned DNA copies of the total mRNA of a particular type of cell or tissue. Unlike genomic DNA clones, cDNA clones contain predominantly protein-coding sequences; they lack introns, regulatory DNA sequences, and promoters. Thus they are useful when the cloned gene is needed to make a protein.
- The polymerase chain reaction (PCR) is a powerful form of DNA amplification that is carried out *in vitro* using a purified DNA polymerase. PCR requires prior knowledge of the sequence to be amplified, because two synthetic oligonucleotide primers must be synthesized that bracket the portion of DNA to be replicated.
- Historically, genes were cloned using hybridization techniques to identify the bacteria carrying the desired sequence in a DNA library. Today, a gene is usually cloned using PCR to specifically amplify it from a sample of DNA or mRNA.
- DNA sequencing techniques have become increasingly fast and cheap, so that the entire genomes of thousands of different organisms are now known, including thousands of individual humans.
- Using recombinant DNA techniques, a protein can be joined to a molecular tag, such as green fluorescent protein (GFP), which allows its movement to be tracked inside a cell and, in some cases, inside a living organism.
- *In situ* nucleic acid hybridization can be used to detect the precise location of genes on chromosomes and of RNAs in cells and tissues.
- DNA microarrays and RNA-Seq can be used to monitor the expression of tens of thousands of genes at once.
- Cloned genes can be altered *in vitro* and stably inserted into the genome of a cell or an organism to study their function. Such mutants are called transgenic organisms.

- The expression of particular genes can be inhibited in cells or organisms by the technique of RNA interference (RNAi), which prevents an mRNA from being translated into protein.
- Bacteria, yeasts, and mammalian cells can be engineered to synthesize large quantities of any protein whose gene has been cloned, making it possible to study proteins that are otherwise rare or difficult to isolate.

KEY TERMS

cDNA
cDNA library
dideoxy (Sanger) DNA sequencing
DNA cloning
DNA library
DNA ligase
DNA microarray
gene knockout
gene replacement
genomic DNA library
green fluorescent protein (GFP)
hybridization
in situ hybridization
plasmid
polymerase chain reaction (PCR)
recombinant DNA
recombinant DNA technology
reporter gene
restriction nuclease
RNA interference (RNAi)
RNA-Seq
transformation
transgenic organism

QUESTIONS

QUESTION 10–5

What are the consequences for a DNA sequencing reaction if the ratio of dideoxyribonucleoside triphosphates to deoxyribonucleoside triphosphates is increased? What happens if this ratio is decreased?

QUESTION 10–6

Almost all the cells in an individual animal contain identical genomes. In an experiment, a tissue composed of several different cell types is fixed and subjected to *in situ* hybridization with a DNA probe to a particular gene. To your surprise, the hybridization signal is much stronger in some cells than in others. How might you explain this result?

QUESTION 10–7

After decades of work, Dr. Ricky M. isolated a small amount of attractase—an enzyme that produces a powerful human pheromone—from hair samples of Hollywood celebrities. To take advantage of attractase for his personal use, he obtained a complete genomic clone of the attractase gene, connected it to a strong bacterial promoter on an expression plasmid, and introduced the plasmid into *E. coli* cells. He was devastated to find that no attractase was produced in the cells. What is a likely explanation for his failure?

QUESTION 10–8

Which of the following statements are correct? Explain your answers.

A. Restriction nucleases cut DNA at specific sites that are always located between genes.

B. DNA migrates toward the positive electrode during electrophoresis.

C. Clones isolated from cDNA libraries contain promoter sequences.

D. PCR utilizes a heat-stable DNA polymerase because for each amplification step, double-stranded DNA must be heat-denatured.

E. Digestion of genomic DNA with AluI, a restriction enzyme that recognizes a four-nucleotide sequence, produces fragments that are all exactly 256 nucleotides in length.

F. To make a cDNA library, both a DNA polymerase and a reverse transcriptase must be used.

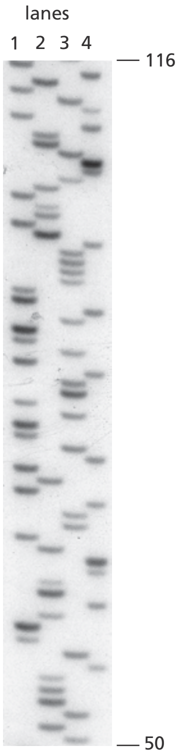
G. DNA fingerprinting by PCR relies on the fact that different individuals have different numbers of repeats in STR regions in their genome.

H. It is possible for a coding region of a gene to be present in a genomic library prepared from a particular tissue but to be absent from a cDNA library prepared from the same tissue.

QUESTION 10–9

A. What is the sequence of the DNA that was used in the sequencing reaction shown in **Figure Q10–9**? The four lanes show the products of sequencing reactions that contained ddG (lane 1), ddA (lane 2), ddT (lane 3), and ddC (lane 4). The numbers to the right of the autoradiograph represent the positions of marker DNA fragments of 50 and 116 nucleotides.

B. This DNA was derived from the middle of a cDNA clone of a mammalian protein. Using the genetic code table (see Figure 7–25), can you determine the amino acid sequence of this portion of the protein?



(Courtesy of Leander Lauffer and Peter Walter.)

Figure Q10–9

QUESTION 10–10

A. How many different DNA fragments would you expect to obtain if you cleaved human genomic DNA with HaeIII? (Recall that there are 3 × 10^9 nucleotide pairs per haploid genome.) How many fragments would you expect with EcoRI?

B. Human genomic libraries used for DNA sequencing are often made from fragments obtained by cleaving human DNA with HaeIII in such a way that the DNA is only partially digested; that is, not all the possible HaeIII sites have been cleaved. What is a possible reason for doing this?

QUESTION 10–11

A molecule of double-stranded DNA was cleaved with restriction nucleases, and the resulting products were separated by gel electrophoresis (**Figure Q10–11**). DNA fragments of known sizes were electrophoresed on the same gel for use as size markers (*left* lane). The size of the DNA markers is given in kilobase pairs (kb), where 1 kb = 1000 nucleotide pairs. Using the size markers as a guide, estimate the length of each restriction fragment obtained. From this information, construct a map of the original DNA molecule indicating the relative positions of all the restriction enzyme cleavage sites.

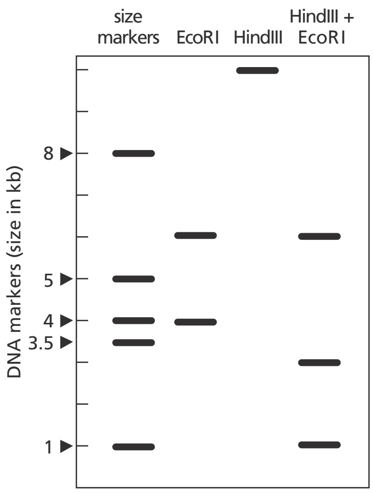


Figure Q10–11

QUESTION 10–12

You have isolated a small amount of a rare protein. You cleaved the protein into fragments using proteases, separated some of the fragments by chromatography, and determined their amino acid sequence. Unfortunately, as is often the case when only small amounts of protein are available, you obtained only three short stretches of amino acid sequence from the protein:

1. Trp-Met-His-His-Lys

2. Leu-Ser-Arg-Leu-Arg

3. Tyr-Phe-Gly-Met-Gln

A. Using the genetic code (see Figure 7–25), design a collection of DNA probes specific for each peptide that could be used to detect the gene in a cDNA library by hybridization. Which of the three collections of oligonucleotide probes would it be preferable to use first? Explain your answer. (Hint: the genetic code is redundant, so each peptide has multiple potential coding sequences.)

B. You have also been able to determine that the Gln of your peptide #3 is the C-terminal (i.e., the final) amino acid of your protein. How would you go about designing oligonucleotide primers that could be used to amplify a portion of the gene from a cDNA library using PCR?

C. Suppose the PCR amplification in (B) yields a DNA that is precisely 300 nucleotides long. Upon determining the nucleotide sequence of this DNA, you find the sequence CTATCACGCCTTAGG approximately in its middle. What would you conclude from these observations?

QUESTION 10–13

Assume that a DNA sequencing reaction is carried out as shown in Figure 10–20, except that the four different dideoxyribonucleoside triphosphates are modified so that each contains a covalently attached dye of a different color (which does not interfere with its incorporation into the DNA chain). What would the products be if you added a mixture of all four of these labeled dideoxyribonucleoside triphosphates along with the four unlabeled deoxyribonucleoside triphosphates into a single sequencing reaction? What would the results look like if you electrophoresed these products in a single lane of a gel?

QUESTION 10–14

Genomic DNA clones are often used to “walk” along a chromosome. In this approach, one cloned DNA is used to isolate other clones that contain overlapping DNA sequences (**Figure Q10–14**). Using this method, it is possible to build up a long stretch of DNA and thus identify new genes in near proximity to a previously cloned gene.

A. Would it be faster to use cDNA clones in this method, because they do not contain any intron sequences?

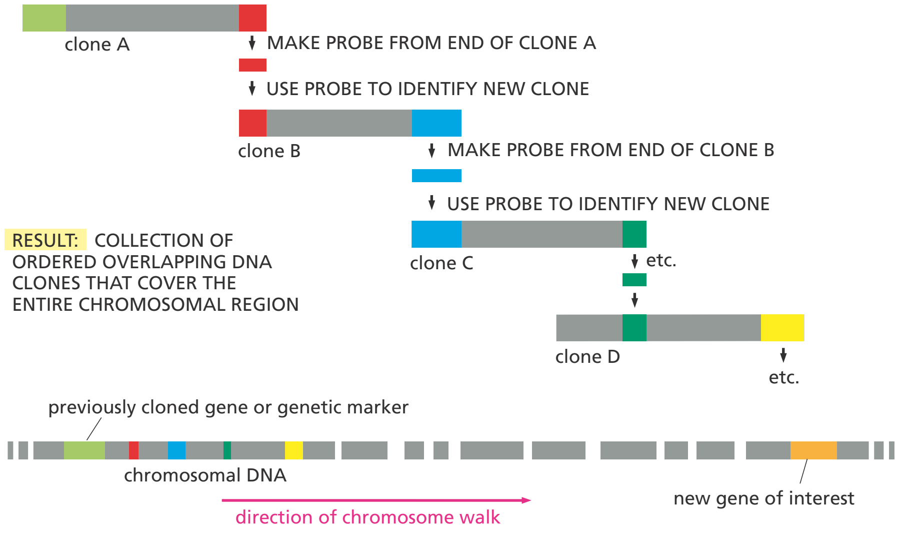

clone A
MAKE PROBE FROM END OF CLONE A
USE PROBE TO IDENTIFY NEW CLONE
clone B
MAKE PROBE FROM END OF CLONE B
USE PROBE TO IDENTIFY NEW CLONE
RESULT: COLLECTION OF ORDERED OVERLAPPING DNA CLONES THAT COVER THE ENTIRE CHROMOSOMAL REGION
clone C
etc.
clone D
etc.
previously cloned gene or genetic marker
chromosomal DNA
direction of chromosome walk
new gene of interest

Figure Q10–14

(Courtesy of John Bedbrook and DNA Plant Technology Corporation.)

Figure Q10–16

B. What would happen if you encountered a repetitive DNA sequence, like the *L1* transposon (see Figure 9–17), which is found in many copies and in many different places in the genome?

QUESTION 10–15

There has been a colossal snafu in the maternity ward of your local hospital. Four sets of male twins, born within an hour of each other, were inadvertently shuffled in the excitement occasioned by that unlikely event. You have been called in to set things straight. As a first step, you would like to match each baby with his twin. (Many newborns look alike so you don't want to rely on appearance alone.) To that end you analyze a small blood sample from each infant using a hybridization probe that detects short tandem repeats (STRs) located in widely scattered regions of the genome. The results are shown in Figure Q10–15.

A. Which infants are twins? Which are identical twins?

B. How could you match a pair of twins to the correct parents?

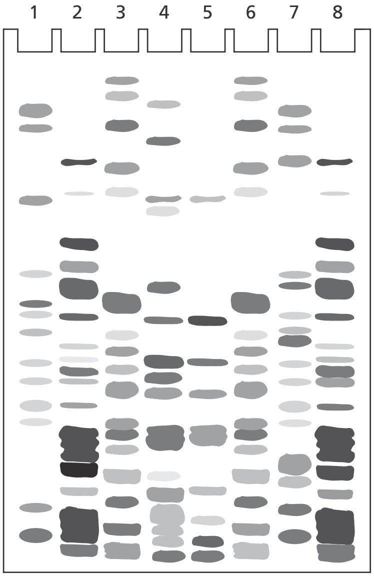


Figure Q10–15

QUESTION 10–16

One of the first organisms that was genetically modified using recombinant DNA technology was a bacterium that normally lives on the surface of strawberry plants. This bacterium makes a protein, called ice-protein, that causes the efficient formation of ice crystals around it when the temperature drops to just below freezing. Thus, strawberries harboring this bacterium are particularly susceptible to frost damage because their cells are destroyed by the ice crystals. Consequently, strawberry farmers have a considerable interest in preventing ice crystallization.

A genetically engineered version of this bacterium was constructed in which the ice-protein gene was knocked out. The mutant bacteria were then introduced in large numbers into strawberry fields, where they displaced the normal bacteria by competition for their ecological niche. This approach has been successful: strawberries bearing the mutant bacteria show a much reduced susceptibility to frost damage.

At the time they were first carried out, the initial open-field trials triggered an intense debate because they represented the first release into the environment of an organism that had been genetically engineered using recombinant DNA technology. Indeed, all preliminary experiments were carried out with extreme caution and in strict containment (Figure Q10–16).

Do you think that bacteria lacking the ice-protein could be isolated without the use of modern DNA technology? Is it likely that such mutations have already occurred in nature? Would the use of a mutant bacterial strain isolated from nature be of lesser concern? Should we be concerned about the risks posed by the application of recombinant DNA techniques in agriculture and medicine? Explain your answers.