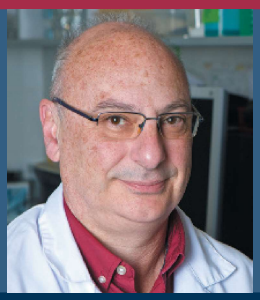
Unit 3 GENETICS



Dr. Francisco Mojica is a Professor of Microbiology at the Universidad de Alicante in Spain, where he studies CRISPR sequences. His parents fostered in him a love of nature, and he realized in high school that he was most interested in biology. In college, when he first looked through a microscope and saw small single-celled "living beings" moving around, he was captivated, and he has spent his career focusing on prokaryotes. While in graduate school in the 1990s, he found the prokaryotic DNA repeats that he later named CRISPR repeats, and he spent the next decade establishing their role as a prokaryotic immune system. Dr. Mojica's lab currently studies the mysteries that still surround CRISPR, trying to understand every single aspect of its mechanism of action and regulation.

AN INTERVIEW WITH

Francisco Mojica

When did you first decide to be a biologist?

In high school, my parents asked me, "What are you thinking you'll do with your life—become a doctor or a lawyer?" I said, "Not at all. I really like biology." In spite of his concern about my future career, my father told me I must study what I loved, so that I could be happy in my life. I really appreciated his words! So I went to university and studied biology. When you like biology, initially you get attracted because of the things you see—animals and plants. But when you see microscopic living beings for the first time, and they're moving around, that's really amazing! So, I decided to focus on microbiology.



▲ Dr. Mojica cultured and analyzed prokaryotes from this salt lake in Spain, leading to his discovery of CRISPR.

What did you study as a graduate student?

I studied prokaryotes that live in salt ponds. These weird organisms can tolerate very high salt concentrations. I was interested in doing molecular biology, which had not been done in that university before—I knew that would be a challenge, and I love challenges. There weren't many genome sequences available at that time, in the early 1990s, but we had cloned a few sequences, and we had indirect evidence they might be related to adaptation to salinity. One of the DNA sequences had these regularly spaced repeats that were partially palindromic (reading the same forward and backward), so they could form secondary structures. We thought maybe these repeats were affected by salinity and would in turn affect gene expression, but we were absolutely wrong! At the end of my Ph.D. thesis, all I could say was, "We have found regularly interspaced repeats that are expressed by the cell, and they could be very important."

"Studying biology is basically just saying 'Wow!' all the time."

What made you continue to study these repeats?

When you find something in your thesis, you believe that it belongs to you—you have to try to answer that question. No one could figure out the role that these repeats played. However, others had found similar repeats in bacteria like *E. coli* and another very distantly-related bacterial species. That was an important finding that pushed me to keep working on the repeats, because it meant the repeats must be doing something important for them to stay in the genomes of various prokaryotes over such long evolutionary time. Evolution is so fundamental to biology, to understanding everything.

How did you figure out what the repeats were doing?

I started wondering about the sequences *between* the repeats, the so-called "spacers." Where did they come from? I compared the spacer sequences to genome database sequences. By 2000 or 2001, many more genome sequences were available. Finally, in 2003, I found a spacer in *E. coli* whose sequence was identical to the sequence of a virus that infects *E. coli*. Even more interesting, the strain of *E. coli* with that spacer could not be infected by that virus. I checked more and more spacer sequences, and when a spacer sequence matched a sequence in the database, the match was always to the sequence of a virus that could no longer infect that species. So I realized these bacteria keep a "memory" of the viruses that infect them, and that the repeat regions were part of an immune system—an unexpected and very surprising finding. I proposed naming the repeats Clustered Regularly Interspaced Short Palindromic Repeats, or CRISPR, and to my surprise, the name took hold.

What is your advice to a student considering a career in biology?

Studying biology is basically just saying "Wow!" all the time. In science, the truth changes very fast. The most important thing is to be open. Be ready to hear something that destroys what you believe—because biology is a fast-moving field, a field that evolves a lot. Knowledge will change.

13 Meiosis and Sexual Life Cycles

KEY CONCEPTS

13.1 Offspring acquire genes from parents by inheriting chromosomes *p. 255*

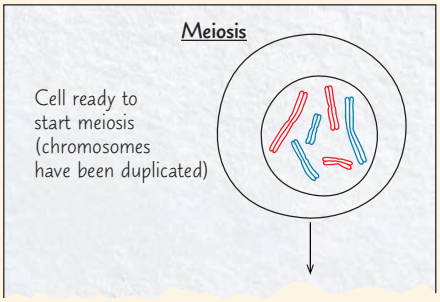
13.2 Fertilization and meiosis alternate in sexual life cycles *p. 256*

13.3 Meiosis reduces the number of chromosome sets from diploid to haploid *p. 259*

13.4 Genetic variation produced in sexual life cycles contributes to evolution *p. 265*

Study Tip

Make a visual study guide: Figure 13.1 presents the events of meiosis as a simplified diagram. As you learn more about meiosis, draw a detailed diagram of the stages of meiosis, starting with the sketch below. Include explanatory labels. Compare your visual study guides of mitosis (from Chapter 12) and meiosis, listing similarities and differences.



Go to Mastering Biology

For Students (in eText and Study Area)
- Figure 13.10 Walkthrough: A Comparison of Mitosis and Meiosis
- BioFlix® Animation: Meiosis
- Animation: Genetic Variation from Crossing Over

For Instructors to Assign (in Item Library)
- Activity: Meiosis Animation
- BioFlix® Tutorial: Meiosis (3 of 3): Determinants of Heredity and Genetic Variation

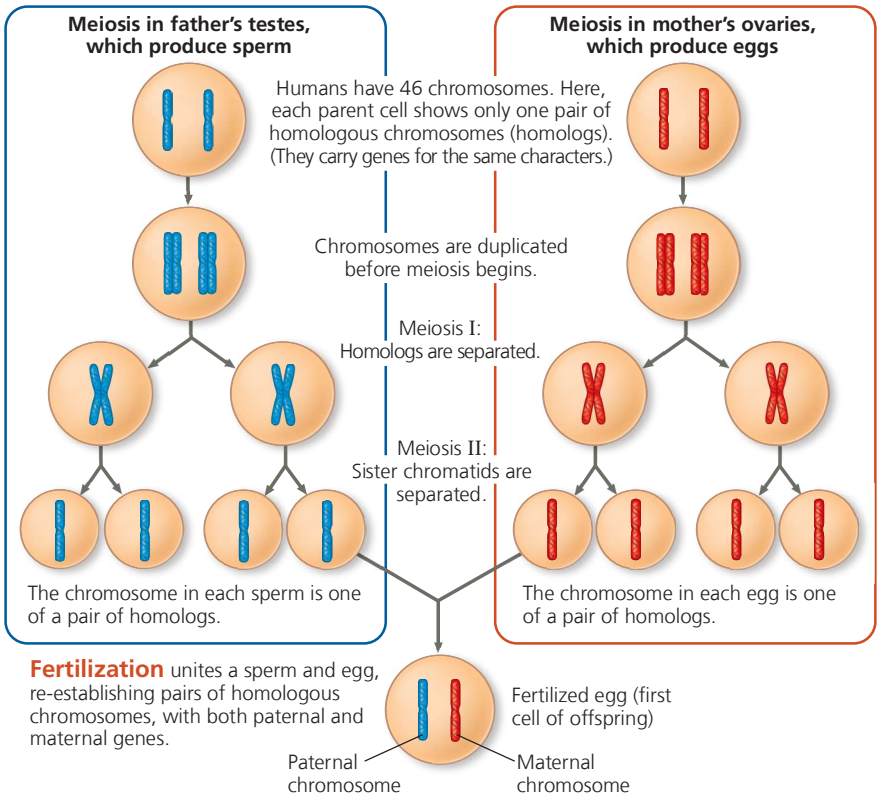
Ready-to-Go Teaching Module (in Instructor Resources)
- Meiosis (Concept 13.3)



Figure 13.1 The family members shown in this photo have some similar features. Offspring resemble their parents more than they do unrelated individuals.

What biological mechanisms account for the resemblance between offspring and their parents?

Meiosis is a special type of cell division that produces cells with half the chromosomes of the parent cell. It occurs only in specialized cells, such as the cells of testes and ovaries in humans.



CONCEPT 13.1

Offspring acquire genes from parents by inheriting chromosomes

Friends may tell you that you have your mother's nose or your father's eyes—although they don't mean that literally, of course. The transmission of traits from one generation to the next is called inheritance, or **heredity** (from the Latin *heres*, heir). At the same time, sons and daughters are not identical copies of either parent or of their siblings. Along with inherited similarity, there is also **variation**. The study of both heredity and inherited variation is called **genetics**.

Inheritance of Genes

Parents endow their offspring with coded information in the form of hereditary units called **genes**. The genes we inherit from our mothers and fathers are our genetic link to our parents, and they account for family resemblances such as shared eye color or freckles. Our genes program specific traits that emerge as we develop from fertilized eggs into adults.

The genetic program is written in the language of DNA, the polymer of four different nucleotides you learned about in Concepts 1.1 and 5.5. Inherited information is passed on in the form of each gene's specific sequence of DNA nucleotides, much as printed information is communicated in the form of meaningful sequences of letters. In both cases, the language is symbolic. Just as your brain translates the word *apple* into a mental image of the fruit, cells translate genes into freckles and other features. Most genes program cells to synthesize specific enzymes and other proteins, whose cumulative action produces an organism's inherited traits. The programming of these traits in the form of DNA is one of the unifying themes of biology.

The transmission of hereditary traits has its molecular basis in the replication of DNA, which produces copies of genes that can be passed from parents to offspring. In animals and plants, reproductive cells called **gametes** are the vehicles that transmit genes from one generation to the next. During fertilization, male and female gametes (sperm and eggs) unite, passing on genes of both parents to their offspring.



▲ Sperm and egg

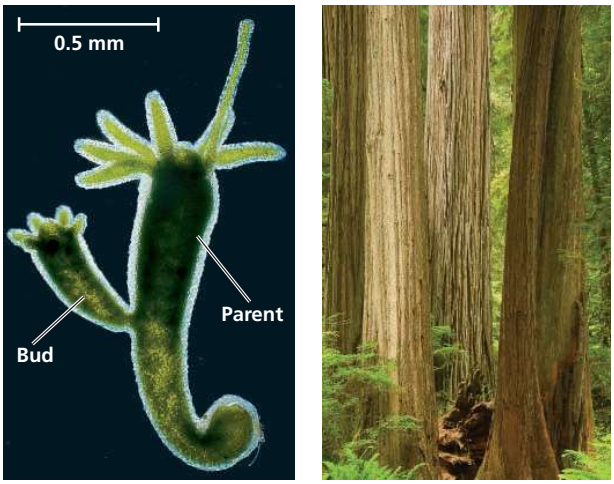
Except for small amounts of DNA in mitochondria and chloroplasts, the DNA of a eukaryotic cell is packaged into chromosomes within the nucleus. Every species has a characteristic number of chromosomes. For example, humans have 46 chromosomes in their **somatic cells**—all cells of the body except the gametes and their precursors. Each chromosome consists of a single long DNA molecule, elaborately coiled in association with various proteins. One chromosome includes several hundred to a few thousand genes, each gene being a precise sequence of nucleotides along the DNA molecule. A gene's specific location along the length of a chromosome is called the gene's **locus** (plural, *loci*; from the Latin, meaning "place"). Our genetic endowment (our genome) consists of the genes and other DNA that make up the chromosomes we inherited from our parents.

Comparison of Asexual and Sexual Reproduction

Only organisms that reproduce asexually have offspring that are exact genetic copies of themselves. In **asexual reproduction**, a single individual (like a yeast cell or an amoeba; see Figure 12.2a) is the sole parent and passes copies of all its genes to its offspring without the fusion of gametes. For example, single-celled eukaryotic organisms can reproduce asexually by mitotic cell division, in which DNA is copied and allocated equally to two daughter cells. The genomes of the offspring are virtually exact copies of the parent's genome. Some multicellular organisms are also capable of reproducing asexually **(Figure 13.2)**. Because the cells of the offspring arise via mitosis in the parent, the offspring is usually genetically identical to its parent. An individual that reproduces asexually gives rise to a **clone**, an individual or group of individuals that are genetically identical to the parent. Genetic differences occasionally arise in asexually reproducing organisms as a result of changes in the DNA called mutations, which we will discuss in Concept 17.5.

In **sexual reproduction**, two parents give rise to offspring that have unique combinations of genes inherited from the two parents. In contrast to a clone, offspring of sexual reproduction vary genetically from their siblings and both parents: They are variations on a common theme of family resemblance, not exact replicas. Genetic variation like that shown in Figure 13.1 is an important consequence of sexual reproduction. What mechanisms generate this genetic variation? The key is the behavior of chromosomes during the sexual life cycle.

▼ Figure 13.2 Asexual reproduction in two multicellular organisms. (a) This relatively simple animal, a hydra, reproduces by budding. The bud, a localized mass of mitotically dividing cells, develops into a small hydra, which detaches from the parent (LM). **(b)** All the trees in this circle of redwoods arose asexually from a single parent tree, whose stump is in the center of the circle.



(a) Hydra

(b) Redwoods

Mastering Biology Video: Hydra Budding
Animation: Asexual Reproduction

CONCEPT CHECK 13.1

1. **MAKE CONNECTIONS** Using your knowledge of gene expression in a cell, explain what causes the traits of parents (such as hair color) to show up in their offspring. (See Concept 5.5.)
2. Describe how an asexually reproducing eukaryotic organism produces offspring that are genetically identical to each other and to their parents.
3. **WHAT IF?** A horticulturalist breeds orchids, trying to obtain a plant with a unique combination of desirable traits. After many years, she finally succeeds, and wants to produce more plants like this one. Discuss whether she should crossbreed it with another plant or cause it to undergo asexual reproduction (forming a clone), and why.

For suggested answers, see Appendix A.

CONCEPT 13.2

Fertilization and meiosis alternate in sexual life cycles

A **life cycle** is the generation-to-generation sequence of stages in the reproductive history of an organism, from conception to production of its own offspring. In this section, we use humans as an example to track the behavior of chromosomes through the sexual life cycle. We begin by considering the chromosome count in human somatic cells and gametes. We will then explore how the behavior of chromosomes relates to the human life cycle and other types of sexual life cycles.

Sets of Chromosomes in Human Cells

In humans, each somatic cell has 46 chromosomes. Before mitosis begins, the chromosomes are duplicated. During mitosis, the chromosomes become condensed enough to be visible under a light microscope. At this point, they can be distinguished from one another by their size, the position of their centromeres, and the pattern of colored bands produced by certain chromatin-binding stains.

Careful examination of a micrograph of the 46 human chromosomes from a single cell in mitosis reveals that there are two chromosomes of each of 23 types. This becomes clear when images of the chromosomes are arranged in pairs, starting with the longest chromosomes. The resulting ordered display is called a **karyotype** **(Figure 13.3)**. The two chromosomes of a pair have the same length, centromere position, and staining pattern: These are called **homologous chromosomes** (or **homologs**). Both chromosomes of each pair carry genes controlling the same inherited characters. For example, if a gene for eye color is situated at a particular locus on a certain chromosome, then its homologous chromosome (its homolog) will also have a version of the eye-color gene at the equivalent locus.

The two chromosomes referred to as X and Y are an important exception to the general pattern of homologous chromosomes in human somatic cells. Typically, human females have a homologous pair of X chromosomes (XX), while males have one X and one Y chromosome (XY; see Figure 13.3). Only small

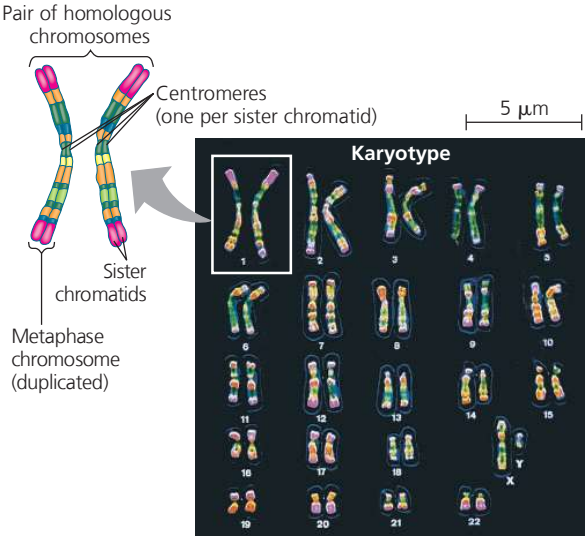
▼ Figure 13.3 Research Method

Preparing a Karyotype

Application A karyotype is a display of condensed chromosomes arranged in pairs. Karyotyping can be used to screen for defective chromosomes or abnormal numbers of chromosomes associated with certain congenital disorders, such as Down syndrome.



Technique Karyotypes are prepared from isolated somatic cells, which are treated with a drug to stimulate mitosis and then grown in culture for several days. Cells arrested when the chromosomes are most highly condensed—at metaphase—are stained and then viewed with a microscope equipped with a digital camera. An image of the chromosomes is displayed on a computer monitor, and digital software is used to arrange them in pairs according to their appearance.



Results This karyotype shows the chromosomes from a human male (as seen by the presence of the XY chromosome pair), colored to emphasize their chromosome banding patterns. The size of the chromosome, position of the centromere, and pattern of stained bands help identify specific chromosomes. Although difficult to discern in the karyotype, each metaphase chromosome consists of two closely attached sister chromatids (see the diagram of the first pair of homologous duplicated chromosomes).

MAKE CONNECTIONS *Explain why each chromosome in the karyotype consists of two sister chromatids. (See Figures 12.6 and 12.7.)*

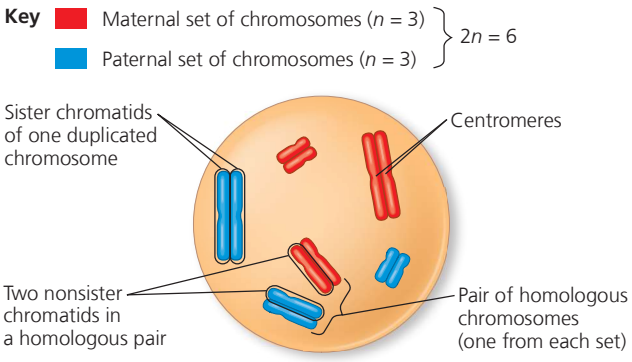
parts of the X and Y are homologous. Most of the genes carried on the X chromosome do not have counterparts on the tiny Y, and the Y chromosome has genes lacking on the X. Due to their role in sex determination, the X and Y chromosomes are called **sex chromosomes**. The other chromosomes are called **autosomes**.

The occurrence of pairs of homologous chromosomes in each human somatic cell is a consequence of our sexual origins. We inherit one chromosome of a pair from each parent. Thus, the 46 chromosomes in our somatic cells are actually two sets of 23 chromosomes—a maternal set (from our mother) and a paternal set (from our father). The number of chromosomes in a single set is represented by *n*. Any cell with two chromosome sets is called a **diploid cell** and has a diploid number of chromosomes, abbreviated 2*n*. For humans, the diploid number is 46 ($2n = 46$), the number of chromosomes in our somatic cells. In a cell in which DNA synthesis has occurred, all the chromosomes are duplicated, and therefore each consists of two identical sister chromatids, associated closely at the centromere and along the arms. (Even though the chromosomes are duplicated, we still say the cell is diploid, or 2*n*. This is because it has only two sets of information regardless of the number of chromatids, which are merely copies of the information in one set.) **Figure 13.4** helps clarify the various terms that we use to describe duplicated chromosomes in a diploid cell.

Unlike somatic cells, gametes contain a single set of chromosomes. Such cells are called **haploid cells**, and each has a haploid number of chromosomes (*n*). For humans, the haploid number is 23 ($n = 23$). The set of 23 consists of the 22 autosomes plus a single sex chromosome. An unfertilized egg contains an X chromosome; a sperm contains either an X or a Y chromosome.

Each sexually reproducing species has a characteristic diploid and haploid number. For example, the fruit fly *Drosophila melanogaster* has a diploid number (2*n*) of 8 and a haploid number (*n*) of 4, while for dogs, 2*n* is 78 and *n* is 39. The chromosome number generally does not correlate with the size or complexity of a species' genome; it simply reflects how many linear pieces of DNA make up the genome, which is a function of the evolutionary history of that species (see Concept 21.5). Now let's consider chromosome behavior during sexual life cycles. We'll use the human life cycle as an example.

▼ Figure 13.4 Describing chromosomes. A cell from an organism with a diploid number of 6 ($2n = 6$) is depicted here following chromosome duplication and condensation. Each of the six duplicated chromosomes consists of two sister chromatids associated closely along their lengths. Each homologous pair is composed of one chromosome from the maternal set (red) and one from the paternal set (blue). Each set is made up of three chromosomes in this example (long, medium, and short). Together, one maternal and one paternal chromatid in a pair of homologous chromosomes are called nonsister chromatids.



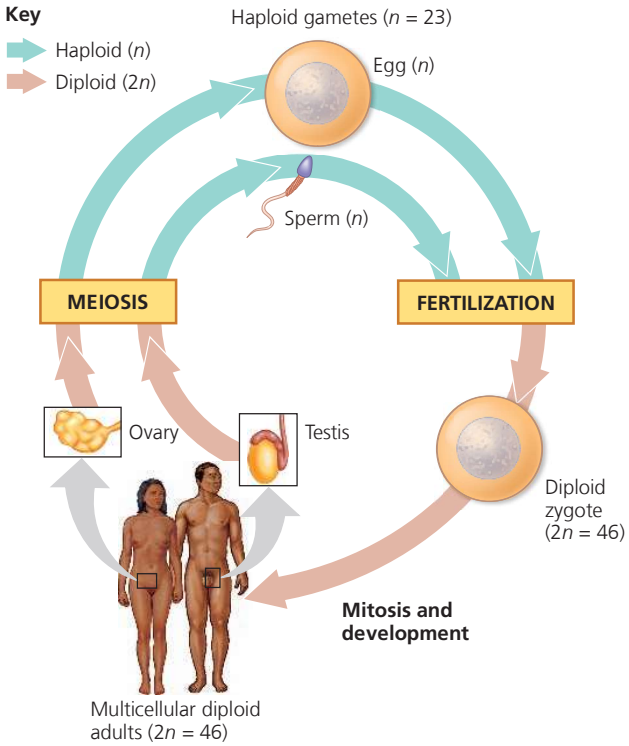
VISUAL SKILLS *How many sets of chromosomes are present in this diagram? How many pairs of homologous chromosomes are present? If the phases of mitosis in Figure 12.7 were redrawn using the color scheme shown in this figure, describe how the six chromosomes would be colored.*

Mastering Biology BioFlix® Animation: Chromosomes

Behavior of Chromosome Sets in the Human Life Cycle

The human life cycle begins when a haploid sperm from the father fuses with a haploid egg from the mother **(Figure 13.5)**. This union of gametes, culminating in fusion of their nuclei, is

▼ Figure 13.5 The human life cycle. In each generation, the number of chromosome sets per cell is halved during meiosis but doubles at fertilization. For humans, the number of chromosomes in a haploid cell is 23, consisting of one set ($n = 23$); the number of chromosomes in the diploid zygote and all somatic cells arising from it is 46, consisting of two sets ($2n = 46$).



This figure introduces a color code that will be used for other life cycles later in this book. The aqua arrows identify haploid stages of a life cycle, and the tan arrows identify diploid stages.

Mastering Biology Animation: The Human Life Cycle

called **fertilization**. The resulting fertilized egg, or **zygote**, is diploid because it contains two haploid sets of chromosomes bearing genes representing the maternal and paternal family lines. As a human develops into a sexually mature adult, mitosis of the zygote and its descendant cells generates all the somatic cells of the body. Both chromosome sets in the zygote and all the genes they carry are passed with precision to the somatic cells.

The only cells of the human body not produced by mitosis are the gametes, which develop from specialized cells called *germ cells* in the gonads—ovaries in females and testes in males (see Figure 13.5). Imagine what would happen if human gametes were made by mitosis: They would be diploid like the somatic cells. At the next round of fertilization, when two gametes fused, the normal chromosome number of 46 would double to 92, and each subsequent generation would double the number of chromosomes yet again. This does not happen, however, because in sexually reproducing organisms, gamete formation involves a type of cell division called **meiosis**. This type of cell division reduces the number of sets of chromosomes from two in the parent cell to one in each gamete, counterbalancing the doubling that occurs at fertilization. As a result of meiosis, each human sperm and egg is haploid ($n = 23$). Fertilization restores the diploid condition by combining two sets of chromosomes, and the human life cycle is repeated, generation after generation (see Figure 13.5).

The steps of the human life cycle are typical of many sexually reproducing animals. Indeed, fertilization and meiosis are also the hallmarks of sexual reproduction in plants, fungi, and protists just as in animals. Fertilization and meiosis alternate in sexual life cycles, maintaining a constant number of chromosomes in a species from one generation to the next.

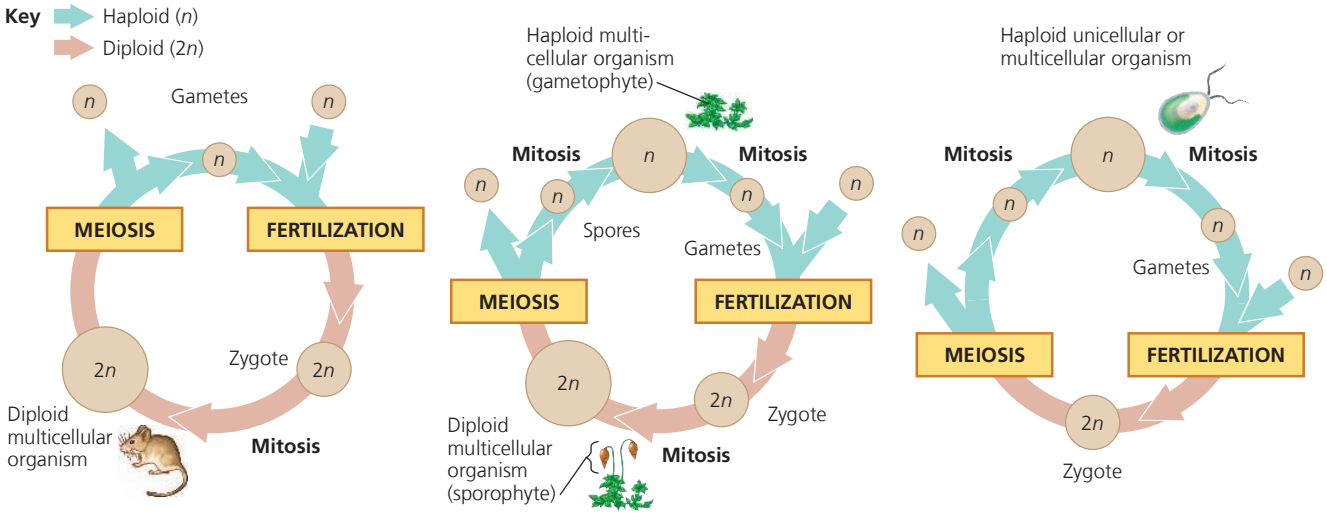
The Variety of Sexual Life Cycles

Although the alternation of meiosis and fertilization is common to all organisms that reproduce sexually, the timing of these two events in the life cycle varies, depending on the species. These variations can be grouped into three main types of life cycles. In the type that occurs in humans and most other animals, gametes are the only haploid cells **(Figure 13.6a)**. Meiosis occurs in germ cells during the production of gametes, which undergo no further cell division prior to fertilization. After fertilization, the diploid zygote divides by mitosis, producing a multicellular organism that is diploid.

Plants and some species of algae exhibit a second type of life cycle called **alternation of generations (Figure 13.6b)**. This type includes both diploid and haploid stages that are multicellular. The multicellular diploid stage is called the *sporophyte*. Meiosis in the sporophyte produces haploid cells called *spores*. Unlike a gamete, a haploid spore doesn't fuse with another cell but divides mitotically, generating a multicellular haploid stage called the *gametophyte*. Cells of the gametophyte give rise to gametes by mitosis. Fusion of two haploid gametes at fertilization results in a diploid zygote, which develops into the next sporophyte generation. Therefore, in this type of life cycle, the sporophyte generation produces a gametophyte as its offspring, and the gametophyte generation produces a sporophyte. The name *alternation of generations* fits well for this type of life cycle.

A third type of life cycle occurs in most fungi and some protists, including some algae **(Figure 13.6c)**. After gametes fuse and form a diploid zygote, meiosis occurs without a multicellular diploid offspring developing. Meiosis produces not gametes but

▼ Figure 13.6 Three types of sexual life cycles. The common feature of all three cycles is the alternation of meiosis and fertilization, key events that contribute to genetic variation among offspring. The cycles differ in the timing of these two key events. (Small circles are cells; large circles are organisms.)



(a) Animals **(b) Plants and some algae** **(c) Most fungi and some protists**

VISUAL SKILLS *For each type of life cycle, indicate whether haploid cells undergo mitosis, and if they do, describe the cells that are formed.*

haploid cells that then divide by mitosis and give rise to either unicellular descendants or a haploid multicellular adult organism. Subsequently, the haploid organism carries out further mitoses, producing the cells that develop into gametes. The only diploid stage found in these species is the single-celled zygote.

Note that *either* haploid or diploid cells can divide by mitosis, depending on the type of life cycle. Only diploid cells, however, can undergo meiosis: Haploid cells can't because they already have a single set of chromosomes that cannot be further reduced. Though the three types of sexual life cycles differ in the timing of meiosis and fertilization, they share a fundamental result: genetic variation among offspring.

CONCEPT CHECK 13.2

1. **MAKE CONNECTIONS** In Figure 13.4, how many DNA molecules (double helices) are present (see Figure 12.5)? What is the haploid number of this cell? Is a set of chromosomes haploid or diploid?
2. **VISUAL SKILLS** In the karyotype in Figure 13.3, how many pairs of chromosomes are present? How many sets?
3. Using shoes as an analogy for chromosomes, how would you describe the collection of "shoes" in human diploid and haploid cells?
4. **WHAT IF?** A certain eukaryote lives as a unicellular organism, but during environmental stress, it produces gametes. The gametes fuse, and the resulting zygote undergoes meiosis, generating new single cells. What type of organism could this be?

For suggested answers, see Appendix A.

CONCEPT 13.3

Meiosis reduces the number of chromosome sets from diploid to haploid

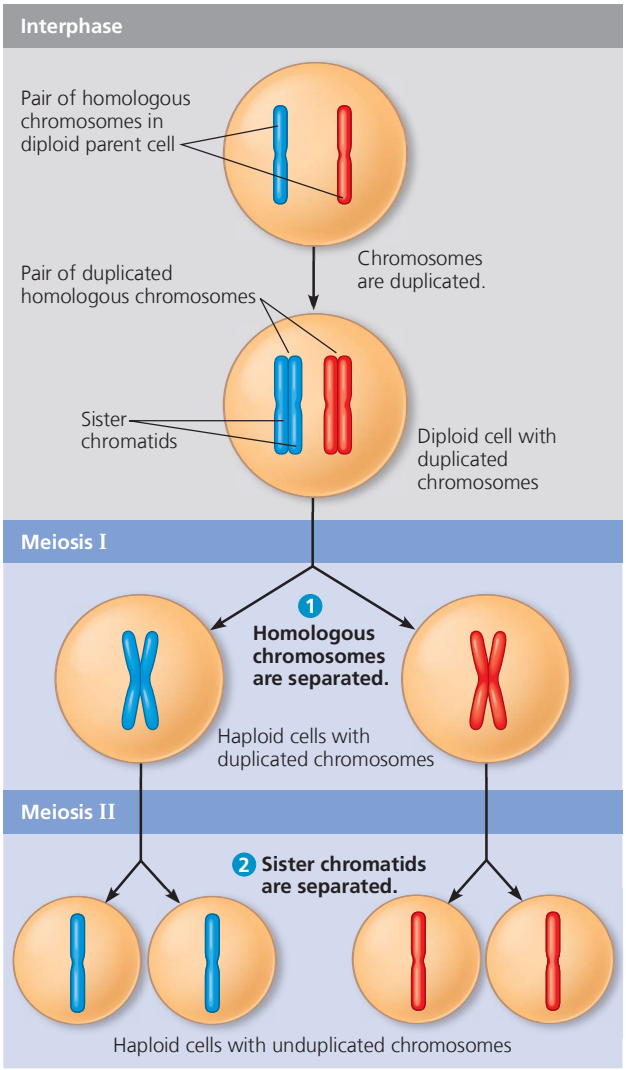
Several steps of meiosis closely resemble corresponding steps in mitosis. Meiosis, like mitosis, is preceded by interphase, which includes S phase (the duplication of chromosomes). However, this is followed by not one but two consecutive cell divisions, called **meiosis I** and **meiosis II**. These two divisions result in four daughter cells (rather than the two daughter cells of mitosis), each with only half as many chromosomes as the parent cell—one set, rather than two.

The Stages of Meiosis

The overview of meiosis in **Figure 13.7** shows, for a single pair of homologous chromosomes in a diploid cell, that both members of the pair are duplicated and the copies sorted into four haploid daughter cells. Recall that sister chromatids are two copies of *one* chromosome, closely associated all along their lengths; this association is called *sister chromatid cohesion*. Together, the sister chromatids make up one duplicated chromosome (see Figure 13.4). In contrast, the two chromosomes of a homologous pair are individual chromosomes that were inherited from each parent. Homologs appear alike in the microscope, but they may have different versions of genes at corresponding loci; each version is called an *allele* of that gene (see Figure 14.4). For example, one chromosome might have an allele for freckles, but the homologous chromosome may have an allele for the absence of freckles at the same locus. Homologs are not associated with each other in any obvious way except during meiosis.

Figure 13.8 describes in detail the stages of the two divisions of meiosis for an animal cell whose diploid number is 6. Study this figure thoroughly before going on.

▼ Figure 13.7 Overview of meiosis: how meiosis reduces chromosome number. After the chromosomes duplicate in interphase, the diploid cell divides *twice*, yielding four haploid daughter cells. This overview tracks just one pair of homologous chromosomes, which for the sake of simplicity are drawn in the condensed state throughout. (They would not normally be condensed during interphase.)

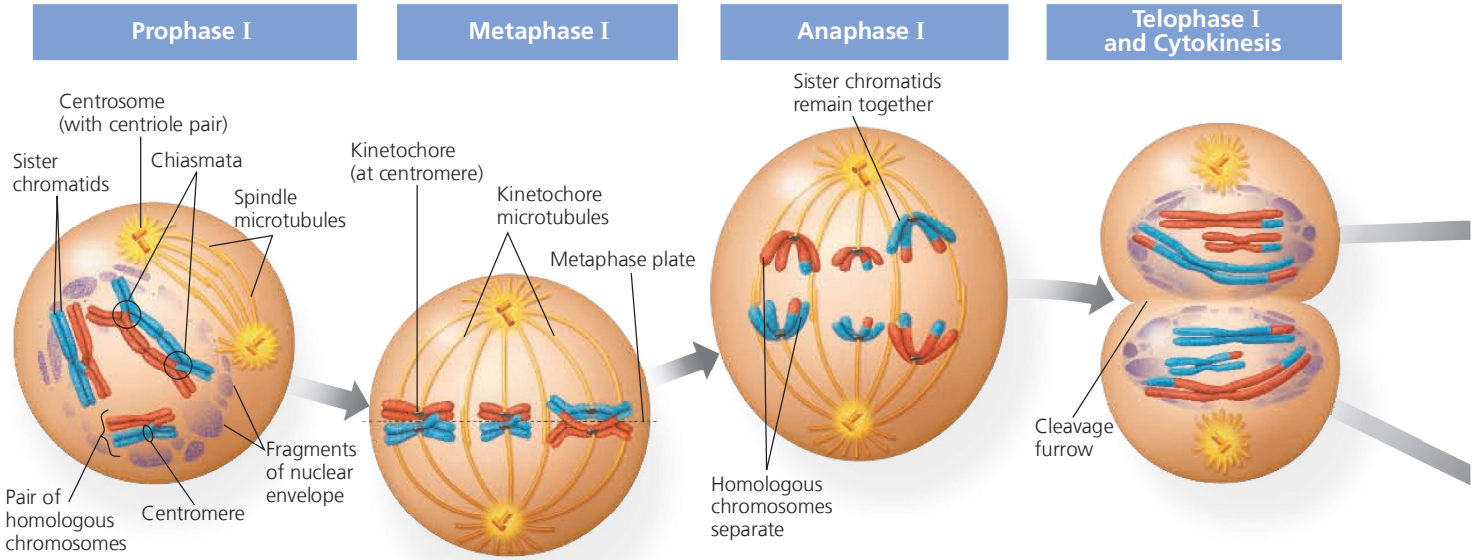


DRAW IT *Redraw the cells in this figure using a simple double helix to represent each DNA molecule.*

Mastering Biology **Animation: Overview of Meiosis**

▼ Figure 13.8 Exploring Meiosis in an Animal Cell

MEIOSIS I: Separates homologous chromosomes



Duplicated homologous chromosomes (red and blue) pair up and exchange segments; 2*n* = 6 in this example.

Chromosomes line up by homologous pairs.

The two homologous chromosomes of each pair separate.

Two haploid cells form; each chromosome still consists of two sister chromatids.

Prophase I

- Centrosome movement, spindle formation, and nuclear envelope breakdown occur as in mitosis. Chromosomes condense progressively throughout prophase I.
- During early prophase I, before the stage shown above, each chromosome pairs with its homolog, aligned gene by gene, and **crossing over** occurs: The DNA molecules of nonsister chromatids are broken (by proteins) and are rejoined to each other.
- At the stage shown above, each homologous pair has one or more X-shaped regions called **chiasmata** (singular, *chiasma*), where crossovers have occurred.
- Later in prophase I, microtubules from one pole or the other attach to the kinetochores, one at the centromere of each homolog. (The two kinetochores on the sister chromatids of a homolog are linked together by proteins and act as a single kinetochore.) Microtubules move the homologous pairs toward the metaphase plate (see the metaphase I diagram).

Metaphase I

- Pairs of homologous chromosomes are now arranged at the metaphase plate, with one chromosome of each pair facing each pole.
- Each pair has lined up independently of other pairs. (This arrangement is called *independent assortment*, to be discussed later.)
- Both chromatids of one homolog are attached to kinetochore microtubules from one pole; the chromatids of the other homolog are attached to microtubules from the opposite pole.

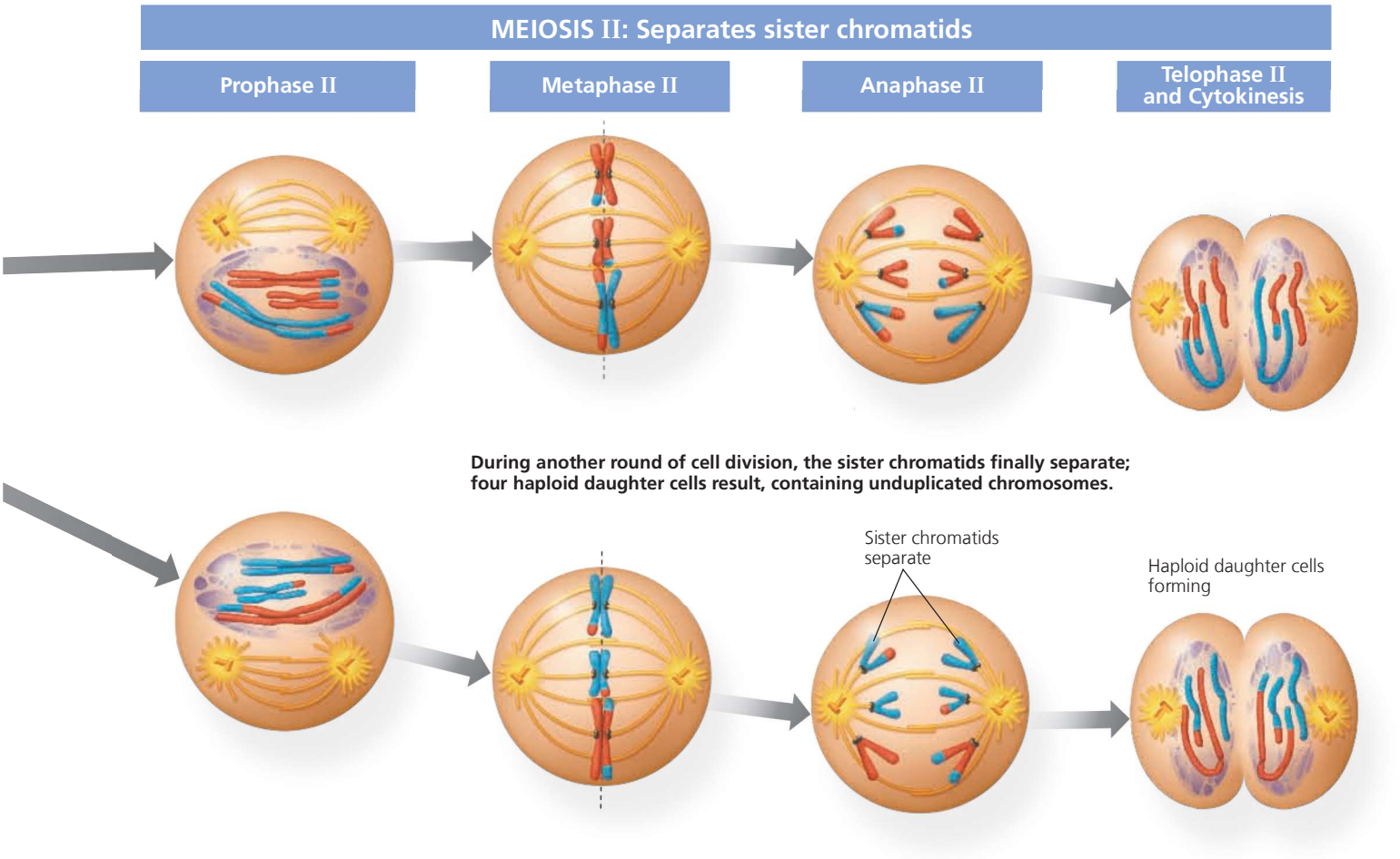
Anaphase I

- Breakdown of proteins that are responsible for sister chromatid cohesion along chromatid arms allows homologs to separate.
- The homologs move toward opposite poles, guided by the spindle apparatus.
- Sister chromatid cohesion persists at the centromere, causing the two chromatids of each chromosome to move as a unit toward the same pole.

Telophase I and Cytokinesis

- When telophase I begins, each half of the cell has a complete haploid set of duplicated chromosomes. Each chromosome is composed of two sister chromatids; one or both chromatids include regions of nonsister chromatid DNA.
- Cytokinesis (division of the cytoplasm) usually occurs simultaneously with telophase I, forming two haploid daughter cells.
- In animal cells like these, a cleavage furrow forms. (In plant cells, a cell plate forms.)
- In some species, chromosomes decondense and nuclear envelopes form.
- No chromosome duplication occurs between meiosis I and meiosis II.

Mastering Biology Video: Meiosis I in Sperm Formation



Prophase II

- A spindle apparatus forms.
- In late prophase II (not shown here), chromosomes, each still composed of two chromatids associated at the centromere, are moved by microtubules toward the metaphase II plate.

Metaphase II

- The chromosomes are positioned at the metaphase plate as in mitosis.
- Because of crossing over in meiosis I, the two sister chromatids of each chromosome are not genetically identical.
- The kinetochores of sister chromatids are attached to microtubules extending from opposite poles.

Anaphase II

- Breakdown of proteins holding the sister chromatids together at the centromere allows the chromatids to separate and move toward opposite poles. Each chromatid has now become an individual chromosome.

Telophase II and Cytokinesis

- Nuclei form, the chromosomes begin decondensing, and cytokinesis occurs.
- The meiotic division of one parent cell produces four daughter cells, each with a haploid set of (unduplicated) chromosomes.
- The four daughter cells are genetically distinct from one another and from the parent cell.

MAKE CONNECTIONS *Look at Figure 12.7 and imagine the two daughter cells undergoing another round of mitosis, yielding four cells. Compare the number of chromosomes in each of those four cells, after mitosis, with the number in each cell in Figure 13.8, after meiosis. Explain how the process of meiosis results in this difference, even though meiosis also includes two cell divisions.*

Mastering Biology BioFlix® Animation: Meiosis
Animation: Meiosis

Crossing Over and Synapsis During Prophase I

Prophase I of meiosis is a very busy time. The prophase I cell shown in Figure 13.8 is at a point fairly late in prophase I, when pairing of homologous chromosomes, crossing over, and chromosome condensation have already taken place. The sequence of events leading up to that point is shown in more detail in **Figure 13.9**.

After interphase, the chromosomes have been duplicated and the sister chromatids are held together by proteins called *cohesins*. (1) Early in prophase I, the two members of a homologous pair associate loosely along their length. Each gene on one homolog is aligned precisely with the corresponding allele of that gene on the other homolog. The DNA of two nonsister chromatids—one maternal and one paternal—is broken by specific proteins at precisely matching points. (2) Next, the formation of a zipper-like structure called the **synaptonemal complex** holds one homolog tightly to the other. (3) During this association, called **synapsis**, the DNA breaks are closed up so that each broken end is joined to the corresponding segment of the *nonsister* chromatid. Thus, a paternal chromatid is joined to a piece of maternal chromatid beyond the crossover point, and vice versa.

(4) These points where crossing over has just occurred become visible as chiasmata (singular, *chiasma*) after the synaptonemal complex disassembles and the homologs move slightly apart from each other. The homologs remain attached because sister chromatids are still held together by sister chromatid cohesion, even though some of the DNA may no longer be attached to its original chromosomal DNA. At least one crossover per chromosome must occur in order for the homologous pair to stay together as it moves to the metaphase I plate, for reasons that will be explained shortly.

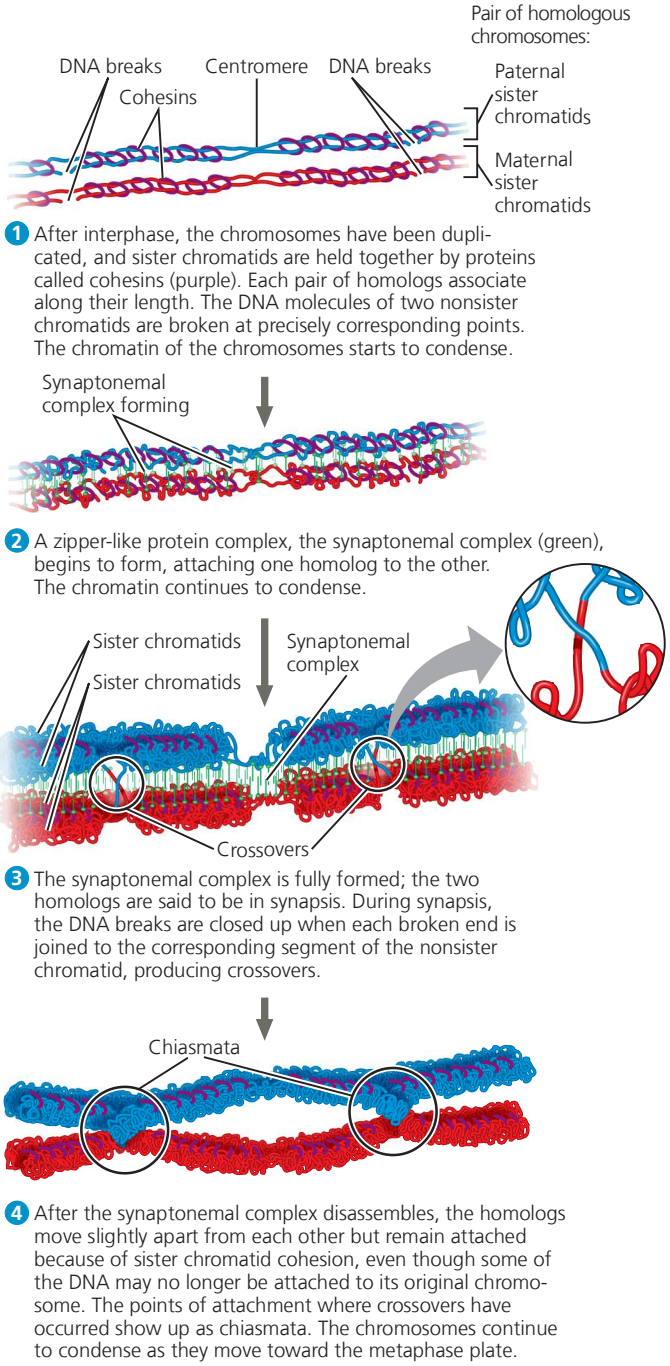
A Comparison of Mitosis and Meiosis

Figure 13.10 summarizes the key differences between meiosis and mitosis in diploid cells. Meiosis produces four cells and reduces the number of chromosome sets from two (diploid) to one (haploid), whereas mitosis produces two cells and conserves the number of chromosome sets. Meiosis produces cells that differ genetically from the parent cell and from each other, whereas mitosis produces daughter cells that are genetically identical to their parent cell and to each other.

Three events unique to meiosis occur during meiosis I:

1. **Synapsis and crossing over.** During prophase I, duplicated homologs pair up and crossing over occurs as described previously and in Figure 13.9. Synapsis and crossing over do not occur during prophase of mitosis.
2. **Alignment of homologous pairs at the metaphase plate.** At metaphase I of meiosis, pairs of homologs are positioned at the metaphase plate, rather than individual chromosomes, as in metaphase of mitosis.

▼ Figure 13.9 Crossing over and synapsis in prophase I: a closer look.



(1) After interphase, the chromosomes have been duplicated, and sister chromatids are held together by proteins called cohesins (purple). Each pair of homologs associate along their length. The DNA molecules of two nonsister chromatids are broken at precisely corresponding points. The chromatin of the chromosomes starts to condense.

(2) A zipper-like protein complex, the synaptonemal complex (green), begins to form, attaching one homolog to the other. The chromatin continues to condense.

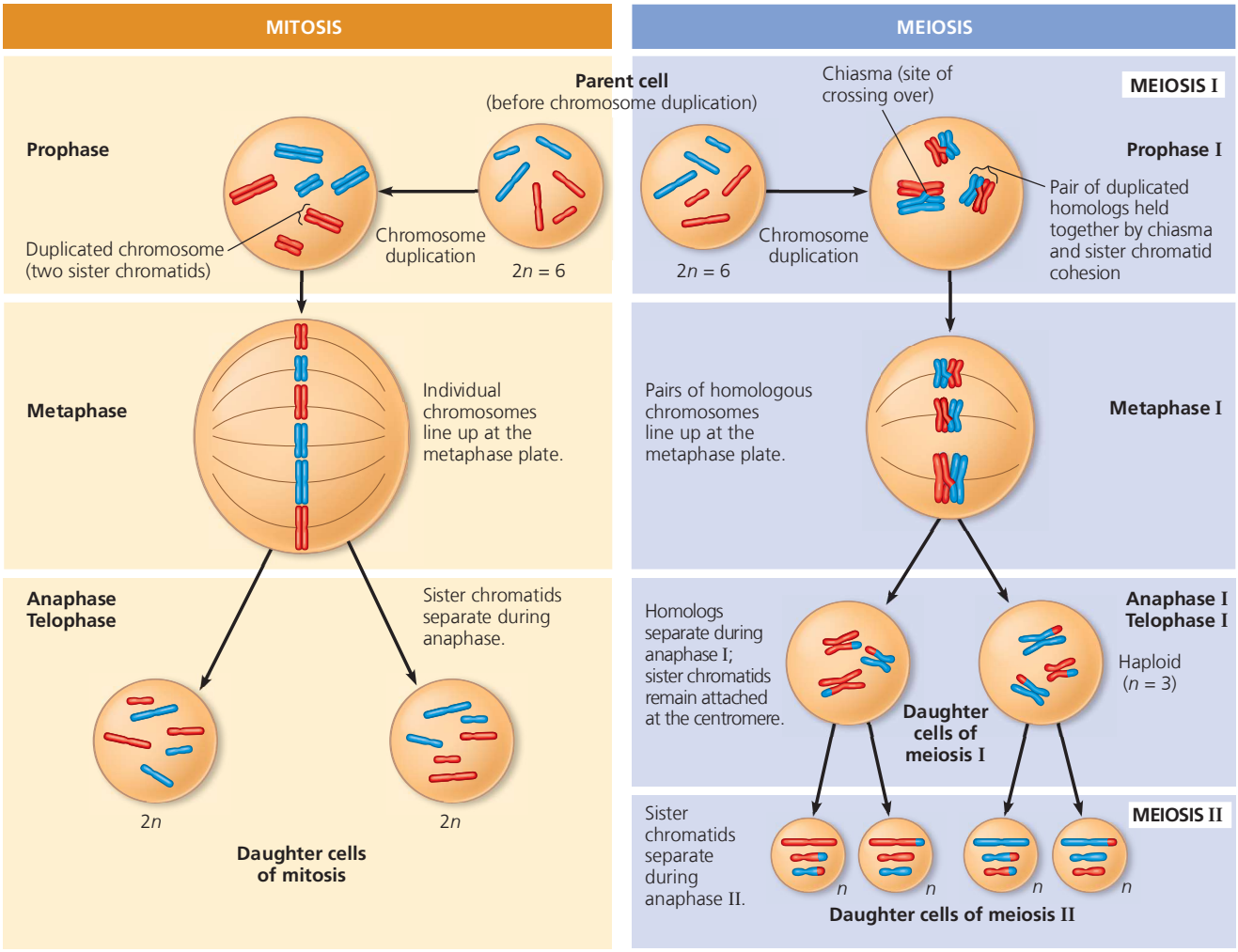
(3) The synaptonemal complex is fully formed; the two homologs are said to be in synapsis. During synapsis, the DNA breaks are closed up when each broken end is joined to the corresponding segment of the nonsister chromatid, producing crossovers.

(4) After the synaptonemal complex disassembles, the homologs move slightly apart from each other but remain attached because of sister chromatid cohesion, even though some of the DNA may no longer be attached to its original chromosome. The points of attachment where crossovers have occurred show up as chiasmata. The chromosomes continue to condense as they move toward the metaphase plate.

VISUAL SKILLS *Label the four DNA breaks in step 2.*

3. **Separation of homologs.** At anaphase I of meiosis, the duplicated chromosomes of each homologous pair move toward opposite poles, but the sister chromatids of each duplicated chromosome remain attached. In anaphase of mitosis, by contrast, sister chromatids separate.

▼ Figure 13.10 A comparison of mitosis and meiosis.



SUMMARY

Property	**Mitosis** (occurs in both diploid and haploid cells)	**Meiosis** (can only occur in diploid cells)
DNA replication	Occurs during interphase, before mitosis begins	Occurs during interphase before meiosis I but not before meiosis II
Number of divisions	One, including prophase, prometaphase, metaphase, anaphase, and telophase	Two, each including prophase, metaphase, anaphase, and telophase
Synapsis of homologous chromosomes	Does not occur	Occurs during prophase I along with crossing over between nonsister chromatids; resulting chiasmata hold pairs together due to sister chromatid cohesion
Number of daughter cells and genetic composition	Two, each genetically identical to the parent cell, with the same number of chromosomes	Four, each haploid (*n*); genetically different from the parent cell and from each other
Role in animals, fungi, and plants	Enables multicellular animal, fungus, or plant (gametophyte or sporophyte) to arise from a single cell; produces cells for growth, repair, and, in some species, asexual reproduction; produces gametes in the plant gametophyte	Produces gametes (in animals) or spores (in fungi and in plant sporophytes); reduces number of chromosome sets by half and introduces genetic variability among the gametes or spores

DRAW IT *Could any other combinations of chromosomes be generated during meiosis II from the specific cells shown in telophase I? Explain. (Hint: Draw the cells as they would appear in metaphase II.)*

Mastering Biology Figure Walkthrough

Sister chromatids stay together due to sister chromatid cohesion, mediated by cohesin proteins. In mitosis, this attachment lasts until the end of metaphase, when enzymes cleave the cohesins, freeing the sister chromatids to move to opposite poles of the cell. In meiosis, sister chromatid cohesion is released in two steps: cohesion of *arms* at the start of anaphase I and cohesion of *centromeres* at anaphase II. In metaphase I, the two homologs of each pair are held together because there is still cohesion between sister chromatid arms in regions beyond points of crossing over, where stretches of sister chromatids now belong to different chromosomes. The combination of crossing over and sister chromatid cohesion along the arms results in the formation of a chiasma. Chiasmata hold homologs together as the spindle forms for the first meiotic division. At the onset of anaphase I, the release of cohesion along sister chromatid arms allows homologs to separate. At anaphase II, the release of sister chromatid cohesion at the centromeres allows the sister chromatids to separate. Thus, sister chromatid cohesion and crossing over, acting together, play an essential role in the lining up of chromosomes by homologous pairs at metaphase I.

Meiosis I reduces the number of chromosome sets: from two (diploid) to one (haploid). During the second meiotic division, sister chromatids separate, producing four haploid daughter cells. The mechanisms for separating sister chromatids in meiosis II and mitosis are virtually identical. The molecular basis of chromosome behavior during meiosis continues to be a focus of intense research. In the **Scientific Skills Exercise**, you can work with data that track the amount of DNA in cells as they progress through meiosis.

Scientific Skills Exercise

Making a Line Graph and Converting Between Units of Data

How Does DNA Content Change as Budding Yeast Cells Proceed Through Meiosis? When nutrients are low, cells of the budding yeast (*Saccharomyces cerevisiae*) exit the mitotic cell cycle and enter meiosis. In this exercise, you will track the DNA content of a population of yeast cells as they progress through meiosis.

How the Experiment Was Done Researchers grew a culture of yeast cells in a nutrient-rich medium and then transferred the cells to a nutrient-poor medium to induce meiosis. At different times after induction, the DNA content per cell was measured in a sample of the cells, and the average DNA content per cell was recorded in femtograms (fg; 1 femtogram $= 1 \times 10^{-15}$ gram).

Data from the Experiment

Time After Induction of Meiosis (hr)	Average Amount of DNA per Cell (fg)
0.0	24.0
1.0	24.0
2.0	40.0
3.0	47.0
4.0	47.5
5.0	48.0
6.0	48.0
7.0	47.5
7.5	25.0
8.0	24.0
9.0	23.5
9.5	14.0
10.0	13.0
11.0	12.5
12.0	12.0
13.0	12.5
14.0	12.0



▶ **Budding yeast cells**

INTERPRET THE DATA

1. First, set up your graph. **(a)** Place the labels for the independent and dependent variables on the appropriate axes, followed by units of measurement in parentheses. Explain your choices. **(b)** Add tick marks and values for each axis. Explain your choices. (For additional information about graphs, see the Scientific Skills Review in Appendix D.)
2. Because the variable on the *x*-axis varies continuously, it makes sense to plot the data on a line graph (if you are using Excel, choose "Scatter"). **(a)** Plot each data point from the table onto the graph. **(b)** Connect the data points with line segments.
3. Most of the yeast cells in the culture were in G_1 of the cell cycle before being moved to the nutrient-poor medium. **(a)** How many femtograms of DNA are there in each yeast cell in G_1? Estimate this value from the data in your graph. **(b)** How many femtograms of DNA should be present in each cell in G_2? (See Concept 12.2 and Figure 12.6.) At the end of meiosis I (MI)? At the end of meiosis II (MII)? (See Figure 13.7.) **(c)** Using these values as a guideline, distinguish the different phases by inserting vertical dashed lines in the graph between phases and label each phase (G_1, S, G_2, MI, MII). You can figure out where to put the dividing lines based on what you know about the DNA content of each phase (see Figure 13.7). **(d)** Think carefully about the point where the line at the highest value begins to slope downward. What specific point of meiosis does this "corner" represent? What stage(s) correspond(s) to the downward sloping line?
4. Given the fact that 1 fg of DNA $= 9.78 \times 10^5$ base pairs (on average), you can convert the amount of DNA per cell to the length of DNA in numbers of base pairs. **(a)** Calculate the number of base pairs of DNA in the haploid yeast genome. Express your answer in millions of base pairs (Mb), a standard unit for expressing genome size. Show your work. **(b)** How many base pairs per minute were synthesized during the S phase of these yeast cells?

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.

Further Reading G. Simchen, Commitment to meiosis: What determines the mode of division in budding yeast? *BioEssays* 31:169–177 (2009).

CONCEPT CHECK 13.3

1. **MAKE CONNECTIONS** Compare the chromosomes in a cell at metaphase of mitosis with those in a cell at metaphase II. (See Figures 12.7 and 13.8.)
2. **WHAT IF?** After the synaptonemal complex disappears, how would any pair of homologous chromosomes be associated if crossing over did not occur? What effect might this have on gamete formation?

For suggested answers, see Appendix A.

CONCEPT 13.4

Genetic variation produced in sexual life cycles contributes to evolution

How do we account for the genetic variation of the family members in Figure 13.1? As you will learn more about in later chapters, mutations are the original source of genetic diversity. These changes in an organism's DNA create the different versions of genes, known as alleles. Once these differences arise, reshuffling of the alleles during sexual reproduction produces the variation that results in each member of a sexually reproducing population having a unique combination of traits.

Origins of Genetic Variation Among Offspring

In species that reproduce sexually, the behavior of chromosomes during meiosis and fertilization is responsible for most of the variation that arises in each generation. Three mechanisms contribute to the genetic variation arising from sexual reproduction: independent assortment of chromosomes, crossing over, and random fertilization.

Independent Assortment of Chromosomes

One aspect of sexual reproduction that generates genetic variation is the random orientation of pairs of homologous chromosomes at metaphase of meiosis I. At metaphase I, the homologous pairs, each consisting of one maternal and one paternal chromosome, are situated at the metaphase plate. (Note that the terms *maternal* and *paternal* refer, respectively, to whether the chromosome in question was contributed by the mother or the father of the individual whose cells are undergoing meiosis.) Each pair may orient with either its maternal or paternal homolog closer to a given pole—its orientation is as random as the flip of a coin. Thus, there is a 50% chance that a particular daughter cell of meiosis I will get the maternal chromosome of a certain homologous pair and a 50% chance that it will get the paternal chromosome.

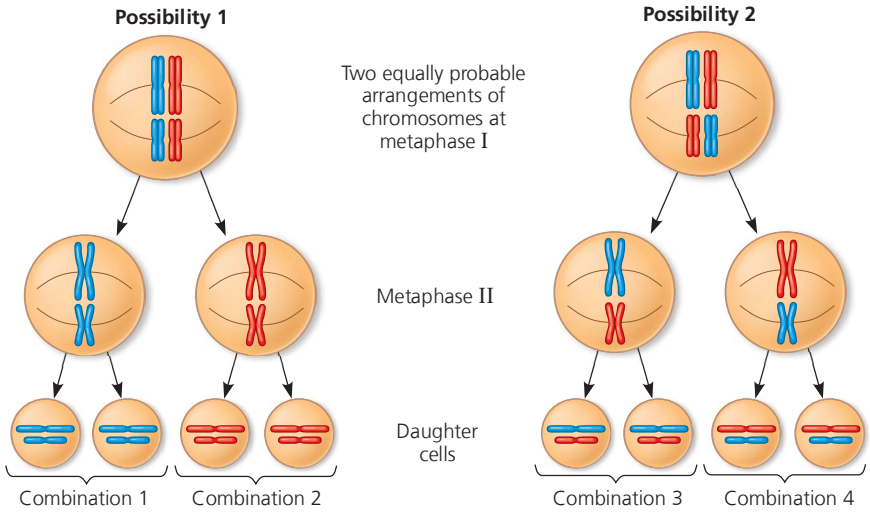
Because each pair of homologous chromosomes is positioned independently of the other pairs at metaphase I, the first meiotic division results in each pair sorting its maternal and paternal homologs into daughter cells independently of every other pair. This is called *independent assortment*. Each daughter cell represents one outcome of all possible combinations of maternal and paternal chromosomes. As shown in **Figure 13.11**, the number of combinations possible for daughter cells formed by meiosis of a diploid cell with two pairs of homologous chromosomes ($n = 2$) is four: two possible arrangements for the first pair times two possible arrangements for the second pair. Note that only two of the four combinations of daughter cells shown in the figure would result from meiosis of a *single* diploid cell because a single parent cell would have one or the other possible chromosomal arrangement at metaphase I, but not both. However, the population of daughter cells resulting from meiosis of a large number of diploid cells contains all four types in approximately equal numbers. In the case of $n = 3$, eight combinations ($2 \times 2 \times 2 = 2^3$) of chromosomes are possible for daughter cells. More generally, the number of possible combinations when chromosomes sort independently during meiosis is 2^n, where n is the haploid number of the species.

In the case of humans ($n = 23$), the number of possible combinations of maternal and paternal chromosomes in the resulting gametes is 2^{23}, or about 8.4 million. Each gamete that you produce in your lifetime contains one of roughly 8.4 million possible combinations of chromosomes. This is an underestimate, because it doesn't take into account crossing over, which we'll consider next.

Crossing Over

As a consequence of the independent assortment of chromosomes during meiosis, each of us produces a collection

▼ Figure 13.11 The independent assortment of homologous chromosomes in meiosis.



Mastering Biology Animation: Genetic Variation from Independent Assortment of Chromosomes

of gametes differing greatly in their combinations of the chromosomes we inherited from our two parents. Figure 13.11 suggests that each chromosome in a gamete is exclusively maternal or paternal in origin. In fact, this is *not* the case, because crossing over produces **recombinant chromosomes**, individual chromosomes that carry genes (DNA) from two different parents **(Figure 13.12)**. In meiosis in humans, an average of one to three crossover events occurs per chromosome pair, depending on the size of the chromosomes and the position of their centromeres.

As you learned in Figure 13.9, crossing over produces chromosomes with new combinations of maternal and paternal alleles. At metaphase II, chromosomes that contain one or more recombinant chromatids can be oriented in two alternative, nonequivalent ways with respect to other chromosomes because their sister chromatids are no longer identical (see Figure 13.12). The different possible arrangements of nonidentical sister chromatids during meiosis II further increase the number of genetic types of daughter cells that can result from meiosis.

You'll learn more about crossing over in Concept 15.3. The important point for now is that crossing over, by combining DNA inherited from two parents into a single chromosome, is an important source of genetic variation in sexual life cycles.

Random Fertilization

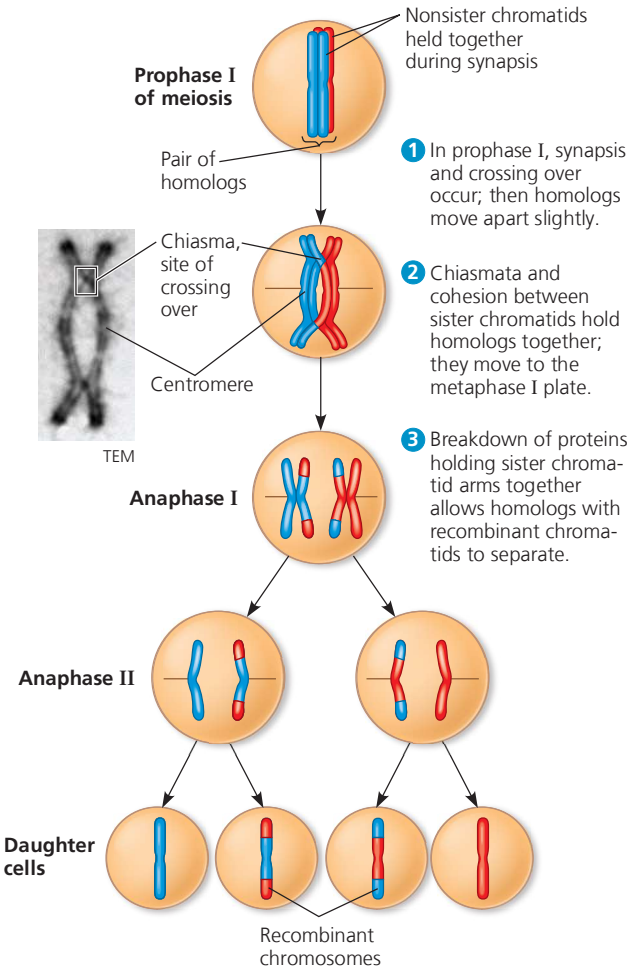
The random nature of fertilization adds to the genetic variation arising from meiosis. In humans, each male and female gamete represents one of about 8.4 million (2^{23}) possible chromosome combinations due to independent assortment. The fusion of a male gamete with a female gamete during fertilization will produce a zygote with any of about 70 trillion ($2^{23} \times 2^{23}$) diploid combinations. If we factor in the variation brought about by crossing over, the number of possibilities is truly astronomical. You really *are* unique.

Mastering Biology **Animation: Genetic Variation from Random Fertilization**

The Evolutionary Significance of Genetic Variation Within Populations

EVOLUTION Now that you've learned how new combinations of genes arise among offspring in a sexually reproducing population, how does the genetic variation in a population relate to evolution? Darwin recognized that a population evolves through the differential reproductive success of its variant members. On average, those individuals best suited to the local environment leave the most offspring, thereby transmitting their genes. Thus, natural selection results in the accumulation of genetic variations favored in a certain environment. As the environment changes, the population may survive if, in each generation, at least some of its members can cope effectively with the new conditions. Mutations are the original source of different alleles, which are then mixed and

▼ Figure 13.12 The results of crossing over during meiosis.



Mastering Biology **Animation: Genetic Variation from Crossing Over**

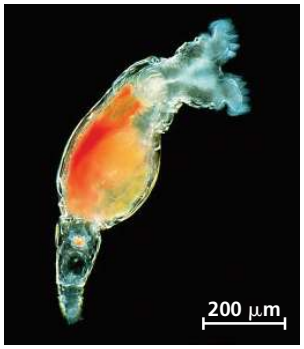
matched during meiosis. New and different combinations of alleles may work better than those that previously prevailed.

In a stable environment, though, sexual reproduction seems as if it would be less advantageous than asexual reproduction, which ensures perpetuation of successful combinations of alleles. Furthermore, sexual reproduction is more expensive energetically than asexual reproduction. In spite of these apparent disadvantages, sexual reproduction is almost universal among animals. Why is this?

The ability of sexual reproduction to generate genetic diversity is the most commonly proposed explanation for the evolutionary persistence of this process. However, consider the unusual case of the bdelloid rotifer **(Figure 13.13)**. It appears that this group may not have reproduced sexually for more than 50 million years of their evolutionary history, a model that has been supported by recent analysis of the genetic sequences in its genome. Does this mean that genetic diversity is not advantageous in this species? It turns out that bdelloid rotifers are

an exception to the "rule" that sex alone generates genetic diversity: Bdelloids have mechanisms other than sexual reproduction for generating genetic diversity. For example, they live in environments that can dry up for long periods of time, during which they can enter a state of suspended animation. In this state, their cell membranes may crack in places, allowing entry of DNA from other rotifer species and even from more distantly related species. Evidence suggests that this foreign DNA can become incorporated into the genome of the bdelloid, leading to increased genetic diversity. In fact, the genomic analysis shows that bdelloid rotifers pick up non-bdelloid DNA at a much higher rate than most other species pick up foreign DNA. The conclusion that bdelloid rotifers have developed other ways of achieving genetic diversity supports the idea that genetic diversity is advantageous, but points out as well that sexual reproduction is not the only way of generating such diversity.

▼ Figure 13.13 A bdelloid rotifer, an animal that reproduces only asexually.



In this chapter, we have seen how sexual reproduction greatly increases the genetic variation present in a population. Although Darwin realized that heritable variation is what makes evolution possible, he could not explain why offspring resemble—but are not identical to—their parents. Ironically, Gregor Mendel, a contemporary of Darwin, published a theory of inheritance that helps explain genetic variation, but his discoveries had no impact on biologists until 1900, more than 15 years after Darwin (1809–1882) and Mendel (1822–1884) had died. In the next chapter, you'll learn how Mendel discovered the basic rules governing the inheritance of specific traits.

CONCEPT CHECK 13.4

1. What is the original source of variation among the different alleles of a gene?
2. The diploid number for fruit flies is 8, and the diploid number for grasshoppers is 46. If no crossing over took place, would the genetic variation among offspring from a given pair of parents be greater in fruit flies or grasshoppers? Explain.
3. **WHAT IF?** If maternal and paternal chromatids have the same two alleles for every gene, will crossing over lead to genetic variation?

For suggested answers, see Appendix A.

13 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 13.1

Offspring acquire genes from parents by inheriting chromosomes *(pp. 255–256)*

- Each **gene** in an organism's DNA exists at a specific **locus** on a certain chromosome.
- In **asexual reproduction**, a single parent produces genetically identical offspring by mitosis. **Sexual reproduction** combines genes from two parents, leading to genetically diverse offspring.

? *Explain why human offspring resemble their parents but are not identical to them.*

CONCEPT 13.2

Fertilization and meiosis alternate in sexual life cycles *(pp. 256–259)*

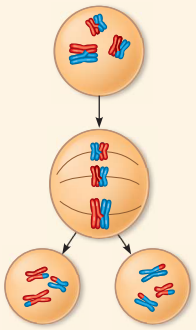
- Normal human **somatic cells** are **diploid**. They have 46 chromosomes made up of two sets of 23 chromosomes, one set from each parent. Human diploid cells have 22 pairs of **homologs** that are **autosomes**, and one pair of **sex chromosomes**; the latter typically determines whether the person is female (XX) or male (XY).
- In humans, ovaries and testes produce **haploid gametes** by **meiosis**, each gamete containing a single set of 23 chromosomes ($n = 23$). During **fertilization**, an egg and sperm unite, forming a diploid ($2n = 46$) single-celled **zygote**, which develops into a multicellular organism by mitosis.
- Sexual life cycles differ in the timing of meiosis relative to fertilization and in the point(s) of the cycle at which a multicellular organism is produced by mitosis.

? *Compare and contrast the life cycles of animals and plants.*

CONCEPT 13.3

Meiosis reduces the number of chromosome sets from diploid to haploid *(pp. 259–265)*

- The two cell divisions of meiosis, **meiosis I** and **meiosis II**, produce four haploid daughter cells. The number of chromosome sets is reduced from two (diploid) to one (haploid) during meiosis I.
- Meiosis is distinguished from mitosis by three events of meiosis I:



Prophase I: Each pair of homologous chromosomes undergoes **synapsis** and **crossing over** between nonsister chromatids with the subsequent appearance of **chiasmata**.

Metaphase I: Chromosomes line up as homologous pairs on the metaphase plate.

Anaphase I: Homologs separate from each other; sister chromatids remain joined at the centromere.

- Meiosis II separates the sister chromatids.
- Sister chromatid cohesion and crossing over allow chiasmata to hold homologs together until anaphase I. Cohesins are cleaved along the arms at anaphase I, allowing homologs to separate, and at the centromeres in anaphase II, releasing sister chromatids from each other.

? *In prophase I, homologous chromosomes pair up and undergo synapsis and crossing over. Can this also occur during prophase II? Explain.*

CONCEPT 13.4

Genetic variation produced in sexual life cycles contributes to evolution *(pp. 265–267)*

- Three events in sexual reproduction contribute to genetic variation in a population: independent assortment of chromosomes during meiosis I, crossing over during meiosis I, and random fertilization of egg cells by sperm. During crossing over, DNA of nonsister chromatids in a homologous pair is broken and rejoined.
- Genetic variation is the raw material for evolution by natural selection. Mutations are the original source of this variation; recombination of variant genes generates additional genetic diversity.

? *Explain how three processes unique to sexual reproduction generate a great deal of genetic variation.*

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. A human cell containing 22 autosomes and a Y chromosome is
(A) a sperm. (C) a zygote.
(B) an egg. (D) a somatic cell of a male.

2. The two homologs of a pair move toward opposite poles of dividing cell during
(A) mitosis. (C) meiosis II.
(B) meiosis I. (D) fertilization.

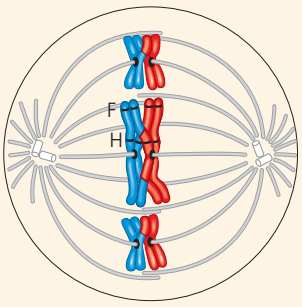
Levels 3-4: Applying/Analyzing

3. Meiosis II is similar to mitosis in that
(A) sister chromatids separate during anaphase.
(B) DNA replicates before the division.
(C) the daughter cells are diploid.
(D) homologous chromosomes synapse.

4. If the DNA content of a diploid cell in the G_1 phase of the cell cycle is x, then the DNA content of the same cell at metaphase of meiosis I will be
(A) 0.25x. (B) 0.5x. (C) x. (D) 2x.

5. If we continue to follow the cell lineage from question 4, then the DNA content of a single cell at metaphase of meiosis II will be
(A) 0.25x. (B) 0.5x. (C) x. (D) 2x.

6. DRAW IT The diagram shows a cell in meiosis.



(a) Label the appropriate structures with these terms: chromosome (label as duplicated or unduplicated), centromere, kinetochore, sister chromatids, nonsister chromatids, homologous pair (use a bracket when labeling), homolog (label each one), chiasma, sister chromatid cohesion, and gene loci. Circle and label the alleles of the F gene.
(b) Precisely describe the makeup of a haploid set and a diploid set in this cell.
(c) Identify the stage of meiosis shown.

Levels 5-6: Evaluating/Creating

7. Explain how you can tell that the cell in question 6 is undergoing meiosis, not mitosis.

8. EVOLUTION CONNECTION Many species can reproduce either asexually or sexually. Explain what you think might be the evolutionary significance of the switch from asexual to sexual reproduction that occurs in some species when the environment becomes unfavorable.

9. SCIENTIFIC INQUIRY The diagram in question 6 represents just a few of the chromosomes of a meiotic cell in a certain person. Assume that the freckles gene is located at the locus marked F, and the hair-color gene is located at the locus marked H, both on the long chromosome. The individual from whom this cell was taken has inherited different alleles for each gene ("freckles" and "black hair" from one parent, and "no freckles" and "blond hair" from the other). Predict allele combinations in the gametes resulting from this meiotic event. (It will help if you draw out the rest of meiosis and label the alleles by name.) List other possible combinations of these alleles in this individual's gametes.

10. WRITE ABOUT A THEME: INFORMATION The continuity of life is based on heritable information in the form of DNA. In a short essay (100–150 words), explain how chromosome behavior during sexual reproduction in animals ensures perpetuation of parental traits in offspring and, at the same time, genetic variation among offspring.

11. SYNTHESIZE YOUR KNOWLEDGE



The Cavendish banana, the world's most popular fruit, is currently threatened by extinction due to a fungus. This banana variety is "triploid" (3n, with three sets of chromosomes) and can only reproduce through cloning by cultivators. Given what you know about meiosis, explain how the banana's triploid number accounts for its inability to form normal gametes. Considering genetic diversity, discuss how the absence of sexual reproduction might make this domesticated species vulnerable to infectious agents.

For selected answers, see Appendix A.

14 Mendel and the Gene Idea

KEY CONCEPTS

14.1 Mendel used the scientific approach to identify two laws of inheritance *p. 270*

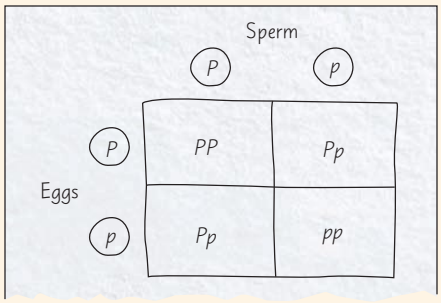
14.2 Probability laws govern Mendelian inheritance *p. 276*

14.3 Inheritance patterns are often more complex than predicted by simple Mendelian genetics *p. 278*

14.4 Many human traits follow Mendelian patterns of inheritance *p. 284*

Study Tip

Solving genetics problems: In Figure 14.5, you'll learn how to predict the offspring of a genetic cross with a Punnett square, like the example shown below. For crosses shown in other figures, cover up the Punnett square and draw your own. Also, check out the Tips for Genetics Problems at the end of the chapter.



Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 14
- Figure 14.4 Walkthrough: Alleles, Alternative Versions of a Gene
- Animation: Cross of One Character in "MendAliens"

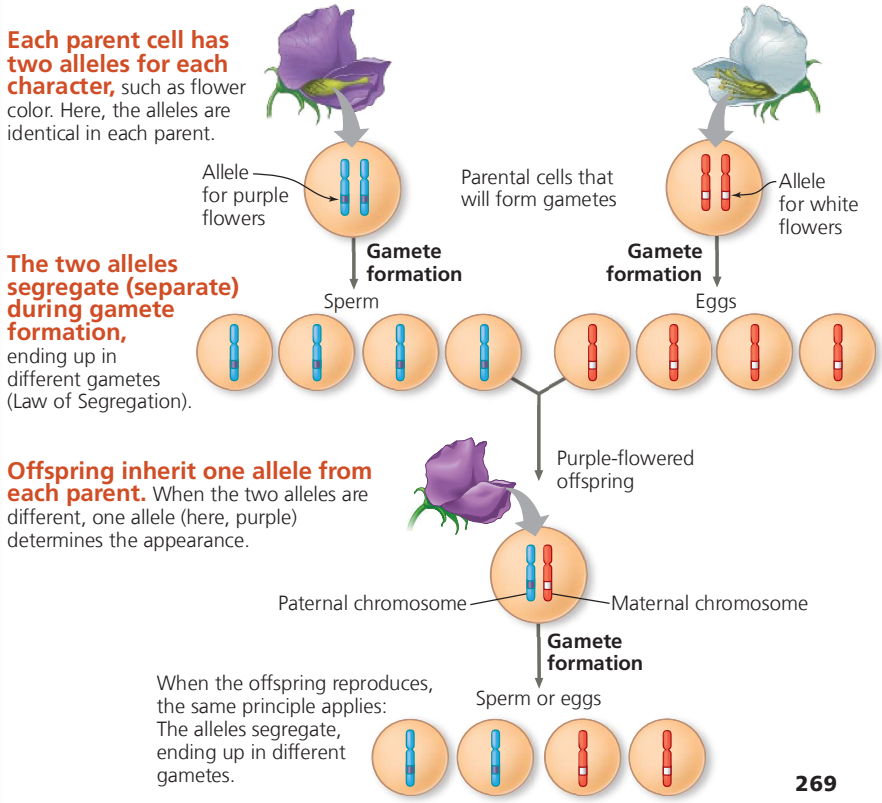
For Instructors to Assign (in Item Library)
- Tutorial: Mendel's Law of Independent Assortment
- Tutorial: Pedigree Analysis: Galactosemia



Figure 14.1 Flower color was one of the many characteristics of pea plants studied by Gregor Mendel. By breeding pea plants over many generations and carefully counting the types of offspring, Mendel developed a theory of inheritance that is the basis of modern genetics.

How are traits, such as the purple or white color of flowers, transmitted from parents to offspring?

Alternative versions of a gene (alleles) account for different traits.



CONCEPT 14.1

Mendel used the scientific approach to identify two laws of inheritance

Modern genetics began during the mid-1800s with a monk named Gregor Mendel, who discovered the basic principles of heredity by breeding garden peas in carefully planned experiments. As we retrace his work, you will recognize the key elements of the scientific process that were introduced in Concept 1.3.

Mendel's Experimental, Quantitative Approach

Mendel grew up on his parents' small farm in a region of Austria that is now part of the Czech Republic. In this agricultural area, Mendel received agricultural training in school along with a basic education. As an adolescent, Mendel overcame financial hardship and illness to excel in high school and, later, at the Olmutz Philosophical Institute.

In 1843, at the age of 21, Mendel entered an Augustinian monastery, a reasonable choice at that time for someone who valued the life of the mind. He considered becoming a teacher but failed the necessary examination. In 1851, he left the monastery to pursue two years of study in physics and chemistry at the University of Vienna. These were very important years for Mendel's development as a scientist, in large part due to the strong influence of two professors. One was the physicist Christian Doppler, who encouraged his students to learn science through experimentation and trained Mendel to use mathematics to help explain natural phenomena. The other was a botanist named Franz Unger, who aroused Mendel's interest in the causes of variation in plants.

After attending university, Mendel returned to the monastery and was assigned to teach at a local school, where several other instructors were enthusiastic about scientific research. In addition, his fellow monks shared a long-standing fascination with the breeding of plants. Around 1857, Mendel began breeding garden peas in the abbey garden to study inheritance. Although the question of heredity had long been a focus of curiosity at the monastery, Mendel's fresh approach allowed him to deduce principles that had remained elusive to others.

One reason Mendel probably chose to work with peas is that there are many varieties. For example, one variety has purple flowers, while another variety has white flowers. A heritable feature that varies among individuals, such as flower color, is called a **character**. Each variant for a character, such as purple or white color for flowers, is called a **trait**.

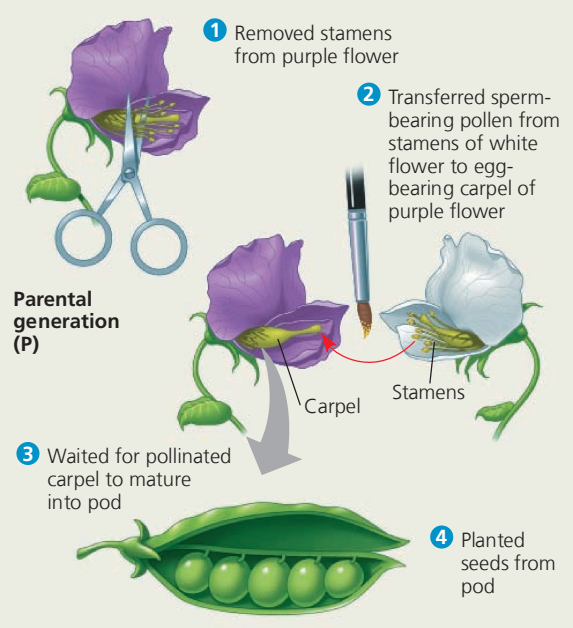
Other advantages of using peas are their short generation time and the large number of offspring from each mating. Furthermore, Mendel could strictly control mating between plants **(Figure 14.2)**. Each pea flower has both pollen-producing organs (stamens) and an egg-bearing organ (carpel). In nature, pea plants usually self-fertilize: Pollen grains from the stamens land on the carpel of the same flower, and sperm released from the pollen grains fertilize eggs present in the carpel.* To achieve

▼ Figure 14.2 Research Method

Crossing Pea Plants

Application By crossing, or mating, two true-breeding varieties of an organism, scientists can study patterns of inheritance. In this example, Mendel crossed pea plants that varied in flower color.

Technique



Results When pollen from a white flower was transferred to a purple flower, the first-generation hybrids all had purple flowers. The result was the same for the reciprocal cross, which involved the transfer of pollen from purple flowers to white flowers.



*As you learned in Figure 13.6b, meiosis in plants produces spores, not gametes. In flowering plants like the pea, each spore develops into a microscopic haploid gametophyte that contains only a few cells and is located on the parent plant. The gametophyte produces sperm, in pollen grains, and eggs, in the carpel. For simplicity, we will not include the gametophyte stage in our discussion of fertilization in plants.

cross-pollination of two plants, Mendel removed the immature stamens of a plant before they produced pollen and then dusted pollen from another plant onto the altered flowers (see Figure 14.2). Each resulting zygote then developed into a plant embryo encased in a seed (a pea). His method allowed Mendel to always be sure of the parentage of new seeds.

Mendel chose to track only those characters that occurred in two distinct, alternative forms, such as purple or white flower color. He also made sure that he started his experiments with varieties that were **true-breeding**—that is, over many generations of self-pollination, these plants had produced only the same variety as the parent plant. For example, a plant with purple flowers is true-breeding if the seeds produced by self-pollination in successive generations all give rise to plants that also have purple flowers.

In a typical breeding experiment, Mendel cross-pollinated two contrasting, true-breeding pea varieties—for example, purple-flowered plants and white-flowered plants (see Figure 14.2). This mating, or *crossing*, of two true-breeding varieties is called **hybridization**. The true-breeding parents are referred to as the **P generation** (parental generation), and their hybrid offspring are the **F_1 generation** (first filial generation, the word *filial* from the Latin word for "son"). Allowing these F_1 hybrids to self-pollinate (or to cross-pollinate with other F_1 hybrids) produces an **F_2 generation** (second filial generation). Mendel usually followed traits for at least the P, F_1, and F_2 generations. Had he stopped his experiments with the F_1 generation, the basic patterns of inheritance would have eluded him. Mendel's quantitative analysis of the F_2 plants from thousands of genetic crosses like these allowed him to deduce two fundamental principles of heredity, now called the law of segregation and the law of independent assortment.

The Law of Segregation

The explanation of heredity most widely in favor during the 1800s was the "blending" hypothesis, the idea that genetic material contributed by the two parents mixes just as blue and yellow paints blend to make green. This hypothesis predicts that over many generations, a freely mating population will give rise to a uniform population of individuals, something we don't see.

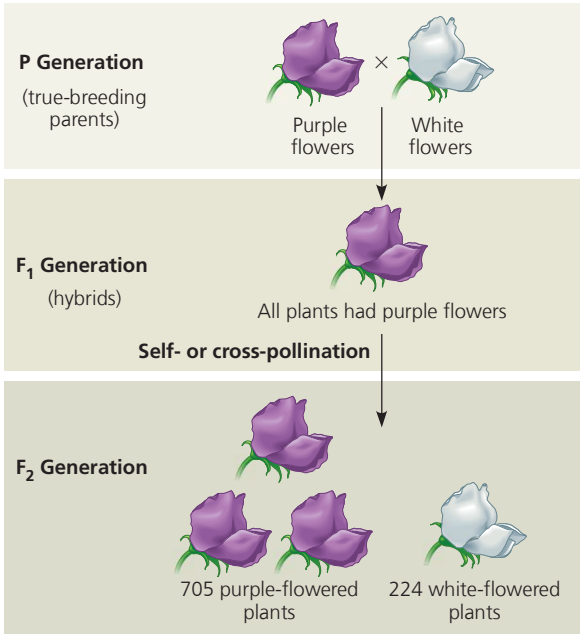
If the blending model were correct, the F_1 hybrids from Mendel's cross between purple-flowered and white-flowered pea plants would have been pale purple flowers, a trait intermediate between those of the P generation. Notice in Figure 14.2 that the experiment produced a very different result: All the F_1 offspring had flowers of the same color as the purple-flowered parents. What happened to the white-flowered plants' genetic contribution to the hybrids? If it were lost, then the F_1 plants could produce only purple-flowered offspring in the F_2 generation. But when Mendel allowed the F_1 plants to self- or cross-pollinate and planted their seeds, the white-flower trait reappeared in the F_2 generation. The blending hypothesis is also inconsistent with this reappearance of traits after they've skipped a generation.

Mendel used very large sample sizes and kept accurate records of his results: 705 of the F_2 plants had purple flowers, and 224 had white flowers. These data fit a ratio of approximately three purple to one white **(Figure 14.3)**. Mendel reasoned that the heritable factor for white flowers did not disappear in the F_1 plants but was somehow hidden, or masked,

▼ Figure 14.3 Inquiry

When F_1 hybrid pea plants self- or cross-pollinate, which traits appear in the F_2 generation?

Experiment Mendel crossed true-breeding purple-flowered plants and white-flowered plants (crosses are symbolized by ×). The resulting F_1 hybrids were allowed to self-pollinate or were cross-pollinated with other F_1 hybrids. The F_2 generation plants were then observed for flower color.



Results Both purple-flowered and white-flowered plants appeared in the F_2 generation, in a ratio of approximately 3:1.

Conclusion The "heritable factor" for the recessive trait (white flowers) had not been destroyed, deleted, or "blended" in the F_1 generation but was merely masked by the presence of the factor for purple flowers, which is the dominant trait.

Data from G. Mendel, Experiments in plant hybridization, *Proceedings of the Natural History Society of Brünn* 4:3–47 (1866).

WHAT IF? *If you mated two purple-flowered plants from the P generation, what ratio of traits would you expect to observe in the offspring? Explain. What might Mendel have concluded if he stopped his experiment after the F_1 generation?*

when the purple-flower factor was present. In Mendel's terminology, purple flower color is a *dominant* trait, and white flower color is a *recessive* trait. The reappearance of white-flowered plants in the F_2 generation was evidence that the heritable factor causing white flowers had not been diluted or destroyed by coexisting with the purple-flower factor in the F_1 hybrids. Mendel observed the same pattern of inheritance in six other characters, each represented by two distinctly different traits **(Table 14.1)**. For example, when Mendel crossed a true-breeding variety that produced smooth, round pea seeds with one that produced wrinkled seeds, all the F_1 hybrids produced round seeds; this is the dominant trait for seed shape. In the F_2 generation, approximately 75% of the seeds were round and 25% were wrinkled—a 3:1 ratio, as in Figure 14.3. Now let's see how Mendel deduced the law of segregation from his experimental results. In the discussion that follows, we will use modern terms instead of some of the terms used by Mendel. (For example, we'll use "gene" instead of Mendel's "heritable factor.")

Table 14.1 The Results of Mendel's F_1 Crosses for Seven Characters in Pea Plants

Character	Dominant Trait	×	Recessive Trait	F_2 Generation Dominant: Recessive	Ratio
Flower color	Purple 	×	White 	705:224	3.15:1
Seed color	Yellow 	×	Green 	6,022:2,001	3.01:1
Seed shape	Round 	×	Wrinkled 	5,474:1,850	2.96:1
Pod color	Green 	×	Yellow 	428:152	2.82:1
Pod shape	Inflated 	×	Constricted 	882:299	2.95:1
Flower position	Axial 	×	Terminal 	651:207	3.14:1
Stem length	Tall 	×	Dwarf 	787:277	2.84:1

Mendel's Model

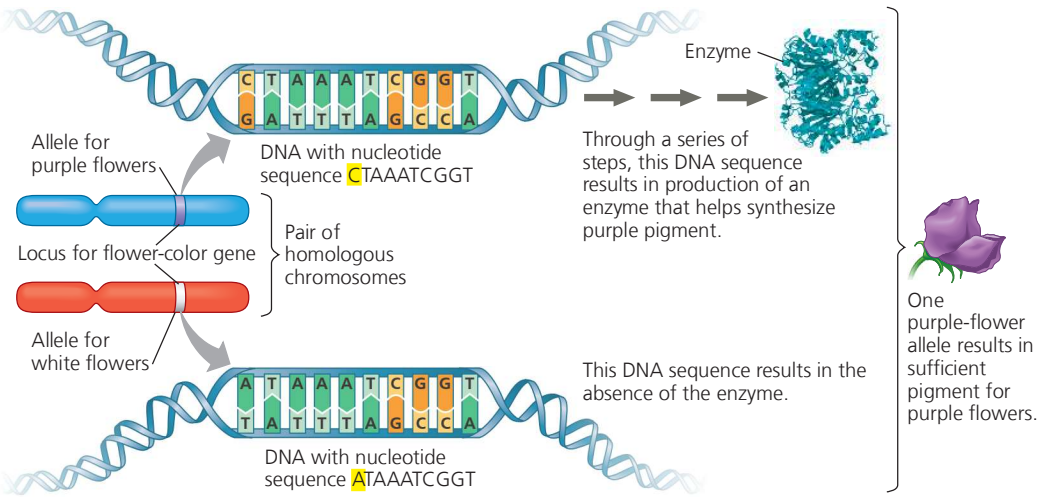
Mendel developed a model to explain the 3:1 inheritance pattern that he consistently observed among the F_2 offspring in his pea experiments. We'll now explore four related concepts making up Mendel's model, the fourth of which is the law of segregation.

First, *alternative versions of genes account for variations in inherited characters.* The gene for flower color in pea plants, for example, exists in two versions, one for purple flowers and the other for white flowers. These alternative versions of a gene are called **alleles**. Today, we can relate this concept to chromosomes and DNA. As shown in **Figure 14.4**, each gene is a sequence of nucleotides at a specific place, or locus, along a particular chromosome. The DNA at that locus, however, can vary slightly in its nucleotide sequence. This variation in information content can affect the function of the encoded protein and thus an inherited character of the organism. The purple-flower allele and the white-flower allele are two DNA sequence variations possible at the flower-color locus on a pea plant's chromosomes. The purple-flower allele sequence allows synthesis of purple pigment, and the white-flower allele sequence does not.

Second, *for each character, an organism inherits two versions (that is, two alleles) of a gene, one from each parent.* Remarkably, Mendel made this deduction without knowing about the role, or even the existence, of chromosomes. Each somatic cell in a diploid organism has two sets of chromosomes, one set inherited from each parent (see Figure 13.4). Thus, a genetic locus is actually represented twice in a diploid cell, once on each homolog of a specific pair of chromosomes. The two alleles at a particular locus may be identical, as in the true-breeding plants of Mendel's P generation. Or the alleles may differ, as in the F_1 hybrids (see Figure 14.4).

Third, *if the two alleles at a locus differ, then one, the* ***dominant allele****, determines the organism's appearance; the other, the* ***recessive allele****, has no noticeable effect on the organism's appearance.* Accordingly, Mendel's F_1 plants had purple flowers because the allele for that trait is dominant and the allele for white flowers is recessive.

The fourth and final part of Mendel's model, the **law of segregation**, states that *the two alleles for a heritable character segregate (in other words, separate from each other) during gamete formation and end up in different gametes.* Thus, an egg or a sperm gets only one of the two alleles that are present in the diploid cells of the organism making the gamete. In terms of chromosomes, this segregation corresponds to the distribution of copies of the two members of a pair of homologous chromosomes to different gametes in meiosis (see Figure 13.7). Note that if an organism has identical alleles for a particular character, then that allele is present in all gametes. Because it is the only allele that can be passed on to offspring when the plant self-pollinates, the offspring always look the same as their parents in regard to that characteristic; this explains why these plants are true-breeding. But if different alleles are



▶ Figure 14.4 Alleles, alternative versions of a gene. This diagram shows a pair of homologous chromosomes in an F_1 hybrid pea plant, with the actual DNA sequence from the flower-color allele of each chromosome. The paternally inherited chromosome (blue) has an allele for purple flowers, which codes for a protein that indirectly controls synthesis of purple pigment. The maternally inherited chromosome (red) has an allele for white flowers, which results in no functional protein being made.

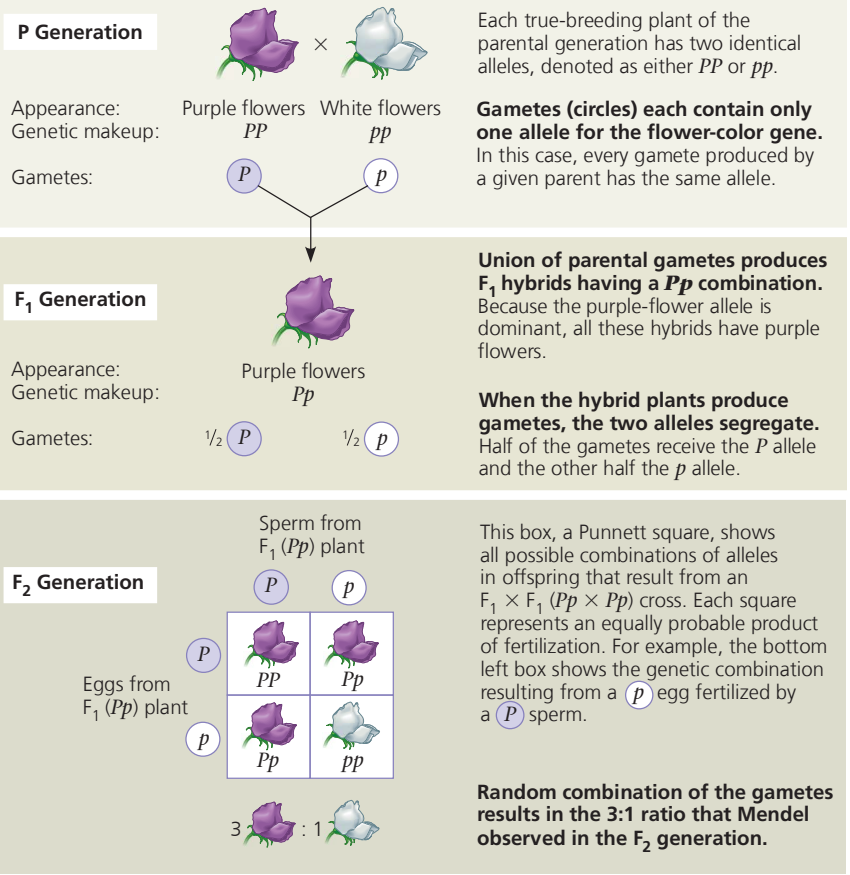
Mastering Biology
Figure Walkthrough

present, as in the F_1 hybrids, then 50% of the gametes receive the dominant allele and 50% receive the recessive allele.

Does Mendel's segregation model account for the 3:1 ratio he observed in the F_2 generation of his numerous crosses? For the flower-color character, the model predicts that the two different alleles present in an F_1 individual will segregate into gametes such that half the gametes will have the purple-flower allele and half will have the white-flower allele. During self-pollination, gametes of each class unite randomly. An egg with a purple-flower allele has an equal chance of being fertilized by a sperm with a purple-flower allele or by a sperm with a white-flower allele. Since the same is true for an egg with a white-flower allele, there are four equally likely combinations of sperm and egg. **Figure 14.5** illustrates these combinations using a **Punnett square**, a handy diagrammatic device for predicting the allele composition of offspring from a cross between individuals of known genetic makeup. Notice that we use a capital letter to symbolize a dominant allele and a lowercase letter for a recessive allele. In our example, *P* is the purple-flower allele, and *p* is the white-flower allele; it is often useful as well to be able to refer to the gene itself as the *P*/*p* gene.

In the F_2 offspring, what color will the flowers be? One-fourth of the

▼ Figure 14.5 Mendel's law of segregation. This diagram shows the genetic makeup of the generations in Figure 14.3. It illustrates Mendel's model for inheritance of the alleles of a single gene. Each plant has two alleles for the gene controlling flower color, one allele inherited from each of the plant's parents. To construct a Punnett square that predicts the F_2 generation offspring, we list all the possible gametes from one parent (here, the F_1 female) along the left side of the square and all the possible gametes from the other parent (here, the F_1 male) along the top. The boxes represent the offspring resulting from all the possible unions of male and female gametes.



Mastering Biology Animation: Mendel's Cross of One Character: Flower Color
Animation: Simplified Crosses of One Character in Humans
Animation: Cross of One Character in "MendAliens"

plants have inherited two purple-flower alleles; clearly, these plants will have purple flowers. One-half of the F_2 offspring have inherited one purple-flower allele and one white-flower allele; these plants will also have purple flowers, the dominant trait. Finally, one-fourth of the F_2 plants have inherited two white-flower alleles and will express the recessive trait. Thus, Mendel's model accounts for the 3:1 ratio of traits that he observed in the F_2 generation.

Useful Genetic Vocabulary

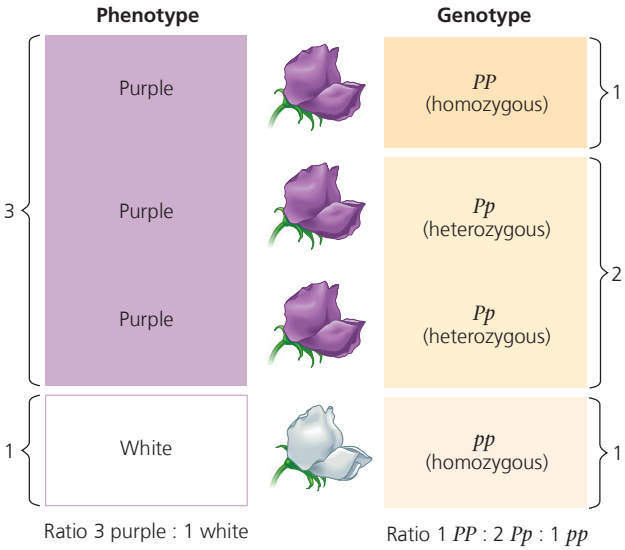
An organism that has a pair of identical alleles for a gene encoding a character is called a **homozygote** and is said to be **homozygous** for that gene. In the parental generation in Figure 14.5, the purple-flowered pea plant is homozygous for the dominant allele (*PP*), while the white plant is homozygous for the recessive allele (*pp*). Homozygous plants "breed true" because all of their gametes contain the same allele—either *P* or *p* in this example. If we cross dominant homozygotes with recessive homozygotes, every offspring will have two different alleles—*Pp* in the case of the F_1 hybrids of our flower-color experiment (see Figure 14.5). An organism that has two different alleles for a gene is called a **heterozygote** and is said to be **heterozygous** for that gene. Unlike homozygotes, heterozygotes produce gametes with different alleles, so they are not true-breeding. For example, *P*- and *p*-containing gametes are both produced by our F_1 hybrids. Self-pollination of the F_1 hybrids thus produces both purple-flowered and white-flowered offspring.

Because of the different effects of dominant and recessive alleles, an organism's traits do not always reveal its genetic composition. Therefore, we distinguish between an organism's appearance or observable traits, called its **phenotype**, and its genetic makeup, its **genotype**. As shown in Figure 14.5 for the case of flower color in pea plants, *PP* and *Pp* plants have the same phenotype (purple flowers) but different genotypes. **Figure 14.6** reviews these terms. Note that the term *phenotype* refers not only to traits that relate directly to physical appearance but also to physiological traits. For example, one pea variety lacks the normal ability to self-pollinate, which is a phenotypic trait (called non-self-pollination).

The Testcross

Given a purple-flowered pea plant, we cannot tell if it is homozygous (*PP*) or heterozygous (*Pp*) because both genotypes result in the same purple phenotype. To determine the genotype, we can cross this plant with a white-flowered plant (*pp*), which will make only gametes with the recessive allele (*p*). The allele in the gamete contributed by the purple-flowered plant of unknown genotype will therefore determine the appearance of the offspring **(Figure 14.7)**. If all the offspring of the cross have purple flowers, then the purple-flowered mystery plant must be homozygous for the dominant allele, because a *PP* × *pp* cross produces all *Pp* offspring. But if both the purple and the white phenotypes

▼ Figure 14.6 Phenotype versus genotype. Grouping F_2 offspring from a cross for flower color according to phenotype results in the typical 3:1 phenotypic ratio. In terms of genotype, however, there are actually two categories of purple-flowered plants, *PP* (homozygous) and *Pp* (heterozygous), giving a 1:2:1 genotypic ratio.



appear among the offspring, then the purple-flowered parent must be heterozygous. The offspring of a *Pp* × *pp* cross will be expected to have a 1:1 phenotypic ratio. Breeding an organism of unknown genotype with a recessive homozygote is called a **testcross** because it can reveal the genotype of that organism. The testcross was devised by Mendel and continues to be used by geneticists.

The Law of Independent Assortment

Mendel derived the law of segregation from experiments in which he followed only a *single* character, such as flower color. All the F_1 progeny produced in his crosses of true-breeding parents were **monohybrids**, meaning that they were heterozygous for the one particular character being followed in the cross. We refer to a cross between such heterozygotes as a **monohybrid cross**.

Mendel worked out the second law of inheritance by following *two* characters at the same time, such as seed color and seed shape. Seeds (peas) may be either yellow or green. They also may be either round (smooth) or wrinkled. From single-character crosses, Mendel knew that the allele for yellow seeds is dominant (*Y*), and the allele for green seeds is recessive (*y*). For the seed-shape character, the allele for round is dominant (*R*), and the allele for wrinkled is recessive (*r*).

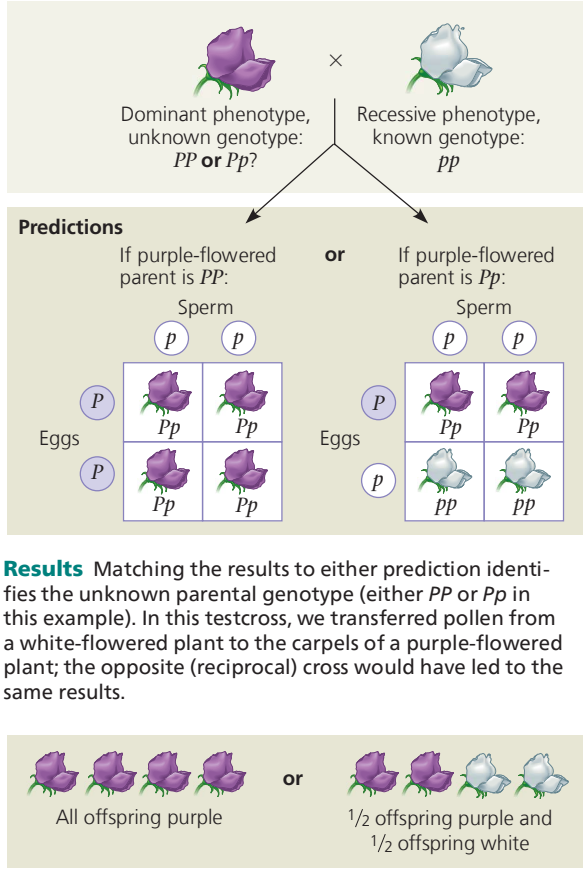
Imagine crossing two true-breeding pea varieties that differ in *both* of these characters—a cross between a plant with yellow round seeds (*YYRR*) and a plant with green wrinkled seeds (*yyrr*). The F_1 plants will be **dihybrids**, individuals heterozygous for the two characters being followed in the cross (*YyRr*). But are these two characters transmitted from parents to

▼ Figure 14.7 Research Method

The Testcross

Application An organism that shows a dominant trait in its phenotype, such as purple flowers in pea plants, can be either homozygous for the dominant allele or heterozygous. To determine the organism's genotype, geneticists can perform a testcross.

Technique In a testcross, the individual with the unknown genotype is crossed with a homozygous individual expressing the recessive trait (white flowers in this example), and Punnett squares are used to predict the possible outcomes.



Results Matching the results to either prediction identifies the unknown parental genotype (either *PP* or *Pp* in this example). In this testcross, we transferred pollen from a white-flowered plant to the carpels of a purple-flowered plant; the opposite (reciprocal) cross would have led to the same results.

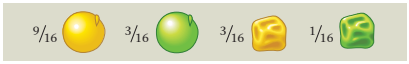
Mastering Biology Animation: Testcross in "MendAliens"

offspring as a package? That is, will the *Y* and *R* alleles always stay together, generation after generation? Or are seed color and seed shape inherited independently? **Figure 14.8** shows how a **dihybrid cross**, a cross between F_1 dihybrids, can determine which of these two hypotheses is correct.

The F_1 plants, of genotype *YyRr*, exhibit both dominant phenotypes, yellow seeds with round shapes, no matter which hypothesis is correct. The key step in the experiment is to see what happens when F_1 plants self-pollinate and produce F_2 offspring. If the hybrids must transmit their alleles in the same combinations in which the alleles were inherited from the P generation, then the F_1 hybrids will produce only two classes of gametes: *YR* and *yr*. As shown on the left side of Figure 14.8, this "dependent assortment" hypothesis predicts that the phenotypic ratio of the F_2 generation will be 3:1, just as in a monohybrid cross:



The alternative hypothesis is that the two pairs of alleles segregate independently of each other. In other words, genes are packaged into gametes in all possible allelic combinations, as long as each gamete has one allele for each gene (see Figure 13.11). In our example, an F_1 plant will produce four classes of gametes in equal quantities: *YR, Yr, yR*, and *yr*. If sperm of the four classes fertilize eggs of the four classes, there will be 16 (4×4) equally probable ways in which the alleles can combine in the F_2 generation, as shown on the right side of Figure 14.8. These combinations result in four phenotypic categories with a ratio of 9:3:3:1 (9 yellow round to 3 green round to 3 yellow wrinkled to 1 green wrinkled):



When Mendel did the experiment and classified the F_2 offspring, his results were close to the predicted 9:3:3:1 phenotypic ratio, supporting the hypothesis that the alleles for one gene—the gene for seed color, for example—segregate into gametes independently of the alleles of any other gene, such as the gene for seed shape.

Mendel tested his seven pea characters in various dihybrid combinations and always observed a 9:3:3:1 phenotypic ratio in the F_2 generation. Is this consistent with the 3:1 phenotypic ratio seen for the monohybrid cross shown in Figure 14.5? To answer this question, count the number of yellow and green peas, ignoring shape, and calculate the ratio. The results of Mendel's dihybrid experiments are the basis for what we now call the **law of independent assortment**, which states that *two or more genes assort independently—that is, each pair of alleles segregates independently of any other pair of alleles—during gamete formation.*

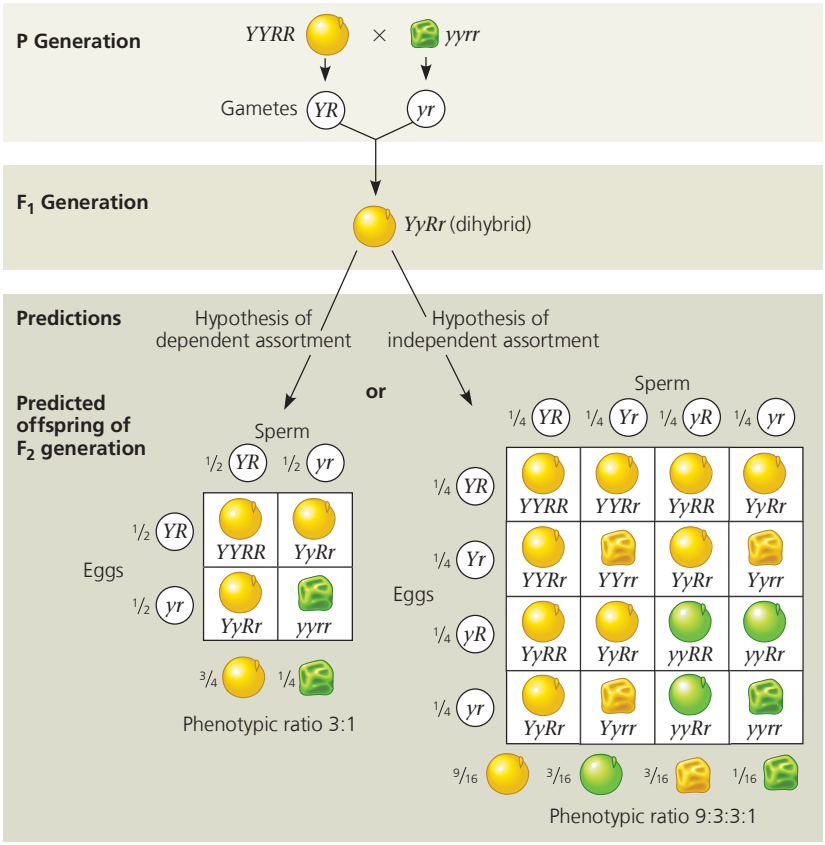
This law applies only to genes (allele pairs) located on different chromosomes (that is, on chromosomes that are not homologous) or, alternatively, to genes that are very far apart on the same chromosome. (This will be explained in Concept 15.3, along with the more complex inheritance patterns of genes located near each other, alleles of which tend to be inherited together.) All the pea characters Mendel chose for analysis were controlled by genes on different chromosomes or were far apart on the same chromosome; this situation greatly simplified interpretation of his multicharacter pea crosses. All the examples we consider in the rest of this chapter involve genes located on different chromosomes.

Mastering Biology Animation: Independent Assortment

▼ Figure 14.8 Inquiry

Do the alleles for one character segregate into gametes dependently or independently of the alleles for a different character?

Experiment To follow the characters of seed color and seed shape through the F_2 generation, Mendel crossed a true-breeding plant with yellow round seeds with a true-breeding plant with green wrinkled seeds, producing dihybrid F_1 plants. Self-pollination of the F_1 dihybrids produced the F_2 generation. The two hypotheses (dependent and independent "assortment" of the two genes) predict different phenotypic ratios.



Results

315 108 101 32 Phenotypic ratio approximately 9:3:3:1

Conclusion The results support the hypothesis of independent assortment, the only one that predicts two newly observed phenotypes: green round seeds and yellow wrinkled seeds (see the right-hand Punnett square). The alleles for each gene segregate independently of those of the other, and the two genes are said to assort independently.

Data from G. Mendel, Experiments in plant hybridization, *Proceedings of the Natural History Society of Brünn* 4:3–47 (1866).

WHAT IF? *Suppose Mendel had transferred pollen from an F_1 plant to the carpel of a plant that was homozygous recessive for both genes. Set up the cross and draw Punnett squares that predict the offspring for both hypotheses. Would this cross have supported the hypothesis of independent assortment equally well?*

Mastering Biology
Animation: Mendel's Cross of Two Characters: Seed Shape and Seed Color
Animation: A Simplified Cross of Two Characters in Humans
Animation: Crosses of Two Characters in "MendAliens"

CONCEPT CHECK 14.1

1. **DRAW IT** Pea plants heterozygous for flower position and stem length (*AaTt*) are allowed to self-pollinate, and 400 of the resulting seeds are planted. Draw a Punnett square for this cross. How many offspring would be predicted to have terminal flowers and be dwarf? (See Table 14.1.)
2. **WHAT IF?** List all gametes that could be made by a pea plant heterozygous for seed color, seed shape, and pod shape (*YyRrIi*; see Table 14.1). How large a Punnett square would you need to draw to predict the offspring of a self-pollination of this "trihybrid"?
3. **MAKE CONNECTIONS** In some pea plant crosses, the plants are self-pollinated. Is self-pollination considered asexual or sexual reproduction? Explain. (See Concept 13.1.)

For suggested answers, see Appendix A.

CONCEPT 14.2

Probability laws govern Mendelian inheritance

Mendel's laws of segregation and independent assortment reflect the same rules of probability that apply to tossing coins, rolling dice, and drawing cards from a deck. The probability scale ranges from 0 to 1. An event that is certain to occur has a probability of 1, while an event that is certain *not* to occur has a probability of 0. With a coin that has heads on both sides, the probability of tossing heads is 1, and the probability of tossing tails is 0. With a normal coin, the chance of tossing heads is $\frac{1}{2}$, and the chance of tossing tails is $\frac{1}{2}$. The probability of drawing the ace of spades from a 52-card deck is $\frac{1}{52}$. The probabilities of all possible outcomes for an event must add up to 1. With a deck of cards, the chance of picking a card other than the ace of spades is $\frac{51}{52}$.

Tossing a coin illustrates an important lesson about probability. For every toss, the probability of heads is $\frac{1}{2}$. The outcome of any particular toss is unaffected by what has happened on previous trials. We refer to phenomena such as coin tosses as independent events. Each toss of a coin, whether done sequentially with one coin or simultaneously with many, is independent of every other toss. And like two

separate coin tosses, the alleles of one gene segregate into gametes independently of another gene's alleles (the law of independent assortment). We'll now look at two basic rules of probability that help us predict the outcome of the fusion of such gametes in simple monohybrid crosses and more complicated crosses as well.

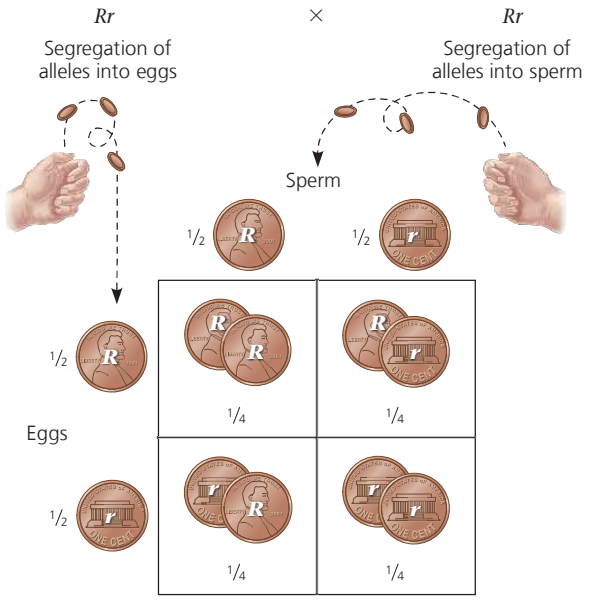
The Multiplication and Addition Rules Applied to Monohybrid Crosses

How do we determine the probability that two or more independent events will occur together in some specific combination? For example, what is the chance that two coins tossed simultaneously will both land heads up? The **multiplication rule** states that to determine the probability of one event *and* the other occurring, we multiply the probability of one event (one coin coming up heads) by the probability of the other event (the other coin coming up heads). By the multiplication rule, then, the probability that both coins will land heads up is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

We can apply the same reasoning to an F_1 monohybrid cross. With seed shape in pea plants as the heritable character, the genotype of F_1 plants is *Rr*. Segregation in a heterozygous plant is like flipping a coin in terms of calculating the probability of each outcome: Each egg produced has a $\frac{1}{2}$ chance of carrying the dominant allele (*R*) and a $\frac{1}{2}$ chance of carrying the recessive allele (*r*). The same odds apply to each sperm cell produced. For a particular F_2 plant to have wrinkled seeds, the recessive trait, both the egg and the sperm that come together must carry the *r* allele. The probability that an *r* allele will be present in the egg *and* in the sperm at fertilization is found by multiplying $\frac{1}{2}$ (the probability that the egg will have an *r*) $\times$ $\frac{1}{2}$ (the probability that the sperm will have an *r*). Thus, the multiplication rule tells us that the probability of an F_2 plant having wrinkled seeds (*rr*) is $\frac{1}{4}$ **(Figure 14.9)**. Likewise, the probability of an F_2 plant carrying both dominant alleles for seed shape (*RR*) is $\frac{1}{4}$.

To figure out the probability that an F_2 plant from a monohybrid cross will be heterozygous rather than homozygous, we need to invoke a second rule. Notice in Figure 14.9 that the dominant allele can come from the egg and the recessive allele from the sperm, or vice versa. That is, F_1 gametes can combine to produce *Rr* offspring in two *mutually exclusive* ways: For any particular heterozygous F_2 plant, the dominant allele can come from the egg *or* the sperm, but not from both. According to the **addition rule**, the probability that any one of two or more mutually exclusive events (one event *or* the other) will occur is calculated by adding their individual probabilities. As we have just seen, the multiplication rule gives us the individual probabilities that we will now add together. The probability for one possible way of obtaining an F_2 heterozygote—the dominant allele from the egg and the recessive allele from the sperm—is $\frac{1}{4}$. The probability for the other possible way—the recessive allele from the egg and the dominant allele from the sperm—is also $\frac{1}{4}$ (see Figure 14.9). Using the rule of addition, then, we can calculate the probability of an F_2 heterozygote as $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

▼ Figure 14.9 Segregation of alleles and fertilization as chance events. When a heterozygote (*Rr*) forms gametes, whether a particular gamete ends up with an *R* or an *r* is like the toss of a coin. We can determine the probability for any genotype among the offspring of two heterozygotes by multiplying together the individual probabilities of an egg and sperm having a particular allele (*R* or *r* in this example).



(Note that use of the word *and* in the statement is a clue that you should use the multiplication rule, while *or* suggests use of the addition rule.)

Solving Complex Genetics Problems with the Rules of Probability

We can also apply the rules of probability to predict the outcome of crosses involving multiple characters. Recall that each allelic pair (gene) segregates independently during gamete formation (the law of independent assortment). Thus, a dihybrid or other multicharacter cross is equivalent to two or more independent monohybrid crosses occurring simultaneously. By applying what we have learned about monohybrid crosses, we can determine the probability of specific genotypes occurring in the F_2 generation without having to construct unwieldy Punnett squares.

Consider the dihybrid cross between *YyRr* heterozygotes shown in Figure 14.8. We will focus first on the seed-color character. For a monohybrid cross of *Yy* plants, we can use a simple Punnett square to determine that the probabilities of the offspring genotypes are $\frac{1}{4}$ for *YY*, $\frac{1}{2}$ for *Yy*, and $\frac{1}{4}$ for *yy*. We can draw a second Punnett square to determine that the same probabilities apply to the offspring genotypes for seed shape: $\frac{1}{4}$ *RR*, $\frac{1}{2}$ *Rr*, and $\frac{1}{4}$ *rr*. Knowing these probabilities, we can simply use the multiplication rule to determine the probability of each of the genotypes in the F_2 generation. To give

two examples, the calculations for finding the probabilities of two of the possible F_2 genotypes (*YYRR* and *YyRR*) are shown below:

Probability of *YYRR* = 1/4 (probability of *YY*) × 1/4 (*RR*) = 1/16

Probability of *YyRR* = 1/2 (*Yy*) × 1/4 (*RR*) = 1/8

The *YYRR* genotype corresponds to the upper left box in the larger Punnett square in Figure 14.8 (one box = 1/16). Looking closely at the larger Punnett square in Figure 14.8, you will see that 2 of the 16 boxes (1/8) correspond to the *YyRR* genotype.

Now let's see how we can combine the multiplication and addition rules to solve even more complex problems in Mendelian genetics. Imagine a cross of two pea varieties in which we track the inheritance of three characters. Let's cross a trihybrid with purple flowers and yellow round seeds (heterozygous for all three genes) with a plant with purple flowers and green wrinkled seeds (heterozygous for flower color but homozygous recessive for the other two characters). Using Mendelian symbols, our cross is *PpYyRr* × *Ppyyrr*. What fraction of offspring from this cross are predicted to exhibit the recessive phenotypes for *at least two* of the three characters?

To answer this question, we can start by listing all genotypes we could get that fulfill this condition: *ppyyRr, ppYyrr, Ppyyrr, PPyyrr*, and *ppyyrr*. (Because the condition is *at least two* recessive traits, it includes the last genotype, which shows all three recessive traits.) Next, we calculate the probability for each of these genotypes resulting from our *PpYyRr* × *Ppyyrr* cross by multiplying together the individual probabilities for the allele pairs, just as we did in our dihybrid example. Note that in a cross involving heterozygous and homozygous allele pairs (for example, *Yy* × *yy*), the probability of heterozygous (*Yy*) offspring is 1/2 and the probability of homozygous (in this case, *yy*) offspring is 1/2. Finally, we use the addition rule to add the probabilities for all the different genotypes that fulfill the condition of at least two recessive traits resulting from our *PpYyRr* × *Ppyyrr* cross, as shown below:

ppyyRr	1/4 (probability of *pp*) × 1/2 (*yy*) × 1/2 (*Rr*)	= 1/16
ppYyrr	1/4 (*pp*) × 1/2 (*Yy*) × 1/2 (*rr*)	= 1/16
Ppyyrr	1/2 (*Pp*) × 1/2 (*yy*) × 1/2 (*rr*)	= 2/16
PPyyrr	1/4 (*PP*) × 1/2 (*yy*) × 1/2 (*rr*)	= 1/16
ppyyrr	1/4 (*pp*) × 1/2 (*yy*) × 1/2 (*rr*)	= 1/16
Chance of *at least two* recessive traits		= 6/16 = 3/8

In time, you'll be able to solve genetics problems faster by using the rules of probability than by filling in Punnett squares.

We cannot predict with certainty the exact numbers of progeny of different genotypes resulting from a genetic cross. But the rules of probability give us the *likelihood* of various outcomes. Usually, the larger the sample size, the closer the results will conform to our predictions. Mendel understood this statistical feature of inheritance and had a keen sense of the rules of chance. It was for this reason that he set up his experiments so as to generate, and then count, large numbers of offspring from his crosses.

CONCEPT CHECK 14.2

1. For any gene with a dominant allele *A* and recessive allele *a*, what proportions of the offspring from an *AA* × *Aa* cross are expected to be homozygous dominant, homozygous recessive, and heterozygous?
2. Two organisms, with genotypes *BbDD* and *BBDd*, are mated. Assuming independent assortment of the *B*/*b* and *D*/*d* genes, write the genotypes of all possible offspring from this cross and use the rules of probability to calculate the chance of each genotype occurring.
3. **WHAT IF?** Three characters (flower color, seed color, and pod shape) are considered in a cross between two pea plants: *PpYyIi* × *ppYyIi*. Using the rules of probability, determine the fraction of offspring predicted to be homozygous recessive for at least two of the three characters.

For suggested answers, see Appendix A.

CONCEPT 14.3

Inheritance patterns are often more complex than predicted by simple Mendelian genetics

In the 1900s, geneticists extended Mendelian principles not only to diverse organisms, but also to patterns of inheritance more complex than those described by Mendel. For the work that led to his two laws of inheritance, Mendel chose pea plant characters that turn out to have a relatively simple genetic basis: Each character is determined by one gene, for which there are only two alleles, one completely dominant and the other completely recessive. (There is one exception: Mendel's pod shape character is actually determined by two genes.) Not all heritable characters are determined so simply, and the relationship between genotype and phenotype is rarely so straightforward. Mendel himself realized that he could not explain the more complicated patterns he observed in crosses involving other pea characters or other plant species. This does not diminish the utility of Mendelian genetics, however, because the basic principles of segregation and independent assortment apply even to more complex patterns of inheritance. In this section, we will extend Mendelian genetics to hereditary patterns that were not reported by Mendel.

Extending Mendelian Genetics for a Single Gene

The inheritance of characters determined by a single gene deviates from simple Mendelian patterns when alleles are not completely dominant or recessive, when a particular gene has

more than two alleles, or when a single gene produces multiple phenotypes. We will describe examples of each of these situations in this section.

Degrees of Dominance

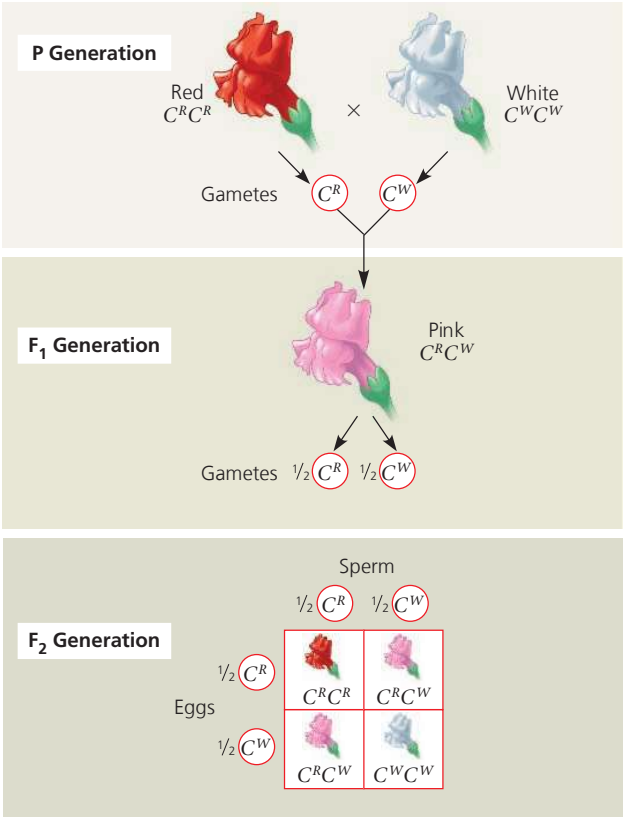
Alleles can show different degrees of dominance and recessiveness in relation to each other. In Mendel's classic pea crosses, the F_1 offspring always looked like one of the two parental varieties because one allele in a pair showed **complete dominance** over the other. In such situations, the phenotypes of the heterozygote and the dominant homozygote are indistinguishable (see Figure 14.6).

For some genes, however, neither allele is completely dominant, and the F_1 hybrids have a phenotype somewhere between those of the two parental varieties. This phenomenon, called **incomplete dominance**, is seen when red snapdragons are crossed with white snapdragons: All the F_1 hybrids have pink flowers **(Figure 14.10)**. This third, intermediate phenotype results from flowers of the heterozygotes having less red pigment than the red homozygotes. (This is unlike the case of Mendel's pea plants, where the *Pp* heterozygotes make enough pigment for the flowers to be purple, indistinguishable from those of *PP* plants.)

At first glance, incomplete dominance of either allele seems to provide evidence for the blending hypothesis of inheritance, which would predict that the red or white trait could never reappear among offspring of the pink hybrids. In fact, interbreeding F_1 hybrids produces F_2 offspring with a phenotypic ratio of one red to two pink to one white. (Because heterozygotes have a separate phenotype from homozygotes, the genotypic and phenotypic ratios for the F_2 generation are the same, 1:2:1.) The segregation of the red-flower and white-flower alleles in the gametes produced by the pink-flowered plants confirms that the alleles for flower color are heritable factors that maintain their identity in the hybrids; that is, that the factors are discrete rather than "blendable."

Another variation on dominance relationships between alleles is called **codominance**; in this variation, the two alleles each affect the phenotype in separate, distinguishable ways. For example, the human MN blood group is determined by codominant alleles for two specific molecules located on the surface of red blood cells, the M and N molecules. A single gene (L), for which two allelic variations are possible (L^M or L^N), determines the phenotype of this blood group. Individuals homozygous for the L^M allele (L^ML^M) have red blood cells with only M molecules; individuals homozygous for the L^N allele (L^NL^N) have red blood cells with only N molecules. But *both* M and N molecules are present on the red blood cells of individuals heterozygous for the *M* and *N* alleles (L^ML^N). Note that the MN phenotype is *not* intermediate between the M and N phenotypes, which distinguishes codominance from incomplete dominance. Rather, *both* M and N phenotypes are exhibited by heterozygotes, since both molecules are present.

▼ Figure 14.10 Incomplete dominance in snapdragon color. When red snapdragons are crossed with white ones, the F_1 hybrids have pink flowers. Segregation of alleles into gametes of the F_1 plants results in an F_2 generation with a 1:2:1 ratio for both genotype and phenotype. Neither allele is dominant, so rather than using upper- and lowercase letters, we use the letter *C* with a superscript to indicate an allele for flower color: C^R for red and C^W for white.



? *Suppose a classmate argues that this figure supports the blending hypothesis for inheritance. What might your classmate say, and how would you respond?*

Mastering Biology **Animation: Incomplete Dominance in "MendAliens"**

The Relationship Between Dominance and Phenotype We've now seen that the relative effects of two alleles range from complete dominance of one allele to incomplete dominance of either allele to codominance of both alleles. It is important to understand that an allele is called *dominant* because it is seen in the phenotype, not because it somehow subdues a recessive allele. Alleles are simply variations in a gene's nucleotide sequence (see Figure 14.4). When a dominant allele coexists with a recessive allele in a heterozygote, they do not actually interact at all. It is in the pathway from genotype to phenotype that dominance and recessiveness come into play.

To illustrate the relationship between dominance and phenotype, we can use one of the characters Mendel

studied—round versus wrinkled pea seed shape. The dominant allele (round) codes for an enzyme that helps convert an unbranched form of starch to a branched form in the seed. The recessive allele (wrinkled) codes for a defective form of this enzyme, leading to an accumulation of unbranched starch, which causes excess water to enter the seed by osmosis. Later, when the seed dries, it wrinkles. If a dominant allele is present, no excess water enters the seed and it does not wrinkle when it dries. One dominant allele results in enough of the enzyme to synthesize adequate amounts of branched starch, which means that dominant homozygotes and heterozygotes have the same phenotype: round seeds.

A closer look at the relationship between dominance and phenotype reveals an intriguing fact: For any character, the observed dominant/recessive relationship of alleles depends on the level at which we examine phenotype. **Tay-Sachs disease**, an inherited disorder in humans, is an example. The brain cells of a child with Tay-Sachs disease cannot metabolize certain lipids because a crucial enzyme does not work properly. As these lipids accumulate in brain cells, the child begins to suffer seizures, blindness, and degeneration of motor and mental performance and dies within a few years.

Only children who inherit two copies of the Tay-Sachs allele (homozygotes) have the disease. Thus, at the *organismal* level, the Tay-Sachs allele qualifies as recessive. However, the activity level of the lipid-metabolizing enzyme in heterozygotes is intermediate between the activity level in individuals homozygous for the normal allele and the activity level in individuals with Tay-Sachs disease. (The term *normal* is used in the genetic sense to refer to the allele coding for the enzyme that functions properly.) The intermediate phenotype observed at the *biochemical* level is characteristic of incomplete dominance of either allele. Fortunately, the heterozygote condition does not lead to disease symptoms, apparently because half the normal enzyme activity is sufficient to prevent lipid accumulation in the brain. Extending our analysis to yet another level, we find that heterozygous individuals produce equal numbers of normal and dysfunctional enzyme molecules. Thus, at the *molecular* level, the normal allele and the Tay-Sachs allele are codominant. As you can see, whether alleles appear to be completely dominant, incompletely dominant, or codominant depends on the level at which the phenotype is analyzed.

Frequency of Dominant Alleles While you might assume that the dominant allele for a particular character would be more common than the recessive allele, this is not always the case. For an example of a rare dominant allele, about one baby out of 400 in the United States is born with extra fingers or toes, a condition known as polydactyly. Some cases are caused by the presence of a dominant allele. The low frequency of these types of polydactyly indicates that the recessive allele, which results in five digits per appendage, is far more prevalent than the dominant allele in the population.

Multiple Alleles

Only two alleles exist for the pea characters that Mendel studied, but most genes have more than two alleles. The ABO blood groups in humans, for instance, are determined by the two alleles a person has of the blood group gene; the three possible alleles are I^A, I^B, and i. A person's blood group may be one of four types: A, B, AB, or O. These letters refer to two carbohydrates—A and B—that may be found attached to specific cell-surface molecules on red blood cells. An individual's blood cells may have carbohydrate A (type A blood), carbohydrate B (type B), both (type AB), or neither (type O), as shown in **Figure 14.11**, along with the relevant genotypes. Matching compatible blood groups is critical for safe blood transfusions (see Concept 43.3).

▼ Figure 14.11 Multiple alleles for the ABO blood groups. The four blood groups result from different combinations of three alleles.

(a) The three alleles for the ABO blood groups and their carbohydrates. Each allele codes for an enzyme that may add a specific carbohydrate (designated by the superscript on the allele and shown as a triangle or circle) to red blood cells.

Allele	I^A	I^B	i
Carbohydrate	A △	B ○	none

(b) Blood group genotypes and phenotypes. There are six possible genotypes, resulting in four different phenotypes.

Genotype	I^AI^A or I^Ai	I^BI^B or I^Bi	I^AI^B	ii
Red blood cell with surface carbohydrates				
Phenotype (blood group)	A	B	AB	O

VISUAL SKILLS *Based on the surface carbohydrate phenotype in part (b), what are the dominance relationships among the alleles?*

Pleiotropy

So far, we have treated Mendelian inheritance as though each gene affects only one phenotypic character. Most genes, however, have multiple phenotypic effects, a property called **pleiotropy** (from the Greek *pleion*, more). In humans, for example, pleiotropic alleles are responsible for the multiple symptoms associated with certain hereditary diseases, such as cystic fibrosis and sickle-cell disease, discussed later in this chapter. In the garden pea, the gene that determines flower color also affects the color of the coating on the outer surface of the seed, which can be gray or white. Given the intricate molecular and cellular interactions responsible for an organism's development and physiology, it isn't surprising that a single gene can affect a number of characters.

Extending Mendelian Genetics for Two or More Genes

Dominance relationships, multiple alleles, and pleiotropy all have to do with the effects of the alleles of a single gene. We now consider two situations in which two or more genes are involved in determining a particular phenotype. In the first case, called epistasis, one gene affects the phenotype of another because the two gene products interact; in the second case, called polygenic inheritance, multiple genes independently affect a single trait.

Epistasis

In **epistasis** (from the Greek for "standing upon"), the phenotypic expression of a gene at one locus alters that of a gene at a second locus. An example will help clarify this concept. In Labrador retrievers (commonly called "Labs"), black coat color is dominant to brown. Let's designate *B* and *b* as the two alleles for this character. For a Lab to have brown fur, its genotype must be *bb*; these dogs are called chocolate Labs. But there is more to the story. A second gene determines whether or not pigment will be deposited in the hair. The dominant allele, symbolized by *E*, results in the deposition of either black or brown pigment, depending on the genotype at the first locus. But if the Lab is homozygous recessive for the second locus (*ee*), then the coat is yellow, regardless of the genotype at the black/brown locus (yellow Labs). In this case, the gene for pigment deposition (*E*/*e*) is said to be *epistatic to* the gene that codes for black or brown pigment (*B*/*b*).

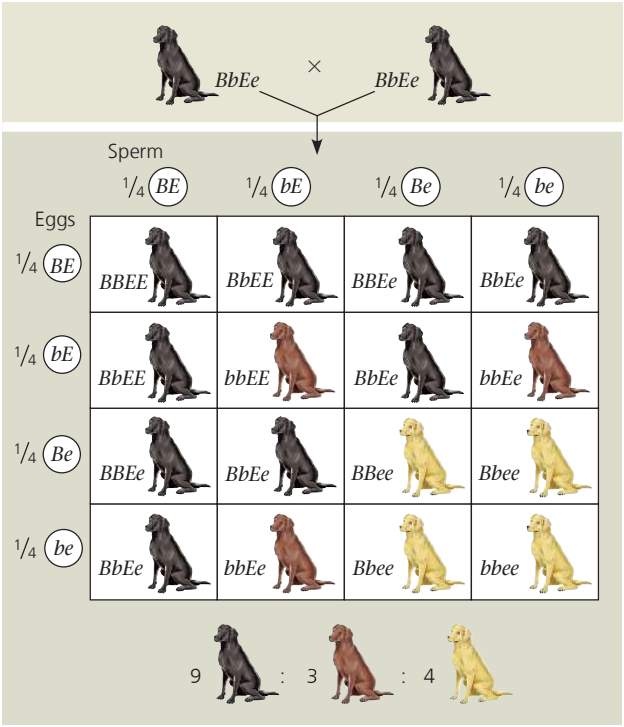
What happens if we mate black Labs that are heterozygous for both genes (*BbEe*)? Although the two genes affect the same phenotypic character (coat color), they follow the law of independent assortment. Thus, our breeding experiment represents an F_1 dihybrid cross, like those that produced a 9:3:3:1 ratio in Mendel's experiments. We can use a Punnett square to represent the genotypes of the F_2 offspring **(Figure 14.12)**. As a result of epistasis, the phenotypic ratio among the F_2 offspring is 9 black to 3 chocolate to 4 yellow Labs. Other types of epistatic interactions produce different ratios, but all are modified versions of 9:3:3:1.

Polygenic Inheritance

Mendel studied characters that could be classified on an either-or basis, such as purple versus white flower color. But many characters, such as human skin color and height, are not one of two discrete characters, but instead vary in the population in gradations along a continuum. These are called **quantitative characters**. Quantitative variation usually indicates **polygenic inheritance**, an additive effect of two or more genes on a single phenotypic character. (In a way, this is the converse of pleiotropy, where a single gene affects several phenotypic characters.) Height is a good example of polygenic inheritance: In 2014, a genomic study of over 250,000 people found almost 700 genetic variations in over 180 genes that affect height. Many variations were in or near genes involved in biochemical pathways affecting growth of the skeleton, but others were associated with genes not obviously related to growth. Another study in 2018, of 300,000 individuals, identified 124 genes that affect hair color. Although simplifying can sometimes be valuable in understanding genetic principles, as we'll see shortly, the classic idea of a single gene determining eye color or hair color or other such traits is an oversimplification. (Earlobe attachment, an example used until a decade ago in this textbook, was shown in 2017 to be affected by nearly 50 genes!)

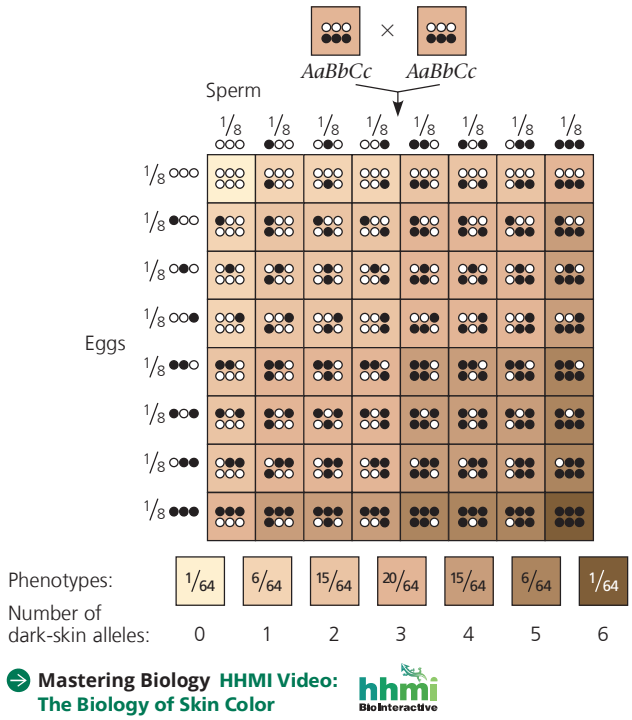
Skin pigmentation in humans is also controlled by many separately inherited genes—378 at latest count, many of which are involved in the production of melanin skin pigments. Here, we'll simplify the story in order to understand the concept of polygenic inheritance. Let's consider three genes, with a dark-skin allele for each gene (*A, B*, or *C*) contributing one "unit" of darkness (also a simplification) to the phenotype and being incompletely dominant to the other, light-skin allele (*a, b*, or *c*). In our model, an *AABBCC* person would be very dark, whereas an *aabbcc* individual would be very light.

▼ Figure 14.12 An example of epistasis. This Punnett square illustrates the genotypes and phenotypes predicted for offspring of matings between two black Labrador retrievers of genotype *BbEe*. The *E*/*e* gene, which is epistatic to the *B*/*b* gene coding for hair pigment, controls whether or not pigment of any color will be deposited in the hair.



VISUAL SKILLS *Compare the four squares in the lower right of this Punnett square with those in Figure 14.8. Explain the genetic basis for the difference between the ratio (9:3:4) of phenotypes seen in this cross and the 9:3:3:1 ratio seen in Figure 14.8.*

▼ Figure 14.13 A simplified model for polygenic inheritance of skin color. In this model, three separately inherited genes affect skin color. (The reported number is actually 378 genes.) The heterozygous individuals (*AaBbCc*) represented by the two rectangles at the top of this figure each carry three dark-skin alleles (black circles, which represent *A*, *B*, or *C*) and three light-skin alleles (white circles, which represent *a*, *b*, or *c*). The Punnett square shows all possible genetic combinations in gametes and offspring of hypothetical matings between these two. The results are summarized by phenotypic frequencies (fractions) under the Punnett square. (The phenotypic ratio is 1:6:15:20:15:6:1.)



Mastering Biology HHMI Video: The Biology of Skin Color

An *AaBbCc* person would have skin of an intermediate shade. Because the alleles have a cumulative effect, the genotypes *AaBbCc* and *AABbcc* would make the same genetic contribution (three units) to skin darkness. There are seven skin color phenotypes that could result from a mating between *AaBbCc* heterozygotes, as shown in **Figure 14.13**. In a large number of such matings, the majority of offspring would be expected to have intermediate phenotypes (skin color in the middle range). You can graph the predictions from the Punnett square in the **Scientific Skills Exercise**. Environmental factors, such as exposure to the sun, also affect the skin color phenotype.

Nature and Nurture: The Environmental Impact on Phenotype

Another departure from simple Mendelian genetics arises when the phenotype for a character depends on environment as well as genotype. A single tree, locked into its inherited genotype, has leaves that vary in size, shape, and greenness, depending on their exposure to wind and sun. For humans, nutrition influences height, exercise alters build, sun-tanning darkens the skin, and experience improves performance on intelligence tests. Even identical twins, who are genetic equals, accumulate phenotypic differences as a result of their unique experiences.

Whether human characters are more influenced by genes or the environment—in everyday terms, nature versus nurture—is a debate that we will not attempt to settle here. We can say, however, that a genotype generally is not associated with a rigidly defined phenotype, but rather with a range of phenotypic possibilities due to environmental influences **(Figure 14.14)**. For some characters, such as the ABO blood group system, the phenotypic range is extremely narrow; that is, a given genotype mandates a very specific phenotype. Other characters, such as a person's blood count of red and white cells, vary quite a bit, depending on such factors as the altitude, the customary level of physical activity, and the presence of infectious agents.

▼ Figure 14.14 The effect of environment on phenotype. The outcome of a genotype lies within a phenotypic range that depends on the environment in which the genotype is expressed. For example, the acidity and free aluminum content of the soil affect the color of hydrangea flowers, which range from pink (basic soil) to blue-violet (acidic soil). Free aluminum is necessary for bluer colors.



(a) Hydrangeas grown in basic soil



(b) Hydrangeas of the same genetic variety grown in acidic soil with free aluminum

Generally, the phenotypic range is broadest for polygenic characters. Environment contributes to the quantitative nature of these characters, as we have seen in the continuous variation of skin color. Geneticists refer to such characters as **multifactorial**, meaning that many factors, both genetic and environmental, collectively influence phenotype.

Mastering Biology BBC Video: Genetics vs. Environment

A Mendelian View of Heredity and Variation

We have now broadened our view of Mendelian inheritance by exploring degrees of dominance as well as multiple alleles, pleiotropy, epistasis, polygenic inheritance, and the phenotypic impact of the environment. How can we integrate these refinements into a comprehensive theory of Mendelian genetics? The key is to make the transition from the reductionist emphasis on single genes and phenotypic characters to the emergent properties of the organism as a whole, one of the themes of this book.

The term *phenotype* can refer not only to specific characters, such as flower color and blood group, but also to

Scientific Skills Exercise

Making a Histogram and Analyzing a Distribution Pattern



What Is the Distribution of Phenotypes Among Offspring of Two Parents Who Are Both Heterozygous for Three Additive Genes? Human skin color is a polygenic trait that is determined by the additive effects of many different genes. In this exercise, you will work with a simplified model of skin color genetics where only three genes are assumed to affect the darkness of skin color and where each gene has two alleles—dark or light (see Figure 14.13). In this model, each dark allele contributes equally to the darkness of skin color, and each pair of alleles segregates independently of any other pair. Using a type of graph called a histogram, you will determine the distribution of phenotypes of offspring with different numbers of dark-skin alleles. (For additional information about graphs, see the Scientific Skills Review in Appendix D.)

How This Model Is Analyzed To predict the phenotypes of the offspring of parents heterozygous for the three genes in our simplified model, we can use the Punnett square in Figure 14.13. The heterozygous individuals (*AaBbCc*) represented by the two rectangles at the top of that figure each carry three dark-skin alleles (black circles, which represent *A, B*, or *C*) and three light-skin alleles (white circles, which represent *a, b*, or *c*). The Punnett square shows all the possible genetic combinations in gametes and in offspring of a large number of hypothetical matings between these heterozygotes.

Predictions from the Punnett Square If we assume that each square in the Punnett square represents one offspring of the heterozygous *AaBbCc* parents, then the squares below show the possible skin color phenotypes and their predicted frequencies. Below the squares is the number of dark-skin alleles for each phenotype.

Phenotypes:	1/64	6/64	15/64	20/64	15/64	6/64	1/64
Number of dark-skin alleles:	0	1	2	3	4	5	6

INTERPRET THE DATA

1. A histogram is a bar graph that shows the distribution of numeric data (here, the number of dark-skin alleles). To make a histogram of the allele distribution, put skin color (as the number of dark-skin alleles) along the *x*-axis and predicted number of offspring (out of 64) with each phenotype on the *y*-axis. There are no gaps in these allele data, so draw the bars next to each other with no space in between.
2. You can see that the skin-color phenotypes are not distributed uniformly. **(a)** Which phenotype has the highest frequency? Draw a vertical dashed line through that bar. **(b)** Distributions of values like this one tend to show one of several common patterns. Sketch a rough curve that approximates the values and look at its shape. Is it symmetrically distributed around a central peak value (a "normal distribution," sometimes called a bell curve); is it skewed to one end of the *x*-axis or the other (a "skewed distribution"); or does it show two apparent groups of frequencies (a "bimodal distribution")? Explain the reason for the curve's shape. (It will help to read the text description that supports Figure 14.13.)

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.

Further Reading R. A. Sturm, A golden age of human pigmentation genetics, *Trends in Genetics* 22:464–468 (2006).

an organism in its entirety—*all* aspects of its physical appearance, internal anatomy, physiology, and behavior. Similarly, the term *genotype* can refer to an organism's entire genetic makeup, not just its alleles for a single genetic locus. In most cases, a gene's impact on phenotype is affected by other genes and by the environment. In this integrated view of heredity and variation, an organism's phenotype reflects its overall genotype and unique environmental history.

Considering all that can occur in the pathway from genotype to phenotype, it is indeed impressive that Mendel could uncover the fundamental principles governing the transmission of individual genes from parents to offspring. Mendel's laws of segregation and of independent assortment explain heritable variations in terms of alternative forms of genes (hereditary "particles," now known as the alleles of genes) that are passed along, generation after generation, according to simple rules of probability. This theory of inheritance is equally valid for peas, flies, fishes, birds, and human beings—indeed, for any organism with a sexual life cycle. Furthermore, by extending the principles of segregation and independent assortment to help explain such hereditary patterns as epistasis and quantitative characters, we begin to see how broadly Mendelian genetics applies. From Mendel's abbey garden came a theory of inheritance that anchors modern genetics. In the last section of this chapter, we will apply Mendelian genetics to human inheritance, with emphasis on the transmission of hereditary diseases.

CONCEPT CHECK 14.3

1. *Incomplete dominance* and *epistasis* are both terms that define genetic relationships. What is the most basic distinction between these terms?
2. If a man with type AB blood marries a woman with type O, what blood types would you expect in their children? What fraction would you expect of each type?
3. **WHAT IF?** A rooster with gray feathers and a hen of the same phenotype produce 15 gray, 6 black, and 8 white chicks. What is the simplest explanation for the inheritance of these colors in chickens? What phenotypes would you expect in the offspring of a cross between a gray rooster and a black hen?

For suggested answers, see Appendix A.

CONCEPT 14.4

Many human traits follow Mendelian patterns of inheritance

The study of human genetics is fueled by our desire to understand our own inheritance and to develop treatments for genetically based diseases. While peas are convenient subjects for genetic research, humans are not. The human generation span is long (about 20 years), human parents produce many fewer offspring than peas, and breeding experiments would be unethical!

Pedigree Analysis

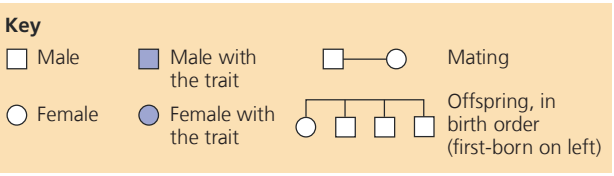
In place of breeding experiments, geneticists analyze the results of human matings that have already occurred. They collect information about a family's history for a particular trait and assemble this information into a family tree describing the trait across the generations—a family **pedigree**.

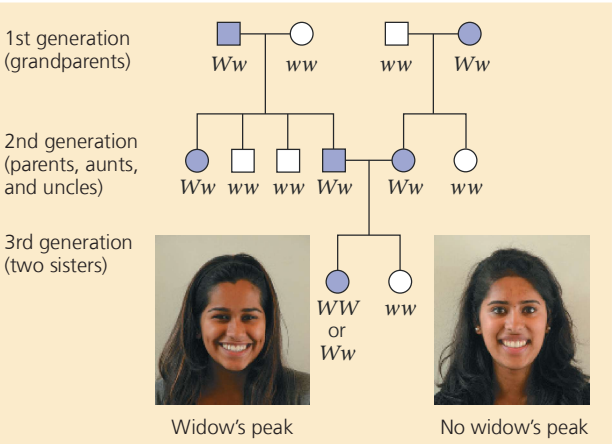
Figure 14.15a shows a three-generation pedigree that traces the occurrence of a pointed contour of the hairline on the forehead. We simplify the genetics of this trait, called a widow's peak, by assuming it is due to a dominant allele, *W*. (In reality, other genes are probably also involved.) In our model, because the widow's peak allele is dominant, all individuals who lack a widow's peak must be homozygous recessive (*ww*). The two grandparents with widow's peaks must have the *Ww* genotype since some of their offspring are homozygous recessive. The offspring in the second generation who *do* have widow's peaks must also be heterozygous because they are the products of $Ww \times ww$ matings. The third generation in this pedigree consists of two sisters. The one who has a widow's peak could be either homozygous (*WW*) or heterozygous (*Ww*), given what we know about the genotypes of her parents (both *Ww*).

Figure 14.15b is a pedigree of the same family, but this time we focus on a recessive trait, the inability of individuals to taste a chemical called PTC (phenylthiocarbamide). Compounds similar to PTC are found in broccoli, brussels sprouts, and related vegetables and account for the bitter taste some people report when eating these foods. We'll use *t* for the recessive allele and *T* for the dominant allele, which results in the ability to taste PTC. As you work your way through the pedigree, notice once again that you can apply what you have learned about Mendelian inheritance to understand the genotypes shown for the family members.

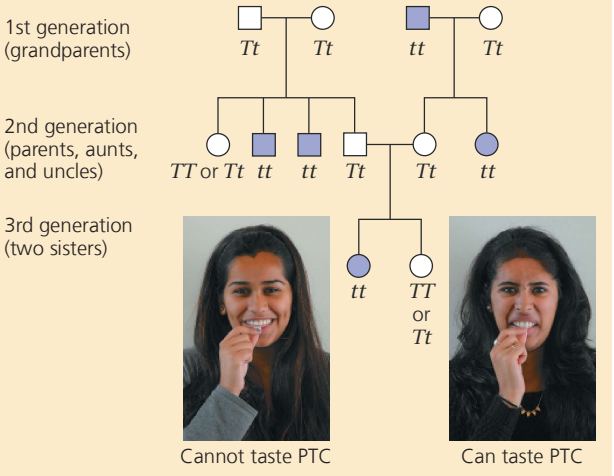
An important application of a pedigree is to help us calculate the probability that a future child will have a particular genotype and phenotype. Suppose that the couple represented in the second generation of Figure 14.15 decides to have one more child. What is the probability that the child will have a widow's peak? This is equivalent to a Mendelian F_1

▼ Figure 14.15 Pedigree analysis. Each pedigree traces a trait through three generations of the same family; the two traits have different inheritance patterns. For the sake of understanding genetic principles, each trait is shown as being determined by two alleles of a single gene—a simplification, since other genes likely also affect these two characters.





(a) Is a widow's peak a dominant or recessive trait?
Tips for pedigree analysis: Notice in the third generation that the second-born daughter lacks a widow's peak, although both of her parents had the trait. Such a pattern indicates that the trait is due to a dominant allele. If it were due to a *recessive* allele, and both parents had the recessive phenotype (straight hairline), *all* of their offspring would also have the recessive phenotype.



(b) Is the inability to taste a chemical called PTC a dominant or recessive trait?
Tips for pedigree analysis: Notice that the first-born daughter in the third generation has the trait (is unable to taste PTC), although both parents lack that trait (they *can* taste PTC). Such a pattern is explained if the non-taster phenotype is due to a recessive allele. If it were due to a *dominant* allele, then at least one parent would also have had the trait.

monohybrid cross ($Ww \times Ww$), and therefore the probability that a child will inherit a dominant allele and have a widow's peak is $\frac{3}{4}$ ($\frac{1}{4}$ WW + $\frac{1}{2}$ Ww). What is the probability that the child will be unable to taste PTC? We can also treat this as a monohybrid cross ($Tt \times Tt$), but this time we want to know the chance that the offspring will be homozygous recessive (tt). That probability is $\frac{1}{4}$. Finally, what is the chance that the child will have a widow's peak *and* be unable to taste PTC? Assuming that the genes for these two characters are on different chromosomes, the two pairs of alleles will assort independently in this dihybrid cross ($WwTt \times WwTt$). Therefore, we can use the multiplication rule: $\frac{3}{4}$ (chance of widow's peak) $\times$ $\frac{1}{4}$ (chance of inability to taste PTC) = $\frac{3}{16}$ (chance of widow's peak and inability to taste PTC).

Pedigrees are a more serious matter when the alleles in question cause disabling or deadly diseases instead of innocuous human variations such as hairline or inability to taste a harmless chemical. However, for disorders inherited as simple Mendelian traits, the same techniques of pedigree analysis apply.

Recessively Inherited Disorders

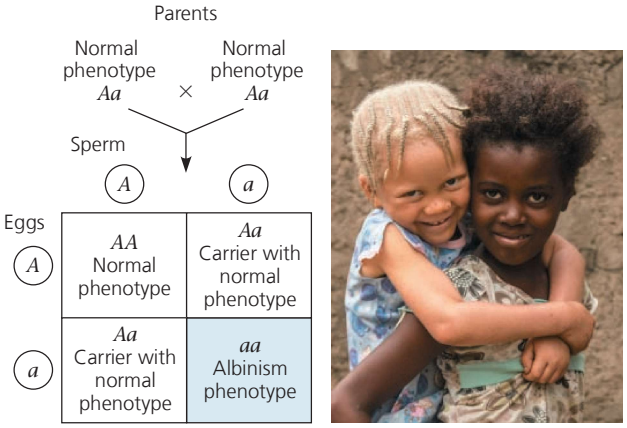
Thousands of genetic disorders are known to be inherited as simple recessive traits. These disorders range in severity from relatively mild, such as albinism (lack of pigmentation, which results in susceptibility to skin cancers and vision problems), to life-threatening, such as cystic fibrosis.

The Behavior of Recessive Alleles

How can we account for the behavior of alleles that cause recessively inherited disorders? Recall that genes code for proteins of specific function. An allele that causes a genetic disorder (let's call it allele *a*) codes for either a malfunctioning protein or no protein at all. In the case of disorders classified as recessive, heterozygotes (*Aa*) typically have the normal phenotype because one copy of the normal allele (*A*) produces a sufficient amount of the specific protein. Thus, a recessively inherited disorder shows up only in the homozygous individuals (*aa*) who inherit a recessive allele from each parent. Although phenotypically normal with regard to the disorder, heterozygotes may transmit the recessive allele to their offspring and thus are called **carriers**. **Figure 14.16** illustrates these ideas using albinism as an example.

Most people who have recessive disorders are born to parents who are carriers of the disorder but have a normal phenotype, as is the case shown in the Punnett square in Figure 14.16. A mating between two carriers corresponds to a Mendelian F_1 monohybrid cross, so the predicted genotypic ratio for the offspring is 1 *AA* : 2 *Aa* : 1 *aa*. Thus, each child has a $\frac{1}{4}$ chance of inheriting a double dose of the recessive allele; in the case of albinism, such a child will have albinism. From the genotypic ratio, we also can see that out of three offspring with the *normal* phenotype (one *AA* plus two *Aa*),

▼ Figure 14.16 Albinism: a recessive trait. One of the two sisters shown here does not have albinism; the other does. Most recessive homozygotes are born to parents who are carriers of the disorder but themselves have a normal phenotype, the case shown in the Punnett square.



? *What is the probability that the sister without albinism is a carrier of the albinism allele?*

two are predicted to be heterozygous carriers, $\frac{2}{3}$ chance. Recessive homozygotes could also result from $Aa \times aa$ and $aa \times aa$ matings, but if the disorder is lethal before reproductive age or results in sterility (neither of which is true for albinism), no *aa* individuals will reproduce. Even if recessive homozygotes are able to reproduce, this will occur relatively rarely because such individuals are much less common in the population than heterozygous carriers (for reasons we'll examine in Concept 23.2).

In general, genetic disorders are not evenly distributed among all groups of people. For example, the incidence of Tay-Sachs disease, which we described earlier in this chapter, is disproportionately high among Ashkenazic Jews, Jewish people whose ancestors lived in central Europe. In that population, Tay-Sachs disease occurs in one out of 3,600 births, an incidence about 100 times greater than that among non-Jews or Mediterranean (Sephardic) Jews. This uneven distribution results from the different genetic histories of the world's peoples during less technological times, when populations were more geographically (and hence genetically) isolated.

When a disease-causing recessive allele is rare, it is relatively unlikely that two carriers of the same harmful allele will meet and mate. The probability of passing on recessive traits increases greatly, however, if the man and woman are close relatives (for example, siblings or first cousins). This is because people with recent common ancestors are more likely to carry the same recessive alleles than are unrelated people. Thus, these consanguineous ("same blood") matings, indicated in pedigrees by double lines, are more likely to produce offspring homozygous for recessive traits—including harmful ones. Such effects can be observed in many types of domesticated and zoo animals that have become inbred.

Although geneticists generally agree that inbreeding causes an increase in autosomal recessive conditions compared to those resulting from matings between unrelated parents, they debate exactly how much human consanguinity increases the risk of inherited diseases. For one thing, many harmful alleles have such severe effects that a homozygous embryo spontaneously aborts long before birth. Most societies and cultures have laws or taboos forbidding marriages between close relatives, some for social or economic reasons. These rules may have also resulted from the empirical observation that in most populations, stillbirths and physical or biochemical abnormalities at birth are more common when parents are closely related.

Cystic Fibrosis

The most common lethal genetic disease in the United States is **cystic fibrosis**, which strikes one out of every 2,500 people of European descent but is much rarer in other groups. Among people of European descent, one out of 25 (4%) are carriers of the cystic fibrosis allele. The normal allele for this gene codes for a membrane protein that functions in the transport of chloride ions between certain cells and the extracellular fluid. These chloride transport channels are defective or absent in the plasma membranes of children who inherit two recessive alleles for cystic fibrosis. The result is an abnormally high concentration of intracellular chloride, which causes an uptake of water due to osmosis. This in turn causes the mucus that coats certain cells to become thicker and stickier than normal and to build up in the pancreas, lungs, digestive tract, and other organs. Multiple (pleiotropic) effects result, including poor absorption of nutrients from the intestines, chronic bronchitis, and recurrent bacterial infections.

Untreated, cystic fibrosis can cause death by the age of 5. Daily doses of antibiotics to stop infection, gentle pounding on the chest to clear mucus from clogged airways, and other therapies can prolong life. In the United States, more than half of those with cystic fibrosis now survive into their 40s.

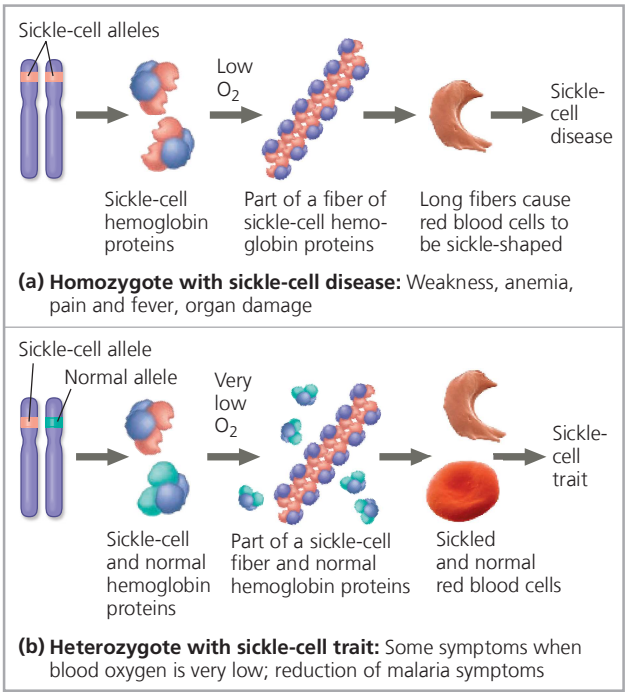
Sickle-Cell Disease: A Genetic Disorder with Evolutionary Implications

EVOLUTION The most common inherited disorder among people of African descent is **sickle-cell disease**, which affects one out of 400 African-Americans. Sickle-cell disease is caused by the substitution of a single amino acid in the hemoglobin protein of red blood cells; in homozygous individuals, all hemoglobin is of the sickle-cell (abnormal) variety. When the oxygen content of an affected individual's blood is low (at high altitudes or under physical stress, for instance), the sickle-cell hemoglobin proteins aggregate into long fibers that deform the red cells into a sickle shape (see Figure 5.19). Sickled cells may clump and clog small blood vessels, often leading to other symptoms throughout the body, including physical weakness, pain, organ damage, and even stroke and paralysis. Regular blood transfusions can ward off brain damage in children with sickle-cell disease, and new drugs can help prevent or treat other problems. There is currently no widely available cure, but the disease is the target of ongoing gene therapy research.

Although two sickle-cell alleles are necessary for an individual to have sickle-cell disease and thus the condition is considered recessive, the presence of one sickle-cell allele can affect the phenotype. Thus, at the organismal level, the normal allele is incompletely dominant to the sickle-cell allele **(Figure 14.17)**. At the molecular level, the two alleles are codominant; both normal and abnormal (sickle-cell) hemoglobins are made in heterozygotes (carriers), who are said to have *sickle-cell trait*. (Here the word "trait" is used to distinguish this condition from sickle-cell disease, not as it was defined earlier—as any variant of a character.) Heterozygotes are usually healthy but may suffer some symptoms during long periods of reduced blood oxygen.

About one out of ten African-Americans have sickle-cell trait, an unusually high frequency of heterozygotes for an allele with severe detrimental effects in homozygotes. Why haven't evolutionary processes resulted in the disappearance of the allele among this population? One explanation is that having a single copy of the sickle-cell allele reduces the frequency and severity of malaria attacks, especially among young children. The malaria parasite spends part of its life cycle in red blood cells (see Figure 28.18), and the presence of even heterozygous amounts of sickle-cell hemoglobin results in lower parasite densities and hence reduced malaria symptoms.

▼ Figure 14.17 Sickle-cell disease and sickle-cell trait.



(a) Homozygote with sickle-cell disease: Weakness, anemia, pain and fever, organ damage

(b) Heterozygote with sickle-cell trait: Some symptoms when blood oxygen is very low; reduction of malaria symptoms

Mastering Biology **HHMI Animation: Sickle-Cell Disease**

Thus, in tropical Africa, where infection with the malaria parasite is common, the sickle-cell allele confers an advantage to heterozygotes even though it is harmful in the homozygous state. (The balance between these two effects will be discussed in Concept 23.4; see Make Connections Figure 23.18.) The relatively high frequency of African-Americans with sickle-cell trait is a vestige of their tropical African ancestry.

Mastering Biology **HHMI Video: The Making of the Fittest: Natural Selection in Humans (Sickle-Cell Disease)**

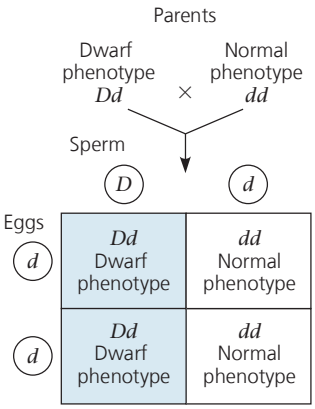
Dominantly Inherited Disorders

Although many harmful alleles are recessive, some disorders are due to dominant alleles. One example is *achondroplasia*, a form of dwarfism that occurs in one of every 25,000 people. Heterozygous individuals have the dwarf phenotype **(Figure 14.18)**. Therefore, all people who do not have achondroplasia—99.99% of the population—are homozygous recessive. Like the presence of extra fingers or toes mentioned earlier, achondroplasia is a trait for which the recessive allele is much more prevalent than the corresponding dominant allele.

Unlike achondroplasia, which is relatively harmless, some dominant alleles cause lethal diseases. Those that do are much less common than recessive alleles that have lethal effects. A lethal recessive allele is only lethal when homozygous; it can be passed from one generation to the next by heterozygous carriers because the carriers themselves have normal phenotypes. A lethal dominant allele, however, often causes the death of afflicted individuals before they can mature and reproduce, and in this case the allele is not passed on to future generations.

A lethal dominant allele may be passed on, though, if the lethal disease symptoms first appear after reproductive age.

▼ Figure 14.18 Achondroplasia: a dominant trait. Dr. Michael C. Ain has achondroplasia, a form of dwarfism caused by a dominant allele. This has inspired his work: He is a specialist in the repair of bone defects caused by achondroplasia and other disorders. The dominant allele (*D*) might have arisen as a mutation in the egg or sperm of a parent or could have been inherited from an affected parent, as shown for an affected father in the Punnett square.





In these cases, the individual may already have transmitted the allele to his or her children. For example, a degenerative disease of the nervous system, called **Huntington's disease**, is caused by a lethal dominant allele that has no obvious phenotypic effect until the individual is about 35 to 45 years old. Once the deterioration of the nervous system begins, it is irreversible and inevitably fatal. As with other dominant traits, a child born to a parent with the Huntington's disease allele has a 50% chance of inheriting the allele and the disorder (see the Punnett square in Figure 14.18). In the United States, this disease afflicts about one in 10,000 people.

At one time, the onset of symptoms was the only way to know if a person had inherited the Huntington's allele, but this is no longer the case. By analyzing DNA samples from a large family with a high incidence of the disorder, geneticists tracked the Huntington's allele to a locus near the tip of chromosome 4, and the gene was sequenced in 1993. This information led to the development of a test that could detect the presence of the Huntington's allele in an individual's genome. (The methods that make such tests possible are discussed in Concepts 20.1 and 20.4.) The availability of this test poses an agonizing dilemma for those with a family history of Huntington's disease. Some individuals may want to be tested for this disease, whereas others may decide it would be too stressful to find out.

Mastering Biology **Interview with Nancy Wexler: Mapping the gene that causes Huntington's disease**



Multifactorial Disorders

The hereditary diseases we have discussed so far are sometimes described as simple Mendelian disorders because they result from abnormality of one or both alleles at a single genetic locus. Many more people are susceptible to diseases that have a multifactorial basis—a genetic component plus a significant environmental influence. Heart disease, diabetes, cancer, alcoholism, certain mental illnesses such as schizophrenia and bipolar disorder, and many other diseases are multifactorial. In these cases, the hereditary component is polygenic. For example, many genes affect cardiovascular health, making some of us more prone than others to heart attacks and strokes. No matter what our genotype, however, our lifestyle has a tremendous effect on phenotype for cardiovascular health and other multifactorial characters. Exercise, a healthful diet, abstinence from smoking, and an ability to handle stressful situations all reduce our risk of heart disease and some types of cancer.

Genetic Testing and Counseling

Avoiding simple Mendelian disorders is possible when the risk of a particular genetic disorder can be assessed before a child is conceived or during the early stages of the pregnancy. Many hospitals have genetic counselors who can provide

information to prospective parents concerned about a family history for a specific disease. Fetal and newborn testing can also reveal genetic disorders.

Counseling Based on Mendelian Genetics and Probability Rules

Consider the case of a hypothetical couple, Tyler and Lily. Each had a brother who died from the same recessively inherited lethal disease. Before conceiving their first child, Tyler and Lily seek genetic counseling to determine the risk of having a child with the disease. From the information about their brothers, we know that both parents of Tyler and both parents of Lily must have been carriers of the recessive allele. Thus, Tyler and Lily are both products of $Aa \times Aa$ crosses, where *a* symbolizes the allele that causes this particular disease. We also know that Tyler and Lily are not homozygous recessive (*aa*) because they do not have the disease. Therefore, their genotypes are either *AA* or *Aa*.

Given a genotypic ratio of 1 *AA* : 2 *Aa* : 1 *aa* for offspring of an $Aa \times Aa$ cross, Tyler and Lily each have a $\frac{2}{3}$ chance of being carriers (*Aa*). According to the rule of multiplication, the overall probability of their firstborn having the disorder is $\frac{2}{3}$ (the chance that Tyler is a carrier) times $\frac{2}{3}$ (the chance that Lily is a carrier) times $\frac{1}{4}$ (the chance of two carriers having a child with the disease), which equals $\frac{1}{9}$. Suppose that Lily and Tyler decide to have a child—after all, there is an $\frac{8}{9}$ chance that their baby will not have the disorder. If, despite these odds, their child is born with the disease, then we would know that *both* Tyler and Lily are, in fact, carriers (*Aa* genotype). If both Tyler and Lily are carriers, there is a $\frac{1}{4}$ chance that any subsequent child this couple has will have the disease. The probability is higher for subsequent children because the diagnosis of the disease in the first child established that both parents are carriers, not because the genotype of the first child affects in any way that of future children.

When we use Mendel's laws to predict possible outcomes of matings, it is important to remember that each child represents an independent event in the sense that its genotype is unaffected by the genotypes of older siblings. Suppose that Tyler and Lily have three more children, and *all three* have the hypothetical hereditary disease. There is only one chance in 64 ($\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$) that such an outcome will occur. Despite these circumstances, the chance that a fourth child of this couple will have the disease remains $\frac{1}{4}$.

Tests for Identifying Carriers

Most children with recessive disorders are born to parents with normal phenotypes. The key to accurately assessing the genetic risk for a particular disease is therefore to find out whether the prospective parents are heterozygous carriers of the recessive allele. Tests are available that can distinguish individuals of normal phenotype who are dominant homozygotes from those who are heterozygous carriers, using blood or cells from the inside of the cheek. Hundreds of different recessive alleles can be detected, including those for Tay-Sachs disease, sickle-cell disease, and cystic fibrosis.

These tests for identifying carriers enable people with family histories of genetic disorders to make informed decisions about having children, including whether to do genetic testing of the fetus, should they decide to become pregnant. The results can also lead carriers to alert family members that they may also be carriers and might choose to be tested. The tests raise other issues as well: Could carriers be denied health or life insurance or lose the jobs providing those benefits, even though they themselves are healthy? The Genetic Information Nondiscrimination Act, in place since 2008, allays these concerns by prohibiting discrimination in employment or health insurance coverage based on genetic test results. This is an important act that protects patients' rights but is unknown to many, including physicians. Genetic counselors, trained in a growing number of professional programs, can help individuals understand their genetic test results. Once test results are clearly understood, those with certain results may face difficult decisions. Advances in biotechnology offer the potential to reduce human suffering, but along with them come ethical issues that require conscientious deliberation.

Fetal Testing

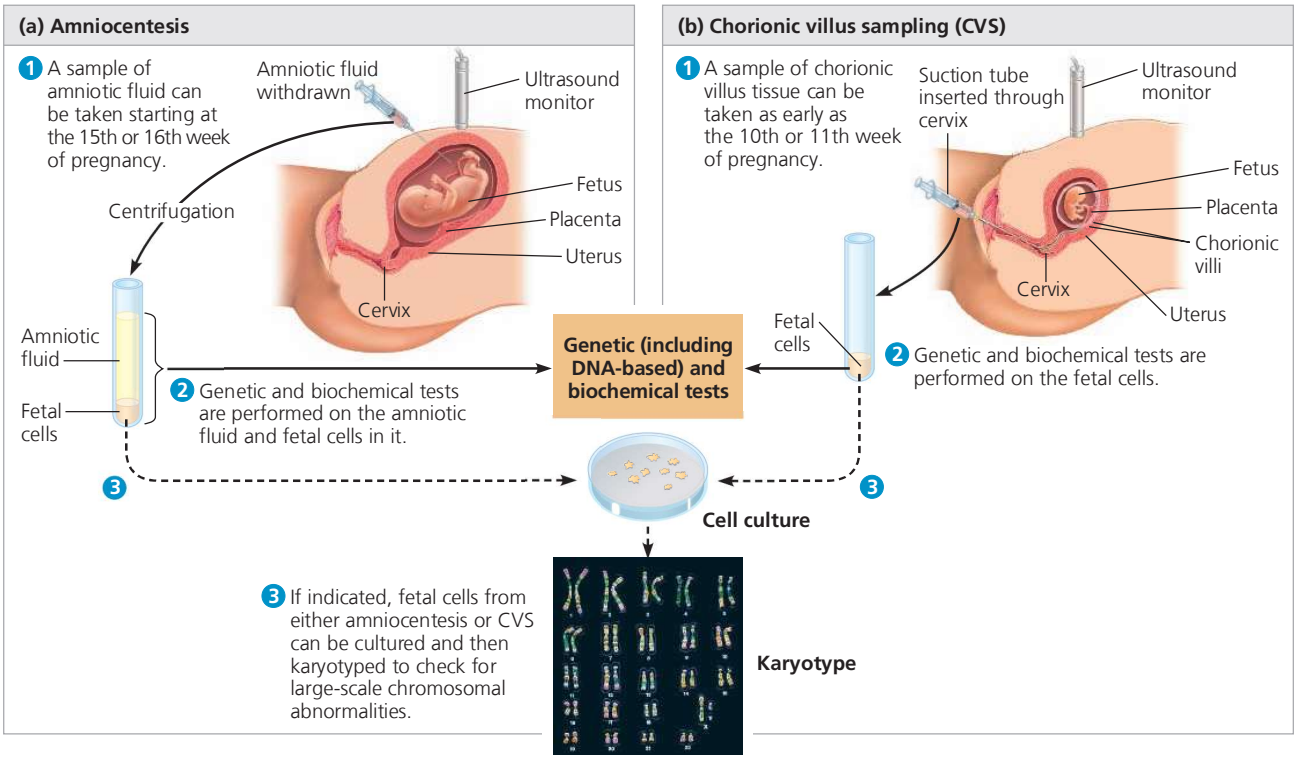
Two types of testing can be done to look for genetic disorders in the fetus: screening tests and diagnostic tests. Screening tests are generally noninvasive and can be followed by diagnostic tests if screening results suggest that a genetic disorder might be present or if there are other risks, as when both parents are carriers.

Screening tests include imaging and blood tests. Imaging techniques allow a physician to examine a fetus directly for major anatomical abnormalities that might not show up in genetic tests. In the *ultrasound* technique, for example, reflected sound waves are used to produce an image of the fetus by a simple noninvasive procedure. An ultrasound is often accompanied by a blood test that looks for fetal proteins that might signal a genetic disorder.

Medical scientists have also developed methods for isolating and analyzing fetal DNA that has escaped into the mother's blood (*cell-free fetal DNA*), using various techniques. Cell-free fetal DNA tests and other blood tests are increasingly being used as noninvasive prenatal screening tests for certain chromosomal defects and disorders; a positive result indicates to the parents that further diagnostic testing should be considered.

One diagnostic test that can determine whether the developing fetus has a serious recessive disease is **amniocentesis**, which can be performed starting at the 15th or 16th week of pregnancy **(Figure 14.19a)**. In this procedure, a physician inserts a needle into the uterus and extracts about 10 mL of amniotic fluid, the liquid that bathes the fetus. Some genetic disorders can be detected from the presence of certain proteins or hormones in the amniotic fluid itself. Tests

▼ Figure 14.19 Diagnostic fetal testing for genetic disorders. Most testing for genetic abnormalities is now DNA-based. In some cases biochemical tests are used to detect particular disorders, and karyotyping may be done to see whether the chromosomes of the fetus are normal in number and appearance.



for thousands of other disorders (such as cystic fibrosis) are performed on DNA from fetal cells sloughed off into the amniotic fluid using targeted DNA sequencing or a technique called a *DNA microarray* (see Figure 20.13) that tests for many alleles at the same time. A karyotype of these cultured cells can also confirm certain chromosomal defects that may not have been suggested by cell-free fetal DNA tests.

In an alternative diagnostic technique that is called **chorionic villus sampling (CVS)**, a physician inserts a narrow tube through the cervix into the uterus and suctions out a tiny sample of tissue from the placenta, the organ that transmits nutrients and fetal wastes between the fetus and the mother **(Figure 14.19b)**. The cells of the chorionic villi of the placenta—the portion sampled—are derived from the fetus and have the same genotype and DNA sequence as the new individual. DNA from these cells can be analyzed for the presence of many genetic disorders. The main advantage of CVS is that it can be performed as early as the 10th or 11th week of pregnancy.

Ultrasound and isolation of fetal cells or DNA from maternal blood pose no known risk to either mother or fetus, while amniocentesis and CVS can cause complications in a small percentage of cases (for both, between 1 in 1,000 and 1 in 500 cases). Amniocentesis or CVS for diagnostic testing is generally offered to women over age 35, due to their increased risk of bearing a child with Down syndrome or other chromosomal disorders, and may also be offered to younger women if there are risk factors or conditions suggested by family history. If the fetal tests reveal a serious disorder, the parents face the difficult choice of either terminating the pregnancy or preparing to care for a child with a genetic disorder, one that might even be fatal. Parental and fetal screening for Tay-Sachs alleles done since 1980 has reduced the number of children born with this incurable disease by 90%.

Newborn Screening

Some genetic diseases and disorders can be detected at birth by simple biochemical tests that are routinely performed on blood from the baby's heel in most hospitals in the United States. One common screening program is for phenylketonuria (PKU), a recessively inherited disorder that occurs in about one out of every 10,000–15,000 births in the United States. Children with this disease cannot properly metabolize the amino acid phenylalanine. This compound and its by-product, phenylpyruvate, can accumulate to toxic levels in the blood, causing severe intellectual disability. However, if PKU is detected in the newborn, a special diet low in phenylalanine will usually allow for typical development. (Among many other substances, this diet excludes the artificial sweetener aspartame, which contains phenylalanine.)

The number of newborn genetic tests, many for metabolic disorders or enzyme deficiencies, is over 100 and ever-increasing;

each state regulates which tests are used. If screening identifies a baby with a disease or disorder for which treatment can be started early on, that individual has improved chances for better life quality. Some diseases, like PKU, are treated by dietary changes and others by replacement of missing enzymes. Unfortunately, there are quite a few diseases for which no treatment is yet known.

Fetal and newborn screening for serious inherited diseases, tests for identifying carriers, and genetic counseling all rely on the Mendelian model of inheritance. We owe the "gene idea"—the concept of heritable factors transmitted according to simple rules of chance—to the elegant quantitative experiments of Gregor Mendel. The importance of his discoveries was overlooked by most biologists until the early 1900s, decades after he reported his findings. In the next chapter, you will learn how Mendel's laws have their physical basis in the behavior of chromosomes during sexual life cycles and how the synthesis of Mendelian genetics and a chromosome theory of inheritance catalyzed progress in genetics.

CONCEPT CHECK 14.4

1. Lucia and Jared each have a sibling with cystic fibrosis, but neither Lucia nor Jared nor any of their parents have the disease. Calculate the probability that if this couple has a child, the child will have cystic fibrosis. What would be the probability if a test revealed that Jared is a carrier but Lucia is not? Explain your answers.
2. **MAKE CONNECTIONS** Explain how the change of a single amino acid in hemoglobin leads to the aggregation of hemoglobin into long fibers. (Review Figures 5.14, 5.18, and 5.19.)
3. Juanita was born with six toes on each foot, a dominant trait called polydactyly. Two of her five siblings and her mother, but not her father, also have extra digits. What is Juanita's genotype for the number-of-digits character? Explain your answer. Use D and d to symbolize the alleles for this character.
4. **MAKE CONNECTIONS** In Table 14.1, note the phenotypic ratio of the dominant to recessive trait in the F_2 generation for the monohybrid cross involving flower color. Then determine the phenotypic ratio for the offspring of the second-generation couple in Figure 14.15b. What accounts for the difference in the two ratios?

For suggested answers, see Appendix A.

14 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 14.1

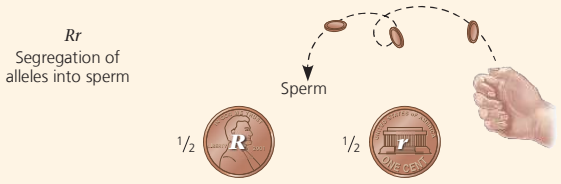
Mendel used the scientific approach to identify two laws of inheritance *(pp. 270–276)*

- Gregor Mendel formulated a theory of inheritance based on experiments with garden peas, proposing that parents pass on to their offspring discrete genes that retain their identity through generations. This theory includes two "laws."
- The **law of segregation** states that genes have alternative forms, or **alleles**. In a diploid organism, the two alleles of a gene segregate (separate) during meiosis and gamete formation; each sperm or egg carries only one allele of each pair. This law explains the 3:1 ratio of F_2 **phenotypes** observed when **monohybrids** cross-or self-pollinate. Each organism inherits one allele for each gene from each parent. In **heterozygotes**, the two alleles are different: Expression of the **dominant allele** masks the phenotypic effect of the **recessive allele. Homozygotes** have identical alleles of a given gene and are therefore **true-breeding**.
- The **law of independent assortment** states that the pair of alleles for a given gene segregates into gametes independently of the pair of alleles for any other gene. In a cross between **dihybrids** (individuals heterozygous for two genes), the offspring have four phenotypes in a 9:3:3:1 ratio.

? *When Mendel did crosses of true-breeding purple- and white-flowered pea plants, the white-flowered trait disappeared from the F_1 generation but reappeared in the F_2 generation. Use genetic terms to explain why that happened.*

CONCEPT 14.2

Probability laws govern Mendelian inheritance *(pp. 276–278)*



- The **multiplication rule** states that the probability of two or more events occurring together (a statement using *and*) is equal to the product of the individual probabilities of the independent single events. The **addition rule** states that the probability of an event that can occur in two or more independent, mutually exclusive ways (using *or*) is the sum of the individual probabilities.
- The rules of probability can be used to solve complex genetics problems. A dihybrid or other multicharacter cross is equivalent to two or more independent monohybrid crosses occurring simultaneously. In calculating the chances of the various offspring genotypes from such crosses, each character is first considered separately and then the individual probabilities are multiplied.

DRAW IT *Redraw the Punnett square on the right side of Figure 14.8 as two smaller monohybrid Punnett squares, one for each gene. Below each square, list the fractions of each phenotype produced. Use the rule of multiplication to compute the overall fraction of each of the possible dihybrid phenotypes. What is the phenotypic ratio?*

CONCEPT 14.3

Inheritance patterns are often more complex than predicted by simple Mendelian genetics *(pp. 278–283)*

- Extensions of Mendelian genetics for a single gene:

Relationship among alleles of a single gene	Description	Example
Complete dominance of one allele	Heterozygous phenotype same as that of homozygous dominant	*PP* *Pp*
Incomplete dominance of either allele	Heterozygous phenotype intermediate between the two homozygous phenotypes	C^RC^R C^RC^W C^WC^W
Codominance	Both phenotypes expressed in heterozygotes	I^AI^B
Multiple alleles	In the population, some genes have more than two alleles	ABO blood group alleles I^A, I^B, i
Pleiotropy	One gene affects multiple phenotypic characters	Sickle-cell disease

- Extensions of Mendelian genetics for two or more genes:

Relationship among two or more genes	Description	Example
Epistasis	The phenotypic expression of one gene affects the expression of another gene	*BbEe* × *BbEe*; *BE*, *bE*, *Be*, *be*; 9 : 3 : 4
Polygenic inheritance	A single phenotypic character is affected by two or more genes	*AaBbCc* × *AaBbCc*

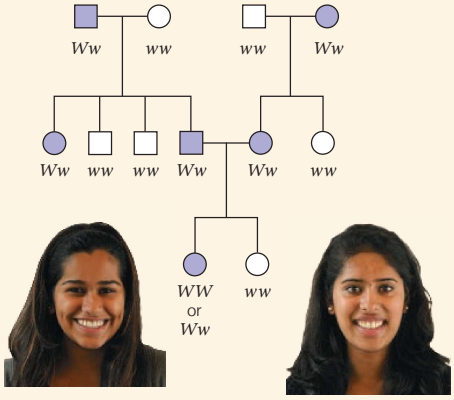
- The expression of a genotype can be affected by environmental influences, resulting in a range of phenotypes. Polygenic characters that are also influenced by the environment are called **multifactorial** characters.
- An organism's overall phenotype reflects its overall genotype and unique environmental history. Even in more complex inheritance patterns, Mendel's fundamental laws still apply.

? *Which genetic relationships listed in the first column of the two tables above are demonstrated by the inheritance pattern of the ABO blood group alleles? For each genetic relationship, explain why this inheritance pattern is or is not an example.*

CONCEPT 14.4

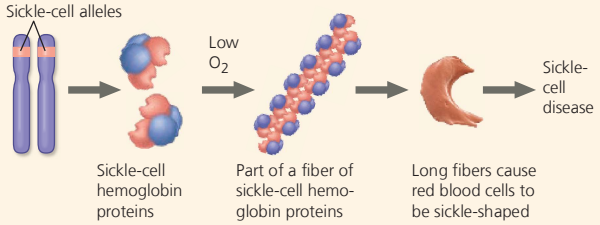
Many human traits follow Mendelian patterns of inheritance *(pp. 284–290)*

- Analysis of family **pedigrees** can be used to deduce the possible genotypes of individuals and make predictions about future offspring. Such predictions are statistical probabilities rather than certainties.



Widow's peak — No widow's peak

- Many genetic disorders are inherited as simple recessive traits. Most affected (homozygous recessive) individuals are children of phenotypically normal, heterozygous **carriers**.
- The sickle-cell allele has probably persisted for evolutionary reasons: Homozygotes have **sickle-cell disease**, but heterozygotes have an advantage because one copy of the sickle-cell allele reduces both the frequency and severity of malaria attacks.



- Lethal dominant alleles are eliminated from the population if affected people die before reproducing. Nonlethal dominant alleles and lethal ones that strike relatively late in life can be inherited in a Mendelian way.
- Many human diseases are multifactorial—that is, they have both genetic and environmental components and do not follow simple Mendelian inheritance patterns.
- Using family histories, genetic counselors help couples determine the probability that their children will have genetic disorders. Genetic testing of prospective parents can reveal whether they are carriers of recessive alleles associated with specific disorders. Ultrasounds and blood tests can screen for certain disorders in a fetus. **Amniocentesis** and **chorionic villus sampling** can diagnose whether a suspected genetic disorder is present in a fetus. Other genetic tests can be performed after birth.

? *Both members of a couple know that they are carriers of the cystic fibrosis allele. None of their three children have cystic fibrosis, but any one of them might be a carrier. The couple would like to have a fourth child but are worried that he or she would very likely have the disease, since the first three do not. What would you tell the couple? Would it remove some uncertainty from their prediction if they could find out whether the three children are carriers?*

TIPS FOR GENETICS PROBLEMS

1. Write down symbols for the alleles. (These may be given in the problem.) When represented by single letters, the dominant allele is uppercase and the recessive is lowercase.
2. Write down the possible genotypes, as determined by the phenotype.
   a. If the phenotype is that of the dominant trait (for example, purple flowers), then the genotype is either homozygous dominant or heterozygous (*PP* or *Pp* in this example).
   b. If the phenotype is that of the recessive trait, the genotype must be homozygous recessive (for example, *pp*).
   c. If the problem says "true-breeding," the genotype is homozygous.
3. Determine what the problem is asking for. If asked to do a cross, write it out in the form [genotype] × [genotype], using the alleles you've decided on.
4. To figure out the outcome of a cross, set up a Punnett square.
   a. Put the gametes of one parent at the top and those of the other on the left. To determine the allele(s) in each gamete for a given genotype, set up a systematic way to list all the possibilities. (Remember, each gamete has one allele of each gene.) Note that there are 2^n possible types of gametes, where n is the number of gene loci that are heterozygous. For example, an individual with genotype *AaBbCc* would produce $2^3 = 8$ types of gametes. Write the genotypes of the gametes in circles above the columns and to the left of the rows.
   b. Fill in the Punnett square as if each possible sperm were fertilizing each possible egg, making all of the possible offspring. In a cross of *AaBbCc* × *AaBbCc*, for example, the Punnett square would have 8 columns and 8 rows, so there are 64 different offspring; you would know the genotype of each and thus the phenotype. Count genotypes and phenotypes to obtain the genotypic and phenotypic ratios. Because the Punnett square is so large, this method is not the most efficient. See tip 5.
5. You can use the rules of probability if a Punnett square would be too big. (For example, see the question at the end of the summary for Concept 14.2 and question 7 below.) You can consider each gene separately (see the section Solving Complex Genetics Problems with the Rules of Probability in Concept 14.2).
6. If the problem gives you the phenotypic ratios of offspring but not the genotypes of the parents in a given cross, the phenotypes can help you deduce the parents' unknown genotypes.
   a. For example, if ½ of the offspring have the recessive phenotype and ½ the dominant, you know that the cross was between a heterozygote and a homozygous recessive.
   b. If the ratio is 3:1, the cross was between two heterozygotes.
   c. If two genes are involved and you see a 9:3:3:1 ratio in the offspring, you know that each parent is heterozygous for both genes. Caution: Don't assume that the reported numbers will exactly equal the predicted ratios. For example, if there are 13 offspring with the dominant trait and 11 with the recessive, assume that the ratio is one dominant to one recessive.
7. For pedigree problems, use the tips in Figure 14.15 and below to determine what kind of trait is involved.
   a. If parents without the trait have offspring with the trait, the trait must be recessive and the parents both carriers.
   b. If the trait is seen in every generation, it is most likely dominant (see the next possibility, though).
   c. If both parents have the trait, then in order for it to be recessive, all offspring must show the trait.
   d. To determine the likely genotype of a certain individual in a pedigree, first label the genotypes of all the family members you can. Even if some of the genotypes are incomplete, label what you do know. For example, if an individual has the dominant phenotype, the genotype must be *AA* or *Aa*; you can write this as *A*–. Try different possibilities to see which fits the results. Use the rules of probability to calculate the probability of each possible genotype being the correct one.

TEST YOUR UNDERSTANDING

For multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. **DRAW IT** A pea plant heterozygous for inflated pods (*Ii*) is crossed with a plant homozygous for constricted pods (*ii*). Draw a Punnett square for this cross to predict genotypic and phenotypic ratios. Assume that pollen comes from the *ii* plant.
2. A man with type A blood marries a woman with type B blood. Their child has type O blood. What are the genotypes of these three individuals? What genotypes, and in what frequencies, would you expect in future offspring from this marriage?
3. A man has six fingers on each hand and six toes on each foot. His wife and their daughter have the typical number of digits. Remember that extra digits is a dominant trait. What fraction of this couple's children would be expected to have extra digits?
4. **DRAW IT** Two pea plants heterozygous for the characters of pod color and pod shape are crossed. (See Table 14.1.) Draw a Punnett square to determine the phenotypic ratios of the offspring.

Levels 3-4: Applying/Analyzing

5. Flower position, stem length, and seed shape are three characters that Mendel studied. Each is controlled by an independently assorting gene and has dominant and recessive expression as indicated in Table 14.1. If a plant that is heterozygous for all three characters is allowed to self-fertilize, what proportion of the offspring would you expect to be each of the following? (Note: Use the rules of probability instead of a huge Punnett square.)
   (a) homozygous for the three dominant traits
   (b) homozygous for the three recessive traits
   (c) heterozygous for all three characters
   (d) homozygous for axial and tall, heterozygous for seed shape
6. Hemochromatosis is an inherited disease caused by a recessive allele. If a woman and her husband, who are both carriers, have three children, what is the probability of each of the following?
   (a) All three children are of normal phenotype.
   (b) One or more of the three children have the disease.
   (c) All three children have the disease.
   (d) At least one child is phenotypically normal.

   (*Note*: It will help to remember that the probabilities of all possible outcomes always add up to 1.)
7. The genotype of F_1 individuals in a tetrahybrid cross is *AaBbCcDd*. Assuming independent assortment of these four genes, what are the probabilities that F_2 offspring will have the following genotypes?
   (a) *aabbccdd*
   (b) *AaBbCcDd*
   (c) *AABBCCDD*
   (d) *AaBBccDd*
   (e) *AaBBCCdd*

8. What is the probability that each of the following pairs of parents will produce the indicated offspring? (Assume independent assortment of all gene pairs.)
   (a) *AABBCC* × *aabbcc* → *AaBbCc*
   (b) *AABbCc* × *AaBbCc* → *AAbbCC*
   (c) *AaBbCc* × *AaBbCc* → *AaBbCc*
   (d) *aaBbCC* × *AABbcc* → *AaBbCc*

9. Keisha and Jerome each have a sibling with sickle-cell disease. Neither Keisha nor Jerome nor any of their parents have the disease, and none of them have ever been tested to see if they carry the sickle-cell allele. Based on this incomplete information, calculate the probability that if this couple has a child, the child will have sickle-cell disease.

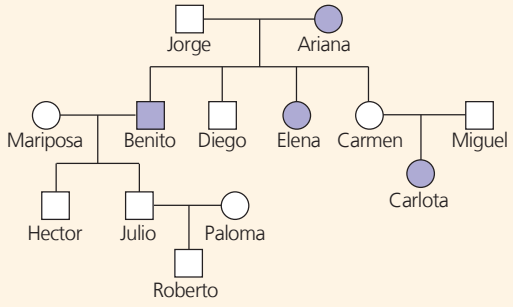
10. In 1981, a stray black cat with unusual rounded, curled-back ears was adopted by a family in California. Thousands of descendants of the cat have since been born, and cat fanciers have developed the curl cat into a show breed. Suppose you owned the first curl cat and wanted to develop a true-breeding variety. How would you determine whether the curl allele is dominant or recessive? How would you obtain true-breeding curl cats? How could you be sure they are true-breeding?



11. In tigers, a recessive allele of a particular gene causes both an absence of fur pigmentation (a white tiger) and a cross-eyed condition. If two phenotypically normal tigers that are heterozygous at this locus are mated, what percentage of their offspring will be cross-eyed? What percentage of tigers with crossed eyes will also be white?

12. In maize (corn) plants, a dominant allele *I* inhibits kernel color, while the recessive allele *i* permits color when homozygous. At a different locus, the dominant allele *P* causes purple kernel color, while the homozygous recessive genotype *pp* causes red kernels. If plants heterozygous at both loci are crossed, what will be the phenotypic ratio of the offspring?

13. The pedigree below traces the inheritance of alkaptonuria, a biochemical disorder. Affected individuals, indicated here by the colored circles and squares, are unable to metabolize a substance called alkapton, which colors the urine and stains body tissues. Does alkaptonuria appear to be caused by a dominant allele or by a recessive allele? Fill in the genotypes of the individuals whose genotypes can be deduced. What genotypes are possible for each of the other individuals?



14. Imagine that you are a genetic counselor, and a couple planning to start a family comes to you for information. Charles was married once before, and he and his first wife had a child with cystic fibrosis. The brother of his current wife, Elaine, died of cystic fibrosis. What is the probability that Charles and Elaine will have a baby with cystic fibrosis? (Neither Charles, Elaine, nor their parents have cystic fibrosis.)

Levels 5-6: Evaluating/Creating

15. **EVOLUTION CONNECTION** Over the past half century, there has been a trend in the United States and other developed countries for people to marry and start families later in life than did their parents and grandparents. What effects might this trend have on the incidence (frequency) of late-acting dominant lethal alleles in the population?

16. **SCIENTIFIC INQUIRY** You are handed a mystery pea plant with tall stems and axial flowers and asked to determine its genotype as quickly as possible. You know that the allele for tall stems (*T*) is dominant to that for dwarf stems (*t*) and that the allele for axial flowers (*A*) is dominant to that for terminal flowers (*a*).
   (a) Identify all the possible genotypes for your mystery plant.
   (b) Describe the one cross you would do in your garden to determine the exact genotype of your mystery plant.
   (c) While waiting for the results of your cross, you predict the results for each possible genotype listed in part a. Explain how you do this and why this is not called "performing a cross."
   (d) Explain how the results of your cross and your predictions will help you learn the genotype of your mystery plant.

17. **WRITE ABOUT A THEME: INFORMATION** The continuity of life is based on heritable information in the form of DNA. In a short essay (100–150 words), explain how the passage of genes from parents to offspring, in the form of particular alleles, ensures perpetuation of parental traits in offspring and, at the same time, genetic variation among offspring. Use genetic terms in your explanation.

18. **SYNTHESIZE YOUR KNOWLEDGE**



Just for fun, imagine that "shirt-striping" is a phenotypic character caused by a single gene. Construct a genetic explanation for the appearance of the family in the above photograph, consistent with their "shirt phenotypes." Include in your answer the presumed allele combinations for "shirt-striping" in each family member. Identify the inheritance pattern shown by the child.

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
What genes affect skin color? Go to "A Golden Fish Reveals Pigmentation Loss in Europeans" at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.

15 The Chromosomal Basis of Inheritance

KEY CONCEPTS

15.1 Mendelian inheritance has its physical basis in the behavior of chromosomes *p. 295*

15.2 Sex-linked genes exhibit unique patterns of inheritance *p. 298*

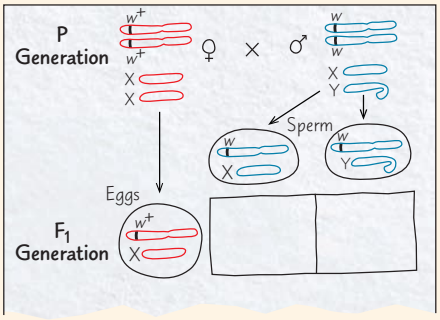
15.3 Linked genes tend to be inherited together because they are located near each other on the same chromosome *p. 301*

15.4 Alterations of chromosome number or structure cause some genetic disorders *p. 306*

15.5 Some inheritance patterns are exceptions to standard Mendelian inheritance *p. 310*

Study Tip

Draw chromosomes: As you work on the figure legend questions in this chapter, draw the chromosomes for the crosses described in the questions. Below is a starting sketch for the What If? question that follows Figure 15.3. (Here you are including sex chromosomes to keep track of the offspring.)



Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 15
- Animation: The Chromosomal Basis of Independent Assortment
- Animation: Linked Genes and Crossing Over

For Instructors to Assign (in Item Library)
- Tutorial: Pedigrees and Sex-Linkage
- Tutorial: Recombination and Linkage Mapping



Figure 15.1 The four yellow dots mark the locations of a specific gene, tagged with a fluorescent yellow dye, on a pair of homologous chromosomes. The chromosomes have duplicated, so each chromosome has two sister chromatids, each with a copy of the gene. This provides a visual demonstration that genes—Mendel's "factors"—are segments of DNA located along chromosomes.

What is the relationship between genes and chromosomes?

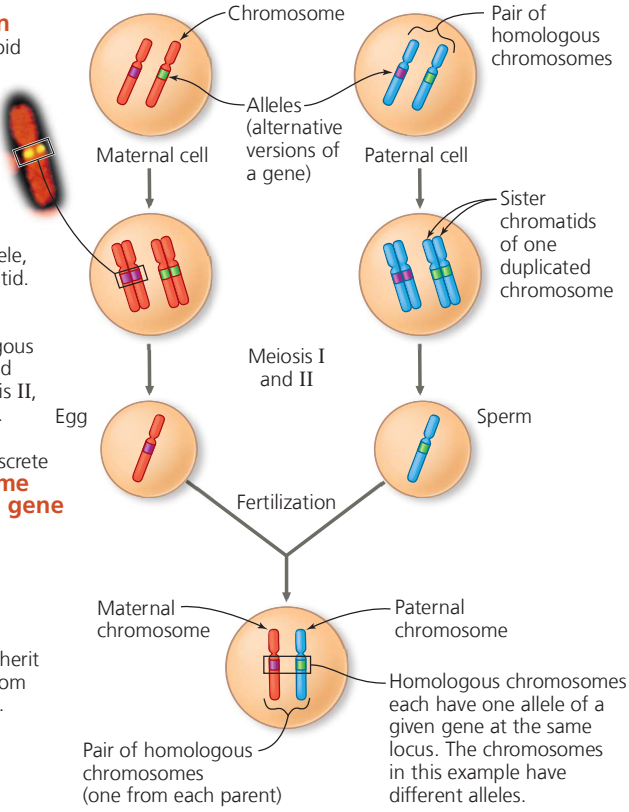
Genes are located on chromosomes. In diploid cells, chromosomes and genes are present in pairs.

Chromosomes duplicate before cell division. Each duplicated chromosome has two copies of each allele, one on each sister chromatid.

During meiosis I, homologous chromosomes separate and alleles segregate. In meiosis II, sister chromatids separate.

Genes are passed on as discrete units. **Each chromosome has one version of a gene (one allele).**

Offspring inherit one allele from each parent.



CONCEPT 15.1

Mendelian inheritance has its physical basis in the behavior of chromosomes

When Gregor Mendel proposed the existence of "hereditary factors" in 1860, no cellular structures had been identified that could house these imaginary units, and most biologists were skeptical. When the processes of mitosis and meiosis were worked out later that century, biologists saw parallels between the behavior of Mendel's proposed factors (genes) during sexual life cycles and the behavior of chromosomes: As shown in **Figure 15.1**, chromosomes and genes are both present in pairs in diploid cells, and homologous chromosomes separate and alleles segregate during the process of meiosis. After meiosis, fertilization restores the paired condition for both chromosomes and genes. Around 1902, Walter S. Sutton, Theodor Boveri, and others independently noted these parallels and began to develop the **chromosome theory of inheritance**. According to this theory, Mendelian genes have specific loci (sites) along chromosomes, and it is the chromosomes that undergo segregation and independent assortment.

The first solid evidence associating a specific gene with a specific chromosome came early in the 1900s from the work of Thomas Hunt Morgan, an experimental embryologist at Columbia University. Although Morgan was initially doubtful about both Mendelian genetics and the chromosome theory, his early experiments provided convincing evidence that chromosomes are indeed the location of Mendel's heritable factors.

Morgan's Choice of Experimental Organism

Many times in the history of biology, important discoveries have come to those insightful or lucky enough to choose an experimental organism suitable for the research problem being tackled. Mendel chose the garden pea because a number of distinct varieties were available. For his work, Morgan selected a species of fruit fly, *Drosophila melanogaster*, a common insect that feeds on the fungi growing on fruit. Fruit flies are prolific breeders; a single mating will produce hundreds of offspring, and a new generation can be bred every two weeks. Morgan's laboratory began using this convenient organism for genetic studies in 1907 and soon became known as "the fly room."

Another advantage of the fruit fly is that it has only four pairs of chromosomes, which are easily distinguishable with a light microscope. There are three pairs of autosomes and one pair of sex chromosomes. Female fruit flies have a pair of homologous X chromosomes, and males have one X chromosome and one Y chromosome.

While Mendel could readily obtain different pea varieties from seed suppliers, Morgan was probably the first person to want different varieties of the fruit fly. He faced the tedious task of carrying out many matings and then microscopically inspecting large numbers of offspring in search of naturally occurring variant individuals. After many months of this, he complained, "Two years' work wasted. I have been breeding those flies for all that time and I've got nothing out of it." Morgan persisted, however, and was finally rewarded with the discovery of a single male fly with white eyes instead of the usual red. The phenotype for a character most commonly observed in natural populations, such as red eyes in *Drosophila*, is called the **wild type** **(Figure 15.2)**. Traits that are alternatives to the wild type, such as white eyes in *Drosophila*, are called *mutant phenotypes* because they are due to alleles assumed to have originated as changes, or mutations, in the wild-type allele.

▼ **Figure 15.2 Morgan's first mutant.** Wild-type *Drosophila* flies have red eyes (left). Among his flies, Morgan discovered a mutant male with white eyes (right). This variation made it possible for Morgan to trace a gene for eye color to a specific chromosome.



Wild type (red eyes) **Mutant** (white eyes)

Morgan and his students invented a notation for symbolizing alleles in *Drosophila* that is still widely used for fruit flies. For a given character in flies, the gene takes its symbol from the first mutant (non–wild type) discovered. Thus, the allele for white eyes in *Drosophila* is symbolized by *w*. A superscript + identifies the allele for the wild-type trait: w^+ for the allele for red eyes, for example. Over the years, a variety of gene notation systems have been developed for different organisms. For example, human genes are usually written in all capitals, such as *HTT* for the gene involved in Huntington's disease. (More than one letter may be used for an allele, as in *HTT*.)

Correlating Behavior of a Gene's Alleles with Behavior of a Chromosome Pair: *Scientific Inquiry*

Morgan mated his white-eyed male fly with a red-eyed female. All the F_1 offspring had red eyes, suggesting that the wild-type allele is dominant. When Morgan bred the F_1 flies to each other, he observed the classical 3:1 phenotypic ratio among the F_2 offspring. However, there was a surprising additional result: The white-eye trait showed up only in males. All the F_2 females had red eyes, while half the males had red eyes and half had white eyes. Therefore, Morgan concluded

that somehow a fly's eye color was linked to its sex. (If the eye color gene were unrelated to sex, half of the white-eyed flies would have been female.)

Recall that a female fly has two X chromosomes (XX), while a male fly has an X and a Y (XY). The correlation between the trait of white eye color and the male sex of the affected F_2 flies suggested to Morgan that the gene involved in his white-eyed mutant was located exclusively on the X chromosome, with no corresponding allele present on the Y chromosome. His reasoning can be followed in **Figure 15.3**. For a male, a single copy of the mutant allele would confer white eyes; since a male has only one X chromosome, there can be no wild-type allele (w^+) present to mask the recessive allele. However, a female could have white eyes only if both her X chromosomes carried the recessive mutant allele (w). This was impossible for the F_2 females in Morgan's experiment because all the F_1 fathers had red eyes, so each F_2 female received a w^+ allele on the X chromosome inherited from her father.

Morgan's finding of the correlation between a particular trait and an individual's sex provided support for the chromosome theory of inheritance: namely, that a specific gene is carried on a specific chromosome. In this case, an eye color gene is carried on the X chromosome.

Figure 15.4 illustrates the relationship between the chromosome theory of inheritance and Mendel's laws. The separation of homologs during anaphase I accounts for the segregation of the two alleles of a gene into separate gametes, and the random arrangement of chromosome pairs at metaphase I accounts for independent assortment of the alleles for two or more genes located on different homolog pairs. This figure traces the same dihybrid pea cross you learned about in Figure 14.8. By carefully studying Figure 15.4, you can see how the behavior of chromosomes during meiosis in the F_1 generation and subsequent random fertilization give rise to the F_2 phenotypic ratio observed by Mendel.

In addition to showing that a specific gene is carried on a specific chromosome, Morgan's work indicated that genes located on a sex chromosome exhibit unique inheritance patterns, which we will discuss in the next section. Recognizing the importance of Morgan's early work, many bright students were attracted to his fly room.

CONCEPT CHECK 15.1

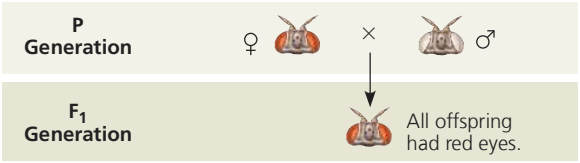
1. **WHAT IF?** Propose a possible reason that the first naturally occurring mutant fruit fly Morgan saw involved a gene on a sex chromosome and was found in a male.
2. Which one of Mendel's laws describes the inheritance of alleles for a single character? Which law relates to the inheritance of alleles for two characters in a dihybrid cross?
3. **MAKE CONNECTIONS** Review the description of meiosis (see Figure 13.8) and Mendel's laws of segregation and independent assortment (see Concept 14.1). What is the physical basis for each of Mendel's laws?

For suggested answers, see Appendix A.

▼ Figure 15.3 Inquiry

In a cross between a wild-type female fruit fly and a mutant white-eyed male, what color eyes will the F_1 and F_2 offspring have?

Experiment Thomas Hunt Morgan wanted to analyze the behavior of two alleles of a fruit fly eye color gene. In crosses similar to those done by Mendel with pea plants, Morgan and his colleagues mated a wild-type (red-eyed) female with a mutant white-eyed male.

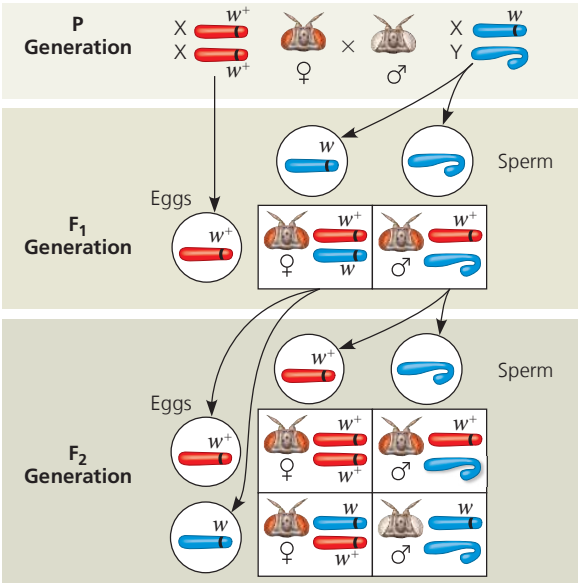


Morgan then bred an F_1 red-eyed female to an F_1 red-eyed male to produce the F_2 generation.

Results The F_2 generation showed a typical Mendelian ratio of 3 red-eyed flies: 1 white-eyed fly. However, all white-eyed flies were males; no females displayed the white-eye trait.



Conclusion All F_1 offspring had red eyes, so the mutant white-eye trait (w) must be recessive to the wild-type red-eye trait (w^+). Since the recessive trait—white eyes—was expressed only in males in the F_2 generation, Morgan deduced that this eye color gene is located on the X chromosome and that there is no corresponding locus on the Y chromosome.



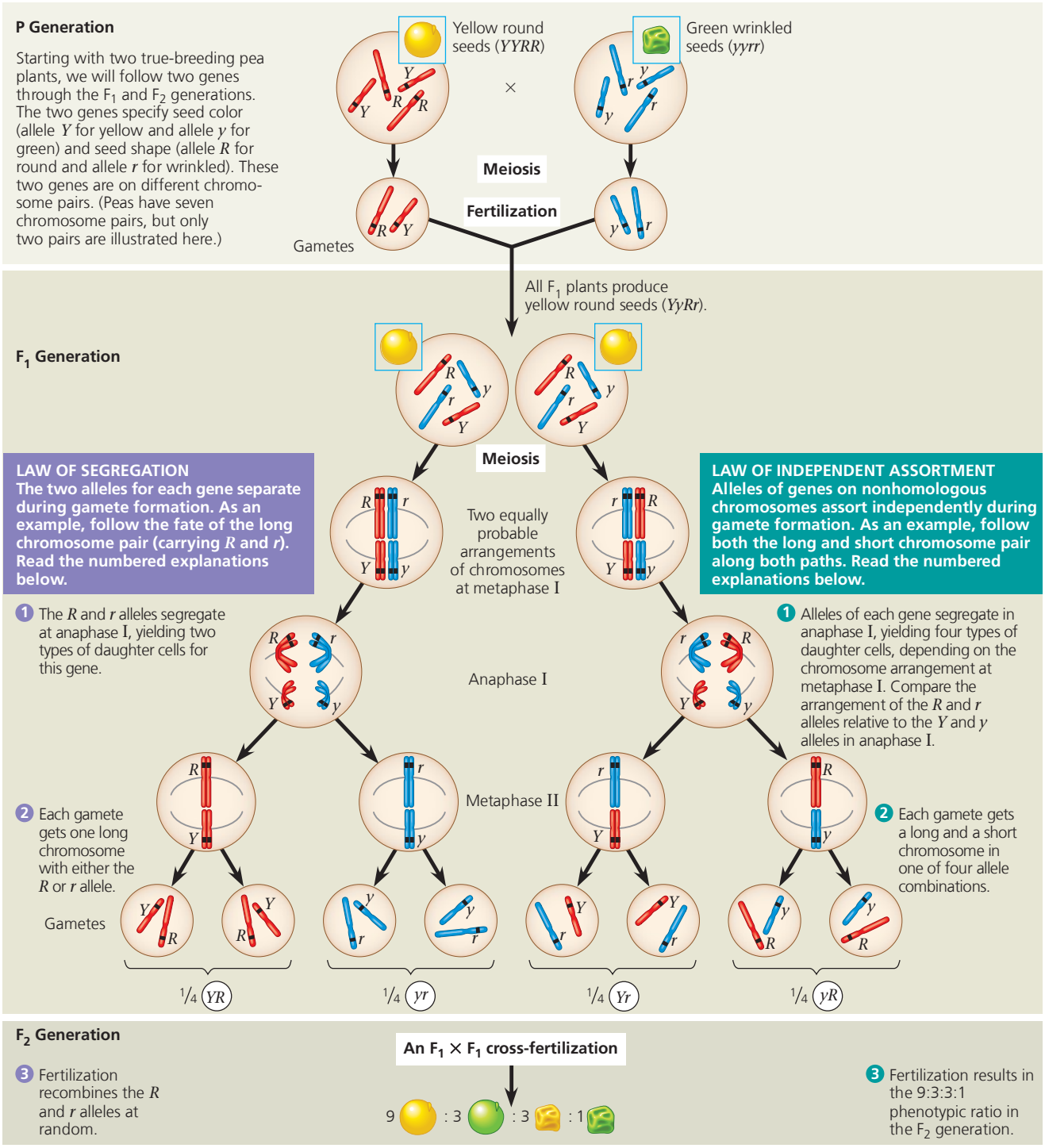
Data from T. H. Morgan, Sex-limited inheritance in *Drosophila*, *Science* 32: 120–122 (1910).

Instructors: A related Experimental Inquiry Tutorial can be assigned in **Mastering Biology**.

WHAT IF? *Suppose this eye color gene were located on an autosome rather than a sex chromosome. Predict the phenotypes (including sex) of the F_2 flies in this hypothetical cross. (Hint: Draw Punnett squares, including chromosomes, for the F_1 and F_2 generations; see the Study Tip.)*

▼ Figure 15.4 The chromosomal basis of Mendel's laws. Here we correlate the results of one of Mendel's dihybrid crosses (see Figure 14.8) with the behavior of chromosomes during meiosis (see Figure 13.8). The arrangement of chromosomes at metaphase I of meiosis and their movement during anaphase I account for, respectively, the independent assortment and segregation of the alleles for seed color and shape. Each cell that undergoes meiosis in an F_1 plant produces two kinds of gametes. If we count the results for all cells, however, each F_1 plant produces equal numbers of all four kinds of gametes because the alternative chromosome arrangements at metaphase I are equally likely.

Mastering Biology
Animation: The Chromosomal Basis of Independent Assortment
MP3 Tutor: The Chromosomal Basis of Inheritance



? *If you crossed an F_1 plant above with a plant that was homozygous recessive for both genes (*yyrr*), how would the phenotypic ratio of the offspring compare with the 9:3:3:1 ratio seen here?*

CONCEPT 15.2

Sex-linked genes exhibit unique patterns of inheritance

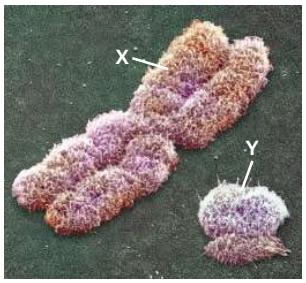
Morgan's discovery of a trait (white eyes) that correlated with the sex of flies provided important support for the chromosome theory of inheritance. Because the identity of the sex chromosomes in an individual could be inferred by observing the sex of the fly, the behavior of the two members of the pair of sex chromosomes could be correlated with the behavior of the two alleles of the eye color gene. In this section, we'll take a closer look at the role of sex chromosomes in inheritance.

The Chromosomal Basis of Sex

Although sex has traditionally been described as binary—male or female—we are coming to understand that this classification may be too simplistic. Here, we use the term *sex* to refer to classification into a group with a shared set of anatomical and physiological traits. In this sense, sex in many species is determined largely by inheritance of sex chromosomes. (The term *gender*, previously used as a synonym of sex, is now more often used to refer to an individual's own experience of identifying as male, female, or otherwise.)

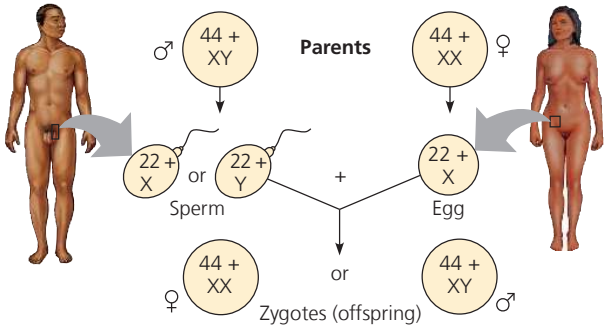
Humans and other mammals have two types of sex chromosomes, designated X and Y. The Y chromosome is much smaller than the X chromosome **(Figure 15.5)**. A person who inherits two X chromosomes, one from each parent, usually develops anatomy we associate with the "female" sex, while "male" properties are associated with the inheritance of one X chromosome and one Y chromosome **(Figure 15.6a)**. Short segments at either end of the Y chromosome are the only regions that are homologous with regions on the X. These homologous regions allow the X and Y chromosomes in males to pair and behave like homologs during meiosis in the testes.

▼ Figure 15.5 Human sex chromosomes (duplicated).

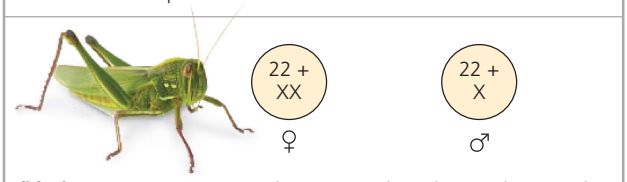


In mammalian testes and ovaries, the two sex chromosomes segregate during meiosis. Each egg receives one X chromosome. In contrast, sperm fall into two categories: Half the sperm cells a male produces receive an X chromosome, and half receive a Y chromosome. We can trace the sex of each offspring to the events of conception: If a sperm cell bearing an X chromosome fertilizes an egg, the zygote is XX, a female; if a sperm cell containing a Y chromosome fertilizes an egg, the zygote is XY, a male (see Figure 15.6a). Thus, in general, sex determination is a matter of chance—a fifty-fifty chance. Note that the mammalian X-Y system isn't the only chromosomal system for determining sex. **Figure 15.6b–d** illustrates three other systems.

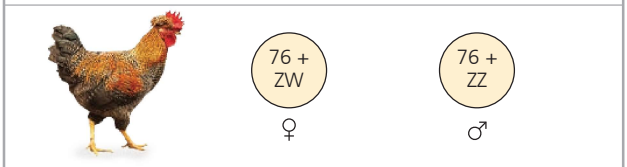
▼ Figure 15.6 Some chromosomal systems of sex determination. Numerals indicate the number of autosomes in the species pictured. In *Drosophila*, males are XY, but sex depends on the ratio of the number of X chromosomes to the number of autosome sets, not simply on the presence of a Y chromosome. In some species (not shown here), sex is determined not by chromosomes but by environmental factors such as temperature.



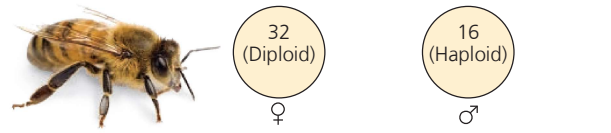
(a) The X-Y system. In mammals, the sex of an offspring depends on whether the sperm cell contains an X chromosome or a Y.



(b) The X-0 system. In grasshoppers, cockroaches, and some other insects, there is only one type of sex chromosome, the X. Females are XX; males have only one sex chromosome (X0). Sex of the offspring is determined by whether the sperm cell contains an X chromosome or no sex chromosome.



(c) The Z-W system. In birds, some fishes, and some insects, the sex chromosomes present in the egg (not the sperm) differ, and thus determine the sex of offspring. The sex chromosomes are designated Z and W. Females are ZW and males are ZZ.



(d) The haplo-diploid system. There are no sex chromosomes in most species of bees and ants. Females develop from fertilized eggs and are thus diploid. Males develop from unfertilized eggs and are haploid; they have no fathers.

In humans, the anatomical signs of sex begin to emerge when the embryo is about 2 months old. Before then, the rudiments of the gonads are generic—they can develop into either testes or ovaries, depending on whether or not a Y chromosome is present, and depending on what genes are active. A gene on the Y chromosome—called *SRY*, for sex-determining region of Y—is required for the development of testes. In the absence of *SRY*, the gonads develop into ovaries, even in an XY embryo.

A gene located on either sex chromosome is called a **sex-linked gene**. The human X chromosome contains approximately 1,100 genes, which are called **X-linked genes**, while genes located on the Y chromosome are called *Y-linked genes.* On the human Y chromosome, researchers have identified 78 genes that code for about 25 proteins (some genes are duplicates). About half of these genes are expressed only in the testis, and some are required for normal testicular functioning and the production of normal sperm. The Y chromosome is passed along virtually intact from a father to all his sons. Because there are so few Y-linked genes, very few disorders are transferred from father to son on the Y chromosome.

The development of female gonads in humans requires a gene called *WNT4* (on chromosome 1, an autosome), which encodes a protein that promotes ovary development. An embryo that is XY but has extra copies of the *WNT4* gene can develop rudimentary female gonads. Overall, sex is determined by the interactions of a network of gene products like these.

The biochemical, physiological, and anatomical features associated with "males" and "females" are turning out to be more complicated than previously thought, with many genes involved in their development. Because of the complexity of this process, many variations exist: Some individuals vary in the number of sex chromosomes in their cells (see Concept 15.4), and others are born with intermediate sexual (*intersex*) characteristics, or with anatomical features that do not match an individual's sense of their own gender (*transgender* individuals). Sex determination is an active area of research that will likely yield a more sophisticated understanding in years to come.

Inheritance of X-Linked Genes

The fact that males and females inherit a different number of X chromosomes leads to a pattern of inheritance different from that produced by genes located on autosomes. While there are very few Y-linked genes, many of which help determine sex, the X chromosomes have numerous genes for characters unrelated to sex. X-linked genes in humans follow the same pattern of inheritance that Morgan observed for the eye color locus he studied in *Drosophila* (see Figure 15.3). Fathers pass X-linked alleles to all of their daughters but to none of their sons. In contrast, mothers can pass X-linked alleles to both sons and daughters, as shown in **Figure 15.7** for the inheritance of a mild X-linked disorder, red-green color blindness.

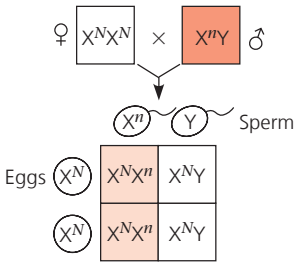
If an X-linked trait is due to a recessive allele, a female will express the phenotype only if she is homozygous for that allele. Because males have only one locus, the terms *homozygous* and *heterozygous* lack meaning for describing their X-linked genes; the term *hemizygous* is used in such cases. Any male receiving the recessive allele from his mother will express the trait. For this reason, far more males than females have X-linked recessive disorders. However, even though the chance of a female inheriting a double dose of the mutant allele is much less than the probability of a male inheriting a single dose, there *are* females with X-linked disorders. For instance, color blindness is almost always inherited as an X-linked trait. A color-blind daughter may be born to a color-blind father whose mate is a carrier (see Figure 15.7c). Because the X-linked allele for color blindness is relatively rare, though, the probability that such a man and woman will mate is low.

A number of human X-linked disorders are much more serious than color blindness, such as **Duchenne muscular dystrophy**, which affects about one out of 5,000 males born in the United States. The disease is characterized by a progressive weakening of the muscles and loss of coordination. Affected individuals' life expectancy is into the mid-20s. Researchers have traced the disorder to the absence of a key muscle protein called dystrophin and have mapped the gene for this protein to a specific locus on the X chromosome. Since the gene is known, gene therapy is being explored.

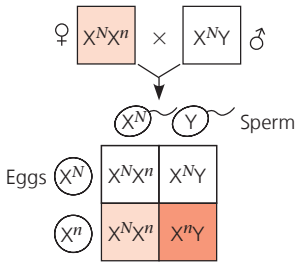
▼ Figure 15.7 The transmission of X-linked recessive traits. In this diagram, red-green color blindness is used as an example. The superscript *N* represents the dominant allele for normal color vision carried on the X chromosome, while *n* represents the recessive allele, which has a mutation causing color blindness. White boxes indicate unaffected individuals, light orange boxes indicate carriers, and dark orange boxes indicate color-blind individuals.

? *If a color-blind woman married a man who had normal color vision, what would be the probable phenotypes of their children?*

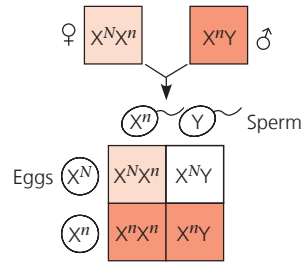
Mastering Biology Animation: X-Linked Genes in "MendAliens"



(a) A father with color-blindness will transmit the mutant allele to all daughters but to no sons. When the mother is a dominant homozygote, the daughters will have the normal phenotype but will be carriers of the mutation.



(b) If a carrier mates with a male who has normal color vision, there is a 50% chance that each daughter will be a carrier like her mother and a 50% chance that each son will have the disorder.



(c) If a carrier mates with a color-blind male, there is a 50% chance that each child born to them will have the disorder, regardless of sex. Daughters who have normal color vision will be carriers, whereas males who have normal color vision will be free of the recessive allele.

Hemophilia is an X-linked recessive disorder defined by the absence of one or more of the proteins required for blood clotting. When a person with hemophilia is injured, bleeding is prolonged because a firm clot is slow to form. Small cuts in the skin are usually not a problem, but bleeding in the muscles or joints can be painful and can lead to serious damage. In the 1800s, hemophilia was widespread among the royal families of Europe. Queen Victoria of England is known to have passed the allele to several of her descendants. Subsequent intermarriage with royal family members of other nations, such as Spain and Russia, further spread this X-linked trait, and its incidence is well documented in royal pedigrees. A few years ago, new genomic techniques allowed sequencing of DNA from tiny amounts isolated from the buried remains of royal family members. The genetic basis of the mutation, and how it resulted in a nonfunctional blood-clotting factor, is now understood. Today, people with hemophilia are treated as needed with intravenous injections of the protein that is missing.

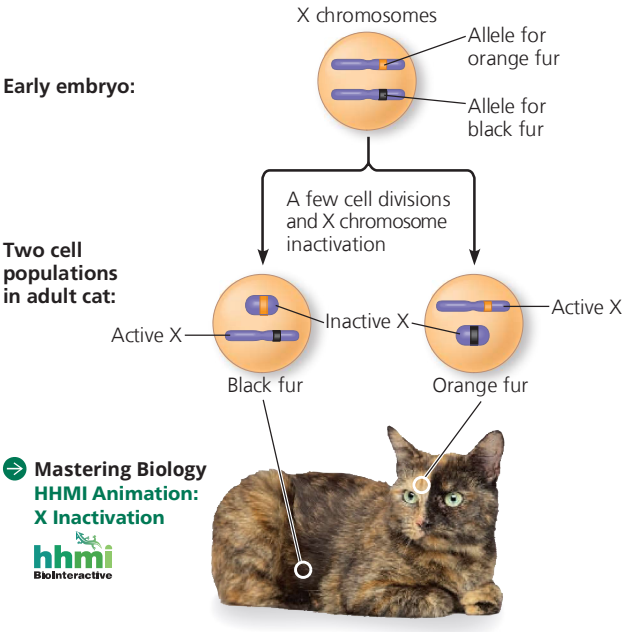
X Inactivation in Female Mammals

Female mammals, including human females, inherit two X chromosomes—twice the number inherited by males—so you may wonder whether females make twice as many of the proteins encoded by X-linked genes as males. In fact, almost all of one X chromosome in each cell in female mammals becomes inactivated during early embryonic development. As a result, the cells of females and males have the same effective dose (one active copy) of most X-linked genes. The inactive X in each cell of a female condenses into a compact object called a **Barr body** (discovered by Canadian anatomist Murray Barr), which lies along the inside of the nuclear envelope. Most of the genes of the X chromosome that forms the Barr body are not expressed. In the ovaries, however, Barr body chromosomes are reactivated in the cells that give rise to eggs, resulting in every female gamete (egg) having an active X after meiosis.

British geneticist Mary Lyon demonstrated that the selection of which X chromosome will form the Barr body occurs randomly and independently in each embryonic cell present at the time of X inactivation. As a consequence, females consist of a *mosaic* of two types of cells: those with the active X derived from the father and those with the active X derived from the mother. After an X chromosome is inactivated in a particular cell, all mitotic descendants of that cell have the same inactive X. Thus, if a female is heterozygous for a sex-linked trait, about half of her cells will express one allele, while the others will express the alternate allele. **Figure 15.8** shows how this mosaicism results in the patchy coloration of a tortoiseshell cat. In humans, mosaicism can be observed in a recessive X-linked mutation that prevents the development of sweat glands. A woman who is heterozygous for this trait has patches of normal skin and patches of skin lacking sweat glands.

Inactivation of an X chromosome involves modification of the DNA and proteins bound to it called histones, including attachment of methyl groups ($-CH_3$) to DNA nucleotides.

▼ Figure 15.8 X inactivation and the tortoiseshell cat. The tortoiseshell gene is on the X chromosome, and the tortoiseshell phenotype requires the presence of two different alleles, one for orange fur and one for black fur. Normally, only females can have both alleles because only they have two X chromosomes. If a female cat is heterozygous for the tortoiseshell gene, she is tortoiseshell. Orange patches are formed by populations of cells in which the X chromosome with the orange allele is active; black patches have cells in which the X chromosome with the black allele is active. ("Calico" cats also have white areas, which are determined by another gene.)



Mastering Biology
HHMI Animation: X Inactivation
hhmi BioInteractive

(The regulatory role of DNA methylation is discussed in Concept 18.2.) A particular region of each X chromosome contains several genes involved in the inactivation process. The two regions, one on each X chromosome, associate briefly with each other in each cell at an early stage of embryonic development. Then one of the genes, called *XIST* (for X-inactive specific transcript), becomes active *only* on the chromosome that will become the Barr body (the inactive X). Multiple copies of the RNA product of this gene apparently attach to the X chromosome on which they are made, eventually almost covering it. Interaction of this RNA with the chromosome initiates X inactivation, and the RNA products of nearby genes help to regulate the process.

CONCEPT CHECK 15.2

1. A white-eyed female *Drosophila* is mated with a red-eyed (wild-type) male, the reciprocal cross of the one shown in Figure 15.3. What phenotypes and genotypes do you predict for the offspring from this cross?
2. Neither Tim nor Shonda has Duchenne muscular dystrophy, but their firstborn son does. What is the probability that a second child will have the disease? What is the probability if the second child is a boy? A girl?
3. **MAKE CONNECTIONS** Consider what you learned about dominant and recessive alleles in Concept 14.1. If a disorder were caused by a dominant X-linked allele, how would the inheritance pattern differ from what we see for recessive X-linked disorders?

For suggested answers, see Appendix A.

CONCEPT 15.3

Linked genes tend to be inherited together because they are located near each other on the same chromosome

The number of genes in a cell is far greater than the number of chromosomes; in fact, each chromosome (except the Y) has hundreds or thousands of genes. Genes located near each other on the same chromosome tend to be inherited together in genetic crosses; such genes are said to be genetically linked and are called **linked genes**. When geneticists follow linked genes in breeding experiments, the results deviate from those expected from Mendel's law of independent assortment.

How Linkage Affects Inheritance

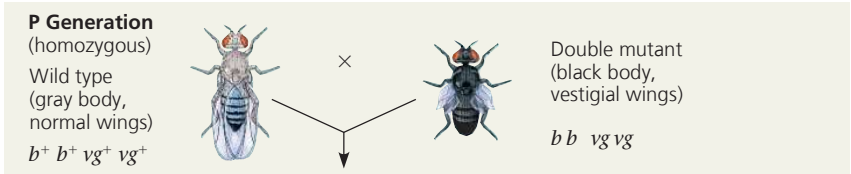
To see how linkage between genes affects the inheritance of two different characters, let's examine another of Morgan's *Drosophila* experiments. In this case, the characters are body color and wing size, each with two different phenotypes. Wild-type flies have gray bodies and normal-sized wings. In addition to these flies, Morgan had managed to obtain, through breeding, doubly mutant flies with black bodies and wings much smaller than normal, called vestigial wings. The mutant alleles are recessive to the wild-type alleles, and neither gene is on a sex chromosome. In his investigation of these two genes, Morgan carried out the crosses shown in **Figure 15.9**. The first was a P generation cross to generate F_1 dihybrid flies, and the second was essentially a testcross.

▼ Figure 15.9 Inquiry

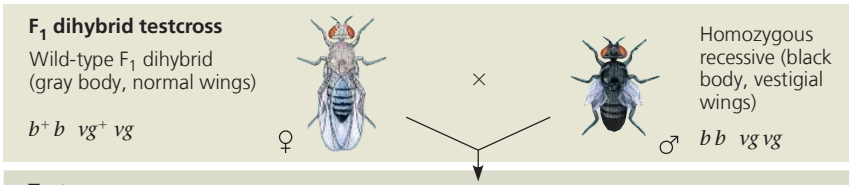
How does linkage between two genes affect inheritance of characters?

Experiment Morgan wanted to know whether the genes for body color and wing size are genetically linked and, if so, how this affects their inheritance. The alleles for body color are b^+ (gray) and b (black), and those for wing size are vg^+ (normal) and vg (vestigial).

Morgan mated true-breeding P (parental) generation flies—wild-type flies with black, vestigial-winged flies—to produce heterozygous F_1 dihybrids ($b^+\ b\ \ vg^+\ vg$), all of which are wild-type in appearance.

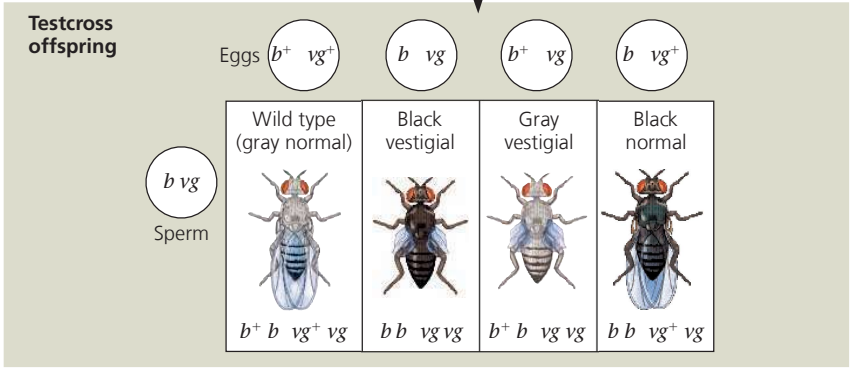


He then mated wild-type F_1 dihybrid females with homozygous recessive males. This testcross will reveal the genotype of the eggs made by the dihybrid female.



The male's sperm contributes only recessive alleles, so the phenotype of the offspring reflects the genotype of the female's eggs.

Note: Although only females (with pointed abdomens) are shown, half the offspring in each class would be males (with rounded abdomens).



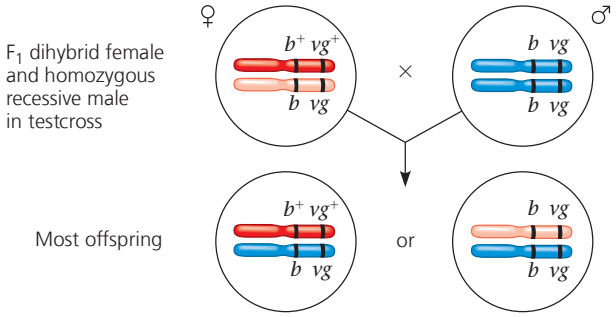
Predicted ratios of testcross offspring								
	if genes are located on different chromosomes:	1	:	1	:	1	:	1
	if genes are located on the same chromosome *and* parental alleles are always inherited together:	1	:	1	:	0	:	0
Results	Data from Morgan's experiment:	965	:	944	:	206	:	185

Conclusion Since most offspring had a parental (P generation) phenotype, Morgan concluded that the genes for body color and wing size are genetically linked on the same chromosome. However, the production of a relatively small number of offspring with nonparental phenotypes indicated that some mechanism occasionally breaks the linkage between specific alleles of genes on the same chromosome.

Data from T. H. Morgan and C. J. Lynch, The linkage of two factors in *Drosophila* that are not sex-linked, *Biological Bulletin* 23:174–182 (1912).

WHAT IF? *If the parental (P generation) flies had been true-breeding for gray body with vestigial wings and black body with normal wings, which phenotypic class(es) would be largest among the testcross offspring?*

The resulting flies had a much higher proportion of the combinations of traits seen in the P generation flies (called parental phenotypes) than would be expected if the two genes assorted independently. Morgan thus concluded that body color and wing size are usually inherited together in specific (parental) combinations because the genes are linked; they are near each other on the same chromosome:



As you proceed, be sure to keep in mind the distinction between the terms *linked genes* (two or more genes on the same chromosome that tend to be inherited together) and *sex-linked gene* (a single gene on a sex chromosome).

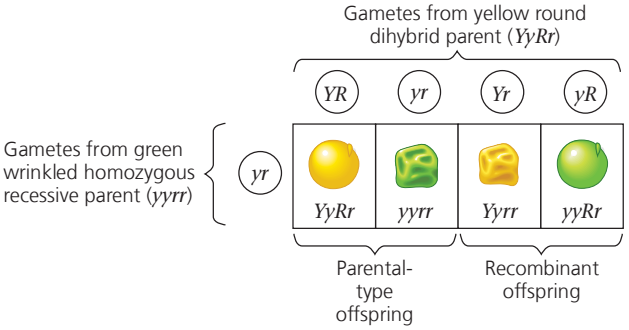
As Figure 15.9 shows, both of the combinations of traits not seen in the P generation (called nonparental phenotypes) were also produced in Morgan's experiments, suggesting that the body color and wing size alleles are not always linked genetically. To understand this conclusion, we need to further explore **genetic recombination**, the production of offspring with combinations of traits that differ from those found in either P generation parent.

Genetic Recombination and Linkage

Meiosis and random fertilization generate genetic variation among offspring of sexually reproducing organisms due to independent assortment of chromosomes, crossing over in meiosis I, and the possibility of any sperm fertilizing any egg (see Concept 13.4). Here we'll examine the chromosomal basis of recombination of alleles in relation to the genetic findings of Mendel and Morgan.

Recombination of Unlinked Genes: Independent Assortment of Chromosomes

Mendel learned from crosses in which he followed two characters that some offspring have combinations of traits that do not match those of either parent. For example, consider a cross of a dihybrid pea plant with yellow round seeds, heterozygous for both seed color and seed shape (*YyRr*), with a plant homozygous for both recessive alleles (with green wrinkled seeds, *yyrr*). (This cross acts as a testcross because the results will reveal the genotype not only of the dihybrid *YyRr* plant, which we know, but of the gametes made in that plant.) Let's represent the cross by the following Punnett square:



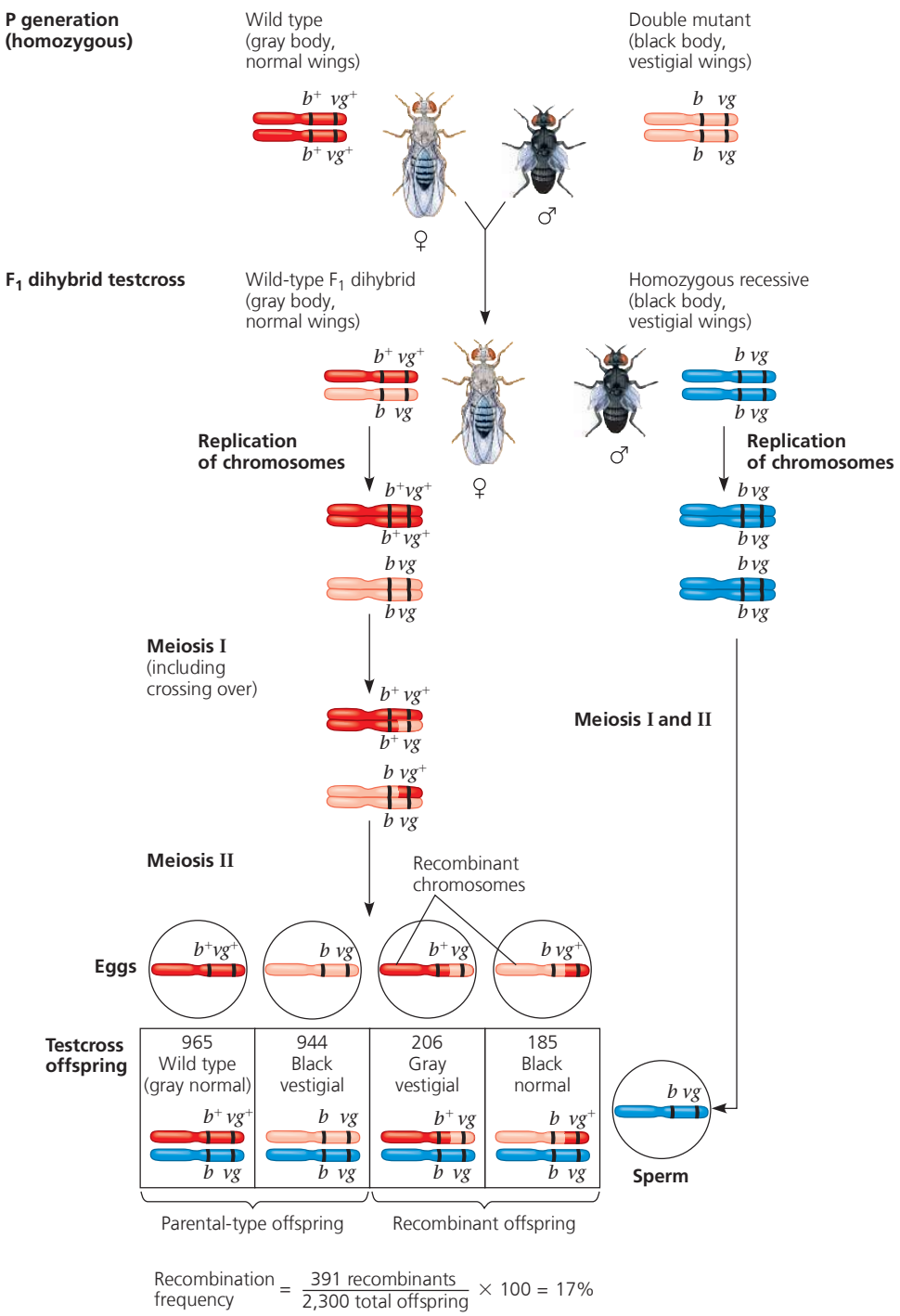
Notice in this Punnett square that one-half of the offspring are expected to inherit a phenotype that matches either of the phenotypes of the P (parental) generation originally crossed to produce the F_1 dihybrid (see Figure 15.2). These matching offspring are called **parental types** (short for phenotypes). But two nonparental phenotypes are also found among the offspring. Because these offspring have new combinations of seed shape and color, they are called **recombinant types**, or **recombinants** for short. When 50% of all offspring are recombinants, as in this example, geneticists say that there is a 50% frequency of recombination. The predicted phenotypic ratios among the offspring are similar to what Mendel actually found in his *YyRr* × *yyrr* crosses.

A 50% frequency of recombination in such testcrosses is observed for any two genes that are located on different chromosomes and thus cannot be linked. The physical basis of recombination between unlinked genes is the random orientation of homologous chromosomes at metaphase I of meiosis, which leads to the independent assortment of the two unlinked genes (see Figure 13.11 and the question in the Figure 15.4 legend).

Recombination of Linked Genes: Crossing Over

Now, let's explain the results of the *Drosophila* testcross in Figure 15.9. Recall that most (83%) of the offspring from the testcross for body color and wing size had parental phenotypes. That suggested that the two genes were on the same chromosome, since the occurrence of parental types with a frequency greater than 50% indicates that the genes are linked. About 17% of offspring, however, were recombinants.

Seeing these results, Morgan proposed that some process must occasionally break the physical connection between specific alleles of genes on the same chromosome. Later experiments showed that this process, now called **crossing over**, accounts for the recombination of linked genes. In crossing over, which occurs while replicated homologous chromosomes are paired during prophase of meiosis I, a set of proteins breaks the DNA molecules of one maternal and one paternal chromatid and rejoins each to the other



▶ Figure 15.10 Chromosomal basis for recombination of linked genes. In these diagrams re-creating the testcross in Figure 15.9, we track chromosomes as well as genes. The maternal chromosomes (those present in the wild-type F_1 dihybrid) are color-coded red and pink to distinguish one homolog from the other before any meiotic crossing over has occurred. Because crossing over between the b^+/b and vg^+/vg loci occurs in some, but not most, egg-producing cells, more eggs with parental-type chromosomes than with recombinant ones are produced in the mating females. Fertilization of the eggs by sperm of genotype $b\ vg$ gives rise to some recombinant offspring. The recombination frequency is the percentage of recombinant flies in the total pool of offspring.

DRAW IT *Suppose, as in the question at the bottom of Figure 15.9, the parental (P generation) flies were true-breeding for gray body with vestigial wings and black body with normal wings. Draw the chromosomes in each of the four possible kinds of eggs from an F_1 female, and label each chromosome as "parental" or "recombinant."*

Mastering Biology Animation: Linked Genes and Crossing Over

(see Figure 13.9). In effect, when a single crossover occurs, end portions of two nonsister chromatids trade places.

Figure 15.10 shows how crossing over in a dihybrid female fly resulted in recombinant eggs and ultimately recombinant offspring in Morgan's testcross. Most eggs had a chromosome with either the $b^+\ vg^+$ or $b\ vg$ parental genotype, but some eggs had a recombinant chromosome ($b^+\ vg$ or $b\ vg^+$). Fertilization of all classes of eggs by homozygous recessive sperm ($b\ vg$) produced an offspring population in which 17% exhibited a nonparental, recombinant phenotype, reflecting combinations of alleles not seen before in either P generation parent. In the **Scientific Skills Exercise**, you can use a statistical test to analyze the results from an F_1 dihybrid testcross and see whether the two genes assort independently or are linked.

Scientific Skills Exercise

Using the Chi-Square (χ^2) Test

Are Two Genes Linked or Unlinked? Genes that are in close proximity on the same chromosome will result in the linked alleles being inherited together more often than not. But how can you tell if certain alleles are inherited together due to linkage or whether they just happen to assort together randomly? In this exercise, you'll use a simple statistical test, the chi-square (χ^2) test, to analyze phenotypes of F_1 testcross progeny in order to see whether two genes are linked or unlinked.

How These Experiments Are Done If genes are unlinked and assorting independently, the phenotypic ratio of offspring from an F_1 testcross is expected to be 1:1:1:1 (see Figure 15.9). If the two genes are linked, however, the observed phenotypic ratio of the offspring will not match that ratio. Given that random fluctuations in the data do occur, how much must the observed numbers deviate from the expected numbers for us to conclude that the genes are not assorting independently but may instead be linked?

To answer this question, scientists use a statistical test. This test, called a chi-square (χ^2) test, compares an observed data set with an expected data set predicted by a hypothesis (here, that the genes are unlinked) and measures the discrepancy between the two, thus determining the "goodness of fit." If the discrepancy between the observed and expected data sets is so large that it is unlikely to have occurred by random fluctuation, we say there is statistically significant evidence against the hypothesis (or, more specifically, evidence for the genes being linked). If the discrepancy is small, then our observations are well explained by random variation alone. In this case, we say that the observed data are consistent with our hypothesis, or that the discrepancy is statistically insignificant. Note, however, that consistency with our hypothesis is not the same as proof of our hypothesis. Also, the size of the experimental data set is important: With small data sets like this one, even if the genes are linked, discrepancies might be small by chance alone if the linkage is weak. For simplicity, we overlook the effect of sample size here.

Data from the Simulated Experiment In cosmos plants, purple stem (*A*) is dominant to green stem (*a*), and short petals (*B*) is dominant to long petals (*b*). In a simulated cross, *AABB* plants were crossed with *aabb* plants to generate F_1 dihybrids (*AaBb*), which were then testcrossed (*AaBb* × *aabb*). A total of 900 offspring plants were scored for stem color and flower petal length.

Offspring from testcross of *AaBb* (F_1) × *aabb*	Purple stem/short petals (*A–B–*)*	Green stem/short petals (*aaB–*)	Purple stem/long petals (*A–bb*)	Green stem/long petals (*aabb*)
Expected ratio if the genes are unlinked	1	1	1	1
Expected number of offspring (of 900)				
Observed number of offspring (of 900)	220	210	231	239

*If the phenotype is dominant, a dash is used for the second allele; it could be either the dominant or recessive allele.



➤ **Cosmos plants**

INTERPRET THE DATA

1. The results in the data table are from a simulated F_1 dihybrid testcross. The hypothesis that the two genes are unlinked predicts that the offspring phenotypic ratio will be 1:1:1:1. Using this ratio, calculate the expected number of each phenotype out of the 900 total offspring, and enter the values in that data table.
2. The goodness of fit is measured by χ^2. This statistic measures the amounts by which the observed values differ from their respective predictions to indicate how closely the two sets of values match. The formula for calculating this value is

$$\chi^2 = \Sigma \frac{(o - e)^2}{e}$$

where Σ = sum of, o = observed and e = expected. Calculate the χ^2 value for the data using the table below. Fill out that table, carrying out the operations indicated in the top row. Then add up the entries in the last column to find the χ^2 value.

Testcross Offspring	Expected (e)	Observed (o)	Deviation ($o - e$)	$(o - e)^2$	$(o - e)^2/e$
A–B–		220			
aaB–		210			
A–bb		231			
aabb		239			
				χ^2 = Sum	

3. The χ^2 value means nothing on its own—it is used to find the probability that, assuming the hypothesis is true, the observed data set could have resulted from random fluctuations. A low probability suggests that the observed data are not consistent with the hypothesis and thus the hypothesis should be rejected. A standard cutoff point used by biologists is a probability of 0.05 (5%). If the probability corresponding to the χ^2 value is 0.05 or less, the differences between observed and expected values are considered statistically significant and the hypothesis (that the genes are unlinked) should be rejected. If the probability is above 0.05, the results are not statistically significant; the observed data are consistent with the hypothesis that the genes are unlinked.

   To find the probability, locate your χ^2 value in the χ^2 distribution table in Appendix D. The "degrees of freedom" (df) of your data set is the number of categories (here, 4 phenotypes) minus 1, so df = 3. **(a)** Determine which values on the df = 3 line of the table your calculated χ^2 value lies between. **(b)** The column headings for these values show the probability range for your χ^2 number. Based on whether there are nonsignificant ($p > 0.05$) or significant ($p \leq 0.05$) differences between the observed and expected values, are the data consistent with the hypothesis that the two genes are unlinked and assorting independently, or is there enough evidence to reject this hypothesis?

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.

New Combinations of Alleles: Variation for Natural Selection

EVOLUTION The physical behavior of chromosomes during meiosis contributes to the generation of variation in offspring (see Concept 13.4). Each pair of homologous chromosomes lines up independently of other pairs during metaphase I, and crossing over prior to that, during prophase I, can mix and match parts of maternal and paternal homologs. Mendel's elegant experiments show that the behavior of the abstract entities known as genes—or, more concretely, alleles of genes—also leads to variation in offspring (see Concept 14.1). Now, putting these different ideas together, you can see that the recombinant chromosomes resulting from crossing over may bring alleles together in new combinations, and the subsequent events of meiosis distribute to gametes the recombinant chromosomes in a multitude of combinations, such as the new variants seen in Figures 15.9 and 15.10. Random fertilization then increases even further the number of variant allele combinations that can be created.

This abundance of genetic variation provides the raw material on which natural selection works. If the traits conferred by particular combinations of alleles are better suited for a given environment, organisms possessing those genotypes will be expected to thrive and leave more offspring, ensuring the continuation of their genetic complement. In the next generation, of course, the alleles will be shuffled anew. Ultimately, the interplay between environment and phenotype (and thus genotype) will determine which genetic combinations persist over time.

Mapping the Distance Between Genes Using Recombination Data: *Scientific Inquiry*

The discovery of linked genes and recombination due to crossing over motivated one of Morgan's students, Alfred H. Sturtevant, to work out a method for constructing a **genetic map**, an ordered list of the genetic loci along a particular chromosome.

Sturtevant hypothesized that the percentage of recombinant offspring, the *recombination frequency*, calculated from experiments like the one in Figures 15.9 and 15.10, depends on the distance between genes on a chromosome. He assumed that crossing over is a random event, with the chance of crossing over approximately equal at all points along a chromosome. Based on these assumptions, Sturtevant predicted that *the farther apart two genes are, the higher the probability that a crossover will occur between them and therefore the higher the recombination frequency*. His reasoning was simple: The greater the distance between two genes, the more points there are between them where crossing over can occur. Using recombination data from various fruit fly crosses, Sturtevant proceeded to assign relative positions to genes on the same chromosomes—that is, to *map* genes.

A genetic map based on recombination frequencies is called a **linkage map**. **Figure 15.11** shows Sturtevant's linkage map of three genes: the body color (*b*) and wing size (*vg*) genes depicted in Figure 15.10 and a third gene, called cinnabar (*cn*). Cinnabar is one of many *Drosophila* genes affecting eye color. Cinnabar eyes, a mutant phenotype, are a brighter red than the wild-type color. The recombination frequency between *cn* and *b* is 9%; that between *cn* and *vg*, 9.5%; and that between *b* and *vg*, 17%. In other words, crossovers between *cn* and *b* and between *cn* and *vg* are about half as frequent as crossovers between *b* and *vg*. Only a map that locates *cn* about midway between *b* and *vg* is consistent with these data, as you can prove to yourself by drawing alternative maps. Sturtevant expressed the distances between genes in **map units**, defining one map unit as equivalent to a 1% recombination frequency.

In practice, the interpretation of recombination data is more complicated than this example suggests. Some genes

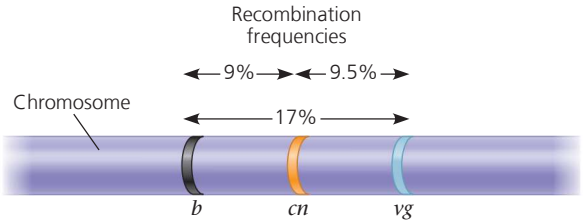
▼ Figure 15.11 Research Method

Constructing a Linkage Map

Application A linkage map shows the relative locations of genes along a chromosome.

Technique A linkage map is based on the assumption that the probability of a crossover between two genetic loci is proportional to the distance separating the loci. The recombination frequencies used to construct a linkage map for a particular chromosome are obtained from experimental crosses, such as the cross depicted in Figures 15.9 and 15.10. The distances between genes are expressed as map units, with one map unit equivalent to a 1% recombination frequency. Genes are arranged on the chromosome in the order that best fits the data.

Results In this example, the observed recombination frequencies between three *Drosophila* gene pairs (*b* and *cn*, 9%; *cn* and *vg*, 9.5%; and *b* and *vg*, 17%) best fit a linear order in which *cn* is positioned about halfway between the other two genes:



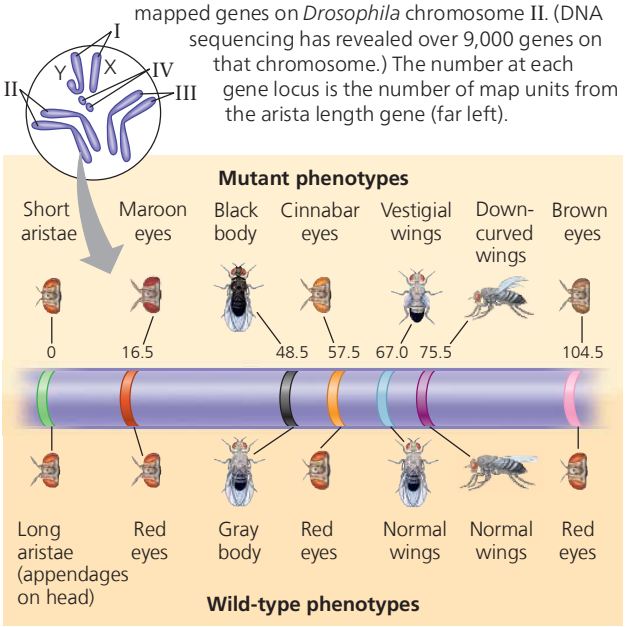
The *b*-*vg* recombination frequency (17%) is slightly less than the sum of the *b*-*cn* and *cn*-*vg* frequencies (9 + 9.5 = 18.5%) because of the few times that one crossover occurs between *b* and *cn* and another crossover occurs between *cn* and *vg*. The second crossover would "cancel out" the first, reducing the observed *b*-*vg* recombination frequency while contributing to the frequency between each of the closer pairs of genes. The value of 18.5% (18.5 map units) is closer to the actual distance between the genes. In practice, a geneticist would add the smaller distances in constructing a map.

on a chromosome are so far from each other that a crossover between them is virtually certain. The observed frequency of recombination in crosses involving two such genes can have a maximum value of 50%, a result indistinguishable from that for genes on different chromosomes. In this case, the physical connection between genes on the same chromosome is not reflected in the results of genetic crosses. Despite being on the same chromosome and thus being *physically connected*, the genes are *genetically unlinked*; alleles of such genes assort independently, as if they were on different chromosomes. In fact, at least two of the genes for pea characters that Mendel studied are now known to be on the same chromosome, but the distance between them is so great that linkage is not observed in genetic crosses. Consequently, the two genes behaved as if they were on different chromosomes in Mendel's experiments. Genes located far apart on a chromosome are mapped by adding the recombination frequencies from crosses involving closer pairs of genes lying between the two distant genes.

Using recombination data, Sturtevant and his colleagues were able to map numerous *Drosophila* genes in linear arrays. They found that the genes clustered into four groups of linked genes (*linkage groups*). Light microscopy had revealed four pairs of chromosomes in *Drosophila*, so the linkage map provided additional evidence that genes are located on chromosomes. Each chromosome has a linear array of specific genes, each gene with its own locus **(Figure 15.12)**.

Because a linkage map is based strictly on recombination frequencies, it gives only an approximate picture of a chromosome. The frequency of crossing over is not actually uniform over the length of a chromosome, as Sturtevant assumed, and therefore map units do not correspond to actual physical distances (in nanometers, for instance). A linkage map does portray the order of genes along a chromosome, but it does not accurately portray the precise locations of those genes. Other methods enable geneticists to construct *cytogenetic maps* of chromosomes, which locate genes with respect to chromosomal features, such as stained bands, that can be seen in the microscope. Technical advances over recent decades have enormously increased the rate and affordability of DNA sequencing. Therefore, today, most researchers sequence whole genomes to map the locations of genes of a given species. The entire nucleotide sequence is the ultimate physical map of a chromosome, revealing the physical distances in DNA nucleotides between gene loci (see Concept 21.1). Comparing a linkage map with such a physical map or with a cytogenetic map of the same chromosome, we find that the linear order of genes is identical in all the maps, but the spacing between genes is not.

▼ Figure 15.12 A partial genetic (linkage) map of a *Drosophila* chromosome. This partial map shows just seven mapped genes on *Drosophila* chromosome II. (DNA sequencing has revealed over 9,000 genes on that chromosome.) The number at each gene locus is the number of map units from the arista length gene (far left).



CONCEPT CHECK 15.3

1. When two genes are located on the same chromosome, what is the physical basis for the production of recombinant offspring in a testcross between a dihybrid parent and a double-mutant (recessive) parent?
2. **VISUAL SKILLS** For each type of offspring of the testcross in Figure 15.9, explain the relationship between its phenotype and the alleles contributed by the female parent. (It will be useful to draw out the chromosomes of each fly and follow the alleles throughout the cross.)
3. **WHAT IF?** Genes *A*, *B*, and *C* are located on the same chromosome. Testcrosses show that the recombination frequency between *A* and *B* is 28% and that between *A* and *C* is 12%. Can you determine the linear order of these genes? Explain.

For suggested answers, see Appendix A.

CONCEPT 15.4

Alterations of chromosome number or structure cause some genetic disorders

As you have learned so far in this chapter, the phenotype of an organism can be affected by small-scale changes involving individual genes. Random mutations are the source of all new alleles, which can lead to new phenotypic traits.

Large-scale chromosomal changes can also affect an organism's phenotype. Physical and chemical disturbances, as well as errors during meiosis, can damage chromosomes in major ways or alter their number in a cell. Large-scale chromosomal alterations in humans and other mammals often lead to spontaneous abortion (miscarriage) of a fetus, and individuals born with these types of genetic defects commonly exhibit various developmental disorders. Plants appear to tolerate such genetic defects better than animals do.

Mastering Biology
Interview with Terry Orr-Weaver: Why are humans so bad at meiosis?

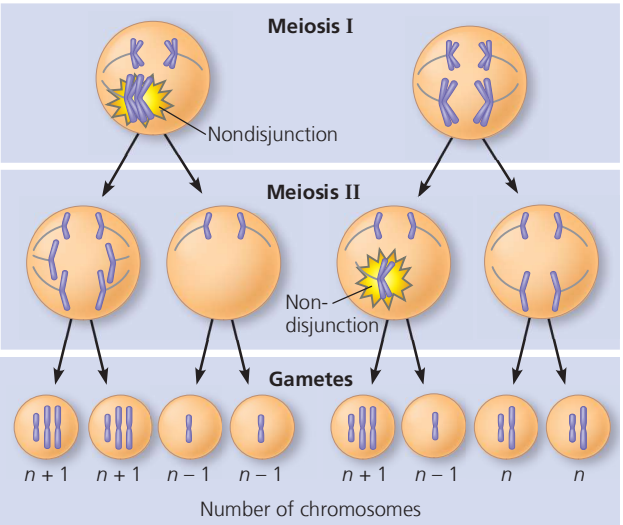


Abnormal Chromosome Number

Ideally, the meiotic spindle distributes chromosomes to daughter cells without error. But there is an occasional mishap, called a **nondisjunction**, in which the members of a pair of homologous chromosomes do not move apart properly during meiosis I or sister chromatids fail to separate during meiosis II **(Figure 15.13)**. In nondisjunction, one gamete receives two of the same type of chromosome and another gamete receives no copy. The other chromosomes are usually distributed normally.

If either of the aberrant gametes unites with a normal one at fertilization, the zygote will also have an abnormal number of a particular chromosome, a condition known as **aneuploidy**. Fertilization involving a gamete that has no copy of a particular chromosome will lead to a missing chromosome in the zygote (so that the cell has $2n - 1$ chromosomes); the aneuploid zygote is said to be **monosomic** for that chromosome. If a chromosome is present in triplicate in the zygote (so that the cell has $2n + 1$ chromosomes), the aneuploid cell is **trisomic** for that chromosome. Mitosis will subsequently transmit the anomaly to all embryonic cells. Monosomy and trisomy are estimated to occur in 10–25% of human conceptions and are the main reason for pregnancy loss. If the organism survives, it usually has a set of traits caused by the abnormal dose of the genes associated with the extra or missing chromosome. Down syndrome is an example of trisomy in humans that will be discussed later.

▼ Figure 15.13 Meiotic nondisjunction. Gametes with an abnormal chromosome number can arise by nondisjunction in either meiosis I or meiosis II. For simplicity, the figure does not show the spores formed by meiosis in plants. Ultimately, spores form gametes that have the defects shown. (See Figure 13.6.)



(a) Nondisjunction of homologous chromosomes in meiosis I

(b) Nondisjunction of sister chromatids in meiosis II

Nondisjunction can also occur during mitosis. If such an error takes place early in embryonic development, then the aneuploid condition is passed along by mitosis to a large number of cells and is likely to have a substantial effect on the organism.

Some organisms have more than two complete chromosome sets in all somatic cells. The general term for this chromosomal alteration is **polyploidy**; the specific terms *triploidy* ($3n$) and *tetraploidy* ($4n$) indicate three and four chromosomal sets, respectively. One way a triploid cell may arise is by the fertilization of an abnormal diploid egg produced by nondisjunction of all its chromosomes. Tetraploidy could result from the failure of a $2n$ zygote to divide after replicating its chromosomes. Subsequent normal mitotic divisions would then produce a $4n$ embryo.

Polyploidy is fairly common in the plant kingdom. The spontaneous origin of polyploid individuals plays an important role in plant evolution (see Concept 24.2). Many species we eat are polyploid: Bananas are triploid, wheat hexaploid ($6n$), and strawberries octoploid ($8n$). Polyploid animal species are much less common, but there are a few fishes and amphibians known to be polyploid. In general, polyploids are more nearly normal in appearance than aneuploids. One extra (or missing) chromosome apparently disrupts genetic balance more than does an entire extra set of chromosomes.

Mastering Biology **Animation: Polyploid Plants**

Alterations of Chromosome Structure

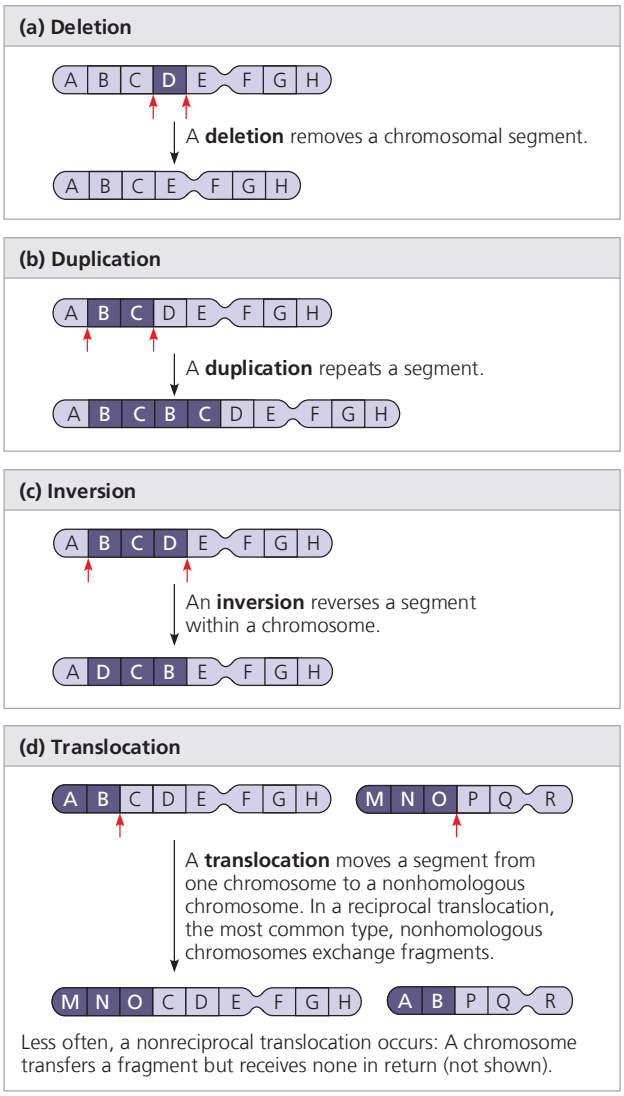
Errors in meiosis or damaging agents such as radiation can cause breakage of a chromosome, which can lead to four types of changes in chromosome structure **(Figure 15.14)**. A **deletion** occurs when a chromosomal fragment is lost. The affected chromosome is then missing certain genes. A broken fragment may become reattached as an extra segment to a sister or nonsister chromatid, producing a **duplication** of a portion of that chromosome. A chromosomal fragment may also reattach to the original chromosome but in the reverse orientation, producing an **inversion**. A fourth possible result of chromosomal breakage is for the fragment to join a nonhomologous chromosome, a rearrangement called a **translocation**.

Deletions and duplications are especially likely to occur during meiosis. In crossing over, nonsister chromatids sometimes exchange unequal-sized segments of DNA, so that one partner gives up more genes than it receives (see Figure 21.13). The products of such an unequal crossover are one chromosome with a deletion and one chromosome with a duplication.

A diploid embryo that is homozygous for a large deletion (both homologs have the deletion) or, if a male, has a single X chromosome with a large deletion is usually missing a number of essential genes. This condition is typically lethal.

Duplications and translocations also tend to be harmful. In reciprocal translocations, in which segments are exchanged between nonhomologous chromosomes, and in

▼ Figure 15.14 Alterations of chromosome structure. Red arrows indicate breakage points. Dark purple highlights the chromosomal parts affected by the rearrangements.



Mastering Biology **Animation: Alterations of Chromosome Structure**

inversions, the balance of genes is not abnormal—all genes are present in their normal doses. Nevertheless, translocations and inversions can alter phenotype because a gene's expression can be influenced by its location among neighboring genes, which can have devastating effects.

Human Conditions Due to Chromosomal Alterations

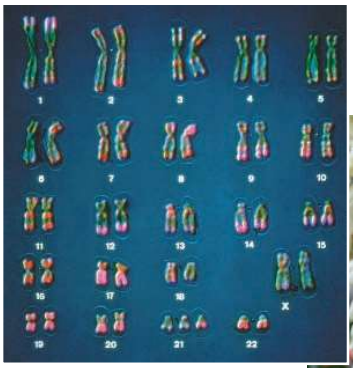
Alterations of chromosome number and structure are associated with a number of human conditions, which vary in severity. As described earlier, nondisjunction in meiosis results in aneuploidy in gametes and any resulting zygotes. Although the frequency of aneuploid zygotes may be quite high in humans, most of these chromosomal alterations are so disastrous to development that the affected embryos are spontaneously aborted long before birth. However, some types of aneuploidy appear to upset the genetic balance less than others, so individuals with certain aneuploid conditions can survive to birth and beyond. These individuals have a set of traits—a *syndrome*—characteristic of the type of aneuploidy. Genetic disorders caused by aneuploidy can be diagnosed before birth by fetal testing (see Figure 14.19).

Down Syndrome (Trisomy 21)

One aneuploid condition, **Down syndrome**, affects approximately one out of every 830 children born in the United States **(Figure 15.15)**. Down syndrome is usually the result of an extra chromosome 21, so that each body cell has a total of 47 chromosomes. Because the cells are trisomic for chromosome 21, Down syndrome is often called *trisomy 21*. Down syndrome includes characteristic facial features, short stature, correctable heart defects, and developmental delays. Individuals with Down syndrome have an increased chance of developing leukemia and Alzheimer's disease but have a lower rate of high blood pressure, atherosclerosis (hardening of the arteries), stroke, and many types of solid tumors. People with Down syndrome, on average, have a life span shorter than is typical; however, most live to middle age and beyond with proper medical care. Many live independently or at home with their families, are employed, and are valuable contributors to their communities. Almost all males and about half of females with Down syndrome are sexually underdeveloped and sterile.

The frequency of Down syndrome increases with the age of the mother. While the syndrome occurs in just 0.1% (1 out of 900) births of children born to women at age 30, the risk

▼ Figure 15.15 Down syndrome. The karyotype shows trisomy 21, the most common cause of Down syndrome. The child exhibits the facial features characteristic of this syndrome.





climbs to 1% (1 out of 100) for mothers at age 40 and is even higher for older mothers. The correlation of Down syndrome with maternal age has not yet been explained. Most cases result from nondisjunction during meiosis I, and some research points to an age-dependent abnormality in meiosis. Trisomies of some other chromosomes (13 and 18) also increase in incidence with maternal age, although infants with other autosomal trisomies rarely survive for long. Medical experts recommend that prenatal screening for trisomies in the embryo be offered to all pregnant women, due to its low risk and useful results. In 2008, a law was passed stipulating that medical practitioners give accurate, up-to-date information about any prenatal or postnatal diagnosis received by parents and that they connect parents with appropriate support services.

Aneuploidy of Sex Chromosomes

Aneuploid conditions involving sex chromosomes appear to upset the genetic balance less than those involving autosomes. This may be because the Y chromosome carries relatively few genes. Also, extra copies of the X chromosome simply become inactivated as Barr bodies.

An extra X chromosome in a male, producing XXY, occurs approximately once in every 650 live male births. People with this condition, called *Klinefelter syndrome*, have male sex organs, but their testes are small and produce little or no sperm. Signs and symptoms vary widely, but may include taller than average height, less muscle mass, and enlarged breast tissue. Affected individuals may have learning disabilities. Testosterone replacement therapy helps stimulate changes that typically occur at puberty in males. About one of every 1,000 males is born with an extra Y chromosome (XYY). These males undergo typical sexual development and do not exhibit any well-defined syndrome, but tend to be taller than average.

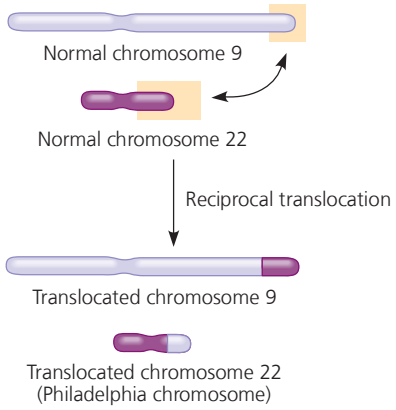
Females with trisomy X (XXX), which occurs once in approximately 1,000 live female births, are generally healthy and have no unusual physical features other than being slightly taller than average. Females with trisomy X are at risk for learning disabilities. Monosomy X, which is called *Turner syndrome*, occurs about once in every 2,500 female births and is the only known viable monosomy in humans. Although these X0 individuals are phenotypically female, they are usually sterile because their sex organs do not mature. When provided with estrogen replacement therapy, girls with Turner syndrome do develop secondary sex characteristics. Most have typical intelligence.

Disorders Caused by Structurally Altered Chromosomes

Many deletions in human chromosomes, even in a heterozygous state, cause severe problems. One such syndrome, known as *cri du chat* ("cry of the cat"), results from a specific deletion in chromosome 5. A child born with this deletion has a small head with unusual facial features, severe intellectual disabilities, and a cry that sounds like the mewing of a distressed cat. Such individuals usually die in infancy or early childhood.

Chromosomal translocations can also occur during mitosis; some have been implicated in certain cancers, including *chronic myelogenous leukemia* (*CML*). This disease occurs when a reciprocal translocation happens during mitosis of cells that are precursors of white blood cells. The exchange of a large portion of chromosome 22 with a small fragment from a tip of chromosome 9 produces a much shortened, easily recognized chromosome 22, called the *Philadelphia chromosome* **(Figure 15.16)**. Such an exchange causes cancer by creating a new "fused" gene that leads to uncontrolled cell cycle progression. (The mechanism of gene activation will be discussed in Concept 18.5.)

▼ Figure 15.16 Translocation associated with chronic myelogenous leukemia (CML). The cancerous cells in nearly all CML patients contain an abnormally short chromosome 22, the so-called Philadelphia chromosome, and an abnormally long chromosome 9. These altered chromosomes result from the reciprocal translocation shown here, which presumably occurred in a single white blood cell precursor undergoing mitosis and was then passed along to all descendant cells.



CONCEPT CHECK 15.4

1. About 5% of individuals with Down syndrome have a chromosomal translocation in which a third copy of chromosome 21 is attached to chromosome 14. If this translocation occurred in a parent's gonad, how could it lead to Down syndrome in a child?
2. **MAKE CONNECTIONS** The ABO blood type locus has been mapped on chromosome 9. A father who has type AB blood and a mother who has type O blood have a child with trisomy 9 and type A blood. Using this information, can you tell in which parent the nondisjunction occurred? Explain your answer. (See Figures 14.11 and 15.13.)
3. **MAKE CONNECTIONS** The gene that is activated on the Philadelphia chromosome codes for an intracellular tyrosine kinase. Review the discussion of cell cycle control in Concept 12.3, and explain how the activation of this gene could contribute to the development of cancer.

For suggested answers, see Appendix A.

CONCEPT 15.5

Some inheritance patterns are exceptions to standard Mendelian inheritance

In the previous section, you learned about deviations from the usual patterns of chromosomal inheritance due to abnormal events in meiosis and mitosis. We conclude this chapter by describing two normally occurring exceptions to Mendelian genetics, one involving genes located in the nucleus and the other involving genes located outside the nucleus. In both cases, the sex of the parent contributing an allele is a factor in the pattern of inheritance.

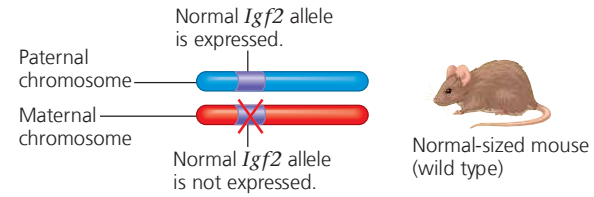
Genomic Imprinting

Throughout our discussions of Mendelian genetics and the chromosomal basis of inheritance, we have assumed that a given allele will have the same effect whether it was inherited from the mother or the father. This is probably a safe assumption most of the time. For example, when Mendel crossed purple-flowered pea plants with white-flowered pea plants, he observed the same results regardless of whether the purple-flowered parent supplied the eggs or the sperm. In recent years, however, geneticists have identified a number of traits in placental mammals (and some flowering plants) that depend on which parent passed along the alleles for those traits. Such variation in phenotype depending on whether an allele is inherited from the male or female parent is called **genomic imprinting**. (Note that unlike sex-linked genes, most imprinted genes are on autosomes.) Using newer DNA sequence-based methods, about 100 imprinted genes have been identified in humans and 120–180 in mice.

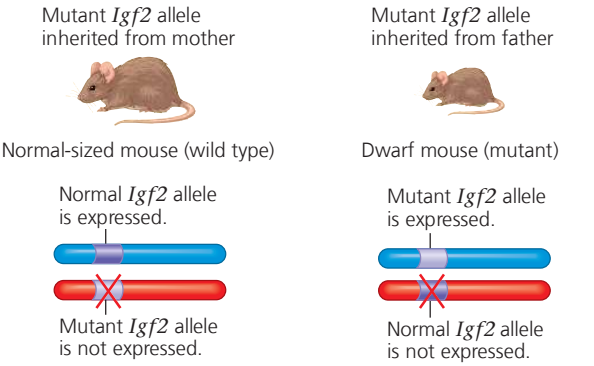
Genomic imprinting occurs during gamete formation and results in the silencing of a particular allele of certain genes. Because these genes are imprinted differently in sperm and eggs, the offspring expresses only one allele of an imprinted gene, the one that has been inherited from a specific parent—either the female parent or the male parent, depending on the particular gene. The imprints are then transmitted to all body cells during early development. In each generation, the old imprints are "erased" in gamete-producing cells, and the chromosomes of the developing gametes are newly imprinted according to the sex of the individual forming the gametes. In a given species, the imprinted genes are always imprinted in the same way. For instance, a gene imprinted for maternal allele expression is always imprinted this way, generation after generation.

Consider, for example, the mouse gene for insulin-like growth factor 2 (*Igf2*), one of the first imprinted genes to be identified. Although this growth factor is required for normal prenatal growth, only the paternal allele is expressed **(Figure 15.17a)**. Evidence that the *Igf2* gene is imprinted came initially from crosses between normal-sized (wild-type) mice and dwarf (mutant) mice homozygous for a recessive mutation in the *Igf2* gene. The phenotypes of heterozygous offspring (with one normal allele and one mutant) differed depending on whether the mutant allele came from the mother or the father **(Figure 15.17b)**.

▼ **Figure 15.17 Genomic imprinting of the mouse *Igf2* gene.**



(a) Homozygote. A mouse homozygous for the wild-type *Igf2* allele is normal-sized. Only the paternal allele of this gene is expressed.



(b) Heterozygotes. Heterozygotes have different phenotypes depending on which allele is contributed by each parent. Matings between wild-type mice and those homozygous for the recessive mutant *Igf2* allele produce heterozygous offspring. Since the maternal *Igf2* allele is not expressed, the dwarf (mutant) phenotype is seen only when the father contributes the mutant allele.

What exactly is a genomic imprint? It turns out that imprinting can involve either silencing an allele in one type of gamete (egg or sperm) or activating it in the other. In many cases, the imprint seems to consist of methyl ($-CH_3$) groups that are added to cytosine nucleotides of one of the alleles. Such methylation may silence the allele, an effect consistent with evidence that heavily methylated genes are usually inactive (see Concept 18.2). However, for a few genes, methylation has been shown to *activate* expression of the allele. This is the case for the *Igf2* gene: Methylation of certain cytosines on the paternal chromosome leads to expression of the paternal *Igf2* allele, by an indirect mechanism involving chromatin structure and protein-DNA interactions.

Genomic imprinting may affect only a small fraction of the genes in mammalian genomes, but most of the known imprinted genes are critical for embryonic development. Recent research has revealed that imprinted genes are involved in regulation of body temperature, sleep, and other metabolic functions. Some evolutionary biologists have proposed that

genomic imprinting represents a kind of evolved "competition" between males (who benefit from larger offspring more likely to pass on the father's genes) and female mammals (who benefit from smaller offspring during birth). In experiments with mice, embryos engineered to inherit both copies of certain chromosomes from the same parent usually die before birth, whether that parent is male or female. Normal development seems to require that embryonic cells have exactly one active copy—not zero, not two—of certain genes. The association of improper imprinting with abnormal development and certain cancers has stimulated ongoing studies of how different genes are imprinted.

Mastering Biology
Interview with Shirley Tilghman: Discovering imprinted genes



Inheritance of Organelle Genes

Although our focus in this chapter has been on the chromosomal basis of inheritance, we end with an important amendment: Not all of a eukaryotic cell's genes are located on nuclear chromosomes, or even in the nucleus; some genes are located in organelles in the cytoplasm. Because they are outside the nucleus, these genes are sometimes called *extranuclear genes* or *cytoplasmic genes*. Mitochondria, as well as chloroplasts and other plastids in plants, contain small circular DNA molecules that carry a number of genes. These organelles reproduce themselves and transmit their genes to daughter organelles. Genes on organelle DNA are not distributed to offspring according to the same rules that direct the distribution of nuclear chromosomes during meiosis, so they do not display Mendelian inheritance.

The first hint that extranuclear genes exist came from studies by the German scientist Carl Correns on the inheritance of yellow or white patches on the leaves of an otherwise green plant. In 1909, he observed that the coloration of the offspring was determined only by the maternal parent (the source of eggs) and not by the paternal parent (the source of sperm). Subsequent research showed that such coloration patterns, or variegation **(Figure 15.18)**, are due to mutations in genes controlling pigmentation that are located in the DNA circles in the organelles called plastids (see Concept 6.5). In most plants, a zygote receives all its plastids from the cytoplasm of the egg and none from the sperm, which contributes little more than a haploid set of chromosomes. An egg may contain plastids with different alleles for a pigmentation gene. As the zygote develops, plastids containing wild-type or mutant pigmentation genes are distributed randomly to daughter cells. The pattern of leaf coloration exhibited by a plant depends on the ratio of wild-type to mutant plastids in its various tissues.



◀ Figure 15.18 A painted nettle coleus plant. The variegated (patterned) leaves on this coleus plant (*Plectranthus scutellarioides*) result from mutations that affect expression of pigment genes located in plastids, which generally are inherited from the maternal parent.

Similar maternal inheritance is also the rule for mitochondrial genes in most animals and plants, because the mitochondria passed on to a zygote come from the cytoplasm of the egg. (The few mitochondria contributed by the sperm appear to be destroyed in the egg.) The products of most mitochondrial genes help make up some of the protein complexes of the electron transport chain and ATP synthase (see Figure 9.14). Defects in one or more of these proteins, therefore, reduce the amount of ATP the cell can make and have been shown to cause a number of human disorders in as many as one out of every 5,000 births. Because the parts of the body most susceptible to energy deprivation are the nervous system and the muscles, most mitochondrial diseases primarily affect these systems. For example, *mitochondrial myopathy* causes weakness, intolerance of exercise, and muscle deterioration. Another mitochondrial disorder is *Leber's hereditary optic neuropathy*, which can produce sudden blindness in people as young as their 20s or 30s. The four mutations found thus far to cause this disorder affect oxidative phosphorylation during cellular respiration, a crucial function for the cell (see Concept 9.4).

The fact that mitochondrial disorders are inherited only from the mother suggests a way to avoid passing along these disorders. The chromosomes from the egg of an affected mother could be transferred to an egg of a healthy donor that has had its own chromosomes removed. This "two-mother" egg could then be fertilized by a sperm from the prospective father and transplanted into the womb of the prospective mother, becoming an embryo with three parents. After optimizing conditions for this approach in monkeys, researchers reported in 2013 that they have successfully carried out this procedure on human eggs. In 2016, clinics in Mexico and China reported the births of such "three-parent" babies. In the United Kingdom, this procedure was approved for use in 2017. In the United States, approval by the relevant federal agencies has not yet occurred.

In addition to the rarer diseases clearly caused by defects in mitochondrial DNA, mitochondrial mutations inherited from a person's mother may contribute to at least some types of diabetes and heart disease, as well as to other disorders that commonly debilitate the elderly, such as Alzheimer's disease. In the course of a lifetime, new mutations gradually accumulate in our mitochondrial DNA, and some researchers think that these mutations play a role in the normal aging process.

CONCEPT CHECK 15.5

1. Gene dosage—the number of copies of a gene that are actively being expressed—is important to proper development. Identify and describe two processes that establish the proper dosage of certain genes.
2. Reciprocal crosses between two primrose varieties, A and B, produced the following results: A female × B male → offspring with all green (nonvariegated) leaves; B female × A male → offspring with patterned (variegated) leaves. Explain these results.
3. **WHAT IF?** Mitochondrial genes are critical to the energy metabolism of cells, but mitochondrial disorders caused by mutations in these genes are generally not lethal. Why not?

For suggested answers, see Appendix A.

15 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 15.1

Morgan showed that Mendelian inheritance has its physical basis in the behavior of chromosomes *(pp. 295–297)*

- Morgan's work with an eye color gene in *Drosophila* led to the **chromosome theory of inheritance**, which states that genes are located on chromosomes and that the behavior of chromosomes during meiosis accounts for Mendel's laws.

? *What characteristic of the sex chromosomes allowed Morgan to correlate their behavior with that of the alleles of the eye color gene?*

CONCEPT 15.2

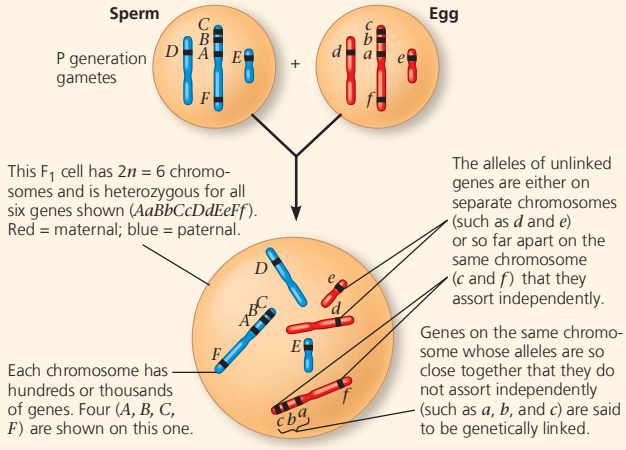
Sex-linked genes exhibit unique patterns of inheritance *(pp. 298–300)*

- Sex is often chromosomally based. Humans and other mammals have an X-Y system in which sex is largely determined by whether a Y chromosome is present. Other systems are found in birds, fishes, and insects.
- The sex chromosomes carry **sex-linked genes**, virtually all of which are on the X chromosome (**X-linked**). Any male who inherits a recessive X-linked allele (from his mother) will express the trait, such as color blindness.
- In mammalian females, one of the two X chromosomes in each cell is randomly inactivated during early embryonic development, becoming highly condensed into a **Barr body**.

? *Why are males affected by X-linked disorders much more often than females?*

CONCEPT 15.3

Linked genes tend to be inherited together because they are located near each other on the same chromosome *(pp. 301–306)*



- An F_1 dihybrid testcross yields **parental types** with the same combination of traits as those in the P generation parents and **recombinant types** (**recombinants**) with new combinations of traits not seen in either P generation parent. Because of the independent assortment of chromosomes, unlinked genes exhibit a 50% frequency of recombination in the gametes. For genetically **linked genes, crossing over** between nonsister chromatids during meiosis I accounts for the observed recombinants, always less than 50%.
- The order of genes on a chromosome and the relative distances between them can be deduced from recombination frequencies observed in genetic crosses. These data allow construction of a **linkage map** (a type of **genetic map**). The farther apart genes are, the more likely their allele combinations will be recombined during crossing over.

? *Why are specific alleles of two distant genes more likely to show recombination than those of two closer genes?*

CONCEPT 15.4

Alterations of chromosome number or structure cause some genetic disorders *(pp. 306–309)*

- **Aneuploidy**, an abnormal chromosome number, can result from **nondisjunction** during meiosis. When a normal gamete unites with one containing two copies or no copies of a particular chromosome, the resulting zygote and its descendant cells either have one extra copy of that chromosome (**trisomy**, $2n + 1$) or are missing a copy (**monosomy**, $2n - 1$). **Polyploidy** (extra sets of chromosomes) can result from nondisjunction of all chromosomes.
- Chromosome breakage can result in alterations of chromosome structure: **deletions, duplications, inversions**, and **translocations**. Translocations can be reciprocal or nonreciprocal.
- Changes in the number of chromosomes per cell or in the structure of individual chromosomes can affect the phenotype and, in some cases, lead to disorders. Such alterations cause **Down syndrome** (usually due to trisomy of chromosome 21), certain cancers associated with chromosomal translocations that occur during mitosis, and various other human disorders.

? *Why are inversions and reciprocal translocations less likely to be lethal than are aneuploidy, duplications, deletions, and nonreciprocal translocations?*

CONCEPT 15.5

Some inheritance patterns are exceptions to standard Mendelian inheritance *(pp. 310–311)*

- In mammals, the phenotypic effects of a small number of particular genes depend on which allele is inherited from each parent, a phenomenon called **genomic imprinting**. Imprints are formed during gamete production, with the result that one allele (either maternal or paternal) is not expressed in offspring.
- The inheritance of traits controlled by the genes present in mitochondria and plastids depends solely on the maternal parent because the zygote's cytoplasm containing these organelles comes from the egg. Some diseases affecting the nervous and muscular systems are caused by defects in mitochondrial genes that prevent cells from making enough ATP.

? *Explain how genomic imprinting and inheritance of mitochondrial and chloroplast DNA are exceptions to standard Mendelian inheritance.*

TEST YOUR UNDERSTANDING

For multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. A man with hemophilia (a recessive, sex-linked condition) has a daughter without hemophilia who marries a man without hemophilia. What is the probability of their daughter having hemophilia? Their son? If they have four sons, that all will be affected?
2. Pseudohypertrophic muscular dystrophy is an inherited disorder that causes gradual deterioration of the muscles. It is seen almost exclusively in boys born to unaffected parents and usually results in death in the early teens. Is this disorder caused by a dominant or a recessive allele? Is its inheritance sex-linked or autosomal? How do you know? Explain why this disorder is almost never seen in girls.
3. A wild-type fruit fly (heterozygous for gray body color and normal wings) is mated with a black fly with vestigial wings. The offspring have the following phenotypic distribution: wild-type, 778; black vestigial, 785; black normal, 158; gray vestigial, 162. What is the recombination frequency between these genes for body color and wing size? Is this consistent with the results of the experiment in Figure 15.9? Draw the chromosomes in the wild-type and black parents.
4. A planet is inhabited by creatures that reproduce with the same hereditary patterns seen in humans. Three phenotypic characters are height (T = tall, t = dwarf), head appendages (A = antennae, a = no antennae), and nose shape (S = upturned snout, s = downturned snout). Since the creatures are not "intelligent," Earth scientists are able to do some controlled breeding experiments using various heterozygotes in testcrosses. For tall heterozygotes with antennae, the offspring are tall/antennae, 46; dwarf/antennae, 7; dwarf/no antennae, 42; tall/no antennae, 5. For heterozygotes with antennae and an upturned snout, the offspring are antennae/upturned snout, 47; antennae/downturned snout, 2; no antennae/downturned snout, 48; no antennae/upturned snout, 3. Calculate the recombination frequencies for both experiments.

Levels 3-4: Applying/Analyzing

5. Scientists do a testcross of the creatures from problem 4 using heterozygotes for height and nose shape. The offspring are tall/upturned snout, 41; dwarf/upturned snout, 8; dwarf/downturned snout, 43; tall/downturned snout, 8. Calculate the recombination frequency from these data, and then use your answer from problem 4 to determine the correct order of the three linked genes.
6. A wild-type fruit fly (heterozygous for gray body color and red eyes) is mated with a black fruit fly with purple eyes. The offspring are wild-type, 721; black/purple, 751; gray/purple, 49; black/red, 45. What is the recombination frequency between these genes for body color and eye color? Using information from problem 3, what fruit flies (genotypes and phenotypes) would you mate to determine the order of the body color, wing size, and eye color genes on the chromosome?
7. Assume that genes A and B are 50 map units apart on the same chromosome. An animal heterozygous at both loci is crossed with one that is homozygous recessive at both loci. What percentage of the offspring will show recombinant phenotypes resulting from crossovers? Without knowing these genes are on the same chromosome, how would you interpret the results of this cross?
8. Two genes of a flower, one controlling blue (B) versus white (b) petals and the other controlling round (R) versus oval (r) stamens, are linked and are 10 map units apart. You cross a homozygous blue/oval plant with a homozygous white/round plant. The resulting F_1 progeny are crossed with homozygous white/oval plants, and 1,000 offspring plants are obtained. How many F_2 plants of each of the four phenotypes do you expect?
9. You design *Drosophila* crosses to provide recombination data for gene a, which is located on the chromosome shown in Figure 15.12. Gene a has recombination frequencies of 14% with the vestigial wing locus and 26% with the brown eye locus. Approximately where is a located along the chromosome relative to these two loci?

Levels 5-6: Evaluating/Creating

10. Banana plants, which are triploid, are seedless and therefore sterile. Thinking about meiosis, propose a possible explanation.
11. **EVOLUTION CONNECTION** Crossing over is thought to be evolutionarily advantageous because it continually shuffles genetic alleles into novel combinations. It was thought that Y-linked genes might degenerate because they lack homologous genes on the X chromosome with which to pair up prior to crossing over. However, when the Y chromosome was sequenced, scientists discovered eight large regions that were internally homologous to each other, and several genes are duplicates. (Y chromosome researcher David Page has called it a "hall of mirrors.") How might this be beneficial?
12. **SCIENTIFIC INQUIRY DRAW IT** Assume you are mapping genes *A, B, C,* and *D* in *Drosophila*. You know that these genes are linked on the same chromosome, and you determine the recombination frequencies between each pair of genes to be as follows: *A* and *B*, 8%; *A* and *C*, 28%; *A* and *D*, 25%; *B* and *C*, 20%; *B* and *D*, 33%.
    (a) Describe how you determined the recombination frequency for each pair of genes.
    (b) Draw a chromosome map based on your data.
13. **WRITE ABOUT A THEME: INFORMATION** The continuity of life is based on heritable information in the form of DNA. In a short essay (100–150 words), relate the structure and behavior of chromosomes to inheritance in both asexually and sexually reproducing species.
14. **SYNTHESIZE YOUR KNOWLEDGE**



Butterflies have an X-Y sex determination system that is different from that of flies or humans. Female butterflies may be either XY or X0, while butterflies with two or more X chromosomes are males. This photograph shows a tiger swallowtail *gynandromorph*, which is half male (left side) and half female (right side). Given that the first division of the zygote divides the embryo into the future right and left halves of the butterfly, propose a hypothesis that explains how nondisjunction during the first mitosis might have produced this unusual-looking butterfly.

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
How does smoking affect the Y chromosome?
Go to "Does Smoking Make a Man Less of a Man?"
at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.

16 The Molecular Basis of Inheritance

KEY CONCEPTS

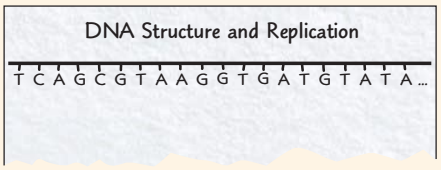
16.1 **DNA is the genetic material** *p. 315*

16.2 **Many proteins work together in DNA replication and repair** *p. 320*

16.3 **A chromosome consists of a DNA molecule packed together with proteins** *p. 330*

Study Tip

Draw DNA replication: Make up a single-stranded DNA sequence that is 40 nucleotides in length. As you go through Concepts 16.1 and 16.2, use your sequence to draw a double-stranded DNA molecule, labeling the ends. Then diagram the process of replication of your DNA molecule.



Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 16
- Figure 16.15 Walkthrough: Addition of a Nucleotide to a DNA Strand
- BioFlix® Animation: DNA Replication

For Instructors to Assign (in Item Library)
- Activity: The Hershey-Chase Experiment
- Tutorial: Visualizing DNA

Ready-to-Go Teaching Module (in Instructor Resources)
- DNA Replication (Concept 16.2)

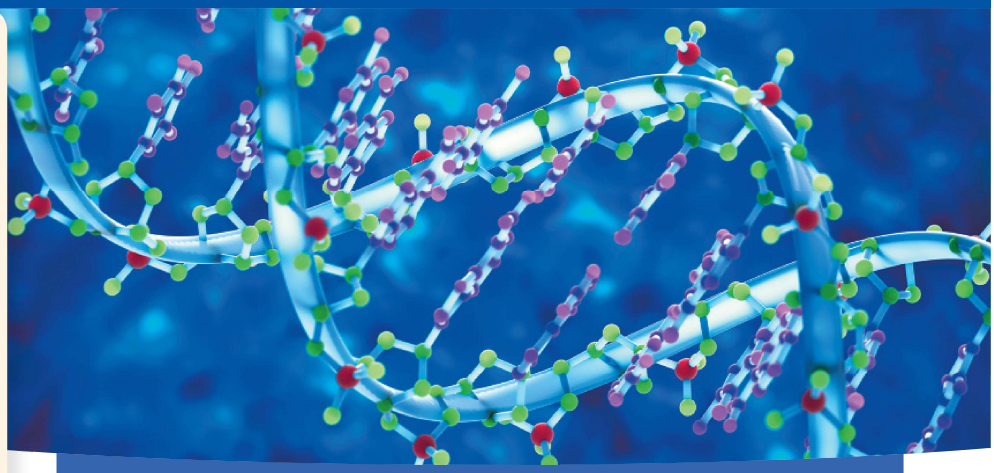
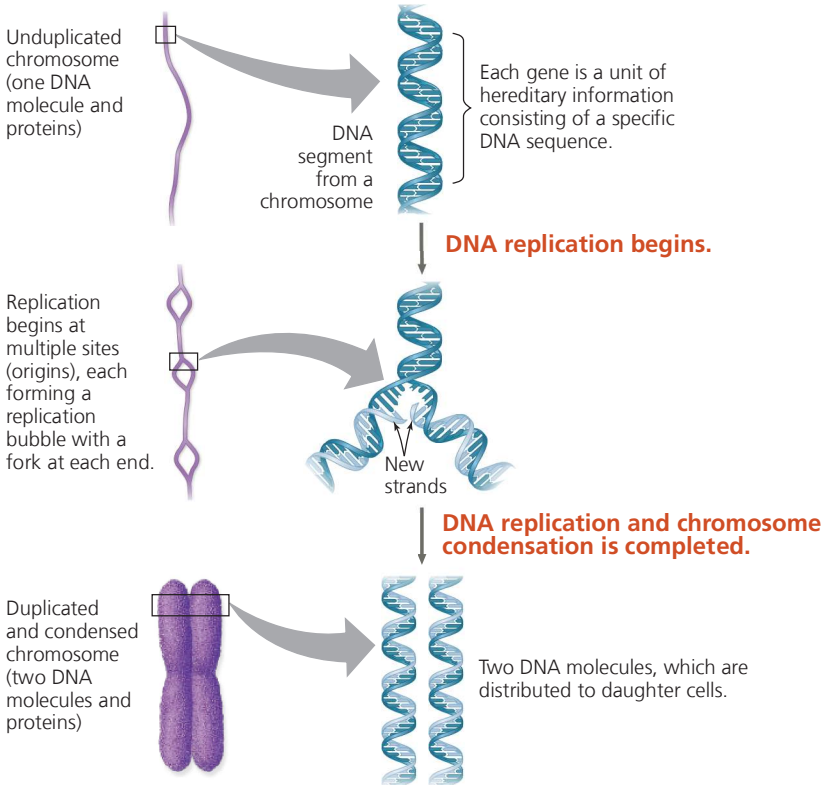


Figure 16.1 The elegant double-helical structure of deoxyribonucleic acid (DNA) shook the scientific world when it was proposed in April 1953 by James Watson and Francis Crick. The DNA you inherited from your parents contains all your genes—your genetic information.

How does DNA replication transmit genetic information?

DNA replication allows genetic information to be inherited from a parent cell to daughter cells (by mitosis) and from generation to generation (starting with meiosis).



CONCEPT 16.1

DNA is the genetic material

Today, even schoolchildren have heard of DNA, and scientists routinely manipulate DNA in the laboratory. Early in the 20th century, however, identifying the molecules of inheritance posed a major challenge to biologists.

The Search for the Genetic Material: *Scientific Inquiry*

Once T. H. Morgan's group showed that genes exist as parts of chromosomes (described in Concept 15.1), the two chemical components of chromosomes—DNA and protein—emerged as the leading candidates for the genetic material. Until the 1940s, the case for proteins seemed stronger: Biochemists had identified proteins as a class of macromolecules with great heterogeneity and specificity of function, essential requirements for the hereditary material. Moreover, little was known about nucleic acids, whose physical and chemical properties seemed far too uniform to account for the multitude of specific inherited traits exhibited by every organism. This view gradually changed as the role of DNA in heredity was worked out in studies of bacteria and the viruses that infect them, systems far simpler than fruit flies or humans. Let's trace the search for the genetic material as a case study in scientific inquiry.

Evidence That DNA Can Transform Bacteria

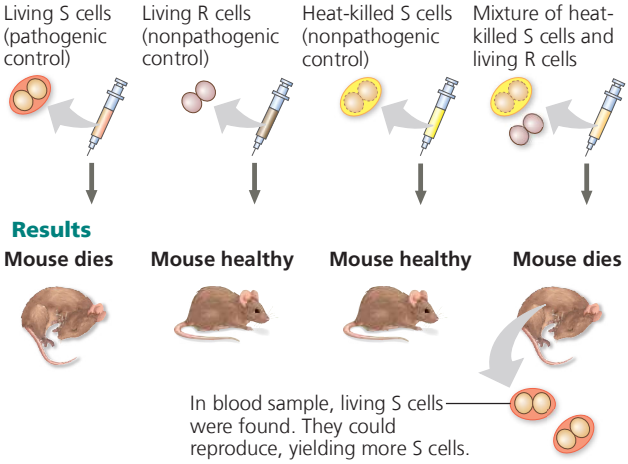
In 1928, a British medical officer named Frederick Griffith was trying to develop a vaccine against pneumonia. He was studying *Streptococcus pneumoniae*, a bacterium that causes pneumonia in mammals. Griffith had two strains (varieties) of the bacterium, one pathogenic (disease-causing) and one nonpathogenic (harmless). He was surprised to find that when he killed the pathogenic bacteria with heat and then mixed the cell remains with living bacteria of the nonpathogenic strain, some of the living cells became pathogenic **(Figure 16.2)**. Furthermore, this newly acquired trait of pathogenicity was inherited by all the descendants of the transformed bacteria. Apparently, some chemical component of the dead pathogenic cells caused this heritable change, although the identity of the substance was not known. Griffith called the phenomenon **transformation**, now defined as a change in genotype and phenotype due to the assimilation of external DNA by a cell. Later work by Oswald Avery, Maclyn McCarty, and Colin MacLeod identified the transforming substance as DNA.

Scientists remained skeptical, however, since many still viewed proteins as better candidates for the genetic material. Also, many biologists were not convinced that bacterial genes would be similar in composition and function to those of more complex organisms. But the major reason for the continued doubt was that so little was known about DNA.

▼ **Figure 16.2** Inquiry

Can a genetic trait be transferred between different bacterial strains?

Experiment Frederick Griffith studied two strains of the bacterium *Streptococcus pneumoniae*. The S (smooth) strain can cause pneumonia in mice; it is pathogenic because the cells have an outer capsule that protects them from an animal's immune system. Cells of the R (rough) strain lack a capsule and are nonpathogenic. To test for the trait of pathogenicity, Griffith injected mice with the two strains:



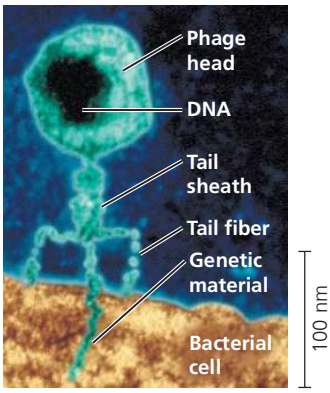
Conclusion The living R bacteria had been transformed into pathogenic S bacteria by an unknown, heritable substance from the dead S cells that enabled the R cells to make capsules.

Data from F. Griffith, The significance of pneumococcal types, *Journal of Hygiene* 27:113–159 (1928).

WHAT IF? *How did this experiment rule out the possibility that the R cells simply used the dead S cells' capsules to become pathogenic?*

Evidence That Viral DNA Can Program Cells

Additional evidence that DNA was the genetic material came from studies of viruses that infect bacteria **(Figure 16.3)**. These viruses are called **bacteriophages** (meaning "bacteria-eaters"), or **phages** for short. Viruses are much simpler than cells. A **virus** is little more than DNA (or sometimes RNA) enclosed by a protective coat, which is often simply protein. To produce more viruses, a virus must infect a cell and take over the cell's metabolic machinery.



▶ **Figure 16.3 A virus infecting a bacterial cell.** A phage called T2 attaches to a host cell and injects its genetic material through the plasma membrane, while the head and tail parts remain on the outer bacterial surface (colorized TEM).

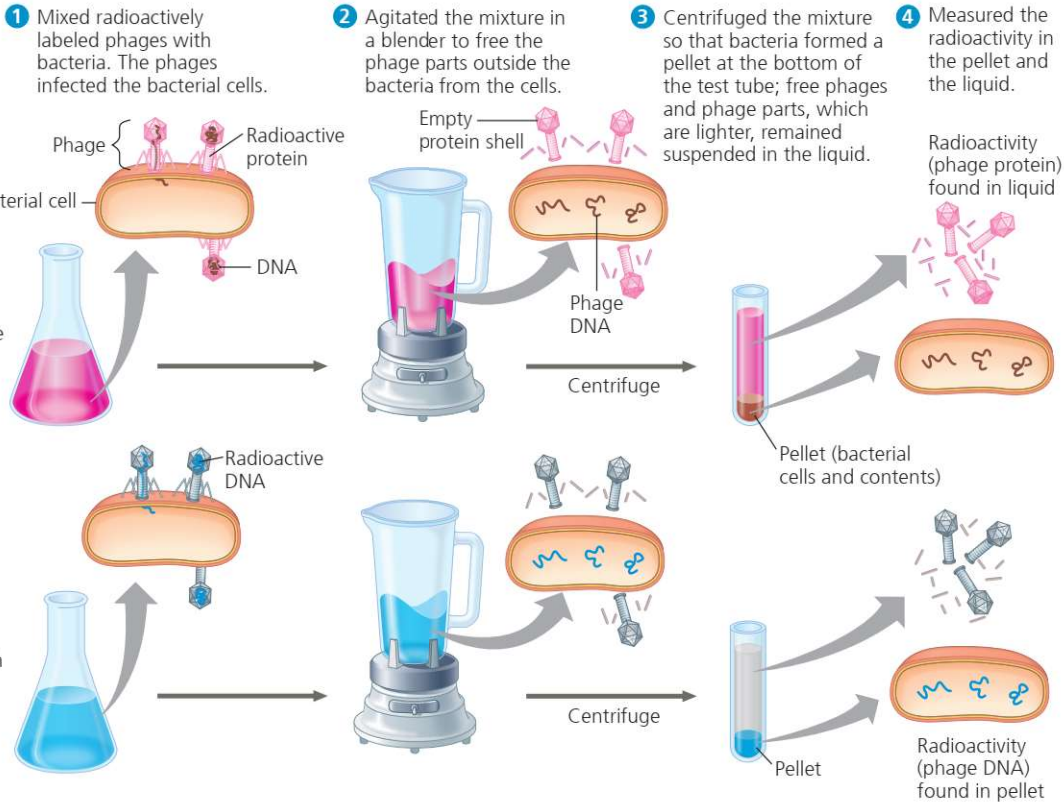
Phages have been widely used as tools by researchers in molecular genetics. In 1952, Alfred Hershey and Martha Chase performed experiments showing that DNA is the genetic material of a phage known as T2. This is one of many phages that infect *Escherichia coli* (*E. coli*), a bacterium that normally lives in the intestines of mammals and is a model organism for molecular biologists. At that time, biologists already knew that T2, like many other phages, was composed almost entirely of DNA and protein. They also knew that the T2 phage could quickly turn an *E. coli* cell into a T2-producing factory that released many copies of new phages when the cell ruptured. Somehow, T2 could reprogram its host cell to produce viruses. But which viral component—protein or DNA—was responsible?

Hershey and Chase answered this question by devising an experiment showing that only one of the two components of T2 actually enters the *E. coli* cell during infection **(Figure 16.4)**. In their experiment, they used a radioactive isotope of sulfur to tag protein in one batch of T2 and a radioactive isotope of phosphorus to tag DNA in a second batch. Because protein, but not DNA, contains sulfur, radioactive sulfur atoms were incorporated only into the protein of the phage. In a similar way, the atoms of radioactive phosphorus labeled only the DNA, not the protein, because nearly all the phage's phosphorus is in its DNA. In the experiment, separate samples of nonradioactive *E. coli* cells were infected separately with the protein-labeled and DNA-labeled batches of T2. The researchers then tested the two samples

▼ Figure 16.4 Inquiry

Is protein or DNA the genetic material of phage T2?

Experiment Alfred Hershey and Martha Chase used radioactive sulfur and phosphorus to trace the fates of protein and DNA, respectively, of T2 phages that infected bacterial cells. They wanted to see which of these molecules entered the cells and could reprogram them to make more phages.



Results When proteins were labeled (batch 1), radioactivity remained outside the cells, but when DNA was labeled (batch 2), radioactivity was found inside the cells. Cells containing radioactive phage DNA released new phages with some radioactive phosphorus.

Conclusion Phage DNA entered bacterial cells, but phage proteins did not. Hershey and Chase concluded that DNA, not protein, functions as the genetic material of phage T2.

Data from A. D. Hershey and M. Chase, Independent functions of viral protein and nucleic acid in growth of bacteriophage, *Journal of General Physiology* 36:39–56 (1952).

WHAT IF? *How would the results have differed if proteins carried the genetic information?*

Mastering Biology **Animation: The Hershey-Chase Experiment**

shortly after the onset of infection to see which type of radioactively labeled molecule—protein or DNA—had entered the bacterial cells and would therefore be capable of reprogramming them.

Hershey and Chase found that the phage DNA entered the host cells, but the phage protein did not. Moreover, when these bacteria were returned to a culture medium and the infection ran its course, the *E. coli* released phages that contained some radioactive phosphorus. This result further showed that the DNA inside the cell played an ongoing role during the infection process. They concluded that the DNA injected by the phage must be the molecule carrying the genetic information that makes the cells produce new viral DNA and proteins. The Hershey-Chase experiment was a landmark study because it provided powerful evidence that nucleic acids, rather than proteins, are the hereditary material, at least for certain viruses.

Additional Evidence That DNA Is the Genetic Material

Further evidence that DNA is the genetic material came from the laboratory of biochemist Erwin Chargaff. DNA was known to be a polymer of nucleotides, each having three components: a nitrogenous (nitrogen-containing) base, a pentose sugar called deoxyribose, and a phosphate group **(Figure 16.5)**. The base can be adenine (A), thymine (T), guanine (G), or cytosine (C). Chargaff analyzed the base composition of DNA from a number of different organisms. In 1950, he reported that the base composition of DNA varies from one species to another. For example, he found that 32.8% of sea urchin DNA nucleotides have the base A, whereas only 24.7% of those from the bacterium *E. coli* have an A. This evidence of molecular diversity among species, which most scientists had presumed to be absent from DNA, made DNA a more credible candidate for the genetic material.

Chargaff also noticed a peculiar regularity in the ratios of nucleotide bases. In the DNA of each species he studied, the number of adenines approximately equaled the number of thymines, and the number of guanines approximately equaled the number of cytosines. In sea urchin DNA, for example, the four bases are present in these percentages: A = 32.8% and T = 32.1%; G = 17.7% and C = 17.3%. (The percentages are not exactly the same because of limitations in Chargaff's techniques.)

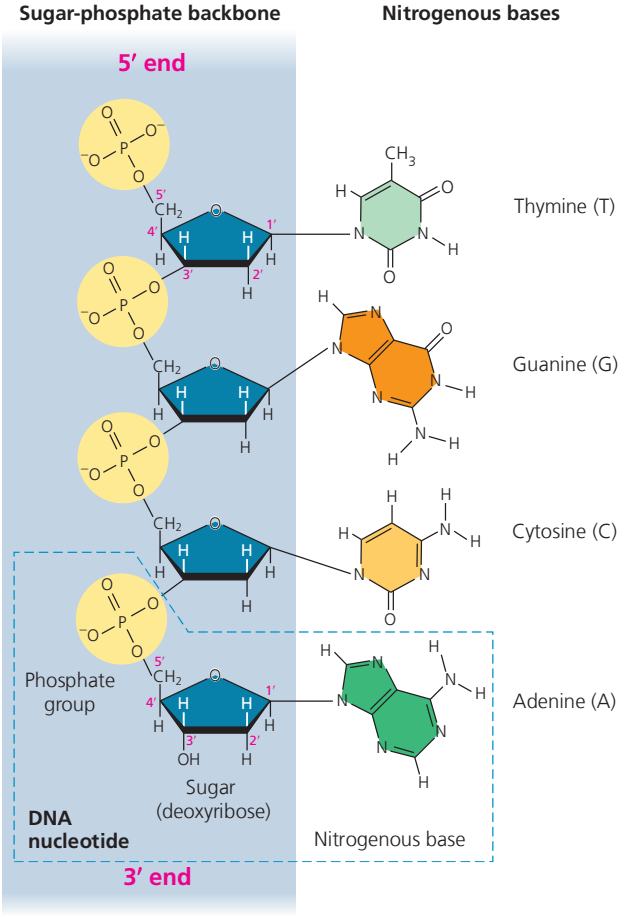
These two findings became known as *Chargaff's rules*: (1) DNA base composition varies between species, and (2) for each species, the percentages of A and T bases are roughly equal, as are those of G and C bases. In the **Scientific Skills Exercise**, you can use Chargaff's rules to predict percentages of nucleotide bases. The basis for these rules remained unexplained until the discovery of the double helix.

Mastering Biology HHMI Animation: Chargaff's Ratio

Building a Structural Model of DNA

Once most biologists were convinced that DNA was the genetic material, the challenge was to determine how the structure

▼ Figure 16.5 The structure of a DNA strand. Each DNA nucleotide monomer consists of the sugar deoxyribose (blue) attached to both a nitrogenous base (A, T, G, or C), and a phosphate group (yellow). The phosphate group of one nucleotide is attached to the sugar of the next by a covalent bond, forming a "backbone" of alternating phosphates and sugars from which the bases project. A polynucleotide strand has directionality, from the 5′ end (with the phosphate group) to the 3′ end (with the —OH group of the sugar). 5′and 3′ refer to the numbers assigned to the carbons in the sugar ring (see magenta numbers).



Mastering Biology Animation: DNA and RNA Structure

of DNA could account for its role in inheritance. By the early 1950s, the arrangement of covalent bonds in a nucleic acid polymer was well established (see Figure 16.5), and researchers focused on discovering the three-dimensional structure of DNA. Among the scientists working on the problem were Linus Pauling, at the California Institute of Technology, and Maurice Wilkins and Rosalind Franklin, at King's College in London. First to come up with the complete answer, however, were two scientists who were relatively unknown at the time—the American James Watson and the Englishman Francis Crick.

The collaborative work that solved the puzzle of DNA structure began soon after Watson went to Cambridge University, where Crick was studying protein structure with a technique called X-ray crystallography (see Figure 5.21).

Scientific Skills Exercise

Working with Data in a Table

Given the Percentage Composition of One Nucleotide in a Genome, Can We Predict the Percentages of the Other Three Nucleotides? Even before the structure of DNA was elucidated, Erwin Chargaff and his co-workers noticed a pattern in the base composition of nucleotides from different species: The percentage of adenine (A) bases roughly equaled that of thymine (T) bases, and the percentage of cytosine (C) bases roughly equaled that of guanine (G) bases. Further, the percentage of each pair (A–T or C–G) varied from species to species. We now know that the 1:1 A/T and C/G ratios are due to complementary base pairing between A and T and between C and G in the DNA double helix, and differences between species are due to the unique sequences of bases along a DNA strand. In this exercise, you will apply Chargaff's rules to predict the composition of bases in a genome.

How the Experiments Were Done In Chargaff's experiments, DNA was extracted from the given species, hydrolyzed to break apart the nucleotides, and then analyzed chemically. These studies provided approximate values for each type of nucleotide. (Today, whole-genome sequencing allows base composition analysis to be done more precisely directly from the sequence data.)

Data from the Experiments Tables are useful for organizing sets of data representing a common set of values (here, percentages of A, G, C, and T) for a number of different samples (in this case, from different species). You can apply the patterns that you see in the known data to predict unknown values. In the table, complete base distribution data are given for sea urchin DNA and salmon DNA; you will use Chargaff's rules to fill in the rest of the table with predicted values.

Source of DNA	Base Percentage			
	Adenine	Guanine	Cytosine	Thymine
Sea urchin	32.8	17.7	17.3	32.1
Salmon	29.7	20.8	20.4	29.1
Wheat	28.1	21.8	22.7	
E. coli	24.7	26.0		
Human	30.4			30.1
Ox	29.0			
Average %				

Data from several papers by Chargaff: for example, E. Chargaff et al., Composition of the desoxypentose nucleic acids of four genera of sea-urchin, *Journal of Biological Chemistry* 195: 155–160 (1952).



▶ **Sea urchin**

INTERPRET THE DATA

1. Explain how the sea urchin and salmon data demonstrate both of Chargaff's rules.
2. Using Chargaff's rules, fill in the table with your predictions of the missing percentages of bases, starting with the wheat genome and proceeding through *E. coli*, human, and ox. Show how you arrived at your answers.
3. If Chargaff's rule—that the amount of A equals the amount of T and the amount of C equals the amount of G—is valid, then hypothetically we could extrapolate this to the combined DNA of all species on Earth (like one huge Earth genome). To see whether the data in the table support this hypothesis, calculate the average percentage for each base in your completed table by averaging the values in each column. Does Chargaff's equivalence rule still hold true?

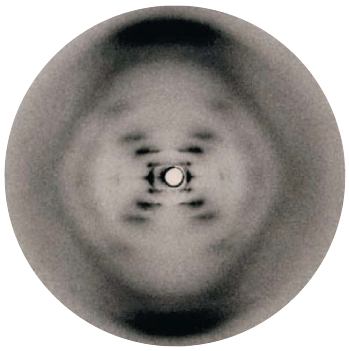
Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology.**

While visiting the laboratory of Maurice Wilkins, Watson saw an X-ray diffraction image of DNA produced by Wilkins's accomplished colleague Rosalind Franklin **(Figure 16.6)**. Images produced by X-ray crystallography are not actually pictures of molecules. The spots in the image were produced by X-rays that were diffracted (deflected) as they passed through aligned fibers of purified DNA. Watson was familiar with the type of X-ray diffraction pattern of helical molecules, and seeing the photo that Wilkins showed him confirmed that DNA was helical in shape. The photo also added to earlier data obtained by Franklin and others suggesting the width of the helix and the spacing of the nitrogenous bases along it. The pattern in this photo implied that the helix was made up of two strands, contrary to a three-stranded model proposed by Linus Pauling a short time earlier. The presence of two strands accounts for the now-familiar term **double helix**. DNA is shown in some of its many different representations in **Figure 16.7**.

▼ **Figure 16.6 Rosalind Franklin and her X-ray diffraction photo of DNA.**



(a) Rosalind Franklin



(b) Franklin's X-ray diffraction photograph of DNA

Watson and Crick began building models of a double helix that would conform to the X-ray measurements and what was then known about the chemistry of DNA, including Chargaff's rule of base equivalences. Having also read an unpublished annual report summarizing Franklin's work, they knew she had concluded that the sugar-phosphate backbones were on the outside of the DNA molecule, contrary to their working model. Franklin's arrangement was appealing because it put the negatively charged phosphate groups facing the aqueous surroundings, while the relatively hydrophobic nitrogenous bases were hidden in the interior. Watson constructed such a model, in which the two sugar-phosphate backbones are **antiparallel**—that is, their subunits run in opposite directions (see Figure 16.7). You can imagine the overall arrangement as a rope ladder with rigid rungs. The side ropes represent

▼ Figure 16.7 VISUALIZING DNA

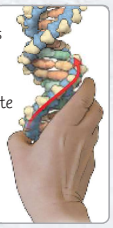
DNA can be illustrated in many ways, but all diagrams represent the same basic structure. The level of detail shown depends on the process or the type of information being conveyed.

Structural Images

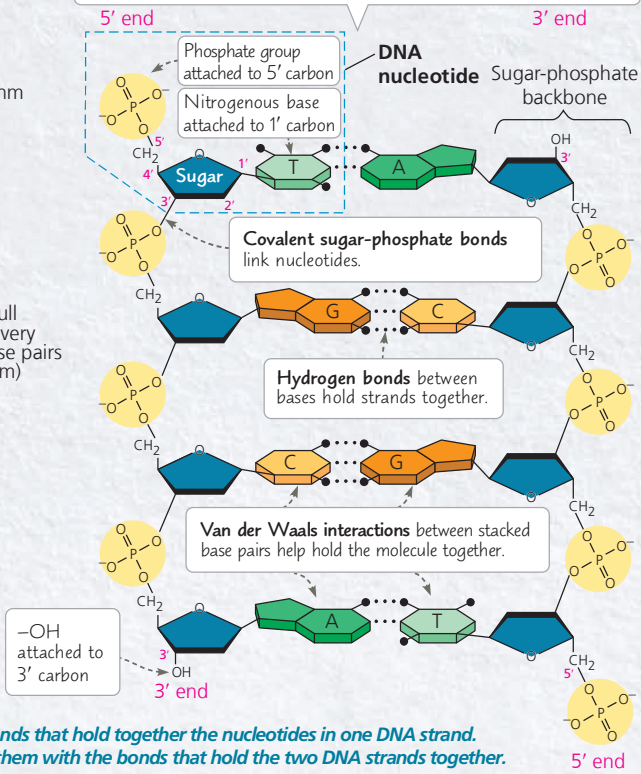
The space-filling model on the left shows the 3-D shape of the DNA double helix; the diagram on the right shows chemical details of DNA's structure. In both images, phosphate groups are yellow, deoxyribose sugars are blue, and nitrogenous bases are shades of green and orange.



The DNA double helix is "right-handed." Use your right hand as shown to follow the sugar-phosphate backbone up the helix and around to the back. (It won't work with your left hand.)



Here, the two DNA strands are shown untwisted so it's easier to see the chemical details. Note that the strands are **antiparallel**—they are oriented in opposite directions, like the lanes of a divided street.



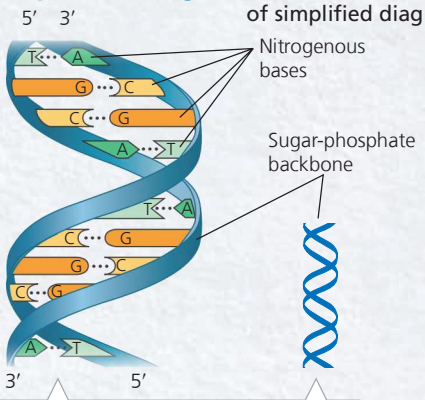
Mastering Biology Animation: DNA Double Helix

1. ***Describe the bonds that hold together the nucleotides in one DNA strand. Then compare them with the bonds that hold the two DNA strands together.***

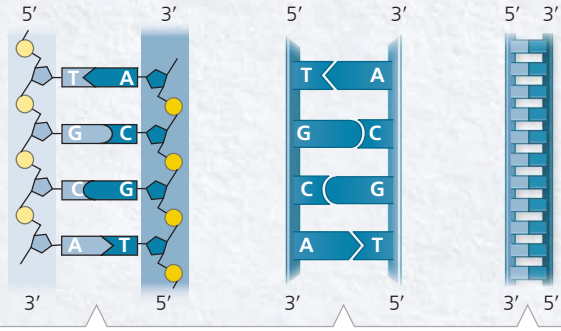
2. ***How do the two ends of a DNA strand differ in structure?***

Simplified Images

When molecular detail is not necessary, DNA is portrayed in a range of simplified diagrams, depending on the focus of the figure.



The "ribbons" in these simplified double helix diagrams represent the sugar-phosphate backbones; these models emphasize the 3-D nature of DNA.



These flattened "ladder style" diagrams of DNA depict the sugar-phosphate backbones like the side rails of a ladder, with the base pairs as rungs. Light blue is used to indicate the more recently synthesized strand.

Sometimes the double-stranded DNA molecule is shown simply as two straight lines.

3. ***Compare the information conveyed in the three ladder diagrams.***

DNA Sequences

Genetic information is carried in DNA as a linear sequence of nucleotides that may be transcribed into mRNA and translated into a polypeptide. When focusing on the DNA sequence, each nucleotide can be represented simply as the letter of its base: A, T, C, or G.

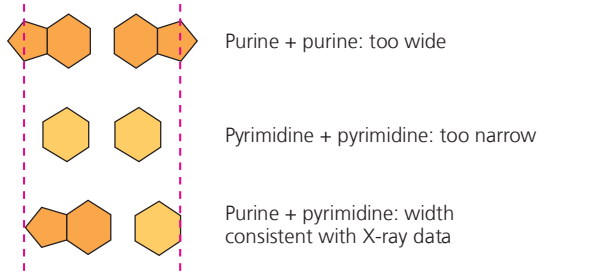
3′ - A C G T A A G C G G T T A A T - 5′
5′ - T G C A T T C G C C A A T T A - 3′

Instructors: Additional questions related to this Visualizing Figure can be assigned in **Mastering Biology**.

the sugar-phosphate backbones, and the rungs represent pairs of nitrogenous bases. Now imagine twisting the ladder to form a helix. Franklin's X-ray data indicated that the helix makes one full turn every 3.4 nm along its length. With the bases stacked just 0.34 nm apart, there are ten layers of base pairs, or rungs of the ladder, in each full turn of the helix.

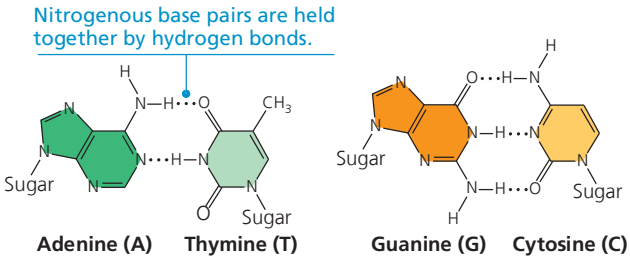
The nitrogenous bases of the double helix are paired in specific combinations: adenine (A) with thymine (T), and guanine (G) with cytosine (C). It was mainly by trial and error that Watson and Crick arrived at this key feature of DNA. At first, Watson imagined that the bases paired like with like—for example, A with A and C with C. But this model did not fit the X-ray data, which suggested that the double helix had a uniform diameter. Why is this requirement inconsistent with like-with-like pairing of bases? Adenine and guanine are purines, nitrogenous bases with two organic rings, while cytosine and thymine are nitrogenous bases called pyrimidines, which have a single ring. Pairing a purine with a pyrimidine is the only combination that results in a uniform diameter for the double helix **(Figure 16.8)**.

▼ Figure 16.8 Possible base pairings in the DNA double helix.



Watson and Crick reasoned that there must be additional specificity of pairing dictated by the structure of the bases. Each base has chemical side groups that can form hydrogen bonds with its appropriate partner: Adenine can form two hydrogen bonds with thymine and only thymine; guanine forms three hydrogen bonds with cytosine and only cytosine. In shorthand, A pairs with T, and G pairs with C **(Figure 16.9)**.

▼ Figure 16.9 Base pairing in DNA.



The Watson-Crick model took into account Chargaff's ratios and ultimately explained them. Wherever one strand of a DNA molecule has an A, the partner strand has a T. Similarly, a G in one strand is always paired with a C in the complementary strand. Therefore, in the DNA of any organism, the amount of adenine equals the amount of thymine, and the amount of guanine equals the amount of cytosine. (Modern DNA-sequencing techniques have confirmed that the amounts are exactly equal.) Although the base-pairing rules dictate the combinations of nitrogenous bases that form the "rungs" of the double helix, they do not restrict the sequence of nucleotides *along* each DNA strand. The linear sequence of the four bases can be varied in countless ways, and each gene has a unique base sequence.

In April 1953, Watson and Crick surprised the scientific world with a succinct, one-page paper that reported their molecular model for DNA: the double helix, which has since become an icon of molecular biology. Watson and Crick, along with Maurice Wilkins, were awarded the Nobel Prize in 1962 for this work. (Sadly, Rosalind Franklin had died at the age of 37 in 1958 and was thus ineligible for the prize.) The beauty of the double helix model was that the structure of DNA suggested the basic mechanism of its replication.

Mastering Biology HHMI Video: Great Discoveries in Science: The Double Helix

CONCEPT CHECK 16.1

1. Given a polynucleotide sequence such as GAATTC, explain what further information you would need in order to identify which is the 5′ end. (See Figure 16.5.)
2. **VISUAL SKILLS** While trying to develop a vaccine for *S. pneumonia*, Griffith was surprised to discover the phenomenon of bacterial transformation. Based on the results in the second and third panels of Figure 16.2, what result was he expecting in the fourth panel? Explain.

For suggested answers, see Appendix A.

CONCEPT 16.2

Many proteins work together in DNA replication and repair

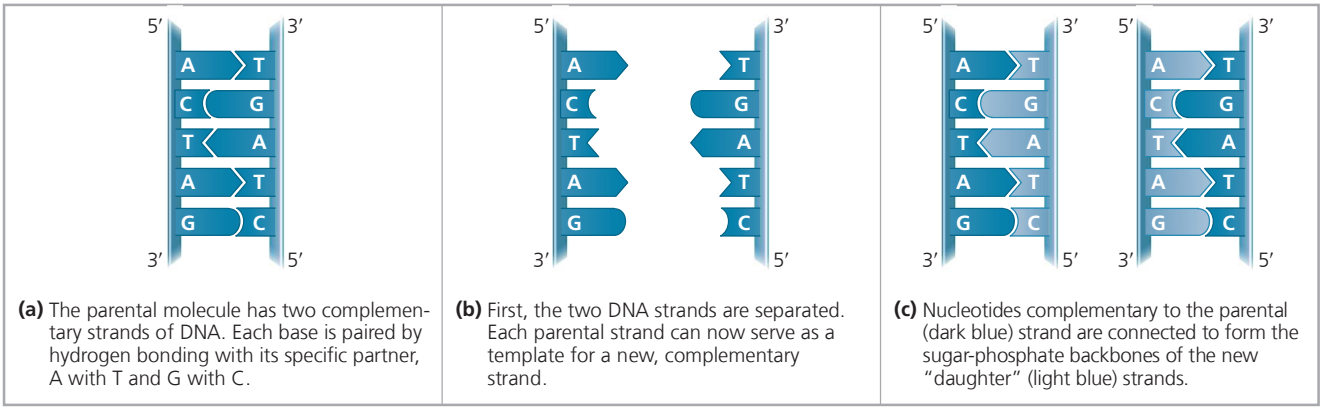
Hereditary information in DNA directs the development of your biochemical, anatomical, physiological, and, to some extent, behavioral traits. Your resemblance to your parents has its basis in the accurate replication of DNA prior to meiosis and therefore its transmission from your parents' generation to yours. Replication prior to mitosis ensures faithful transmission of genetic information from a parent cell to two daughter cells.

Of all nature's molecules, nucleic acids are unique in their ability to dictate their own replication from monomers. The relationship between structure and function is evident in the double helix: The specific complementary pairing of nitrogenous bases in DNA has a functional significance. Watson and Crick ended their classic paper with this statement: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."[1] In this section, you will learn about the basic principle of **DNA replication**, the copying of DNA, as well as some important details of the process.

[1] J. D. Watson and F. H. C. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acids, *Nature* 171:737–738 (1953).

▼ Figure 16.10 A model for DNA replication: the basic concept. In this simplified illustration, a short segment of DNA has been untwisted. Simple shapes symbolize the four kinds of bases. Dark blue represents DNA strands present in the parental molecule; light blue represents newly synthesized DNA.

Mastering Biology Animation: DNA Replication: An Overview



(a) The parental molecule has two complementary strands of DNA. Each base is paired by hydrogen bonding with its specific partner, A with T and G with C.

(b) First, the two DNA strands are separated. Each parental strand can now serve as a template for a new, complementary strand.

(c) Nucleotides complementary to the parental (dark blue) strand are connected to form the sugar-phosphate backbones of the new "daughter" (light blue) strands.

The Basic Principle: Base Pairing to a Template Strand

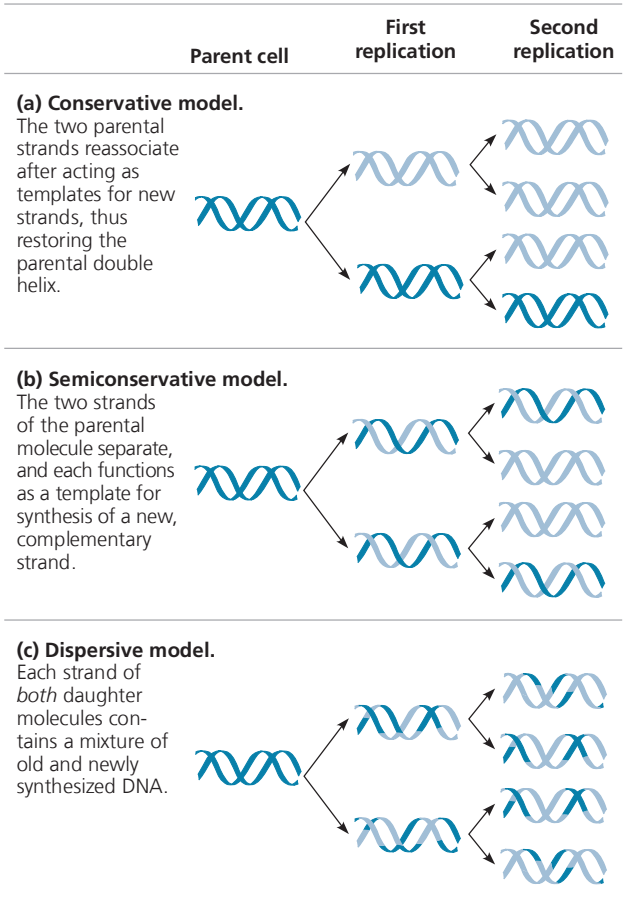
In a second paper, Watson and Crick stated their hypothesis for how DNA replicates: "Now our model for deoxyribonucleic acid is, in effect, a pair of templates, each of which is complementary to the other. We imagine that prior to duplication the hydrogen bonds are broken, and the two chains unwind and separate. Each chain then acts as a template for the formation on to itself of a new companion chain, so that eventually we shall have two pairs of chains, where we only had one before. Moreover, the sequence of the pairs of bases will have been duplicated exactly."[2]

Figure 16.10 illustrates the basic idea. If you cover a DNA strand in Figure 16.10a, its linear sequence of nucleotides is revealed by applying the base-pairing rules to the uncovered strand. The two strands are complementary; each stores the information necessary to reconstruct the other. When a cell copies a DNA molecule, each strand serves as a template for ordering nucleotides into a new, complementary strand. Nucleotides line up along the template strand according to the base-pairing rules and are linked to form the new strands. One double-stranded DNA molecule becomes two, each an exact replica of the "parental" molecule.

This model of DNA replication remained untested for several years following publication of the DNA structure. The necessary experiments were simple in concept but difficult to perform. Watson and Crick's model predicts that when a double helix replicates, each of the two daughter molecules will have one old strand, from the parental molecule, and one newly made strand. This **semiconservative model** can be distinguished from a *conservative model* of replication, in which the two parental strands somehow come back together after the process (that is, the parental molecule is conserved). In yet a third model, called the *dispersive model*, all four strands of DNA following replication have a mixture of old and new DNA **(Figure 16.11)**.

[2]J. D. Watson and F. H. C. Crick, Genetical implications of the structure of deoxyribonucleic acid, *Nature* 171:964–967 (1953).

▼ Figure 16.11 DNA replication: three alternative models. Each short segment of double helix symbolizes the DNA within a cell. Beginning with a parent cell, we follow the DNA for two more generations of cells—two rounds of DNA replication. Parental DNA is dark blue; newly made DNA is light blue.



Although mechanisms for conservative or dispersive DNA replication are not easy to devise, these models remained possibilities until they could be ruled out. After two years of preliminary work at the California Institute of Technology in the late 1950s, Matthew Meselson and Franklin Stahl devised a clever experiment that distinguished between the three models, described in **Figure 16.12**. Their results supported the semiconservative model of DNA replication, as predicted by Watson and Crick, and their experiment is widely recognized among biologists as a classic example of elegant design.

The basic principle of DNA replication is conceptually simple. However, the actual process involves some complicated biochemical gymnastics, as we will now see.

DNA Replication: *A Closer Look*

The bacterium *E. coli* has a single chromosome of about 4.6 million nucleotide pairs. In a favorable environment, an *E. coli* cell can copy all of this DNA and divide to form two genetically identical daughter cells in considerably less than an hour. Each of *your* somatic cells has 46 DNA molecules in its nucleus, one long double-helical molecule per chromosome. In all, that represents about 6 billion nucleotide pairs, or over 1,000 times more DNA than is found in most bacterial cells. If we were to print the one-letter symbols for these bases (A, G, C, and T) the size of the type you are now reading, the 6 billion nucleotide pairs of information in a diploid human cell would fill about 1,400 biology textbooks. Yet it takes one of your cells just a few hours to copy all of this DNA during S phase of interphase. This replication of an enormous amount of genetic information is achieved with very few errors—only about one per 10 billion nucleotides. The copying of DNA is remarkable in its speed and accuracy.

More than a dozen enzymes and other proteins participate in DNA replication. Much more is known about how this "replication machine" works in bacteria (such as *E. coli*) than in eukaryotes, and we will describe the basic steps of the process for *E. coli*, except where otherwise noted. What scientists have learned about eukaryotic DNA replication suggests, however, that most of the process is fundamentally similar for prokaryotes and eukaryotes.

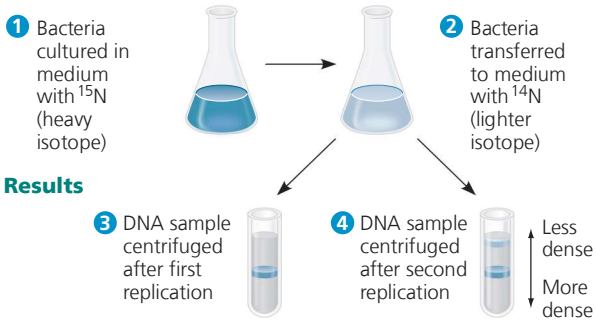
Getting Started

The replication of chromosomal DNA begins at particular sites called **origins of replication**, short stretches of DNA that have a specific sequence of nucleotides. The *E. coli* chromosome, like many other bacterial chromosomes, is circular and has a single origin. Proteins that initiate DNA replication recognize this sequence and attach to the DNA, separating the two strands and opening up a replication "bubble" **(Figure 16.13a)**. Replication of DNA then proceeds in both directions until the entire molecule is copied. In contrast to a bacterial chromosome, a eukaryotic chromosome may have hundreds or even a few thousand replication origins. Multiple replication bubbles form and eventually fuse, thus speeding up the copying of the very

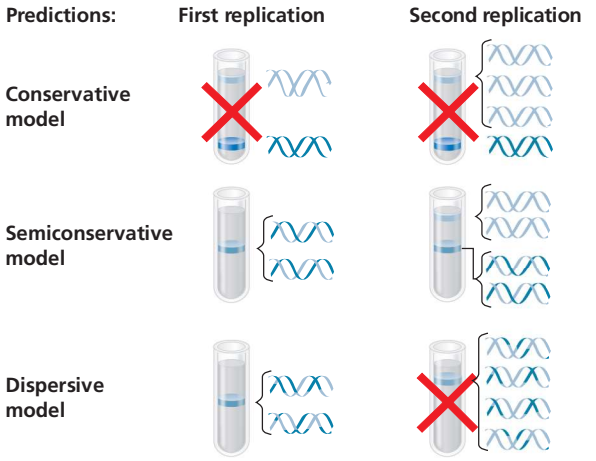
▼ Figure 16.12 Inquiry

Does DNA replication follow the conservative, semiconservative, or dispersive model?

Experiment Matthew Meselson and Franklin Stahl cultured *E. coli* for several generations in a medium containing nucleotide precursors labeled with a heavy isotope of nitrogen, ^{15}N. They then transferred the bacteria to a medium with only ^{14}N, a lighter isotope. They took one sample after the first DNA replication and another after the second replication. They extracted DNA from the bacteria in the samples and then centrifuged each DNA sample to separate DNA of different densities.



Conclusion Meselson and Stahl compared their results to those predicted by each of the three models in Figure 16.11, as shown below. The first replication in the ^{14}N medium produced a band of many molecules of hybrid (^{15}N-^{14}N) DNA. This result eliminated the conservative model. The second replication produced both light and hybrid DNA, a result that refuted the dispersive model and supported the semiconservative model. They therefore concluded that DNA replication is semiconservative.



Data from M. Meselson and F. W. Stahl, The replication of DNA in *Escherichia coli*, *Proceedings of the National Academy of Sciences USA* 44:671–682 (1958).

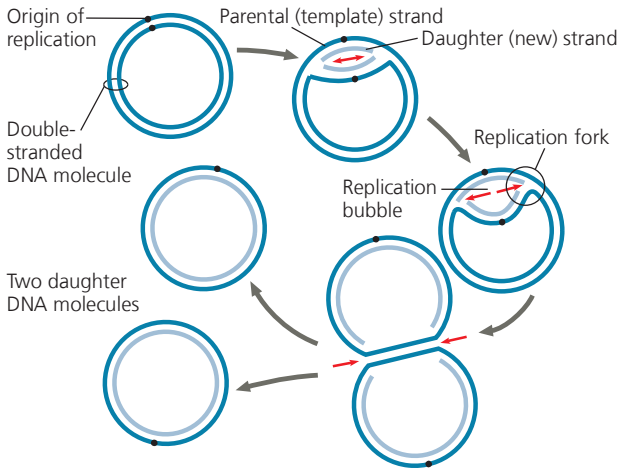
INQUIRY IN ACTION Read and analyze the original paper in *Inquiry in Action: Interpreting Scientific Papers.*

Instructors: A related Experimental Inquiry Tutorial can be assigned in **Mastering Biology**.

WHAT IF? *If Meselson and Stahl had first grown the cells in ^{14}N-containing medium and then moved them into ^{15}N-containing medium before taking samples, what would have been the result after each of the two replications?*

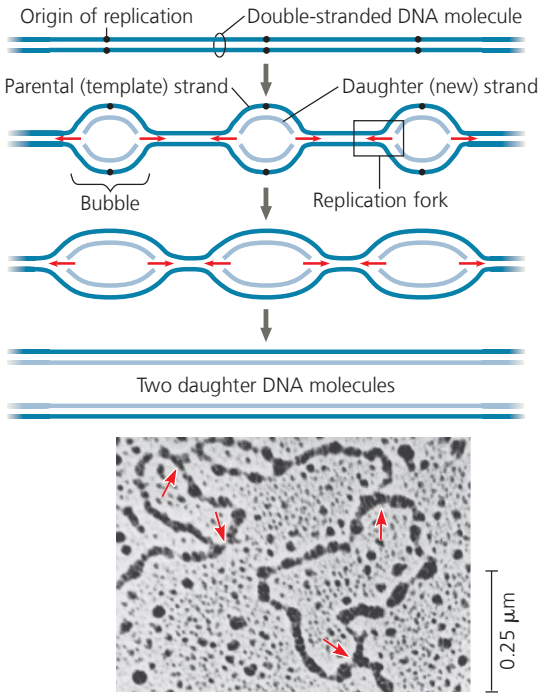
▼ Figure 16.13 Origins of replication in *E. coli* and eukaryotes. The red arrows indicate the movement of the replication forks and thus the overall directions of DNA replication within each bubble.

(a) Origin of replication in an *E. coli* cell



The circular chromosome of *E. coli* and other bacteria has only one origin of replication. The parental strands separate there, forming a replication bubble with two forks (red arrows). Replication proceeds in both directions until the forks meet on the other side, resulting in two daughter DNA molecules. The TEM shows a bacterial chromosome with a replication bubble.

(b) Origins of replication in a eukaryotic cell



In a linear chromosome of a eukaryote, replication bubbles form at many sites along the giant DNA molecule during S phase of interphase. The bubbles expand as replication proceeds in both directions (red arrows). Eventually, the bubbles fuse and synthesis of the daughter strands is complete. The TEM shows three replication bubbles along the DNA of a cultured Chinese hamster cell.

DRAW IT *In the TEM, add arrows in the forks of the third bubble.*

Mastering Biology BioFlix® Animation: The Replication Fork in *E. coli*

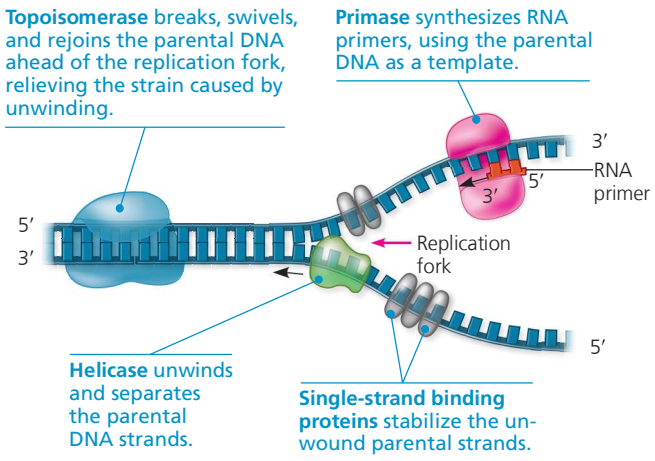
long DNA molecules **(Figure 16.13b)**. As in bacteria, eukaryotic DNA replication proceeds in both directions from each origin.

At each end of a replication bubble is a **replication fork**, a Y-shaped region where the parental strands of DNA are being unwound. Several kinds of proteins participate in the unwinding **(Figure 16.14)**. **Helicases** are enzymes that untwist the double helix at the replication forks, separating the two parental strands and making them available as template strands. After the parental strands separate, **single-strand binding proteins** bind to the unpaired DNA strands, keeping them from re-pairing. The untwisting of the double helix causes tighter twisting and strain ahead of the replication fork. **Topoisomerase** is an enzyme that helps relieve this strain by breaking, swiveling, and rejoining DNA strands.

Synthesizing a New DNA Strand

The unwound sections of parental DNA strands are now available to serve as templates for the synthesis of new complementary DNA strands. However, the enzymes that

▼ Figure 16.14 Some of the proteins involved in the initiation of DNA replication. The same proteins function at both replication forks in a replication bubble. For simplicity, only the left-hand fork is shown, and the DNA bases are drawn much larger in relation to the proteins than they are in reality.



synthesize DNA cannot *initiate* the synthesis of a polynucleotide; they can only add DNA nucleotides to the end of an already existing chain that is base-paired with the template strand. An initial nucleotide chain that can be used as a pre-existing chain is produced during DNA synthesis; this is actually a short stretch of RNA, not DNA. The RNA chain is called a **primer** and is synthesized by the enzyme **primase** (see Figure 16.14). Primase starts a complementary RNA chain with a single RNA nucleotide and adds RNA nucleotides one at a time, using the parental DNA strand as a template. The completed primer, generally five to ten nucleotides long, is thus base-paired to the template strand. The new DNA strand will start from the 3′ end of the RNA primer.

Enzymes called **DNA polymerases** catalyze the synthesis of new DNA by adding nucleotides to the 3′ end of a pre-existing chain. In *E. coli*, there are several DNA polymerases, but two of them appear to play the major roles in DNA replication: DNA polymerase III and DNA polymerase I. The situation in eukaryotes is more complicated, with at least 11 different DNA polymerases discovered so far, although the general principles are the same.

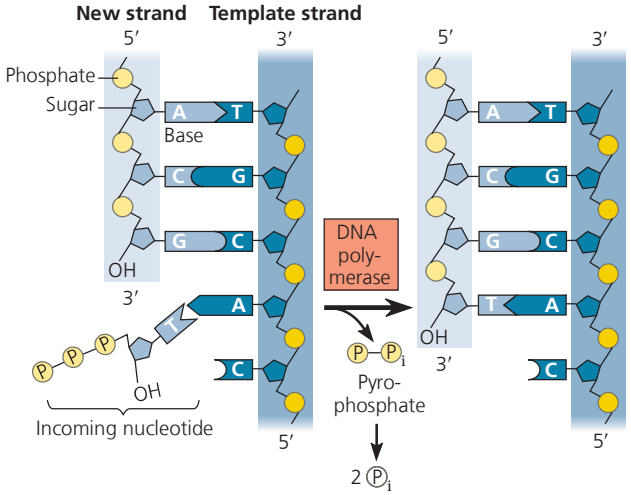
Most DNA polymerases require a primer and a DNA template strand, along which complementary DNA nucleotides are lined up, one by one. In *E. coli*, DNA polymerase III (abbreviated DNA pol III) adds a DNA nucleotide to the RNA primer and then continues adding DNA nucleotides, which are complementary to the parental DNA template strand, to the growing end of the new DNA strand. The rate of elongation is about 500 nucleotides per second in bacteria and 50 per second in human cells.

Each nucleotide to be added to a growing DNA strand consists of a sugar attached to a base and to three phosphate groups. You have already encountered such a molecule—ATP (adenosine triphosphate; see Figure 8.9). The only difference between the ATP of energy metabolism and dATP, the adenine nucleotide used to make DNA, is the sugar component, which is deoxyribose in the building block of DNA but ribose in ATP. Like ATP, the nucleotides used for DNA synthesis are chemically reactive, partly because their triphosphate tails have an unstable cluster of negative charge. DNA polymerase catalyzes the addition of each monomer to the growing end of a DNA strand by a condensation reaction in which two phosphate groups are lost as a molecule of pyrophosphate (Ⓟ—Ⓟ$_i$). Subsequent hydrolysis of the pyrophosphate to two molecules of inorganic phosphate (Ⓟ$_i$) is a coupled exergonic reaction that helps drive the polymerization reaction **(Figure 16.15)**.

Antiparallel Elongation

As we have noted previously, the two ends of a DNA strand are different, giving each strand directionality, like a one-way street (see Figure 16.5). In addition, the two strands of DNA in a double helix are antiparallel, meaning that they are oriented in opposite directions to each other, like the two sides

▼ Figure 16.15 Addition of a nucleotide to a DNA strand.
DNA polymerase catalyzes addition of a nucleotide to the 3′ end of a growing DNA strand, with the release of two phosphates.



DRAW IT *Circle the area where the new bond was made.*

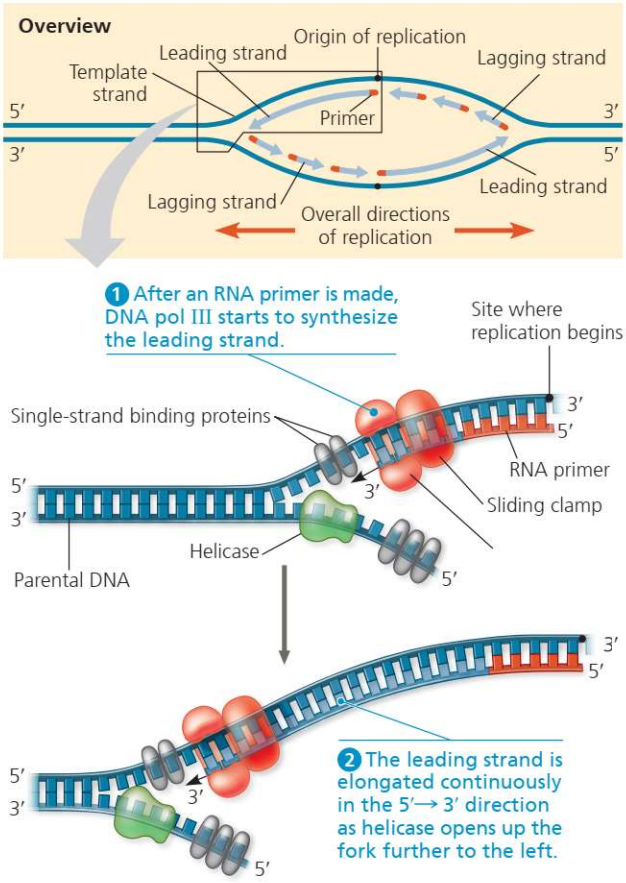
Mastering Biology Figure Walkthrough

of a divided street (see Figure 16.15). Therefore, the two new strands formed during DNA replication must also be antiparallel to their template strands.

The antiparallel arrangement of the double helix, together with a constraint on the function of DNA polymerases, has an important effect on how replication occurs. Because of their structure, DNA polymerases can add nucleotides only to the free 3′ end of a primer or growing DNA strand, never to the 5′ end (see Figure 16.15). Thus, a new DNA strand can elongate only in the 5′ → 3′ direction. With this in mind, let's examine one of the two replication forks in a bubble **(Figure 16.16)**. Along one template strand, DNA polymerase III can synthesize a complementary strand continuously by elongating the new DNA in the mandatory 5′ → 3′ direction. DNA pol III remains in the replication fork on that template strand and continuously adds nucleotides to the new complementary strand as the fork progresses. The DNA strand made by this mechanism is called the **leading strand**. Only one primer is required for DNA pol III to synthesize the entire leading strand (see Figure 16.16).

To elongate the other new strand of DNA in the mandatory 5′ → 3′ direction, DNA pol III must work along the other template strand in the direction *away from* the replication fork. The DNA strand elongating in this direction is called the **lagging strand**. In contrast to the leading strand, which elongates continuously, the lagging strand is synthesized discontinuously, as a series of segments. These segments of the lagging strand are called **Okazaki fragments**, after Reiji Okazaki, the Japanese scientist who discovered them. The fragments are about 1,000–2,000 nucleotides long in *E. coli* and 100–200 nucleotides long in eukaryotes.

▼ Figure 16.16 Synthesis of the leading strand during DNA replication. This diagram focuses on the left replication fork shown in the overview box. DNA polymerase III (DNA pol III), shaped like a cupped hand, is shown closely associated with a protein called the "sliding clamp" that encircles the newly synthesized double helix like a doughnut. The sliding clamp moves DNA pol III along the DNA template strand.

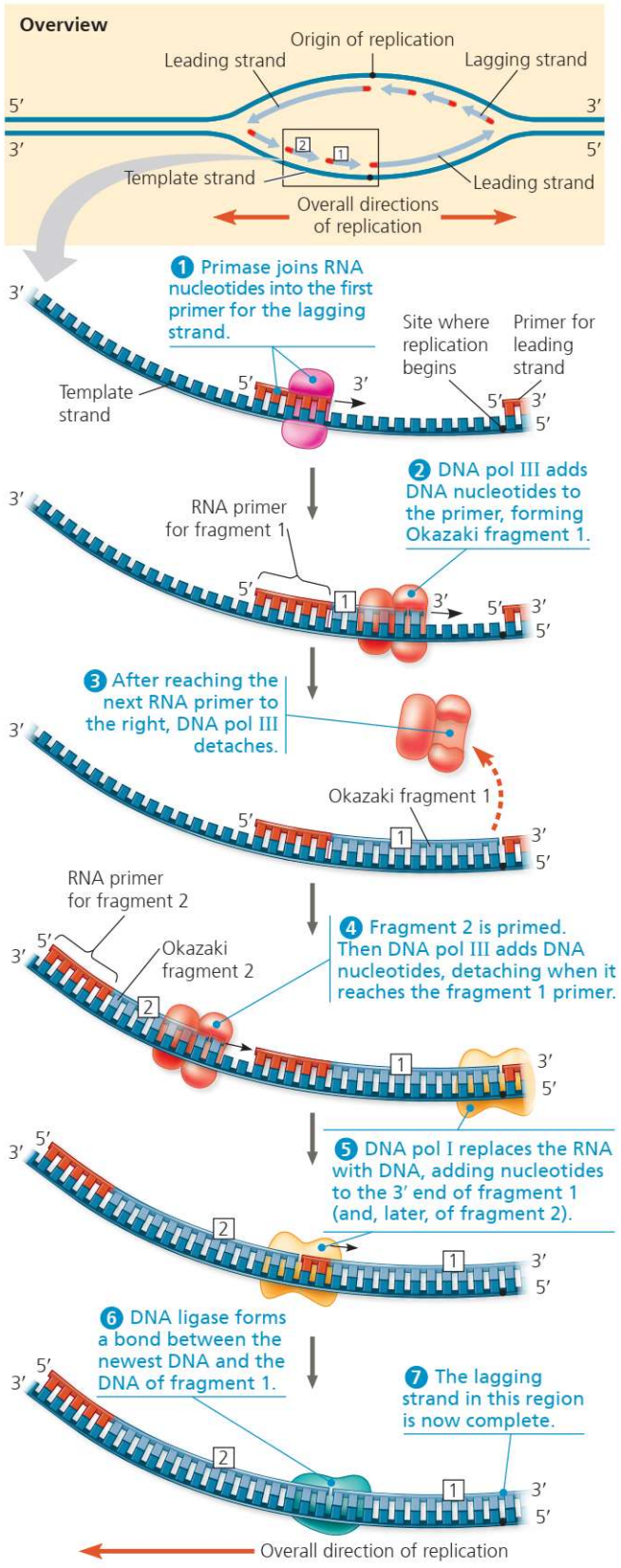


Mastering Biology BioFlix® Animation: Synthesis of the Leading Strand

Figure 16.17 illustrates the steps in the synthesis of the lagging strand at one fork. Whereas only one primer is required on the leading strand, each Okazaki fragment on the lagging strand must be primed separately (steps 1 and 4). After DNA pol III forms an Okazaki fragment (steps 2 to 4), another DNA polymerase, DNA pol I, replaces the RNA nucleotides of the adjacent primer with DNA nucleotides one at a time (step 5). But DNA pol I cannot join the final nucleotide of this replacement DNA segment to the first DNA nucleotide of the adjacent Okazaki fragment. Another enzyme, **DNA ligase**, accomplishes this task, joining the sugar-phosphate backbones of all the Okazaki fragments into a continuous DNA strand (step 6).

Mastering Biology BioFlix® Animation: Synthesis of the Lagging Strand

▼ Figure 16.17 Synthesis of the lagging strand.



Synthesis of the leading strand and synthesis of the lagging strand occur concurrently and at the same rate. The lagging strand is so named because its synthesis is delayed slightly relative to synthesis of the leading strand; each new fragment of the lagging strand cannot be started until enough template has been exposed at the replication fork.

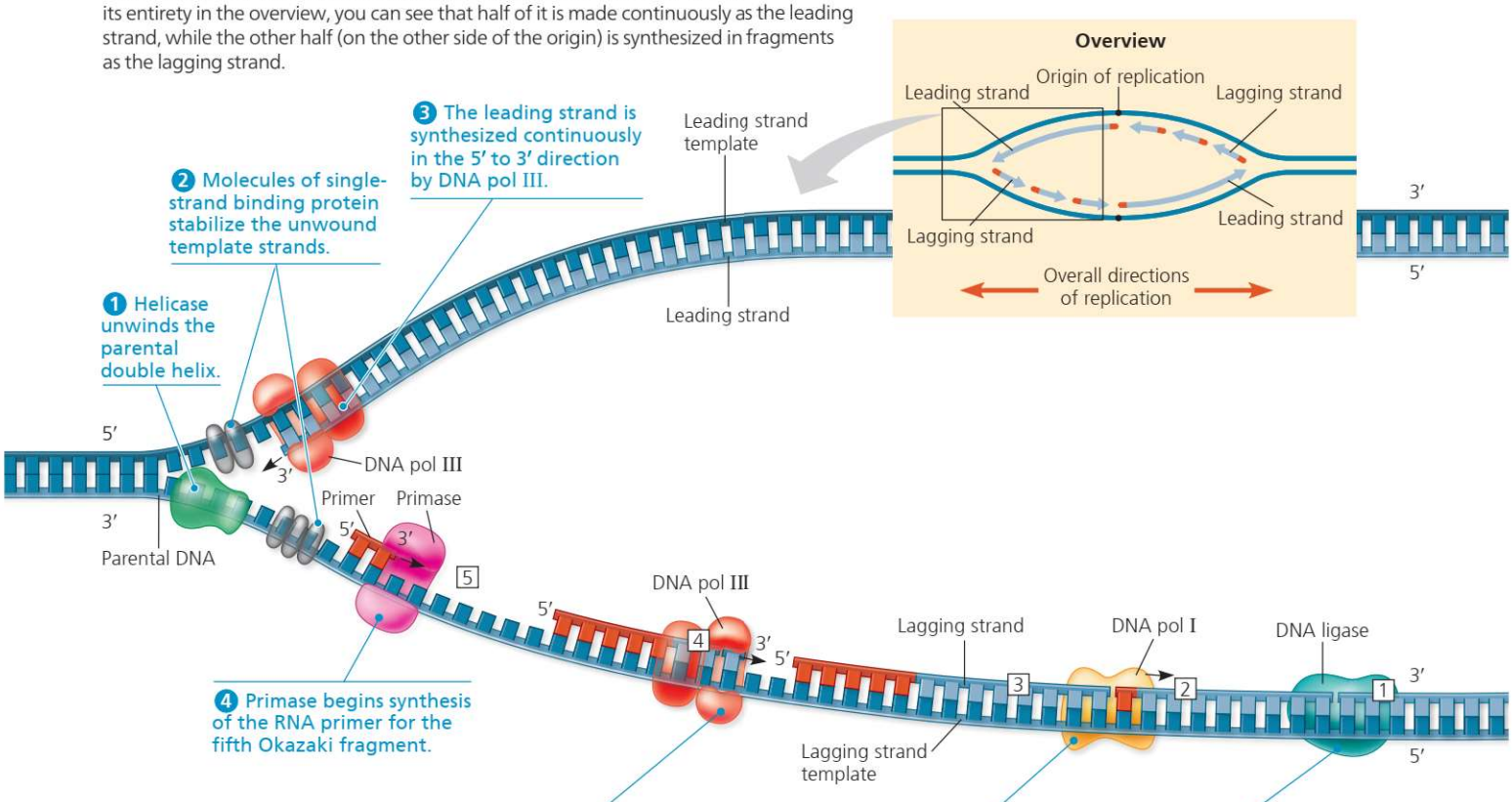
Figure 16.18 and **Table 16.1** summarize DNA replication. Please study them carefully before proceeding.

The DNA Replication Complex

It is traditional—and convenient—to represent DNA polymerase molecules as locomotives moving along a DNA railroad track, but such a model is inaccurate in two important ways. First, the various proteins that participate in DNA replication actually form a single large complex, a "DNA replication machine." Many protein-protein interactions facilitate the efficiency of this complex. For example, by interacting with other proteins at the fork, primase apparently acts as a molecular brake, slowing progress of the replication fork and coordinating the placement of primers and the rates of replication on the leading and lagging strands. Second, the DNA replication complex may not move along the DNA; rather, the DNA may move through the complex during the replication process. In eukaryotic cells, multiple copies of the complex, perhaps grouped into "factories," may be anchored to the nuclear matrix, a framework of fibers extending through the interior of the nucleus.

Some experimental evidence supports a model in which two DNA polymerase molecules, one on each template strand, "reel in" the parental DNA and extrude newly made daughter DNA molecules. In this so-called *trombone model*, the lagging strand is also looped back through the complex **(Figure 16.19)**. Whether the complex moves along the DNA or whether the DNA moves through the complex, either anchored or not, are still open, unresolved questions that are under active investigation.

▼ Figure 16.18 A summary of bacterial DNA replication. The detailed diagram shows the left-hand replication fork of the replication bubble shown in the overview (upper right). Viewing each daughter strand in its entirety in the overview, you can see that half of it is made continuously as the leading strand, while the other half (on the other side of the origin) is synthesized in fragments as the lagging strand.



DRAW IT *Draw a diagram showing the right-hand fork of the bubble in this figure. Number the Okazaki fragments, and label all 5′ and 3′ ends.*

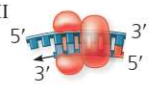
5 DNA pol III is completing synthesis of fragment 4. When it reaches the RNA primer on fragment 3, it will detach and begin adding DNA nucleotides to the 3′ end of the fragment 5 primer in the replication fork.

6 DNA pol I removes the primer from the 5′ end of fragment 2, replacing it with DNA nucleotides added one by one to the 3′ end of fragment 3. After the last addition, the backbone is left with a free 3′ end.

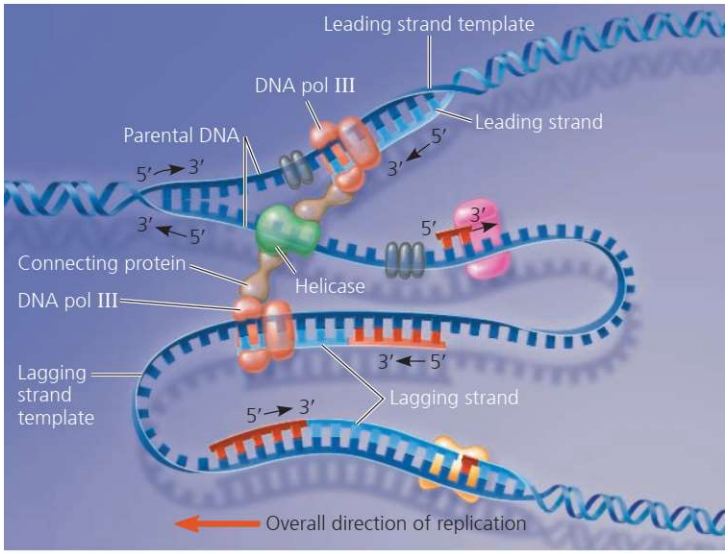
7 DNA ligase joins the 3′ end of fragment 2 to the 5′ end of fragment 1.

Mastering Biology **Animation: DNA Replication: A Closer Look**

Table 16.1 Bacterial DNA Replication Proteins and Their Functions

Protein	Function
Helicase 	Unwinds parental double helix at replication forks
Single-strand binding protein 	Binds to and stabilizes single-stranded DNA until it is used as a template
Topoisomerase 	Relieves overwinding strain ahead of replication forks by breaking, swiveling, and rejoining DNA strands
Primase 	Synthesizes an RNA primer at 5′ end of leading strand and at 5′ end of each Okazaki fragment of lagging strand
DNA pol III 	Using parental DNA as a template, synthesizes new DNA strand by adding nucleotides to an RNA primer or a pre-existing DNA strand
DNA pol I 	Removes RNA nucleotides of primer from 5′ end and replaces them with DNA nucleotides added to 3′ end of adjacent fragment
DNA ligase 	Joins Okazaki fragments of lagging strand; on leading strand, joins 3′ end of DNA that replaces primer to rest of leading strand DNA

▼ Figure 16.19 The "trombone" model of the DNA replication complex. In this proposed model, two molecules of DNA polymerase III work together in a complex, one on each strand, with helicase and other proteins. The lagging strand template DNA loops through the complex, resembling the slide of a trombone.



DRAW IT *Draw a line tracing the lagging strand template in this figure.*

Mastering Biology BioFlix® Animation: DNA Replication

Proofreading and Repairing DNA

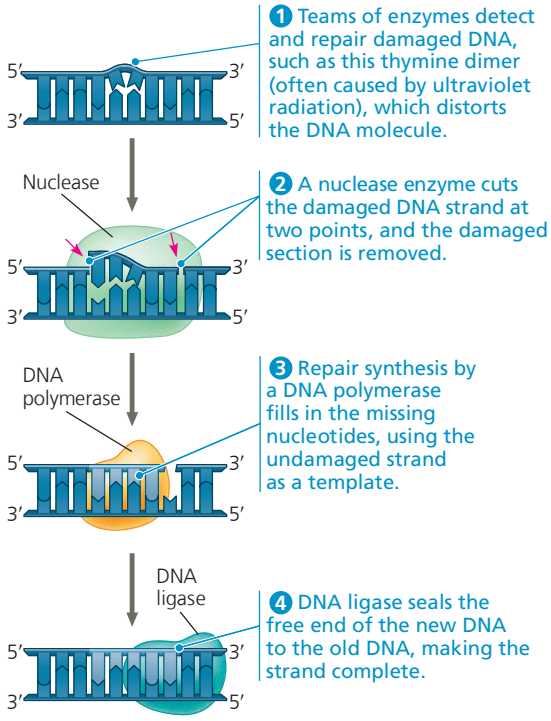
We cannot attribute the accuracy of DNA replication solely to the specificity of base pairing. Initial pairing errors between incoming nucleotides and those in the template strand occur at a rate of one in 10^5 nucleotides. However, errors in the completed DNA molecule amount to only one in 10^{10} (10 billion) nucleotides, an error rate that is 100,000 times lower. This is because during DNA replication, many DNA polymerases proofread each nucleotide against its template as soon as it is covalently bonded to the growing strand. Upon finding an incorrectly paired nucleotide, the polymerase removes the nucleotide and then resumes synthesis. (This action is similar to fixing a texting error by deleting the wrong letter and then entering the correct one.)

Mismatched nucleotides sometimes evade proofreading by a DNA polymerase. In **mismatch repair**, other enzymes remove and replace incorrectly paired nucleotides that have resulted from replication errors. Researchers highlighted the importance of such repair enzymes when they found that a hereditary defect in one of them is associated with a form of colon cancer. Apparently, this defect allows cancer-causing errors to accumulate in the DNA faster than normal.

Incorrectly paired or altered nucleotides can also arise after replication. In fact, maintenance of the genetic information encoded in DNA requires frequent repair of various kinds of damage to existing DNA. DNA molecules are constantly subjected to potentially harmful chemical and physical agents, such as X-rays, as we'll discuss in Concept 17.5. In addition, DNA bases may undergo spontaneous chemical changes under normal cellular conditions. However, these changes in DNA are usually corrected before they become permanent changes—*mutations*—perpetuated through successive replications. Each cell continuously monitors and repairs its genetic material. Because repair of damaged DNA is so important to the survival of an organism, it is no surprise that many different DNA repair enzymes have evolved. Almost 100 are known in *E. coli*, and about 170 have been identified so far in humans.

Most cellular systems for repairing incorrectly paired nucleotides, whether they are due to DNA damage or to replication errors, use a mechanism that takes advantage of the base-paired structure of DNA. In many cases, a segment of the strand containing the damage is cut out (excised) by a DNA-cutting enzyme—a **nuclease**—and the resulting gap is then filled in with nucleotides, using the undamaged strand as a template. The enzymes involved in filling the gap are a DNA polymerase and DNA ligase. There are several such DNA repair systems; one is called **nucleotide excision repair**.

▼ **Figure 16.20** **Nucleotide excision repair of DNA damage.**



An example is shown in **Figure 16.20**. An important function of the DNA repair enzymes in our skin cells is to repair genetic damage caused by the UV rays of sunlight: For example, adjacent thymine bases on a DNA strand can become covalently linked into *thymine dimers*, causing the DNA to buckle and interfere with DNA replication. The importance of repairing this kind of damage is underscored by a disorder called xeroderma pigmentosum (XP), which in most cases is caused by an inherited defect in a nucleotide excision repair enzyme. Individuals with XP are hypersensitive to sunlight; mutations in their skin cells caused by ultraviolet light are left uncorrected, often resulting in skin cancer. The effects are extreme: Without sun protection, children who have XP can develop skin cancer by age 10.

Evolutionary Significance of Altered DNA Nucleotides

EVOLUTION Faithful replication of the genome and repair of DNA damage are important for the functioning of the organism and for passing on a complete, accurate genome to the next generation. The error rate after proofreading and repair is extremely low, but rare mistakes do slip through. Once a mismatched nucleotide pair is replicated, the sequence change is permanent in the daughter molecule that has the incorrect nucleotide as well as in any subsequent copies. As we mentioned earlier, a permanent change in the DNA sequence is called a mutation.

Mutations can change the phenotype of an organism (see Concept 17.5). And if they occur in germ cells, which give rise to gametes, mutations can be passed on from generation to generation. The vast majority of such changes either have no effect or are harmful, but a very small percentage can be beneficial. In either case, mutations are the original source of the variation on which natural selection operates during evolution and are ultimately responsible for the appearance of new species. (You'll learn more about this process in Unit Four.) The balance between complete fidelity of DNA replication or repair and a low mutation rate has resulted in new proteins that contribute to different phenotypes. Ultimately, over long periods of time, this process leads to new species and thus to the rich diversity of life on Earth today.

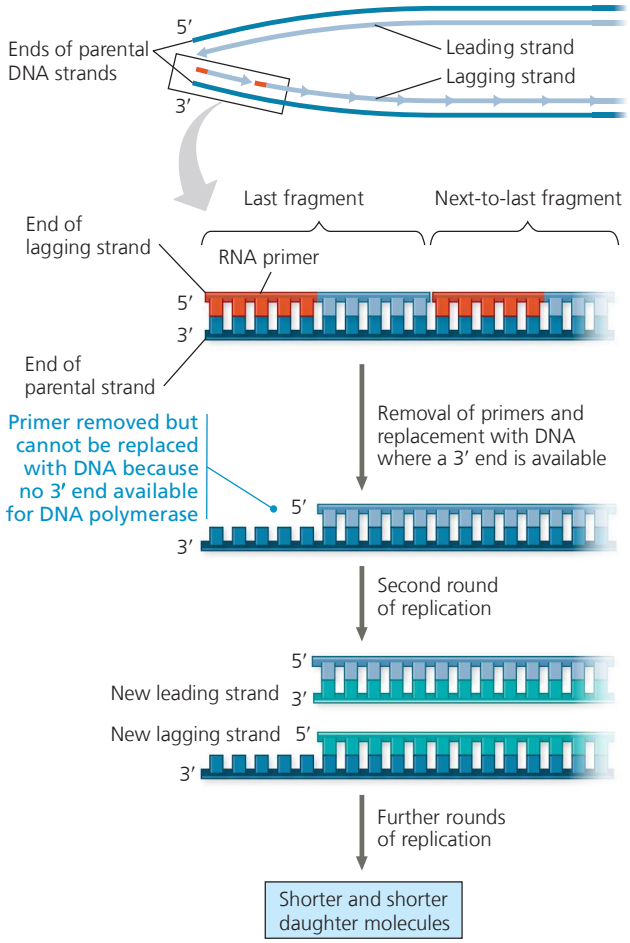
Replicating the Ends of DNA Molecules

For linear DNA, such as the DNA of eukaryotic chromosomes, the usual replication machinery cannot complete the 5′ ends of daughter DNA strands because there is no 3′ end of a preexisting polynucleotide for DNA polymerase to add onto. This is another consequence of the enzyme's requirements. Even if an Okazaki fragment can be started with an RNA primer hydrogen-bonded to the very end of the template strand, once that primer is removed, it cannot be replaced with DNA because there is no 3′ end available for nucleotide addition **(Figure 16.21)**. As a result, repeated rounds of replication produce shorter and shorter DNA molecules with uneven ("staggered") ends.

Most prokaryotes have a circular chromosome, with no ends, so the shortening of DNA does not occur. But what protects the genes of linear eukaryotic chromosomes from being eroded away during successive rounds of DNA replication? Eukaryotic chromosomal DNA molecules have special nucleotide sequences called **telomeres** at their ends **(Figure 16.22)**. Telomeres do not contain genes; instead, the DNA typically consists of multiple repetitions of one short nucleotide sequence. In each human telomere, for example, the six-nucleotide sequence TTAGGG is repeated between 100 and 1,000 times.

Telomeres have two protective functions. First, specific proteins associated with telomeric DNA prevent the staggered ends of the daughter molecule from activating the cell's systems for monitoring DNA damage. (Staggered ends of a DNA molecule, which often result from double-strand breaks, can trigger signal transduction pathways leading to cell cycle arrest or cell death.) Second, telomeric DNA acts as a kind of buffer zone that provides some protection against the organism's genes shortening, somewhat like how the plastic-wrapped ends of a shoelace slow down its unraveling. Telomeres do not prevent the erosion of genes near the ends of chromosomes; they merely postpone it.

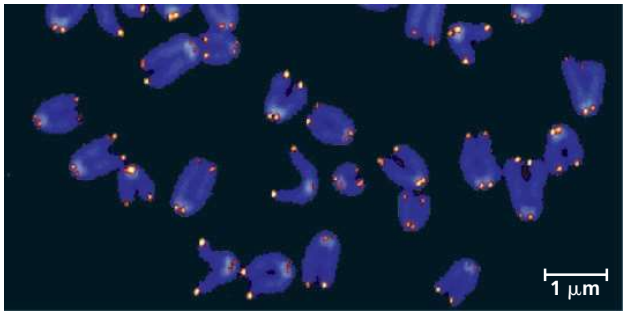
▼ Figure 16.21 Shortening of the ends of linear DNA molecules. Here we follow the left end of one DNA molecule through two rounds of replication. After the first round, the new lagging strand is shorter than its template. After a second round, both the leading and lagging strands have become shorter than the original parental DNA. Although not shown here, the other ends of these chromosomal DNA molecules (not shown) also become shorter.



As shown in Figure 16.21, telomeres become shorter during every round of replication. Thus, as expected, telomeric DNA tends to be shorter in dividing somatic cells of older individuals and in cultured cells that have divided many times. It has been proposed that shortening of telomeres is somehow connected to the aging process of certain tissues and even to aging of the organism as a whole.

But what about germ cells, whose genome must persist virtually unchanged from an organism to its offspring over many generations? If the chromosomes of germ cells became shorter in every cell cycle, essential genes would eventually be missing from the gametes they produce. However, this does not occur: An enzyme called *telomerase* catalyzes the lengthening of telomeres in eukaryotic germ cells, thus restoring their original length and compensating for the shortening that occurs during DNA replication. This enzyme contains its own RNA molecule that it uses as a template to artificially "extend" the leading strand, allowing the lagging strand to maintain a given length. Telomerase is not active in most human somatic cells, but its activity varies from tissue to tissue. The activity of telomerase in germ cells results in telomeres of maximum length in the zygote.

▼ Figure 16.22 Telomeres. Eukaryotes have repetitive, noncoding sequences called telomeres at the ends of their DNA. Telomeres are stained orange in these mouse chromosomes (LM).



Mastering Biology BBC Video: Can Humans Live Forever?

Normal shortening of telomeres may protect organisms from cancer by limiting the number of divisions that somatic cells can undergo. Cells from large tumors often have unusually short telomeres, as we would expect for cells that have undergone many cell divisions. Further shortening would presumably lead to self-destruction of the tumor cells. Telomerase activity is abnormally high in cancerous somatic cells, suggesting that its ability to stabilize telomere length may allow these cancer cells to persist. Many cancer cells do seem capable of unlimited cell division, as do immortal strains of cultured cells (see Concept 12.3). For several years, researchers have studied inhibition of telomerase as a possible cancer therapy. While studies that inhibited telomerase in mice with tumors have led to the death of cancer cells, eventually the cells have restored the length of their telomeres by an alternative pathway. This is an area of ongoing research that may eventually yield useful cancer treatments.

CONCEPT CHECK 16.2

1. What role does complementary base pairing play in the replication of DNA?
2. Identify two major functions of DNA pol III in DNA replication.
3. **MAKE CONNECTIONS** What is the relationship between DNA replication and the S phase of the cell cycle? See Figure 12.6.
4. **VISUAL SKILLS** If the DNA pol I in a given cell were nonfunctional, how would that affect the synthesis of a *leading* strand? In the overview box in Figure 16.18, point out where DNA pol I would normally function on the top leading strand.

For suggested answers, see Appendix A.

CONCEPT 16.3

A chromosome consists of a DNA molecule packed together with proteins

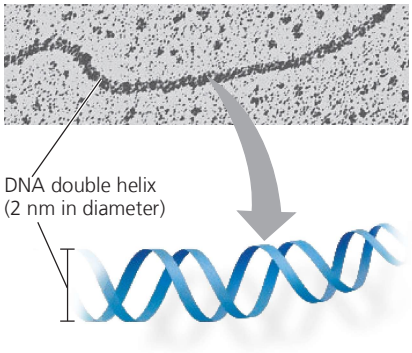
Now let's examine how DNA is packaged into chromosomes, the structures that carry genetic information. The main component of the genome in most bacteria is a double-stranded, circular DNA molecule that is associated with specific proteins. A bacterial chromosome differs from a eukaryotic chromosome in that a eukaryotic chromosome consists of a single linear DNA molecule associated with a large number of proteins. In *E. coli*, the chromosomal DNA consists of about 4.6 million nucleotide pairs, representing about 4,400 genes. This is 100 times more DNA than is found in a typical virus, but only about one-thousandth as much DNA as in a human somatic cell. Even so, that is a tremendous amount of DNA to be packaged in such a small container.

Stretched out, the DNA of an *E. coli* cell would measure about a millimeter in length, 1,000 times longer than the region it occupies in the cell. Within a bacterium, however, certain proteins cause the chromosome to coil and "supercoil," densely packing it so that it fills only part of the cell.

▼ Figure 16.23 Exploring Chromatin Packing in a Eukaryotic Chromosome

DNA, the Double Helix

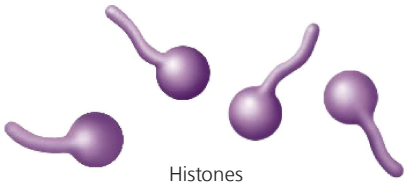
Shown below is a ribbon model of DNA, with each ribbon representing one of the polynucleotide strands. The TEM shows a molecule of naked (protein-free) DNA; the double helix is 2 nm across.



Histones

Proteins called **histones** are responsible for the main level of DNA packing in interphase chromatin. More than a fifth of a histone's amino acids are positively charged (Lys or Arg) and therefore bind tightly to the negatively charged DNA.

Four types of histones are most common in chromatin. The histones are very similar among eukaryotes, probably reflecting their important role in chromatin structure.

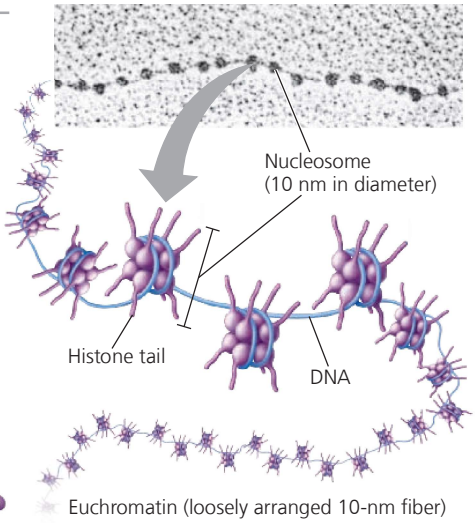

Histones

Nucleosomes in a 10-nm Fiber

In electron micrographs, unfolded chromatin is roughly 10 nm in diameter (the *10-nm fiber*). Such chromatin resembles beads on a string (see the TEM). Each "bead" is a **nucleosome**, the basic unit of DNA packing; the "string" between beads is called *linker DNA*.

A nucleosome consists of DNA wound twice around a protein core of eight histones, two each of the four main histone types. The amino end of each histone (the *histone tail*) extends outward from the nucleosome and is involved in regulation of gene expression.

In the cell cycle, the histones leave the DNA only briefly during DNA replication. Generally, they do the same during the process of transcription, which requires access to the DNA by transcription proteins.



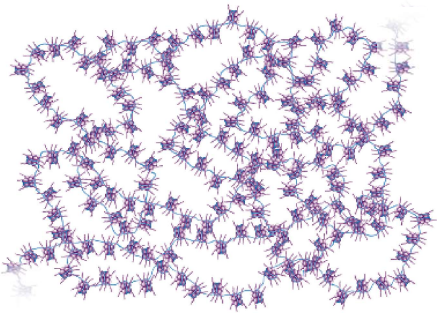
Euchromatin (loosely arranged 10-nm fiber)

Euchromatin/Heterochromatin

A recent technique called ChromEMT allows scientists to visualize chromatin in intact cells. ChromEMT and other new techniques have shown that the 10-nm fiber is the basic constituent of interphase chromatin.

During interphase, different regions of a chromosome may exist as euchromatin (below left) or heterochromatin (below). In euchromatin, the 10-nm fiber is loosely arranged in a more open configuration than heterochromatin; higher degrees of organization, including the 30-nm fiber in one older model, may exist in specific cells or at certain times. (This is an area of very active research.) The DNA in euchromatin is accessible to the proteins that carry out transcription, and its genes can be expressed. In heterochromatin, the 10-nm fiber is more densely arranged and less accessible to these proteins; genes in heterochromatin are generally not expressed.

Euchromatin and heterochromatin are organized into regions by other proteins not shown here; this organization is dynamic but disappears once mitosis begins.


Heterochromatin (densely arranged 10-nm fiber)

Unlike the nucleus of a eukaryotic cell, this dense region of DNA in a bacterium, called the nucleoid, is not bounded by membrane (see Figure 6.5).

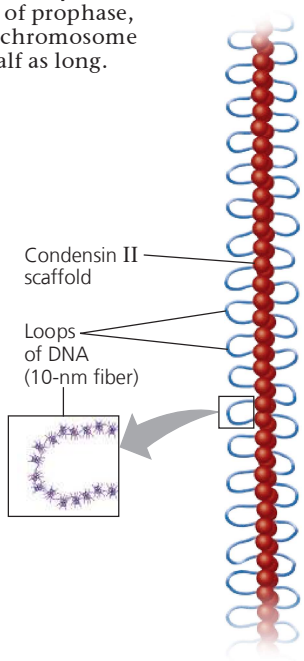
Each eukaryotic chromosome contains a single linear DNA double helix that, in humans, averages about 1.5×10^8 nucleotide pairs. This is an enormous amount of DNA relative to a chromosome's condensed length. If completely stretched out, such a DNA molecule would be about 4 cm long, thousands of times the diameter of a cell nucleus—and that's not even considering the DNA of the other 45 human chromosomes!

In the cell, eukaryotic DNA is precisely combined with a large amount of protein. The exact way in which this complex of DNA and protein, called **chromatin**, fits into the nucleus has long been debated. **Figure 16.23** outlines the current view of the organization of chromatin during interphase and how it is condensed during mitosis into the metaphase chromosome. Study this figure carefully before reading further.

Chromatin undergoes striking changes in how densely it is packed during the course of the cell cycle (see Figure 12.7). In interphase cells stained for light microscopy, the

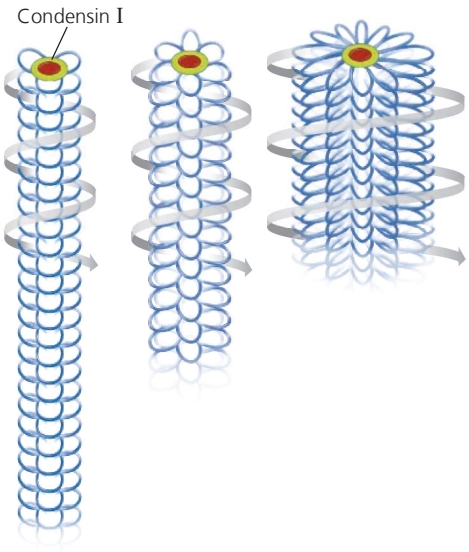
Prophase

When mitosis begins, DNA replication has already occurred, so each chromosome consists of two sister chromatids. During prophase, the chromatin of each sister chromatid begins to condense. Two related proteins called condensin II and condensin I play important roles. First, condensin II proteins (red) bind to the 10-nm fiber of DNA and form DNA loops that get larger and larger. The condensin II proteins form a central scaffold from which the loops extend. As the loops grow, the chromosome gets wider and shorter. By the end of prophase, the chromosome is half as long.



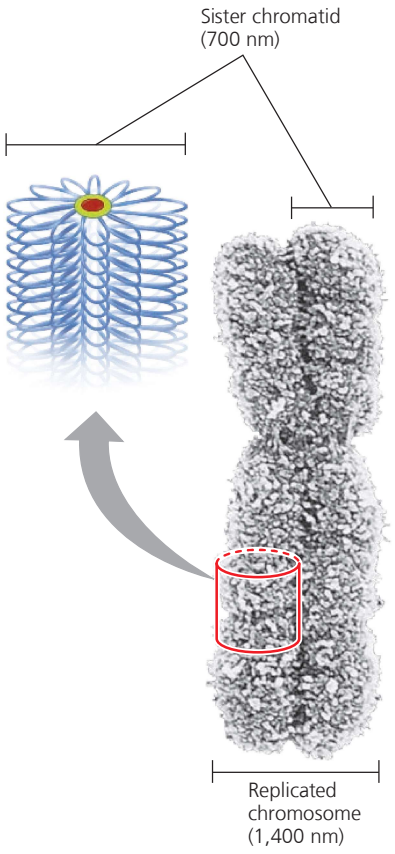
Prometaphase

When prometaphase begins, condensin I proteins (green) bind to DNA outside the central scaffold, making smaller loops (not shown) out of the larger loops generated by condensin II. The process continues, with more and more loops extending outward, and the chromosome getting denser, shorter, and wider. The scaffold itself also begins to twist into a helix (suggested by the curved gray arrows), allowing even more loops per given length of chromosome.



Metaphase

At metaphase, the chromosome is at its most dense, with the most loops per turn, and therefore is at its shortest length. The two sister chromatids are fully condensed.



Mastering Biology **Animation: DNA Packing**

chromatin usually appears as a diffuse mass within the nucleus, with some denser clumps, including in regions of centromeres and telomeres. The less compacted, more dispersed interphase chromatin is called **euchromatin** ("true chromatin") to distinguish it from the more compacted, denser-appearing **heterochromatin** (see Figure 16.23). Recent research has clarified our understanding of chromatin structure: For both types of chromatin, the basic organizing unit is the 10-nm fiber—nucleosomes joined by linker DNA. In heterochromatin, this 10-nm fiber is folded and bent back on itself to a much greater degree than in euchromatin, accounting for its denser appearance. Because heterochromatin is so compacted, it is largely inaccessible to the proteins responsible for transcribing the genetic information, a crucial early step in gene expression. In contrast, the looser packing of euchromatin makes its DNA accessible to those proteins, and the genes present in euchromatin are available for transcription.

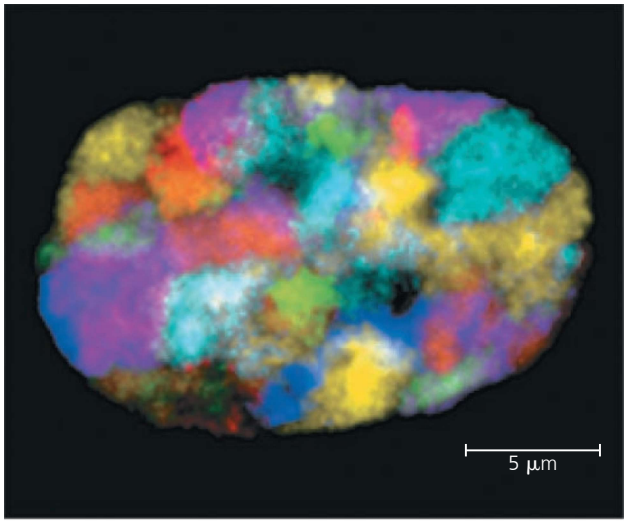
Early on, biologists assumed that interphase chromatin was a tangled mass in the nucleus, like a bowl of spaghetti, but this is far from the case. Although an interphase chromosome lacks an obvious scaffold, there are proteins that further organize the 10-nm fiber into larger compartments and smaller looped domains. Some of the looped domains appear to be attached to the nuclear lamina, on the inside of the nuclear envelope, and perhaps also to fibers of the nuclear matrix. These attachments may help organize regions of chromatin where genes are active. The chromatin of each chromosome occupies a specific restricted area within the interphase nucleus, and the chromatin fibers of different chromosomes do not appear to be entangled **(Figure 16.24a)**.

As a cell prepares for mitosis, its chromatin becomes organized into loops and coils, eventually condensing into a characteristic number of short, thick metaphase chromosomes that are distinguishable from each other with the light microscope **(Figure 16.24b)**.

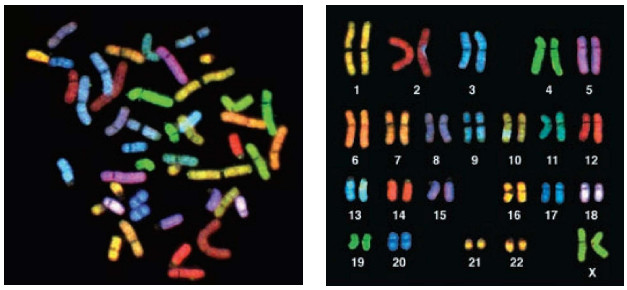
The chromosome is a dynamic structure that is condensed, loosened, modified, and remodeled as necessary for various cell processes, including DNA replication, mitosis, meiosis, and gene expression. Certain chemical modifications of histones affect the state of chromatin condensation and also have multiple effects on gene expression, as you'll see in Concept 18.2.

In this chapter, you have learned how DNA molecules are arranged in chromosomes and how DNA replication provides the copies of genes that parents pass to offspring. However, it is not enough that genes be copied and transmitted; the information they carry must be used by the cell. In other words, genes must also be expressed. In the next chapter, we will examine how the cell expresses the genetic information encoded in DNA.

▼ Figure 16.24 "Painting" chromosomes. Researchers can treat ("paint") human chromosomes with molecular tags that cause each chromosome pair to appear a different color.



(a) The ability to visually distinguish among chromosomes makes it possible to see how the chromosomes are arranged in the interphase nucleus. Each chromosome appears to occupy a specific territory during interphase. In general, the two homologs of a pair are not located together.



(b) These metaphase chromosomes have been "painted" so that the two homologs of a pair are the same color. On the left is a spread of treated chromosomes; on the right, they have been organized into a karyotype.

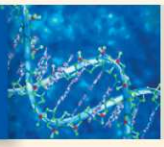
MAKE CONNECTIONS *If you arrested a human cell in metaphase I of meiosis and applied this technique, what would you observe? How would this differ from what you would see in metaphase of mitosis? Review Figure 13.8 and Figure 12.7.*

CONCEPT CHECK 16.3

1. Describe the structure of a nucleosome, the basic unit of DNA packing in eukaryotic cells.
2. How does euchromatin differ from heterochromatin in structure and function?
3. **MAKE CONNECTIONS** Interphase chromosomes appear to be attached to the nuclear lamina and perhaps also the nuclear matrix. Describe these two structures. See Figure 6.9 and the associated text.

For suggested answers, see Appendix A.

16 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

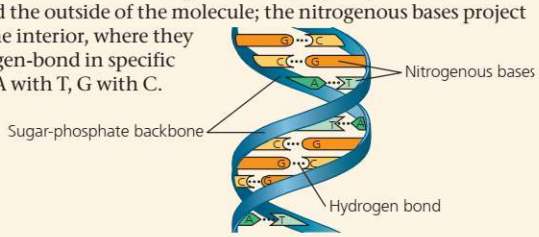
SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 16.1

DNA is the genetic material *(pp. 315–320)*

- Experiments with bacteria and with **phages** provided the first strong evidence that the genetic material is DNA.
- Watson and Crick deduced that DNA is a **double helix** and built a structural model. Two **antiparallel** sugar-phosphate chains wind around the outside of the molecule; the nitrogenous bases project into the interior, where they hydrogen-bond in specific pairs, A with T, G with C.

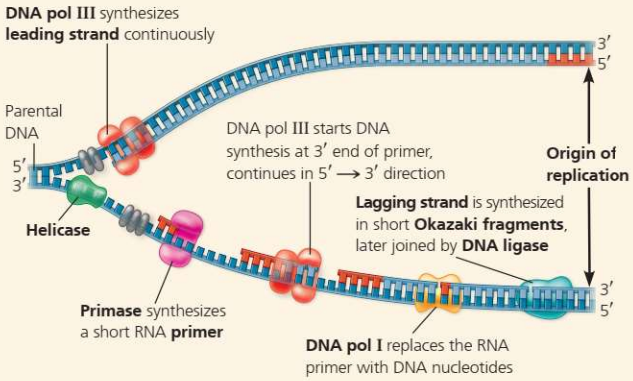


? *What does it mean when we say that the two DNA strands in the double helix are antiparallel? What would an end of the double helix look like if the strands were parallel?*

CONCEPT 16.2

Many proteins work together in DNA replication and repair *(pp. 320–329)*

- The Meselson-Stahl experiment showed that **DNA replication** is **semiconservative**: The parental molecule unwinds, and each strand then serves as a template for the synthesis of a new strand according to base-pairing rules.
- DNA replication at one **replication fork** is summarized here:



- **DNA polymerases** proofread new DNA, replacing incorrect nucleotides. In **mismatch repair**, enzymes correct errors that persist. **Nucleotide excision repair** is a process by which **nucleases** cut out and other enzymes replace damaged stretches of DNA.
- The ends of eukaryotic chromosomal DNA get shorter with each round of replication. The presence of **telomeres**, repetitive sequences at the ends of linear DNA molecules, postpones the erosion of genes. Telomerase catalyzes the lengthening of telomeres in germ cells.

? *Compare DNA replication on the leading and lagging strands, including similarities and differences.*

CONCEPT 16.3

A chromosome consists of a DNA molecule packed together with proteins *(pp. 330–332)*

- The chromosome of most bacterial species is a circular DNA molecule with some associated proteins, making up the nucleoid.
- The **chromatin** making up a eukaryotic chromosome is composed of DNA, **histones**, and other proteins. The histones bind to each other and to the DNA to form **nucleosomes**, the most basic units of DNA packing, which exist as part of the 10-nm fiber. Histone tails extend outward from each bead-like nucleosome core. Additional coiling and folding lead ultimately to the highly condensed chromatin of the metaphase chromosome.
- Chromosomes occupy restricted areas in the interphase nucleus. In interphase cells, most 10-nm fiber chromatin is loosely arranged (**euchromatin**), but some is more densely arranged (**heterochromatin**). Euchromatin, but not heterochromatin, is generally accessible for gene transcription.

? *Describe the levels of chromatin packing you'd expect to see in an interphase nucleus.*

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. In his work with pneumonia-causing bacteria and mice, Griffith found that
   (A) the protein coat from pathogenic cells was able to transform nonpathogenic cells.
   (B) heat-killed pathogenic cells caused pneumonia.
   (C) some substance from pathogenic cells was transferred to nonpathogenic cells, making them pathogenic.
   (D) the polysaccharide coat of bacteria caused pneumonia.
2. What is the basis for the difference in how the leading and lagging strands of DNA molecules are synthesized?
   (A) The origins of replication occur only at the 5′ end.
   (B) Helicases and single-strand binding proteins work at the 5′ end.
   (C) DNA polymerase can join new nucleotides only to the 3′ end of a pre-existing strand, and the strands are antiparallel.
   (D) DNA ligase works only in the 3′ → 5′ direction.
3. In analyzing the number of different bases in a DNA sample, which result would be consistent with the base-pairing rules?
   (A) A = G
   (B) A + G = C + T
   (C) A + T = G + C
   (D) A = C
4. The elongation of the leading strand during DNA synthesis
   (A) progresses away from the replication fork.
   (B) occurs in the 3′ → 5′ direction.
   (C) produces Okazaki fragments.
   (D) depends on the action of DNA polymerase.
5. In a nucleosome, the DNA is wrapped around
   (A) histones.
   (B) ribosomes.
   (C) polymerase molecules.
   (D) a thymine dimer.

Levels 3-4: Applying/Analyzing

6. *E. coli* cells grown on ^{15}N medium are transferred to ^{14}N medium and allowed to grow for two more generations (two rounds of DNA replication). DNA extracted from these cells is centrifuged. What density distribution of DNA would you expect in this experiment?
(A) one high-density and one low-density band
(B) one intermediate-density band
(C) one high-density and one intermediate-density band
(D) one low-density and one intermediate-density band

7. A student isolates, purifies, and combines in a test tube a variety of molecules needed for DNA replication. When she adds some DNA to the mixture, replication occurs, but each DNA molecule consists of a normal strand paired with numerous segments of DNA a few hundred nucleotides long. What has she probably left out of the mixture?
(A) DNA polymerase
(B) DNA ligase
(C) Okazaki fragments
(D) primase

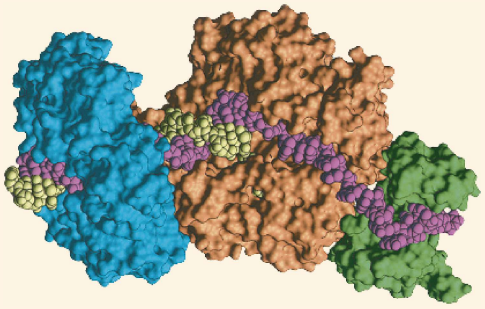
8. The spontaneous loss of amino groups from adenine in DNA results in hypoxanthine, an uncommon base, opposite thymine. What combination of proteins could repair such damage?
(A) nuclease, DNA polymerase, DNA ligase
(B) telomerase, primase, DNA polymerase
(C) telomerase, helicase, single-strand binding protein
(D) DNA ligase, replication fork proteins, adenylyl cyclase

9. MAKE CONNECTIONS Although the proteins that cause the *E. coli* chromosome to coil are not histones, what property would you expect them to share with histones, given their ability to bind to DNA (see Figure 5.14)?

Levels 5-6: Evaluating/Creating

10. EVOLUTION CONNECTION Some bacteria may be able to respond to environmental stress by increasing the rate at which mutations occur during cell division. How might this be accomplished? Might there be an evolutionary advantage to this ability? Explain.

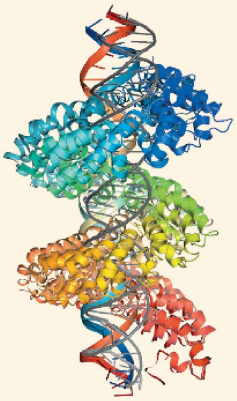
11. SCIENTIFIC INQUIRY



DRAW IT Model building can be an important part of the scientific process. The illustration shown above is a computer-generated model of a DNA replication complex. The parental and newly synthesized DNA strands are color-coded differently, as are each of the following three proteins: DNA pol III, the sliding clamp, and single-strand binding protein.
(a) Using what you've learned in this chapter to clarify this model, label each DNA strand and each protein.
(b) Draw an arrow to indicate the overall direction of DNA replication.

12. WRITE ABOUT A THEME: ORGANIZATION The continuity of life is based on heritable information in the form of DNA, and structure and function are correlated at all levels of biological organization. In a short essay (100–150 words), describe how the structure of DNA is correlated with its role as the molecular basis of inheritance.

13. SYNTHESIZE YOUR KNOWLEDGE



This image shows DNA (gray) interacting with a computer-generated model of a TAL protein (multicolored), one of a family of proteins found only in a species of the bacterium *Xanthomonas*. The bacterium uses proteins like this one to find specific gene sequences in cells of the organisms it infects, such as tomatoes, rice, and citrus fruits. Given what you know about DNA structure and considering the image above, discuss how the TAL protein's structure suggests that it functions.

For selected answers, see Appendix A.

17 Gene Expression: From Gene to Protein

KEY CONCEPTS

17.1 Genes specify proteins via transcription and translation *p. 336*

17.2 Transcription is the DNA-directed synthesis of RNA: ***A Closer Look*** *p. 342*

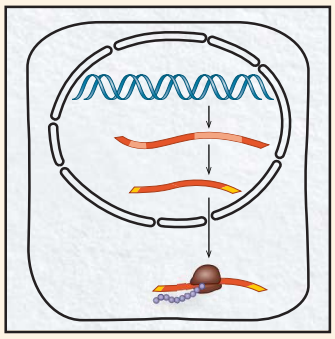
17.3 Eukaryotic cells modify RNA after transcription *p. 345*

17.4 Translation is the RNA-directed synthesis of a polypeptide: ***A Closer Look*** *p. 347*

17.5 Mutations of one or a few nucleotides can affect protein structure and function *p. 357*

Study Tip

Make a visual study guide: Sketch the process shown below, and add labels and details as you read the chapter. (In this exercise, assume all processes take place in a eukaryotic cell.)



Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 17
- BioFlix® Animation: Protein Synthesis
- Figure 17.27 Walkthrough: Types of Small-Scale Mutations that Affect mRNA Sequence

For Instructors to Assign (in Item Library)
- BioFlix® Tutorial: Protein Synthesis (1 of 3): Overview
- Tutorial: CRISPR: A Revolution in Genome Editing

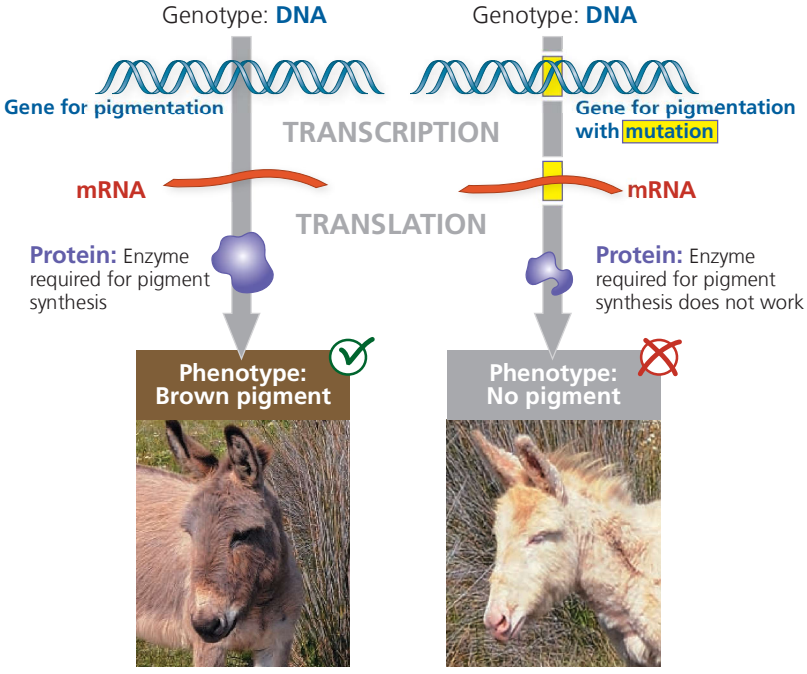
Ready-to-Go Teaching Module (in Instructor Resources)
- Gene Expression: Mutations (Concept 17.5)



Figure 17.1 A population of albino donkeys grazes on vegetation on the hillsides of Asinara, an Italian island. Several centuries ago, a recessive mutation that disables pigment synthesis arose in the DNA of one donkey and was passed down through the generations. Inbreeding has resulted in a large number of homozygous albino donkeys living on the island today.

How can one change in DNA result in such a dramatic change in appearance?

Proteins are the link between genotype and phenotype. Gene expression is the process by which DNA directs the synthesis of proteins:



CONCEPT 17.1

Genes specify proteins via transcription and translation

Inherited traits like albinism are determined by genes, which have information content in the form of specific sequences of nucleotides along stretches of DNA, the genetic material. The DNA inherited by an organism leads to specific traits by dictating the synthesis of proteins and of RNA molecules involved in protein synthesis. In the case of coat and skin color, the gene containing the information to synthesize an enzyme that makes pigment allows normal coloring, and a faulty version of that gene leads to lack of color, or the albino phenotype. Proteins are the link between genotype and phenotype. **Gene expression** is the process by which DNA directs the synthesis of proteins (or, in some cases, just RNAs). The expression of genes that code for proteins includes two stages: transcription and translation.

Before going into the details of how genes direct protein synthesis, let's step back and examine how the fundamental relationship between genes and proteins was discovered.

Evidence from Studying Metabolic Defects

In 1902, British physician Archibald Garrod was the first to suggest that genes dictate phenotypes through enzymes, proteins that catalyze specific chemical reactions in the cell. He postulated that the symptoms of an inherited disease reflect an inability to make a particular enzyme. He later referred to such diseases as "inborn errors of metabolism." For example, people with a disease called alkaptonuria have black urine because it contains a chemical called alkapton, which darkens upon exposure to air. Garrod reasoned that these people cannot make an enzyme that breaks down alkapton, so alkapton is expelled in their urine.

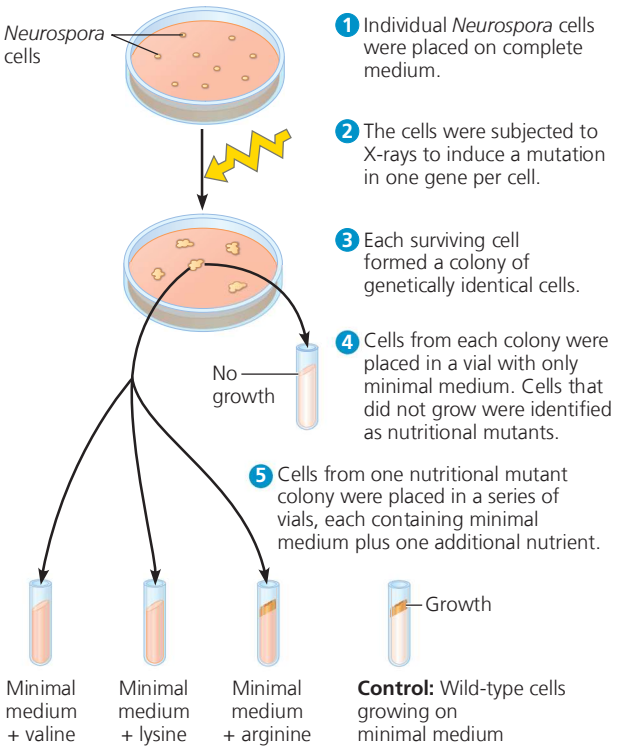
Several decades later, research supported Garrod's hypothesis that a gene dictates the production of a specific enzyme, later named the *one gene–one enzyme hypothesis*. Biochemists learned that cells synthesize and degrade most organic molecules via metabolic pathways, in which each chemical reaction in a sequence is catalyzed by a specific enzyme (see Concept 8.1). Such metabolic pathways lead, for instance, to the synthesis of the pigments that give the brown donkey in **Figure 17.1** its fur color or fruit flies (*Drosophila*) their eye color (see Figure 15.3). In the 1930s, the American biochemist and geneticist George Beadle and his French colleague Boris Ephrussi speculated that in *Drosophila*, each mutation affecting eye color blocks pigment synthesis at a specific step by preventing production of the enzyme that catalyzes that step. But neither the chemical reactions nor the enzymes that catalyze them were known at the time.

Nutritional Mutants in Neurospora: Scientific Inquiry

A breakthrough in demonstrating the relationship between genes and enzymes came a few years later at Stanford University, where Beadle and Edward Tatum began working with a bread mold, *Neurospora crassa*. Unlike the diploid organisms studied by Mendel (peas) and T. H. Morgan (fruit flies), *Neurospora* is a haploid species. Because its genome contains only one copy of each gene, that single copy determines the individual's expressed phenotype. Therefore, when Beadle and Tatum wanted to discover the function of any gene, they only needed to mutate and disable that one allele. They chose to study the protein-coding genes required for a specific nutritional activity. They caused mutations in genes by bombarding *Neurospora* with X-rays and looked among the survivors for mutants that differed in their nutritional needs from the wild-type bread mold.

Wild-type *Neurospora* has modest food requirements. It can grow in the laboratory on *minimal medium*—a simple solution containing minimal nutrients for growth—inorganic salts, glucose, and the vitamin biotin—incorporated into agar, a support medium. From these basics, wild-type mold cells use their metabolic pathways to produce all the other nutrient molecules (like amino acids) they need for growth, dividing repeatedly and forming visible colonies of genetically identical cells. As shown in **Figure 17.2**, Beadle and Tatum generated different "nutritional mutants" of *Neurospora* cells, each of which had a mutation in one gene and was unable to synthesize a particular

▼ Figure 17.2 Beadle and Tatum's experimental approach. To obtain nutritional mutants, Beadle and Tatum exposed *Neurospora* cells to X-rays, inducing mutations, then screened for (tested for) mutants that had new nutritional requirements, such as arginine, as shown here.



essential nutrient. Such cells could not grow on minimal medium but could grow on *complete medium*, which contains all nutrients needed for growth. For *Neurospora*, the complete medium consists of the minimal medium supplemented with all 20 amino acids and a few other nutrients. Beadle and Tatum hypothesized that in each nutritional mutant, the gene for the enzyme that synthesizes a particular nutrient had been disabled by the mutation. They were able to determine experimentally which nutrient each mutant strain (original cell and its descendants) was unable to synthesize.

This approach resulted in a valuable collection of mutant strains of *Neurospora*, each catalogued by its defect in a particular pathway. Two colleagues, Adrian Srb and Norman Horowitz, used a collection of arginine-requiring mutants to investigate the biochemical pathway for arginine synthesis in *Neurospora* **(Figure 17.3)**. Srb and Horowitz pinned down each mutant's defect more specifically, using additional tests to distinguish among three classes of arginine-requiring mutants. Mutants in each class required a different set of compounds along the arginine-synthesizing pathway, which has three steps. These results, and those of many similar experiments done by Beadle and Tatum, suggested that each class was blocked at a different step in this pathway because mutants in that class lacked the enzyme that catalyzes the blocked step.

Because Beadle and Tatum set up their experimental conditions so that each mutant was defective in a single gene, the collected results, taken together, provided strong support for the *one gene–one enzyme hypothesis*, as they dubbed it: that the function of a gene is to dictate the production of a specific enzyme. Further support for this hypothesis came from experiments that identified the specific enzymes lacking in the mutants. Beadle and Tatum shared a Nobel Prize in 1958 for "their discovery that genes act by regulating definite chemical events" (in the words of the Nobel committee).

Today, we know of countless examples in which a mutation in a gene causes a faulty enzyme that in turn leads to an identifiable condition. The albino donkeys in Figure 17.1 lack a key enzyme called tyrosinase in the metabolic pathway that produces melanin, a dark pigment. The absence of melanin causes white fur and other effects throughout the donkey's body. Its nose, ears, and hooves, as well as its eyes, are pink because no melanin is present to mask the reddish color of the blood vessels that run through those structures.

The Products of Gene Expression: A Developing Story

As researchers learned more about proteins, they made revisions to the one gene–one enzyme hypothesis. First of all, not all proteins are enzymes. Keratin, the structural protein of animal hair, and the hormone insulin are two examples of nonenzyme proteins. Because proteins that are not enzymes are nevertheless gene products, molecular biologists began to think in terms of one gene–one protein. However, many proteins are constructed from two or more different polypeptide chains, and each polypeptide is specified by its own gene. For example, hemoglobin—the oxygen-transporting protein of vertebrate red blood cells—contains two kinds of polypeptides, and thus two genes code for this protein, one for each type of polypeptide (see Figure 5.18). Beadle and Tatum's idea was therefore restated as the *one gene–one polypeptide hypothesis*. Even this description is not entirely accurate, though. First, in many cases, a eukaryotic gene can code for a set of closely related polypeptides via a process called alternative splicing, which you will learn about later in this chapter. Second, quite a few genes code for RNA molecules that have important functions in cells even though they are never translated into protein. For now, we will focus on genes that do code for polypeptides. (Note that it is common to refer to these gene products as proteins—a practice you will encounter in this book—rather than more precisely as polypeptides.)

Basic Principles of Transcription and Translation

Genes provide the instructions for making specific proteins, but a gene does not build a protein directly. The bridge between DNA and protein synthesis is the nucleic acid RNA. RNA is chemically similar to DNA except that it contains ribose instead of deoxyribose as its sugar and has the nitrogenous base uracil rather than thymine (see Figure 5.23). Thus, each nucleotide along a DNA strand has A, G, C, or T as its base, and each nucleotide along an RNA strand has A, G, C, or U as its base. An RNA molecule usually consists of a single strand.

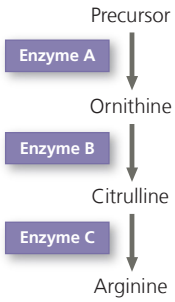
We often describe the flow of information from gene to protein in terms of languages. Just as specific sequences of letters communicate information in a language such as English, specific sequences of monomers convey information in polymers like nucleic acids and proteins. In DNA or RNA, the monomers are the four types of nucleotides listed previously, which differ in their nitrogenous bases. Genes are typically hundreds or thousands of nucleotides long, each gene having a specific sequence of nucleotides. Each polypeptide of a protein also has monomers arranged in a particular linear order (the protein's primary structure; see Figure 5.18), but its monomers are amino acids. Thus, nucleic acids and proteins contain information written in two different chemical languages. Getting from DNA to protein involves two major stages: transcription and translation.

Transcription is the synthesis (production) of RNA using information in the DNA. The two nucleic acids are written in different forms of the same language, and the information is simply transcribed, or "rewritten," from DNA to RNA. Just as a DNA strand provides a template for making a new complementary strand during DNA replication (see Concept 16.2), it also can serve as a template for assembling a complementary sequence of RNA nucleotides. For a protein-coding gene, the resulting RNA molecule is a faithful transcript of the gene's

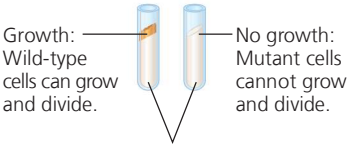
▼ Figure 17.3 Inquiry

Do individual genes specify the enzymes that function in a biochemical pathway?

Experiment Working with the mold *Neurospora crassa*, Adrian Srb and Norman Horowitz, then at Stanford University, used Beadle and Tatum's experimental approach (see Figure 17.2) to isolate mutants that required arginine in their growth medium. The researchers showed that these mutants fell into three classes, each defective in a different gene. From studies by others on mammalian liver cells, they suspected that the metabolic pathway of arginine biosynthesis involved a precursor nutrient and the intermediate molecules ornithine and citrulline, as shown in the diagram on the right.



Their most famous experiment, shown here, tested both the *one gene–one enzyme hypothesis* and their postulated arginine-synthesizing pathway. In this experiment, they grew their three classes of mutants under the four different conditions shown in the results table below. They included minimal medium (MM) as a control, knowing that wild-type cells could grow on MM but mutant cells could not. (See test tubes below.)



Control: Minimal medium

Results As shown in the table on the right, the wild-type strain was capable of growth under all experimental conditions, requiring only the minimal medium. The three classes of mutants each had a specific set of growth requirements. For example, class II mutants could not grow when ornithine alone was added but could grow when either citrulline or arginine was added.

Conclusion From the growth requirements of the mutants, Srb and Horowitz deduced that each class of mutant was unable to carry out one step in the pathway for synthesizing arginine, presumably because it lacked the necessary enzyme, as shown in the table on the right. Because each of their mutants was mutated in a single gene, they concluded that each mutated gene must normally dictate the production of one enzyme. Their results supported the one gene–one enzyme hypothesis, proposed by Beadle and Tatum, and also confirmed that the arginine pathway described in the mammalian liver also operates in *Neurospora*. (Notice in the results table that a mutant can grow only if supplied with a compound made *after* the defective step because this bypasses the defect.)

Data from A. M. Srb and N. H. Horowitz, The ornithine cycle in *Neurospora* and its genetic control, *Journal of Biological Chemistry* 154:129–139 (1944).

Results Table

Condition	Classes of *Neurospora crassa*			
	Wild type	**Class I mutants**	**Class II mutants**	**Class III mutants**
Minimal medium (MM) (control)				
MM + ornithine				
MM + citrulline				
MM + arginine (control)				
Summary of results	Can grow with or without any supplements	Can grow on ornithine, citrulline, or arginine	Can grow only on citrulline or arginine	Require arginine to grow

Gene (codes for enzyme)	**Wild type**	**Class I mutants** (mutation in gene *A*)	**Class II mutants** (mutation in gene *B*)	**Class III mutants** (mutation in gene *C*)
Gene *A* → Enzyme A Gene *B* → Enzyme B Gene *C* → Enzyme C	Precursor → Enzyme A → Ornithine → Enzyme B → Citrulline → Enzyme C → Arginine	Precursor → ~~Enzyme A~~ → Ornithine → Enzyme B → Citrulline → Enzyme C → Arginine	Precursor → Enzyme A → Ornithine → ~~Enzyme B~~ → Citrulline → Enzyme C → Arginine	Precursor → Enzyme A → Ornithine → Enzyme B → Citrulline → ~~Enzyme C~~ → Arginine

WHAT IF? *Suppose the experiment had shown that class I mutants could grow only in MM supplemented by ornithine or arginine and that class II mutants could grow in MM supplemented by citrulline, ornithine, or arginine. What conclusions would the researchers have drawn from those results regarding the biochemical pathway and the defect in class I and class II mutants?*

protein-building instructions. This type of RNA molecule is called **messenger RNA (mRNA)** because it carries a genetic message from the DNA to the protein-synthesizing machinery of the cell. (Transcription is the general term for the synthesis of *any* kind of RNA on a DNA template. Later, you will learn about some other types of RNA produced by transcription.)

Translation is the synthesis of a polypeptide using the information in the mRNA. During this stage, there is a change in language: The cell must translate the nucleotide sequence of an mRNA molecule into the amino acid sequence of a polypeptide. The sites of translation are **ribosomes**, molecular complexes that facilitate the orderly linking of amino acids into polypeptide chains.

Transcription and translation occur in all organisms. Because most studies have involved bacteria and eukaryotic cells, they are our main focus in this chapter. Our understanding of transcription and translation in archaea lags behind, but we do know that archaeal cells share some features of gene expression with bacteria and others with eukaryotes.

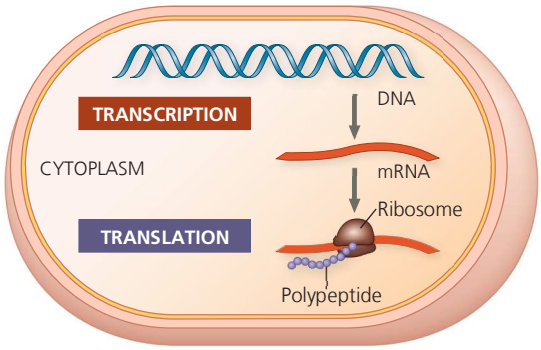
The basic mechanics of transcription and translation are similar for bacteria and eukaryotes, but there is an important difference in the flow of genetic information within the cells. Bacteria do not have nuclei. Therefore, nuclear membranes do not separate bacterial DNA and mRNA from ribosomes and the other protein-synthesizing equipment **(Figure 17.4a)**. This lack of compartmentalization allows translation of an mRNA to begin while its transcription is still in progress, as you'll see later. By contrast, eukaryotic cells have nuclei. The presence of a nuclear envelope separates transcription from translation in space and time **(Figure 17.4b)**. Transcription occurs in the nucleus, but the mRNA must be transported to the cytoplasm for translation. In eukaryotes, before RNA transcripts from protein-coding genes can leave the nucleus, they are modified in various ways to produce the final, functional mRNA. The transcription of a protein-coding eukaryotic gene results in *pre-mRNA*, and further RNA processing yields the finished mRNA. The initial RNA transcript from any gene, including those specifying RNA that is not translated into protein (such as tRNA and rRNA, which will be described in Concept 17.4), is more generally called a **primary transcript**.

To summarize: Genes program protein synthesis via genetic messages in the form of messenger RNA. Put another way, cells are governed by a molecular chain of command with a directional flow of genetic information:

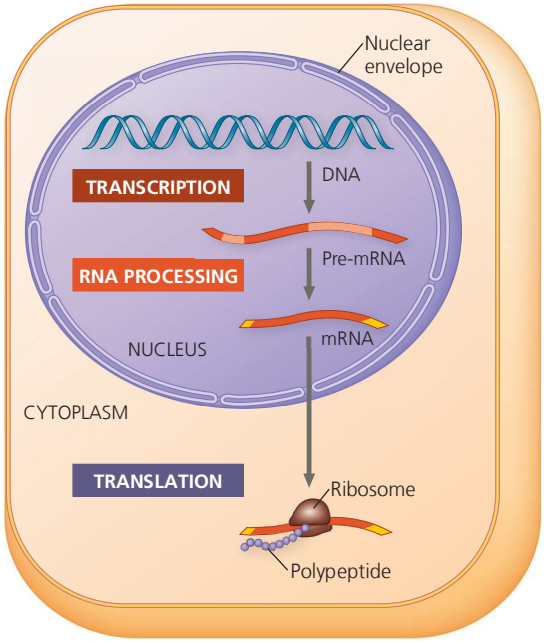


This idea, that the flow of information went only one way, was named the *central dogma* by Francis Crick in 1956. In the 1970s, however, scientists were surprised to discover some enzymes that use RNA molecules as templates for DNA synthesis, an example of information flow from RNA to DNA (see Concept 19.2). Still, these exceptions do not discredit the idea that, in general, genetic information flows from DNA to RNA to protein. In the next section, we discuss how the instructions for assembling amino acids into a specific order are encoded in nucleic acids.

▼ Figure 17.4 Overview: the roles of transcription and translation in the flow of genetic information. In a cell, inherited information flows from DNA to RNA to protein. The two main stages of information flow are transcription and translation. A miniature version of part (a) or (b) accompanies several figures later in the chapter as an orientation diagram to help you see where a particular figure fits into the overall scheme of gene expression.



(a) Bacterial cell. In a bacterial cell, which lacks a nucleus, mRNA produced by transcription is immediately translated without additional processing.



(b) Eukaryotic cell. The nucleus provides a separate compartment for transcription. The original RNA transcript, called pre-mRNA, is processed in various ways before leaving the nucleus as mRNA.

Mastering Biology
Animation: Overview of Protein Synthesis in Bacteria
Animation: Overview of Protein Synthesis in Eukaryotes

The Genetic Code

When biologists began to suspect that the instructions for protein synthesis were encoded in DNA, they recognized a problem: There are only four nucleotide bases to specify 20 amino acids. Thus, the genetic code cannot be a language like Chinese, where each written symbol corresponds to a word. How many nucleotides, then, would turn out to correspond to an amino acid?

Codons: Triplets of Nucleotides

If each kind of nucleotide base were translated into an amino acid, only four amino acids could be specified, one per nucleotide base. Would a language of two-letter code words suffice? The two-nucleotide sequence AG, for example, could specify one amino acid, and GT could specify another. Since there are four possible nucleotide bases in each position, this would give us 16 (that is, 4×4, or 4^2) possible arrangements—still not enough to code for all 20 amino acids.

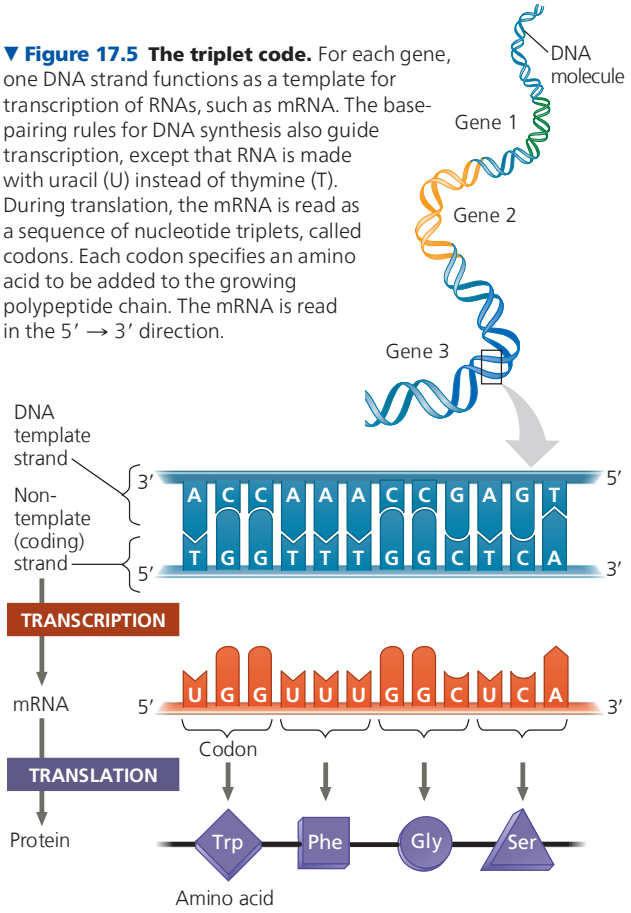
Triplets of nucleotide bases are the smallest units of uniform length that can code for all the amino acids. If each arrangement of three consecutive nucleotide bases specifies an amino acid, there can be 64 (that is, 4^3) possible code words—more than enough to specify all the amino acids. Experiments have verified that the flow of information from gene to protein is based on a **triplet code**: The genetic instructions for a polypeptide chain are written in the DNA as a series of nonoverlapping, three-nucleotide words. The series of words in a gene is transcribed into a complementary series of nonoverlapping, three-nucleotide words in mRNA, which is then translated into a chain of amino acids **(Figure 17.5)**.

During transcription, the gene determines the sequence of nucleotide bases along the length of the RNA molecule that is being synthesized. For each gene, only one of the two DNA strands is transcribed. This strand is called the **template strand** because it provides the pattern, or template, for the sequence of nucleotides in an RNA transcript. For any given gene, the same strand is used as the template every time that gene is transcribed. However, farther along on the same chromosomal DNA molecule, the opposite strand may function as the template for a different gene. Particular DNA sequences associated with a gene determine how the enzyme that transcribes genes is oriented when it binds, and this establishes which strand will be used as the template.

An mRNA molecule is complementary rather than identical to its DNA template because RNA nucleotides are assembled on the template according to base-pairing rules (see Figure 17.5). The pairs are similar to those that form during DNA replication, except that U (the RNA substitute for T) pairs with A and the mRNA nucleotides contain ribose instead of deoxyribose. Like a new strand of DNA, the RNA molecule is synthesized in an antiparallel direction to the template strand of DNA. (To review what is meant by "antiparallel" and the 5′ and 3′ ends of a nucleic acid chain, see Figure 16.7.) In the example in Figure 17.5, the nucleotide triplet ACC along the DNA template strand (written as 3′-ACC-5′) provides a template for 5′-UGG-3′ in the mRNA molecule. The mRNA nucleotide triplets are called **codons**, and they are customarily written in the 5′ → 3′ direction. In our example, UGG is the codon for the amino acid tryptophan (abbreviated Trp, or W). The term *codon* is also used for the DNA nucleotide triplets along the *nontemplate* strand. These codons are complementary to the template strand and thus identical in sequence to the mRNA (except for T's instead of U's). For this reason, the nontemplate DNA strand is often called the **coding strand**; by convention, the sequence of the coding strand is used when a gene's sequence is reported.

During translation, the sequence of codons along an mRNA molecule is decoded, or translated, into a sequence of amino acids making up a polypeptide chain. The codons are read by the translation machinery in the 5′ → 3′

▼ Figure 17.5 The triplet code. For each gene, one DNA strand functions as a template for transcription of RNAs, such as mRNA. The base-pairing rules for DNA synthesis also guide transcription, except that RNA is made with uracil (U) instead of thymine (T). During translation, the mRNA is read as a sequence of nucleotide triplets, called codons. Each codon specifies an amino acid to be added to the growing polypeptide chain. The mRNA is read in the 5′ → 3′ direction.



VISUAL SKILLS *By convention, the nontemplate strand, also called the coding strand, is used to represent a DNA sequence. Write the sequence of the mRNA strand and the nontemplate strand—in both cases reading from 5′ to 3′—and compare them. Why do you think this convention was adopted? (Hint: Why is this called the coding strand?)*

Mastering Biology MP3 Tutor: DNA to RNA to Protein

direction along the mRNA. Each codon specifies which one of the 20 amino acids will be incorporated at the corresponding position along a polypeptide. Because codons are nucleotide triplets, the number of nucleotides making up a genetic message must be three times the number of amino acids in the protein product. For example, 300 nucleotides along an mRNA strand code for 100 amino acids in the polypeptide it encodes.

Cracking the Code

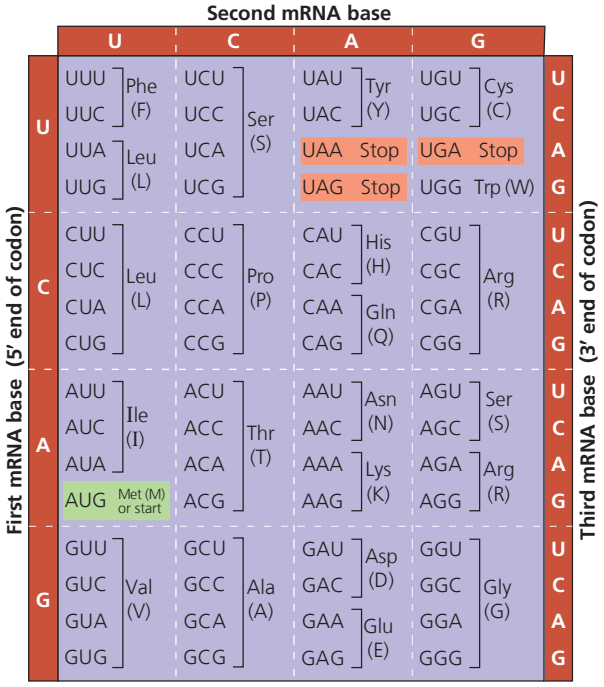
Molecular biologists cracked the genetic code of life in the early 1960s when a series of elegant experiments disclosed the amino acid translations of each of the mRNA codons. The first codon was deciphered in 1961 by Marshall Nirenberg, of the National Institutes of Health, and colleagues. They synthesized an artificial mRNA made up of only uracil-bearing nucleotides, thus containing only one codon (UUU) over and over. When added to a test-tube mixture that contained all 20 amino acids, ribosomes, and the other components required for protein synthesis, the "poly-U" mRNA was translated into a polypeptide containing many units of the phenylalanine (Phe, or F) amino acids, strung together as a polyphenylalanine chain. Thus, Nirenberg determined that the mRNA codon UUU specifies the amino acid phenylalanine. Soon, the amino acids specified by the codons AAA, GGG, and CCC were similarly identified.

Although more elaborate techniques were required to decode mixed triplets such as AUA and CGA, all 64 codons were deciphered by the mid-1960s. As **Figure 17.6** shows, 61 of the 64 triplets code for amino acids. The other three codons act as "stop" signals, or termination codons, marking the end of translation. Notice that the codon AUG has a dual function: It codes for the amino acid methionine (Met, or M) and also functions as a "start" signal, or initiation codon. Genetic messages usually begin with the mRNA codon AUG, which signals the protein-synthesizing machinery to begin translating the mRNA at that location. (Because AUG also stands for methionine, polypeptide chains all begin with methionine when they are synthesized. However, an enzyme may subsequently remove this starter amino acid from the chain.)

Notice in Figure 17.6 that there is redundancy in the genetic code, but no ambiguity. For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them ever specifies any other amino acid (no ambiguity). The redundancy in the code is not altogether random. In many cases, codons that are synonyms for a particular amino acid differ only in the third nucleotide base of the triplet. We will see why this is later in the chapter.

Our ability to extract the intended message from a written language depends on reading the symbols in the correct groupings—that is, in the correct **reading frame**. Consider this statement: "The red dog ate the bug." Group the letters incorrectly by starting at the wrong point, and the result will probably be gibberish: for example, "her edd oga tet heb ug." The reading frame is also important in the molecular language of cells. The short stretch of polypeptide shown in Figure 17.5, for instance, will be made correctly only if the mRNA nucleotides are read from left to right (5′ → 3′) in the groups of three shown in the figure: UGG UUU GGC UCA. Although a genetic message is written with no spaces between the codons, the cell's protein-synthesizing machinery reads the message as a series of nonoverlapping three-letter words. The message is *not* read as a series of overlapping words—UGGUUU, and so on—which would convey a very different message.

▼ Figure 17.6 The codon table for mRNA. The three nucleotide bases of an mRNA codon are designated here as the first, second, and third bases, reading in the 5′ → 3′ direction along the mRNA. The codon AUG not only stands for the amino acid methionine (Met, or M) but also functions as a "start" signal for ribosomes to begin translating the mRNA at that point. Three of the 64 codons function as "stop" signals, marking where ribosomes end translation. Both one- and three-letter codes are shown for the amino acids; see Figure 5.14 for their full names.



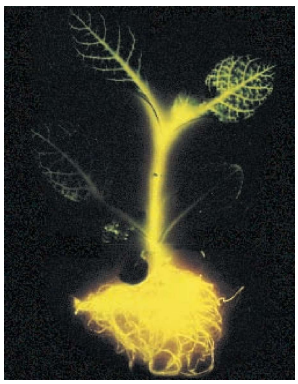
VISUAL SKILLS *A segment in the middle of an mRNA has the sequence 5′-AGAGAACCGCGA-3′. Using the codon table, translate this sequence, assuming the first three nucleotides are a codon.*

Mastering Biology **Animation: Translation: The Genetic Code**

Evolution of the Genetic Code

EVOLUTION The genetic code is nearly universal, shared by organisms from the simplest bacteria to the most complex plants and animals. The mRNA codon CCG, for instance, is translated as the amino acid proline in all organisms whose

▼ Figure 17.7 Evidence for evolution: expression of genes from different species. Because diverse forms of life share a common genetic code due to their shared ancestry, one species can be programmed to produce proteins characteristic of a second species by introducing DNA from the second species into the first.



(a) Tobacco plant expressing a firefly gene. The yellow glow is produced by a chemical reaction catalyzed by the protein product of the firefly gene.



(b) Mosquito larva expressing a jellyfish gene. The gene codes for green fluorescent protein (GFP) and is inserted into organisms as a *reporter gene* so researchers can tell if a gene of interest is being expressed.

Mastering Biology Video: GFP Transgenic Mice

genetic code has been examined. In laboratory experiments, genes can be transcribed and translated after being transplanted from one species to another, sometimes with quite striking results, as shown in **Figure 17.7**. Bacteria can be programmed by the insertion of human genes to synthesize certain human proteins for medical use, such as insulin. Such applications have produced many exciting developments in the area of biotechnology (see Concept 20.4).

The evolutionary significance of the code's near universality is clear. A language shared by all living things must have been operating very early in the history of life—early enough to be present in the common ancestor of all present-day organisms. A shared genetic vocabulary is a reminder of the kinship of all life.

CONCEPT CHECK 17.1

1. **MAKE CONNECTIONS** In a research article about alkaptonuria published in 1902, Garrod suggested that humans inherit two "characters" (alleles) for a particular enzyme and that both parents must contribute a faulty version for the offspring to have alkaptonuria. Today, would this disorder be called dominant or recessive? (See Concept 14.4.)
2. Describe the polypeptide product you would expect from a poly-G mRNA that is 30 nucleotides long.
3. **DRAW IT** The template strand of a gene contains the sequence 3′-TTCAGTCGT-5′. Suppose that the nontemplate sequence was transcribed instead of the template sequence. Draw the nontemplate sequence in 3′ to 5′ order. Then draw the mRNA sequence and translate it using Figure 17.6. Predict how well the protein synthesized from the nontemplate strand would function, if at all.

For suggested answers, see Appendix A.

CONCEPT 17.2

Transcription is the DNA-directed synthesis of RNA: *A Closer Look*

Now that we have considered the linguistic logic and evolutionary significance of the genetic code, we are ready to reexamine transcription, the first stage of gene expression, in greater detail.

Molecular Components of Transcription

Messenger RNA, the carrier of information from DNA to the cell's protein-synthesizing machinery, is transcribed from the template strand of a gene. An enzyme called an **RNA polymerase** pries the two strands of DNA apart and joins together RNA nucleotides complementary to the DNA template strand, thus elongating the RNA polynucleotide **(Figure 17.8)**. Like the DNA polymerases that function in DNA replication, RNA polymerases can assemble a polynucleotide only in its 5′ → 3′ direction, adding onto its 3′ end. Unlike DNA polymerases, however, RNA polymerases are able to start a chain from scratch; they don't need to add the first nucleotide onto a pre-existing primer.

Specific sequences of nucleotides along the DNA mark where transcription of a gene begins and ends. The DNA sequence where RNA polymerase attaches and initiates transcription is known as the **promoter**; in bacteria, the sequence that signals the end of transcription is called the **terminator**. (The termination mechanism in eukaryotes will be described later.) Molecular biologists refer to the direction of transcription as "downstream" and the other direction as "upstream." These terms are also used to describe the positions of nucleotide sequences within the DNA or RNA. Thus, the promoter sequence in DNA is said to be upstream from the terminator. The stretch of DNA downstream from the promoter that is transcribed into an RNA molecule is called a **transcription unit**.

Bacteria have a single type of RNA polymerase that synthesizes not only mRNA but also other types of RNA that function in gene expression, such as ribosomal RNA. In contrast, eukaryotes have at least three types of RNA polymerase in their nuclei; the one used for pre-mRNA synthesis is called RNA polymerase II. The other RNA polymerases transcribe RNA molecules that are not translated into protein. In the discussion that follows, we start with the features of mRNA synthesis common to both bacteria and eukaryotes and then describe some key differences.

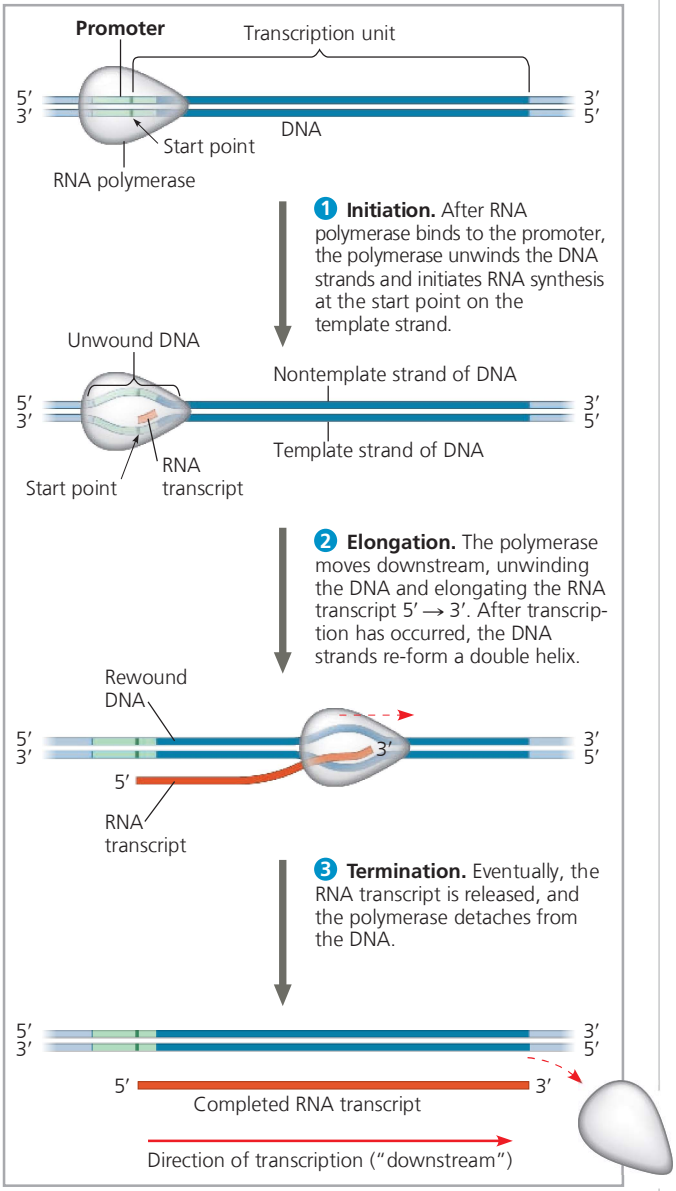
Synthesis of an RNA Transcript

The three stages of transcription, as shown in Figure 17.8 and described next, are initiation, elongation, and termination of the RNA chain. Study Figure 17.8 to familiarize yourself with the stages and the terms used to describe them.

RNA Polymerase Binding and Initiation of Transcription

The promoter sequence of a gene includes within it the transcription **start point**—the nucleotide where RNA polymerase actually begins synthesizing the mRNA—and typically extends several dozen or so nucleotide pairs upstream from the start point **(Figure 17.9)**. Based on interactions with proteins (transcription factors), RNA polymerase binds in a precise location and orientation on the promoter. This binding determines where transcription starts and the direction it will travel, thus which strand of DNA is used as the template.

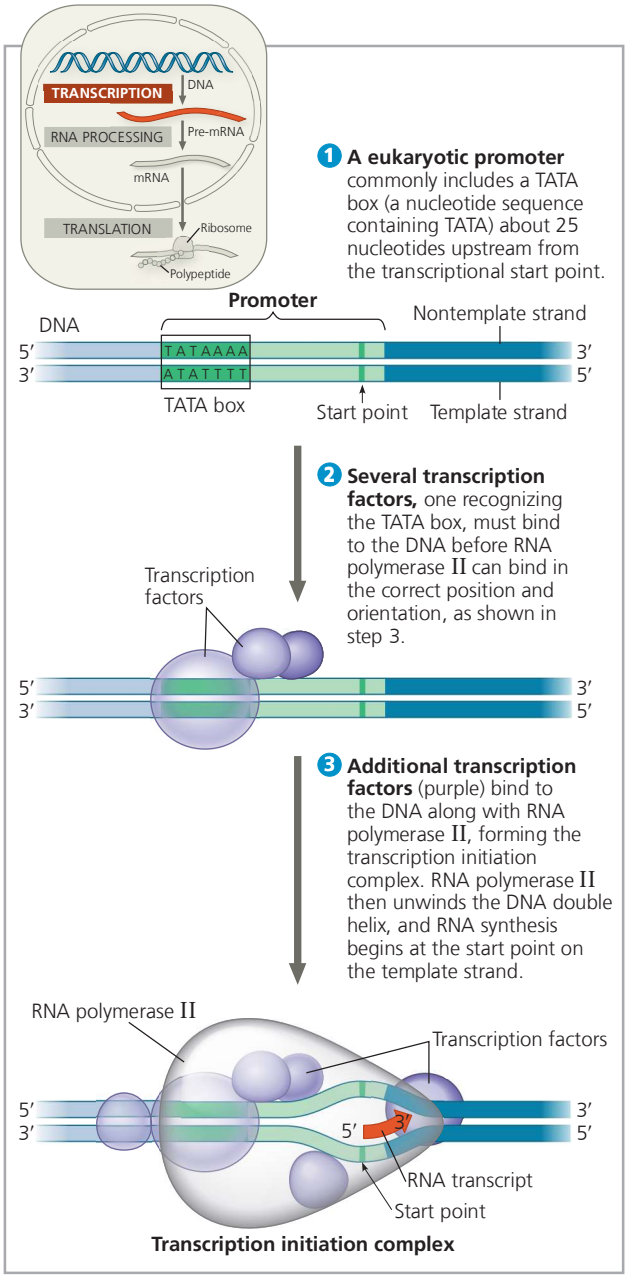
▼ Figure 17.8 The stages of transcription: initiation, elongation, and termination. This general depiction of transcription applies to both bacteria and eukaryotes, but the details of termination differ, as described in the text. Also, in a bacterium, the RNA transcript is immediately usable as mRNA; in a eukaryote, the RNA transcript must first undergo processing.



MAKE CONNECTIONS *Compare the use of a template strand during transcription and replication. See Figure 16.18.*

Mastering Biology **Animation: Overview of Transcription**
Animation: Overview of Transcription in Bacteria

▼ Figure 17.9 The initiation of transcription at a eukaryotic promoter. In eukaryotic cells, proteins called transcription factors mediate the initiation of transcription by RNA polymerase II.



? *Explain how the interaction of RNA polymerase with the promoter would differ if the figure showed transcription initiation for bacteria.*

Certain sections of a promoter sequence are especially important for binding RNA polymerase in a way that ensures that transcription will begin at the correct start point. In bacteria, part of the RNA polymerase itself specifically recognizes and binds to the promoter. In eukaryotes, a collection of proteins called **transcription factors** (purple proteins in Figure 17.9), help guide the binding of RNA polymerase and the initiation of transcription. Only after transcription factors are attached to the promoter does RNA polymerase II bind to it. The whole complex of transcription factors and RNA polymerase II bound to the promoter is called a **transcription initiation complex**. Figure 17.9 shows the role of transcription factors and a crucial promoter DNA sequence called the **TATA box** in forming the initiation complex at a eukaryotic promoter.

The interaction between eukaryotic RNA polymerase II and transcription factors is an example of the importance of protein-protein interactions in controlling eukaryotic transcription. Once the appropriate transcription factors are firmly attached to the promoter DNA and the polymerase is bound to them in the correct orientation on the DNA, the enzyme unwinds the two DNA strands and begins transcribing the template strand at the start point.

Elongation of the RNA Strand

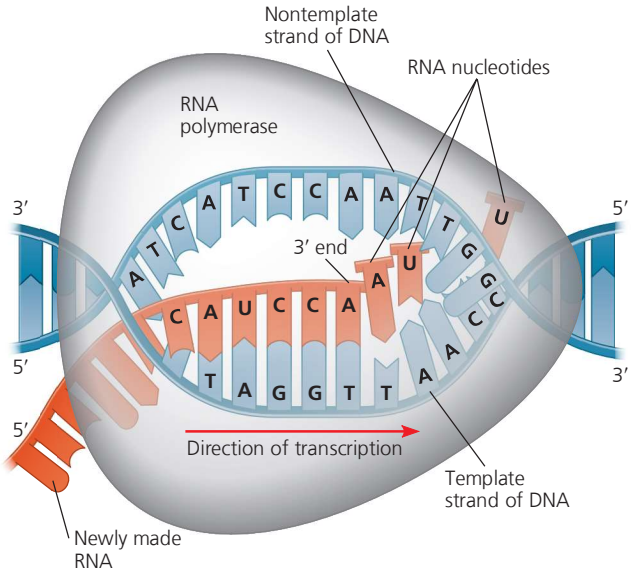
As RNA polymerase moves along the DNA, it untwists the double helix, exposing about 10–20 DNA nucleotides at a time for pairing with RNA nucleotides **(Figure 17.10)**. The enzyme adds nucleotides to the 3′ end of the growing RNA molecule as it moves along the double helix. As transcription proceeds forward, the newly synthesized RNA molecule behind the RNA polymerase peels away from its DNA template, and the DNA double helix re-forms. Transcription progresses at a rate of about 40 nucleotides per second in eukaryotes.

A single gene can be transcribed simultaneously by several molecules of RNA polymerase following each other like trucks in a convoy. A growing strand of RNA trails off from each polymerase, with the length of each new strand reflecting how far along the template the enzyme has traveled from the start point (see the mRNA molecules in Figure 17.23). The congregation of many polymerase molecules simultaneously transcribing a single gene increases the amount of mRNA transcribed from it, which helps the cell make the encoded protein in large amounts.

Termination of Transcription

Bacteria and eukaryotes differ in the way they terminate transcription. In bacteria, transcription proceeds through a terminator sequence in the DNA. The transcribed terminator (an RNA sequence) functions as the termination signal, causing the polymerase to detach from the DNA and release the transcript, which requires no further modification before translation. In eukaryotes, RNA polymerase II

▼ Figure 17.10 Transcription elongation. RNA polymerase moves along the DNA template strand, joining complementary RNA nucleotides to the 3′ end of the growing RNA transcript. Behind the polymerase, the new RNA peels away from the template strand, which re-forms a double helix with the nontemplate strand.



Mastering Biology BioFlix® Animation: Transcription
Animation: Elongation of the RNA Strand

transcribes a sequence on the DNA called the polyadenylation signal sequence, which specifies a polyadenylation signal (AAUAAA) in the pre-mRNA. This is called a "signal" because once this stretch of six RNA nucleotides appears, it is immediately bound by certain proteins in the nucleus. Then, at a point about 10–35 nucleotides downstream from the AAUAAA, these proteins cut the RNA transcript free from the polymerase, releasing the pre-mRNA. The pre-mRNA then undergoes processing, the topic of the next section. Although that cleavage marks the end of the mRNA, the RNA polymerase II continues to transcribe. Enzymes begin to degrade the RNA made after cleavage, starting at its newly exposed 5′ end. The polymerase continues transcribing, pursued by the enzymes, until they catch up to the polymerase and it dissociates from the DNA.

CONCEPT CHECK 17.2

1. What is a promoter? Is it located at the upstream or downstream end of a transcription unit?
2. What enables RNA polymerase to start transcribing a gene at the right place on the DNA in a bacterial cell? In a eukaryotic cell?
3. **WHAT IF?** Suppose X-rays caused a sequence change in the TATA box of a particular gene's promoter. How would that affect transcription of the gene? (See Figure 17.9.)

For suggested answers, see Appendix A.

CONCEPT 17.3

Eukaryotic cells modify RNA after transcription

Enzymes in the eukaryotic nucleus modify pre-mRNA in specific ways before the genetic message is dispatched to the cytoplasm. During this **RNA processing**, both ends of the primary transcript are altered. Also, in most cases, certain interior sections of the RNA molecule are cut out and the remaining parts spliced together. These modifications produce an mRNA molecule ready for translation.

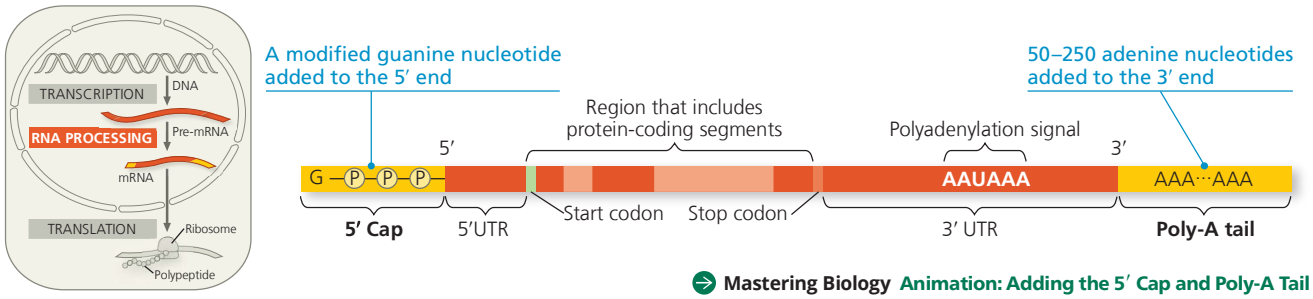
Alteration of mRNA Ends

Each end of a pre-mRNA molecule is modified in a particular way **(Figure 17.11)**. The 3′ end, which is synthesized first, receives a **5′ cap**, a modified form of a guanine (G) nucleotide added onto the 5′ end after transcription of the first 20–40 nucleotides have been transcribed. The 3′ end of the pre-mRNA molecule is also modified before the mRNA exits the nucleus. Recall that the pre-mRNA is cut and released soon after the polyadenylation signal, AAUAAA, is transcribed. At the 3′ end, an enzyme then adds 50–250 more adenine (A) nucleotides, forming a **poly-A tail**. The 5′ cap and poly-A tail share several important functions. First, they seem to facilitate the export of the mature mRNA from the nucleus. Second, they help protect the mRNA from degradation by hydrolytic enzymes. And third, they help ribosomes attach to the 5′ end of the mRNA once it reaches the cytoplasm. In addition to the cap and tail on a eukaryotic mRNA molecule, Figure 17.11 shows the untranslated regions (UTRs) at the 5′ and 3′ ends of the mRNA (referred to as the 5′ UTR and 3′ UTR). The UTRs are parts of the mRNA that will not be translated into protein, but they have other functions, such as ribosome binding.

Split Genes and RNA Splicing

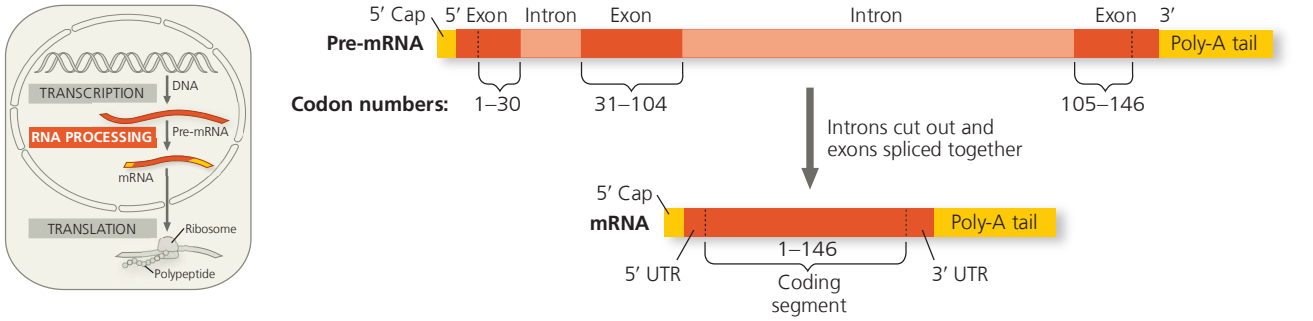
A remarkable stage of RNA processing in the eukaryotic nucleus is called **RNA splicing (Figure 17.12)**, in which large portions of the RNA primary transcript molecules are removed and the remaining portions are reconnected. This

▼ Figure 17.11 RNA processing: addition of the 5′ cap and poly-A tail. Enzymes modify the two ends of a eukaryotic pre-mRNA molecule. The modified ends may promote the export of mRNA from the nucleus, and they help protect the mRNA from degradation. When the mRNA reaches the cytoplasm, the modified ends, in conjunction with certain cytoplasmic proteins, help with ribosome attachment. The 5′ cap and poly-A tail are not translated into protein, nor are the regions called the 5′ untranslated region (5′ UTR) and 3′ untranslated region (3′ UTR). The pink segments are introns, which will be described shortly (see Figure 17.12).



Mastering Biology Animation: Adding the 5′ Cap and Poly-A Tail

▼ Figure 17.12 RNA processing: RNA splicing. The RNA molecule shown here codes for β-globin, one of the polypeptides of hemoglobin. The numbers under the RNA refer to codons; β-globin is 146 amino acids long. The β-globin gene and its pre-mRNA transcript have three exons, corresponding to sequences that will exit the nucleus as mRNA. (The 5′ UTR and 3′ UTR do not code for proteins, but they are parts of exons because they remain in the mRNA.) During RNA processing, the introns are cut out and the exons spliced together. In many genes, the introns are much longer than the exons.



DRAW IT *On the mRNA, indicate the sites of the start and stop codons.*

Mastering Biology BioFlix® Animation: RNA Processing

cut-and-paste job is similar to editing a movie. The average length of a transcription unit along a human DNA molecule is about 27,000 nucleotide pairs, so the primary RNA transcript is also that long. However, the average-sized protein of 400 amino acids requires only 1,200 nucleotides in RNA to code for it. (Remember, each amino acid is encoded by a *triplet* of nucleotides.) This is because most eukaryotic genes and their RNA transcripts have long noncoding stretches of nucleotides, regions that are not translated. Even more surprising is that most of these noncoding sequences are interspersed between coding segments of the gene and thus between coding segments of the pre-mRNA. In other words, the sequence of DNA nucleotides that codes for a eukaryotic polypeptide is usually not continuous; it is split into segments. The noncoding segments of nucleic acid that lie between coding regions are called intervening sequences, or **introns**. The other regions are called **exons**, because they are eventually expressed, usually by being translated into amino acid sequences. (Exceptions include the UTRs of the exons at the ends of the RNA, which make up part of the mRNA but are not translated into protein. Because of these exceptions, you may prefer to think of exons as sequences of RNA that *ex*it the nucleus.) The terms *intron* and *exon* are used for both RNA sequences and the DNA sequences that specify them.

In making a primary transcript from a gene, RNA polymerase II transcribes both introns and exons from the DNA, but the mRNA molecule that enters the cytoplasm is an abridged version. In the process of RNA splicing, the introns are cut out from the molecule and the exons joined together, forming an mRNA molecule with a continuous coding sequence.

How is pre-mRNA splicing carried out? The removal of introns is accomplished by a large complex made of proteins and small RNAs called a **spliceosome**. This complex binds to several short nucleotide sequences along an intron, including key sequences at each end **(Figure 17.13)**. The intron is then released (and rapidly degraded), and the spliceosome joins together the two exons that flanked the intron. The small RNAs in the spliceosome not only participate in spliceosome assembly and splice site recognition, but also catalyze the splicing reaction; like proteins, RNAs can act as catalysts.

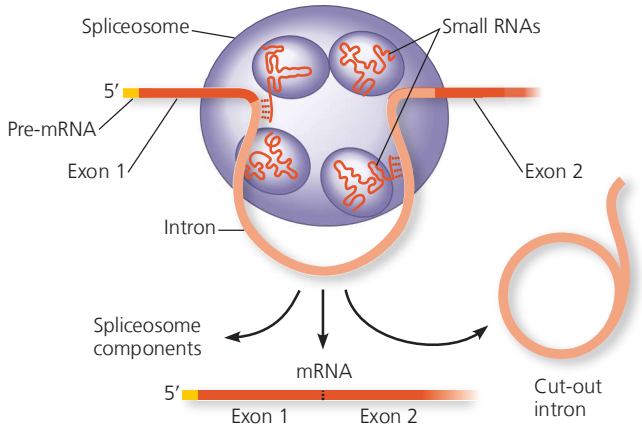
Mastering Biology
Interview with Joan Steitz: Studying RNA, her "favorite molecule"



Ribozymes

The idea of a catalytic role for the RNAs in the spliceosome arose from the discovery of **ribozymes**, RNA molecules that function as enzymes. In some organisms, RNA splicing can occur without proteins or even additional RNA molecules: The intron RNA functions as a ribozyme and catalyzes its own removal. For example, in the ciliate protist *Tetrahymena*, self-splicing occurs in the production of ribosomal RNA (rRNA),

▼ Figure 17.13 A spliceosome splicing a pre-mRNA. The diagram shows a portion of a pre-mRNA transcript, with an intron (pink) flanked by two exons (red). Small RNAs within the spliceosome base-pair with nucleotides at specific sites along the intron. Next, small spliceosome RNAs catalyze cutting of the pre-mRNA and the splicing together of the exons, releasing the intron for rapid degradation.



Mastering Biology **Animation: A Spliceosome**

a component of the organism's ribosomes. The pre-rRNA actually removes its own introns! The discovery of ribozymes invalidated the idea that all biological catalysts are proteins.

Three properties of RNA enable some RNA molecules to function as enzymes. First, because RNA is single-stranded, a region of an RNA molecule may base-pair, in an antiparallel arrangement, with a complementary region elsewhere in the same molecule; this gives the molecule a particular three-dimensional structure. A specific structure is essential to the catalytic function of ribozymes, just as it is for enzymatic proteins. Second, like certain amino acids in an enzymatic protein, some of the bases in RNA contain functional groups that can participate in catalysis. Third, the ability of RNA to hydrogen-bond with other nucleic acid molecules (either RNA or DNA) adds specificity to its catalytic activity. For example, complementary base pairing between the RNA of the spliceosome and the RNA of a primary RNA transcript precisely locates the region where the ribozyme catalyzes splicing. Later in this chapter, you will see how these properties of RNA also allow it to perform important noncatalytic roles in the cell, such as recognition of the three-nucleotide codons on mRNA.

The Functional and Evolutionary Importance of Introns

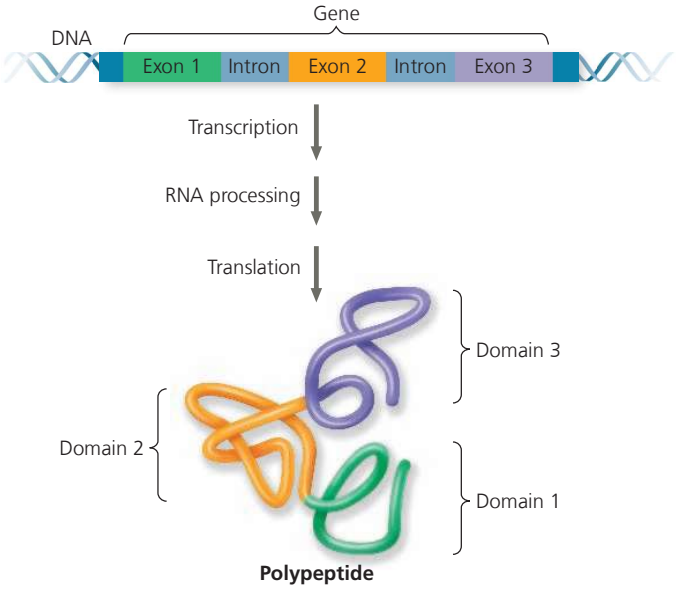
EVOLUTION Whether or not RNA splicing and the presence of introns have provided selective advantages during evolutionary history is a matter of some debate. In any case, it is informative to consider their possible adaptive benefits. Specific functions have not been identified for most introns, but at least some contain sequences that regulate gene expression, and many affect gene products.

One important consequence of the presence of introns in genes is that a single gene can encode more than one kind of polypeptide. Many genes are known to give rise to two or more different polypeptides, depending on which segments are treated as exons during RNA processing; this is called **alternative RNA splicing** (see Figure 18.14). Results from the Human Genome Project (discussed in Concept 21.1) suggest that alternative RNA splicing is one reason humans can get along with about the same number of genes as a nematode (roundworm). Because of alternative splicing, the number of different protein products an organism produces can be much greater than its number of genes.

Proteins often have a modular architecture consisting of discrete structural and functional regions called **domains**. One domain of an enzyme, for example, might include the active site, while another might allow the enzyme to bind to a cellular membrane. In quite a few cases, different exons code for the different domains of a protein **(Figure 17.14)**.

The presence of introns in a gene may facilitate the evolution of new and potentially beneficial proteins as a result of a process known as *exon shuffling* (see Figure 21.16). Introns increase the probability of crossing over between the exons of alleles of a gene—simply by providing more terrain for crossovers without interrupting coding sequences. This might result in new combinations of exons and proteins with altered structure and function. We can also imagine the occasional mixing and matching of exons between completely different (nonallelic) genes. Exon shuffling of either sort could lead to new proteins with novel combinations of functions. While most of the shuffling would result in nonbeneficial changes, occasionally a beneficial variant might arise.

▼ Figure 17.14 Correspondence between exons and protein domains.



CONCEPT CHECK 17.3

1. Given that there are about 20,000 human protein-coding genes, how can human cells make 75,000–100,000 different proteins?
2. Compare RNA splicing to how you would watch a prerecorded television show. What would introns correspond to in this analogy?
3. **WHAT IF?** What would be the effect of treating cells with an agent that removed the 5′ cap from mRNAs?

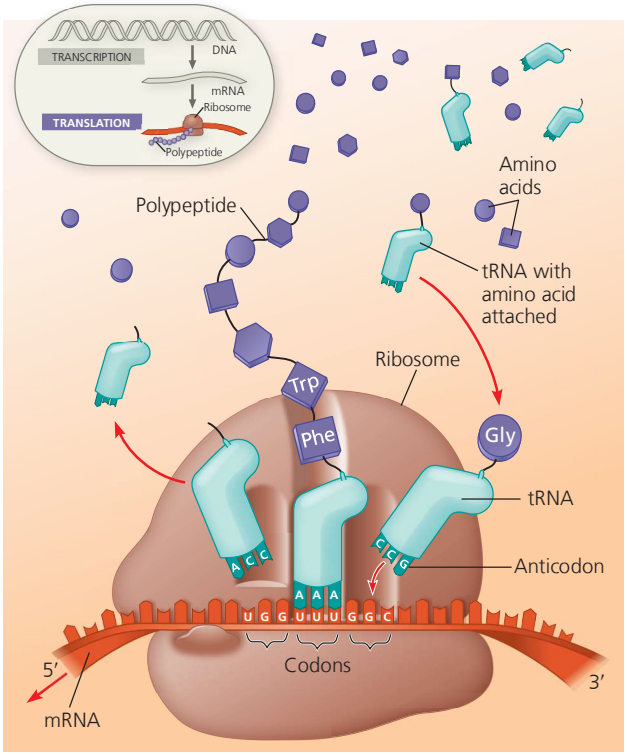
For suggested answers, see Appendix A.

CONCEPT 17.4

Translation is the RNA-directed synthesis of a polypeptide: *A Closer Look*

We will now examine how genetic information flows from mRNA to protein—the process of translation **(Figure 17.15)**. We'll focus on the basic steps of translation that occur in both bacteria and eukaryotes, while pointing out key differences.

▼ Figure 17.15 Translation: the basic concept. As a molecule of mRNA is moved through a ribosome, codons are translated into amino acids, one by one. The translators, or interpreters, are tRNA molecules, each type with a specific anticodon at one end and a corresponding amino acid at the other end. A tRNA adds its amino acid cargo to a growing polypeptide chain when the anticodon hydrogen-bonds to the complementary codon on the mRNA.



Mastering Biology **BioFlix® Animation: Translation**
Animation: Overview of Translation

Molecular Components of Translation

In the process of translation, a cell "reads" a genetic message and builds a polypeptide accordingly. The message is a series of codons along an mRNA molecule, and the translator is called a **transfer RNA (tRNA)**. The function of a tRNA is to transfer an amino acid from the cytoplasmic pool of amino acids to a growing polypeptide in a ribosome. A cell keeps its cytoplasm stocked with all 20 amino acids, either by synthesizing them from other compounds or by taking them up from the surrounding solution. The ribosome, a structure made of proteins and RNAs, adds each amino acid brought to it by a tRNA to the growing end of a polypeptide chain (see Figure 17.15).

Translation is simple in principle but complex in its biochemistry and mechanics, especially in the eukaryotic cell. In dissecting translation, we'll focus on the slightly less complicated version of the process that occurs in bacteria. We'll first look at the major players in this process.

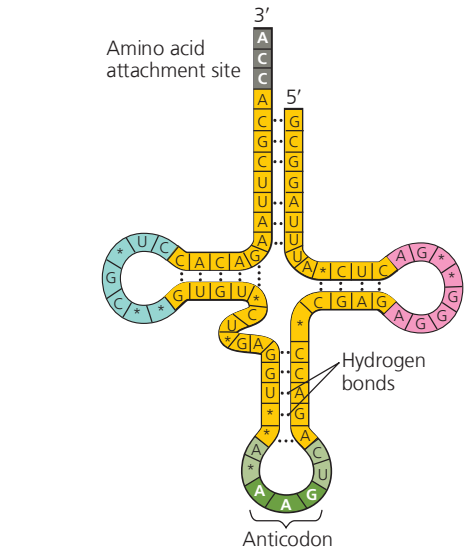
The Structure and Function of Transfer RNA

The key to translating a genetic message into a specific amino acid sequence is the fact that each tRNA molecule enables translation of a given mRNA codon into a certain amino acid. This is possible because a tRNA bears a specific amino acid at one end of its three-dimensional structure, while at the other end is a nucleotide triplet that can base-pair with the complementary codon on mRNA.

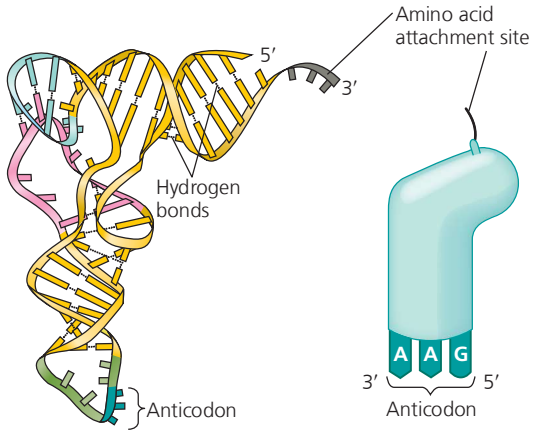
A tRNA molecule consists of a single RNA strand that is only about 80 nucleotides long (compared to hundreds of nucleotides for most mRNA molecules). The presence of complementary stretches of nucleotide bases that can hydrogen-bond to each other allows this single strand to fold back on itself and form a molecule with a particular 3-D structure. Flattened into one plane to clarify this base pairing, a tRNA molecule looks like a cloverleaf **(Figure 17.16a)**. The tRNA actually twists and folds into a compact 3-D structure that is roughly L-shaped **(Figure 17.16b)**, with the 5′ and 3′ ends of the linear tRNA both located near one end of the structure. The protruding 3′ end acts as the attachment site for an amino acid. The loop extending from the other end of the L includes the **anticodon**, the particular nucleotide triplet that base-pairs to a specific mRNA codon. Thus, the structure of a tRNA molecule fits its function.

Anticodons are conventionally written 3′ → 5′ to align properly with codons written 5′ → 3′ (see Figure 17.15). (For base pairing, RNA strands must be antiparallel, like DNA.) As an example of how tRNAs work, consider the mRNA codon 5′-GGC-3′, which is translated as the amino acid glycine. The tRNA that base-pairs with this codon by hydrogen bonding has 3′-CCG-5′ as its anticodon and carries glycine at its other end (see the incoming tRNA approaching the ribosome in Figure 17.15). As an mRNA molecule is moved through

▼ Figure 17.16 The structure of transfer RNA (tRNA).



(a) Two-dimensional structure. The four base-paired regions and three loops are characteristic of all tRNAs, as is the base sequence of the amino acid attachment site at the 3′ end. The anticodon triplet is unique to each tRNA type, as are some sequences in the other two loops. (The asterisks mark bases that have been chemically modified, a characteristic of tRNA. The modified bases contribute to tRNA function in a way that is not yet understood.)



(b) Three-dimensional structure

(c) Symbol used in this book

VISUAL SKILLS *Look at the tRNA shown in this figure. Based on its anticodon, identify the codon it would bind to, as well as the amino acid that it would carry.*

Mastering Biology **HHMI Video: RNA Folding**

a ribosome, glycine will be added to the polypeptide chain whenever the codon 5′-GGC-3′ is presented for translation. Codon by codon, the genetic message is translated as tRNAs position each amino acid in the order prescribed, and the ribosome adds that amino acid onto the growing polypeptide

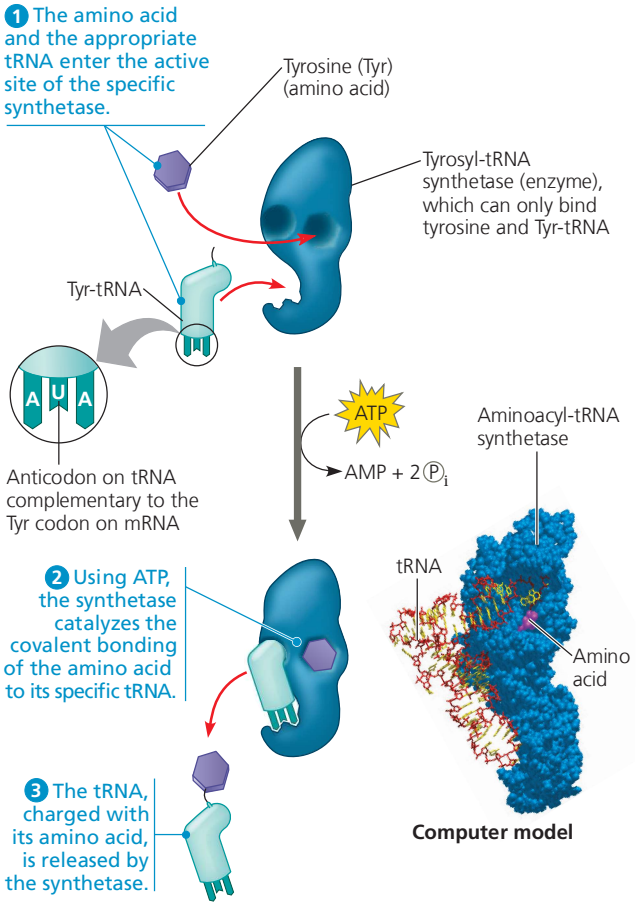
chain. The tRNA molecule is a translator in the sense that, in the context of the ribosome, it can read a nucleic acid word (the mRNA codon) and interpret it as a protein word (the amino acid).

Like mRNA and other types of cellular RNA, transfer RNA molecules are transcribed from DNA templates. In a eukaryotic cell, tRNA, like mRNA, is made in the nucleus and then travels to the cytoplasm, where it will participate in the process of translation. In both bacterial and eukaryotic cells, each tRNA molecule is used repeatedly, picking up its designated amino acid in the cytosol, depositing this cargo onto a polypeptide chain at the ribosome, and then leaving the ribosome, ready to pick up another of the same amino acid.

The accurate translation of a genetic message requires two instances of molecular recognition. First, a tRNA that binds to an mRNA codon specifying a particular amino acid must carry that amino acid, and no other, to the ribosome. The correct matching up of tRNA and amino acid is carried out by a family of related enzymes that are aptly named **aminoacyl-tRNA synthetases** **(Figure 17.17)**. The active site of each type of aminoacyl-tRNA synthetase fits only a specific combination of amino acid and tRNA. (Regions of both the amino acid attachment end and the anticodon end of the tRNA ensure the specific fit.) There are 20 different synthetases, one for each amino acid. A synthetase joins a given amino acid to an appropriate tRNA; one synthetase is able to bind to all the different tRNAs for its particular amino acid. The synthetase catalyzes the covalent attachment of the amino acid to its tRNA in a process driven by the hydrolysis of ATP. The resulting aminoacyl tRNA, also called a *charged tRNA*, is released from the enzyme and is then available to deliver its amino acid to a growing polypeptide chain on a ribosome.

The second instance of molecular recognition is the pairing of the tRNA anticodon with the appropriate mRNA codon. If one tRNA type existed for each mRNA codon specifying an amino acid, there would be 61 tRNAs (see Figure 17.6). In bacteria, however, there are only about 45 tRNAs, signifying that some tRNAs must be able to bind to more than one codon. This is possible because the rules for base pairing between the third nucleotide base of a codon and the corresponding base of a tRNA anticodon are relaxed compared to those at other codon positions. For example, the nucleotide base U at the 5′ end of a tRNA anticodon can pair with either A or G in the third position (at the 3′ end) of an mRNA codon. The flexible base pairing at this codon position is called **wobble**. Wobble explains why the synonymous codons for a given amino acid most often differ in their third nucleotide base. Accordingly, a tRNA with the anticodon 3′-UCU-5′ can base-pair with either the mRNA codon 5′-AGA-3′ or 5′-AGG-3′, both of which code for arginine (see Figure 17.6).

▼ Figure 17.17 Aminoacyl-tRNA synthetases provide specificity in joining amino acids to their tRNAs. Linkage of a tRNA to its amino acid is an endergonic process that occurs at the expense of ATP, which loses two phosphate groups, becoming AMP.



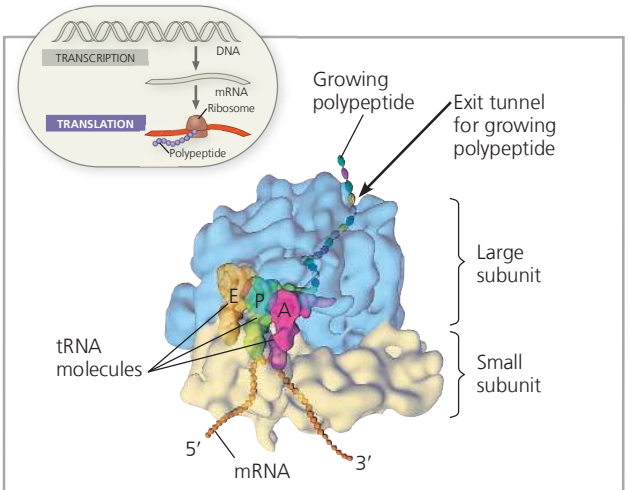
Mastering Biology **Animation: Transfer RNA**

The Structure and Function of Ribosomes

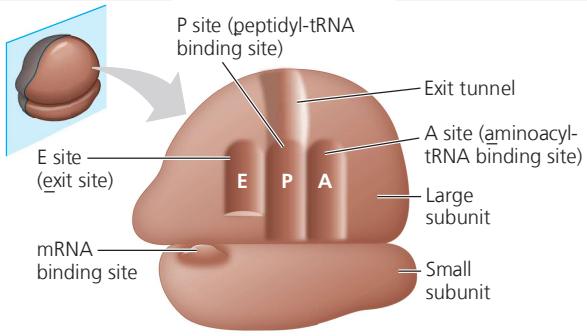
Although the ribosomes of bacteria and eukaryotes are very similar in structure and function, eukaryotic ribosomes are slightly larger and differ somewhat from bacterial ribosomes in their molecular composition. The differences are medically significant. Certain antibiotic drugs can inactivate bacterial ribosomes without affecting the ability of eukaryotic ribosomes to make proteins. These drugs, including tetracycline and streptomycin, are used to combat bacterial infections.

Ribosomes facilitate the specific coupling of tRNA anticodons with mRNA codons during protein synthesis. A ribosome consists of a large subunit and a small subunit, each made up of proteins and one or more **ribosomal RNAs**, or **rRNAs** **(Figure 17.18)**. In eukaryotes, the subunits are made in the nucleolus. Ribosomal RNA genes are transcribed, and the RNA is processed and assembled with proteins imported from the

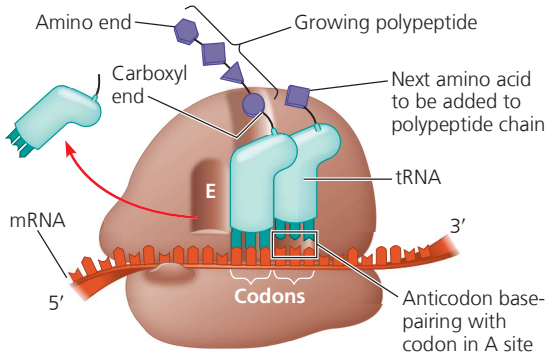
▼ Figure 17.18 The anatomy of a functioning ribosome.



(a) Computer model of functioning ribosome. This is a model of a bacterial ribosome, showing its overall shape. The eukaryotic ribosome is roughly similar. A ribosomal subunit is a complex of ribosomal RNA molecules and proteins.



(b) Schematic model showing binding sites. A ribosome has an mRNA binding site and three tRNA binding sites, known as the A, P, and E sites. This schematic ribosome will appear in later diagrams.



(c) Schematic model with mRNA and tRNA. A tRNA fits into the A site when its anticodon base-pairs with an mRNA codon. The P site holds the tRNA attached to the growing polypeptide. The A site holds the tRNA carrying the next amino acid to be added to the polypeptide chain. Discharged tRNAs leave from the E site. The polypeptide grows at its carboxyl end.

Mastering Biology Animation: Ribosomes

cytoplasm. Completed ribosomal subunits are then exported via nuclear pores to the cytoplasm. In both bacteria and eukaryotes, a large and a small subunit join to form a functional ribosome only when attached to an mRNA molecule. About one-third of the mass of a ribosome is made up of proteins; the rest consists of three rRNA molecules (in bacteria) or four (in eukaryotes). Because most cells contain thousands of ribosomes, rRNA is the most abundant type of cellular RNA.

The structure of a ribosome reflects its function of bringing an mRNA molecule together with tRNAs carrying amino acids. The mRNA itself has a binding site for the ribosome. (In the **Scientific Skills Exercise**, you can work with DNA sequences representing this binding site in a group of *Escherichia coli* genes.) The ribosome, in turn, has a binding site for mRNA, as well as three binding sites for tRNA (see Figure 17.18). The **P site** (peptidyl-tRNA binding site) holds the tRNA carrying the growing polypeptide chain, while the **A site** (aminoacyl-tRNA binding site) holds the tRNA carrying the next amino acid to be added to the chain. Discharged tRNAs leave the ribosome from the **E site** (exit site). The ribosome holds the tRNA and mRNA in close proximity and positions the new amino acid so that it can be added to the carboxyl end of the growing polypeptide. It then catalyzes the formation of the peptide bond. As the polypeptide becomes longer, it passes through an *exit tunnel* in the ribosome's large subunit. When the polypeptide is complete, it is released through the exit tunnel.

The widely accepted model is that rRNAs, rather than ribosomal proteins, are primarily responsible for both the structure and the function of the ribosome. The proteins, which are largely on the exterior, support the shape changes of the rRNA molecules as they carry out catalysis during translation. Ribosomal RNA is the main constituent of the A and P sites and of the interface between the two subunits; it also acts as the catalyst of peptide bond formation. Thus, a ribosome could actually be considered one colossal ribozyme!

Building a Polypeptide

We can divide translation, the synthesis of a polypeptide, into three stages: initiation, elongation, and termination. All three require protein "factors" that aid in the translation process. Some steps of initiation and elongation also require energy, provided by the hydrolysis of guanosine triphosphate (GTP).

Ribosome Association and Initiation of Translation

The initiation stage of translation brings together an mRNA, a tRNA bearing the first amino acid of the polypeptide, and the two subunits of a ribosome. First, a small ribosomal subunit binds to both the mRNA and a specific initiator tRNA, which carries the amino acid methionine. In bacteria, the small subunit can bind the two in either order; it binds the mRNA at a specific RNA sequence, just upstream of the AUG start codon. In eukaryotes, the small subunit, with the initiator tRNA already bound, binds to the 5′ cap of the mRNA and then moves, or *scans*,

Scientific Skills Exercise

Interpreting a Sequence Logo

How Can a Sequence Logo Be Used to Identify Ribosome Binding Sites on Bacterial mRNAs? When initiating translation, ribosomes bind to an mRNA at a ribosome binding site upstream of the AUG start codon. Because mRNAs from different genes all bind to a ribosome, the genes encoding these mRNAs are likely to have a similar base sequence where the ribosomes bind. Therefore, candidate ribosome binding sites on mRNA can be identified by comparing DNA sequences (and thus the mRNA sequences) of multiple genes in a species, searching the region upstream of the start codon for shared (conserved) stretches of bases. In this exercise, you will analyze DNA sequences from multiple such genes, represented by a visual graphic called a sequence logo.

How the Experiment Was Done The DNA sequences of 149 genes from the *E. coli* genome were aligned using computer software. The aim was to identify similar base sequences—at the appropriate location in each gene—as potential ribosome binding sites. Rather than presenting the data as a series of 149 sequences aligned in a column (a sequence alignment), the researchers used a sequence logo.

Data from the Experiment To show how sequence logos are made, the potential ribosome binding regions from ten *E. coli* genes are shown below in a sequence alignment, followed by the sequence logo derived from the aligned sequences. Note that the DNA shown is the nontemplate (coding) strand, which is how DNA sequences are typically presented. (All data from Thomas D. Schneider.)

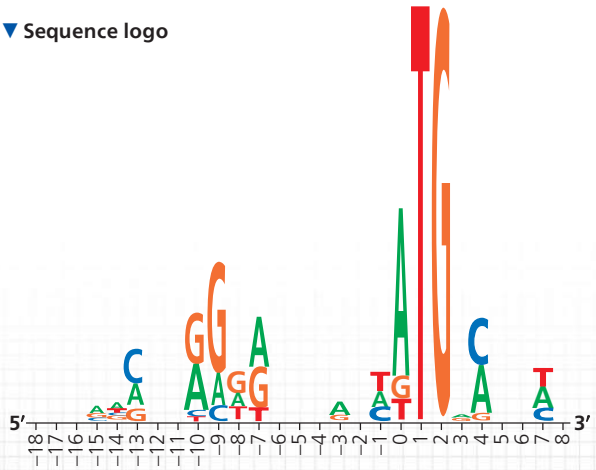
thrA	G	G	T	A	A	C	G	A	G	G	T	A	A	C	A	A	C	C	A	T	G	C	G	A	G	T	G
lacA	C	A	T	A	A	C	G	G	A	G	T	G	A	T	C	G	C	A	T	T	G	A	A	C	A	T	G
lacY	C	G	C	G	T	A	A	G	G	A	A	A	T	C	C	A	T	T	A	T	G	T	A	C	T	A	T
lacZ	T	T	C	A	C	A	C	A	G	G	A	A	A	C	A	G	C	T	A	T	G	A	C	C	A	T	G
lacI	C	A	A	T	T	C	A	G	G	G	T	G	G	T	G	A	A	T	G	T	G	A	A	A	C	C	A
recA	G	G	C	A	T	G	A	C	A	G	G	A	G	T	A	A	A	A	A	T	G	G	C	T	A	T	C
galR	A	C	C	C	A	C	T	A	A	G	G	T	A	T	T	T	T	C	A	T	G	G	C	G	A	C	C
metJ	A	A	G	A	G	G	A	T	T	A	A	G	T	A	T	C	T	C	A	T	G	G	C	T	G	A	A
lexA	A	T	A	C	A	C	C	C	A	G	G	G	G	G	C	G	G	A	A	T	G	A	A	A	G	C	G
trpR	T	A	A	C	A	A	T	G	G	C	G	A	C	A	T	A	T	T	A	T	G	G	C	C	C	A	A

5′ … 3′

−18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8

▲ **Sequence alignment**

▼ **Sequence logo**

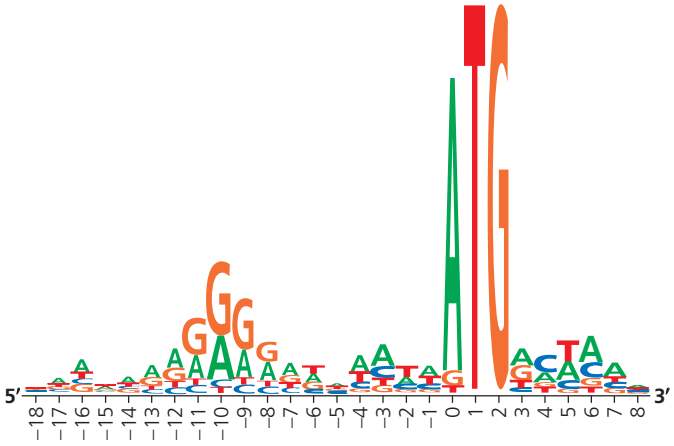


Further Reading T. D. Schneider and R. M. Stephens, Sequence logos: A new way to display consensus sequences, *Nucleic Acids Research* 18:6097–6100 (1990).



INTERPRET THE DATA

1. In the sequence logo, the horizontal axis shows the primary sequence of the DNA by nucleotide position. Letters for each base are stacked on top of each other according to their relative frequency at that position among the aligned sequences, with the most common base as the largest letter at the top of the stack. The height of each letter represents the relative frequency of that base *at that position*. **(a)** In the sequence alignment, count the number of each base at position –9 and order them from most to least frequent. Compare this to the size and placement of each base at –9 in the logo. **(b)** Do the same for positions 0 and 1.
2. The height of a stack of letters in a logo indicates the predictive power of that stack (determined statistically). If the stack is tall, we can be more confident in predicting what base will be in that position if a new sequence is added to the logo. For example, at position 2 in the sequence alignment, all ten sequences have a G; the probability of finding a G there in a new sequence is very high, as is the stack in the sequence logo. For short stacks, the bases all have about the same frequency, so it's hard to predict a base at those positions. **(a)** Looking at the sequence logo, which two positions have the most predictable bases? What bases do you predict would be at those positions in a newly sequenced gene? **(b)** Which 12 positions have the least predictable bases? How do you know? How does this reflect the relative frequencies of the bases shown at these positions in the sequence alignment? Use the two leftmost positions of the 12 as examples in your answer.
3. In the actual experiment, the researchers used 149 sequences to build their sequence logo, which is shown below. There is a stack at each position, even if short, because the sequence logo includes more data. **(a)** Which three positions in this sequence logo have the most predictable bases? Name the most frequent base at each. **(b)** Which four positions have the least predictable bases? How can you tell?



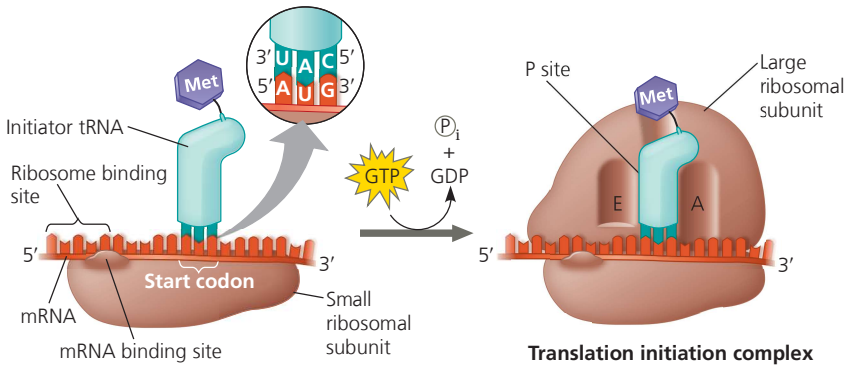
4. A consensus sequence identifies the base occurring most often at each position in the set of sequences. **(a)** Write out the consensus sequence of this (the nontemplate) strand. In any position where the base can't be determined, put a dash. **(b)** Which provides more information—the consensus sequence or the sequence logo? What is lost in the less informative method?
5. **(a)** Based on the logo, what five adjacent base positions in the 5′ UTR region are most likely to be involved in ribosome binding? Explain. **(b)** What is represented by the bases in positions 0–2?

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.

downstream along the mRNA until it reaches the AUG start codon, where the initiator tRNA then hydrogen-bonds, as shown in step 1 of **Figure 17.19**. In either bacteria or eukaryotes, the AUG signals the start of translation; this is important because it establishes the codon reading frame for the mRNA.

The union of mRNA, initiator tRNA, and the small ribosomal subunit is followed by the attachment of a large ribosomal subunit, completing the *translation initiation complex* (see Figure 17.19). Proteins called *initiation factors* are required to bring all these components together. The cell also expends energy obtained by hydrolysis of a GTP molecule to form the initiation complex. At the completion of the initiation process, the initiator tRNA sits in the P site of the ribosome, and the vacant A site is ready for the next aminoacyl tRNA. Note that a polypeptide is always synthesized in one direction, from the initial methionine at the amino end, also called the N-terminus, toward the final amino acid at the carboxyl end, also called the C-terminus (see Figure 5.15).

▼ Figure 17.19 The initiation of translation.



1 A small ribosomal subunit binds to an mRNA. In a bacterial cell, the mRNA binding site on this subunit recognizes a specific nucleotide sequence on the mRNA just upstream of the start codon (see the Scientific Skills Exercise). An initiator tRNA, with the anticodon UAC, base-pairs with the start codon, AUG. This tRNA carries the amino acid methionine (Met).

2 The arrival of a large ribosomal subunit completes the initiation complex. Proteins called initiation factors (not shown) are required to bring all the translation components together. Hydrolysis of GTP provides the energy for the assembly. The initiator tRNA is in the P site; the A site is available to the tRNA bearing the next amino acid.

Mastering Biology Animation: Initiation of Translation

Elongation of the Polypeptide Chain

In the elongation stage of translation, amino acids are added one by one to the previous amino acid at the C-terminus of the growing chain. Each addition involves several proteins called *elongation factors* and occurs in a three-step cycle described in **Figure 17.20**. Energy expenditure occurs in the first and third steps. Codon recognition requires hydrolysis of one molecule of GTP, which increases the accuracy and efficiency of this step. One more GTP is hydrolyzed (broken up) to provide energy for the translocation step.

The mRNA is moved through the ribosome in one direction only, 5′ end first; this is equivalent to the ribosome moving 5′ → 3′ on the mRNA. The main point is that the ribosome and the mRNA move relative to each other, unidirectionally, codon by codon. The elongation cycle takes less than a tenth of a second in bacteria and is repeated as each amino acid is added until the polypeptide is completed. The empty tRNAs that are released from the E site return to the cytoplasm, where they will be reloaded with the appropriate amino acid (see Figure 17.17).

Termination of Translation

The final stage of translation is termination **(Figure 17.21)**. Elongation continues until a stop codon in the mRNA reaches the A site. The nucleotide base triplets UAG, UAA, and UGA (all written 5′ → 3′) do not code for amino acids but instead act as signals to stop translation. A *release factor*, a protein shaped like an aminoacyl tRNA, binds directly to the stop codon in the A site. The release factor causes the addition of a water molecule instead of an amino acid to the polypeptide chain. (Water molecules are abundant in the cytosol.) This reaction hydrolyzes (breaks) the bond between the completed polypeptide and the tRNA in the P site, releasing the polypeptide through the exit tunnel of the ribosome's large subunit. The remainder of the translation assembly then comes apart in a multistep process, aided by other protein factors. Breakdown of the translation assembly requires the hydrolysis of two more GTP molecules.

Completing and Targeting the Functional Protein

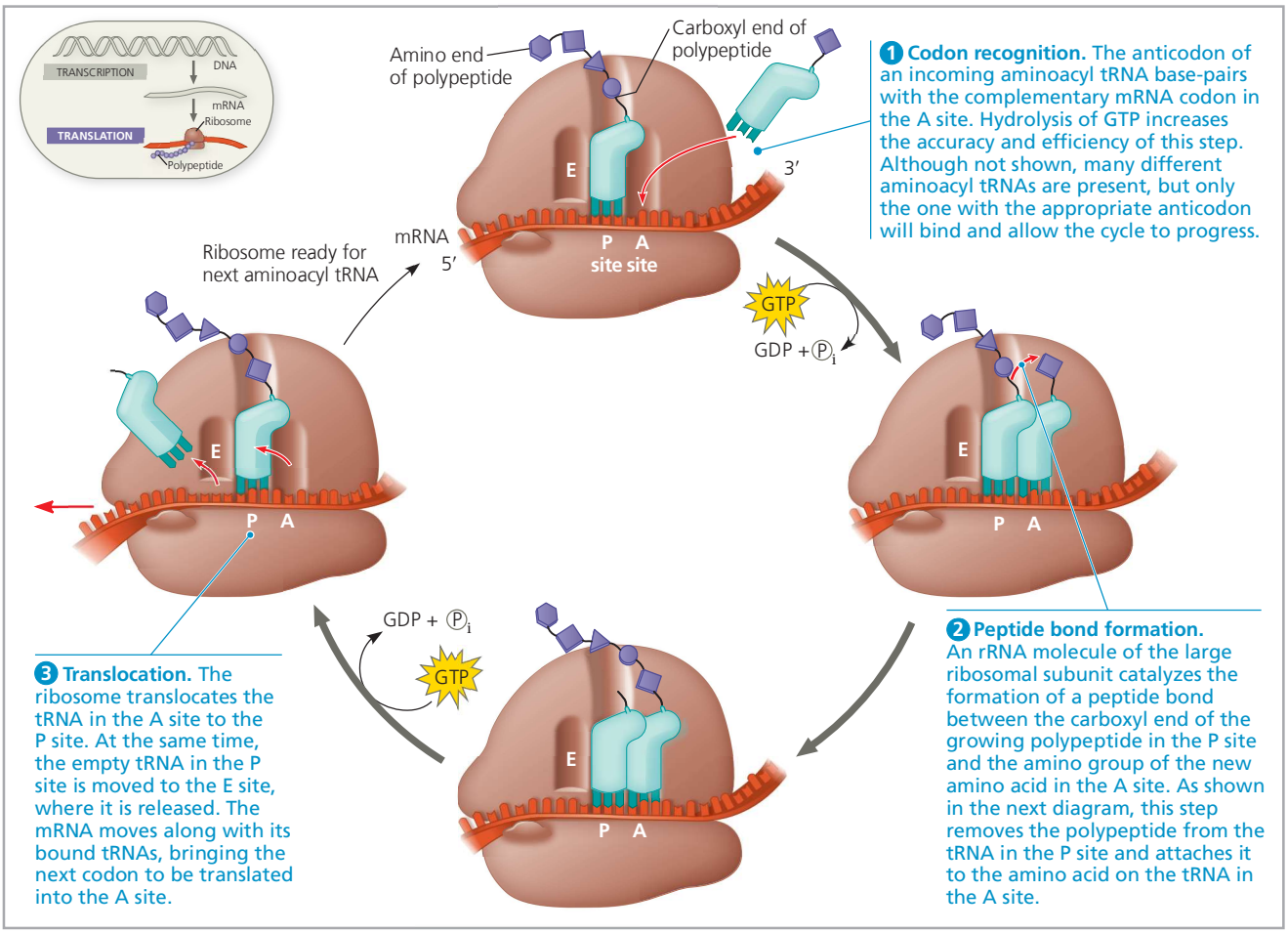
The process of translation is often not sufficient to make a functional protein. In this section, you will learn about modifications that polypeptide chains undergo after the translation process as well as some of the mechanisms used to target completed proteins to specific sites in the cell.

Protein Folding and Post-translational Modifications

During its synthesis, a polypeptide chain begins to coil and fold spontaneously as a consequence of its amino acid sequence (primary structure), forming a protein with a specific shape: a three-dimensional molecule with secondary and tertiary structure (see Figure 5.18). Thus, a gene determines primary structure, which in turn determines shape.

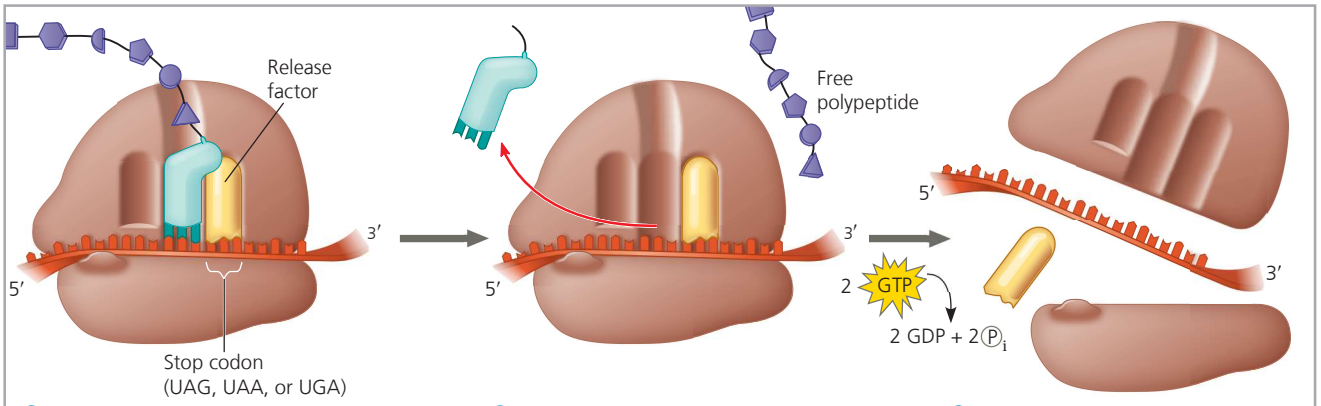
▼ Figure 17.20 The elongation cycle of translation. The hydrolysis of GTP plays an important role in the elongation process; elongation factors are not shown.

Mastering Biology **Animation: Elongation Cycle of Translation**



▼ Figure 17.21 The termination of translation. Like elongation, termination requires GTP hydrolysis as well as additional protein factors, which are not shown here.

Mastering Biology **Animation: Termination of Translation**



1 When a ribosome reaches a stop codon on mRNA, the A site of the ribosome accepts a "release factor"—a protein shaped like a tRNA—instead of an aminoacyl tRNA.

2 The release factor promotes hydrolysis of the bond between the tRNA in the P site and the last amino acid of the polypeptide, thus freeing the polypeptide from the ribosome.

3 The two ribosomal subunits and the other components of the assembly dissociate with the hydrolysis of two GTP molecules.

Additional steps—*post-translational modifications*—may be required before the protein can begin doing its particular job in the cell. Certain amino acids may be chemically modified by the attachment of sugars, lipids, phosphate groups, or other additions. Enzymes may remove one or more amino acids from the leading (amino) end of the polypeptide chain. In some cases, a polypeptide chain may be enzymatically cleaved into two or more pieces. In other cases, two or more polypeptides that are synthesized separately may come together, if the protein has quaternary structure; an example is hemoglobin (see Figure 5.18).

Mastering Biology BioFlix® Animation: Protein Processing

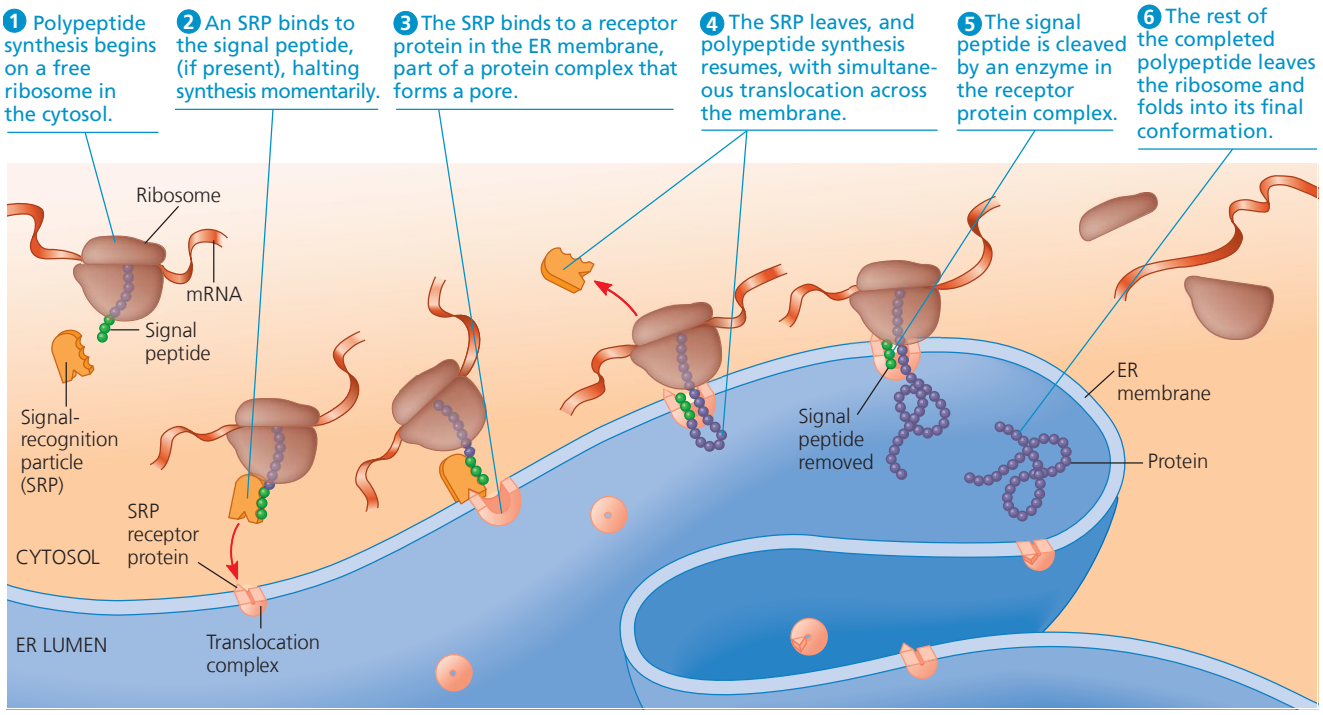
Targeting Polypeptides to Specific Locations

In electron micrographs of eukaryotic cells active in protein synthesis, two populations of ribosomes are evident: free and bound (see Figure 6.10). Free ribosomes are suspended in the cytosol and mostly synthesize proteins that stay in the cytosol and function there. In contrast, bound ribosomes are attached to the cytosolic side of the endoplasmic reticulum (ER) or to the nuclear envelope. Bound ribosomes make proteins of the endomembrane system (see Figure 6.15) as well as proteins secreted from the cell, such as insulin. It is important to note that the ribosomes themselves are identical and can alternate between being free ribosomes one time they are used and being bound ribosomes the next.

What determines whether a ribosome is free in the cytosol or bound to rough ER? Polypeptide synthesis always begins in the cytosol as a free ribosome starts to translate an mRNA molecule. There, the process continues to completion—*unless* the growing polypeptide itself cues the ribosome to attach to the ER. The polypeptides of proteins destined for the endomembrane system or for secretion are marked by a **signal peptide**, which targets the protein to the ER **(Figure 17.22)**. The signal peptide, a sequence of about 20 amino acids at or near the leading end (N-terminus) of the polypeptide, is recognized as it emerges from the ribosome by a protein-RNA complex called a **signal-recognition particle (SRP)**. This particle escorts the ribosome to a receptor protein built into the ER membrane. The receptor is part of a multiprotein translocation complex. Polypeptide synthesis continues there, and the growing polypeptide snakes across the membrane into the ER lumen via a protein pore. The signal peptide is removed by an enzyme. The rest of the completed polypeptide, if it is to be secreted from the cell, is released into solution within the ER lumen (see Figure 17.22). Or, if the polypeptide is to be a membrane protein, amino acid sequences further in the chain cause that part to remain embedded in the ER membrane. In either case, it travels in a transport vesicle to its destination (see Figure 7.9).

Other kinds of signal peptides are used to target polypeptides to mitochondria, chloroplasts, the interior of the nucleus, and other organelles that are not part of the endomembrane system. The critical difference in these cases is that translation

▼ Figure 17.22 The signal mechanism for targeting proteins to the ER.



MAKE CONNECTIONS *If this protein were destined for secretion, what would happen to it after its synthesis was completed? See Figure 7.9.*

Mastering Biology Animation: Cotranslational Translocation

is completed in the cytosol before the polypeptide is imported into the organelle. Translocation mechanisms also vary, but in all cases studied to date, the "postal zip codes" that address proteins for secretion or to cellular locations are signal peptides of some sort. Bacteria also employ signal peptides to target proteins to the plasma membrane or for secretion.

Making Multiple Polypeptides in Bacteria and Eukaryotes

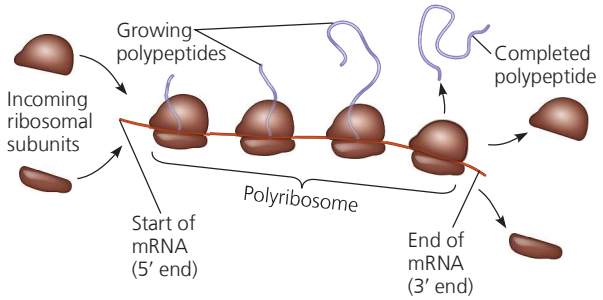
In previous sections, you learned how a single polypeptide is synthesized using the information encoded in an mRNA molecule. When a polypeptide is required in a cell, though, the need is for many copies, not just one.

In both bacteria and eukaryotes, multiple ribosomes translate an mRNA at the same time **(Figure 17.23)**; that is, a single mRNA is used to make many copies of a polypeptide simultaneously. Once a ribosome is far enough past the start codon, a second ribosome can attach to the mRNA, eventually resulting in a number of ribosomes trailing along the mRNA. Such strings of ribosomes, called **polyribosomes** (or **polysomes**), can be seen with an electron microscope; they can be either free or bound. They enable a cell to rapidly make many copies of a polypeptide.

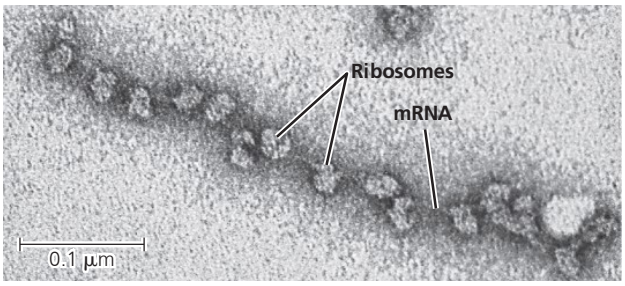
Another way both bacteria and eukaryotes increase the number of copies of a polypeptide is by transcribing multiple mRNAs from the same gene. However, the coordination of the two processes—transcription and translation—differs in the two groups. The most important differences between bacteria and eukaryotes arise from the bacterial cell's lack of compartmental organization. Like a one-room workshop, a bacterial cell ensures a streamlined operation by coupling the two processes. With no nuclear envelope, it can simultaneously transcribe and translate the same gene **(Figure 17.24)**, and the newly made protein can quickly diffuse to its site of function.

In contrast, the eukaryotic cell's nuclear envelope segregates transcription from translation and provides a compartment for extensive RNA processing. This processing stage includes additional steps, discussed earlier, the regulation of which can help coordinate the eukaryotic cell's elaborate activities. **Figure 17.25** summarizes the path from gene to polypeptide in a eukaryotic cell.

▼ Figure 17.23 Polyribosomes.

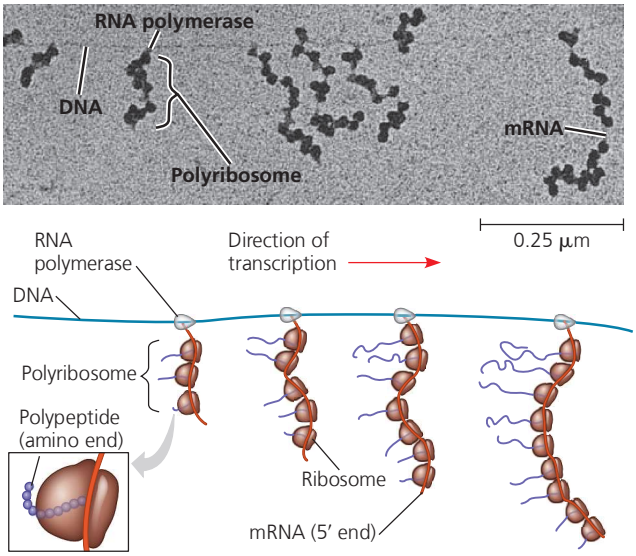


(a) An mRNA molecule is generally translated simultaneously by several ribosomes in clusters called polyribosomes.



(b) This micrograph shows a large polyribosome in a bacterial cell. Growing polypeptides are not visible here (TEM).

▼ Figure 17.24 Coupled transcription and translation in bacteria. In bacterial cells, the translation of mRNA can begin as soon as the leading (5′) end of the mRNA molecule peels away from the DNA template. The micrograph (TEM) shows a strand of *E. coli* DNA being transcribed by RNA polymerase molecules. Attached to each RNA polymerase molecule is a growing strand of mRNA, which is already being translated by ribosomes. The newly synthesized polypeptides are not visible in the micrograph but are shown in the diagram.



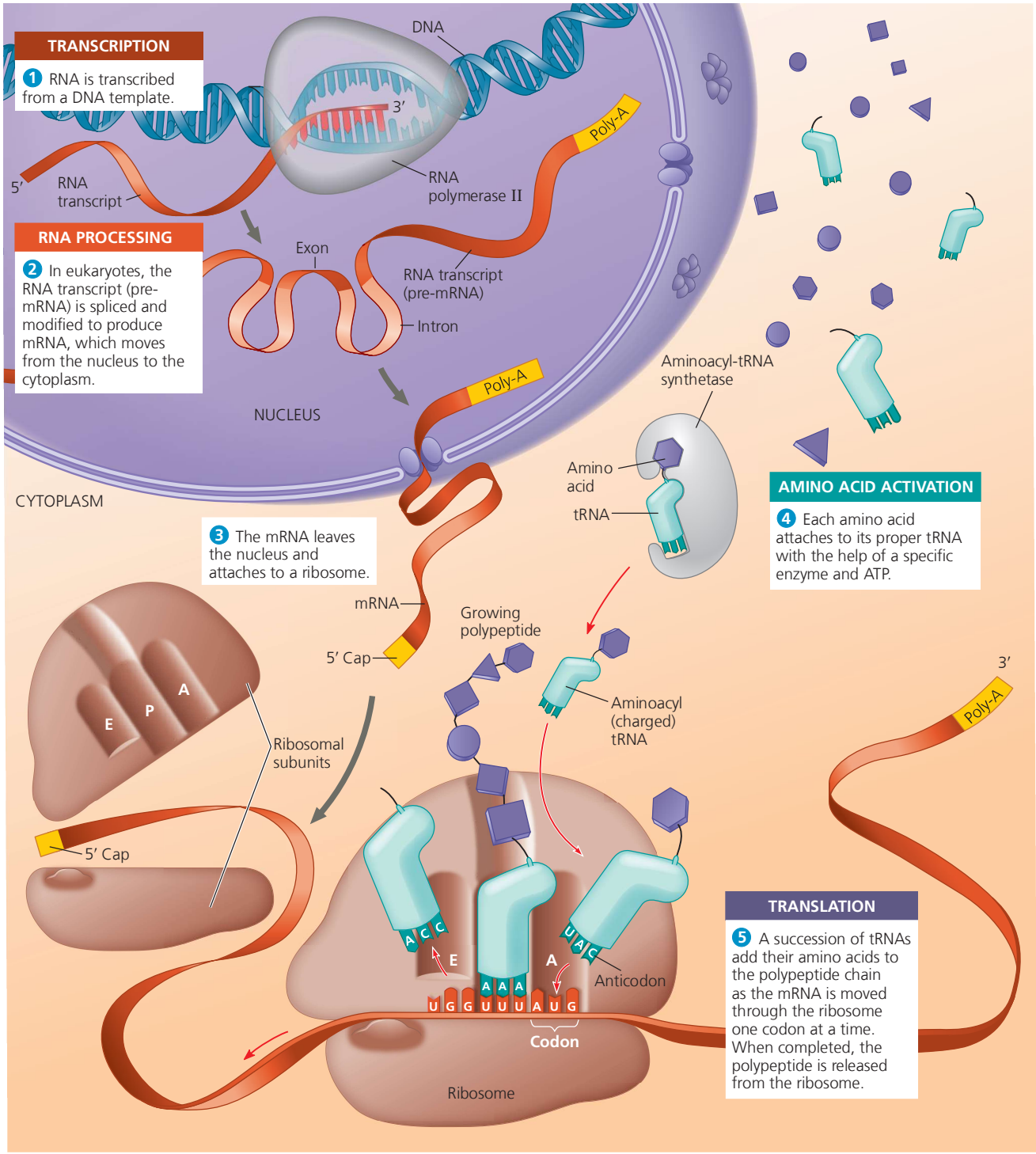
VISUAL SKILLS *Which one of the mRNA molecules started being transcribed first? On that mRNA, which ribosome started translating the mRNA first?*

CONCEPT CHECK 17.4

1. What two processes ensure that the correct amino acid is added to a growing polypeptide chain?
2. Describe how a polypeptide to be secreted reaches the endomembrane system.
3. **DRAW IT** Draw a tRNA with the anticodon 3′-CGU-5′. Given wobble, what two different codons could it bind to? Draw each codon on an mRNA, labeling all 5′ and 3′ ends, the tRNA, and the amino acid it carries.
4. **WHAT IF?** In eukaryotic cells, mRNAs have been found to have a circular arrangement in which proteins hold the poly-A tail near the 5′ cap. How might this increase translation efficiency?

For suggested answers, see Appendix A.

▼ Figure 17.25 A summary of transcription and translation in a eukaryotic cell. This diagram shows the path from one gene to one polypeptide. Each gene in the DNA can be transcribed repeatedly into many identical RNA molecules, and each mRNA can be translated repeatedly to yield many identical polypeptide molecules. (Remember that the final products of some genes are not polypeptides but RNA molecules that don't get translated, including tRNAs and rRNAs.) In general, the steps of transcription and translation are similar in bacterial, archaeal, and eukaryotic cells. The major difference is the occurrence of RNA processing in the eukaryotic nucleus. Other significant differences are found in the initiation stages of both transcription and translation and in the termination of transcription. To visualize these processes in their cellular context, see Figure 6.32d-f.



Mastering Biology BioFlix® Animation: Protein Synthesis

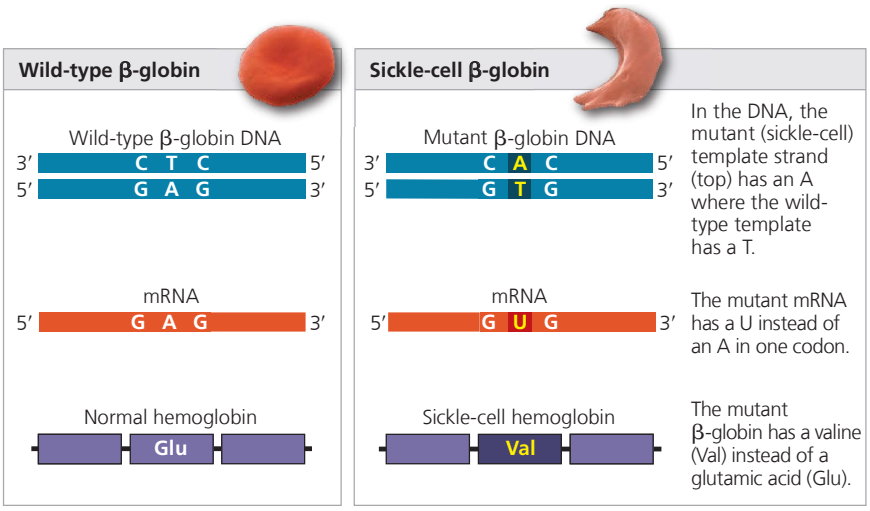
CONCEPT 17.5

Mutations of one or a few nucleotides can affect protein structure and function

Now that you have explored the process of gene expression, you can consider the effects of changes to the genetic information of a cell. Because these changes, called **mutations**, are the ultimate source of new genes, they are responsible for the huge diversity of genes found among organisms. Earlier, we considered chromosomal rearrangements that affect long segments of DNA (see Figure 15.14); these are considered large-scale mutations. Here we examine small-scale mutations of one or a few nucleotide pairs, including **point mutations**, changes in a single nucleotide pair of a gene.

If a point mutation occurs in a gamete or in a cell that gives rise to gametes, it may be transmitted to offspring and to future generations. If the mutation has an adverse effect on the phenotype of a person, the mutant condition is referred to as a genetic disorder or hereditary disease. For example, we can trace the genetic basis of sickle-cell disease to the mutation of a single nucleotide pair in the gene that encodes the β-globin polypeptide of hemoglobin. The change of a single nucleotide in the DNA's template strand leads to an altered mRNA and the production of an abnormal protein **(Figure 17.26**; also see Figure 5.19). In individuals who are homozygous for the mutant allele, the sickling of red blood cells caused by the altered hemoglobin produces the multiple symptoms associated with sickle-cell disease (see Concept 14.4 and Figure 23.18). Another disorder caused by a point mutation is a heart condition called familial cardiomyopathy, which is responsible for some of the tragic incidents of sudden death in young athletes. Point mutations in several genes encoding muscle proteins have been identified, any of which can lead to this disorder.

▼ **Figure 17.26 The molecular basis of sickle-cell disease: a point mutation.** The allele that causes sickle-cell disease differs from the wild-type (normal) allele by a single DNA nucleotide pair. The micrographs are SEMs of a normal red blood cell (on the left) and a sickled red blood cell (right) from individuals homozygous for wild-type and mutant alleles, respectively.



Mastering Biology **HHMI Animation: Sickle-Cell Disease**

Types of Small-Scale Mutations

Many mutations occur in regions outside of protein-coding genes, and any potential effect they have on the phenotype of the organism may be subtle and hard to detect. For this reason, here we'll concentrate on mutations within protein-coding genes. Small-scale mutations within a gene can be divided into two general categories: (1) single nucleotide-pair substitutions and (2) nucleotide-pair insertions or deletions. Insertions and deletions can involve one or more nucleotide pairs.

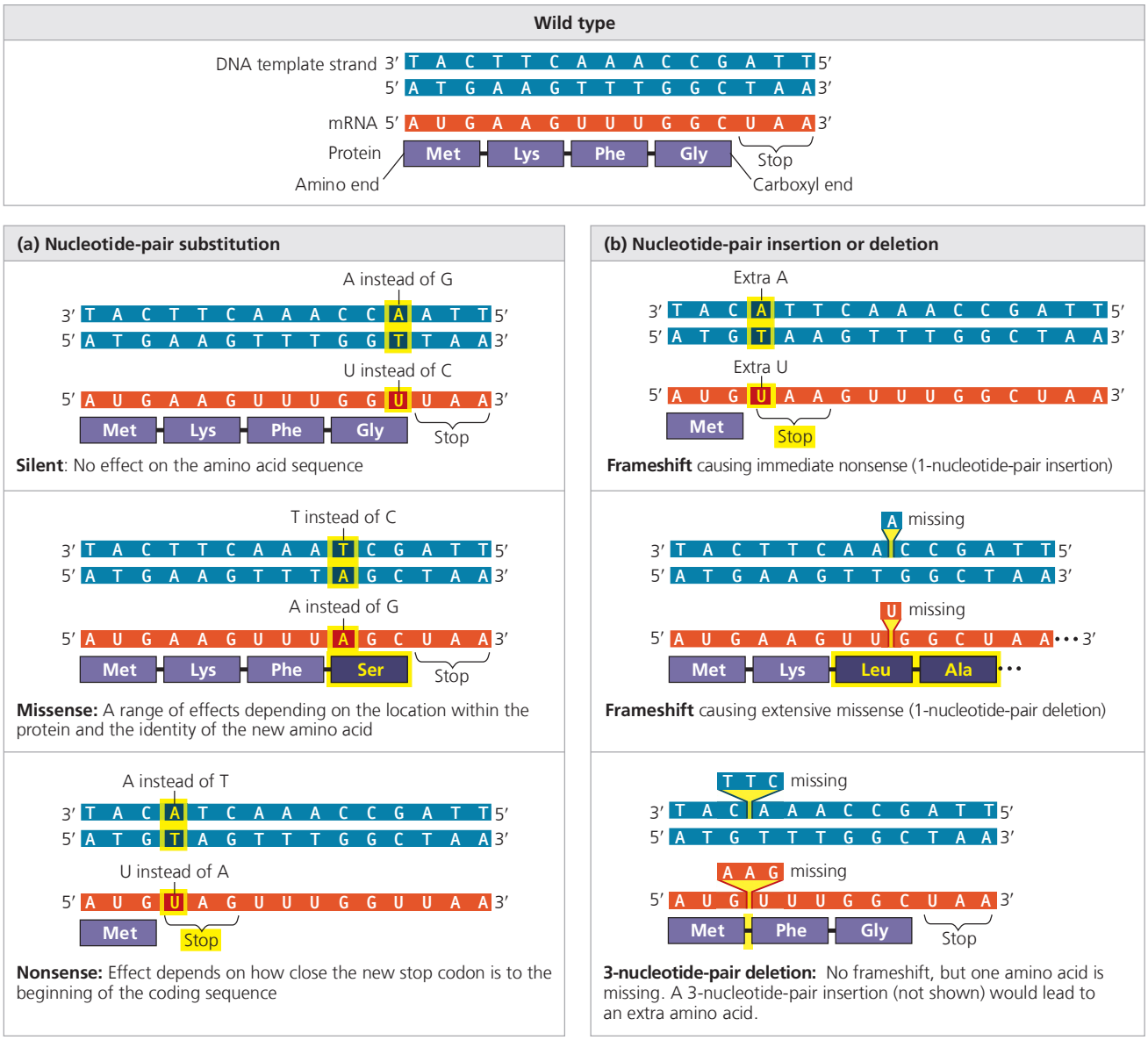
Substitutions

A **nucleotide-pair substitution** is the replacement of one nucleotide and its partner with another pair of nucleotides **(Figure 17.27a)**. Some substitutions have no effect on the encoded protein, owing to the redundancy of the genetic code. For example, if 3′-CCG-5′ on the template strand mutated to 3′-CCA-5′, the mRNA codon that used to be GGC would become GGU, but a glycine would still be inserted at the proper location in the protein (see Figure 17.6). In other words, a change in a nucleotide pair may transform one codon into another that is translated into the same amino acid. Such a change is an example of a **silent mutation**, which has no observable effect on the phenotype. (Silent mutations can occur outside genes as well.) Interestingly, there is evidence that some silent mutations may indirectly affect where or at what level the gene gets expressed, even though the actual protein is the same.

Substitutions that change one amino acid to another one are called **missense mutations**. Such a mutation may have little effect on the protein: The new amino acid may have properties similar to those of the amino acid it replaces, or it may be in a region of the protein where the exact sequence of amino acids is not essential to the protein's function.

However, the nucleotide-pair substitutions of greatest interest are those that cause a major change in a protein. The alteration of a single amino acid in a crucial area of a protein—such as in the part of the β-globin subunit of

▼ Figure 17.27 Types of small-scale mutations that affect mRNA sequence. All but one of the types shown here also affect the amino acid sequence of the encoded polypeptide.



Mastering Biology Figure Walkthrough • Animation: Mutation Types

hemoglobin shown in Figure 17.26 or in the active site of an enzyme as shown in Figure 8.19—can significantly alter protein activity. Occasionally, such a mutation leads to an improved protein or one with novel capabilities, but much more often such mutations are neutral or detrimental, leading to a useless or less active protein that impairs cellular function. A second example of a missense mutation is a mutation in the tyrosinase gene that causes the albino phenotype in the Asinara donkey. Recently, Italian researchers sequenced the tyrosinase gene in donkeys from this wild population and showed that a recessive mutation is responsible for the lack of pigment. A C is changed to a G in this gene, resulting in a histidine being added to the protein instead of an aspartic acid. The changed amino acid is in a site in tyrosinase that normally binds a copper atom, and it makes the enzyme incapable of binding copper, rendering it nonfunctional. As in the case of sickle-cell disease, just one change in a nucleotide base results in a dramatic phenotypic change.

Substitution mutations are usually missense mutations; that is, the altered codon still codes for an amino acid and thus makes sense, although not necessarily the *right* sense. But a point mutation can also change a codon for an amino acid into a stop codon. This is called a **nonsense mutation**, and it causes translation to be terminated prematurely; the

PROBLEM-SOLVING EXERCISE

Are insulin mutations the cause of three infants' neonatal diabetes?



Insulin is a hormone that acts as a key regulator of blood glucose level, and insulin deficiency can lead to diabetes. In some cases of neonatal (newborn) diabetes, the gene coding for the insulin protein has a nucleotide-pair substitution mutation that alters the protein structure enough to cause it to malfunction. Doctors can use DNA sequence information to diagnose diabetes and other diseases. For example, the insulin gene sequence of a patient with neonatal diabetes can be analyzed to determine if it has a mutation and, if so, its effect.

Instructors: A version of this Problem-Solving Exercise can be assigned in **Mastering Biology**. Or a more extensive investigation called "Solve It: Which Insulin Mutations May Result in Disease?" can be assigned.

In this exercise, you will determine the effect of mutations within insulin gene sequences in infant diabetes patients.

Your Approach Suppose you are a medical geneticist presented with three infant patients, all of whom have a nucleotide-pair substitution in their insulin gene. It is your job to analyze each mutation to figure out its effect on the amino acid sequence of the insulin protein. To identify the mutation in each patient, you will compare his or her individual insulin complementary DNA (cDNA) sequence to that of the wild-type cDNA. (cDNA is a double-stranded DNA molecule that is based on the mRNA sequence and thus contains only the portion of a gene that is translated—introns are not included. cDNA sequences are commonly used to compare the coding regions of genes.) Identifying the codons that have been changed will tell you which, if any, amino acids are altered in the patient's insulin protein.

Your Data You will analyze the cDNA codons for amino acids 35–54 (of the 110 amino acids) of each patient's insulin protein, so the start codon (AUG) is not present. The sequences of the wild-type cDNA and the patients' cDNA are shown below, arranged in codons.

Wild-type cDNA	5′-CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TTC TTC TAC ACA CCC AAG ACC-3′
Patient 1 cDNA	5′-CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TGC TTC TAC ACA CCC AAG ACC-3′
Patient 2 cDNA	5′-CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TCC TTC TAC ACA CCC AAG ACC-3′
Patient 3 cDNA	5′-CTG GTG GAA GCT CTC TAC CTA GTG TGC GGG GAA CGA GGC TTC TTG TAC ACA CCC AAG ACC-3′

Data from N. Nishi and K. Nanjo, Insulin gene mutations and diabetes, *Journal of Diabetes Investigation* 2:92–100 (2011).

Your Analysis

1. Compare each patient's cDNA sequence to the wild-type cDNA sequence. Circle the codons where a nucleotide-pair substitution mutation has occurred.
2. Use a codon table (see Figure 17.6) to identify the amino acid encoded by the codon with the mutation in each patient's insulin sequence, and compare it to the amino acid encoded by the codon in the corresponding wild-type sequence. As is standard practice with DNA sequences, the *coding* (nontemplate) strand of the cDNA has been provided, so to convert it to mRNA for use with the codon table, you only need to change T to U. Classify each patient's nucleotide-pair substitution mutation: Is it a silent, missense, or nonsense mutation? Explain, for each answer.
3. Compare the structure of the amino acid you identified in each patient's insulin sequence to that of the corresponding amino acid in the wild-type sequence (see Figure 5.14). Given that each patient has neonatal diabetes, discuss how the change of amino acid in each might have affected the insulin protein and thus resulted in the disease. (Consider the chemical nature of the side chains.)

resulting polypeptide will be shorter than the polypeptide encoded by the normal gene. Most nonsense mutations lead to nonfunctional proteins.

In the **Problem-Solving Exercise**, you'll work with a few common single nucleotide-pair substitution mutations in the gene encoding insulin, some or all of which may lead to diabetes. You will classify these mutations into one of the types we just described and characterize the change in the amino acid sequence.

Insertions and Deletions

Insertions and **deletions** are additions or losses of nucleotide pairs in a gene **(Figure 17.27b)**. These mutations have a disastrous effect on the resulting protein more often than substitutions do. Insertion or deletion of nucleotides may alter the reading frame of the genetic message, the triplet grouping of nucleotides on the mRNA that is read during translation. Such a mutation, called a **frameshift mutation**, occurs whenever the number of nucleotides inserted or deleted is not a multiple of three. All nucleotides downstream of the deletion or insertion will be improperly grouped into codons; the result will be extensive missense mutations, usually ending sooner or later in a nonsense mutation that leads to premature termination. Unless the frameshift is very near the end of the gene, the protein is almost certain to be nonfunctional. Insertions and deletions also occur outside of coding regions; these are not called frameshift mutations, but can have effects on the phenotype—for instance, they can affect how a gene is expressed.

New Mutations and Mutagens

Mutations can arise in a number of ways. Errors during DNA replication or recombination can lead to nucleotide-pair substitutions, insertions, or deletions, as well as to mutations affecting longer stretches of DNA. If an incorrect nucleotide is added to a growing chain during replication, for example, the base on that nucleotide will then be mismatched with the nucleotide base on the other strand. In many cases, the error will be corrected by DNA proofreading and repair systems (see Concept 16.2). Otherwise, the incorrect base will be used as a template in the next round of replication, resulting in a mutation. Such mutations are called *spontaneous mutations*. It is difficult to calculate the rate at which such mutations occur. Rough estimates have been made of the rate of mutation during DNA replication for both *E. coli* and eukaryotes, and the numbers are similar: About one nucleotide in every 10^{10} is altered, and the change is passed on to the next generation of cells.

A number of physical and chemical agents, called **mutagens**, interact with DNA in ways that cause mutations. In the 1920s, Hermann Muller discovered that X-rays caused genetic changes in fruit flies, and he used X-rays to make *Drosophila* mutants for his genetic studies (just as Beadle and Tatum did with their *Neurospora* cells). But Muller also recognized an alarming implication of his discovery: X-rays and other forms of high-energy radiation pose hazards to the genetic material of people as well as laboratory organisms. Mutagenic radiation, a physical mutagen, includes ultraviolet (UV) light, which can cause disruptive thymine dimers in DNA (see Figure 16.20).

Chemical mutagens fall into several categories. Nucleotide analogs are chemicals similar to normal DNA nucleotides but that pair incorrectly during DNA replication. Other chemical mutagens interfere with correct DNA replication by inserting themselves into the DNA and distorting the double helix. Still other mutagens cause chemical changes in bases that change their pairing properties.

Researchers have developed a variety of methods to test the mutagenic activity of chemicals. A major application of these tests is the preliminary screening of chemicals to identify those that may cause cancer. This approach makes sense because most carcinogens (cancer-causing chemicals) are mutagenic, and conversely, most mutagens are carcinogenic.

Using CRISPR to Edit Genes and Correct Disease-Causing Mutations

Ever since biologists have understood how disease-causing proteins were the result of mutations in genes, they have sought techniques for **gene editing**—altering genes in a specific, predictable way. Their aim has been to use such a technique to change specific genes in living cells, in part to try to correct genes that cause disease. Over the past 15 years, biologists have developed a powerful new technique for gene editing, called the **CRISPR-Cas9 system**, that is transforming the field of *genetic engineering*, the direct manipulation of genes for practical purposes. Cas9 is a bacterial protein that helps defend bacteria against the viruses that infect them (bacteriophages). In bacterial cells, Cas9 acts together with "guide RNA" made from the CRISPR region of the bacterial genome. (Figure 19.7 explains how this defense system works in bacteria.)

Mastering Biology HHMI Video: Jennifer Doudna: CRISPR Provides a Tool for Changing Genes

Mastering Biology Interview with Francisco Mojica: Discovering CRISPR in bacteria (see the interview before Chapter 13)

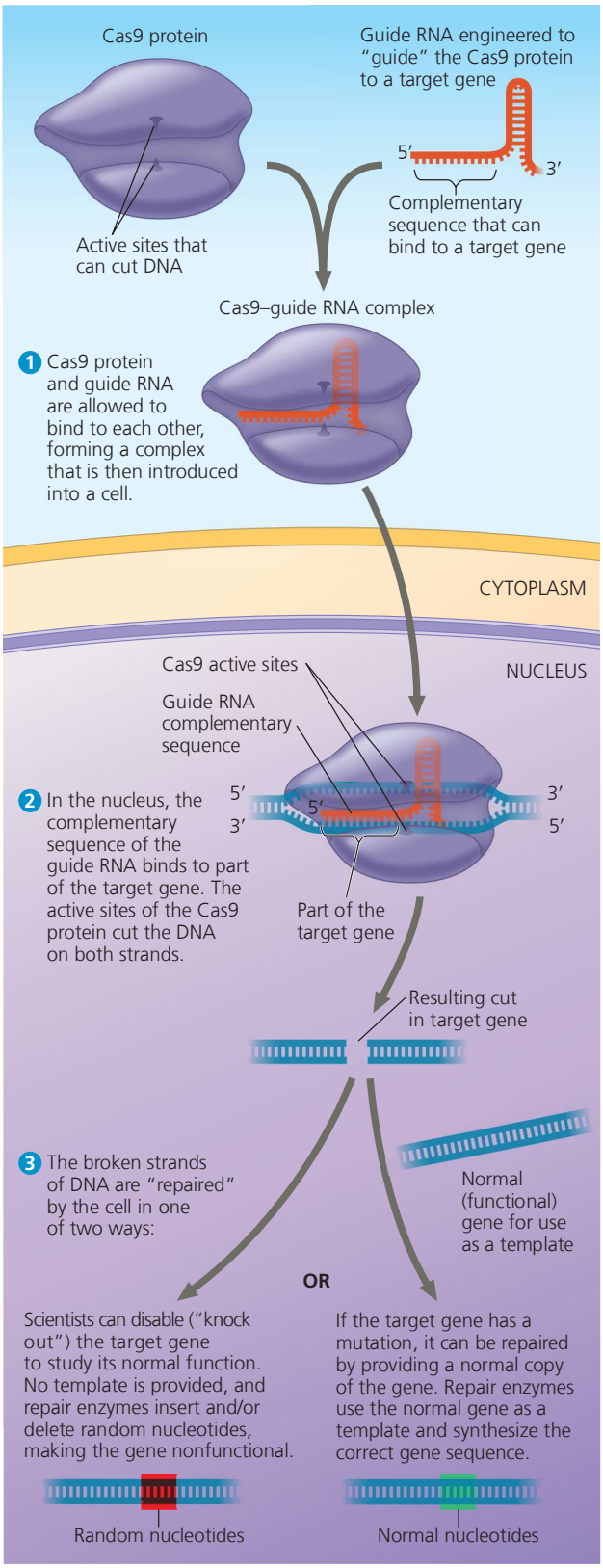


Similar to other DNA-cutting enzymes involved in DNA repair (described in Concept 16.2), Cas9 is a nuclease that cuts double-stranded DNA molecules. The power of Cas9 for gene editing is that the Cas9 protein will cut any sequence to which it is directed. Cas9 is directed to its target by a guide RNA molecule that it binds and uses as a homing device, cutting both strands of any DNA sequence that is complementary to the guide RNA. Scientists have been able to exploit the function of Cas9 by introducing a Cas9–guide RNA complex into a cell they wish to alter **(Figure 17.28)**. The guide RNA in the complex is engineered to be complementary to the "target" gene. Cas9 cuts both strands of the target DNA, and the resulting broken ends of DNA trigger a DNA repair system (similar to that shown in Figure 16.20). When there is no undamaged DNA for the enzymes of the repair system to use as a template, as shown at the bottom left of Figure 17.28, the repair enzymes introduce or remove random nucleotides while rejoining the ends. Generally, this process alters the DNA sequence so that the gene no longer works properly. This technique is a highly successful way for researchers to "knock out" (disable) a given gene to study what that gene does in an organism.

But what about using this system to help treat genetic diseases? Researchers have modified the technique so that the CRISPR-Cas9 system can be used to repair a gene that has a harmful mutation. They introduce a segment from the normal (functional) gene along with the CRISPR-Cas9 system. After Cas9 cuts the target DNA, repair enzymes can use the normal DNA as a template to repair the target DNA at the break point. In this way, the CRISPR-Cas9 system edits the defective gene so that it is corrected (see the bottom right of Figure 17.28).

In 2018, researchers reported promising results using the CRISPR-Cas 9 system in an attempt to correct the genetic

▼ **Figure 17.28** **Gene editing using the CRISPR-Cas9 system.**



defect that causes sickle-cell disease. They edited human cells from patients with sickle-cell disease and injected them into bone marrow in mice. After 19 weeks, the gene remained corrected in 20–40% of the injected cells. Researchers, physicians, and patients alike are excited about the potential of CRISPR technology to treat or even cure human diseases that have a genetic basis, such as sickle-cell, Alzheimer's, and Parkinson's diseases, as well as some types of cancer. However, there are still concerns about using CRISPR-treated cells in humans because of possible effects on genes that are not being targeted; this is an active area of research.

Jennifer Doudna, a co-discoverer of CRISPR-Cas9, recognized not only its incredible potential but also the danger of its misapplication. After having a dream in which Adolf Hitler asked her what could be done with this gene-editing tool, Doudna realized it was imperative to spend some time reflecting on ethical considerations. In 2015, she convened a conference of biologists who agreed at the end of the meeting to urge extreme caution as the field moves forward—and the discussion continues.

What Is a Gene? *Revisiting the Question*

Our definition of a gene has evolved over the past few chapters, as it has through the history of genetics. We began with the Mendelian concept of a gene as a discrete unit of inheritance that affects a phenotypic character (Chapter 14). We saw that Morgan and his colleagues assigned such genes to specific loci on chromosomes (Chapter 15). We went on to view a gene as a region of specific nucleotide sequence along the length of the DNA molecule of a chromosome (Chapter 16). Finally, in this chapter, we have considered a functional definition of a gene as a DNA sequence that codes for a specific polypeptide chain or a functional RNA molecule, such as a tRNA. All these definitions are useful, depending on the context in which genes are being studied.

We have noted that merely saying a gene codes for a polypeptide is an oversimplification. Most eukaryotic genes contain noncoding segments (such as introns), so large portions of these genes have no corresponding segments in polypeptides. Molecular biologists also often include promoters and certain other regulatory regions of DNA within the boundaries of a gene. These DNA sequences are not transcribed, but they can be considered part of the functional gene because they must be present for transcription to occur. Our definition of a gene must also be broad enough to include the DNA that is transcribed into rRNA, tRNA, and other RNAs that are not translated. These genes have no polypeptide products but play crucial roles in the cell. Thus, we arrive at the following definition: *A gene is a region of DNA that can be expressed to produce a final functional product that is either a polypeptide or an RNA molecule.*

When considering phenotypes, however, it is often useful to start by focusing on genes that code for polypeptides. In this chapter, you have learned in molecular terms how a typical gene is expressed—by transcription into RNA and then translation into a polypeptide that forms a protein of specific structure and function. Proteins, in turn, bring about an organism's observable phenotype.

A given type of cell expresses only a subset of its genes. This is an essential feature in multicellular organisms: You'd be in trouble if the lens cells in your eyes started expressing the genes for hair proteins, which are normally expressed only in hair follicle cells! Gene expression is precisely regulated, which we'll explore in the next chapter, beginning with the simpler case of bacteria and continuing with eukaryotes.

CONCEPT CHECK 17.5

1. What happens when one nucleotide pair is lost from the middle of the coding sequence of a gene?
2. **MAKE CONNECTIONS** Individuals heterozygous for the sickle-cell allele are generally healthy but show phenotypic effects of the allele under some circumstances (see Figure 14.17). Explain in terms of gene expression.
3. **DRAW IT** The template strand of a gene includes this sequence: 3′-TACTTGTCCGATATC-5′. The sequence is mutated to 3′-TACTTGTCCAATATC-5′. For both wild-type and mutant sequences, draw the double-stranded DNA, the resulting mRNA, and the amino acid sequence each encodes. What is the effect of the mutation on the amino acid sequence?
4. **WHAT IF?** How could you use the CRISPR-Cas9 system to correct the mutation described in question 3? Based on the answer to question 3, would this be worth doing?

For suggested answers, see Appendix A.

17 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 17.1

Genes specify proteins via transcription and translation *(pp. 336–342)*

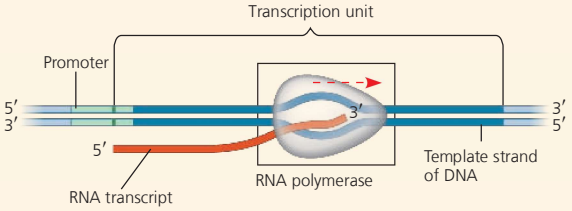
- Beadle and Tatum's studies of mutant strains of *Neurospora* led to the one gene–one polypeptide hypothesis. During **gene expression**, the information encoded in genes is used to make specific polypeptide chains (enzymes and other proteins) or RNA molecules.
- **Transcription** is the synthesis of RNA complementary to a **template strand** of DNA. **Translation** is the synthesis of a polypeptide whose amino acid sequence is specified by the nucleotide sequence in **messenger RNA (mRNA)**.
- Genetic information is encoded as a sequence of nonoverlapping nucleotide triplets, or **codons**. A codon in mRNA either is translated into an amino acid (61 of the 64 codons) or serves as a stop signal (3 codons). Codons must be read in the correct **reading frame**.

? *Describe the process of gene expression, by which a gene affects the phenotype of an organism.*

CONCEPT 17.2

Transcription is the DNA-directed synthesis of RNA: *A Closer Look* *(pp. 342–344)*

- RNA synthesis is catalyzed by **RNA polymerase**, which links together RNA nucleotides complementary to a DNA template strand. Transcription follows the same base-pairing rules as DNA replication, except that in RNA, uracil (U) substitutes for thymine (T).



- The three stages of transcription are initiation, elongation, and termination. A **promoter** sequence, often including a **TATA box** in eukaryotes, establishes where RNA synthesis is initiated. **Transcription factors** help eukaryotic RNA polymerase recognize promoter sequences, forming a **transcription initiation complex**. Termination differs in bacteria and eukaryotes.

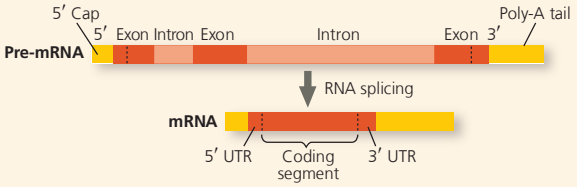
? *Compare the initiation of transcription in bacteria and eukaryotes.*

CONCEPT 17.3

Eukaryotic cells modify RNA after transcription *(pp. 345–347)*

- Eukaryotic mRNAs undergo **RNA processing**, which includes RNA splicing, the addition of a modified nucleotide **5′ cap** to the 5′ end, and the addition of a **poly-A tail** to the 3′ end. The processed mRNA includes an untranslated region (5′ UTR or 3′ UTR) at each end of the coding segment.
- Most eukaryotic genes are split into segments: They have **introns** interspersed among the **exons** (the regions included in the mRNA). In **RNA splicing**, introns are removed and exons joined. RNA splicing is typically carried out by **spliceosomes**, but in

some cases, RNA alone catalyzes its own splicing. The properties of RNA allow some RNAs (called **ribozymes**) to act as catalysts. The presence of introns allows for **alternative RNA splicing**.

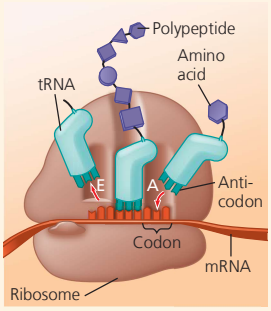


What function do the 5′ cap and the poly-A tail serve on a eukaryotic mRNA?

CONCEPT 17.4

Translation is the RNA-directed synthesis of a polypeptide: *A Closer Look* *(pp. 347–356)*

- A cell translates an mRNA message into protein using **transfer RNAs (tRNAs)**. After being bound to a specific amino acid by an **aminoacyl-tRNA synthetase**, a tRNA lines up via its **anticodon** at the complementary codon on mRNA. A **ribosome**, made up of **ribosomal RNAs (rRNAs)** and proteins, facilitates this coupling with binding sites for mRNA and tRNA.
- Ribosomes coordinate the three stages of translation: initiation, elongation, and termination. The formation of peptide bonds between amino acids is catalyzed by rRNAs as tRNAs move through the **A** and **P sites** and exit through the **E site**.



- After translation, during protein processing, proteins may be modified by cleavage or by attachment of sugars, lipids, phosphates, or other chemical groups.
- Free ribosomes in the cytosol initiate synthesis of all proteins, but proteins with a **signal peptide** are synthesized on the ER.
- A gene can be transcribed by multiple RNA polymerases simultaneously. Also, a single mRNA molecule can be translated simultaneously by a number of ribosomes, forming a **polyribosome**. In bacteria, these processes are coupled, but in eukaryotes they are separated in space and time by the nuclear membrane.

How do tRNAs function in the context of the ribosome during translation?

CONCEPT 17.5

Mutations of one or a few nucleotides can affect protein structure and function *(pp. 357–362)*

- Small-scale **mutations** include **point mutations**, changes in one DNA nucleotide pair, which may lead to production of nonfunctional proteins. **Nucleotide-pair substitutions** can cause **missense** or **nonsense mutations**. Nucleotide-pair **insertions** or **deletions** may produce **frameshift mutations**.
- Spontaneous mutations can occur during DNA replication, recombination, or repair. Chemical and physical **mutagens** cause DNA damage that can alter genes.
- The **CRISPR-Cas9 system** is a powerful new **gene-editing** technique with the potential to correct genetic mutations that cause disease. Significant technical and ethical questions have been raised about how this could or should be used.

What will be the results of chemically modifying one nucleotide base of a gene? What role is played by DNA repair systems in the cell?

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. In eukaryotic cells, transcription cannot begin until
   (A) the two DNA strands have completely separated and exposed the promoter.
   (B) several transcription factors have bound to the promoter.
   (C) the 5′ caps are removed from the mRNA.
   (D) the DNA introns are removed from the template.
2. Which of the following is true of a codon?
   (A) It never codes for the same amino acid as another codon.
   (B) It can code for more than one amino acid.
   (C) It can be either in DNA or in RNA
   (D) It is the basic unit of protein structure.
3. The anticodon of a particular tRNA molecule is
   (A) complementary to the corresponding mRNA codon.
   (B) complementary to the corresponding triplet in rRNA.
   (C) the part of tRNA that bonds to a specific amino acid.
   (D) catalytic, making the tRNA a ribozyme.
4. Which of the following is true of RNA processing?
   (A) Exons are cut out before mRNA leaves the nucleus.
   (B) Nucleotides are added at both ends of the RNA.
   (C) Ribozymes may function in the addition of a 5′ cap.
   (D) RNA splicing adds a poly-A tail to the mRNA.
5. Which component is directly involved in translation?
   (A) RNA polymerase (C) spliceosome
   (B) ribosome (D) DNA

Levels 3-4: Applying/Analyzing

6. Using Figure 17.6, identify a 5′ → 3′ sequence of nucleotides in the DNA template strand for an mRNA coding for the polypeptide sequence Phe-Pro-Lys.
   (A) 5′-UUUCCCAAA-3′
   (B) 5′-GAACCCCTT-3′
   (C) 5′-CTTCGGGAA-3′
   (D) 5′-AAACCCUUU-3′
7. Which of the following mutations would be *most* likely to have a harmful effect on an organism?
   (A) a deletion of three nucleotides near the middle of a gene
   (B) a single nucleotide deletion in the middle of an intron
   (C) a single nucleotide deletion near the end of the coding sequence
   (D) a single nucleotide insertion downstream of, and close to, the start of the coding sequence

8. Would the coupling of the processes shown in Figure 17.24 be found in a eukaryotic cell? Explain why or why not.

9. Complete the following table:

Type of RNA	Functions
Messenger RNA (mRNA)	
Transfer RNA (tRNA)	
	In a ribosome, plays a structural role; as a ribozyme, plays a catalytic role (catalyzes peptide bond formation)
Primary transcript	
Small RNAs in the spliceosome	

Levels 5-6: Evaluating/Creating

10. **EVOLUTION CONNECTION** Most amino acids are coded for by a set of similar codons (see Figure 17.6). Propose at least one evolutionary explanation to account for this pattern.

11. **SCIENTIFIC INQUIRY** Knowing that the genetic code is almost universal, a scientist uses molecular biological methods to insert the human β-globin gene (shown in Figure 17.12) into bacterial cells, hoping the cells will express it and synthesize functional β-globin protein. Instead, the protein produced is nonfunctional and is found to contain many fewer amino acids than does β-globin made by a eukaryotic cell. Explain why.

12. **WRITE ABOUT A THEME: INFORMATION** Evolution accounts for the unity and diversity of life, and the continuity of life is based on heritable information in the form of DNA. In a short essay (100–150 words), discuss how the fidelity with which DNA is inherited is related to the processes of evolution. (Review the discussion of proofreading and DNA repair in Concept 16.2.)

13. **SYNTHESIZE YOUR KNOWLEDGE**



Some mutations result in proteins that function well at one temperature but are nonfunctional at a different (usually higher) temperature. Siamese cats have such a "temperature-sensitive" mutation in a gene encoding an enzyme that makes dark pigment in the fur. The mutation results in the breed's distinctive point markings and lighter body color (see the photo). Using this information and what you learned in the chapter, explain the pattern of the cat's fur pigmentation.

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
How is green fluorescent protein (GFP) used as a marker for gene expression? Go to "Lighting up Life" at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.

18 Regulation of Gene Expression

KEY CONCEPTS

18.1 Bacteria often respond to environmental change by regulating transcription *p. 366*

18.2 Eukaryotic gene expression is regulated at many stages *p. 370*

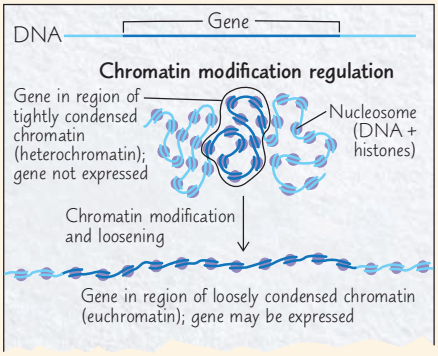
18.3 Noncoding RNAs play multiple roles in controlling gene expression *p. 379*

18.4 A program of differential gene expression leads to the different cell types in a multicellular organism *p. 381*

18.5 Cancer results from genetic changes that affect cell cycle control *p. 388*

Study Tip

Make a visual study guide: Draw a region of DNA representing a gene. For each level of gene regulation that you read about, draw a sketch of your gene being regulated at that level. A drawing of chromatin modification regulation is shown as an example.



Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 18
- Animation: Control of Gene Expression
- Animation: Causes of Cancer

For Instructors to Assign (in Item Library)
- Tutorial: Regulation of Gene Expression in Eukaryotes
- Scientific Skills Exercise: Analyzing DNA Deletion Experiments

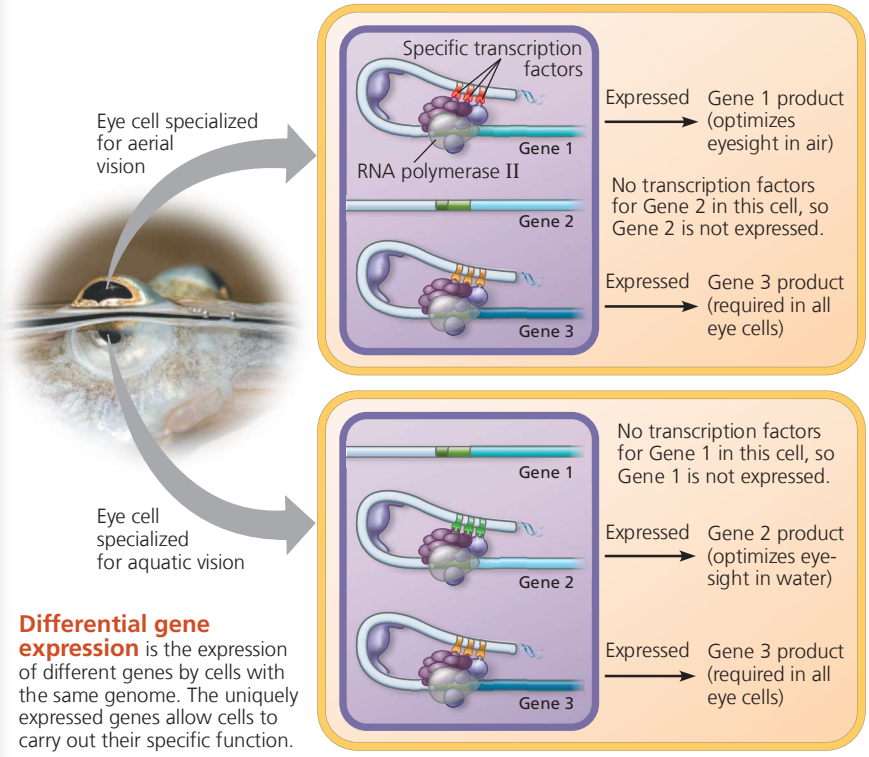
Ready-to-Go Teaching Module (in Instructor Resources)
- The *trp* and *lac* Operons (Concept 18.1)



Figure 18.1 *Anableps anableps*, or "cuatro ojos" ("four eyes"), glides through lakes and ponds in South America, the upper half of each eye protruding from the water. The eye's upper half is well-suited for aerial vision and the lower half for aquatic vision. All eye cells, however, contain the same genes.

How can two cells with the same set of genes function differently?

To be expressed, each gene requires a particular set of transcription factors. Different cells have different sets of specific transcription factors.



Differential gene expression is the expression of different genes by cells with the same genome. The uniquely expressed genes allow cells to carry out their specific function.

CONCEPT 18.1

Bacteria often respond to environmental change by regulating transcription

Bacterial cells that can conserve resources and energy have a selective advantage over cells that are unable to do so. Thus, natural selection has favored bacteria that express only the genes whose products are needed by the cell.

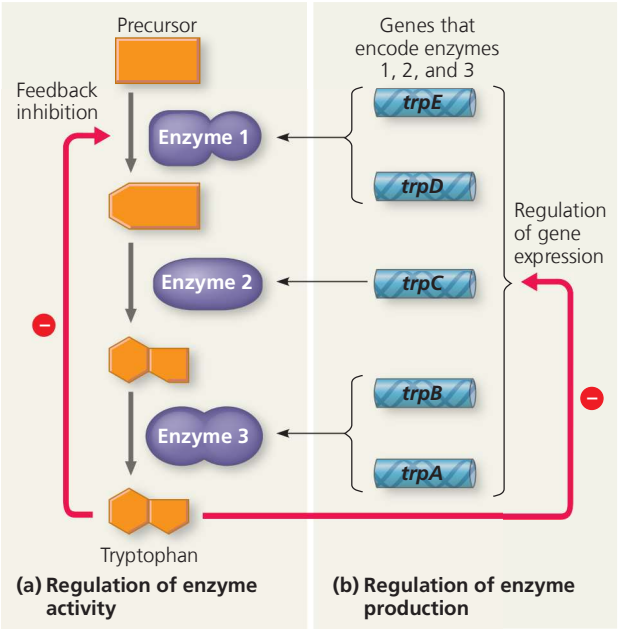
Consider, for instance, an individual *Escherichia coli* (*E. coli*) cell living in a human colon, dependent for its nutrients on the whimsical eating habits of its host. If the environment is lacking in the amino acid tryptophan, which the bacterium needs to survive, the cell responds by activating a metabolic pathway that makes tryptophan from another compound. If the human host later eats a tryptophan-rich meal, the bacterial cell stops producing tryptophan, thus avoiding wasting resources to produce a substance that is readily available from the surrounding solution.

A metabolic pathway can be controlled on two levels, as shown for the synthesis of tryptophan in **Figure 18.2**. First, cells can adjust the activity of enzymes already present. This is a fairly rapid physiological response, which relies on the sensitivity of many enzymes to chemical cues that increase or decrease their catalytic activity (see Concept 8.5). The activity of the first enzyme in the pathway is inhibited by the pathway's end product—tryptophan, in this case (Figure 18.2a). Thus, if tryptophan accumulates in a cell, it shuts down the synthesis of more tryptophan by inhibiting enzyme activity. Such *feedback inhibition*, typical of anabolic (biosynthetic) pathways, allows a cell to adapt to short-term fluctuations in the supply of a substance it needs (see Figure 8.21).

Second, cells can adjust the production level of certain enzymes via a genetic mechanism; that is, they can regulate the expression of the genes encoding the enzymes. If, in our example, the environment provides all the tryptophan the cell needs, the cell stops making the enzymes that catalyze the synthesis of tryptophan (Figure 18.2b). In this case, the control of enzyme production occurs at the level of transcription, the synthesis of messenger RNA from the genes that code for these enzymes.

Regulation of the tryptophan synthesis pathway is just one example of how bacteria tune their metabolism to changing environments. Many genes of the bacterial genome are switched on or off by changes in the metabolic status of the cell; some genes are regulated singly and others as groups of related genes. One basic mechanism for this type of regulation of groups of genes in bacteria, described as the *operon model*, was discovered in 1961 by François Jacob and Jacques Monod at the Pasteur Institute in Paris. Let's see what an operon is and how it works.

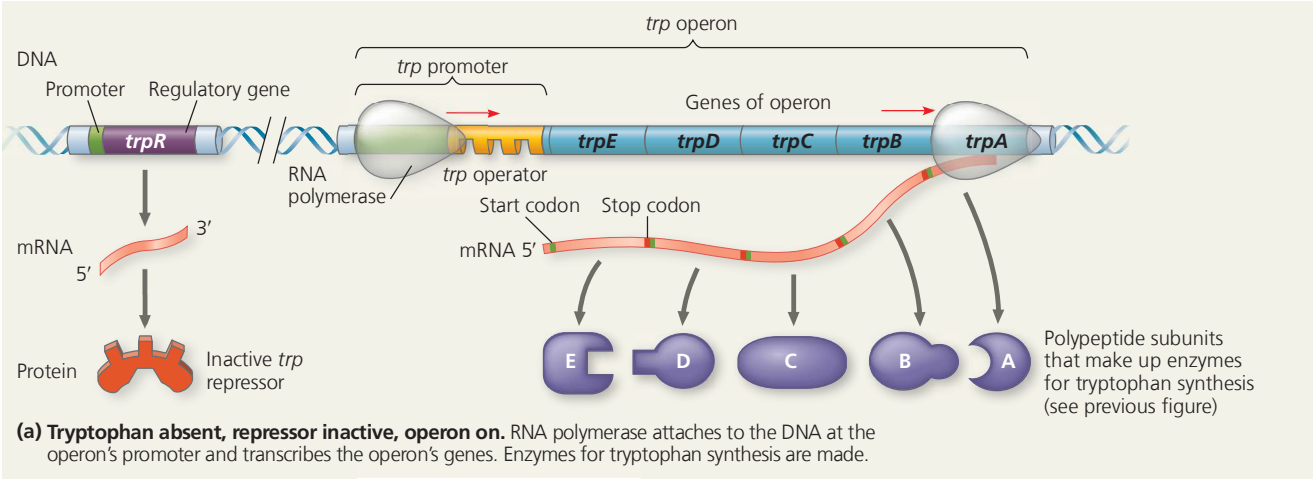
▼ Figure 18.2 Regulation of a metabolic pathway. In the pathway for tryptophan synthesis, an abundance of tryptophan can both **(a)** inhibit the activity of the first enzyme in the pathway (feedback inhibition), a rapid response, and **(b)** repress expression of the genes encoding all subunits of the enzymes in the pathway, a longer-term response. The symbol ⊖ stands for inhibition.



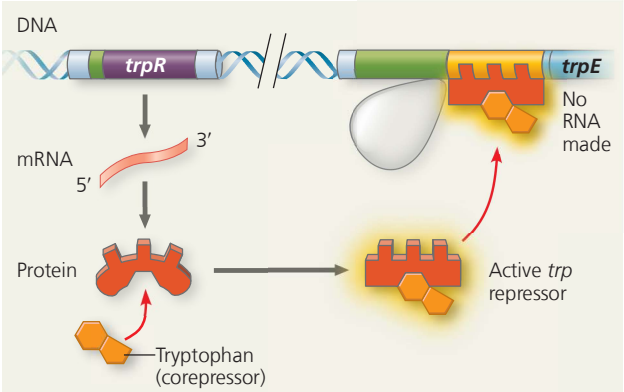
Operons: The Basic Concept

E. coli synthesizes the amino acid tryptophan from a precursor molecule in the three-step pathway shown in Figure 18.2. Each reaction in the pathway is catalyzed by a specific enzyme, and the five genes that code for the subunits of these enzymes are clustered together on the bacterial chromosome. A single promoter serves all five genes, which together constitute a transcription unit. (Recall that a promoter is a site where RNA polymerase can bind to DNA and begin transcription; see Figure 17.8.) Thus, transcription gives rise to one long mRNA molecule that codes for the five polypeptides making up the enzymes in the tryptophan pathway **(Figure 18.3a)**. The cell can translate this one mRNA into five separate polypeptides because the mRNA is punctuated with start and stop codons that signal where the coding sequence for each polypeptide begins and ends.

A key advantage of grouping genes of related function into one transcription unit is that a single "on-off switch" can control the whole cluster of functionally related genes; in other words, these genes are *coordinately controlled*. When an *E. coli* cell must make tryptophan for itself because its surrounding environment lacks this amino acid, all the enzymes for the metabolic pathway are synthesized at the same time. The on-off switch is a segment of DNA called an **operator**. Both its location and name suit its function: Positioned within the promoter or, in some cases, between the promoter and the enzyme-coding genes, the operator controls the access of



(a) Tryptophan absent, repressor inactive, operon on. RNA polymerase attaches to the DNA at the operon's promoter and transcribes the operon's genes. Enzymes for tryptophan synthesis are made.



(b) Tryptophan present, repressor active, operon off. As tryptophan accumulates, it inhibits its own production by activating the repressor protein, which binds to the operator, blocking transcription. Enzymes for tryptophan synthesis are not made.

▲ Figure 18.3 The *trp* operon in *E. coli*: regulated synthesis of repressible enzymes. Tryptophan is an amino acid produced by an anabolic pathway catalyzed by three enzymes (see Figure 18.2). **(a)** The five genes encoding the polypeptide subunits of the enzymes in this pathway are grouped, along with a promoter, into the *trp* operon. The *trp* operator (the repressor binding site) is located within the *trp* promoter (the RNA polymerase binding site). **(b)** Accumulation of tryptophan, the end product of the pathway, represses transcription of the *trp* operon, thus blocking synthesis of all the enzymes in the pathway and shutting down tryptophan production.

VISUAL SKILLS *Describe what happens to the* trp *operon as the cell uses up its store of tryptophan.*

RNA polymerase to the genes. Together, the operator, the promoter, and the genes they control—the entire stretch of DNA required for enzyme production for the tryptophan pathway—constitute an **operon**. The *trp* operon (*trp* for tryptophan) is one of many operons in the *E. coli* genome (see Figure 18.3a).

If the operator is the operon's switch for controlling transcription, how does this switch work? By itself, the *trp* operon is turned on; that is, RNA polymerase can bind to the promoter and transcribe the genes of the operon. The *trp* operon can be switched off by a protein that is called the *trp* repressor. A **repressor** binds to the operator, preventing RNA polymerase from transcribing the genes, often by preventing RNA polymerase from binding **(Figure 18.3b)**. A repressor protein is specific for the operator of a particular operon. For example, the *trp* repressor, which switches off the *trp* operon by binding to the *trp* operator, has no effect on other operons in the *E. coli* genome.

A repressor protein is encoded by a **regulatory gene**—in this case, a gene called *trpR; trpR* is located some distance from the *trp* operon and has its own promoter. Regulatory genes are among the bacterial genes that are expressed continuously, although at a low rate, and a few *trp* repressor molecules are always present in *E. coli* cells. Why, then, is the *trp* operon not switched off permanently? First, the binding of repressors to operators is reversible. An operator alternates between two states: one with the repressor bound and one without the repressor bound. The relative duration of the repressor-bound state is higher when more active repressor molecules are present. Second, the *trp* repressor, like most regulatory proteins, is an allosteric protein, with two alternative shapes: active and inactive (see Figure 8.20). The *trp* repressor is synthesized in the inactive form, which has little affinity for the *trp* operator. Only when a tryptophan molecule binds to the *trp* repressor at an allosteric site does the repressor protein change to the active form that can attach to the operator, turning the operon off.

Tryptophan functions in this system as a **corepressor**, a small molecule that cooperates with a repressor protein to switch an operon off. As tryptophan accumulates, more tryptophan molecules associate with *trp* repressor molecules, which can then bind to the *trp* operator and shut down production of the tryptophan pathway enzymes. If the cell's tryptophan level drops, many fewer *trp* repressor proteins would have tryptophan bound, rendering them inactive; they would dissociate from the operator, allowing transcription of the operon's genes to resume. The *trp* operon is one example of how gene expression can respond to changes in the cell's internal and external environment.

Repressible and Inducible Operons: Two Types of Negative Gene Regulation

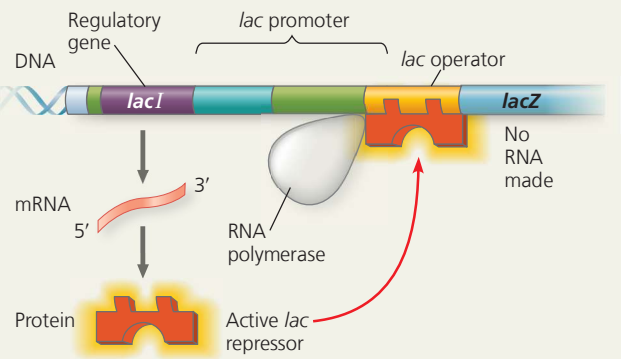
The *trp* operon is said to be a *repressible operon* because its transcription is usually on but can be inhibited (repressed) when a specific small molecule (in this case, tryptophan) binds allosterically to a regulatory protein. In contrast, an *inducible operon* is usually off but can be stimulated (induced) to be on when a specific small molecule interacts with a different regulatory protein. The classic example of an inducible operon is the *lac* operon (*lac* stands for "lactose").

The disaccharide lactose (milk sugar) is available to *E. coli* when the bacterium is in contact with any dairy product. Lactose metabolism by *E. coli* begins with hydrolysis of the disaccharide into its component monosaccharides (glucose and galactose), a reaction catalyzed by the enzyme β-galactosidase. Only a few molecules of this enzyme are present in an *E. coli* cell growing in the absence of lactose. If lactose is added to the bacterium's environment, however, the number of β-galactosidase molecules in the cell increases 1,000-fold within about 15 minutes. How can a cell ramp up enzyme production this quickly?

The gene for β-galactosidase (*lacZ*) is part of the *lac* operon, which includes two other genes coding for enzymes that function in the use of lactose **(Figure 18.4)**. The entire transcription unit is under the command of one main operator and promoter. The regulatory gene, *lacI*, located outside the *lac* operon, codes for an allosteric repressor protein that can switch off the *lac* operon by binding to the *lac* operator. So far, this sounds just like regulation of the *trp* operon, but there is one important difference. Recall that the *trp* repressor protein is inactive by itself and requires tryptophan as a corepressor in order to bind to the operator. The *lac* repressor, in contrast, is active by itself, binding to the operator and switching the *lac* operon off. In this case, a specific small molecule, called an **inducer**, *inactivates* the repressor.

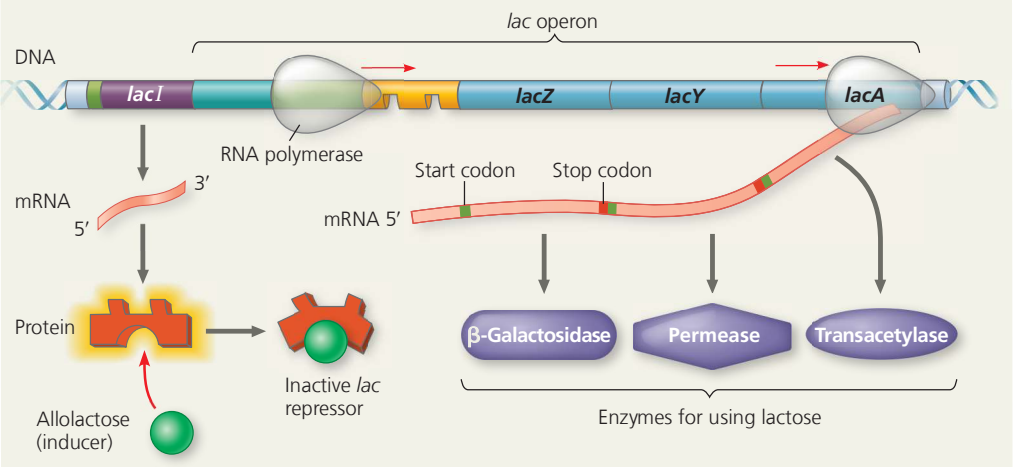
For the *lac* operon, the inducer is allolactose, an isomer of lactose formed in small amounts from lactose that enters the cell. In the absence of lactose (and therefore allolactose), the *lac* repressor is in its active shape and binds to the operator; thus, the genes of the *lac* operon are silenced (Figure 18.4a). If lactose is added to the cell's surroundings, allolactose binds to the *lac* repressor and alters its shape so the repressor can no longer bind to the operator. Without the *lac* repressor bound, the *lac* operon is transcribed into mRNA, and the enzymes for using lactose are made (Figure 18.4b).

In the context of gene regulation, the enzymes of the lactose pathway are referred to as *inducible enzymes* because their synthesis is induced by a chemical signal (allolactose, in this case). Analogously, the enzymes for tryptophan synthesis are said to be repressible. *Repressible enzymes* generally function in anabolic pathways, which synthesize essential end



(a) Lactose absent, repressor active, operon off. The *lac* repressor is innately active, and in the absence of lactose it switches off the operon by binding to the operator. The enzymes for using lactose are not made.

Mastering Biology
Animation: The *lac* Operon in *E. coli*



(b) Lactose present, repressor inactive, operon on. Allolactose, an isomer of lactose, binds to the repressor, inactivating it and "derepressing" the operon. The inactive repressor cannot bind to the operator, and so the genes of the *lac* operon are transcribed, and the enzymes for using lactose are made.

◀ Figure 18.4 The *lac* operon in *E. coli*: regulated synthesis of inducible enzymes. *E. coli* uses three enzymes to take up and metabolize lactose, the genes for which are clustered in the *lac* operon. The first gene, *lacZ*, codes for β-galactosidase, which hydrolyzes lactose to glucose and galactose. The second, *lacY*, codes for a permease, the membrane protein that transports lactose into the cell. The third, *lacA*, codes for transacetylase, an enzyme that detoxifies other molecules entering the cell via the permease. Unusually, the gene for the *lac* repressor, *lacI*, is adjacent to the *lac* operon. The function of the teal region within the promoter will be revealed in Figure 18.5.

products from raw materials (precursors). By suspending production of an end product when it is already present in sufficient quantity, the cell can allocate its organic precursors and energy for other uses. In contrast, inducible enzymes usually function in catabolic pathways, which break down a nutrient to simpler molecules. By producing the appropriate enzymes only when the nutrient is available, the cell avoids wasting energy and precursors making proteins that are not needed.

Regulation of both the *trp* and *lac* operons involves the *negative* control of genes because the operons are switched off by the active form of their respective repressor protein. It may be easier to see this for the *trp* operon, but it is also true for the *lac* operon. In the case of the *lac* operon, allolactose induces enzyme synthesis not by directly activating the *lac* operon, but by freeing it from the negative effect of the repressor (see Figure 18.4b). Gene regulation is said to be *positive* only when a regulatory protein interacts directly with the genome to increase transcription.

Positive Gene Regulation

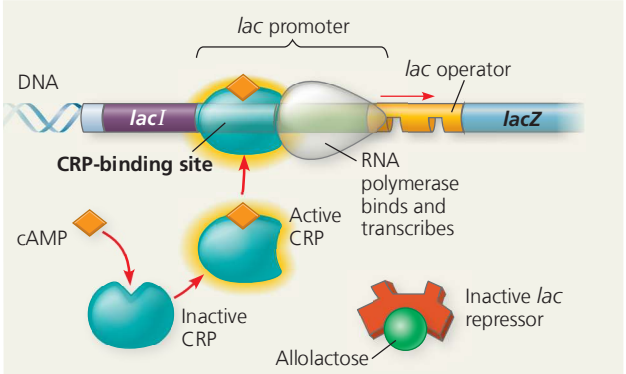
When glucose and lactose are both present in its environment, *E. coli* preferentially uses glucose. The enzymes for glucose breakdown in glycolysis (see Figure 9.8) are continually present. Only when lactose is present *and* glucose is in short supply does *E. coli* use lactose as an energy source, and only then does it synthesize appreciable quantities of the enzymes for lactose breakdown.

How does the *E. coli* cell sense the glucose concentration and relay this information to the *lac* operon? Again, the mechanism depends on the interaction of an allosteric regulatory protein with a small organic molecule, **cyclic AMP (cAMP)** in this case, which accumulates when glucose is scarce (see Figure 11.11 for the structure of cAMP). The regulatory protein, called *cAMP receptor protein* (*CRP*), is an **activator**, a protein that binds to DNA and stimulates transcription of a gene. (CRP is also called *catabolite activator protein*, or CAP.) When cAMP binds to this regulatory protein, CRP assumes its active shape and can attach to a specific site at the upstream end of the *lac* promoter **(Figure 18.5a)**. This attachment increases the affinity of RNA polymerase for the *lac* promoter, which is actually rather low even when no *lac* repressor is bound to the operator. By facilitating the binding of RNA polymerase to the promoter and thereby increasing the rate of transcription of the *lac* operon, the attachment of CRP to the promoter directly stimulates gene expression. Therefore, this mechanism qualifies as positive regulation.

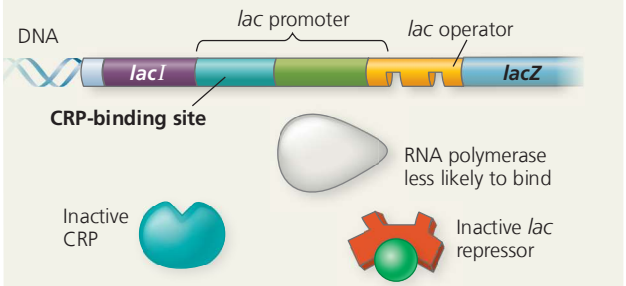
If the amount of glucose in the cell increases, the cAMP concentration falls, and without cAMP, CRP detaches from the *lac* operon. Because CRP is inactive, RNA polymerase binds less efficiently to the promoter, and transcription of the *lac* operon proceeds only at a low level, even when lactose is present **(Figure 18.5b)**. Thus, the *lac* operon is under dual control: negative control by the *lac* repressor and positive control by CRP. Whether or not transcription occurs is controlled by allolactose: Without allolactose, the *lac* repressor is active and the operon is off (transcription does not occur; see Figure 18.4a); with allolactose, the *lac* repressor is inactive and the operon is on (transcription occurs; see Figure 18.4b). The *rate* of transcription is controlled by whether CRP has cAMP bound to it: With bound cAMP, the rate is high; without it, the rate is low. It is as though the operon has both an on-off switch and a volume control.

In addition to regulating the *lac* operon, CRP helps regulate other operons that encode enzymes used in catabolic pathways. All told, it may affect the expression of more than 100 genes in *E. coli*. When glucose is plentiful and CRP is inactive, the synthesis of enzymes that catabolize compounds other than glucose generally slows down. The ability to catabolize other compounds, such as lactose, enables a cell deprived of

▼ Figure 18.5 Positive control of the lac operon by cAMP receptor protein (CRP). RNA polymerase has high affinity for the *lac* promoter only when CRP is bound to a DNA site at the upstream end of the promoter. CRP, in turn, attaches to its DNA site only when associated with cyclic AMP (cAMP), whose concentration in the cell rises when the glucose concentration falls. Thus, when glucose is present, even if lactose is also available, the cell preferentially catabolizes glucose and makes very low levels of the enzymes for using lactose.



(a) Lactose present, glucose scarce (cAMP level high): abundant *lac* mRNA synthesized. If glucose is scarce, the high level of cAMP activates CRP, which binds to the promoter and increases RNA polymerase binding there. The *lac* operon produces large amounts of mRNA coding for the enzymes that the cell needs for use of lactose.



(b) Lactose present, glucose present (cAMP level low): little *lac* mRNA synthesized. When glucose is present, cAMP is scarce, and CRP is unable to stimulate transcription at a significant rate, even though no repressor is bound.

glucose to survive. The compounds present in any given cell at a certain moment determine which operons are switched on—the result of simple interactions of activator and repressor proteins with the promoters of the genes in question.

CONCEPT CHECK 18.1

1. How does binding of the *trp* corepressor to the *trp* repressor alter repressor function and transcription? How does binding of the *lac* inducer alter the function of the *lac* repressor?
2. Describe the binding of repressors and activators to the *lac* operon when both lactose and glucose are scarce. What is the effect of these scarcities on transcription of the *lac* operon?
3. **WHAT IF?** A certain mutation in *E. coli* changes the *lac* operator so that the active repressor cannot bind. How would this affect the cell's production of β-galactosidase?

For suggested answers, see Appendix A.

CONCEPT 18.2

Eukaryotic gene expression is regulated at many stages

All organisms, whether prokaryotes or eukaryotes, must regulate which genes are expressed at any given time. Both unicellular organisms and the cells of multicellular organisms continually turn genes on and off in response to signals from their external and internal environments. Regulation of gene expression is also essential for cell specialization in multicellular organisms, which are made up of different types of cells. To perform its own distinct role, each cell type must maintain a specific program of gene expression in which certain genes are expressed and others are not (see Figure 18.1).

Differential Gene Expression

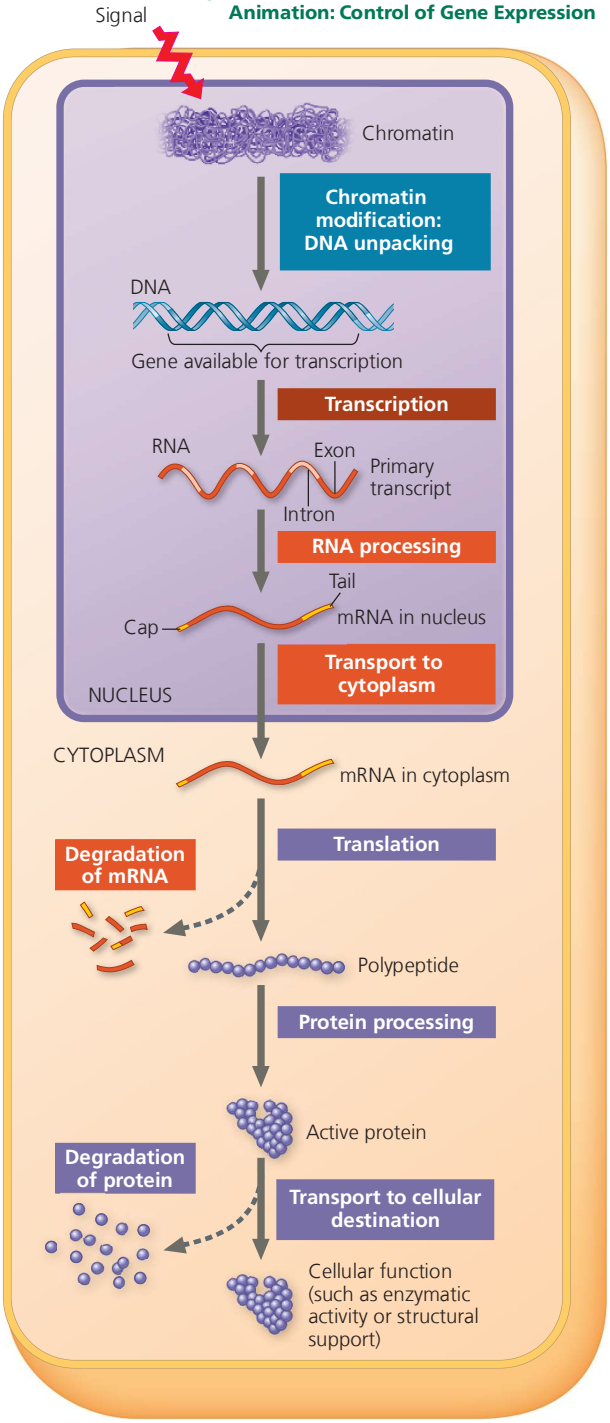
A typical human cell might express about a third to a half of its protein-coding genes at any given time. Highly differentiated cells, such as muscle or nerve cells, express a smaller fraction of their genes. Almost all the cells in a multicellular organism contain an identical genome. (Cells of the immune system are one exception, as you will see in Figure 43.13.) A subset of genes is expressed in each cell type; some of these—about 35%—are "housekeeping" genes, expressed by many cell types, while others are unique to that cell type. The uniquely expressed genes allow these cells to carry out their specific function. The differences between cell types, therefore, are due not to different genes being present, but to **differential gene expression**, the expression of different genes by cells with the same genome.

The function of any cell depends on its expressing the appropriate set of genes. The transcription factors of a cell must locate the right genes at the right time, a task like finding a needle in a haystack. Abnormal gene expression can cause serious imbalances and diseases, including cancer.

Figure 18.6 summarizes the process of gene expression in a eukaryotic cell, highlighting key stages in the expression of a protein-coding gene. Each stage depicted in Figure 18.6

▼ Figure 18.6 Stages in gene expression that can be regulated in eukaryotic cells. In this diagram, the colored boxes indicate the processes most often regulated; each color indicates the type of molecule that is affected (blue = DNA, red/orange = RNA, purple = protein). The nuclear envelope separating transcription from translation in eukaryotic cells allows for RNA processing, a form of post-transcriptional control that is not possible in prokaryotes. Eukaryotes also have a greater variety of control mechanisms operating before transcription and after translation. Some genes are regulated at multiple stages.

Mastering Biology
Animation: Control of Gene Expression



is a potential control point at which gene expression can be turned on or off, accelerated, or slowed down. Many genes have more than one control point.

When the structure of DNA was determined in 1953, an understanding of the mechanisms that control gene expression in eukaryotes seemed almost hopelessly out of reach. Since then, advances in DNA technology (see Chapter 20) have enabled molecular biologists to uncover many details of eukaryotic gene regulation. In all organisms, gene expression is commonly controlled at transcription; regulation at this stage often occurs in response to signals coming from outside the cell, such as hormones or other signaling molecules. For this reason, the term *gene expression* is often equated with transcription for both bacteria and eukaryotes. While this may be the case for bacteria, the greater complexity of eukaryotic cell structure and function provides opportunities for regulating gene expression at many stages besides transcription (see Figure 18.6). Let's examine some of the important control points of eukaryotic gene expression more closely.

Regulation of Chromatin Structure

Recall that the DNA of eukaryotic cells is packaged with proteins in an elaborate complex known as chromatin, the basic unit of which is the nucleosome (see Figure 16.23). The structural organization of chromatin not only packs a cell's DNA into a compact form that fits inside the nucleus, but also helps regulate gene expression in several ways. Genes within heterochromatin, which is more densely arranged than euchromatin, are usually not expressed. In euchromatin, whether or not a gene is transcribed is affected by the location of nucleosomes along a gene's promoter and also the sites where the DNA attaches to the protein scaffolding of the chromosome.

Chromatin structure and gene expression can be influenced by chemical modifications of both the histone proteins of the nucleosomes around which DNA is wrapped and the nucleotides that make up that DNA. Here we examine the effects of these modifications, which are catalyzed by specific enzymes.

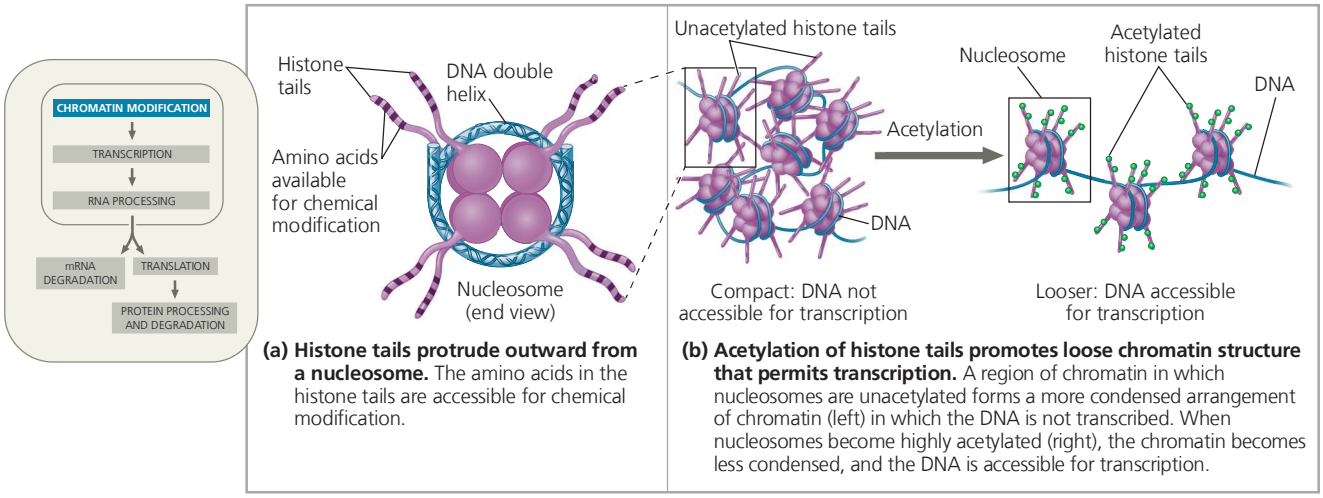
Histone Modifications and DNA Methylation

Chemical modifications to histones, found in all eukaryotic organisms, play a direct role in the regulation of gene transcription. The N-terminus of each histone protein in a nucleosome protrudes outward from the nucleosome **(Figure 18.7a)**. These so-called *histone tails* are accessible to various modifying enzymes that catalyze the addition or removal of specific chemical groups, such as acetyl (—$COCH_3$), methyl, and phosphate groups (see Figure 4.9). Generally, **histone acetylation**—the addition of an acetyl group to an amino acid in a histone tail—appears to promote transcription by opening up chromatin structure **(Figure 18.7b)**, while the addition of methyl groups to histones can lead to the condensation of chromatin and reduced transcription. Often, the addition of a particular chemical group may create a new binding site for enzymes that further modify chromatin structure.

Rather than modifying histone proteins, a different set of enzymes can methylate the DNA itself on certain bases, usually cytosine. Such **DNA methylation** occurs in most plants, animals, and fungi. Long stretches of inactive DNA, such as that of inactivated mammalian X chromosomes (see Figure 15.8), are generally more methylated than regions of actively transcribed DNA. On a smaller scale, the DNA of individual genes is usually more heavily methylated in cells in which those genes are not expressed. Removal of the extra methyl groups can turn on some of these genes.

▼ Figure 18.7 A simple model of histone tails and the effect of histone acetylation. Amino acids in histone tails may be chemically modified by addition of acetyl groups (green balls) or other groups (such as methyl or phosphate groups). Such modifications affect chromatin structure in a region, sometimes by providing binding sites for other chromatin-modifying enzymes.

Mastering Biology **HHMI Animation: Role of MECP2 in Gene Silencing**



(a) Histone tails protrude outward from a nucleosome. The amino acids in the histone tails are accessible for chemical modification.

(b) Acetylation of histone tails promotes loose chromatin structure that permits transcription. A region of chromatin in which nucleosomes are unacetylated forms a more condensed arrangement of chromatin (left) in which the DNA is not transcribed. When nucleosomes become highly acetylated (right), the chromatin becomes less condensed, and the DNA is accessible for transcription.

Once methylated, genes usually stay that way through successive cell divisions in a given individual. At DNA sites where one strand is already methylated, enzymes methylate the correct daughter strand after each round of DNA replication. Methylation patterns are thus passed on to daughter cells, and cells forming specialized tissues keep a chemical record of what occurred during embryonic development. A methylation pattern maintained in this way also accounts for *genomic imprinting* in mammals, where methylation permanently regulates expression of either the maternal or paternal allele of particular genes at the start of development (see Figure 15.17). DNA methylation and histone modification are believed to be coordinated in their regulation.

Epigenetic Inheritance

The chromatin modifications that we just discussed do not change the DNA sequence, yet they still may be passed along to future generations of cells. Inheritance of traits transmitted by mechanisms not involving the nucleotide sequence itself is called *epigenetic inheritance*, the study of which is called **epigenetics**. Whereas mutations in the DNA are permanent changes, modifications to the chromatin can be reversed. For example, DNA methylation patterns are largely erased during gamete formation and reestablished during embryonic development. Furthermore, they are changeable, thus responding more rapidly to environmental conditions.

Research on epigenetics has skyrocketed over the past 20 years. The importance of epigenetic information in regulating gene expression is now widely accepted. One key study by Robert Waterland and Randy Jirtle at Duke University, published in 2003, used a mouse mutant whose genome had been altered so that a gene called *agouti* that determines coat color, normally expressed only briefly during fur formation, was instead expressed throughout development. This overexpression resulted in yellow mice (**Figure 18.8a**, left mouse) rather than the usual brownish ("agouti") color. Previous work had shown that merely supplementing the diet of pregnant mothers with methyl group–containing compounds (like folic acid) could shift the range of coat colors of the offspring back to normal (see Figure 18.8a, right mouse). Waterland and Jirtle reproduced this result, analyzing the state of methylation of the DNA. They showed that the extent of the color shift correlated with the level of DNA methylation. In other words, feeding methyl groups to the mothers at a key time during gestation led to a change in gene expression in the offspring's phenotype. Further studies showed that the effects were even observed in the next generation—the "grandpups" of the original female mouse.

A similar epigenetic effect due to changes in methylation occurs in humans as well. Near the end of World War II, during the winter of 1944–45, Dutch railway workers went on strike to try to prevent the Nazis from bringing in more troops. In retaliation, the Nazis blocked all deliveries of food. Over 20,000 Dutch died in the "Dutch Hunger Winter"

▼ Figure 18.8 Examples of epigenetic inheritance.



(a) Effects of maternal diet on genetically identical mice. These two mice are genetically identical (*agouti* mutants), but their mothers were fed different diets. A maternal diet without compounds that donate methyl groups resulted in the obese, yellow offspring on the left. A diet supplemented with methyl-donating compounds, such as folic acid, led to the speckled brown fur (normal *agouti* phenotype) and normal weight of the mouse on the right.



(b) The Dutch Hunger Winter. During the Nazi blockade of food transport into the Netherlands near the end of World War II, the population was starving, including this mother and her two children. Children of mothers who were pregnant during that time have experienced long-term effects.

(Figure 18.8b). Over time, doctors found that the offspring of women in early pregnancy at that time experienced adverse health effects as adults: higher rates of obesity, high triglyceride and cholesterol levels, type 2 diabetes, and schizophrenia. Furthermore, the affected individuals had a 10% higher mortality rate after the age of 68 than their siblings who were *in utero* when food was readily available. A collaboration between labs in the Netherlands and the United States compared those adults with their siblings and published their findings in 2018. Statistical analysis enabled the researchers to conclude that differences between the siblings in DNA methylation of certain genes caused these long-term adverse medical conditions—examples of epigenetic inheritance.

Epigenetic variations might help explain cases where one identical twin acquires a genetically based disease, such as schizophrenia, but the other does not, despite their identical genomes. Alterations in normal patterns of DNA methylation

are seen in some cancers, where they are associated with inappropriate gene expression. Evidently, enzymes that modify chromatin structure are integral parts of the eukaryotic cell's machinery for regulating transcription.

Regulation of Transcription Initiation

Chromatin-modifying enzymes provide initial control of gene expression by making a region of DNA either more or less able to bind the transcription machinery. Once the chromatin of a gene is optimally modified for expression, the initiation of transcription is the next major step at which gene expression is regulated. As in bacteria, the regulation of transcription initiation in eukaryotes involves proteins that bind to DNA and either facilitate or inhibit binding of RNA polymerase. The process is more complicated in eukaryotes, however. Before looking at how eukaryotic cells control their transcription, let's review the structure of a eukaryotic gene.

Organization of a Typical Eukaryotic Gene and Its Transcript

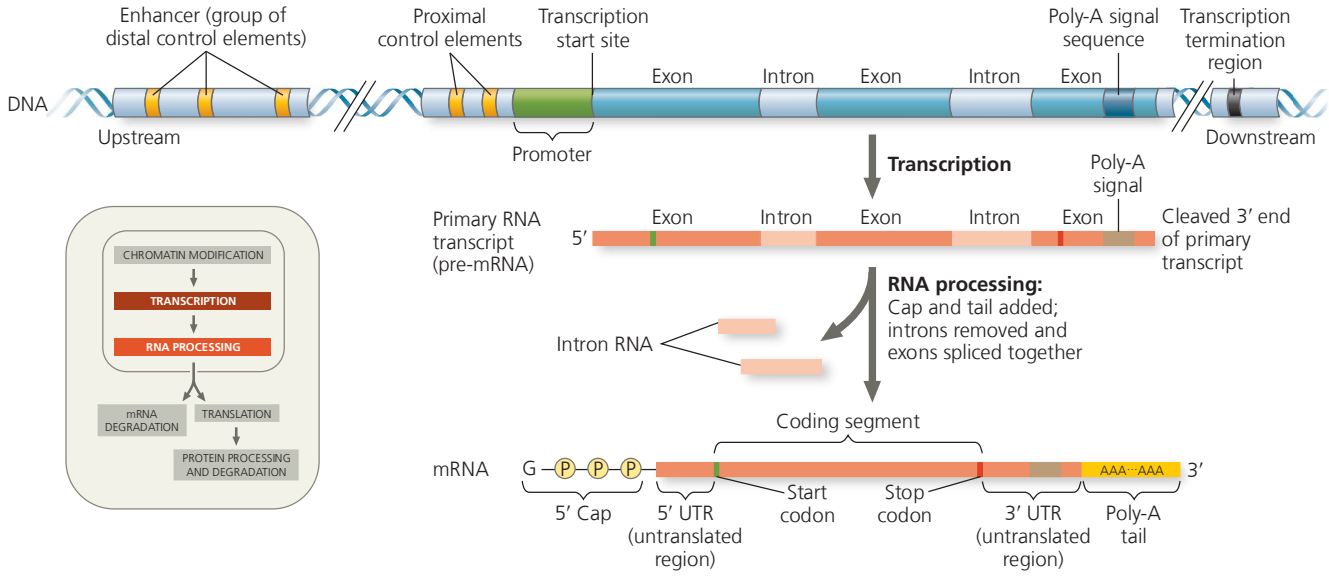
A eukaryotic gene and the DNA elements (segments) that control it are typically organized as shown in **Figure 18.9**, which extends what you learned about eukaryotic genes in Chapter 17. Recall that a cluster of proteins called a *transcription initiation complex* assembles on the promoter sequence at the "upstream" end of the gene (see Figure 17.9). One of these proteins, RNA polymerase II, then proceeds to transcribe the gene, synthesizing a primary RNA transcript (pre-mRNA). RNA processing includes enzymatic addition of a 5′ cap and a poly-A tail, as well as splicing out of introns, to yield a mature mRNA. Associated with most eukaryotic genes are multiple **control elements**, segments of noncoding DNA that serve as binding sites for the proteins called transcription factors, which bind to the control elements and regulate transcription. Control elements on the DNA and the transcription factors that bind to them are critical to the precise regulation of gene expression seen in different cell types.

The Roles of General and Specific Transcription Factors

There are two types of transcription factors: General transcription factors act at the promoter of all genes, while some genes require specific transcription factors that bind to control elements that may be close to or farther away from the promoter.

General Transcription Factors at the Promoter To initiate transcription, eukaryotic RNA polymerase II requires the assistance of transcription factors. Some transcription factors (such as those illustrated in Figure 17.9) are essential for the transcription of *all* protein-coding genes; therefore, they are often called *general transcription factors*. A few general transcription factors bind to a DNA sequence, such as the TATA box in most promoters, but many bind to proteins, including

▼ Figure 18.9 A eukaryotic gene and its transcript. Each eukaryotic gene has a promoter—a DNA sequence where RNA polymerase II binds and starts transcription, proceeding "downstream." A number of control elements (gold) are involved in regulating the initiation of transcription; these are DNA sequences located near (proximal to) or far from (distal to) the promoter. Distal control elements can be grouped together as enhancers, one of which is shown for this gene. At the other end of the gene, a polyadenylation (poly-A) signal sequence in the last exon of the gene is transcribed into an RNA sequence that signals where the transcript is cleaved and the poly-A tail added. Transcription may continue for hundreds of nucleotides beyond the poly-A signal before terminating. RNA processing of the primary transcript into a functional mRNA involves three steps: addition of the 5′ cap, addition of the poly-A tail, and splicing. In the cell, the 5′ cap is added soon after transcription is initiated, and splicing occurs while transcription is still under way (see Figures 17.11 and 17.12).



other transcription factors as well as RNA polymerase II. Protein-protein interactions are crucial to the initiation of eukaryotic transcription. Only when the complete initiation complex has assembled can the polymerase begin to move along the DNA template strand and transcribe.

Some genes are expressed all the time, but others are not; instead, they are regulated. For these genes, the interaction of general transcription factors and RNA polymerase II with a promoter usually leads to a low rate of initiation and production of few RNA transcripts from genes that are not expressed at significant levels all the time or in all cells. In eukaryotes, high levels of transcription of these particular genes at the appropriate time and place depend on the interaction of control elements with another set of proteins, which can be thought of as *specific transcription factors*.

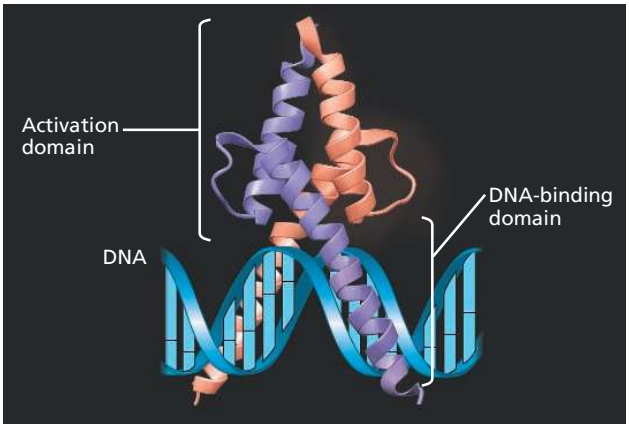
Enhancers and Specific Transcription Factors As you can see in Figure 18.9, some control elements, named *proximal control elements*, are close to the promoter. The more distant *distal control elements*, groupings of which are called **enhancers**, may be thousands of nucleotides upstream or downstream of a gene or even within an intron. A given gene may have multiple enhancers, each active at a different time, cell type, or location in the organism. Each enhancer, however, is generally associated with only that gene and no other.

In eukaryotes, the rate of gene expression can be strongly increased or decreased by the binding of specific transcription factors, either activators or repressors, to the control elements of enhancers. Hundreds of transcription activators have been discovered in eukaryotes; the structure of one is shown in **Figure 18.10**. Researchers have identified two types of structural domains that are commonly found in a large number of transcription activators: a *DNA-binding domain*—a part of the protein's three-dimensional structure that binds to DNA—and one or more *activation domains*. Activation domains bind other regulatory proteins or components of the transcription machinery, facilitating a series of protein-protein interactions that result in enhanced transcription of a given gene.

Figure 18.11 shows the currently accepted model for how binding of activators to an enhancer located far from the promoter influences transcription. Protein-mediated bending of the DNA brings the bound activators into contact with a group of *mediator proteins*, which in turn interact with general transcription factors at the promoter. These protein-protein interactions help assemble and position the initiation complex on the promoter. One of the studies supporting this model shows that proteins regulating one of the mouse globin genes contact both the gene's promoter and an enhancer located about 50,000 nucleotides upstream. Protein interactions allow these two DNA regions to come together in a very specific fashion, in spite of the many nucleotide pairs between them.

Specific transcription factors that function as repressors can inhibit gene expression in several different ways. Some repressors bind directly to control element DNA (in enhancers or elsewhere), blocking activator binding. Other repressors interfere with the activator itself so it can't bind the DNA.

▼ Figure 18.10 The structure of MyoD, a transcriptional activator. The MyoD protein is made up of two polypeptide subunits (purple and salmon) with extensive regions of α helix. Each subunit has one DNA-binding domain (lower half) and one activation domain (upper half). The latter includes binding sites for the other subunit and for other proteins. MyoD is involved in muscle development in vertebrate embryos (see Concept 18.4).



VISUAL SKILLS *Describe how the two functional domains of the MyoD protein relate to the two polypeptide subunits.*

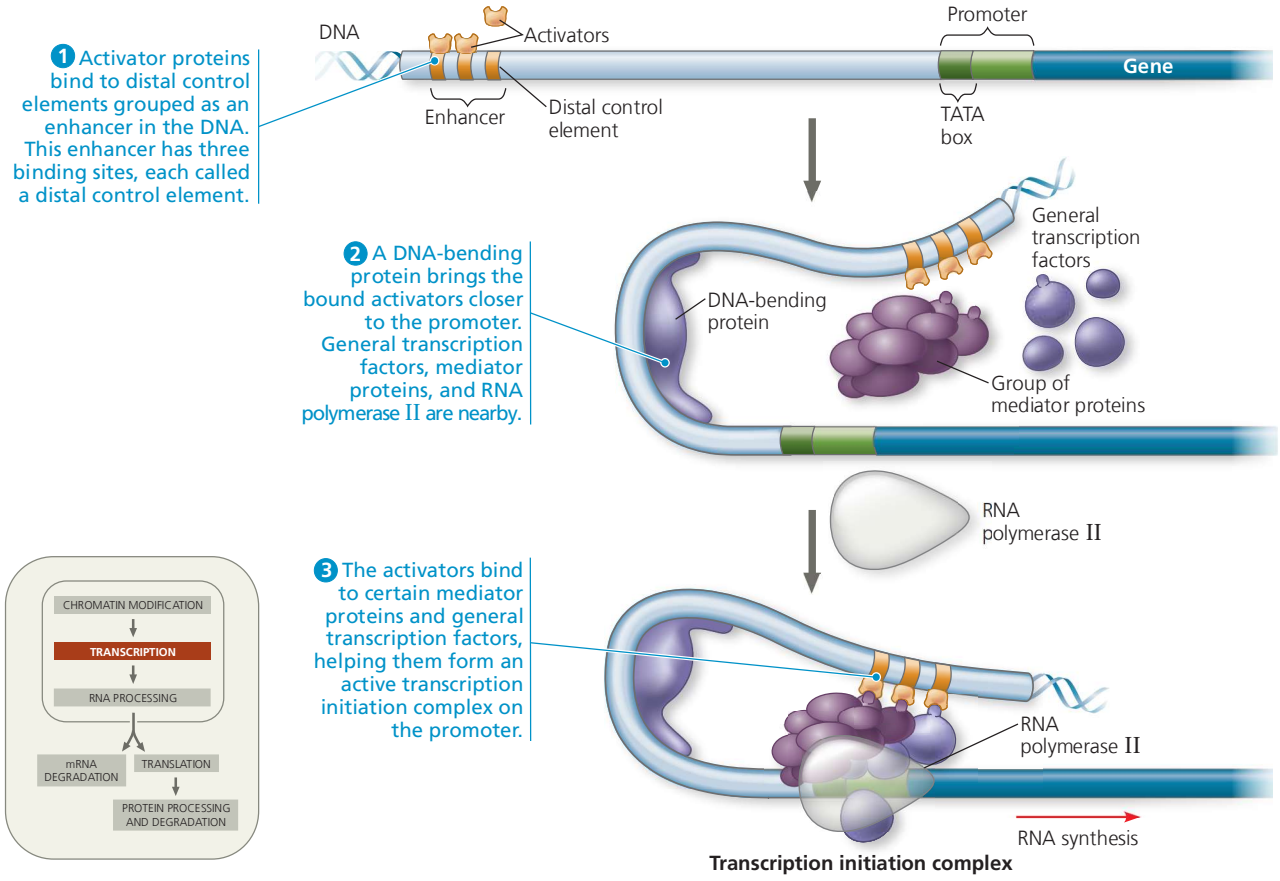
In addition to influencing transcription directly, some activators and repressors indirectly affect chromatin structure. Studies using yeast and mammalian cells show that some activators recruit proteins that acetylate histones near the promoters of specific genes, thus promoting transcription (see Figure 18.7). Similarly, some repressors recruit proteins that remove acetyl groups from histones, leading to reduced transcription, a phenomenon referred to as *silencing*. Indeed, recruitment of chromatin-modifying proteins seems to be the most common mechanism of repression in eukaryotic cells.

Mastering Biology HHMI Animation: Regulation of Eukaryotic DNA Transcription

Combinatorial Control of Gene Activation In eukaryotes, the precise control of transcription depends largely on the binding of activators to DNA control elements. Considering that many genes must be regulated in a typical animal or plant cell, the number of different nucleotide sequences in control elements is surprisingly small. A dozen or so short nucleotide sequences appear again and again in the control elements for different genes. On average, each enhancer is composed of about ten control elements, each binding only one or two specific transcription factors. It is the particular *combination* of control elements in an enhancer associated with a gene, rather than a single unique control element, that is important in regulating transcription of the gene.

Even with only a dozen control element sequences available, many combinations are possible. Each combination of control elements can activate transcription only when the appropriate

▼ Figure 18.11 A model for the action of enhancers and transcription activators. Bending of the DNA by a protein enables enhancers to influence a promoter hundreds or even thousands of nucleotides away. Specific transcription factors (activators) bind to the enhancer DNA sequences and then to a group of mediator proteins. These in turn bind to general transcription factors and then RNA polymerase II, thus assembling the transcription initiation complex. These protein-protein interactions lead to correct positioning of the complex on the promoter and the initiation of RNA synthesis. Only one enhancer (with three gold control elements) is shown here, but a gene may have several enhancers that act at different times or in different cell types.



Mastering Biology
Animation: Regulation of Transcription by Transcription Activators and Enhancers

transcription activators are present, which may occur at a precise time during development or in a particular cell type. In the **Scientific Skills Exercise**, you can work with data from an experiment that identified the control elements in an enhancer of a particular human gene. **Figure 18.12** illustrates how the use of different combinations of just a few control elements can allow differential regulation of transcription in two representative cell types—liver cells and lens cells. This can occur because each cell type contains a different group of transcription activators. How cell types come to differ during this process, even though they all arise from one cell (the fertilized egg), will be explored in Concept 18.4.

Coordinately Controlled Genes in Eukaryotes

How does the eukaryotic cell deal with a group of genes of related function that need to be turned on or off at the same time? Earlier in this chapter, you learned that in bacteria, such *coordinately controlled* genes are often clustered into an operon, which is regulated by a single promoter and transcribed into a single mRNA molecule. Thus, the genes are expressed together, and the encoded proteins are produced at the same time. With a few exceptions, operons that work in this way have *not* been found in eukaryotic cells.

Eukaryotic genes that are co-expressed, such as genes coding for the enzymes of a metabolic pathway, are typically scattered over different chromosomes. Here, coordinate gene expression depends on every gene of a dispersed group having a specific combination of control elements. Transcription activators in the nucleus that recognize the control elements bind to them, promoting simultaneous transcription of the genes, no matter where they are in the genome.

Coordinate control of dispersed genes in a eukaryotic cell often occurs in response to chemical signals from outside the cell. A steroid hormone, for example, enters a cell and binds to a specific

intracellular receptor protein, forming a hormone-receptor complex that serves as a transcription activator (see Figure 11.9). Every gene whose transcription is stimulated by a given steroid hormone, regardless of its chromosomal location, has a control element recognized by that hormone-receptor complex. This is how estrogen activates a group of genes that stimulate cell division in uterine cells, preparing the uterus for pregnancy.

Many signaling molecules, such as nonsteroid hormones and growth factors, bind to receptors on a cell's surface and never actually enter the cell. Such molecules can control gene expression indirectly by triggering signal transduction pathways that activate particular transcription factors (see Figure 11.15). Coordinate regulation in such pathways is the same as for steroid hormones: Genes with the same sets of control elements are activated by the same chemical signals. Because this system for coordinating gene regulation is so widespread, biologists think that it probably arose early in evolutionary history.

Scientific Skills Exercise

Analyzing DNA Deletion Experiments

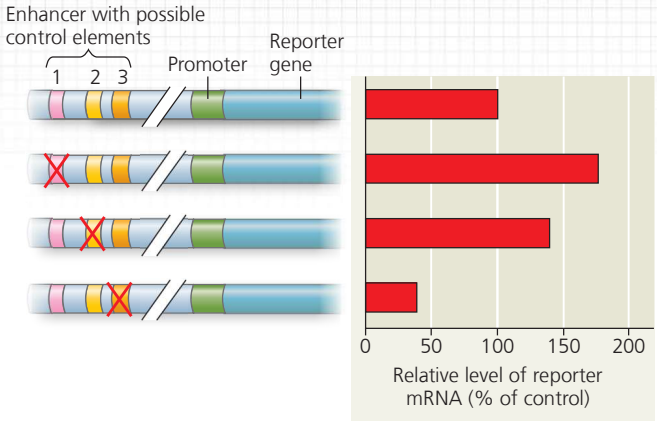


▲ Tissue inflammation

What Control Elements Regulate Expression of the *mPGES-1* Gene? The promoter of a gene includes the DNA immediately upstream of the transcription start site, but the control elements (grouped in an enhancer) that regulate the level of transcription of the gene may be thousands of base pairs upstream of the promoter. Because the distance and spacing of control elements make them difficult to identify, scientists begin by deleting possible control elements and measuring the effect on gene expression. In this exercise, you will analyze data obtained from DNA deletion experiments that tested possible control elements for the human gene *mPGES-1*. This gene codes for an enzyme that synthesizes a type of prostaglandin, a chemical made during tissue inflammation.

How the Experiment Was Done The researchers hypothesized that there were three possible control elements in an enhancer region located 8–9 kilobases (kb) upstream of the *mPGES-1* gene. Control elements regulate whatever gene is in the appropriate downstream location. Thus, to test the activity of the possible elements, researchers first synthesized molecules of DNA *(DNA constructs)* that had the intact enhancer region upstream of a *reporter gene*, a gene whose mRNA product could be easily measured experimentally. Next, they made three more DNA constructs, with one of the three proposed control elements deleted in each (see the left side of the figure). The researchers then introduced each DNA construct into a separate human cell culture, where the cells took up the DNA. After 48 hours, the amount of reporter gene mRNA made by the cells was measured. Comparing these amounts allowed researchers to determine if any of the deletions had an effect on expression of the reporter gene, mimicking the effect of deletions on *mPGES-1* gene expression. (The *mPGES-1* gene itself couldn't be used to measure expression levels because the cells express their own *mPGES-1* gene, and the mRNA from that gene would confuse the results.)

Data from the Experiment The diagrams on the left side of the figure show the intact DNA sequence (top) and the three experimental DNA constructs. A red X is located on the possible control element (1, 2, or 3) that was deleted in each experimental DNA construct. The area between the slashes represents the approximately 8 kb of DNA located between the promoter and the enhancer region. The horizontal bar graph on the right shows the amount of reporter gene mRNA that was present in each cell culture after 48 hours relative to the amount that was in the culture containing the intact enhancer region (top bar = 100%).

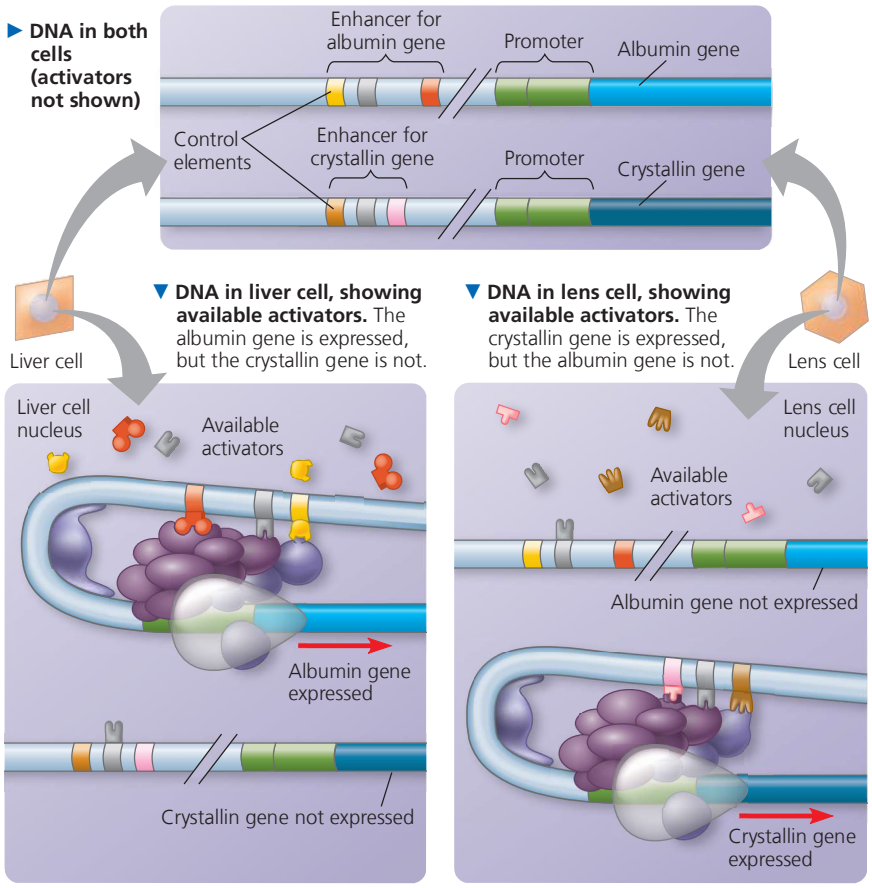


Data from J. N. Walters et al., Regulation of human microsomal prostaglandin E synthase-1 by IL-1b requires a distal enhancer element with a unique role for C/EBPb, *Biochemical Journal* 443:561–571 (2012).

INTERPRET THE DATA

1. **(a)** What is the independent variable in the graph? **(b)** What is the dependent variable? **(c)** What was the control treatment in this experiment? Label it on the diagram.
2. Do the data suggest that any of these possible control elements are actual control elements? Explain.
3. **(a)** Did deletion of any of the possible control elements cause a *reduction* in reporter gene expression? If so, which one(s), and how can you tell? **(b)** If loss of a control element causes a reduction in gene expression, what must be the normal role of that control element? Provide a biological explanation for how the loss of such a control element could lead to a reduction in gene expression.
4. **(a)** Did deletion of any of the possible control elements cause an *increase* in reporter gene expression relative to the control? If so, which one(s), and how can you tell? **(b)** If loss of a control element causes an increase in gene expression, what must be the normal role of that control element? Propose a biological explanation for how the loss of such a control element could lead to an increase in gene expression.

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.



▶ Figure 18.12 Cell type–specific transcription. Both liver cells and lens cells have the genes for making the proteins albumin and crystallin, but only liver cells make albumin (a blood protein) and only lens cells make crystallin (the main protein of the lens of the eye). The specific transcription factors made in a cell determine which genes are expressed. In this example, the genes for albumin and crystallin are shown at the top, each with an enhancer made up of three different control elements. Although the enhancers for the two genes both have a gray control element, each enhancer has a unique *combination* of elements. All the activator proteins required for high-level expression of the albumin gene are present in liver cells only (left), whereas the activators needed for expression of the crystallin gene are present in lens cells only (right). For simplicity, we consider only the role of specific transcription factors that are activators here, although repressors may also influence transcription in certain cell types.

VISUAL SKILLS *Describe the enhancer for the albumin gene in each type of cell. How would the nucleotide sequence of this enhancer in the liver cell compare with that in the lens cell?*

Nuclear Architecture and Gene Expression

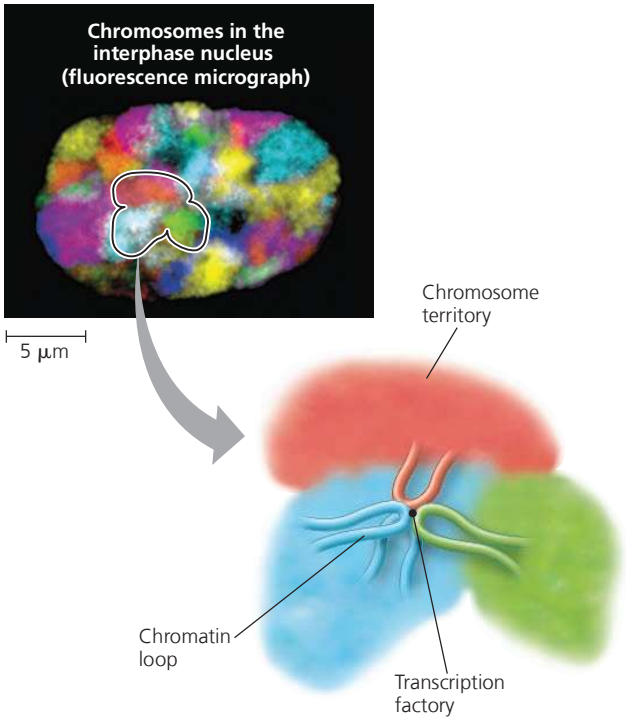
Each chromosome in the interphase nucleus of animal cells occupies a distinct territory (see Figure 16.24a). Recently, *chromosome conformation capture* techniques (*3C* and others) have been developed that allow researchers to cross-link and identify regions of chromosomes associating with each other during interphase. These studies reveal two organizational details: First, the territory of each chromosome is divided into regions of chromatin loops (*topologically associated domains,* or *TADs*), within which chromatin sites associate mainly with each other. Second, loops of chromatin, each likely a TAD, extend from individual chromosomal territories into specific sites in the nucleus **(Figure 18.13)**. Different loops from the same chromosome and loops from other chromosomes may congregate in such sites, some of which are rich in RNA polymerases and other transcription-associated proteins. Like a recreation center that draws members from many different neighborhoods, these so-called *transcription factories* are thought to be areas specialized for a common function.

The old view that the nuclear contents are like a bowl of amorphous chromosomal spaghetti has given way to a new model of a nucleus with a defined architecture and regulated movements of chromatin. Several lines of evidence suggest that unexpressed genes are located in the outer edges of the nucleus, while those that are being expressed are found in its interior region. Relocation of particular genes from their chromosomal territories to transcription factories in the interior may be part of the process of readying genes for transcription. In 2017, a consortium of researchers funded by the National Institutes of Health began investigating genome organization over time and its relationship to genome function.

Mechanisms of Post-transcriptional Regulation

Transcription alone does not constitute gene expression. The expression of a protein-coding gene is ultimately measured by the amount of functional protein a cell makes, and much happens between the synthesis of the RNA transcript and the activity of the protein in the cell. Many regulatory mechanisms operate at the various stages after transcription (see Figure 18.6). These mechanisms allow a cell to rapidly fine-tune gene expression in response to environmental changes without altering its transcription patterns. Here we explore how cells regulate gene expression after transcription.

▼ Figure 18.13 Chromosomal interactions in the interphase nucleus. Although each chromosome has its own territory (see Figure 16.24a), loops of chromatin may extend into other sites in the nucleus. (Each loop may be a topologically associated domain, or TAD.) Some of these sites are transcription factories that are occupied by multiple chromatin loops from the same chromosome (blue loops) or other chromosomes (red and green loops).



RNA Processing

RNA processing in the nucleus and the export of mature RNA to the cytoplasm provide opportunities for regulating gene expression not available in prokaryotes. One example of regulation at the RNA-processing level is **alternative RNA splicing**, in which different mRNA molecules are produced from the same primary transcript, depending on which RNA segments are treated as exons and which as introns. Regulatory proteins specific to a cell type control intron/exon choices by binding to regulatory sequences within the primary transcript.

A simple example of alternative RNA splicing is shown in **Figure 18.14** for the troponin T gene, which encodes multiple closely related proteins with slightly different effects on muscle contraction. Other genes code for many more products. For instance, researchers have found a *Drosophila* gene with enough alternatively spliced exons to generate about 19,000 membrane proteins that have different extracellular domains. At least 17,500 (94%) of the alternative mRNAs are actually synthesized. Each developing nerve cell in the fly appears to synthesize a different form of the protein, which acts as a unique identifier on the cell surface and helps prevent excessive overlap of nerve cells during development of the nervous system.

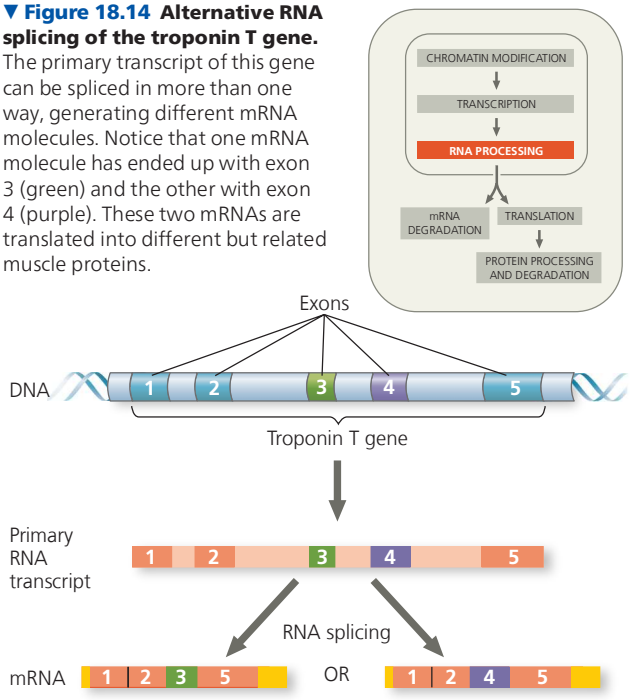
Alternative RNA splicing can significantly expand the repertoire of a eukaryotic genome. In fact, alternative splicing was proposed as one explanation for the surprisingly low number of human genes counted when the human genome was sequenced. The number of human genes was found to be similar to that of a soil worm (nematode), a mustard plant, or a sea anemone. This discovery prompted questions about what, if not the number of genes, accounts for the more complex form and structure of humans. More than 90% of human protein-coding genes likely undergo alternative splicing. Thus, the extent of alternative splicing greatly multiplies the number of possible human proteins, which may be better correlated with complexity of form than the number of genes.

Initiation of Translation and mRNA Degradation

Translation is another stage where gene expression is regulated, most commonly at the initiation stage (see Figure 17.19). For some mRNAs, the initiation of translation can be blocked by regulatory proteins that bind to specific sequences or structures within the untranslated region (UTR) at the 5′ or 3′ end, preventing the attachment of ribosomes. (Recall from Concept 17.3 that both the 5′ cap and the poly-A tail of an mRNA molecule are important for ribosome binding.)

Alternatively, translation of *all* the mRNAs in a cell may be regulated simultaneously. In a eukaryotic cell, such "global" control usually involves the activation or inactivation of one or more protein factors required to initiate translation. This mechanism plays a role in starting translation of mRNAs that are stored in eggs. Just after

▼ Figure 18.14 Alternative RNA splicing of the troponin T gene. The primary transcript of this gene can be spliced in more than one way, generating different mRNA molecules. Notice that one mRNA molecule has ended up with exon 3 (green) and the other with exon 4 (purple). These two mRNAs are translated into different but related muscle proteins.



fertilization, translation is triggered by the sudden activation of translation initiation factors. The response is a burst of synthesis of the proteins encoded by the stored mRNAs. Some plants and algae store mRNAs during periods of darkness; light then triggers the reactivation of the translational apparatus.

The life span of mRNA molecules in the cytoplasm is important in determining the pattern of protein synthesis in a cell. Bacterial mRNA molecules typically are degraded by enzymes within a few minutes. This short life span of mRNAs is one reason bacteria can change their patterns of protein synthesis so quickly in response to environmental changes. In contrast, some mRNAs in multicellular eukaryotes typically survive for hours, days, or even weeks. For instance, the mRNAs for the hemoglobin polypeptides (α-globin and β-globin) in developing red blood cells are unusually stable, and these long-lived mRNAs are translated repeatedly in red blood cells.

Nucleotide sequences that affect how long an mRNA remains intact are often found in the untranslated region at the 3′ end of the molecule (see Figure 18.9). In one experiment, researchers transferred such a sequence from the short-lived mRNA for a growth factor to the 3′ end of a normally stable globin mRNA. The globin mRNA was quickly degraded.

Other mechanisms that degrade or block expression of mRNA molecules have come to light. They involve a group of recently discovered RNA molecules that regulate gene expression at several levels, as you'll see shortly.

Protein Processing and Degradation

The final opportunities for controlling gene expression occur after translation. Often, eukaryotic polypeptides must be processed to yield functional protein molecules. For instance, cleavage of the initial insulin polypeptide (pro-insulin) forms the active hormone. In addition, many proteins undergo chemical modifications that make them functional. Regulatory proteins are commonly activated or inactivated by the reversible addition of phosphate groups (see Figure 11.10), and proteins destined for the surface of animal cells acquire sugars (see Figure 6.12). Cell-surface proteins and many others must also be transported to target destinations in the cell in order to function (see Figure 17.22). Regulation might occur at any of the steps involved in modifying or transporting a protein.

Finally, the length of time each protein functions in the cell is strictly regulated by selective degradation. Many proteins, such as the cyclins involved in regulating the cell cycle, must be relatively short-lived if the cell is to function appropriately (see Figure 12.16). To mark a protein for destruction, the cell commonly attaches molecules of a small protein called ubiquitin to the protein. Giant protein complexes called proteasomes then recognize the ubiquitin tagged proteins and degrade them.

Mastering Biology
Animation: Post-Transcriptional Control Mechanisms

CONCEPT CHECK 18.2

1. In general, what are the effects of histone acetylation and DNA methylation on gene expression?
2. **MAKE CONNECTIONS** Speculate about whether the same enzyme could methylate both a histone and a DNA base. (See Concept 8.4.)
3. Compare the roles of general and specific transcription factors in regulating gene expression.
4. Once mRNA encoding a particular protein reaches the cytoplasm, what are four mechanisms that can regulate the amount of the protein that is active in the cell?
5. **WHAT IF?** Suppose you compared the nucleotide sequences of the distal control elements in the enhancers of three genes that are expressed only in muscle cells. What would you expect to find? Why?

For suggested answers, see Appendix A.

CONCEPT 18.3

Noncoding RNAs play multiple roles in controlling gene expression

Genome sequencing has revealed that protein-coding DNA accounts for only 1.5% of the human genome and a similarly small percentage of the genomes of many other multicellular eukaryotes. A very small fraction of the non-protein-coding DNA consists of genes for RNAs such as ribosomal RNA and transfer RNA. Until recently, scientists assumed that most of the remaining DNA was not transcribed, thinking that since it didn't specify proteins or the few known types of RNA, such DNA didn't contain meaningful genetic information—in fact, it was called "junk DNA." However, some genomic studies have cast doubt on this description. For example, a massive study showed that roughly 75% of the human genome is transcribed at some point in any given cell. Introns account for only a fraction of this transcribed RNA, most of which is untranslated. The verdict is still out on how much of the transcribed RNA is functional, but at least some of the genome is transcribed into non-protein-coding RNAs (also called *noncoding RNAs*, or *ncRNAs*), including a variety of small RNAs. Researchers are uncovering more evidence of the biological roles of these ncRNAs every day.

These discoveries have revealed a large and diverse population of RNA molecules in the cell that play crucial roles in regulating gene expression and have gone largely unnoticed until recently. The longstanding view that mRNAs are the most important RNAs because they code for proteins needs revision. This represents a major shift in thinking by biologists, one that you are witnessing as students entering this field of study.

Effects on mRNAs by MicroRNAs and Small Interfering RNAs

Regulation by both small and large ncRNAs occurs at several points in the pathway of gene expression, including mRNA translation and chromatin modification. We'll examine

two types of small ncRNAs, the importance of which was acknowledged when their discovery was the focus of the 2006 Nobel Prize in Physiology or Medicine, which was awarded for work completed only eight years earlier.

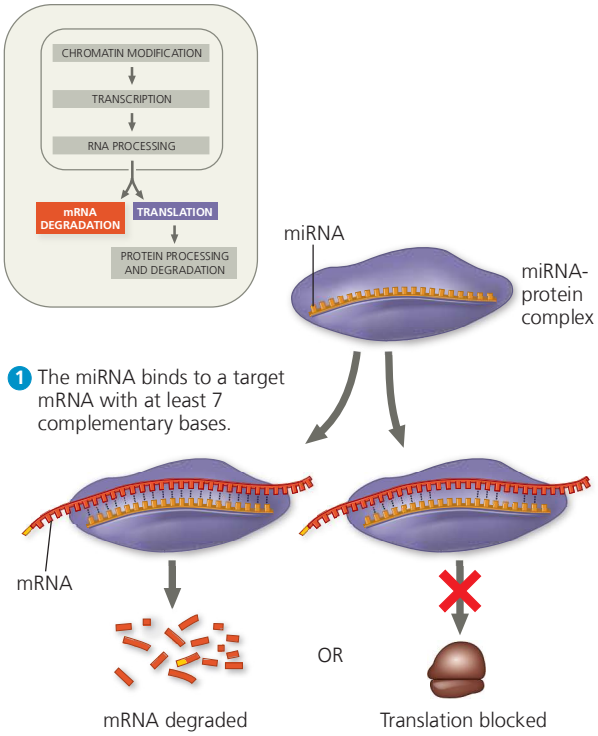
Since 1993, a number of research studies have uncovered **microRNAs (miRNAs)**—small, single-stranded RNA molecules capable of binding to complementary sequences in mRNA molecules. A longer RNA precursor is processed by cellular enzymes into an miRNA, a single-stranded RNA of about 22 nucleotides that forms a complex with one or more proteins **(Figure 18.15)**. The miRNA allows the complex to bind to any mRNA molecule with at least seven or eight nucleotides of complementary sequence. The miRNA-protein complex then degrades the target mRNA or, less often, simply blocks its translation. There are approximately 1,500 genes for miRNAs in the human genome, and biologists estimate that expression of at least one-half of all human genes may be regulated by miRNAs, a remarkable figure given that the existence of miRNAs was unknown until the early 1990s.

Another class of small noncoding RNAs, similar in size and function to miRNAs, is called **small interfering RNAs (siRNAs)**. Both miRNAs and siRNAs can associate with the same proteins, producing similar results. In fact, if siRNA precursor RNA molecules are injected into a cell, the cell's machinery can process them into siRNAs that turn off expression of genes with related sequences, similarly to how miRNAs function. The distinction between miRNAs and siRNAs is based on subtle differences in the structure of their precursors, which in both cases are RNA molecules that are mostly double-stranded. The blocking of gene expression by siRNAs, referred to as **RNA interference (RNAi)**, is used in the laboratory as a means of disabling specific genes to investigate their function.

EVOLUTION How did the RNAi pathway evolve? As you will learn in Concept 19.2, some viruses have double-stranded RNA genomes. Given that the cellular RNAi pathway can process double-stranded RNAs into homing devices that lead to destruction of related RNAs, some scientists think that this pathway may have evolved as a natural defense against infection by such viruses. However, the fact that RNAi can also affect the expression of nonviral cellular genes may reflect a different evolutionary origin for the RNAi pathway.

While this section has focused on ncRNAs in eukaryotes, small ncRNAs are also used by bacteria as a defense system, called the CRISPR-Cas9 system, against viruses that infect them (see Concept 17.5). The use of ncRNAs thus evolved long ago, but we don't yet know how bacterial ncRNAs are related to those of eukaryotes.

▼ Figure 18.15 Regulation of gene expression by microRNAs (miRNAs). An miRNA of about 22 nucleotides, formed by enzymatic processing of an RNA precursor, associates with one or more proteins in a complex that can degrade or block translation of target mRNAs.



Chromatin Remodeling and Effects on Transcription by ncRNAs

In addition to regulating mRNAs, small noncoding RNAs can cause remodeling of chromatin structure. In the S phase of the cell cycle, for example, the centromeric regions of DNA must be loosened for chromosomal replication and then recondensed into heterochromatin in preparation for mitosis. In some yeasts, siRNAs produced by the yeast cells from the centromeric DNA are required to re-form the heterochromatin at the centromeres. Exactly how the process starts is still debated, but biologists agree on the general idea: The siRNA system in yeast interacts with other, larger noncoding RNAs and with chromatin-modifying enzymes to condense the centromere chromatin into heterochromatin. In most mammalian cells, siRNAs have not been found, and the mechanism for centromere DNA condensation is not yet understood. However, it may also turn out to involve small noncoding RNAs.

A recently discovered class of small ncRNAs called *piwi-interacting RNAs*, or *piRNAs*, also induces formation of heterochromatin, blocking expression of some parasitic DNA elements in the genome known as transposons. (Transposons are discussed in Concept 21.4.) Usually 24–31 nucleotides in length, piRNAs are processed from a longer, single-stranded RNA precursor. They play an indispensable role in the germ

cells of many animal species, where they appear to help reestablish appropriate methylation patterns in the genome during gamete formation.

Mastering Biology **Interview with Haifan Lin: Discovering piwi-interacting RNAs and connections to testicular cancer**



Researchers have also found a relatively large number of **long noncoding RNAs (lncRNAs)**, ranging from 200 to hundreds of thousands of nucleotides in length, that are expressed at significant levels in specific cell types at particular times. The functional significance of these lncRNAs has been debated, but in 2017 a large international research consortium published an atlas of almost 28,000 such RNAs; their analysis supported the idea that almost 20,000 were functional and that some were associated with specific diseases. One lncRNA, long known to be functional, is responsible for X chromosome inactivation, which prevents expression of genes located on one of the X chromosomes in most female mammals (see Figure 15.8). In this case, lncRNAs—transcripts of the *XIST* gene located on the chromosome to be inactivated—bind back to and coat that chromosome. This binding leads to condensation of the entire chromosome into heterochromatin.

The examples just described involve chromatin remodeling in large regions of the chromosome. Because chromatin structure affects transcription and thus gene expression, RNA-based regulation of chromatin structure is sure to play an important role in gene regulation. Additionally, some experimental evidence supports the idea of an alternate role for lncRNAs in which they can act as a scaffold, bringing DNA, proteins, and other RNAs together into complexes. These associations may act either to condense chromatin or, in some cases, to help bring the enhancer of a gene together with mediator proteins and the gene's promoter, activating gene expression in a more direct fashion.

Given the extensive functions of noncoding RNAs, it is not surprising that many of the ncRNAs characterized thus far play important roles in embryonic development—the topic we turn to in the next section. Embryonic development is perhaps the ultimate example of precisely regulated gene expression.

CONCEPT CHECK 18.3

1. Compare miRNAs and siRNAs, including their functions.
2. **WHAT IF?** Suppose the mRNA being degraded in Figure 18.14 coded for a protein that promotes cell division in a multicellular organism. What would happen if a mutation disabled the gene for the miRNA that triggers this degradation?
3. **MAKE CONNECTIONS** Inactivation of one of the X chromosomes in female mammals involves a noncoding RNA (*XIST*). Suggest a model for how *XIST* RNA initiates Barr body formation (see Concept 15.2).

For suggested answers, see Appendix A.

CONCEPT 18.4

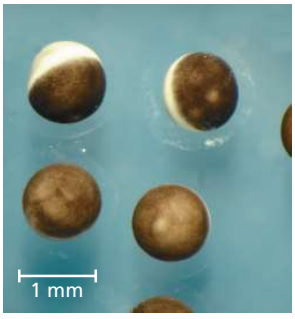
A program of differential gene expression leads to the different cell types in a multicellular organism

In the embryonic development of multicellular organisms, a fertilized egg (a zygote) gives rise to cells of many different types, each with a different structure and corresponding function. Typically, cells are organized into tissues, tissues into organs, organs into organ systems, and organ systems into the whole organism. Thus, any developmental program must produce cells of different types that form higher-level structures arranged in a particular way in three dimensions. The processes that occur during animal development are detailed in Chapter 47; in this chapter, we focus on the program of regulation of gene expression that orchestrates development, using a few animal species as examples.

A Genetic Program for Embryonic Development

The photos in **Figure 18.16** illustrate the dramatic difference between a frog zygote (fertilized egg) and the tadpole it becomes. This remarkable transformation results from three interrelated processes: cell division, cell differentiation, and morphogenesis. Through a succession of mitotic cell divisions, the zygote gives rise to a large number of cells. Cell division alone, however, would merely produce a great ball of identical cells, nothing like a tadpole. During embryonic development, cells not only increase in number, but also undergo cell **differentiation**, the process by which cells become specialized in structure and function. Moreover, the different kinds of cells are not

▼ Figure 18.16 From fertilized egg to animal: What a difference four days makes. It takes just four days for cell division, differentiation, and morphogenesis to transform each of the fertilized frog eggs shown in **(a)** into a tadpole like the one in **(b)**.



(a) Fertilized eggs of a frog



(b) Newly hatched tadpole

randomly distributed but are organized into tissues and organs in a particular three-dimensional arrangement. The physical processes that give an organism its shape constitute **morphogenesis**, the development of the form of an organism and its structures.

All three processes are rooted in cellular behavior. Even morphogenesis, the shaping of the organism, can be traced back to changes in the shape, motility, and other characteristics of the cells that make up various regions of the embryo. As you have seen, the activities of a cell depend on the genes it expresses and the proteins it produces. Almost all cells in an organism have the same genome; therefore, differential gene expression results from the genes being regulated differently in each cell type.

In Figure 18.12, you saw a simplified view of how differential gene expression occurs in two cell types, a liver cell and a lens cell. Each of these fully differentiated cells has a particular mix of specific transcription factor activators that turn on the collection of genes whose products are required in the cell. The fact that both cells arose through a series of mitoses from a common fertilized egg inevitably leads to a question: How do different sets of activators come to be present in the two cells?

It turns out that materials placed into the egg by maternal cells set up a sequential program of gene regulation that is carried out as embryonic cells divide, and this program coordinates cell differentiation during embryonic development. To understand how this works, we will consider two basic developmental processes. First, we'll explore how cells that arise from early embryonic mitoses develop the differences that start each cell along its own differentiation pathway. Second, we'll see how cellular differentiation leads to one particular cell type, using muscle development as an example.

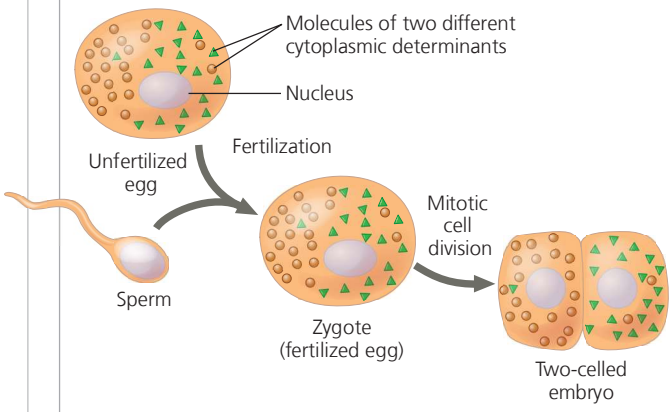
Cytoplasmic Determinants and Inductive Signals

What generates the first differences among cells in an early embryo? And what controls the differentiation of all the various cell types as development proceeds? You can probably deduce the answer: The specific genes expressed in any particular cell of a developing organism determine its path. Two sources of information, used to varying extents in different species, "tell" a cell which genes to express at any given time during embryonic development.

One important source of information early in development is the egg's cytoplasm, which contains both RNA and proteins encoded by the mother's DNA. The cytoplasm of an unfertilized egg is not homogeneous. Messenger RNAs, proteins, other substances, and organelles are distributed unevenly in the unfertilized egg, and this unevenness has a profound impact on the development of the future embryo in many species. Maternal substances in the egg that influence the course of early development are called **cytoplasmic determinants** (Figure 18.17a). After fertilization, early mitotic divisions distribute the zygote's cytoplasm into separate cells. The nuclei of these cells may thus be exposed to different cytoplasmic determinants, depending on which portions of the zygotic cytoplasm a cell received. The combination of cytoplasmic determinants in a cell helps determine its developmental fate by regulating expression of the cell's genes during the course of cell differentiation.

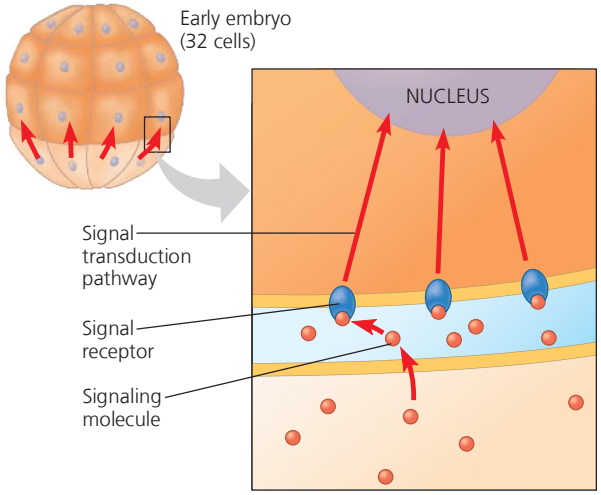
▼ Figure 18.17 Sources of developmental information for the early embryo.

(a) Cytoplasmic determinants in the egg



The unfertilized egg has molecules in its cytoplasm, encoded by the mother's genes, that influence development. Many of these cytoplasmic determinants, like the two shown here, are unevenly distributed in the egg. After fertilization and mitotic division, the cell nuclei of the embryo are exposed to different combinations of cytoplasmic determinants and, as a result, express different genes.

(b) Induction by nearby cells



Cells in the lowest tier of cells are releasing molecules that signal (induce) nearby cells to change their gene expression.

The other major source of developmental information, which becomes increasingly important as the number of embryonic cells increases, is the environment around a particular cell. Most influential are the signals conveyed to an embryonic cell from other embryonic cells in the vicinity, including contact with cell-surface molecules on neighboring cells and the binding of growth factors secreted by neighboring cells (see Concept 11.1). Such signals cause changes in the target cells, a process called **induction** **(Figure 18.17b)**. The molecules that transmit these signals within the target cell are cell-surface receptors and other signaling pathway proteins. In general, the signal sends a cell down a specific developmental path by causing changes in its gene expression that lead to observable cellular changes. Thus, interactions between embryonic cells help induce differentiation into the many specialized cell types making up a new organism.

Sequential Regulation of Gene Expression During Cellular Differentiation

The earliest changes that set a cell on its path to specialization are subtle ones, showing up only at the molecular level. Before biologists knew much about the molecular changes occurring in embryos, they coined the term **determination** to refer to the point at which an embryonic cell is irreversibly committed to becoming a particular cell type. Once it has undergone determination, an embryonic cell can be experimentally placed in another location in the embryo and it will still differentiate into the cell type that is its normal fate. Differentiation, then, is the process by which a cell attains its determined fate. As the tissues and organs of an embryo develop and their cells differentiate, the cells become more noticeably different in structure and function.

Today we understand determination in terms of molecular changes that result in observable cell differentiation, marked by the expression of genes for *tissue-specific proteins*. Such proteins are found only in a specific cell type and give the cell its characteristic structure and function. The first sign of differentiation is the appearance of mRNAs for tissue-specific proteins. Later, differentiation is observable with a microscope as changes in cellular structure. On the molecular level, different sets of genes are sequentially expressed in a regulated manner as new cells arise when their precursors divide. Multiple steps in gene expression may be regulated during differentiation, transcription being the most common. In the fully differentiated cell, transcription remains the principal regulatory point for maintaining appropriate gene expression.

Differentiated cells are specialists at making tissue-specific proteins. For example, as a result of transcriptional regulation, liver cells specialize in making albumin, and lens cells specialize in making crystallin (see Figure 18.12). Skeletal muscle cells in vertebrates are another instructive example. Each of these cells is a long fiber containing many nuclei within a single plasma membrane. Skeletal muscle cells have high concentrations of muscle-specific versions of the contractile proteins myosin and actin, as well as membrane receptor proteins that detect signals from nerve cells.

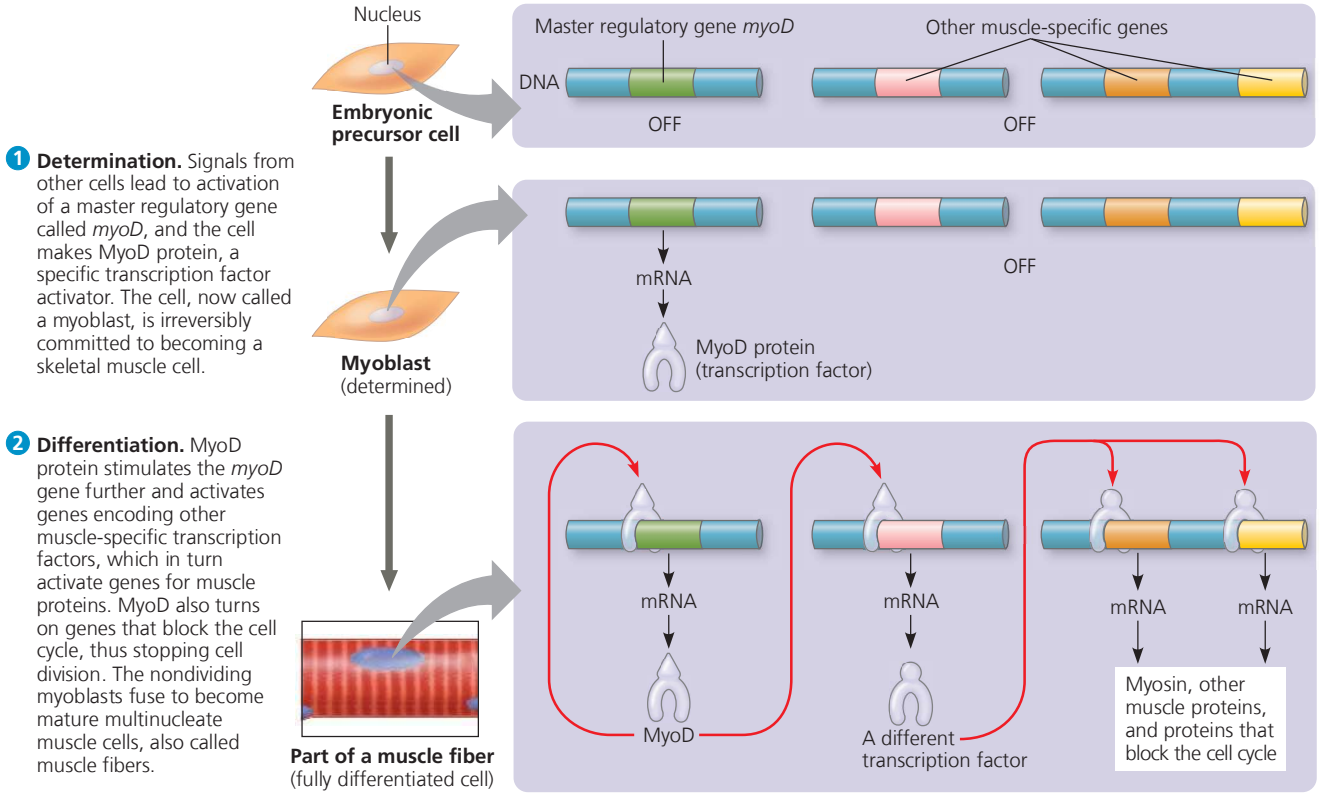
Muscle cells develop from embryonic precursor cells that have the potential to develop into a number of cell types, including cartilage cells and fat cells, but particular conditions commit them to becoming muscle cells. Although the committed cells appear unchanged under the microscope, determination has occurred, and they are now a cell type called *myoblasts*. Eventually, myoblasts start to churn out large amounts of muscle-specific proteins and fuse to form mature, elongated, multinucleate skeletal muscle cells.

Researchers have worked out what happens at the molecular level during muscle cell determination. To do so, they grew embryonic precursor cells in culture and analyzed them using molecular techniques you will learn about in Concepts 20.1 and 20.2. In a series of experiments, they isolated different genes, caused each to be expressed in a separate embryonic precursor cell, and then looked for differentiation into myoblasts and muscle cells. In this way, they identified several so-called "master regulatory genes" whose protein products commit the cells to becoming skeletal muscle cells. Thus, in the case of muscle cells, the molecular basis of determination is the expression of one or more of these master regulatory genes.

To understand more about how determination occurs in muscle cell differentiation, let's focus on the master regulatory gene called *myoD*. The *myoD* gene deserves its designation as a master regulatory gene. Researchers have shown that the MyoD protein it encodes is capable of changing some kinds of fully differentiated nonmuscle cells, such as fat cells and liver cells, into muscle cells. Why doesn't MyoD work on *all* kinds of cells? One likely explanation is that activation of muscle-specific genes is not solely dependent on MyoD but requires a particular *combination* of regulatory proteins, some of which are lacking in cells that do not respond to MyoD.

What is the molecular basis for muscle cell differentiation? The MyoD protein is a transcription factor (see Figure 18.10) that binds to specific control elements in the enhancers of various target genes and stimulates their expression **(Figure 18.18)**. Some target genes for MyoD encode still other muscle-specific transcription factors. MyoD also stimulates expression of the *myoD* gene itself, an example of positive feedback that perpetuates MyoD's effect in maintaining the cell's differentiated state. Presumably, all the genes activated by MyoD have enhancer control elements recognized by MyoD and are thus coordinately controlled. Finally, the secondary transcription factors activate

▼ Figure 18.18 Determination and differentiation of muscle cells. Skeletal muscle cells arise from embryonic cells as a result of changes in gene expression. (In this depiction, the process of gene activation is greatly simplified.)



WHAT IF? *What would happen if a mutation in the* myoD *gene resulted in the production of an altered MyoD protein that could not activate the* myoD *gene?*

the genes for proteins such as myosin and actin that confer the unique properties of skeletal muscle cells.

The determination and differentiation of other kinds of tissues may play out in a similar fashion. Experimental results support the idea that master regulatory proteins like MyoD might function by opening the chromatin in certain regions. This allows access to transcription machinery for activation of the next set of cell-type-specific genes.

Pattern Formation: Setting Up the Body Plan

We have now seen how different programs of gene expression that are activated in the fertilized egg can result in differentiated cells and tissues. But for the tissues to function effectively in the organism as a whole, the organism's *body plan*—its overall three-dimensional arrangement—must be established and superimposed on the differentiation process. Let's look at the molecular basis for the establishment of the body plan, using the well-studied fruit fly *Drosophila melanogaster* as an example.

Cytoplasmic determinants and inductive signals both contribute to spatially organizing the tissues and organs of an organism in their characteristic places. This developmental process is referred to as **pattern formation**.

Just as the locations of the front, back, and sides of a new building are determined before construction begins, pattern formation in animals begins in the early embryo, when the major axes of an animal are established. In a bilaterally symmetrical animal, the relative positions of head and tail, right and left sides, and back and front—the three major body axes—are set up before the organs appear. The molecular cues that control pattern formation, collectively called **positional information**, are provided by cytoplasmic determinants and inductive signals (see Figure 18.17). These cues tell a cell its location relative to the body axes and to neighboring cells, and determine how the cell and its descendants will respond to future molecular signals.

During the early 20th century, classical embryologists made detailed anatomical observations of embryonic development in a number of species and performed experiments in which they manipulated embryonic tissues. Although this research laid the groundwork for understanding the mechanisms of development, it did not reveal the specific

molecules that guide development or determine how patterns are established.

In the 1940s, scientists began using the genetic approach—the study of mutants—to investigate *Drosophila* development. That approach has had spectacular success. These studies have established that genes control development and have led to an understanding of the key roles that specific molecules play in defining position and directing differentiation. By combining anatomical, genetic, and biochemical approaches to the study of *Drosophila* development, researchers have discovered developmental principles common to many other species, including humans.

The Life Cycle of Drosophila

Fruit flies and other arthropods have a modular construction, an ordered series of segments. These segments make up the body's three major parts: the head, the thorax (the mid-body, from which the wings and legs extend), and the abdomen **(Figure 18.19a)**. Like other bilaterally symmetrical animals, *Drosophila* has an anterior-posterior (head-to-tail) axis, a dorsal-ventral (back-to-belly) axis, and a right-left axis. In *Drosophila*, cytoplasmic determinants that are localized in the unfertilized egg provide positional information for the placement of anterior-posterior and dorsal-ventral axes even before fertilization. We'll focus here on the molecules involved in establishing the anterior-posterior axis as a case in point.

The *Drosophila* egg develops in one of the female's ovaries, next to the nurse cells, which supply the egg with nutrients, mRNAs, and other substances needed for development. The egg and nurse cells are surrounded by follicle cells, which make the eggshell (**Figure 18.19b**, top). After fertilization and laying of the egg, embryonic development results in the formation of a segmented larva, which goes through three larval stages. Then, in a process much like that by which a caterpillar becomes a butterfly, the fly larva forms a pupa in which it metamorphoses into the adult fly pictured in Figure 18.19a.

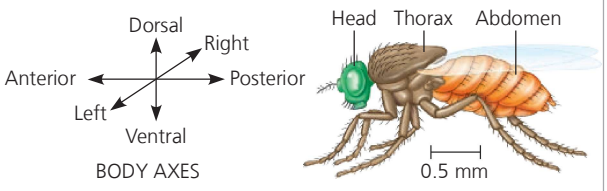
Genetic Analysis of Early Development: Scientific Inquiry

Edward B. Lewis was a visionary American biologist who, in the 1940s, first showed the value of the genetic approach to studying embryonic development in *Drosophila*. Lewis studied bizarre mutant flies with developmental defects that led to extra wings or legs in the wrong place **(Figure 18.20)**. He located the mutations on the fly's genetic map, thus connecting the developmental abnormalities to specific genes. This research supplied the first concrete evidence that genes somehow direct the developmental processes studied by embryologists. The genes Lewis discovered, called **homeotic genes**, are regulatory genes that control pattern formation in the fly.

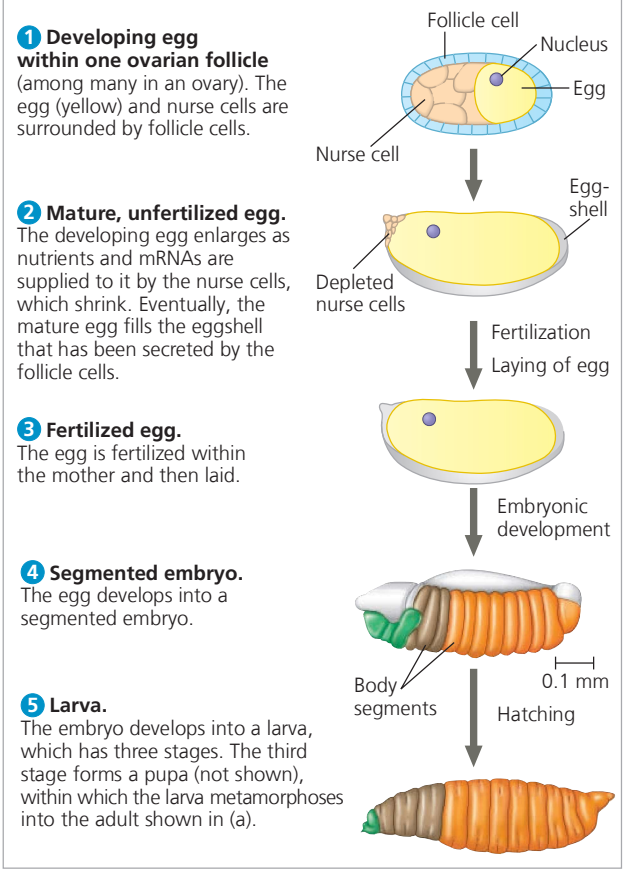
Further insight into pattern formation during early embryonic development did not come for another 30 years,

▼ Figure 18.19 Key events in *Drosophila* development.

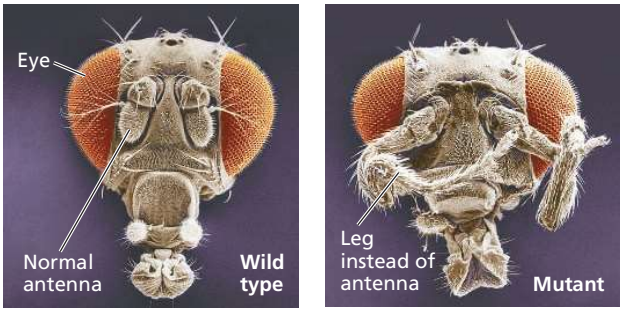
(a) Adult. The adult fly is segmented, and multiple segments make up each of the three main body parts—head, thorax, and abdomen. The body axes are shown by arrows.



(b) Development from egg to larva.



▼ Figure 18.20 Abnormal pattern formation in *Drosophila*. Mutations in homeotic genes cause misplacement of structures in an animal, such as the legs extending from the mutant fly's head in place of antenna (colorized SEMs).



when two researchers in Germany, Christiane Nüsslein-Volhard and Eric Wieschaus, set out to identify *all* the genes that affect segment formation in *Drosophila*. The project was daunting for three reasons. The first was the sheer number of *Drosophila* genes, now known to total about 14,000. The genes affecting segmentation might be just a few needles in a haystack or might be so numerous and varied that the scientists would be unable to make sense of them. Second, mutations affecting a process as fundamental as segmentation would surely be **embryonic lethals**, mutations with phenotypes causing death at the embryonic or larval stage. Because organisms with embryonic lethal mutations never reproduce, they cannot be bred for study. The researchers dealt with this problem by looking for recessive mutations, which can be propagated in heterozygous flies that act as genetic carriers. Third, cytoplasmic determinants in the egg were known to play a role in axis formation, so the researchers knew they would have to study the mother's genes as well as those of the embryo. It is the mother's genes that we will discuss further as we focus on how the anterior-posterior body axis is set up in the developing egg.

Nüsslein-Volhard and Wieschaus began their search for segmentation genes by exposing flies to mutagenic agents and scanning their descendants for dead embryos or larvae with abnormal segmentation or other defects. For example, to find genes that might set up the anterior-posterior axis, they looked for embryos or larvae with abnormal ends, such as two heads or two tails, predicting that such abnormalities would arise from mutations in maternal genes required for correctly setting up the offspring's head or tail end.

Using this approach, Nüsslein-Volhard and Wieschaus eventually identified about 1,200 genes essential for pattern formation during embryonic development. Of these, about 120 were essential for normal segmentation patterns. Next, the researchers were able to group these segmentation genes by general function and to isolate many of them for further study. The result was a detailed molecular understanding of the early steps in pattern formation in *Drosophila*.

When the results of Nüsslein-Volhard and Wieschaus were combined with Lewis's earlier work, a coherent picture of *Drosophila* development emerged. In recognition of their discoveries, the three researchers were awarded a Nobel Prize in 1995. Let's consider a specific example of the genes that Nüsslein-Volhard, Wieschaus, and co-workers found.

Axis Establishment

As we mentioned earlier, cytoplasmic determinants in the egg are the substances that initially establish the axes of the *Drosophila* body. These substances are encoded by genes of the mother, fittingly called maternal effect genes. A gene classified as a **maternal effect gene** is one that, when mutant in the mother, results in a mutant phenotype in the offspring, regardless of the offspring's own genotype. In fruit fly development, the mRNA or protein products of maternal effect genes are placed in the egg while it is still in the mother's ovary. When the mother has a mutation in such a gene, she makes a defective gene product (or none at all), and her eggs are abnormal; when these eggs are fertilized, they fail to develop properly.

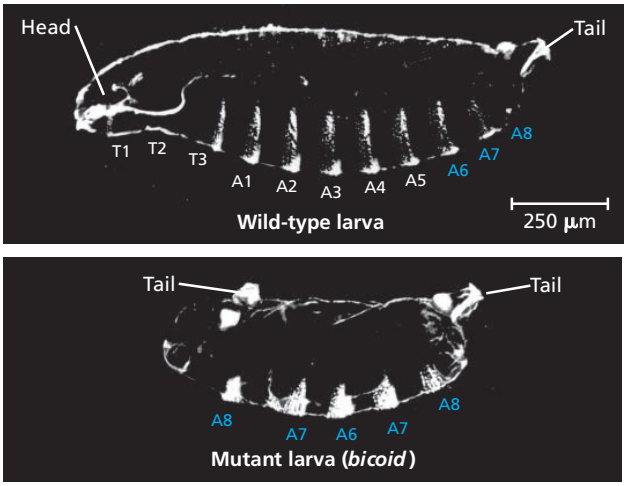
Because maternal effect genes control the orientation (polarity) of the egg and consequently that of the fly, they are also called *egg-polarity genes*. Two groups of these genes set up the anterior-posterior and dorsal-ventral axes of the embryo. Like mutations in segmentation genes, mutations in maternal effect genes are generally embryonic lethals.

Bicoid: A Morphogen That Determines Head Structures

To see how maternal effect genes determine the body axes of the offspring, we will focus on one such gene, called *bicoid*, a term meaning "two-tailed." An embryo or larva whose mother has two mutant *bicoid* alleles lacks the front half of its body and has posterior structures at both ends **(Figure 18.21)**. This phenotype suggested to Nüsslein-Volhard and her colleagues that the product of the mother's *bicoid* gene is essential for setting up the anterior end of the fly and might be concentrated at the future anterior end of the embryo. This hypothesis is an example of the *morphogen gradient hypothesis* first proposed by embryologists a century ago, in which gradients of substances called **morphogens** establish an embryo's axes and other features of its form.

DNA technology and other modern biochemical methods enabled the researchers to test whether the *bicoid* product, a protein called Bicoid, is in fact a morphogen that determines the anterior end of the fly. First, they asked whether the location of the mRNA and protein products of this gene in the egg was consistent with the hypothesis. They found that

▼ Figure 18.21 Effect of the *bicoid* gene on *Drosophila* development. A wild-type fruit fly larva has a head, three thoracic (T) segments, eight abdominal (A) segments, and a tail. A larva whose mother has two mutant alleles of the *bicoid* gene has two tails and lacks all anterior structures (LMs).



bicoid mRNA is highly concentrated at the extreme anterior end of the mature egg **(Figure 18.22)**. After the egg is fertilized, the mRNA is translated into protein. The Bicoid protein then diffuses from the anterior end toward the posterior, resulting in a gradient of protein within the early embryo, most highly concentrated at the anterior end. These results are consistent with the hypothesis that Bicoid protein specifies the fly's anterior end. To test this more specifically, scientists injected pure *bicoid* mRNA into various regions of early embryos. The protein that resulted from its translation caused anterior structures to form at the injection sites.

The *bicoid* research was groundbreaking for several reasons. First, it led to the identification of a specific protein required for some of the earliest steps in pattern formation. It thus helped us understand how different regions of the egg can give rise to cells that go down different developmental pathways. Second, it increased our understanding of the mother's critical role in the initial phases of embryonic development. Third, the principle that a gradient of morphogens can determine polarity and position has proved to be a key developmental concept for a number of species, just as early embryologists had hypothesized.

Maternal mRNAs are crucial during development of many species. In *Drosophila*, gradients of specific proteins encoded by maternal mRNAs not only determine the posterior and anterior ends but also establish the dorsal-ventral axis. As the fly embryo grows, it reaches a point when the embryonic program of gene expression takes over, and the maternal mRNAs must be destroyed. Later, positional information encoded by the embryo's genes, operating on an ever finer scale, establishes a specific number of correctly oriented segments and triggers the formation of each segment's characteristic structures. Mutant genes affecting this final step can lead to an abnormal adult pattern, as you saw in Figure 18.20.

Mastering Biology **Interview with Nancy Hopkins: Studying the genetic basis of development in zebrafish**



Evolutionary Developmental Biology ("Evo-Devo")

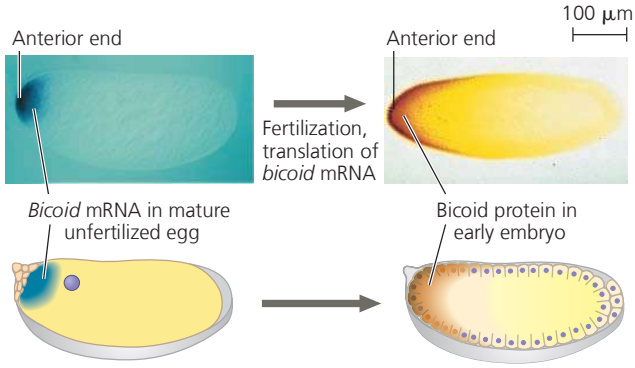
EVOLUTION The fly with legs emerging from its head in Figure 18.20 is the result of a single mutation in one gene, a homeotic gene. The gene does not encode any antenna protein, however. Instead, it encodes a transcription factor that regulates other genes, and its malfunction leads to misplaced structures, such as legs instead of antennae. The observation that a change in gene regulation during development could lead to such a fantastic change in body form prompted some scientists to consider whether these types of mutations could contribute to evolution by generating novel body shapes. Ultimately, this line of inquiry gave rise to the field of evolutionary developmental biology, so-called "evo-devo," which will be further discussed in Concept 21.6.

▼ Figure 18.22 Inquiry

Could Bicoid be a morphogen that determines the anterior end of a fruit fly?

Experiment Using a genetic approach to study *Drosophila* development, Christiane Nüsslein-Volhard and colleagues at two research institutions in Germany analyzed expression of the *bicoid* gene. The researchers hypothesized that *bicoid* normally codes for a morphogen that specifies the head (anterior) end of the embryo. To begin to test this hypothesis, they used molecular techniques to determine whether the mRNA and protein encoded by this gene were found in the anterior end of the fertilized egg and early embryo of wild-type flies.

Results *Bicoid* mRNA (dark blue in the light micrographs and drawings) was confined to the anterior end of the unfertilized egg. Later in development, Bicoid protein (dark orange) was seen to be concentrated in cells at the anterior end of the embryo.



Conclusion The location of *bicoid* mRNA and the diffuse gradient of Bicoid protein seen later are consistent with the hypothesis that Bicoid protein is a morphogen specifying formation of head-specific structures.

Further Reading C. Nüsslein-Volhard et al., Determination of anteroposterior polarity in *Drosophila*, *Science* 238:1675–1681 (1987); W. Driever and C. Nüsslein-Volhard, A gradient of Bicoid protein in *Drosophila* embryos, *Cell* 54:83–93 (1988); T. Berleth et al., The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo, *EMBO Journal* 7:1749–1756 (1988).

WHAT IF? *The researchers needed further evidence, so they injected bicoid mRNA into the anterior end of an egg from a female with a mutation disabling the bicoid gene. Given that the hypothesis was supported, what must their results have been?*

Mastering Biology **Animation: Role of *bicoid* Gene in *Drosophila* Development**

In this section, we have seen how a carefully orchestrated program of sequential gene regulation controls the transformation of a fertilized egg into a multicellular organism. The program is carefully balanced between turning on the genes for differentiation in the right place and turning off other genes. Even when an organism is fully developed, gene expression is regulated in a similarly fine-tuned manner. In the final section of the chapter, we'll consider how fine this tuning is by looking at how specific changes in expression of just a few genes can lead to the development of cancer.

CONCEPT CHECK 18.4

1. **MAKE CONNECTIONS** As you learned in Chapter 12, mitosis gives rise to two daughter cells that are genetically identical to the parent cell. Yet you, the product of many mitotic divisions, are not composed of identical, zygote-like cells. Why?
2. **MAKE CONNECTIONS** Explain how the signaling molecules released by an embryonic cell can induce changes in a neighboring cell without entering the cell. (See Figures 11.15 and 11.16.)
3. How do fruit fly maternal effect genes determine the polarity of the egg and the embryo?
4. **WHAT IF?** In Figure 18.17b, the lower cell is synthesizing signaling molecules, whereas the upper cell is expressing receptors for these molecules. In terms of gene regulation and cytoplasmic determinants, explain how these cells came to synthesize different molecules.

For suggested answers, see Appendix A.

CONCEPT 18.5

Cancer results from genetic changes that affect cell cycle control

In Concept 12.3, we considered cancer as a disease in which cells escape from their normal cell division control mechanisms. Now that we have discussed the molecular basis of gene expression and its regulation, we can look at cancer more closely. The gene regulation systems that go wrong during cancer turn out to be the very same systems that play important roles in embryonic development, the immune response, and many other biological processes. Thus, research into the molecular basis of cancer has both benefited from and informed many other fields of biology.

Types of Genes Associated with Cancer

The genes that normally regulate cell growth and division during the cell cycle include genes for growth factors, their receptors, and the intracellular molecules of signaling pathways. (To review cell signaling, see Concept 11.2; for regulation of the cell cycle, see Concept 12.3.) Mutations that alter any of these genes in somatic cells can lead to cancer. The agent of such change can be random spontaneous mutation. Many cancer-causing mutations likely also result from environmental influences: chemical carcinogens like tobacco, X-rays and other high-energy radiation, and some viruses.

Cancer research uncovered cancer-causing genes called **oncogenes** (from the Greek *onco*, tumor) in certain types of viruses. Later, related versions of viral oncogenes were found in the genomes of humans and other animals. The normal versions of the cellular genes, called **proto-oncogenes**, code for proteins that stimulate normal cell growth and division.

How might a proto-oncogene—a gene that has an essential function in normal cells—become an oncogene, a cancer-causing gene? In general, an oncogene arises from a genetic change that leads to an increase either in the amount of the proto-oncogene's protein product or in the intrinsic activity of each protein molecule. The genetic changes that convert proto-oncogenes to oncogenes fall into four main categories **(Figure 18.23)**: epigenetic changes, translocations, gene amplification, and point mutations.

First, alterations in epigenetic modifications that can lead to abnormal chromatin condensation in a cell are often found in tumor cells. If a mutation in a gene for a chromatin-modifying enzyme leads to loosening of chromatin in a region that is normally not being expressed, a proto-oncogene in that region could be expressed at abnormally high levels (see Figure 18.23a). For example, a gene for one such enzyme has been shown to be mutated in 20% of tumor cells analyzed.

Second, cancer cells are frequently found to contain chromosomes that have broken and rejoined incorrectly, translocating fragments from one chromosome to another (see Figure 15.14). If a translocated proto-oncogene ends up near an especially active promoter (or other control element), its transcription may increase, making it an oncogene (see Figure 18.23b).

The third main type of genetic change, amplification, increases the number of copies of the proto-oncogene in the cell through repeated gene duplication (discussed in Concept 21.5; see Figure 18.23c).

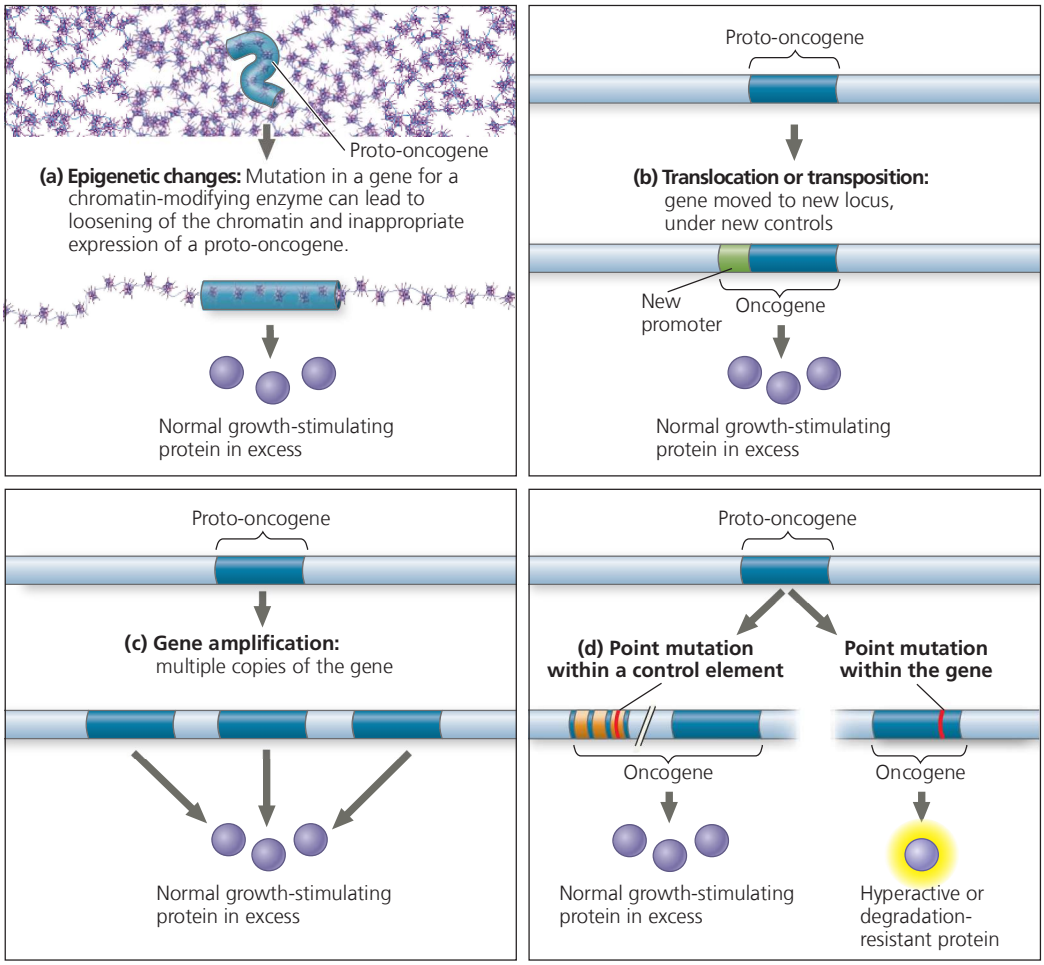
Fourth, a point mutation either in the promoter or an enhancer that controls a proto-oncogene, could cause an increase in its expression. A point mutation in the coding sequence of the proto-oncogene could change the gene's product to a protein that is more active or more resistant to degradation than the normal protein (see Figure 18.23d).

Any of these four mechanisms can lead to abnormal stimulation of the cell cycle and put the cell on the path to becoming a cancer cell.

In addition to genes whose products normally promote cell division, cells contain genes whose normal products *inhibit* cell division. Such genes are called **tumor-suppressor genes** since the proteins they encode help prevent uncontrolled cell growth. Any mutation that decreases the normal activity of a tumor-suppressor protein may contribute to the onset of cancer, in effect stimulating growth through the absence of suppression.

The protein products of tumor-suppressor genes have various functions. Some repair damaged DNA, which prevents the cell from accumulating cancer-causing mutations. Other tumor-suppressor proteins control adhesion of cells to each other or to the extracellular matrix; proper cell anchorage is crucial in normal tissues and is often absent in cancers. Still other tumor-suppressor proteins are components of cell-signaling pathways that inhibit the cell cycle.

▼ **Figure 18.23** **Genetic changes that can turn proto-oncogenes into oncogenes.**



Interference with Normal Cell-Signaling Pathways

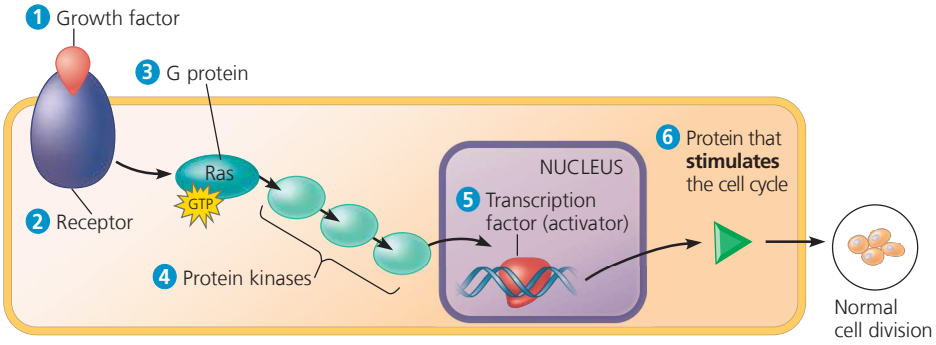
Let's consider how protein components of cell-signaling pathways function in normal cells and what goes wrong with their function in cancer cells. We will focus on the products of two key genes, the *ras* proto-oncogene and the *p53* tumor-suppressor gene. Mutations in *ras* occur in about 30% of human cancers and mutations in *p53* in more than 50%.

The Ras protein, encoded by the ***ras* gene** (named for rat sarcoma, a connective tissue cancer), is a G protein that relays a signal from a growth factor receptor on the plasma membrane to a cascade of protein kinases (see Figures 11.8 and 11.10). The cellular response at the end of the pathway is the synthesis of a protein that stimulates the cell cycle **(Figure 18.24a)**. Normally, such a pathway will not operate unless triggered by the appropriate growth factor. But certain mutations in the *ras* gene can lead to production of a hyperactive Ras protein that triggers the kinase cascade even in the absence of growth factor, resulting in increased cell division **(Figure 18.24b)**. In fact, hyperactive versions or excess amounts of any of the pathway's components can have the same outcome: excessive cell division.

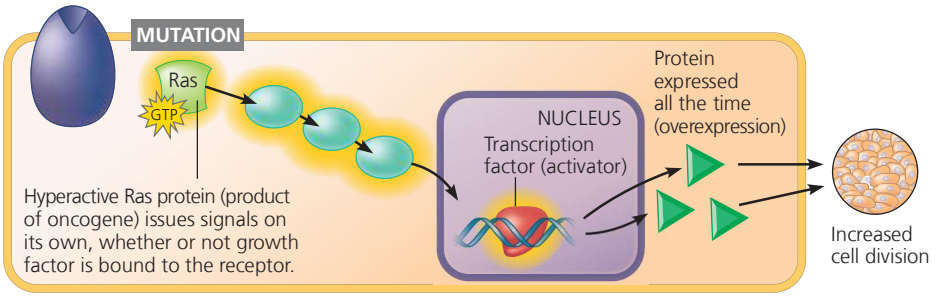
Figure 18.25a shows a pathway in which an intracellular signal leads to the synthesis of a protein that suppresses the cell cycle. In this case, the signal is damage to the cell's DNA, perhaps as the result of exposure to ultraviolet light. Operation of this signaling pathway blocks the cell cycle until the damage has been repaired. Otherwise, the damage might contribute to tumor formation by causing mutations or chromosomal abnormalities. Thus, the genes for the components of the pathway act as tumor-suppressor genes. The ***p53* gene**, named for the apparent molecular weight of its protein product, is a tumor-suppressor gene. The protein it encodes is a specific transcription factor that promotes the synthesis of cell cycle–inhibiting proteins. That is why a mutation that knocks out the *p53* gene or in a gene required to activate the p53 protein (for example, a gene

▼ Figure 18.24 Normal and mutant cell cycle–stimulating pathway.

(a) Normal cell cycle–stimulating pathway. The normal pathway is triggered by 1 a growth factor that binds to 2 its receptor in the plasma membrane. The signal is relayed to 3 a G protein called Ras. Like all G proteins, Ras is active when GTP is bound to it. Ras passes the signal to 4 a series of protein kinases. The last kinase activates 5 a transcription factor (activator) that turns on one or more genes for 6 a protein that stimulates the cell cycle.

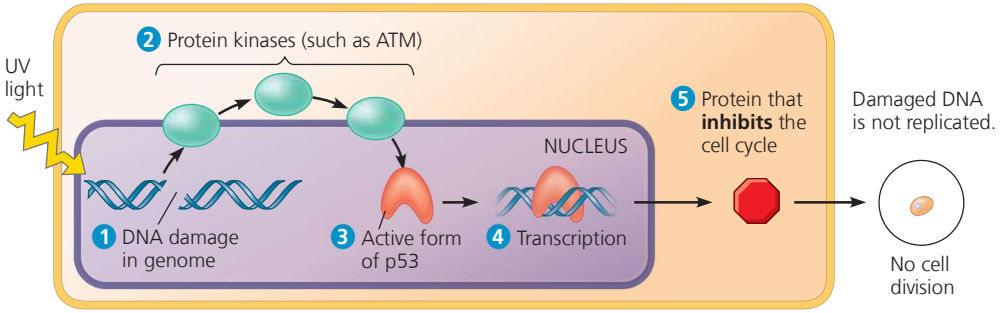


(b) Mutant cell cycle–stimulating pathway. If a mutation makes Ras or any other pathway component abnormally active, excessive cell division and cancer may result.



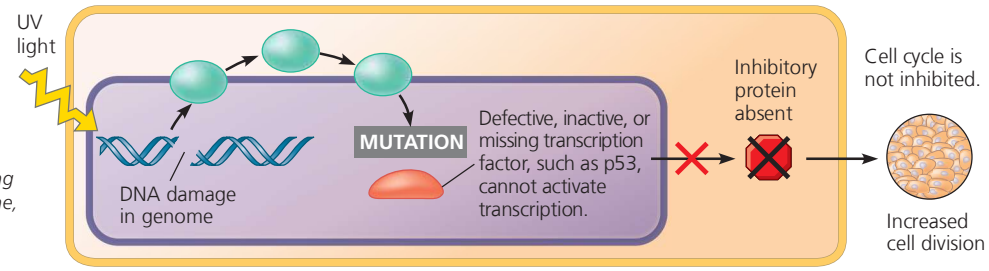
▼ Figure 18.25 Normal and mutant cell cycle–inhibiting pathway.

(a) Normal cell cycle–inhibiting pathway. In the normal pathway, 1 DNA damage is an intracellular signal that is passed via 2 protein kinases, leading to activation of 3 p53. 4 Activated p53 promotes transcription of the gene for 5 a protein that inhibits the cell cycle. The resulting suppression of cell division ensures that the damaged DNA is not replicated. If the DNA damage is irreparable, then the p53 signal leads to apoptosis.



(b) Mutant cell cycle–inhibiting pathway. Mutations causing deficiencies in any pathway component can contribute to the development of cancer.

? *Explain whether a cancer-causing mutation in a tumor-suppressor gene, such as* p53, *is more likely to be a recessive or a dominant mutation.*



called *ATM*) can lead to excessive cell growth and cancer **(Figure 18.25b)**.

The *p53* gene has been called the "guardian angel of the genome." Once the p53 protein is activated—say, by the ATM protein, a protein kinase, after DNA damage—p53 activates several other genes, such as *p21*. The p21 protein halts the cell cycle by binding to cyclin-dependent kinases, allowing time for the cell to repair the DNA. Researchers recently showed that p53 also activates expression of a group of miRNAs that inhibit the cell cycle. The p53 protein can also turn on genes directly involved in DNA repair. If DNA damage is irreparable, p53 activates "suicide" genes, whose protein products bring about programmed cell death (apoptosis; see Figure 11.20). Thus, p53 acts in several ways to prevent a cell from passing

on mutations due to DNA damage. If mutations do accumulate and the cell survives through many divisions—as is more likely if the *p53* tumor-suppressor gene is defective or missing—cancer may ensue. The many functions of p53 suggest a complex picture of regulation in normal cells, one that we do not yet fully understand.

A recent study of elephants may underscore the protective role of *p53*. The incidence of cancer among elephants in zoo-based studies has been estimated at about 3%, compared with closer to 30% for humans. Genome sequencing revealed that elephants have 20 copies of the *p53* gene, compared to one copy in humans, other mammals, and even manatees, elephants' closest living relatives. There are undoubtedly other underlying reasons, but the correlation between low cancer rate and extra copies of the *p53* gene bears further investigation.

Mastering Biology HHMI Animation: Using *p53* to Fight Cancer

For the present, the diagrams in Figure 18.24 and Figure 18.25 are an accurate view of how mutations can contribute to cancer, but we still don't know exactly how a particular cell becomes a cancer cell. Recent studies have shown, for instance, that DNA methylation and histone modification patterns in normal cells differ from those in cancer cells and that miRNAs probably participate in cancer development. There is still a lot to learn, and you and your classmates may be the ones to make important discoveries about cancer biology.

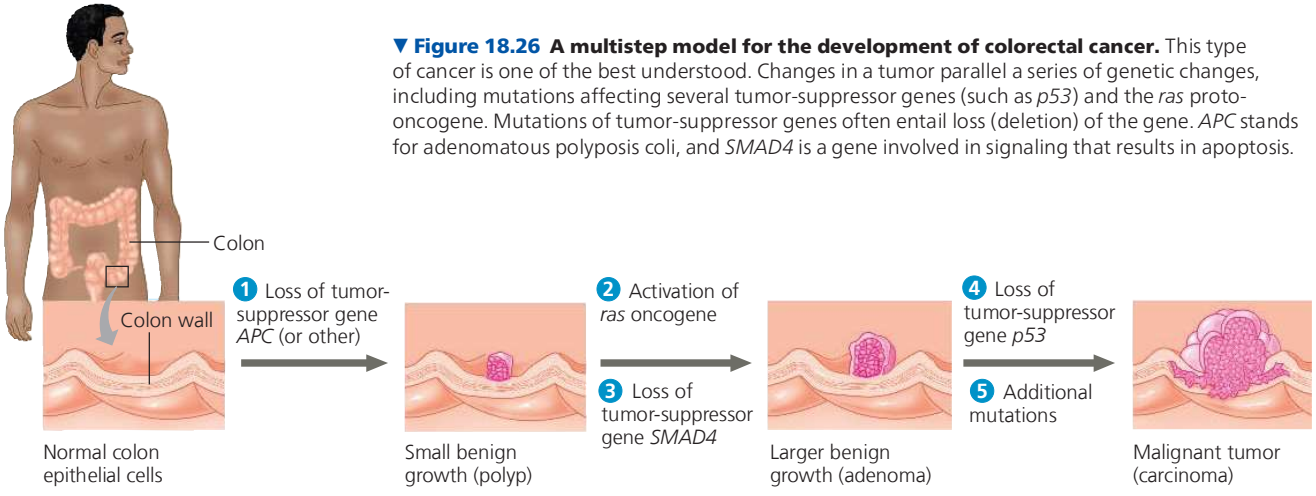
The Multistep Model of Cancer Development

More than one somatic mutation or epigenetic change is generally needed to produce all the changes characteristic of a full-fledged cancer cell. This may help explain why the incidence of cancer increases greatly with age. If cancer results from an accumulation of mutations that occur throughout life, then the longer we live, the more likely we are to develop cancer.

The model of a multistep path to cancer is well supported by studies of one of the best-understood types of human cancer: colorectal cancer, which affects the colon and/or rectum. About 140,000 new cases of colorectal cancer are diagnosed each year in the United States, and the disease causes 50,000 deaths per year. Like most cancers, colorectal cancer develops gradually **(Figure 18.26)**. The first sign is often a polyp, a small, benign growth in the colon lining. The cells of the polyp look normal, although they divide unusually frequently. The tumor grows and may eventually become malignant, invading other tissues. The development of a malignant tumor is paralleled by a gradual accumulation of mutations that convert proto-oncogenes to oncogenes and knock out tumor-suppressor genes. A *ras* oncogene and a mutated *p53* tumor-suppressor gene are often involved.

About half a dozen changes must occur at the DNA level for a cell to become fully cancerous. These changes usually include the appearance of at least one active oncogene and the mutation or loss of several tumor-suppressor genes. Furthermore, since mutant tumor-suppressor alleles are usually recessive, in most cases mutations must knock out *both* alleles in a cell's genome to block tumor suppression. (Most oncogenes, on the other hand, behave as dominant alleles.)

Since we understand the progression of this type of cancer, routine screenings (colonoscopies, for example) are recommended to identify and remove any suspicious polyps. The colorectal cancer mortality rate has been declining for the past 20 years due to increased screening and improved treatments. Treatments for other cancers have improved as well. Advances in the sequencing of DNA and mRNA allow medical researchers to compare the genes expressed by different types of tumors and by the same type in different people. These comparisons have led to personalized treatments based on the molecular characteristics of a person's tumor; **Figure 18.27** shows how this approach has been applied to breast cancer.



▼ Figure 18.26 A multistep model for the development of colorectal cancer. This type of cancer is one of the best understood. Changes in a tumor parallel a series of genetic changes, including mutations affecting several tumor-suppressor genes (such as *p53*) and the *ras* proto-oncogene. Mutations of tumor-suppressor genes often entail loss (deletion) of the gene. *APC* stands for adenomatous polyposis coli, and *SMAD4* is a gene involved in signaling that results in apoptosis.

▼ Figure 18.27

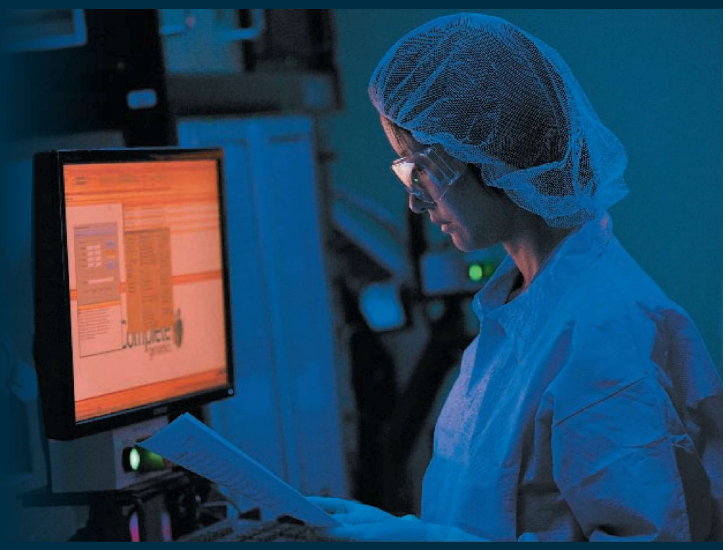
MAKE CONNECTIONS

Genomics, Cell Signaling, and Cancer

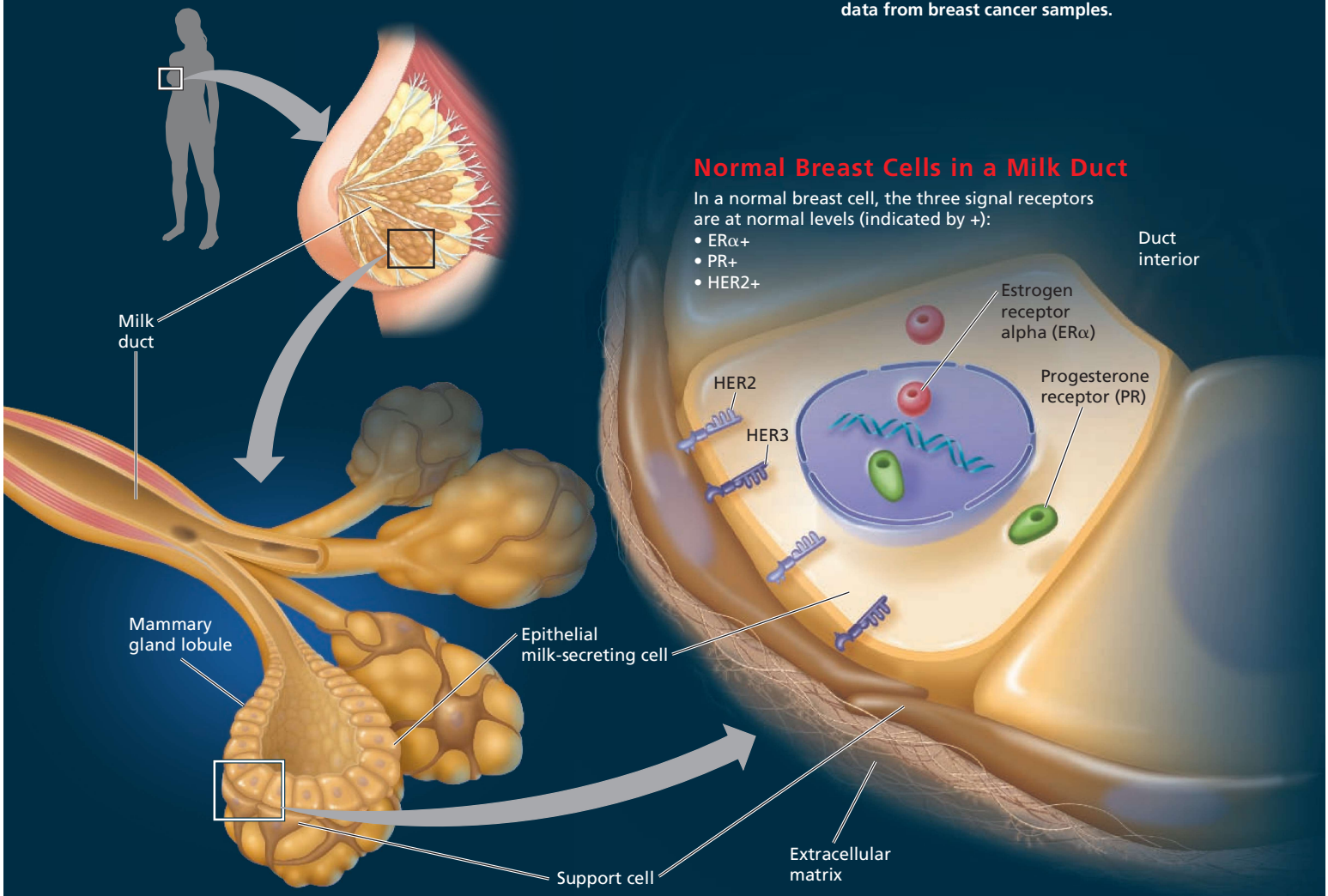
Modern medicine that melds genome-wide molecular studies with cell-signaling research is transforming the treatment of many diseases, such as breast cancer. Using microarray analysis (see Figure 20.13) and other techniques, researchers measured the relative levels of mRNA transcripts for every gene in hundreds of breast cancer tumor samples. They identified four major subtypes of breast cancer that differ in their expression of three signal receptors involved in regulating cell growth and division:

- Estrogen receptor alpha (ERα)
- Progesterone receptor (PR)
- HER2, a type of receptor called a receptor tyrosine kinase (RTK); other RTKs are also present, such as HER3

(ERα and PR are steroid receptors; see Figure 11.9.) The absence or excess expression of these receptors can cause aberrant cell signaling, leading in some cases to inappropriate cell division, which can contribute to cancer (see Figure 18.24).



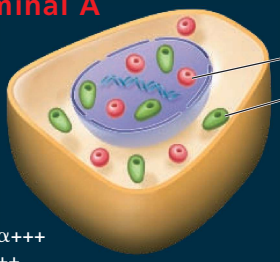
▲ A research scientist examines DNA-sequencing data from breast cancer samples.



Breast Cancer Subtypes

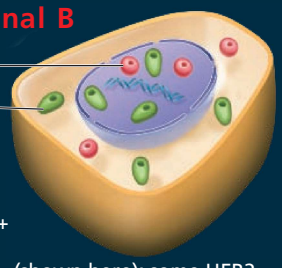
Each breast cancer subtype is characterized by the overexpression (indicated by ++ or +++) or absence (–) of three signal receptors: ERα, PR, and HER2. Breast cancer treatments are becoming more effective because they can be tailored to the specific cancer subtype.

Luminal A



- ERα+++
- PR++
- HER2–
- 40% of breast cancers
- Best prognosis

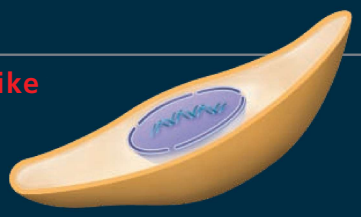
Luminal B



- ERα++
- PR++
- HER2– (shown here); some HER2++
- Divide rapidly
- 15–20% of breast cancers
- Poorer prognosis than luminal A

Both luminal subtypes overexpress ERα (luminal A more than luminal B) and PR, and usually lack expression of HER2. Both can be treated with drugs that target ERα and inactivate it, the most well-known drug being tamoxifen. These subtypes can also be treated with drugs that inhibit estrogen synthesis.

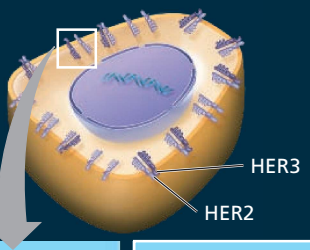
Basal-like



- ERα–
- PR–
- HER2–
- 15–20% of breast cancers
- More aggressive; poorer prognosis than other subtypes

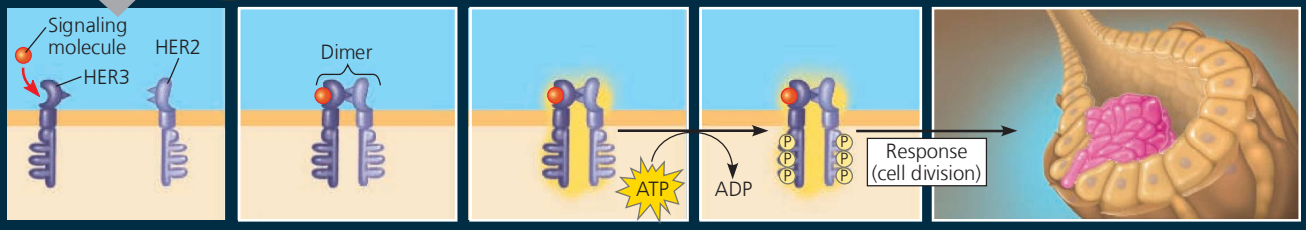
The basal-like subtype is "triple negative"—it usually does not express ERα, PR, or HER2. It often has a mutation in the tumor-suppressor gene *BRCA1* (see Concept 18.5). Treatments that target ERα, PR, or HER2 are not effective, but new treatments are being developed. Currently, patients are treated with cytotoxic chemotherapy, which selectively kills fast-growing cells.

HER2



- ERα– or ERα+
- PR–
- HER2++
- 10–15% of breast cancers
- Poorer prognosis than luminal A subtype

The HER2 subtype overexpresses HER2. Because it does not express PR (and, in many cases, ERα) at normal levels, the cells are unresponsive to therapies that target PR (and, for ERα– cells, ERα). However, patients with the HER2 subtype can be treated with Herceptin, an antibody protein that inactivates HER2 (see Concept 12.3).



1 A signaling molecule (such as a growth factor) binds as a ligand to a single HER3 (a monomer), allowing it to associate with another HER.

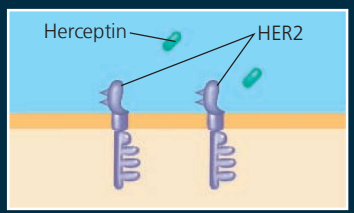
2 Even without a ligand bound, HER2 can associate with another HER. Here, HER3 and HER2 associate closely together, forming a dimer.

3 Formation of a dimer activates each monomer.

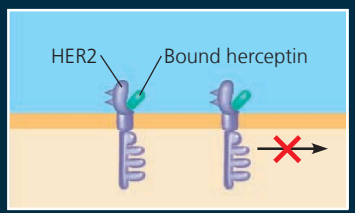
4 Each monomer adds phosphate from ATP to the other monomer, triggering a signal transduction pathway.

5 The signal is transduced through the cell, which leads to a cellular response—in this case, turning on genes that trigger cell division. HER2 cells have up to 100 times as many HER2 receptors as normal cells, so they undergo uncontrolled cell division.

Treatment with Herceptin for the HER2 subtype



1 Patients with the HER2 subtype are treated with the drug Herceptin, which can bind to HER2.



2 When bound to HER2, Herceptin blocks signaling and excessive cell division in certain patients.

MAKE CONNECTIONS

When researchers compared gene expression in normal breast cells and cells from breast cancers, they found that the genes showing the most significant differences in expression encoded signal receptors, as shown here. Given what you learned in Chapters 11, 12, and this chapter, explain why this result is not surprising.

Breast cancer is the second most common form of cancer in the United States, and the first among women. Each year, this cancer strikes over 230,000 women (and some men) in the United States and kills 40,000 (450,000 worldwide). A major problem with understanding breast cancer is its heterogeneity: Tumors differ in significant ways. Identifying differences between types of breast cancer is expected to improve treatment and decrease the mortality rate. In 2012, the Cancer Genome Atlas Network, sponsored by the National Institutes of Health, published the results of a multi-team effort that used a genomics approach to profile subtypes of breast cancer based on their molecular signatures. Four major types of breast cancer were identified (Figure 18.27). It is now routine to screen for the presence of particular signaling receptors in any breast cancer tumors, and individuals with breast cancer, along with their physicians, can now make more informed decisions about their treatments.

Inherited Predisposition and Environmental Factors Contributing to Cancer

The fact that multiple genetic changes are required to produce a cancer cell helps explain the observation that cancers can run in families. An individual inheriting an oncogene or a mutant allele of a tumor-suppressor gene is one step closer to accumulating the necessary mutations for cancer to develop than is an individual without any such mutations.

Geneticists are working to identify inherited cancer alleles so that predisposition to certain cancers can be detected early in life. About 15% of colorectal cancers, for example, involve inherited mutations. One syndrome, called hereditary nonpolyposis colon cancer (HNPCC), increases an individual's lifetime risk of colon cancer to 50–70%. HNPCC, also known as Lynch syndrome, is caused by an autosomal dominant allele of any one of a group of DNA repair genes, underscoring the importance of DNA repair systems. This syndrome is responsible for 2–5% of colon cancers. Other inherited mutations that cause colon cancer affect the tumor-suppressor gene called *adenomatous polyposis coli*, or *APC* (see Figure 18.26). This gene has multiple functions in the cell, including regulation of cell migration and adhesion. Even in patients with no family history of the disease, the *APC* gene is mutated in 60% of colorectal cancers. In these individuals, new mutations must have occurred in both *APC* alleles before the gene's function is lost. Currently, only 15% of colorectal cancers are associated with known inherited mutations, so researchers continue to try to identify "markers" that could predict the risk of developing this type of cancer.

Given the prevalence and significance of breast cancer, it is not surprising that it was one of the first cancers for which the role of inheritance was investigated. It turns out that for 5–10% of patients with breast cancer, there is evidence of a strong inherited predisposition. Geneticist Mary-Claire King began working on this problem in the mid-1970s. After 16 years of research, she convincingly demonstrated that mutations in one gene—*BRCA1*—were associated with increased susceptibility to breast cancer, a finding that flew in the face of medical opinion at the time. (*BRCA* stands for breast cancer.) Mutations in that gene or a gene called *BRCA2* are found in at least half of inherited breast cancers, and tests using DNA sequencing can detect these mutations. A woman who inherits one mutant *BRCA1* allele has a 60% probability of developing breast cancer before the age of 50, compared with only a 2% probability for an individual homozygous for the normal allele.

Mastering Biology
Interview with Mary-Claire King: Discovering breast cancer genes



BRCA1 and *BRCA2* are considered tumor-suppressor genes because their wild-type alleles protect against breast cancer and their mutant alleles are recessive. (Mutations in *BRCA1* are commonly found in the genomes of cells from basal-like breast cancers; see Figure 18.27.) The BRCA1 and BRCA2 proteins both appear to function in the cell's DNA damage repair pathway. More is known about BRCA2: Along with another protein, it helps repair breaks that occur in both strands of DNA, a function crucial for maintaining undamaged DNA.

Because DNA breakage can contribute to cancer, it makes sense that the risk of cancer can be lowered by minimizing exposure to DNA-damaging agents, such as the ultraviolet radiation in sunlight and chemicals found in cigarette smoke. Novel genomics-based analyses of specific cancers, such as the approach described in Figure 18.27, are contributing to both early diagnosis and development of treatments that interfere with expression of key genes in tumors. Ultimately, such approaches are expected to lower the death rate from cancer.

The Role of Viruses in Cancer

The study of genes associated with cancer, inherited or not, increases our basic understanding of how disruption of normal gene regulation can result in this disease. In addition to the mutations and other genetic alterations described in this section, a number of *tumor viruses* can cause cancer in various animals, including humans. In fact, one of the earliest breakthroughs in understanding cancer came in 1911, when Peyton Rous, an American pathologist, discovered a virus that causes cancer in chickens. Also, the Epstein-Barr virus, which causes infectious mononucleosis, has been linked to several types of cancer in humans, notably Burkitt's lymphoma. Papillomaviruses cause cancer of the cervix, and a virus called HTLV-1 causes a type of adult leukemia. Viruses play a role in about 15% of the cases of human cancer.

Viruses may at first seem very different from mutations as a cause of cancer. However, we now know that

viruses can interfere with gene regulation in several ways if they integrate their genetic material into the DNA of a cell. Viral integration may donate an oncogene to the cell, disrupt a tumor-suppressor gene, or convert a proto-oncogene to an oncogene. Some viruses produce proteins that inactivate p53 and other tumor-suppressor proteins, making the cell more prone to becoming cancerous. Viruses are powerful biological agents; you'll learn more about them in Chapter 19.

Mastering Biology **Animation: Causes of Cancer**

CONCEPT CHECK 18.5

1. Cancer-promoting mutations are likely to have different effects on the activity of proteins encoded by proto-oncogenes than they do on proteins encoded by tumor-suppressor genes. Explain.
2. Under what circumstances is cancer considered to have a hereditary component?
3. **MAKE CONNECTIONS** The p53 protein can activate genes involved in apoptosis. Review Concept 11.5, and discuss how mutations in genes coding for proteins that function in apoptosis could contribute to cancer.

For suggested answers, see Appendix A.

18 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

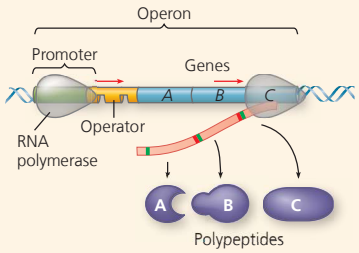
SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

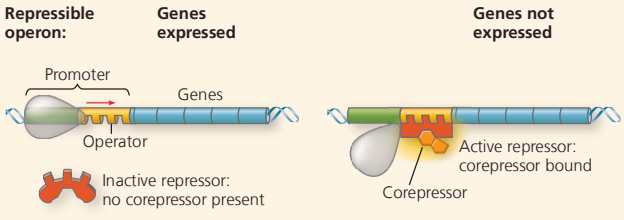
CONCEPT 18.1

Bacteria often respond to environmental change by regulating transcription *(pp. 366–370)*

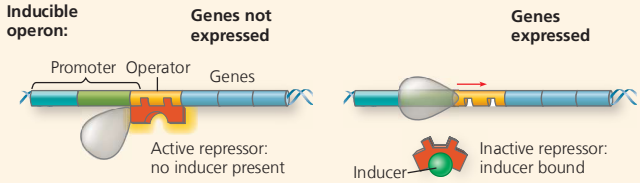
- Cells control metabolism by regulating enzyme activity or the expression of genes coding for enzymes. In bacteria, genes are often clustered into **operons**, with one promoter serving several adjacent genes. An **operator** site on the DNA switches the operon on or off, resulting in coordinate regulation of the genes.



- Both repressible and inducible operons are examples of negative gene regulation. In either type of operon, binding of a specific **repressor** protein to the operator shuts off transcription. (The repressor is encoded by a separate **regulatory gene**.) In a repressible operon (usually encoding anabolic enzymes), the repressor is active when bound to a **corepressor**, often the end product of the pathway.



- In an inducible operon (usually encoding catabolic enzymes), binding of an **inducer** to an innately active repressor inactivates the repressor and turns on transcription.

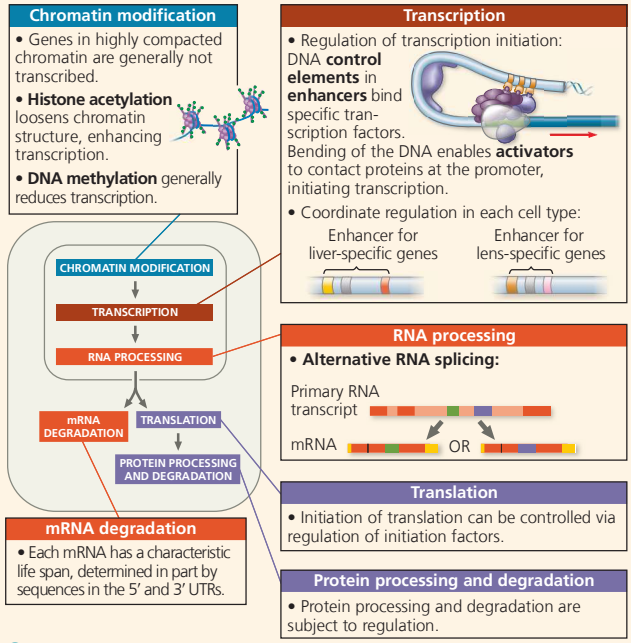


- Some operons have positive gene regulation via a stimulatory **activator** protein (such as CRP, when activated by **cyclic AMP**), binds to a site within the promoter and stimulates transcription.

? *Compare and contrast the roles of a corepressor and an inducer in negative regulation of an operon.*

CONCEPT 18.2

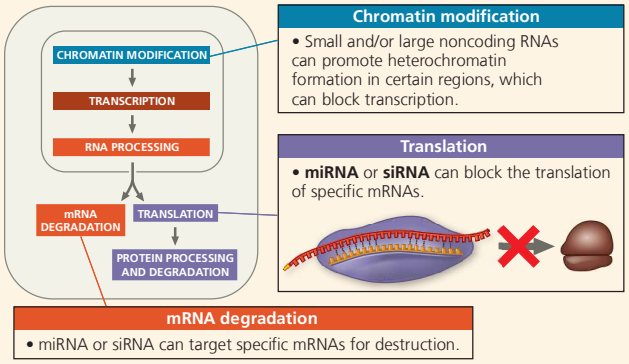
Eukaryotic gene expression is regulated at many stages *(pp. 370–379)*



? *Describe what must happen in a cell for a gene specific to that cell type to be transcribed.*

CONCEPT 18.3

Noncoding RNAs play multiple roles in controlling gene expression *(pp. 379–381)*



? *Why are miRNAs called noncoding RNAs? Explain how they participate in gene regulation.*

CONCEPT 18.4

A program of differential gene expression leads to the different cell types in a multicellular organism *(pp. 381–388)*

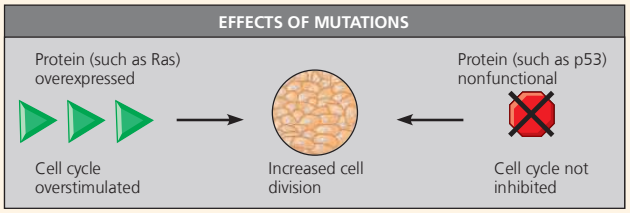
- Embryonic cells become committed to a certain fate (**determination**), and undergo **differentiation**, becoming specialized in structure and function for their determined fate. Cells differ in structure and function not because they contain different genomes but because they express different genes. **Morphogenesis** encompasses the processes that give shape to the organism and its various structures.
- Localized **cytoplasmic determinants** in the unfertilized egg are distributed differentially to daughter cells, where they regulate the expression of those cells' developmental fates. In the process called **induction**, signaling molecules from embryonic cells cause transcriptional changes in nearby target cells.
- Differentiation is marked by the appearance of tissue-specific proteins, which enable differentiated cells to carry out their specialized roles.
- In animals, **pattern formation**, the development of a spatial organization of tissues and organs, begins in the early embryo. **Positional information**, the molecular cues that control pattern formation, tells a cell its location relative to the body's axes and to other cells. In *Drosophila*, gradients of **morphogens** encoded by **maternal effect genes** determine the body axes. For example, the gradient of Bicoid protein determines the anterior-posterior axis.

? *Describe the two main processes that cause embryonic cells to head down different pathways to their final fates.*

CONCEPT 18.5

Cancer results from genetic changes that affect cell cycle control *(pp. 388–395)*

- The products of **proto-oncogenes** and **tumor-suppressor genes** control cell division. A DNA change that makes a proto-oncogene overly active converts it to an **oncogene**, which may promote extra cell division and cancer. A tumor-suppressor gene encodes a protein that inhibits abnormal cell division. A mutation that reduces the activity of its protein product may lead to excessive cell division and cancer.
- Many proto-oncogenes and tumor-suppressor genes encode components of growth-stimulating and growth-inhibiting signaling pathways, respectively, and mutations in them can interfere with normal cell-signaling pathways. A hyperactive version of a protein in a stimulatory pathway, such as **Ras** (a G protein), functions as an oncogene protein. A defective version of a protein in an inhibitory pathway, such as **p53** (a transcription activator), fails to act as a tumor suppressor.



- In the multistep model of cancer development, normal cells are converted to cancer cells by the accumulation of mutations affecting proto-oncogenes and tumor-suppressor genes. Technical advances in DNA and mRNA sequencing are enabling cancer treatments that are more individually based.
- Genomics-based studies have resulted in researchers proposing four subtypes of breast cancer, based on expression of genes by tumor cells.
- Individuals who inherit a mutant allele of a proto-oncogene or tumor-suppressor gene have a predisposition to develop a particular cancer. Certain viruses promote cancer by integration of viral DNA into a cell's genome.

? *Compare the usual functions of proteins encoded by proto-oncogenes with those of proteins encoded by tumor-suppressor genes.*

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. If a particular operon encodes enzymes for making an essential amino acid and is regulated like the *trp* operon, then
   (A) the amino acid inactivates the repressor.
   (B) the repressor is active in the absence of the amino acid.
   (C) the amino acid acts as a corepressor.
   (D) the amino acid turns on transcription of the operon.
2. Muscle cells differ from nerve cells mainly because they
   (A) express different genes.
   (B) contain different genes.
   (C) use different genetic codes.
   (D) have unique ribosomes.
3. The functioning of enhancers is an example of
   (A) a eukaryotic equivalent of prokaryotic promoter functioning.
   (B) transcriptional control of gene expression.
   (C) the stimulation of translation by initiation factors.
   (D) post-translational control that activates certain proteins.
4. Cell differentiation always involves
   (A) transcription of the *myoD* gene.
   (B) the movement of cells.
   (C) the production of tissue-specific proteins.
   (D) the selective loss of certain genes from the genome.
5. Which of the following is an example of post-transcriptional control of gene expression?
   (A) the addition of methyl groups to cytosine bases of DNA
   (B) the binding of transcription factors to a promoter
   (C) the removal of introns and alternative splicing of exons
   (D) gene amplification contributing to cancer

Levels 3-4: Applying/Analyzing

6. What would occur if the repressor of an inducible operon were mutated so it could not bind the operator?
(A) irreversible binding of the repressor to the promoter
(B) reduced transcription of the operon's genes
(C) buildup of a substrate for the pathway controlled by the operon
(D) continuous transcription of the operon's genes

7. Absence of *bicoid* mRNA from a *Drosophila* egg leads to the absence of anterior larval body parts and mirror-image duplication of posterior parts. This is evidence that the product of the *bicoid* gene
(A) normally leads to formation of head structures.
(B) normally leads to formation of tail structures.
(C) is transcribed in the early embryo.
(D) is a protein present in all head structures.

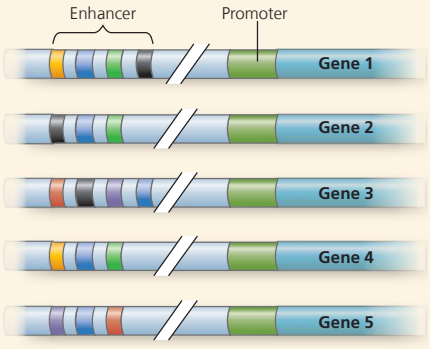
8. Which of the following statements about the DNA in one of your brain cells is true?
(A) Most of the DNA codes for protein.
(B) The majority of genes are likely to be transcribed.
(C) It is the same as the DNA in one of your liver cells.
(D) Each gene lies immediately adjacent to an enhancer.

9. Within a cell, the amount of protein made using a given mRNA molecule in that cell depends partly on
(A) the degree of DNA methylation.
(B) the rate at which the mRNA is degraded.
(C) the number of introns present in the mRNA.
(D) the types of ribosomes present in the cytoplasm.

10. Proto-oncogenes can change into oncogenes that cause cancer. Which of the following best explains the presence of these potential time bombs in eukaryotic cells?
(A) Proto-oncogenes first arose from viral infections.
(B) Proto-oncogenes are mutant versions of normal genes.
(C) Proto-oncogenes are genetic "junk."
(D) Proto-oncogenes normally help regulate cell division.

Levels 5-6: Evaluating/Creating

11. DRAW IT The diagram below shows five genes, including their enhancers, from the genome of a certain species. Imagine that yellow, blue, green, black, red, and purple activator proteins exist that can bind to the appropriately color-coded control elements in the enhancers of these genes.



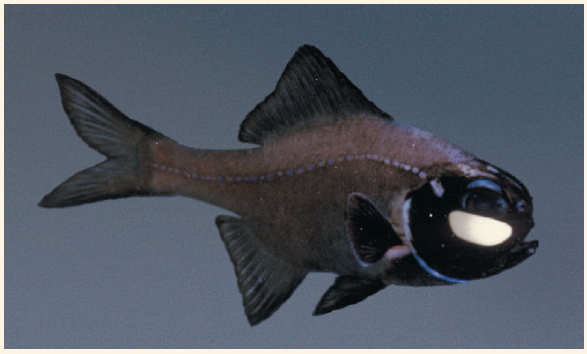
(a) Draw an X above enhancer elements (of all the genes) that would have activators bound in a cell in which only gene 5 is transcribed. Identify which colored activators would be present.
(b) Draw a dot above all enhancer elements that would have activators bound in a cell in which the green, blue, and yellow activators are present. Identify which gene(s) would be transcribed.
(c) Imagine that genes 1, 2, and 4 code for nerve-specific proteins, and genes 3 and 5 are skin-specific. Identify which activators would be present in each cell type to ensure transcription of the appropriate genes.

12. EVOLUTION CONNECTION DNA sequences can act as "tape measures of evolution" (see Concept 5.6). Scientists analyzing the human genome sequence were surprised to find that some of the regions of the human genome that are most highly conserved (similar to comparable regions in other species) don't code for proteins. Propose a possible explanation for this observation.

13. SCIENTIFIC INQUIRY Prostate cells usually require testosterone and other androgens to survive. But some prostate cancer cells thrive despite treatments that eliminate androgens. One hypothesis is that estrogen, often considered a female hormone, may be activating genes normally controlled by an androgen in these cancer cells. Describe one or more experiments to test this hypothesis. (See Figure 11.9 to review the action of these steroid hormones.)

14. WRITE ABOUT A THEME: INTERACTIONS In a short essay (100–150 words), discuss how the processes shown in Figure 18.2 are examples of feedback mechanisms regulating biological systems in bacterial cells.

15. SYNTHESIZE YOUR KNOWLEDGE



The flashlight fish has an organ under its eye that emits light, which serves to startle predators and attract prey, and allows the fish to communicate with other fish. Some species can rotate the organ inside and then out, so the light appears to flash on and off. The light is actually emitted by bacteria (of the genus *Vibrio*) that live in the organ in a mutualistic relationship with the fish. (The bacteria receive nutrients from the fish.) The bacteria must multiply until they reach a certain density in the organ (a "quorum"; see Concept 11.1), at which point they all begin emitting light at the same time. There is a group of six or so genes, called *lux* genes, whose gene products are necessary for light formation. Given that these bacterial genes are regulated together, propose a hypothesis for how the genes are organized and regulated.

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
Can cancer be stopped before it begins?
Go to "Arrested Development: When Cells Make Mistakes"
at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.

19 Viruses

KEY CONCEPTS

19.1 **A virus consists of a nucleic acid surrounded by a protein coat** *p. 399*

19.2 **Viruses replicate only in host cells** *p. 401*

19.3 **Viruses and prions are formidable pathogens in animals and plants** *p. 408*

Study Tip

Make flowcharts: Make a simple flowchart for each viral replicative cycle in this chapter. Below is the beginning of an example for the lysogenic cycle. Note the similarities and differences between the replicative cycles.

Lysogenic cycle

Virus attaches to bacterial cell and injects genome.

↓

Viral genome is integrated into the bacterial chromosome.

↓

Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 19
- HHMI Animation: HIV Replicative Cycle
- HHMI Video: Interview with Katie Walter

For Instructors to Assign (in Item Library)
- Tutorial: Viral Replication
- Activity: Phage Lysogenic and Lytic Cycles
- Tutorial: CRISPR: The Discovery of Bacterial Adaptive Immunity

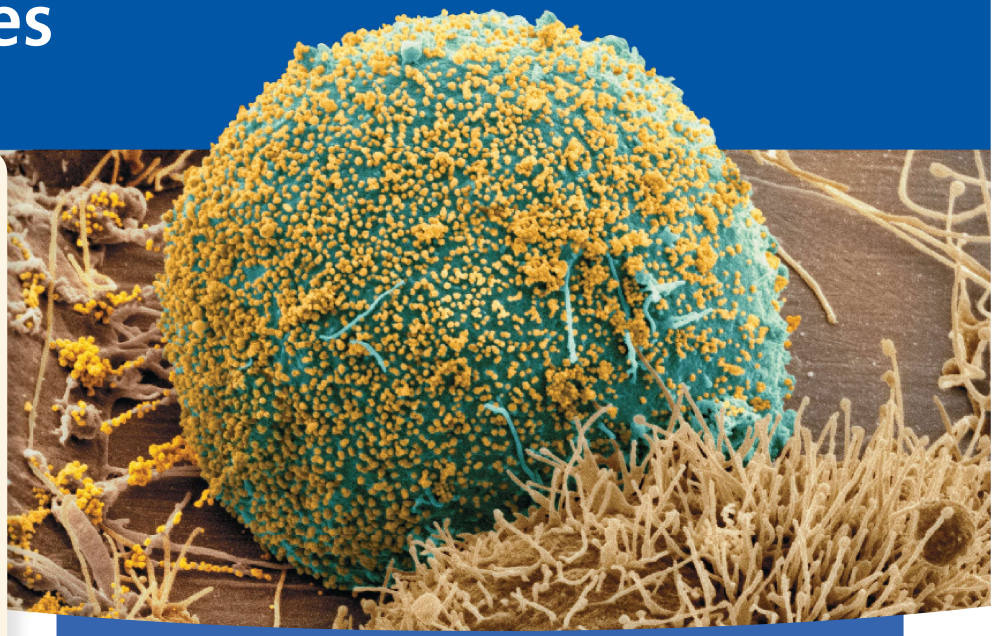
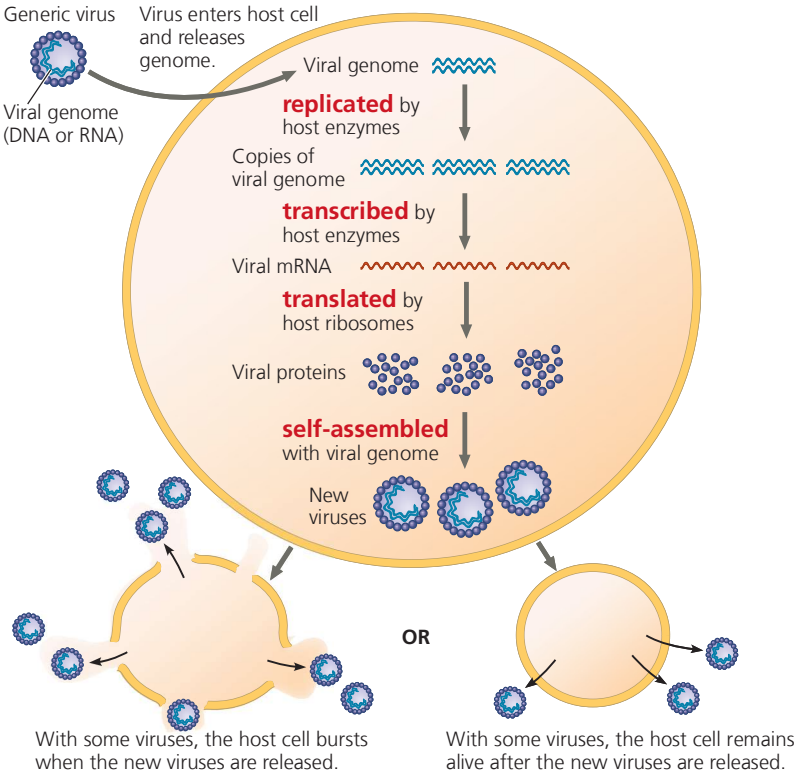


Figure 19.1 A human immune cell (blue) infected by human immunodeficiency virus (HIV) is releasing more HIV viruses (gold dots), which will go on to infect other cells. Left untreated, HIV destroys vital immune system cells, causing AIDS. This disease has killed about 35 million people worldwide.

How does a virus make more viruses?

A virus consists only of nucleic acid, proteins, and sometimes a membranous envelope. After infecting a host cell, it uses the host cell's molecules to make new viruses, as shown in this simplified diagram.



CONCEPT 19.1

A virus consists of a nucleic acid surrounded by a protein coat

Compared with eukaryotic and even prokaryotic cells, most viruses are much smaller and simpler in structure. Lacking the structures and metabolic machinery found in a cell, a **virus** is an infectious particle consisting of little more than genes packaged in a protein coat.

Are viruses living or nonliving? Early on, they were considered biological chemicals; the Latin root for *virus* means "poison." Viruses can cause a wide variety of diseases, so researchers in the late 1800s saw a parallel with bacteria and proposed that viruses were the simplest of living forms. However, viruses cannot reproduce or carry out metabolic activities outside of a host cell. Most biologists would probably agree that viruses are not alive, but instead exist in a shady area between life-forms and chemicals. The simple phrase used by two researchers describes them aptly: Viruses lead "a kind of borrowed life."

The Discovery of Viruses: *Scientific Inquiry*

Scientists detected viruses indirectly long before they were able to see them. The story of how viruses were discovered begins in 1883. A German scientist named Adolf Mayer was studying tobacco mosaic disease, which stunts the growth of tobacco plants and gives their leaves a mottled, or mosaic, coloration. Mayer discovered that he could transmit the disease from plant to plant by rubbing sap extracted from diseased leaves onto healthy plants. After an unsuccessful search for an infectious microorganism in the sap, he suggested that the disease was caused by unusually small bacteria that were invisible under a microscope. This hypothesis was tested a decade later by Dmitri Ivanowsky, a Russian biologist who passed sap from infected tobacco leaves through a filter designed to remove bacteria. After filtration, the sap still produced mosaic disease.

But Ivanowsky reasoned that perhaps the bacteria were small enough to pass through the filter or made a toxin that could do so. The second possibility was ruled out when the Dutch botanist Martinus Beijerinck carried out a classic series of experiments that showed that the infectious agent in the filtered sap could replicate **(Figure 19.2)**.

In fact, the pathogen replicated only within the host it infected. In further experiments, Beijerinck showed that unlike bacteria used in the lab at that time, the mysterious agent of mosaic disease could not be cultivated on nutrient media in test tubes or petri dishes. Beijerinck imagined a replicating particle much smaller and simpler than a bacterium, and he is generally credited with being the first scientist to voice the concept of a virus. His suspicions were confirmed in 1935 when the American scientist Wendell Stanley crystallized the infectious particle, now known as tobacco mosaic virus (TMV). Subsequently, TMV and many other viruses were actually seen with the help of the electron microscope.

▼ Figure 19.2 Inquiry

What causes tobacco mosaic disease?

Experiment In the late 1800s, Martinus Beijerinck, of the Technical School in Delft, the Netherlands, investigated the properties of the agent that causes tobacco mosaic disease (then called spot disease).



1 Extracted sap from tobacco plant with tobacco mosaic disease

2 Passed sap through a porcelain filter known to trap bacteria

3 Rubbed filtered sap on healthy tobacco plants



4 Healthy plants became infected

Results When the filtered sap was rubbed on healthy plants, they became infected. Their sap, extracted and filtered, could then act as a source of infection for another group of plants. Each successive group of plants developed the disease to the same extent as earlier groups.

Conclusion The infectious agent was apparently not a bacterium because it could pass through a bacterium-trapping filter. The pathogen must have been replicating in the plants because its ability to cause disease was undiluted after several transfers from plant to plant.

Data from M. J. Beijerinck, Concerning a *contagium vivum fluidum* as cause of the spot disease of tobacco leaves, *Verhandelingen der Koninkyke akademie Wettenschappen te Amsterdam* 65:3–21 (1898). Translation published in English as Phytopathological Classics Number 7 (1942), American Phytopathological Society Press, St. Paul, MN.

WHAT IF? *If Beijerinck had observed that the infection of each group was weaker than that of the previous group and that ultimately the sap could no longer cause disease, what might he have concluded?*

Structure of Viruses

The tiniest viruses are only 20 nm in diameter—smaller than a ribosome. Millions could easily fit on a pinhead. Even the largest known virus, which has a diameter of 1,500 nanometers (1.5 µm), is barely visible under the light microscope. Stanley's discovery that some viruses could be crystallized was exciting and puzzling news. Not even the simplest of cells can aggregate into regular crystals. But if viruses are not cells, then what are they? Examining the structure of a virus more closely reveals that it is an infectious particle consisting of one or more molecules of a nucleic acid enclosed in a protein coat and, for some viruses, surrounded by a membranous envelope.

The simple structure of viruses make them a useful biological system: To a large extent, molecular biology was born in the laboratories of biologists studying viruses that infect bacteria. Experiments using these viruses provided evidence that genes are made of nucleic acids, and they were critical in working out the molecular mechanisms of the fundamental processes of DNA replication, transcription, and translation.

Viral Genomes

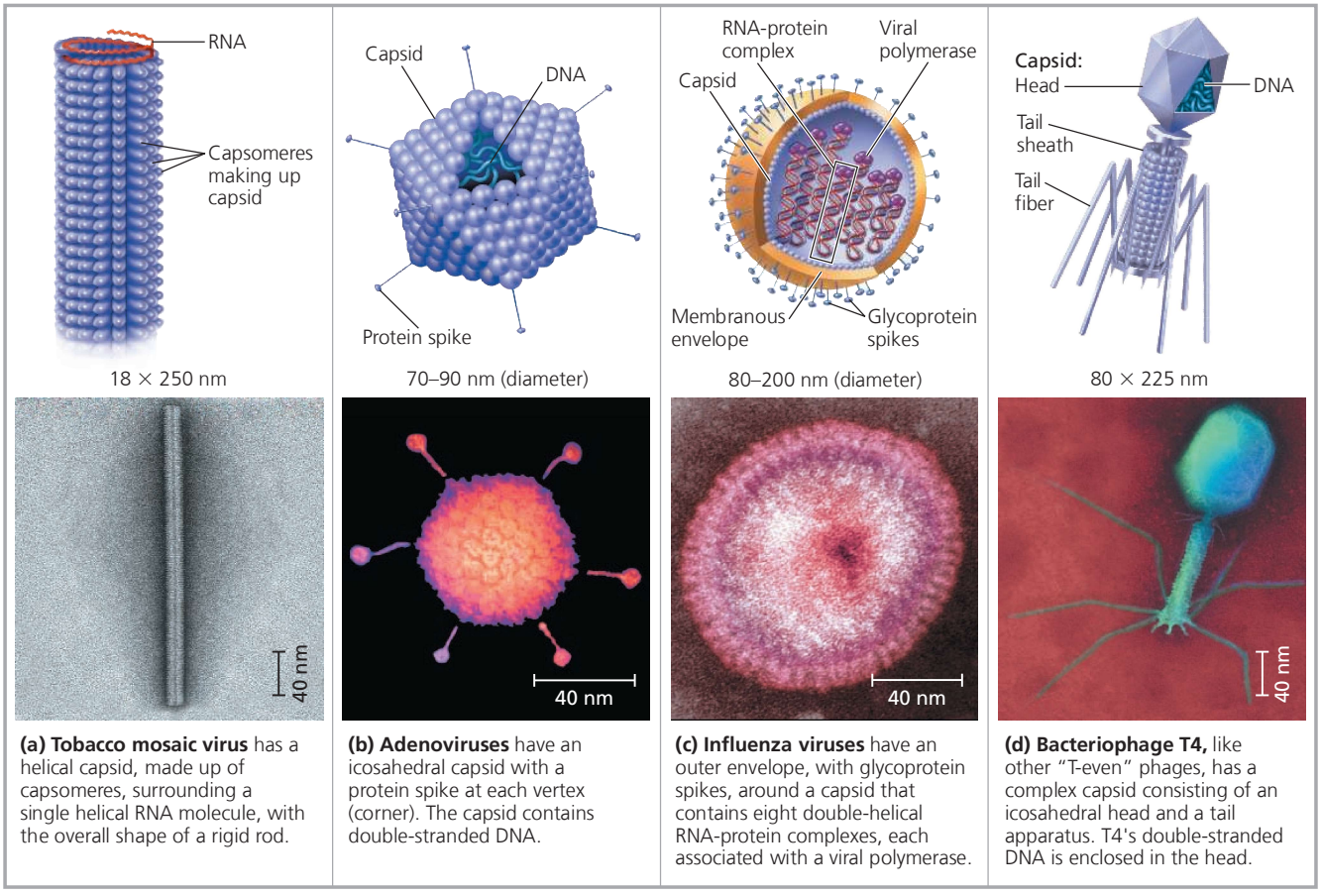
We usually think of genes as being made of double-stranded DNA, but many viruses defy this convention. Their genomes may consist of double-stranded DNA, single-stranded DNA, double-stranded RNA, or single-stranded RNA, depending on the type of virus. A virus is called a DNA virus or an RNA virus based on the kind of nucleic acid that makes up its genome. In either case, the genome is usually organized as a single linear or circular molecule of nucleic acid, although the genomes of some viruses consist of multiple molecules of nucleic acid. The smallest viruses known have only three genes in their genome, while the largest have several hundred to 2,000. For comparison, bacterial genomes contain about 200 to a few thousand genes.

Capsids and Envelopes

The protein shell enclosing the viral genome is called a **capsid**. Depending on the type of virus, the capsid may be rod-shaped, polyhedral, or more complex in shape. Capsids are built from a large number of protein subunits called *capsomeres*, but the number of different *kinds* of proteins in a capsid is usually small. Tobacco mosaic virus has a rigid, rod-shaped capsid made from over 1,000 molecules of a single type of protein arranged in a helix; rod-shaped viruses are commonly called *helical viruses* for this reason **(Figure 19.3a)**. Adenoviruses, which infect the respiratory tracts of animals, have 252 identical protein molecules arranged in a polyhedral capsid with 20 triangular facets—an icosahedron; thus, these and other similarly shaped viruses are referred to as *icosahedral viruses* **(Figure 19.3b)**.

Some viruses have accessory structures that help them infect their hosts. For instance, a membranous envelope surrounds the capsids of influenza viruses and many other viruses found in animals **(Figure 19.3c)**. These **viral envelopes**, which are derived from the membranes of the host cell, contain host cell phospholipids and membrane proteins. They also contain proteins and glycoproteins of viral origin. (Glycoproteins are proteins with

▼ Figure 19.3 Viral structure. Viruses are made up of nucleic acid (DNA or RNA) enclosed in a protein coat (the capsid) and sometimes further wrapped in a membranous envelope. The individual protein subunits making up the capsid are called capsomeres. Although diverse in size and shape, viruses have many common structural features. (All micrographs are colorized TEMs.)



(a) Tobacco mosaic virus has a helical capsid, made up of capsomeres, surrounding a single helical RNA molecule, with the overall shape of a rigid rod.

(b) Adenoviruses have an icosahedral capsid with a protein spike at each vertex (corner). The capsid contains double-stranded DNA.

(c) Influenza viruses have an outer envelope, with glycoprotein spikes, around a capsid that contains eight double-helical RNA-protein complexes, each associated with a viral polymerase.

(d) Bacteriophage T4, like other "T-even" phages, has a complex capsid consisting of an icosahedral head and a tail apparatus. T4's double-stranded DNA is enclosed in the head.

carbohydrates covalently attached.) Some viruses carry a few viral enzymes, such as viral polymerase, within their capsids.

Many of the most complex capsids are found among the viruses that infect bacteria, called **bacteriophages**, or simply **phages**. The first phages studied included seven that infect *Escherichia coli* (*E. coli*). These seven phages were named type 1 (T1), type 2 (T2), and so forth, in the order of their discovery. (The T2 phage was used in the experiment that established DNA as the genetic material; see Figure 16.4.) The three "T-even" phages (T2, T4, and T6) turned out to be very similar in structure. Their capsids have elongated icosahedral heads enclosing their DNA. Attached to the head is a protein tail piece with fibers by which the phages attach to a bacterial cell **(Figure 19.3d)**. In the next section, we'll examine how these few viral parts function together with cellular components to produce large numbers of viral progeny.

CONCEPT CHECK 19.1

1. **VISUAL SKILLS** Describe one similarity and two differences between the structures of tobacco mosaic virus (TMV) and influenza virus (see Figure 19.3).
2. **MAKE CONNECTIONS** Bacteriophages were used to provide evidence that DNA carries genetic information (see Figure 16.4). Briefly describe the experiment carried out by Hershey and Chase, including in your description why the researchers chose to use phages.

For suggested answers, see Appendix A.

CONCEPT 19.2

Viruses replicate only in host cells

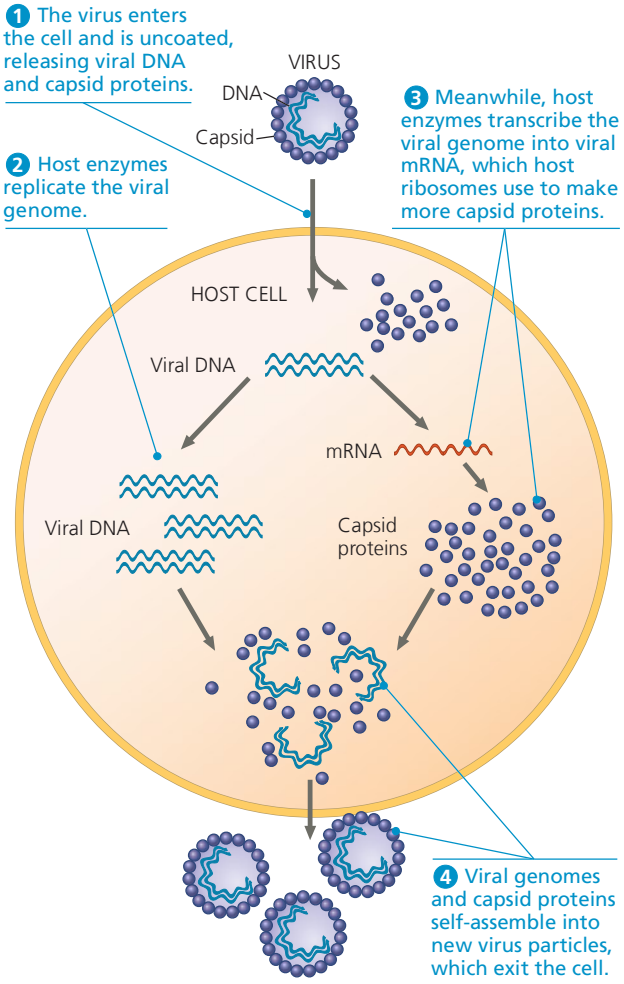
Viruses lack metabolic enzymes and equipment for making proteins, such as ribosomes. They are obligate intracellular parasites; in other words, they can replicate only within a host cell. It is fair to say that viruses in isolation are merely packaged sets of genes in transit from one host cell to another.

Each particular virus can infect cells of only a limited number of host species, called the **host range** of the virus. This host specificity results from the evolution of recognition systems by the virus. Viruses usually identify host cells by a "handshake" fit between viral surface proteins and specific receptor molecules on the outside of cells. Such receptor molecules are most often proteins that carry out necessary functions for the host cell and have been adopted by viruses as portals of entry. Some viruses have broad host ranges. For example, West Nile virus and equine encephalitis virus are distinctly different viruses that can each infect mosquitoes, birds, horses, and humans. Other viruses have host ranges so narrow that they infect only a single species. Measles virus, for instance, can infect only humans. Furthermore, viral infection of multicellular eukaryotes is usually limited to particular tissues. Human cold viruses infect only the cells lining the upper respiratory tract, and the HIV seen in Figure 19.1 binds to receptors present only on certain types of immune cells.

General Features of Viral Replicative Cycles

A viral infection begins when a virus binds to a host cell and the viral genome makes its way inside **(Figure 19.4)**. The mechanism of genome entry depends on the type of virus and the type of host cell. For example, T-even phages use their elaborate tail apparatus to inject DNA into a bacterium (see Figure 19.3d). Other viruses are taken up by endocytosis or, in the case of enveloped viruses, by fusion of the viral envelope with the host's plasma membrane. Once the viral genome is inside, the proteins it encodes can commandeer the host, reprogramming the cell to copy the viral genome and manufacture viral proteins. The host provides the nucleotides for making viral nucleic acids, as well as enzymes, ribosomes,

▼ Figure 19.4 A simplified viral replicative cycle. A virus is an intracellular parasite that uses the equipment and small molecules of its host cell to replicate. In this simplest of viral cycles, the parasite is a DNA virus with a capsid consisting of a single type of protein.



MAKE CONNECTIONS *Label each of the straight gray arrows to identify the process that is occurring. Review Figure 17.25.*

Mastering Biology **Animation: Simplified Viral Replicative Cycle**

tRNAs, amino acids, ATP, and other components needed for making the viral proteins. Many DNA viruses use the DNA polymerases of the host cell to synthesize new genomes along the templates provided by the viral DNA. In contrast, to replicate their genomes, RNA viruses use virally encoded RNA polymerases that can use RNA as a template. (Uninfected cells generally make no enzymes for carrying out this process.)

After the viral nucleic acid molecules and capsomeres are produced, they spontaneously self-assemble into new viruses. In fact, researchers can separate the RNA and capsomeres of TMV and then reassemble complete viruses simply by mixing the components together under the right conditions. The simplest type of viral replicative cycle ends with the exit of hundreds or thousands of viruses from the infected host cell, a process that often damages or destroys the cell. Such cellular damage and death, as well as the body's responses to this destruction, cause many of the symptoms associated with viral infections. The viral progeny that exit a cell have the potential to infect additional cells, spreading the viral infection.

There are many variations on the simplified viral replicative cycle we have just described. We will now take a look at some of these variations in bacterial viruses (phages) and animal viruses; later in the chapter, we will consider plant viruses.

Replicative Cycles of Phages

Phages are the best understood of all viruses, although some of them are also among the most complex. Research on phages led to the discovery that some double-stranded DNA viruses can replicate by two alternative mechanisms: the lytic cycle and the lysogenic cycle.

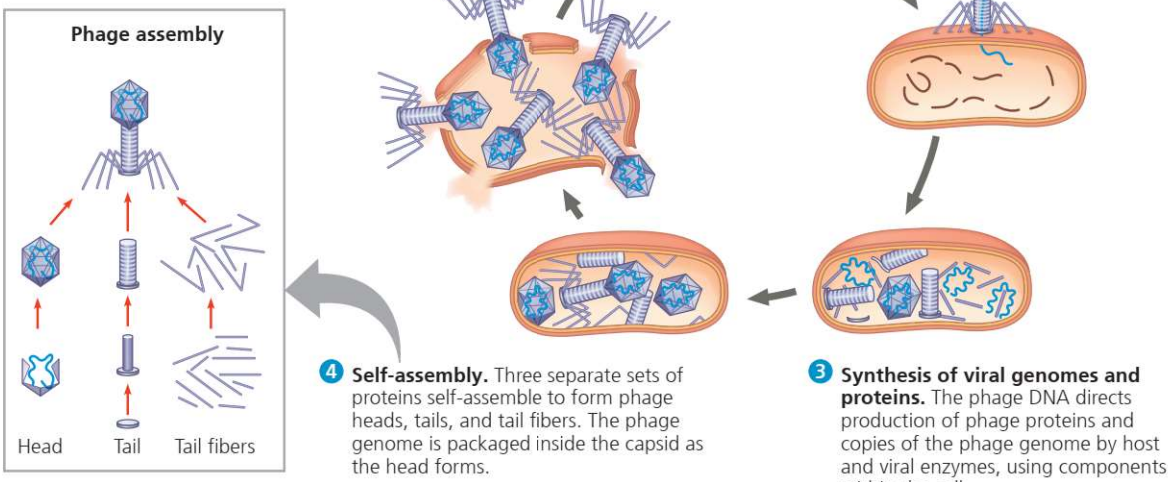
The Lytic Cycle

A phage replicative cycle that culminates in death of the host cell is known as a **lytic cycle**. The term *lytic* refers to the last stage of infection, during which the bacterium lyses (breaks open) and releases the phages that were produced within the cell. Each of these phages can then infect a healthy cell, and a few successive lytic cycles can destroy an entire bacterial population in just a few hours. A phage that replicates only by a lytic cycle is a **virulent phage**. **Figure 19.5** illustrates the major steps in the lytic cycle of T4, a typical virulent phage.

The Lysogenic Cycle

Instead of lysing their host cells, many phages coexist with them in a state called lysogeny. In contrast to the lytic cycle, which kills the host cell, the **lysogenic cycle** allows

► Figure 19.5 The lytic cycle of phage T4, a virulent phage. Phage T4 has almost 300 genes, which are transcribed and translated using the host cell's machinery. One of the first phage genes translated after the viral DNA enters the host cell codes for an enzyme that degrades the host cell's DNA (step 2); the phage DNA is protected from breakdown because it contains a modified form of cytosine that is not recognized by the phage enzyme. The entire lytic cycle, from the phage's first contact with the cell surface to cell lysis, takes only 20–30 minutes at 37°C.



Mastering Biology Animation: Phage Lytic Cycle

replication of the phage genome without destroying the host. Phages capable of using both modes of replicating within a bacterium are called **temperate phages**. A temperate phage called lambda, written with the Greek letter λ, has been widely used in biological research. Phage λ resembles T4, but its tail has only one short tail fiber.

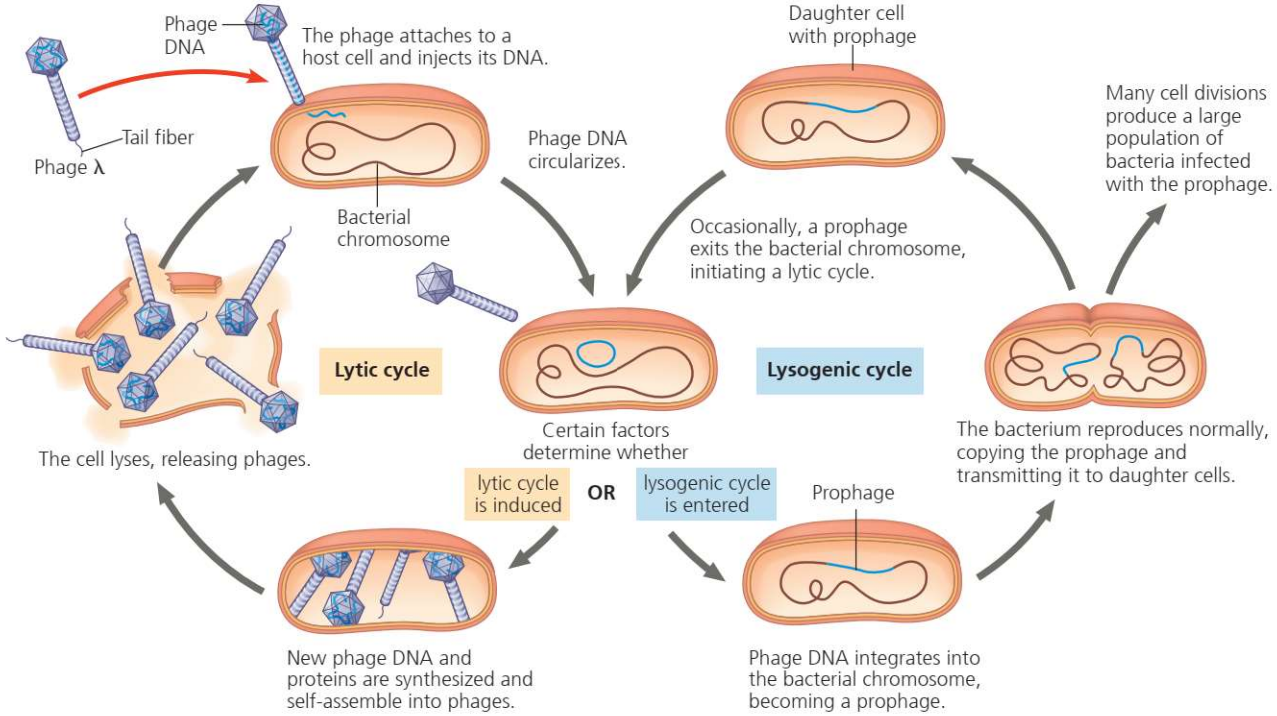
Infection of an *E. coli* cell by phage λ begins when the phage binds to the surface of the cell and injects its linear DNA genome **(Figure 19.6)**. Within the host, the λ DNA molecule forms a circle. What happens next depends on the replicative mode: lytic cycle or lysogenic cycle. During a lytic cycle, the viral genes immediately turn the host cell into a λ-producing factory, and the cell soon lyses and releases its virus progeny. During a lysogenic cycle, however, the λ DNA molecule is incorporated into a specific site on the *E. coli* chromosome by viral proteins that break both circular DNA molecules and join them to each other. When integrated into the bacterial chromosome in this way, the viral DNA is known as a **prophage**. One prophage gene codes for a protein that prevents transcription of most of the other prophage genes. Thus, the phage genome is mostly silent within the bacterium. Every time the *E. coli* cell prepares to divide, it replicates the phage DNA along with its own chromosome such that each daughter cell inherits a prophage. A single infected cell can quickly give rise to a large population of bacteria carrying the virus in prophage form. This mechanism enables viruses to propagate without killing the host cells on which they depend.

The term *lysogenic* signifies that prophages are capable of generating active phages that lyse their host cells. This occurs when the λ genome (or that of another temperate phage) is induced to exit the bacterial chromosome and initiate a lytic cycle. An environmental signal, such as a certain chemical or high-energy radiation, usually triggers the switchover from the lysogenic to the lytic mode.

In addition to the gene for the viral protein that prevents transcription, a few other prophage genes may be expressed during lysogeny. Expression of these genes may alter the host's phenotype, a phenomenon that can have important medical significance. For example, the three species of bacteria that cause the human diseases diphtheria, botulism, and scarlet fever would not be so harmful to humans without certain prophage genes that cause the host bacteria to make toxins. And the difference between the *E. coli* strain in our intestines and the O157:H7 strain that has caused several deaths by food poisoning appears to be the presence of toxin genes of prophages in the O157:H7 strain.

▼ Figure 19.6 The lytic and lysogenic cycles of phage λ, a temperate phage. After entering the bacterial cell and circularizing, the λ DNA can immediately initiate the production of a large number of progeny phages (lytic cycle) or integrate into the bacterial chromosome (lysogenic cycle). In most cases, phage λ follows the lytic pathway, which is similar to that detailed in Figure 19.5. However, once a lysogenic cycle begins, the prophage may be carried in the host cell's chromosome for many generations. Phage λ has one main tail fiber, which is short.

Mastering Biology
Animation: Phage Lysogenic and Lytic Cycles



Bacterial Defenses Against Phages

After reading about the lytic cycle, you may have wondered why phages haven't exterminated all bacteria. Lysogeny is one major reason why bacteria have been spared from extinction caused by phages. Bacteria also have their own defenses against phages. First, natural selection favors bacterial mutants with surface proteins that are no longer recognized as receptors by a particular type of phage. Second, when phage DNA does enter a bacterium, the DNA often is identified as foreign and cut up by cellular enzymes called **restriction enzymes**, which are so named because they *restrict* a phage's ability to replicate within the bacterium. (Restriction enzymes are used in molecular biology and DNA cloning techniques; see Concept 20.1.) The bacterium's own DNA is methylated in a way that prevents attack by its own restriction enzymes. A third defense is a system present in both bacteria and archaea called the *CRISPR-Cas system*, which you learned about in Concept 17.5.

Mastering Biology **HHMI Videos: Jennifer Doudna: CRISPR-Cas9 was First Discovered in Bacteria • Jennifer Doudna: Science is a Serendipitous Endeavor**

The CRISPR-Cas system was discovered during a study of repetitive DNA sequences present in the genomes of many prokaryotes. These sequences, which puzzled scientists, were named clustered regularly interspaced short palindromic repeats (CRISPRs) because each sequence reads the same forward and backward (a palindrome), with different stretches of "spacer DNA" in between the repeats. At first, scientists assumed the spacer DNA sequences were random and meaningless, but analysis by several research groups showed that each spacer sequence corresponded to DNA from a particular phage that had infected the cell. Further studies revealed that particular nuclease proteins interact with the CRISPR region. These nucleases, called Cas (CRISPR-associated) proteins, can identify and cut phage DNA, thereby defending the bacterium against phage infection.

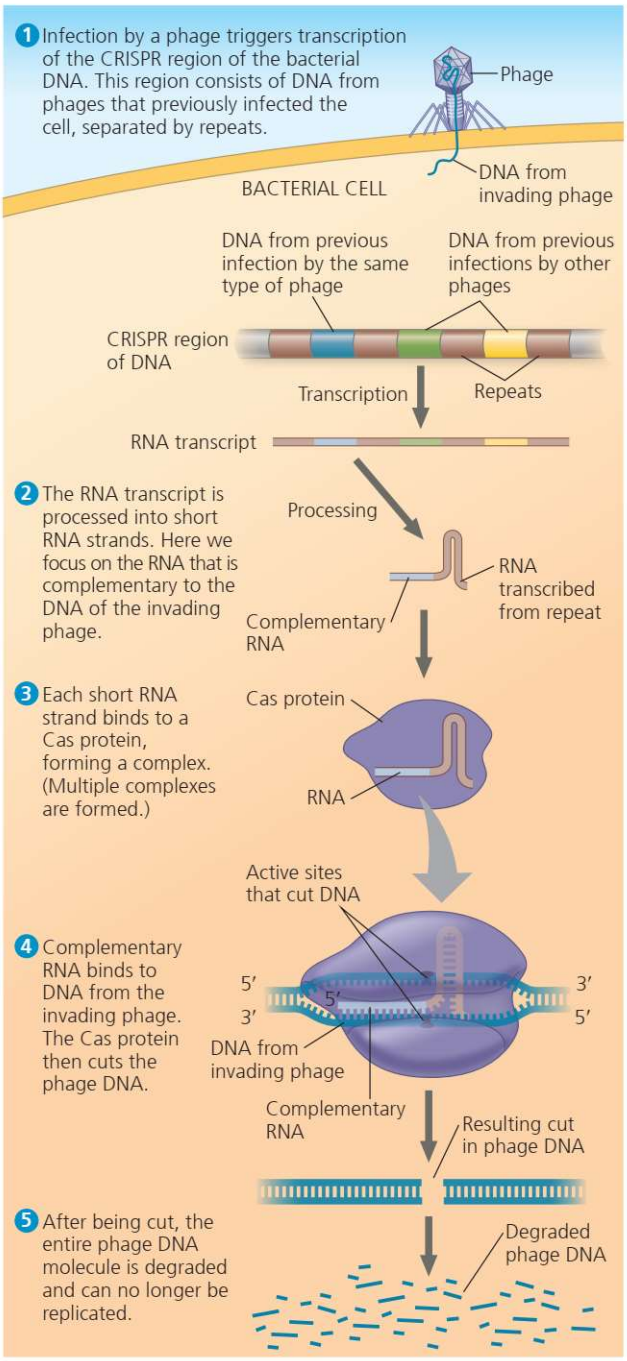
When a phage infects a bacterial cell that has the CRISPR-Cas system, the DNA of the invading phage is stored, integrated into the genome between two repeat sequences **(Figure 19.7)**. If the cell survives the infection, any further attempt by the same type of phage to infect this cell (or its offspring) triggers transcription of the CRISPR region into RNA molecules. These RNAs are cut into pieces and then bound by Cas proteins, such as the Cas9 protein (see Figure 17.28). The Cas protein uses a portion of the phage-related RNA as a homing device to identify the invading phage DNA and cut it, leading to its destruction.

Just as natural selection favors bacteria that have receptors altered by mutation or that have enzymes that cut phage DNA, it also favors phage mutants that can bind to altered receptors or that are resistant to enzymes. Thus, the bacterium-phage relationship is in constant evolutionary flux.

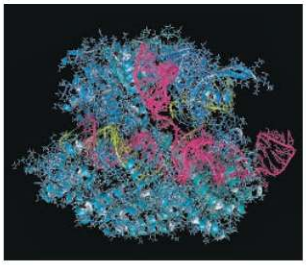
Replicative Cycles of Animal Viruses

Everyone has suffered from viral infections, whether cold sores, influenza, or the common cold. Like all viruses, those

▼ Figure 19.7 The CRISPR-Cas system: a type of bacterial immune system.



▶ Computer model of CRISPR-Cas9 gene editing complex from *Streptococcus pyogenes*





Mastering Biology **Interview with Francisco Mojica: Discovering CRISPR in bacteria (see the interview before Chapter 13)**

that cause illness in humans and other animals can replicate only inside host cells. Many variations on the basic scheme of viral infection and replication are represented among the animal viruses. Key variables are the nature of the viral genome (double- or single-stranded DNA or RNA) and the presence or absence of an envelope.

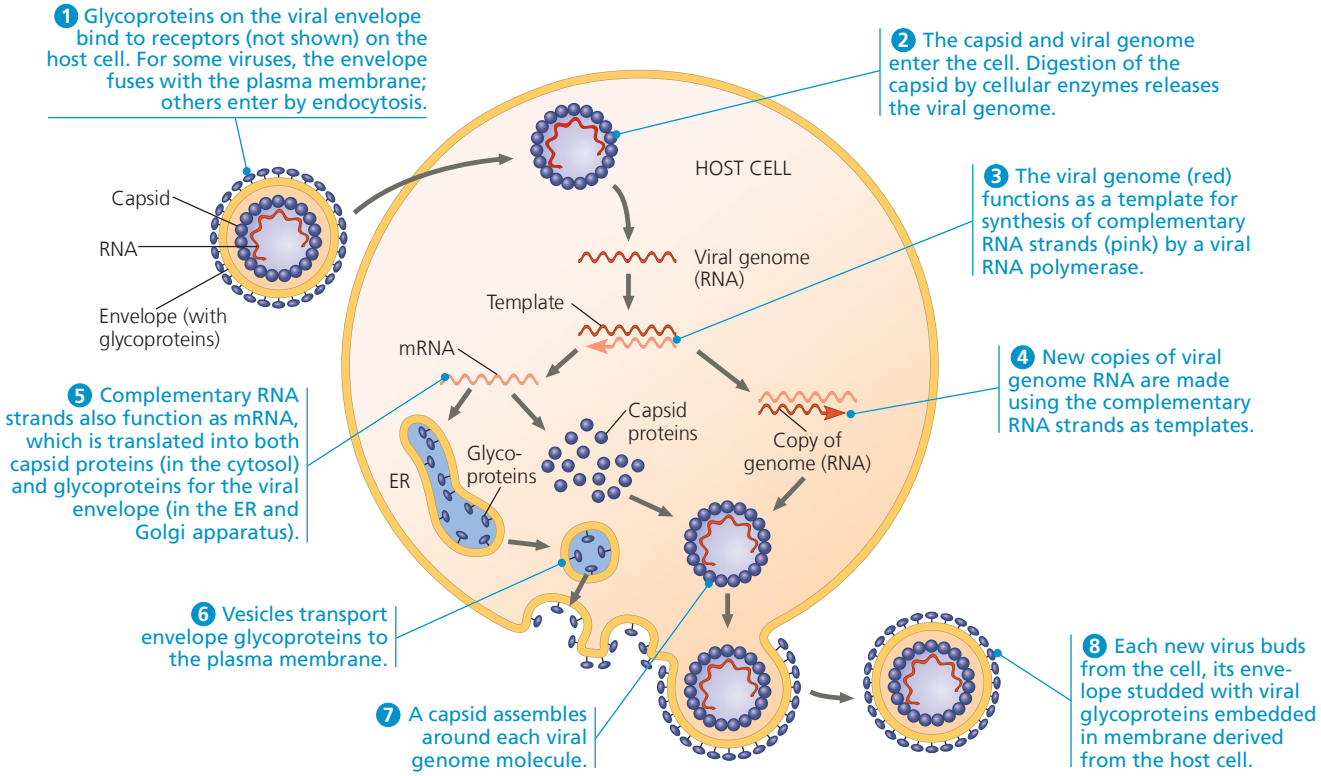
Whereas there are only a few bacteriophages that have an envelope or RNA genome, many animal viruses have both. In fact, nearly all animal viruses with RNA genomes have an envelope, as do some with DNA genomes. Rather than consider all the mechanisms of viral infection and replication, we will focus first on the roles of viral envelopes and then on the functioning of RNA as the genetic material of many animal viruses.

Viral Envelopes

An animal virus equipped with an envelope—that is, a membranous outer layer—uses it to enter the host cell. Protruding from the outer surface of this envelope are viral glycoproteins that bind to specific receptor molecules on the surface of a host cell. **Figure 19.8** outlines the events in the replicative cycle of an enveloped virus with an RNA genome. Ribosomes bound to the endoplasmic reticulum (ER) of the host cell make the protein parts of the envelope glycoproteins; cellular enzymes in the ER and Golgi apparatus then add the sugars. The resulting viral glycoproteins, embedded in membrane derived from the host cell, are transported to the cell surface. In a process much like exocytosis (see Figure 7.20), new viral capsids are wrapped in membrane as they bud from the cell. In other words, the viral envelope is usually derived from the host cell's plasma membrane, although all or most of the molecules of this membrane are specified by viral genes. The enveloped viruses are now free to infect other cells. This replicative cycle does not necessarily kill the host cell, in contrast to the lytic cycles of phages.

Some viruses have envelopes that are not derived from plasma membrane. Herpesviruses, for example, are temporarily cloaked in membrane derived from the nuclear envelope of the host; they then shed this membrane in the cytoplasm and acquire a new envelope made from membrane of the Golgi apparatus. These viruses have a double-stranded DNA genome and replicate within the host cell nucleus, using a combination of viral and cellular enzymes to replicate and transcribe their DNA. In the case of herpesviruses, copies of the viral DNA can remain behind as mini-chromosomes in the nuclei of certain nerve cells. There they remain latent until some sort of physical or emotional stress triggers a new round of active virus production. The infection of other cells by these new viruses causes the blisters characteristic of herpes, such as cold sores or genital sores. Once someone acquires a herpesvirus infection, flare-ups may recur throughout the person's life.

▼ Figure 19.8 The replicative cycle of an enveloped RNA virus. Shown here is a virus with a single-stranded RNA genome that functions as a template for synthesis of mRNA.



Viral Genetic Material

Table 19.1 shows the common classification system for animal viruses, which is based on their genetic material: double- or single-stranded DNA, or double- or single-stranded RNA, with further classification of RNA viruses. Although some phages and most plant viruses are RNA viruses, the broadest variety of RNA genomes is found among the viruses that infect animals. There are three types of single-stranded RNA genomes found in animal viruses (classes IV–VI in Table 19.1). The genome of class IV viruses can directly serve as mRNA and thus can be translated into viral protein immediately after infection. Figure 19.8 shows a virus of class V, in which the RNA genome serves instead as a *template* for mRNA synthesis. The RNA genome is transcribed into complementary RNA strands, which function both as mRNA and as templates for the synthesis of additional copies of genomic RNA. All viruses that use an RNA genome as a template for mRNA transcription require RNA → RNA synthesis. These viruses use a viral enzyme capable of carrying out this process; there are no such enzymes in most cells. The enzyme used in this process is encoded by the viral genome. After the protein is synthesized, it is packaged during viral self-assembly with the genome inside the viral capsid.

Table 19.1 Classes of Animal Viruses

Class/Family	Envelope?	Examples That Cause Human Diseases
I. Double-Stranded DNA (dsDNA)		
Adenovirus (see Figure 19.3b)	No	Respiratory viruses
Papillomavirus	No	Warts, cervical cancer
Polyomavirus	No	Tumors
Herpesvirus	Yes	Herpes simplex I and II (cold sores, genital sores); varicella zoster (shingles, chicken pox); Epstein-Barr virus (mononucleosis, Burkitt's lymphoma)
Poxvirus	Yes	Smallpox virus; cowpox virus
II. Single-Stranded DNA (ssDNA)		
Parvovirus	No	B19 parvovirus (mild rash)
III. Double-Stranded RNA (dsRNA)		
Reovirus	No	Rotavirus (diarrhea); Colorado tick fever virus
IV. Single-Stranded RNA (ssRNA); Serves as mRNA		
Picornavirus	No	Rhinovirus (common cold); poliovirus; hepatitis A virus; other intestinal viruses
Coronavirus	Yes	Severe acute respiratory syndrome (SARS); Middle East respiratory syndrome (MERS)
Flavivirus	Yes	Zika virus (see Figure 19.10c); yellow fever virus; dengue virus; West Nile virus; hepatitis C virus
Togavirus	Yes	Chikungunya virus (see Figure 19.10b); rubella virus; equine encephalitis viruses
V. ssRNA; Serves as Template for mRNA Synthesis		
Filovirus	Yes	Ebola virus (hemorrhagic fever; see Figure 19.10a)
Orthomyxovirus	Yes	Influenza virus (see Figure 19.3c)
Paramyxovirus	Yes	Measles virus; mumps virus
Rhabdovirus	Yes	Rabies virus
VI. ssRNA; Serves as Template for DNA Synthesis		
Retrovirus	Yes	Human immunodeficiency virus (HIV/AIDS; see Figure 19.9); human T-lymphotropic virus type 1 (HTLV-1) (leukemia)

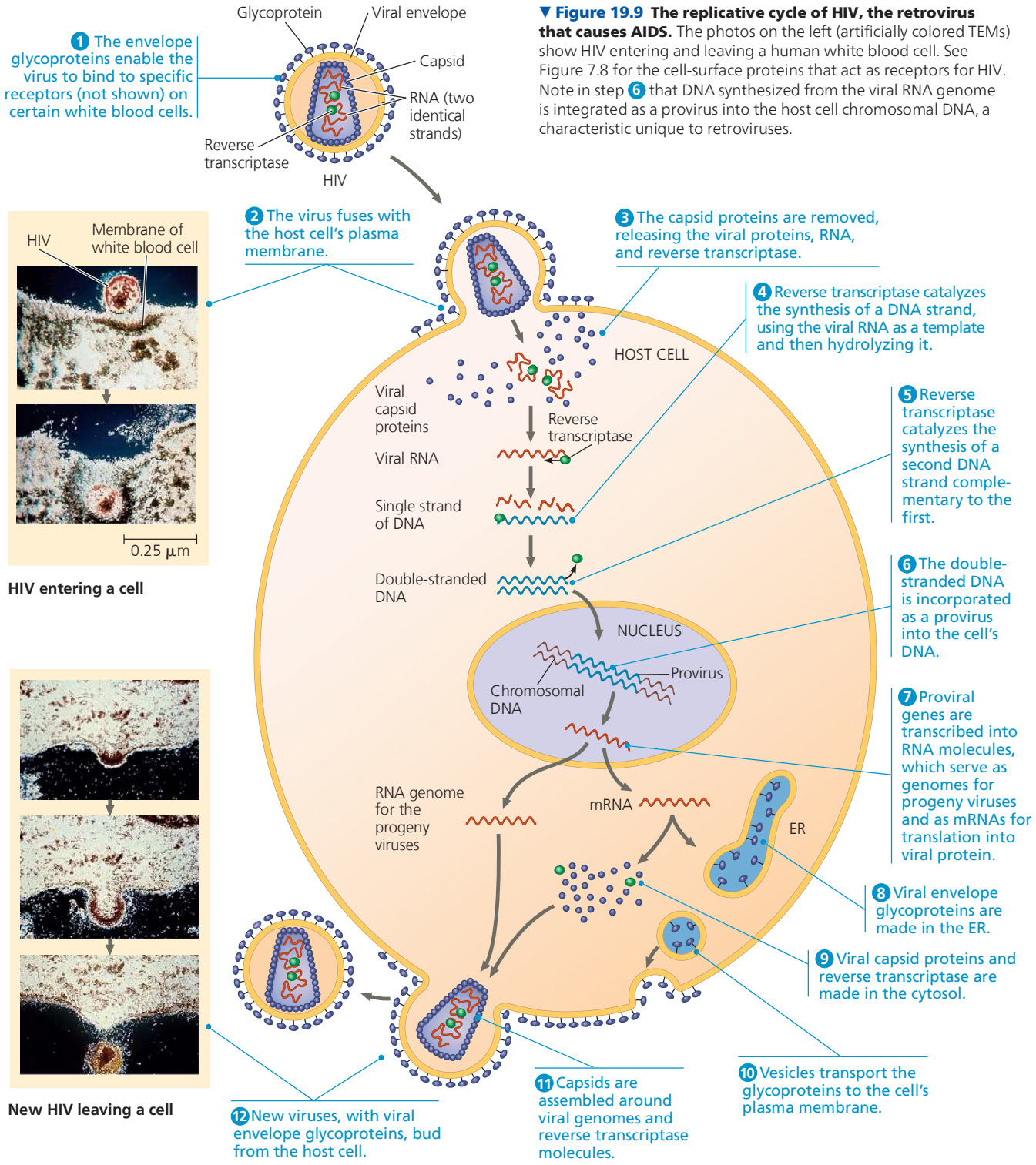
The RNA animal viruses with the most complicated replicative cycles are the **retroviruses** (class VI). These viruses have an enzyme called **reverse transcriptase** that transcribes an RNA template into a DNA copy, an RNA → DNA information flow that is the opposite of the usual direction. This unusual phenomenon is the source of the name retroviruses (*retro* means "backward"). Of particular medical importance is **HIV (human immunodeficiency virus)**, the retrovirus shown in Figure 19.1 that causes **AIDS (acquired immunodeficiency syndrome)**. HIV and other retroviruses are enveloped viruses that contain two identical molecules of single-stranded RNA and two molecules of reverse transcriptase.

The HIV replicative cycle (illustrated in **Figure 19.9**) is typical of a retrovirus. After HIV enters a host cell, its reverse transcriptase molecules are released into the cytoplasm, where they catalyze synthesis of viral DNA. The newly made viral DNA then enters the cell's nucleus and integrates into the DNA of a chromosome. The integrated viral DNA, called a **provirus**, never leaves the host's genome, remaining a permanent resident of the cell. (Recall that a prophage, in contrast, leaves the host's genome at the start of a lytic cycle.) The RNA polymerase of the host transcribes the proviral DNA into RNA molecules, which can function both as mRNA for the synthesis of viral proteins and as genomes for the new viruses that will be assembled and released from the cell. In Concept 43.4, we describe how HIV causes the deterioration of the immune system that occurs in AIDS.

Evolution of Viruses

EVOLUTION We began this chapter by asking whether or not viruses are alive. Viruses do not really fit our definition of living organisms. An isolated virus is biologically inert, unable to replicate its genes or regenerate its own ATP. Yet it has a genetic program written in the universal language of life. Do we think of viruses as nature's most complex associations of molecules or as the simplest forms of life? Either way, we must bend our usual definitions. Although viruses cannot replicate or carry out metabolic activities independently, their use of the genetic code makes it hard to deny their evolutionary connection to the living world.

How did viruses originate? Viruses have been found that infect every form of life—not only bacteria, animals, and plants,



▼ Figure 19.9 The replicative cycle of HIV, the retrovirus that causes AIDS. The photos on the left (artificially colored TEMs) show HIV entering and leaving a human white blood cell. See Figure 7.8 for the cell-surface proteins that act as receptors for HIV. Note in step 6 that DNA synthesized from the viral RNA genome is integrated as a provirus into the host cell chromosomal DNA, a characteristic unique to retroviruses.

MAKE CONECTIONS *Describe what is known about binding of HIV to immune system cells. (See Figure 7.8.) How was this discovered?*

Mastering Biology **HHMI Animation: HIV Replicative Cycle**
Animation: Retrovirus (HIV) Replicative Cycle

but also archaea, fungi, and algae and other protists. Researchers estimate that each milliliter (one-fifth of a teaspoon) of seawater contains between one and 100 million viruses, ten times the number of microorganisms! Because viruses depend on cells for their own propagation, it seems likely that viruses are not the descendants of precellular forms of life but evolved—possibly multiple times—*after* the first cells appeared. Most molecular biologists favor the hypothesis that viruses originated from naked bits of cellular nucleic acids that moved from one cell to another, perhaps via injured cell surfaces. The evolution of genes coding for capsid proteins may have allowed viruses to bind cell membranes, thus facilitating the infection of uninjured cells.

Candidates for the original sources of viral genomes include plasmids and transposons. *Plasmids* are small, circular DNA molecules found in bacteria and in the unicellular fungi called yeasts. Plasmids exist apart from the genome, can replicate independently of the genome, and are occasionally transferred between cells. *Transposons* are DNA segments that can move from one location to another within a cell's genome. Thus, plasmids, transposons, and viruses all share an important feature: They are *mobile genetic elements*. (We'll discuss plasmids in more detail in Concepts 20.1 and 27.2 and transposons in Concept 21.4.)

Consistent with this notion of pieces of DNA shuttling from cell to cell is the observation that a viral genome can have more in common with the genome of its host than with the genomes of viruses that infect other hosts. Indeed, some viral genes are essentially identical to genes of the host.

Debate about the origin of viruses was reinvigorated about 20 years ago when an extremely large virus was discovered: Mimivirus is a double-stranded DNA (dsDNA) virus with an icosahedral capsid that is 400 nm in diameter, the size of a small bacterium. Its genome contains 1.2 million bases (Mb)—about 100 times as many as the influenza virus genome—and an estimated 1,000 genes. Perhaps the most surprising aspect of mimivirus, however, was that its genome included genes previously found only in cellular genomes. Some of these genes code for proteins involved in translation, DNA repair, protein folding, and polysaccharide synthesis. Whether mimivirus evolved *before* the first cells and then developed an exploitative relationship with them or evolved more recently and simply scavenged genes from its hosts is not yet settled.

In the past decade, several even larger viruses have been discovered that cannot be classified with any existing known virus. One such virus is 1 µm (1,000 nm) in diameter, with a dsDNA genome of around 2–2.5 Mb, larger than that of some small eukaryotes. What's more, over 90% of its 2,000 or so genes are unrelated to cellular genes, inspiring the name it was given, pandoravirus. The number of genes in pandoraviruses varies from 1,500 to 2,500 genes. A second virus, called *Pithovirus sibericum*, with a diameter of 1.5 µm and 500 genes, was discovered in permanently frozen soil in Siberia. This virus, once thawed, was able to infect an amoeba after being frozen for 30,000 years! How these and all other viruses fit in the tree of life is an intriguing, unresolved question.

The ongoing evolutionary relationship between viruses and the genomes of their host cells is an association that continues to make viruses very useful experimental systems in molecular biology. Knowledge about viruses also allows many practical applications, since viruses have a tremendous impact on all organisms through their ability to cause disease.

CONCEPT CHECK 19.2

1. Compare the effect on the host cell of a lytic (virulent) phage and a lysogenic (temperate) phage.
2. **MAKE CONNECTIONS** Compare the CRISPR-Cas system with the miRNA system discussed in Concept 18.3, including their mechanisms and their functions.
3. **MAKE CONNECTIONS** The RNA virus in Figure 19.8 has a viral RNA polymerase that functions in step 3 of the virus's replicative cycle. Compare this with a cellular RNA polymerase in terms of template and overall function (see Figure 17.10).
4. Why is HIV called a retrovirus?
5. **VISUAL SKILLS** Looking at Figure 19.9, imagine you are a researcher trying to combat HIV infection. What molecular processes could you attempt to block?

For suggested answers, see Appendix A.

CONCEPT 19.3

Viruses and prions are formidable pathogens in animals and plants

Diseases caused by viral infections afflict humans, agricultural crops, and livestock worldwide. Other smaller, less complex entities known as prions also cause disease in animals. We'll first consider animal viruses.

Viral Diseases in Animals

A viral infection can produce symptoms by a number of different mechanisms. Viruses may damage or kill cells by causing the release of hydrolytic enzymes from lysosomes. Some viruses cause infected cells to produce toxins that lead to disease symptoms, and some have molecular components that are toxic, such as envelope proteins. How much damage a virus causes depends partly on the ability of the infected tissue to regenerate by cell division. People usually recover completely from colds because the epithelium of the respiratory tract, which the viruses infect, can efficiently repair itself. In contrast, damage inflicted by poliovirus to mature nerve cells is permanent because these cells do not divide and usually cannot be replaced. Many of the temporary symptoms associated with viral infections, such as fever and body aches, actually result from the body's own efforts to defend itself against infection rather than from cell death caused by the virus.

The immune system is a critical part of the body's natural defenses (see Chapter 43). It is also the basis for the major medical tool used to prevent viral infections—vaccines. A **vaccine** is a harmless derivative of a pathogen that stimulates the immune system to mount defenses against the harmful pathogen. Smallpox, a viral disease that was once a devastating scourge in

many parts of the world, was eradicated by 1980 due to a vaccination program carried out by the World Health Organization (WHO). The very narrow host range of the smallpox virus—it infects only humans—was a critical factor in the success of this program. Similar worldwide vaccination campaigns are under way to eradicate polio, the incidence of which has dropped by 99%, and measles. Although an effective vaccine exists for measles, a large outbreak occurred in the Pacific Northwest in 2019, correlated with lower vaccination rates in that region. Effective vaccines are also available to protect against rubella, mumps, hepatitis B, and a number of other viral diseases.

Although vaccines can prevent some viral illnesses, medical care can do little, at present, to cure most viral infections once they occur. The antibiotics that help us recover from bacterial infections are powerless against viruses. Antibiotics kill bacteria by inhibiting enzymes specific to bacteria but have little or no effect on eukaryotic or virally encoded enzymes. However, the few enzymes that are encoded only by viruses have provided targets for other drugs. Most antiviral drugs resemble nucleosides and thus interfere with viral nucleic acid synthesis. One such drug is acyclovir, which impedes herpesvirus replication by inhibiting the viral polymerase that synthesizes viral DNA but not the eukaryotic one. Similarly, azidothymidine (AZT) curbs HIV replication by interfering with the synthesis of DNA by reverse transcriptase. In the past 30 years, much effort has gone into developing drugs to treat HIV. Currently, multidrug treatments, sometimes called "cocktails," are considered to be most effective. Such treatments commonly include a combination of two nucleoside mimics and a protease inhibitor, which interferes with an enzyme required for assembly of the viruses. Multidrug treatments originally involved taking up to 20 pills multiple times per day but now usually consist of a single daily tablet. Another effective treatment involves a drug called maraviroc, which blocks a protein on the surface of human immune cells that helps bind the HIV virus (see Figure 7.8). This drug has also been used successfully to prevent infection in individuals who either have been exposed to, or are at risk of exposure to, HIV.

Mastering Biology BBC Video: Know Your Enemy: Bacteria vs. Viruses • HHMI Animation: Treating HIV Infection with a Reverse Transcriptase Inhibitor

Emerging Viral Diseases

Viruses that suddenly become apparent are often referred to as *emerging viruses*. HIV, the AIDS virus, is a classic example: It appeared in San Francisco in the early 1980s, seemingly out of nowhere, although later studies uncovered a case in the Belgian Congo in 1959. Some other dangerous emerging viruses cause encephalitis, inflammation of the brain. One example is the West Nile virus, which appeared in North America in 1999 and has spread to 49 U.S. states, by 2019 resulting in over 50,000 cases and about 2,000 deaths.

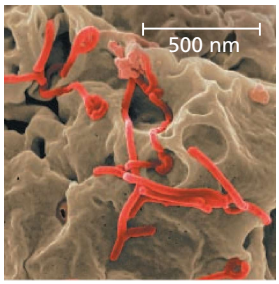
The deadly Ebola virus **(Figure 19.10a)**, recognized initially in 1976 in central Africa, is one of several emerging viruses that cause *hemorrhagic fever*, an often fatal illness characterized by fever, vomiting, massive bleeding, and circulatory system collapse. In 2014, a widespread outbreak, or **epidemic**, of Ebola virus occurred. By 2016, it had resulted in over 11,000 deaths. In 2017, 2018, and 2019, smaller outbreaks occurred in the Democratic Republic of the Congo.

The mosquito-borne chikungunya virus **(Figure 19.10b)** causes an acute illness with fever, rashes, and persistent joint pain. Chikungunya has long been considered a tropical virus, but it has now appeared in Italy, France, and Spain. The Zika virus **(Figure 19.10c)** was first observed in Uganda in 1947, but for decades only a few cases occurred per year. In the spring of 2015, however, it became an emerging virus when it caused a large outbreak in Brazil. Although symptoms of Zika are often mild, the infection of pregnant women was correlated with a striking increase in the number of babies born with abnormally small brains, a condition called microcephaly. Zika is a mosquito-borne flavivirus (like West Nile virus) that infects neural cells, posing a particular danger to fetal brain development.

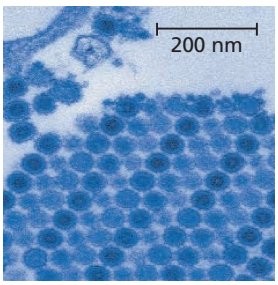
Mastering Biology HHMI Videos: Interview with Katie Walter • Genetically Modified Mosquitoes

Where do these new strains of virus come from? One cause of rapidly emerging viral diseases in humans is mutation of existing viruses into new viruses that can spread more easily. RNA viruses have a high rate of mutation because viral RNA polymerases do not proofread and correct errors in replicating their RNA genomes. Some mutations change existing viruses into new viral strains that can cause disease, even in people immune to the original virus. A well-known related example is how three or four mutations causing changes in a surface protein of a cat virus (feline panleukopenia virus) resulted in the emergence in 1978 of canine parvovirus, a very contagious deadly virus infecting dogs.

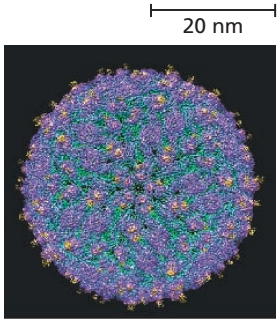
▼ Figure 19.10 Emerging viruses.



(a) Ebola viruses budding from a monkey cell (colorized SEM).



(b) Chikungunya viruses emerging from a cell in the upper left and packing together (colorized TEM).



(c) Computer-generated image of a **Zika virus**, based on a technique called cryo-electron microscopy.

A second cause of the emergence of viral diseases is the spread of a viral disease from a small, isolated human population. HIV, the virus that causes AIDS, went unnamed and virtually unnoticed for decades before spreading around the world. In this case, technological and social factors, including affordable international travel, blood transfusions, unprotected sexual intercourse, and reuse of needles to inject intravenous drugs, allowed a previously rare human disease to become a global scourge.

Mastering Biology
Interview with David Satcher: The role of the CDC in recognizing AIDS and in public health



A third cause of new viral diseases in humans is the spread of existing viruses from other animals. Scientists estimate that about three-quarters of new human diseases originate in this way. Animals infected with a virus that can be transmitted to humans are said to be a natural reservoir for that virus. HIV is an example, as scientists believe that it originated from a version of the virus found in chimpanzees in central Africa, after people ate chimpanzee meat for food and were infected by exposure to chimpanzee blood.

In general, flu epidemics provide an instructive example of these three causes of emerging viruses. There are three types of influenza virus: types B and C, which infect only humans and have never caused an epidemic, and type A, which infects a wide range of animals, including birds, pigs, horses, and humans. The influenza type A viruses present in pigs and wild and domestic birds are potential emerging viruses that represent a long-term threat to human health.

A case in point is the H5N1 strain of avian influenza virus, which is highly contagious and deadly in birds. The first transmission from birds to humans was documented in Hong Kong in 1997. Since then, about 700 people have been infected, with an alarming mortality rate of greater than 50%. The high mortality rate is partly because H5N1 is very different from strains of influenza that have circulated among humans for a long time. Individuals are therefore not able to mount a strong immune response against viruses like H5N1.

Different strains of influenza A are given standardized names; for example, the name H5N1 identifies which forms of two viral surface proteins are present—hemagglutinin and neuraminidase (HA and NA, respectively; see the glycoprotein spikes in Figure 19.3c). These two proteins together help determine the host range and severity of disease caused by each virus. As of 2017, 18 types of hemagglutinin, a protein that helps the flu virus attach to host cells, and 11 types of neuraminidase, an enzyme that helps release new virus particles from infected cells, have been identified. All possible combinations of HA and NA have been found in some waterbirds. In 2018, researchers using advanced microscopy techniques reported that the HA protein is quite dynamic, stretching toward the target host cell, retracting, then refolding to a new shape and re-approaching the target cell.

Although deadly, the H5N1 strain has not yet caused an epidemic because nearly all cases have been the result of transmission from birds to people, rather than person-to-person. Epidemics occur when genetic changes allow a new viral strain to be easily transmitted between humans. An event like this occurred in 2009, when a strain of influenza virus (H1N1) appeared that was very different from the virus that causes the seasonal flu. The H1N1 influenza virus spread rapidly, prompting WHO to declare a global epidemic, or **pandemic**. Within half a year, the disease had reached 207 countries, infecting over 600,000 people and killing almost 8,000. In addition to the 2009 H1N1 pandemic, influenza A strains have caused three other major flu epidemics among humans in the last 100 years. The most notable of these was the "Spanish flu" pandemic of 1918–1919, which killed 40–50 million people worldwide. In the **Scientific Skills Exercise**, you'll analyze genetic changes in variants of the 2009 H1N1 influenza virus and correlate them with spread of the disease.

The disease caused by H1N1 was originally called "swine flu" because parts of the viral genome were very similar to strains of influenza in pigs. However, studies revealed that the virus was not transmitted from pigs to humans. Instead, the story was more complex: H1N1 was a unique combination of swine, avian, and human influenza genes that allowed it to spread among humans.

Influenza viruses can change quickly because they have a genome made up of nine segments of RNA rather than a single RNA molecule. When an animal like a pig or a bird is infected with multiple strains of influenza virus, the RNA molecules making up the viral genomes can mix and match (*reassort*) during viral assembly, resulting in new genetic combinations. If a flu virus from pigs recombines with viruses that circulate widely among humans, it may acquire the ability to spread easily from person to person, dramatically increasing the potential for a major human outbreak. Pigs are believed to have been the main hosts for recombination that led to the 2009 H1N1 flu virus.

Influenza viruses also have a high rate of mutation, for reasons mentioned earlier. Coupled with reassortments, these mutations can lead to the emergence of a viral strain from animals that can infect human cells. Scientists are worried that the H5N1 strain will evolve in a way that enables it to spread as easily as the H1N1 strain. This would represent a major global health threat like that of the 1918 pandemic.

How easily could this capability evolve in the H5N1 strain? In 2011, scientists working with ferrets, small mammals used as model organisms for the human flu, found that only a few mutations of the avian flu virus were sufficient to allow infection of cells in the human nasal cavity and windpipe. Furthermore, when the scientists transferred nasal swabs serially from ferret to ferret, the virus became transmissible through the air. Reports of this startling discovery ignited a firestorm of debate about whether to publish the results. Ultimately, the scientific community decided the benefits of preventing pandemics outweighed the risks of the information being used for harmful purposes, and the work was published in 2012.

Normal seasonal flu viruses (including influenza types A and B) are not considered emerging viruses because variations

Scientific Skills Exercise

Analyzing a Sequence-Based Phylogenetic Tree to Understand Viral Evolution

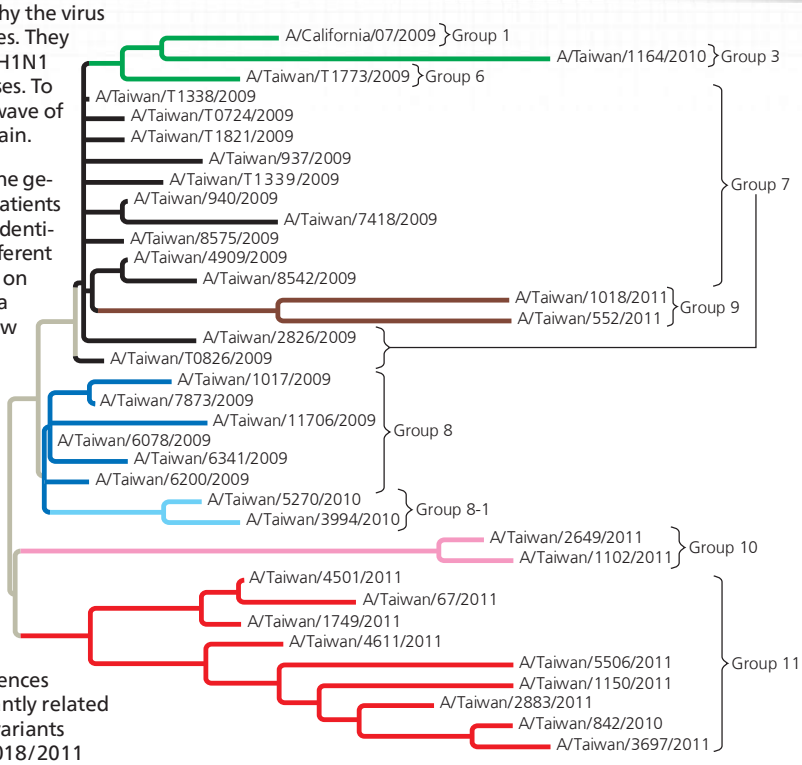


▶ **H1N1 flu vaccination**

How Can Sequence Data Be Used to Track Flu Virus Evolution? In 2009, an influenza A H1N1 virus caused a pandemic, and the virus has continued to resurface in outbreaks across the world. Researchers in Taiwan were curious about why the virus kept appearing despite widespread flu vaccine initiatives. They hypothesized that newly evolved variant strains of the H1N1 virus were able to evade human immune system defenses. To test this hypothesis, they needed to determine if each wave of flu infection was caused by a different H1N1 variant strain.

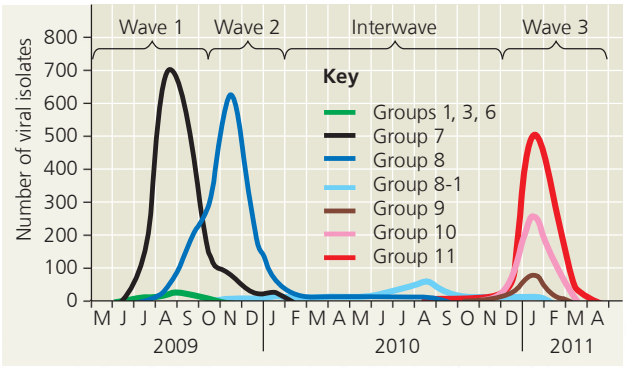
How the Experiment Was Done Scientists obtained the genome sequences for 4,703 virus isolates collected from patients with H1N1 flu in Taiwan, each named by type/location/identifying number/year. They compared the sequences in different strains for the viral hemagglutinin (HA) gene, and based on mutations that had occurred, arranged the isolates into a phylogenetic tree (see Figure 26.5 for information on how to read phylogenetic trees).

Data from the Experiment In the phylogenetic tree, each branch tip is one variant strain of the H1N1 virus with a unique HA gene sequence. The tree is a way to visualize a working hypothesis about the evolutionary relationships between H1N1 variants.



INTERPRET THE DATA

1. The more closely connected two variants are in the tree, the more alike they are in terms of HA gene sequence. Each fork in a branch, called a node, shows where two lineages separate due to different accumulated mutations. The length of the branches is a measure of how many sequence differences there are between the variants, indicating how distantly related they are. Referring to the phylogenetic tree, which variants are more closely related to each other: A/Taiwan/1018/2011 and A/Taiwan/552/2011 or A/Taiwan/1018/2011 and A/Taiwan/8542/2009? Explain your answer.
2. The scientists arranged the branches into groups made up of one ancestral variant and all of its descendant, mutated variants. They are color-coded in the figure. Using group 11 as an example, trace the lineage of its variants. **(a)** Do all of the nodes have the same number of branches or branch tips? **(b)** Are all of the branches in the group the same length? **(c)** What do these results indicate?
3. The graph shows the number of isolates collected (each from an ill patient) on the *y*-axis and the month and year that the isolates were collected on the *x*-axis. Each group of variants is plotted separately with a line color that matches the tree diagram. **(a)** Which group of variants was the earliest to cause the first wave of H1N1 flu in over 100 patients in Taiwan? **(b)** After a group of variants had a peak number of infections, did members of that same group cause another (later) wave of infection? **(c)** One variant in group 1 (green, uppermost branch) was used to make a vaccine that was distributed very early in the pandemic. Based on the graphed data, does it look like the vaccine was effective?
4. Groups 9, 10, and 11 all had H1N1 variants that caused a large number of infections at the same time in Taiwan. Does this mean that the scientists' hypothesis, that new variants cause new waves of infection, was incorrect? Explain your answer.



▲ Scientists graphed the number of isolates by the month and year of isolate collection to show the period in which each viral variant was actively causing illness in people.

Data from J.-R. Yang et al., New variants and age shift to high fatality groups contribute to severe successive waves in the 2009 influenza pandemic in Taiwan, *PLoS ONE* 6(11): e28288 (2011).

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.

of seasonal flu viruses have been circulating among humans for long enough that most components are recognized by the immune system. However, these viruses still undergo mutation and reassortment of genome segments, and variations of the HA protein are used each year to generate vaccines against the strains predicted most likely to occur the following year.

As we have seen, emerging viruses are usually existing viruses that mutate, spread more widely in the current host species, or spread to new host species. Changes in host behavior or environmental changes can increase the viral traffic responsible for emerging diseases. For instance, new roads built through remote areas can allow viruses to spread between previously isolated human populations. Also, the destruction of forests to expand cropland can bring humans into contact with animals that host infectious viruses. Finally, genetic mutations and changes in host ranges can allow viruses to jump between species. Many viruses are transmitted by mosquitoes. A dramatic expansion of the disease caused by the chikungunya virus occurred in the mid-2000s when a mutation allowed it to infect not only the mosquito species *Aedes aegypti* but also the related *Aedes albopictus*. Insecticides and mosquito netting over beds are crucial tools in public health attempts to prevent diseases carried by mosquitoes **(Figure 19.11)**.

▼ Figure 19.11 Netting as protection against virus-carrying mosquitoes.



Recently, scientists have become concerned about the possible effects of climate change on worldwide viral transmission. Dengue fever, also mosquito-borne, has appeared in Florida and Portugal, regions where it had not been seen before. The possibility that global climate change has allowed mosquito species carrying these viruses to expand their ranges and interact more is troubling because of the increased chance of a mutation allowing a virus to jump to a new host. This is an area of active research by scientists applying climate change models to what is known about the habitat requirements of mosquito species.

Viral Diseases in Plants

More than 2,000 types of viral diseases of plants are known, accounting for an annual loss of over $30 billion worldwide due to destruction of crops. Common signs of viral infection include bleached or brown spots on leaves and fruits **(Figure 19.12)**, stunted growth, and damaged flowers or roots, all of which can diminish the yield and quality of crops.

► Figure 19.12 Immature tomato infected by a virus.



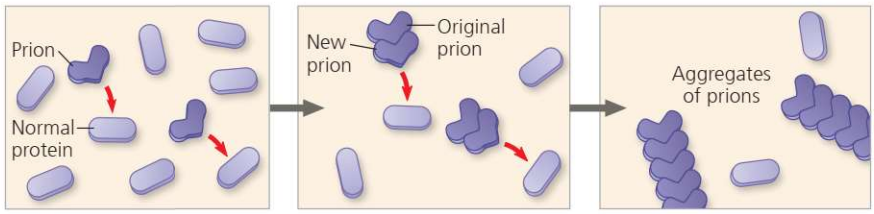
Plant viruses have the same basic structure and mode of replication as animal viruses. Most known plant viruses, including tobacco mosaic virus (TMV), have an RNA genome. Many have a helical capsid, like TMV, while others have an icosahedral capsid (see Figure 19.3b).

Viral diseases of plants spread by two major routes. In the first route, *horizontal transmission*, an external source infects the plant. Because the invading virus must get past the plant's outer protective layer of cells (the epidermis), a plant becomes more susceptible to viral infections if it has been damaged by wind, injury, or herbivores. Herbivores, especially insects, pose a double threat because they can also carry viruses, transmitting disease from plant to plant. Moreover, gardeners may transmit plant viruses inadvertently on pruning shears and other tools. The other route of viral infection is *vertical transmission*, in which a plant inherits a viral infection from a parent. Vertical transmission can occur in asexual propagation (for example, through cuttings) or in sexual reproduction via infected seeds.

Once a virus enters a plant cell and begins replicating, viral genomes and associated proteins can spread throughout the plant through plasmodesmata, the cytoplasmic connections that penetrate the walls between adjacent plant cells (see Figure 36.19). The passage of viral macromolecules from cell to cell is facilitated by virally encoded proteins that cause enlargement of plasmodesmata. Scientists have not yet devised cures for most viral plant diseases, so research efforts are focused largely on reducing disease transmission and on breeding resistant varieties of crop plants.

Prions: Proteins as Infectious Agents

The viruses discussed in this chapter are infectious agents that spread diseases, and their genetic material is composed of nucleic acids, whose ability to be replicated is well known. Surprisingly, there are also *proteins* that are infectious. Proteins called **prions** appear to cause degenerative brain diseases in various animal species. These diseases include scrapie in sheep; mad cow disease, which plagued the European beef industry about 20 years ago; and Creutzfeldt-Jakob disease in humans, which has caused the death of some 175 people in the United Kingdom since 1996. Prions can be transmitted in food, as may occur when people eat beef from cattle with mad cow disease. Kuru, another human disease caused by prions, was identified in the early 1900s among the South Fore indigenous people of New Guinea. When a kuru epidemic peaked there in the 1960s, scientists at first thought the disease had a genetic basis because family members also often contracted the disease. Eventually, however, investigations revealed a



▶ Figure 19.13 Model for how prions propagate. Prions are misfolded versions of normal brain proteins. When a prion contacts a normally folded version of the same protein, it may induce the normal protein to assume the abnormal shape. The resulting chain reaction may continue until high levels of prion aggregation cause cellular malfunction and eventual degeneration of the brain.

Mastering Biology **Animation: Prions: Characteristics • Animation: Prions: Diseases**

different story: After a death, family members practiced ritual cannibalism, eating organs of the deceased, and prions were transmitted primarily in brain tissue. Women got kuru more often than men because men ate the more "prestigious" organs, like the heart, while women and children ate the brains.

Two characteristics of prions are especially alarming. First, prions act very slowly, with an incubation period of at least ten years before symptoms develop. The lengthy incubation period prevents sources of infection from being identified until long after the first cases appear, allowing many more infections to occur. Second, prions are not destroyed or deactivated by heating to normal cooking temperatures. To date, there is no known cure for prion diseases, and the only hope for developing effective treatments lies in understanding the process of infection.

How can a protein, which cannot replicate itself, be a transmissible pathogen? According to the leading model, a prion is a misfolded form of a protein normally present in brain cells. When the prion gets into a cell containing the normal form of the protein, the prion somehow converts normal protein molecules to the misfolded prion versions. Several prions then aggregate into a complex that can convert other normal proteins to prions, which join the chain **(Figure 19.13)**. Prion aggregation interferes with normal cellular functions and causes disease symptoms. This model was greeted with much skepticism when it was first proposed by Stanley Prusiner in the early 1980s, but it is now widely accepted. Prusiner was awarded the Nobel Prize in 1997 for his work on prions. He has also proposed that prions are involved in neurodegenerative diseases such as Alzheimer's and Parkinson's disease. There are many outstanding questions about these small infectious agents.

CONCEPT CHECK 19.3

1. Describe two ways in which a preexisting virus can become an emerging virus.
2. Contrast horizontal and vertical transmission of viruses in plants.
3. **WHAT IF?** TMV has been isolated from virtually all commercial tobacco products. Why, then, is TMV infection not an additional hazard for smokers?

For suggested answers, see Appendix A.

19 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 19.1

A virus consists of a nucleic acid surrounded by a protein coat *(pp. 399–401)*

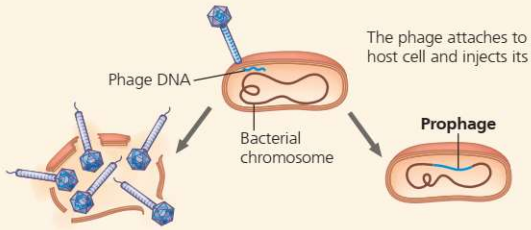
- Researchers discovered viruses in the late 1800s by studying a plant disease, tobacco mosaic disease.
- A **virus** is a small nucleic acid genome enclosed in a protein **capsid** and sometimes a membranous **viral envelope**. The genome may be single- or double-stranded DNA or RNA.

? *Are viruses generally considered living or nonliving? Explain.*

CONCEPT 19.2

Viruses replicate only in host cells *(pp. 401–408)*

- Viruses use enzymes, ribosomes, and small molecules of host cells to synthesize progeny viruses during replication.
- Each type of virus has a characteristic **host range**, affected by whether cell-surface proteins are present that viral surface proteins can bind to.
- **Phages** (viruses that infect bacteria) can replicate by two alternative mechanisms: the **lytic cycle** and the **lysogenic cycle**.



Lytic cycle
- **Virulent** or **temperate phage**
- Destruction of host DNA
- Production of new phages
- Lysis of host cell causes release of progeny phages

Lysogenic cycle
- **Temperate phage** only
- Genome integrates into bacterial chromosome as **prophage**, which (1) is replicated and passed on to daughter cells and (2) can be induced to leave the chromosome and initiate a lytic cycle

- Bacteria have various ways of defending themselves against phage infections, including the CRISPR-Cas system.
- Many animal viruses have an envelope. **Retroviruses** (such as **HIV**) use the enzyme **reverse transcriptase** to copy their RNA genome into DNA, which can be integrated into the host genome as a **provirus**.
- Since viruses can replicate only within cells, they probably evolved after the first cells appeared, perhaps as packaged fragments of cellular nucleic acid.

? *Describe enzymes that are not found in most cells but are necessary for the replication of certain types of viruses.*

CONCEPT 19.3

Viruses and prions are formidable pathogens in animals and plants *(pp. 408–413)*

- Symptoms of viral diseases may be caused by direct viral harm to cells or by the body's immune response. **Vaccines** stimulate the immune system to defend the host against specific viruses.
- An **epidemic**, a widespread outbreak of a disease, can become a **pandemic**, a global epidemic.
- Outbreaks of emerging viral diseases in humans are usually not new, but rather are caused by existing viruses that expand their host territory. The H1N1 2009 flu virus was a new combination of pig, human, and avian viral genes that caused a pandemic. The H5N1 avian flu virus has the potential to cause a high-mortality flu pandemic.
- Viruses enter plant cells through damaged cell walls (horizontal transmission) or are inherited from a parent (vertical transmission).
- **Prions** are slow-acting, virtually indestructible infectious proteins that cause brain diseases in mammals.

? *What aspect of an RNA virus makes it more likely than a DNA virus to become an emerging virus?*

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

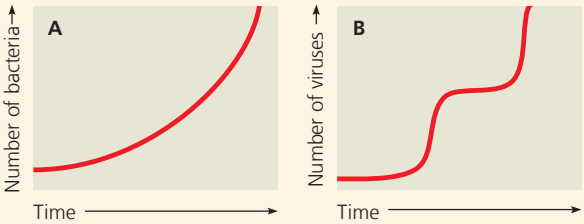
1. Which of the following characteristics, structures, or processes is common to both bacteria and viruses?
   (A) metabolism
   (B) ribosomes
   (C) genetic material composed of nucleic acid
   (D) cell division
2. Emerging viruses arise by
   (A) mutation of existing viruses.
   (B) the spread of existing viruses to new host species.
   (C) the spread of existing viruses more widely within their host species.
   (D) all of the above.
3. To cause a human pandemic, the H5N1 avian flu virus would have to
   (A) spread to primates such as chimpanzees.
   (B) develop into a virus with a different host range.
   (C) become capable of human-to-human transmission.
   (D) become much more pathogenic.

Levels 3-4: Applying/Analyzing

4. A bacterium is infected with an experimentally constructed bacteriophage composed of the T2 phage protein coat and T4 phage DNA. The new phages produced would have
   (A) T2 protein and T4 DNA. (C) T2 protein and T2 DNA.
   (B) T4 protein and T2 DNA. (D) T4 protein and T4 DNA.
5. RNA viruses require their own supply of certain enzymes because
   (A) host cells rapidly destroy the viruses.
   (B) host cells lack enzymes that can replicate the viral genome.
   (C) these enzymes translate viral mRNA into proteins.
   (D) these enzymes penetrate host cell membranes.
6. **DRAW IT** Redraw Figure 19.8 to show the replicative cycle of a virus with a single-stranded genome that can function as mRNA (a class IV virus).

Levels 5-6: Evaluating/Creating

7. **EVOLUTION CONNECTION** The success of some viruses lies in their ability to evolve rapidly within the host. Such viruses evade the host's defenses by mutating and producing many altered progeny viruses before the body can mount an attack. Thus, the viruses present late in infection differ from those that initially infected the body. Discuss this as an example of evolution in microcosm. Which viral lineages tend to predominate?
8. **SCIENTIFIC INQUIRY** When bacteria infect an animal, the number of bacteria in the body increases in an exponential fashion (graph A). After infection by a virulent animal virus with a lytic replicative cycle, there is no evidence of infection for a while. Later, however, the number of viruses rises suddenly and subsequently increases in a series of steps (graph B). Explain the difference in the curves.



9. **WRITE ABOUT A THEME: ORGANIZATION** While viruses are considered by most scientists to be nonliving, they do show some characteristics of life, including the correlation of structure and function. In a short essay (100–150 words), discuss how the structure of a virus correlates with its function.
10. **SYNTHESIZE YOUR KNOWLEDGE**



Oseltamivir (Tamiflu), an antiviral drug prescribed for the flu, inhibits the enzyme neuraminidase. Explain how this drug could prevent infection in someone exposed to the flu or could shorten the course of flu in an infected patient (the reasons for which it is prescribed).

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
How can genomics help track the Ebola virus?
Go to "Ebola Outbreak Traced to the Funeral of a Traditional Healer" at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.

20 DNA Tools and Biotechnology

KEY CONCEPTS

20.1 DNA sequencing and DNA cloning are valuable tools for genetic engineering and biological inquiry *p. 416*

20.2 Biologists use DNA technology to study gene expression and function *p. 423*

20.3 Cloned organisms and stem cells are useful for basic research and other applications *p. 428*

20.4 The practical applications of DNA-based biotechnology affect our lives in many ways *p. 433*

Study Tip

Apply what you've learned: Write down a gene-related question you've wondered about. Then make a table listing techniques you could use to investigate your question and how you would apply them. Below is an example.

Is autism caused by genes, the environment, or both?	
Technique	Application to question
DNA sequencing	-Compare DNA sequences from people with autism with sequences from nonaffected people.

Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 20
- HHMI Video: Genetic Engineering
- Animation: Creating Recombinant DNA
- HHMI Video: Polymerase Chain Reaction (PCR)

For Instructors to Assign (in Item Library)
- Tutorial: Restriction Enzymes, Recombinant DNA, and Gene Cloning
- Activity: Cloning a Gene in Bacteria

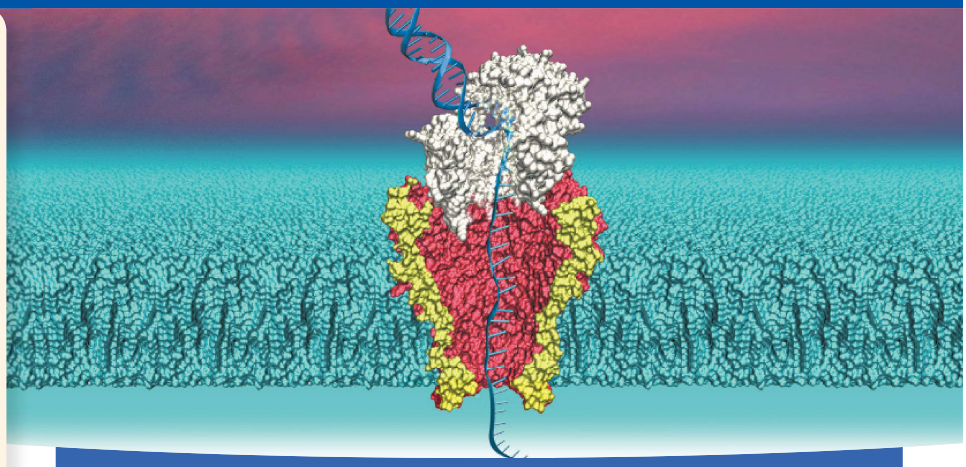
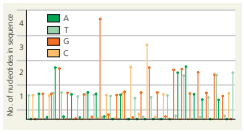


Figure 20.1 This model shows a technique in which a DNA strand is passed through a small pore in a membrane. The resulting changes in an electrical current are used to determine the nucleotide sequence. The first human genome sequence, completed in 2003, took 13 years and cost $1 billion; the time and cost of genome sequencing have been greatly reduced by improved methods like the one shown here.

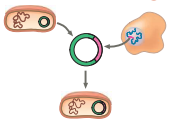
What are the main techniques and applications of biotechnology?

TECHNIQUES

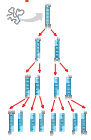
DNA sequencing



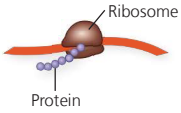
Gene cloning



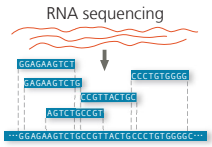
PCR amplification



Expressing genes



Analyzing gene expression



Gene editing



APPLICATIONS

Agriculture



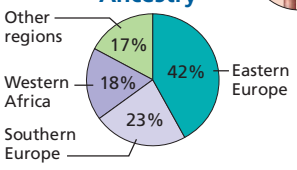
Medicine



Environmental cleanup



Ancestry



Forensics



CONCEPT 20.1

DNA sequencing and DNA cloning are valuable tools for genetic engineering and biological inquiry

The discovery of the structure of the DNA molecule, and specifically the recognition that its two strands are complementary to each other, opened the door for the development of DNA sequencing and other techniques for manipulating DNA—known as **DNA technology—**used in biological research today. Key to these techniques is **nucleic acid hybridization**, the base pairing of one strand of a nucleic acid to a complementary sequence from another nucleic acid strand, either DNA or RNA. Nucleic acid hybridization forms the foundation of virtually every technique used in **genetic engineering**, the direct manipulation of genes for practical purposes. Genetic engineering has launched a revolution in fields as varied as criminal law, medicine, and basic biological research. In this section, we'll explore several important techniques and their uses.

DNA Sequencing

Researchers can exploit the principle of complementary base pairing to determine the complete nucleotide sequence of a DNA molecule, a process called **DNA sequencing**. The first automated procedure, called dideoxy sequencing, was developed in the 1970s by biochemist Frederick Sanger, who received the Nobel Prize in 1980 for this accomplishment. Dideoxy sequencing is still used for routine small-scale sequencing jobs.

Mastering Biology HHMI Video: Sanger Method of DNA Sequencing

During the first decade of this century, "next-generation sequencing" techniques were developed that are rapid and inexpensive **(Figure 20.2)**. DNA fragments are amplified (copied) to yield an enormous number of identical fragments **(Figure 20.3)**. A single template strand of each fragment is immobilized, and the complementary strand is synthesized, one nucleotide at a time. A chemical technique enables electronic monitors to identify in real time which of the four nucleotides is added; this method is thus called *sequencing by synthesis*. Thousands or hundreds of thousands of fragments, each about 300 nucleotides long, are sequenced in parallel in machines like those shown in Figure 20.2, accounting for the high rate of nucleotides sequenced per hour. This is an example of "high-throughput" DNA technology and is currently the method of choice for studies where massive numbers of

▼ **Figure 20.2** Next-generation DNA sequencing machines.



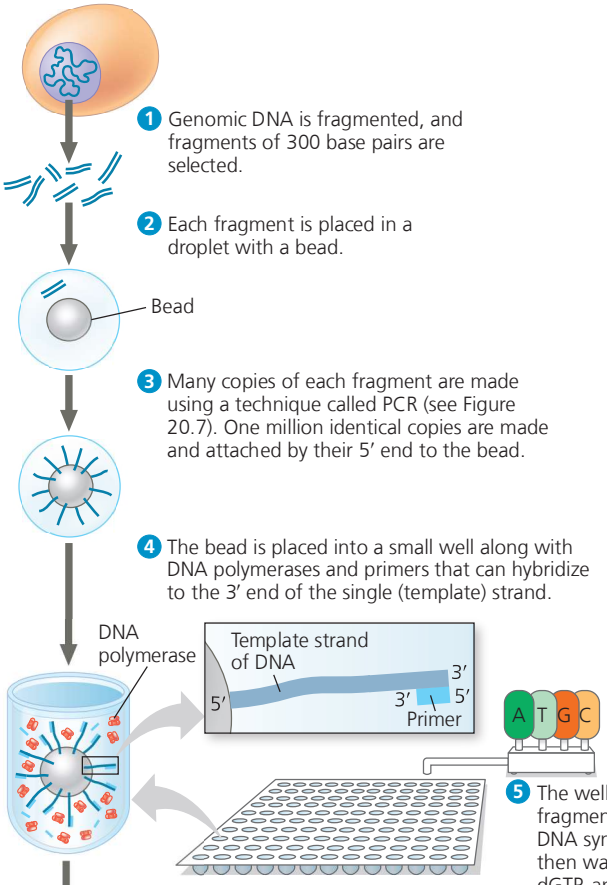
DNA samples—even a set of numerous fragments representing an entire genome—are being sequenced.

More and more often, next-generation sequencing is complemented (or in some cases replaced) by "third-generation sequencing," with each new technique being faster and less expensive than the previous one. In some new methods, the DNA is neither cut into fragments nor amplified. Instead, a single, very long DNA molecule is sequenced on its own. Several groups have developed techniques that move a single strand of a DNA molecule through a very small pore (a *nanopore*) in a membrane, identifying the bases one by one by the distinct way each interrupts an electrical current. One model of this concept is shown in Figure 20.1, in which the pore is a protein channel embedded in a lipid membrane. (Other researchers are using artificial membranes and nanopores.) Each type of base interrupts the electrical current for a slightly different length of time. In 2015, the first nanopore sequencer went on the market; this device is the size of a small candy bar and connects to a computer via a USB port. Associated software allows immediate identification and analysis of the sequence. This is only one of many approaches to further increase the rate and cut the cost of sequencing, while also allowing the methodology to move out of the laboratory and into the field.

Improved DNA-sequencing techniques have transformed the way in which we can explore fundamental biological questions about evolution and how life works (see Make Connections Figure 5.26). Little more than 15 years after the human genome sequence was announced, researchers had completed the sequencing of thousands of genomes, with tens of thousands in progress. Complete genome sequences have been determined for cells from several cancers, for ancient humans, and for the many bacteria that live in the human intestine. In Chapter 21, you'll learn more about how this rapid acceleration of sequencing technology has revolutionized our study of the evolution of species and the genome itself. Now, let's consider how individual genes are studied.

▼ Figure 20.3 Research Method

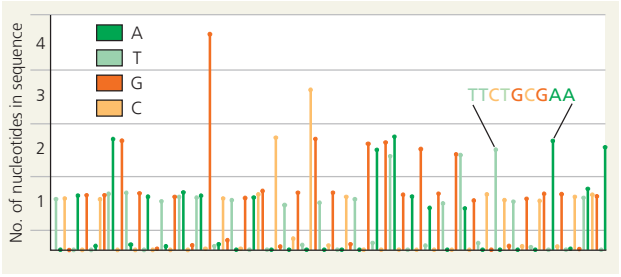
Sequencing by Synthesis: Next-Generation Sequencing



Application In current next-generation sequencing techniques, each fragment is about 300 nucleotides long; by sequencing the fragments in parallel, about 2 billion nucleotides can be sequenced in 24 hours.

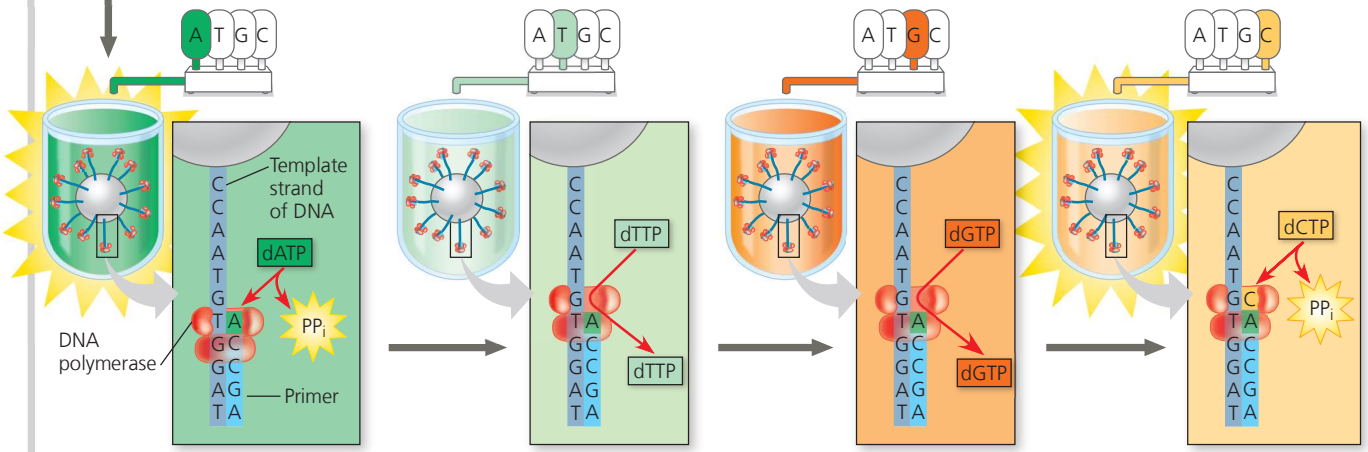
Technique See numbered steps and diagrams.

Results Each of the 2,000,000 wells in the multiwell plate, which holds a different fragment, yields a different sequence. The results for one fragment are shown below as a "flow-gram." The sequences of the entire set of fragments are analyzed using computer software, which "stitches" them together into a whole sequence—here, an entire genome.



INTERPRET THE DATA *If the template strand has two or more identical nucleotides in a row, their complementary nucleotides will be added one after the other in the same flow step. How are two or more of the same nucleotide (in a row) detected in the flow-gram? (See sample on the right.) Write out the sequence of the first 25 nucleotides in the flow-gram above, starting from the left. (Ignore the very short lines.)*

5 The well is one of 2 million on a multiwell plate, each containing a different DNA fragment to be sequenced. A solution of one of the four nucleotides required for DNA synthesis (deoxynucleoside triphosphates, or dNTPs) is added to all wells and then washed off. This is done sequentially for all four nucleotides: dATP, dTTP, dGTP, and then dCTP. The entire process is then repeated over and over again.



6 In each well, if the next base on the template strand (T in this example) is complementary to the added nucleotide (A, here), the nucleotide is joined to the growing strand, releasing PP_i, which causes a flash of light that is recorded.

7 The nucleotide is washed off and a different nucleotide (dTTP, here) is added. If the nucleotide is not complementary to the next template base (G, here), it is not joined to the strand, no reaction occurs, and there is no flash.

8 The process of adding and washing off the four nucleotides is repeated until every fragment has a complete complementary strand. The pattern of flashes reveals the sequence of the original fragment in each well.

Making Multiple Copies of a Gene or Other DNA Segment

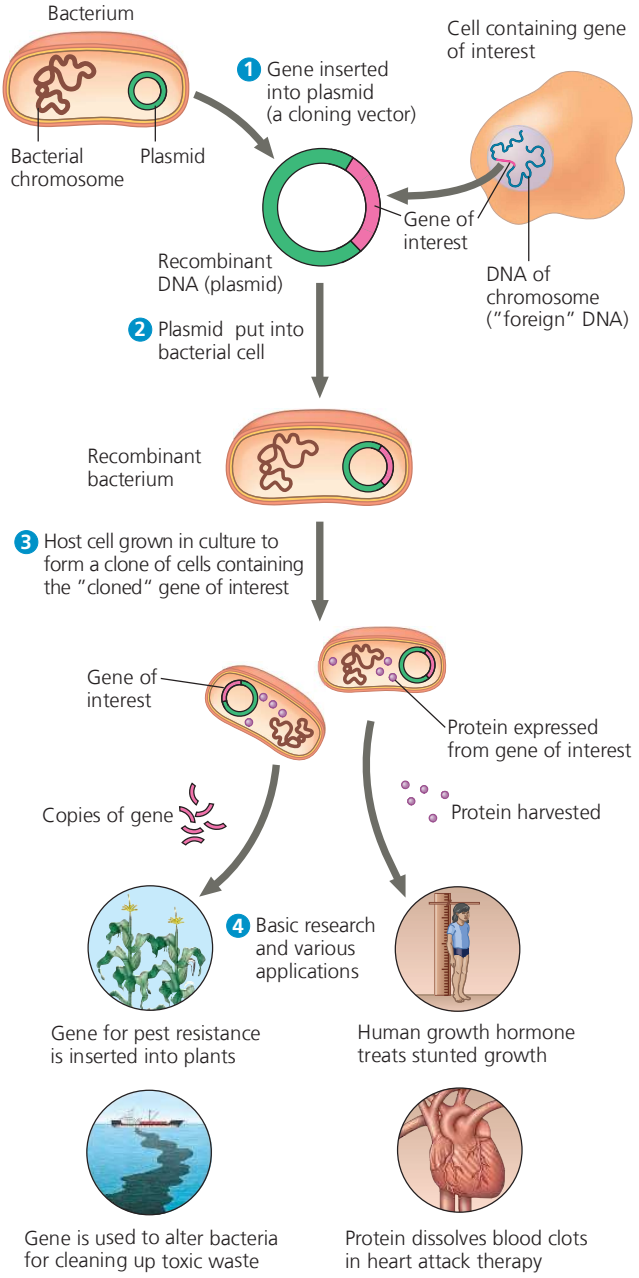
A molecular biologist studying a particular gene or group of genes faces a challenge. Naturally occurring DNA molecules are very long, and a single molecule usually carries hundreds or even thousands of genes. Moreover, in many eukaryotic genomes, protein-coding genes occupy only a small proportion of the chromosomal DNA, the rest being noncoding nucleotide sequences. A single human gene, for example, might constitute only 1/100,000 of a chromosomal DNA molecule. As a further complication, it's not easy to distinguish a gene from the surrounding DNA because they differ only in nucleotide sequence. To study a specific gene, scientists have developed methods to isolate a segment of DNA carrying that gene and make multiple identical copies of it—a process called **DNA cloning**.

Most methods for cloning pieces of DNA in the laboratory share certain general features. One common approach uses bacteria, most often *Escherichia coli*. Recall from Figure 16.13 that the *E. coli* chromosome is a large, circular molecule of DNA. In addition, *E. coli* and many other bacteria also have **plasmids**, small, circular DNA molecules that are replicated separately. A plasmid has only a small number of genes; these genes may be useful when the bacterium is in a particular environment but may not be required for survival or reproduction under most conditions.

To clone pieces of DNA using bacteria, scientists have isolated plasmids from bacterial cells and altered them by genetic engineering. Researchers insert DNA they want to study ("foreign" DNA) into the plasmid **(Figure 20.4)**. The resulting plasmid is now a **recombinant DNA molecule**, a molecule containing DNA from two different sources, very often different species. The plasmid is then returned to a bacterial cell, producing a *recombinant bacterium*. This single cell reproduces through repeated cell divisions to form a clone of cells, a population of genetically identical cells. Because the dividing bacteria replicate the recombinant plasmid and pass it on to their descendants, the foreign DNA and any genes it carries are cloned at the same time. The production of multiple copies of a single gene is a type of DNA cloning called **gene cloning**.

In Figure 20.4, the plasmid acts as a **cloning vector**, a DNA molecule that can carry foreign DNA into a host cell and be replicated there. Bacterial plasmids are widely used as cloning vectors for several reasons: They can be readily obtained from commercial suppliers, manipulated to form recombinant plasmids by insertion of foreign DNA in a test tube (referred to as *in vitro*, from the Latin meaning "in glass"), and then easily introduced into bacterial cells. The foreign DNA in Figure 20.4 is a gene from a eukaryotic cell; we will describe in more detail how the foreign DNA segment was obtained later in this section.

▼ Figure 20.4 Gene cloning and some uses of cloned genes. In this simplified diagram of gene cloning, we start with a plasmid (originally isolated from a bacterial cell) and a gene of interest from another organism. Only one plasmid and one copy of the gene of interest are shown at the top of the figure, but the starting materials would include many of each.



Mastering Biology HHMI Video: Genetic Engineering

Gene cloning is useful for two basic purposes: to make many copies of, or *amplify*, a particular gene and to produce a protein product from it (see Figure 20.4). Researchers can isolate copies of a cloned gene from bacteria for use in basic

research or to endow another organism with a new metabolic capability, such as pest resistance. For example, a resistance gene present in one crop species might be cloned and transferred into plants of another species. (Such organisms are called *genetically modified organisms*, or *GMOs* for short; they will be discussed later in the chapter.) Alternatively, a protein with medical uses, such as human growth hormone, can be harvested in large quantities from cultures of bacteria carrying a cloned gene for the protein. (We'll look at the techniques for expressing cloned genes later.) Since one gene is only a very small part of the total DNA in a cell, the ability to amplify such a DNA fragment is crucial for any application involving a single gene.

Using Restriction Enzymes to Make a Recombinant DNA Plasmid

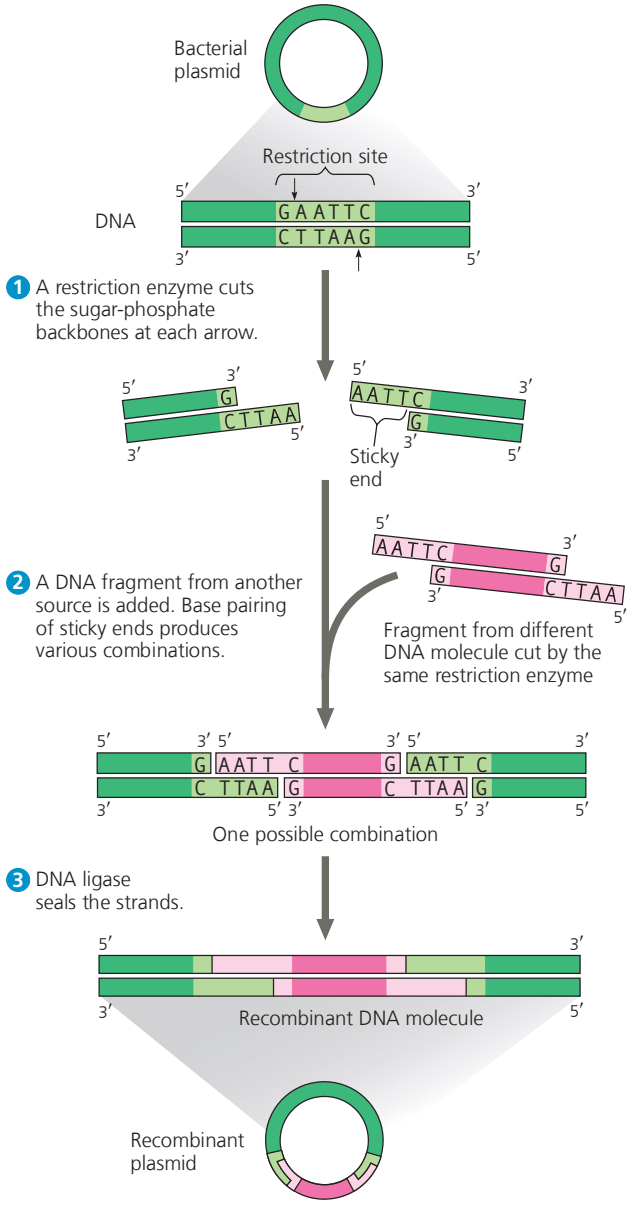
Gene cloning and genetic engineering generally rely on the use of enzymes that cut DNA molecules at a limited number of specific locations. These enzymes, called restriction endonucleases, or **restriction enzymes**, were discovered in the late 1960s by biologists doing basic research on bacteria. Restriction enzymes protect the bacterial cell by cutting up foreign DNA from other organisms or phages (see Concept 19.2).

Hundreds of different restriction enzymes have been identified and isolated. Each restriction enzyme is very specific, recognizing a particular short DNA sequence, or **restriction site**, and cutting both DNA strands at precise points within this restriction site. The DNA of a bacterial cell is protected from the cell's own restriction enzymes by the addition of methyl groups ($—CH_3$) to adenines or cytosines within the sequences recognized by the enzymes.

Mastering Biology Animation: Restriction Enzymes

Figure 20.5 shows how restriction enzymes are used to clone a foreign DNA fragment into a bacterial plasmid. At the top of the figure is a bacterial plasmid (like the one shown in Figure 20.4) that has a single restriction site recognized by a particular restriction enzyme. As shown in this example, most restriction sites are symmetrical. In other words, the sequence of nucleotides is the same on both strands when read in the 5′ → 3′ direction. The most commonly used restriction enzymes recognize sequences containing four to eight nucleotide pairs. Because any sequence that is this short usually occurs (by chance) many times in a long DNA molecule, a restriction enzyme will make many cuts in such a DNA molecule, yielding a set of **restriction fragments**. Since restriction enzymes always cut at the same exact DNA sequence, copies of any given DNA molecule exposed to the same restriction enzyme always yield the same set of restriction fragments.

▼ Figure 20.5 Using a restriction enzyme and DNA ligase to make a recombinant DNA plasmid. The restriction enzyme in this example (called *Eco*RI) recognizes a single six-base-pair restriction site present in this plasmid. It makes staggered cuts in the sugar-phosphate backbones, producing fragments with "sticky ends." Foreign DNA fragments with complementary sticky ends can base-pair with the plasmid ends; the ligated product is a recombinant plasmid. (If the two plasmid sticky ends base-pair, the original nonrecombinant plasmid is reformed.)



DRAW IT *The restriction enzyme* Hind*III recognizes the sequence 5′-AAGCTT-3′, cutting between the two A's. Draw the double-stranded sequence before and after the enzyme cuts it.*

Mastering Biology Animation: Recombinant DNA

The most useful restriction enzymes cleave the sugar-phosphate backbones in the two DNA strands in a staggered manner, as shown in step 1 of Figure 20.5. The resulting double-stranded restriction fragments have at least one single-stranded end, called a **sticky end**. These short extensions can form hydrogen-bonded base pairs with complementary sticky ends on any other DNA molecules cut with the same restriction enzyme, such as the inserted DNA shown in step 2 of Figure 20.5. The associations formed in this way are only temporary but can be made permanent by DNA ligase, an enzyme that catalyzes the formation of covalent bonds that close up the sugar-phosphate backbones of DNA strands (see step 3 of Figure 20.5). At the bottom of Figure 20.5, you can see the stable recombinant DNA molecule that was produced by the ligase-catalyzed joining of DNA from two different sources. The end result, in this example, is the formation of a stable recombinant plasmid containing foreign DNA.

Mastering Biology Animation: Creating Recombinant DNA

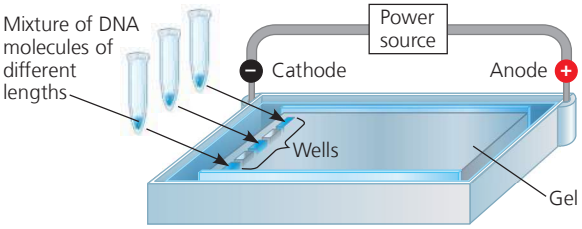
To check the recombinant plasmids after they have been copied many times in host cells to make sure the fragment has been inserted (see Figure 20.4), a researcher might cut the products again using the same restriction enzyme. If the insert is there, there would be two DNA fragments, one the size of the plasmid and one the size of the inserted DNA. To separate and visualize the fragments, researchers carry out a technique called **gel electrophoresis**, which uses a gel made of a polymer that has microscopic holes of different sizes, through which shorter fragments can travel faster. The gel works as a molecular sieve to separate out a mixture of nucleic acid fragments by length **(Figure 20.6)**. Gel electrophoresis is used in conjunction with many different techniques in molecular biology.

Amplifying DNA: The Polymerase Chain Reaction (PCR) and Its Use in DNA Cloning

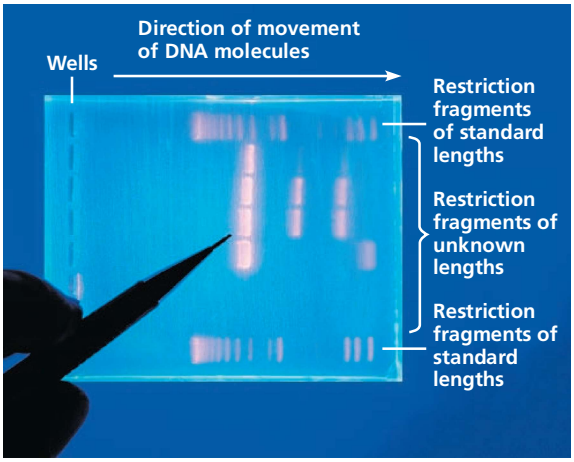
Now that we have examined the cloning vector in some detail, let's consider how biologists obtain the foreign DNA to be inserted. Most researchers have some sequence information about the DNA fragment they want to clone. Using this information, they can start with genomic DNA from the particular species of interest and obtain many copies of the desired gene by using a technique called the **polymerase chain reaction**, or **PCR**. **Figure 20.7** illustrates the steps in PCR. Within a few hours, this technique can make billions of copies of a specific target DNA segment in a sample, even if that segment makes up less than 0.001% of the total DNA in the sample.

Mastering Biology HHMI Video: Polymerase Chain Reaction (PCR)

▼ Figure 20.6 Gel electrophoresis. A gel made of a polymer acts as a molecular sieve to separate nucleic acids or proteins differing in size, electrical charge, or other physical properties as they move in an electric field. In the example shown here, DNA molecules are separated by length in a gel made of a polysaccharide called agarose.



(a) Each sample, a mixture of different DNA molecules, is placed in a separate well near one end of a thin slab of agarose gel. The gel is set into a small plastic support and immersed in an aqueous, buffered solution in a tray with electrodes at each end. The current is then turned on, causing the negatively charged DNA molecules to move toward the positive electrode.



(b) Shorter molecules are slowed down less than longer molecules, so shorter molecules move faster through the gel. After the current is turned off, a DNA-binding dye is added that fluoresces pink in ultraviolet (UV) light. Each pink band corresponds to many thousands of DNA molecules of the same length. The bands at the upper and lower edges of the gel are restriction fragments of standard lengths for comparison with samples of unknown length.

Mastering Biology Animation: Gel Electrophoresis of DNA

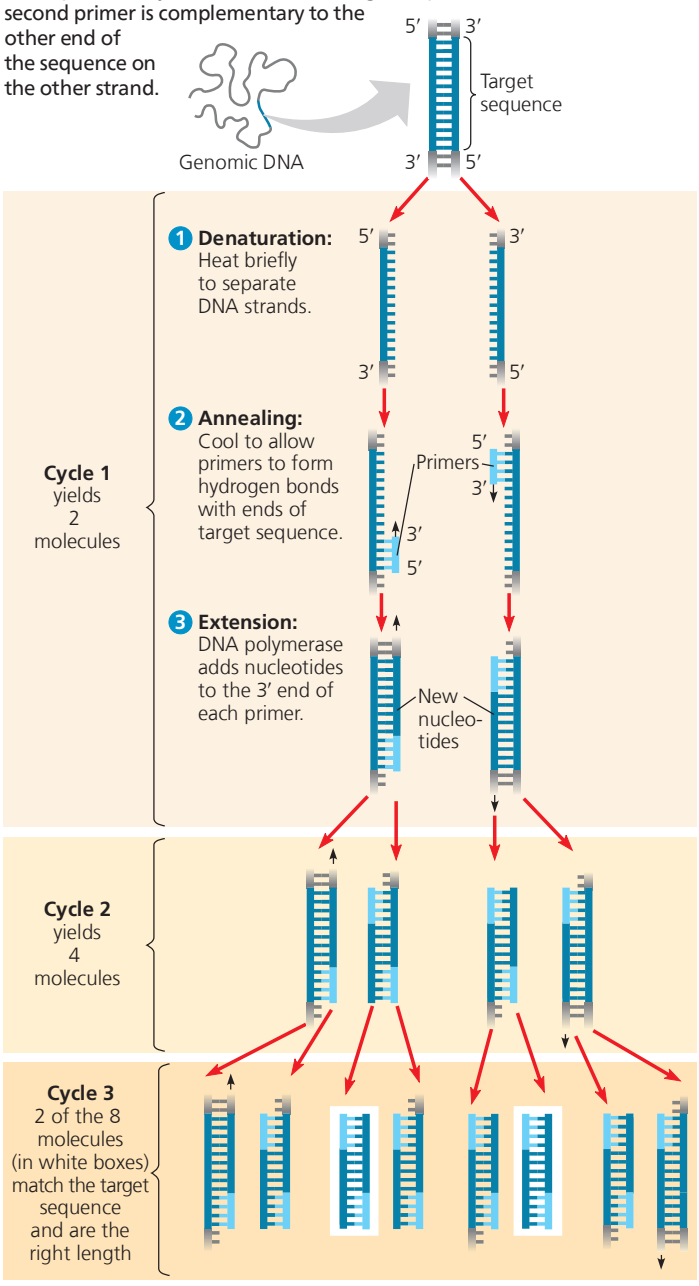
In the PCR procedure, a three-step cycle causes a chain reaction that produces an exponentially growing population of identical DNA molecules. During each cycle, the reaction mixture is 1 heated to high temperatures to denature (separate) the strands of the double-stranded DNA and then 2 cooled to allow annealing (hydrogen bonding) of short, single-stranded DNA primers complementary to sequences on opposite strands at each end of the target sequence; finally, 3 a special DNA polymerase extends the primers in the 5′ → 3′ direction. This cycle is then repeated 30–40 times. If a standard DNA polymerase were used, this enzyme

▼ Figure 20.7 Research Method

The Polymerase Chain Reaction (PCR)

Application With PCR, any specific segment—the so-called target sequence—in a DNA sample can be copied many times (amplified) within a test tube.

Technique PCR requires double-stranded DNA containing the target sequence, a heat-resistant DNA polymerase, all four nucleotides, and two 15- to 20-nucleotide single DNA strands that serve as primers. One primer is complementary to one end of the target sequence on one strand; the second primer is complementary to the other end of the sequence on the other strand.



Results After three cycles, two molecules match the target sequence exactly. After 30 more cycles, over 1 billion (10^9) molecules match.

Mastering Biology Animation: Copying DNA Through PCR

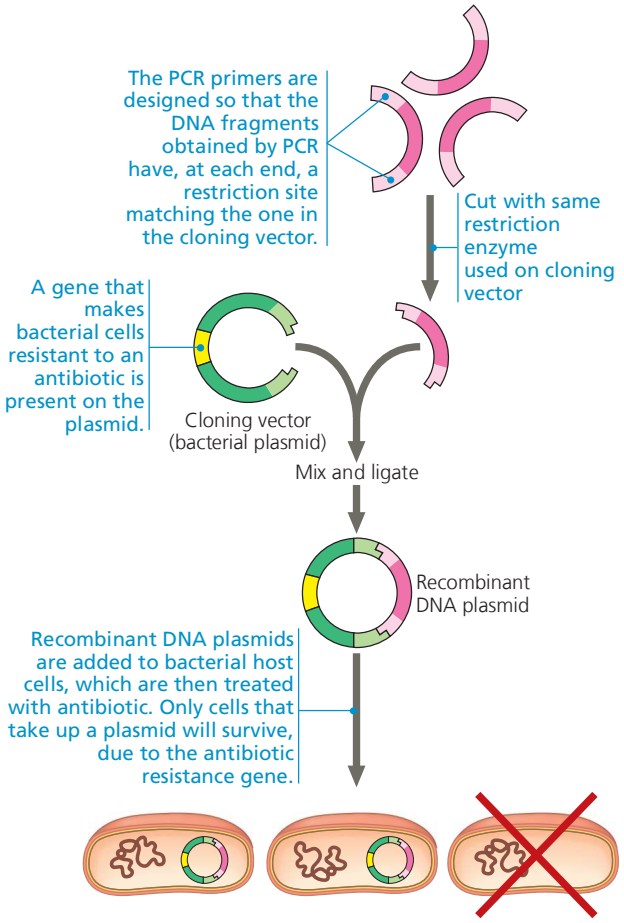
would be denatured along with the DNA during the first heating step and would have to be replaced after each cycle. The key to automating PCR was the discovery of an unusual heat-stable DNA polymerase enzyme called *Taq* polymerase, named after the bacterial species from which it was first isolated. This bacterial species, *Thermus aquaticus*, lives in hot springs, and the stability of its DNA polymerase at high temperatures is an evolutionary adaptation that enables the enzyme to function at temperatures up to 95°C. Today, researchers also use a DNA polymerase from the archaean species *Pyrococcus furiosus*. This enzyme, called *Pfu* polymerase, is more accurate and stable but more expensive than *Taq* polymerase.

PCR is speedy and very specific. Only a minuscule amount of DNA need be present in the starting material, and this DNA can be partially degraded, as long as there are a few copies of the complete target sequence. The key to the high specificity is the pair of primers used for each PCR amplification. The primer sequences are chosen so that they hybridize *only* to sequences at opposite ends of the target segment, one on the 3′ end of each strand. (For high specificity, the primers must be at least 15 nucleotides long.) With each successive cycle, the number of target segment molecules of the correct length doubles, so the number of molecules equals 2^n, where n is the number of cycles. After 30 or so cycles, about a billion copies of the target sequence are present!

Despite its speed and specificity, PCR amplification cannot substitute for gene cloning in cells to make large amounts of a gene. This is because the polymerases that are used have no proofreading function, and occasional errors during PCR replication limit the number of good copies and the length of DNA fragments that can be copied. Instead, PCR is used to provide the specific DNA fragment for cloning. PCR primers are synthesized to include a restriction site at each end of the DNA fragment that matches the site in the cloning vector, and the fragment and vector are cut and ligated together **(Figure 20.8)**. The resulting plasmids are sequenced so that those with error-free inserts can be selected.

Devised in 1985, PCR has had a major impact on biological research and genetic engineering. PCR has been used to amplify DNA from a wide variety of sources: a 40,000-year-old frozen woolly mammoth; fingerprints or tiny amounts of blood, tissue, or semen found at crime scenes; single embryonic cells for rapid prenatal diagnosis of genetic disorders (see Figure 14.19); and cells infected with viruses that are

▼ Figure 20.8 Use of a restriction enzyme and PCR in gene cloning. In a closer look at the process shown at the top of Figure 20.4, PCR is used to produce the DNA fragment or gene of interest that will be ligated into a cloning vector, in this case a bacterial plasmid.



difficult to detect, such as HIV. (To test for HIV, viral genes are amplified.) We'll return to applications of PCR later.

Expressing Cloned Eukaryotic Genes

Once a gene has been cloned in host cells, its protein product can be expressed in large amounts for research or for practical applications, which we'll explore in Concept 20.4. Cloned genes can be expressed in either bacterial or eukaryotic cells; each option has advantages and disadvantages.

Bacterial Expression Systems

Getting a cloned eukaryotic gene to function in bacterial host cells can be difficult because certain aspects of gene expression are different in eukaryotes and bacteria. To overcome differences in promoters and other DNA control sequences (see Concept 17.2), scientists usually employ an **expression vector**, a cloning vector that contains a highly active bacterial promoter just upstream of a restriction site where the eukaryotic gene can be inserted in the correct reading frame. The bacterial host cell will recognize the promoter and proceed to express the foreign gene now linked to that promoter. Such expression vectors allow the synthesis of many eukaryotic proteins in bacterial cells.

Another problem with expressing cloned eukaryotic genes in bacteria is the presence of noncoding regions (introns) in most eukaryotic genes (see Concept 17.3). Introns can make a eukaryotic gene very long and unwieldy, and they prevent correct expression of the gene by bacterial cells, which do not have RNA-splicing machinery. This problem can be surmounted by using a form of the gene that includes only the exons. (This is called *complementary DNA*, or *cDNA*; see Figure 20.10.)

Eukaryotic DNA Cloning and Expression Systems

Molecular biologists can avoid eukaryotic-bacterial incompatibility by using eukaryotic cells such as yeasts as hosts for cloning and expressing eukaryotic genes. Yeasts, single-celled fungi, are as easy to grow as bacteria, and they have plasmids, a rarity among eukaryotes.

In addition to enabling RNA splicing, eukaryotic host cells are advantageous because many eukaryotic proteins will not function unless they are modified after translation—for example, by the addition of carbohydrate groups (glycosylation) or lipid groups in the ER and Golgi. Bacterial cells lack membrane-bound organelles and cannot carry out these modifications, and if the gene product requiring such processing is from a mammalian gene, even yeast cells may not be able to modify the protein correctly. Several cultured cell types have proved successful as host cells for this purpose, including some mammalian cell lines and an insect cell line that can be infected by a virus carrying recombinant DNA.

Scientists have developed other methods for introducing recombinant DNA into eukaryotic cells. In **electroporation**, a brief electrical pulse applied to a solution containing cells creates temporary holes in their plasma membranes, through which DNA can enter. (This technique is now commonly used for bacteria as well.) Alternatively, scientists can inject DNA directly into single eukaryotic cells using microscopically thin needles. Another way to get DNA into plant cells is by using the soil bacterium *Agrobacterium tumefaciens*, as we'll see later. Whatever the method, if the introduced DNA is incorporated into a cell's genome by genetic recombination, then it can be stably expressed by the cell. Expressing different versions of genes in cells allows researchers to study protein function, a topic we'll return to in Concept 20.2.

Cross-Species Gene Expression and Evolutionary Ancestry

EVOLUTION The ability to express eukaryotic proteins in bacteria (even if the proteins can't be modified properly) is quite remarkable when we consider how different eukaryotic and bacterial cells are. In fact, examples abound of genes that are taken from one species and function perfectly well when transferred into another very different species, such as a firefly gene in a tobacco plant and a jellyfish gene in a pig (see Figure 17.7). These observations underscore the shared evolutionary ancestry of species living today.

One example involves a gene called *Pax-6*, which has been found in animals as diverse as vertebrates and fruit flies. The vertebrate *Pax-6* gene product (the PAX-6 protein) triggers a complex program of gene expression resulting in formation of the vertebrate eye, which has a single lens. Expression of the fly *Pax-6* gene leads to formation of the compound fly eye, which is quite different from the vertebrate eye. When the mouse *Pax-6* gene was cloned and introduced into a fly embryo so that it replaced the fly's own *Pax-6* gene, researchers were surprised to see that the mouse version of the gene led to formation of a compound fly eye (see Figure 50.16). Conversely, when the fly *Pax-6* gene was transferred into a vertebrate embryo—a frog, in this case—a frog eye formed. Although the genetic programs triggered in vertebrates and flies generate very different eyes, the two versions of the *Pax-6* gene can substitute for each other to trigger lens development, evidence of their evolution from a gene in a very ancient common ancestor. Because of their ancient evolutionary roots, all living organisms share the same basic mechanisms of gene expression. This commonality is the basis of many recombinant DNA techniques described in this chapter.

CONCEPT CHECK 20.1

1. **MAKE CONNECTIONS** The restriction site for an enzyme called *Pvu*I is the following sequence:

   5′–CGATCG–3′
   3′–GCTAGC–5′

   Staggered cuts are made between the T and C on each strand. What type of bonds are being cleaved? (See Concept 5.5.)
2. **DRAW IT** One strand of a DNA molecule has the following sequence:

   5′–CTTGACGATCGTTACCG–3′

   Draw the other strand. Will *Pvu*I (see question 1) cut this molecule? If so, draw the products.
3. What are some potential difficulties in using plasmid vectors and bacterial host cells to produce large quantities of proteins from cloned eukaryotic genes?
4. **VISUAL SKILLS** Compare Figure 20.7 with Figure 16.21. How does replication of DNA ends during PCR proceed without shortening the fragments each time?

For suggested answers, see Appendix A.

CONCEPT 20.2

Biologists use DNA technology to study gene expression and function

To see how a biological system works, scientists seek to understand the functioning of the system's component parts. Analysis of when and where a gene or group of genes is expressed can provide important clues about their function and how they contribute to the organism as a whole.

Analyzing Gene Expression

Biologists driven to understand the assorted cell types of a multicellular organism, cancer cells, or the developing tissues of an embryo first try to discover which genes are expressed by the cells of interest. The most straightforward way to do this is usually to identify the mRNAs being made. We'll first examine techniques that look for patterns of expression of specific individual genes. Next, we'll explore ways to characterize groups of genes being expressed by cells or tissues of interest. As you will see, all of these procedures depend in some way on base pairing between complementary nucleotide sequences.

Studying the Expression of Single Genes

Suppose we have cloned a gene that we suspect plays an important role in the embryonic development of *Drosophila melanogaster* (the fruit fly). The first thing we might wonder is which embryonic cells express the gene—in other words, where in the embryo is the corresponding mRNA found? We can detect the mRNA in an embryonic sample using nucleic acid hybridization with molecules of complementary sequence to the mRNA we want to follow. Using our cloned gene as a template, we can synthesize a short, single-stranded nucleic acid (either RNA or DNA) complementary to the mRNA of interest; this is called a **nucleic acid probe**. For example, if part of the sequence on the mRNA were

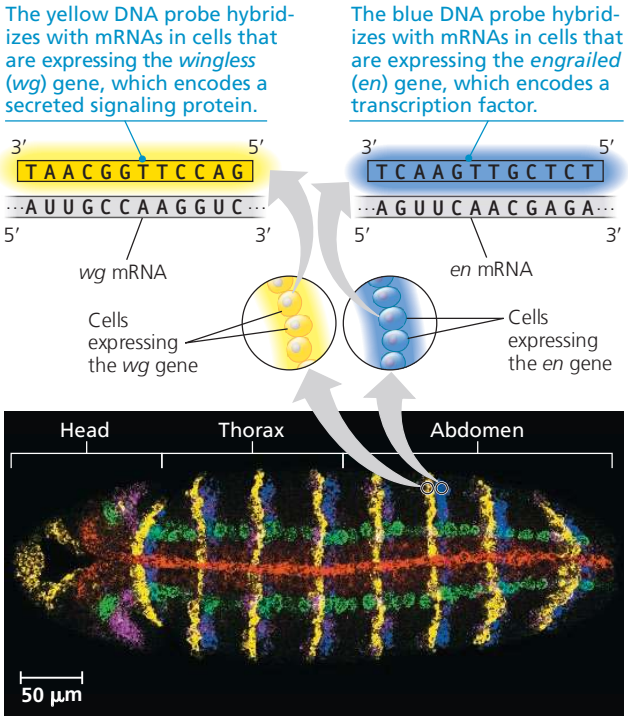
5′ ···C U C A U C A C C G G C··· 3′

then we would synthesize this single-stranded DNA probe:

3′ G A G T A G T G G C C G 5′

Each probe molecule is labeled during synthesis with a fluorescent tag so we can follow it. A solution containing probe molecules is applied to *Drosophila* embryos, allowing the probe to hybridize specifically with any complementary sequences on the many mRNAs in embryonic cells in which the gene is being transcribed. Because this technique allows us to see the mRNA in place (or *in situ*, in Latin) in the intact organism, this technique is called ***in situ* hybridization**.

▼ Figure 20.9 Determining where single genes are expressed by *in situ* hybridization analysis. This *Drosophila* embryo was incubated in a solution containing DNA probes for five different mRNAs, each probe labeled with a different fluorescently colored tag. The embryo was then viewed using fluorescence microscopy. Each color marks where a specific gene is expressed as mRNA. The arrows from the groups of yellow and blue cells above the micrograph show a magnified view of nucleic acid hybridization of the appropriately colored probe to the mRNA. The thorax (trunk) and abdomen are made up of repeating segments. Yellow cells (expressing the *wg* gene) interact with blue cells (expressing the *en* gene); their interaction helps establish the pattern in a body segment.

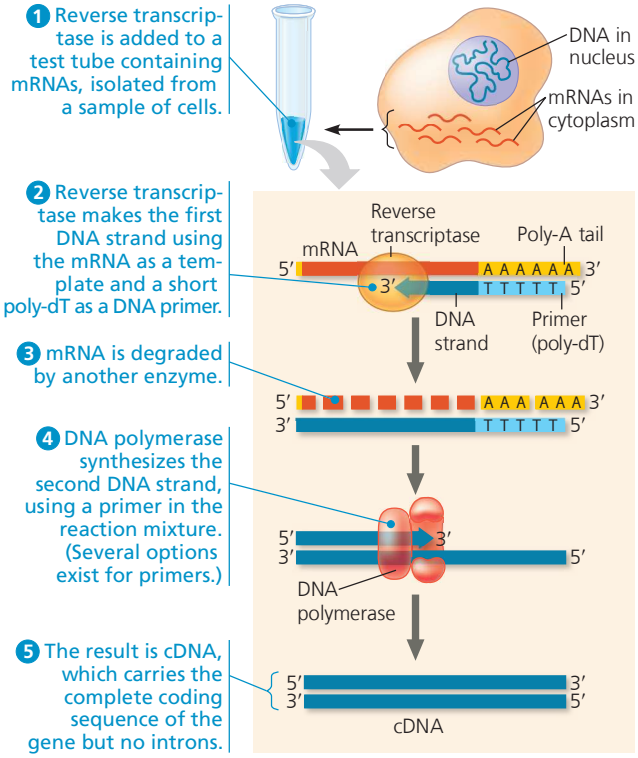


Different probes can be labeled with different fluorescent dyes, sometimes with strikingly beautiful results **(Figure 20.9)**.

Other mRNA detection techniques may be preferable for comparing the amounts of a specific mRNA in several samples at the same time—for example, in different cell types or in embryos of different stages. One method that is widely used is called the **reverse transcriptase polymerase chain reaction**, or **RT-PCR**.

RT-PCR begins by turning sample sets of mRNAs into double-stranded DNAs with the corresponding sequences. First, the enzyme reverse transcriptase (from a retrovirus; see Figure 19.9) is used to synthesize a complementary DNA copy (a *reverse transcript*) of each mRNA in the sample **(Figure 20.10)**. The mRNA is then degraded by addition of a specific enzyme, and a second DNA strand, complementary to the first, is synthesized by DNA polymerase. The resulting double-stranded DNA is called **complementary DNA (cDNA)**. (Made from mRNA, cDNA lacks introns and can be used for protein expression in bacteria, as mentioned earlier.) To analyze the timing of expression of the *Drosophila* gene of interest, for example, we would first isolate all the mRNAs from different stages of *Drosophila* embryos and make cDNA from each stage **(Figure 20.11)**.

▼ Figure 20.10 Making complementary DNA (cDNA) from eukaryotic genes. Complementary DNA is made in a test tube using mRNA as a template for the first strand. Only one mRNA is shown after step 1, but the final collection of cDNAs would reflect all the mRNAs present in the cell.



Next in RT-PCR is the PCR step (see Figure 20.7). As you will recall, PCR is a way of rapidly making many copies of one specific stretch of double-stranded DNA, using primers that hybridize to the opposite ends of the segment that we are interested in studying. In our case, we would add primers corresponding to a segment of our *Drosophila* gene, using the cDNA from each embryonic stage as a template for PCR amplification in separate samples. When the products are analyzed on a gel, copies of the amplified region will be observed as bands only in samples that originally contained mRNA from the gene being studied. This method can tell researchers whether an mRNA is present. However, accurate measurement of mRNA levels requires newer, more quantitative PCR machines that use a dye that fluoresces only when bound to a double-stranded PCR product. This technique,

called *quantitative RT-PCR* (*qRT-PCR*), can detect the light and measure the PCR product, thus avoiding the need for electrophoresis while providing quantitative data, a distinct advantage. RT-PCR or qRT-PCR can also be carried out with mRNAs collected from different tissues at one time to discover which tissue is producing a specific mRNA.

Instructors: The Scientific Skills Exercise "Analyzing Quantitative and Spatial Gene Expression Data" can be assigned in **Mastering Biology**. Students can work with data from an experiment that investigated mRNA expression using both *in situ* hybridization and quantitative RT-PCR.

Studying the Expression of Interacting Groups of Genes

A major goal of biologists is to learn how genes act together to produce and maintain a functioning organism. Now that the genomes of a number of species have been sequenced, it is possible to study the expression of large groups of genes—an approach called the *systems approach*. Researchers use what is known about the whole genome to investigate which genes are transcribed in different tissues or at different stages of development. One aim is to identify networks of gene expression across an entire genome.

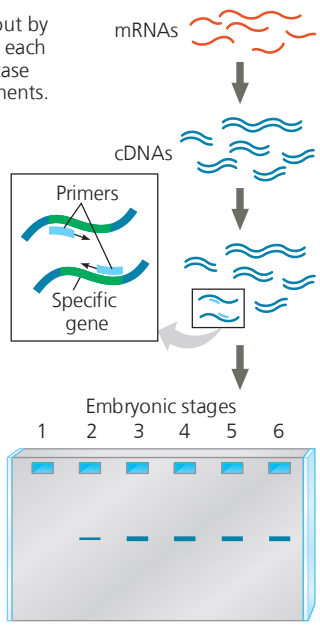
To accomplish such genome-wide expression studies, today's rapid, inexpensive DNA-sequencing methods allow researchers to discover which genes are expressed by simply sequencing the cDNA samples from different tissues or different embryonic stages. This straightforward method is called **RNA sequencing**, or **RNA-seq** (pronounced "RNA-seek"), even though it is the cDNA that is actually sequenced. In RNA-seq, the mRNA (or other RNA) samples are isolated, cut into shorter, similar-sized fragments, and converted into cDNAs **(Figure 20.12)**. These short cDNA stretches are sequenced, and a computer program reassembles them, either mapping them onto the genome of the species in question (when available) or simply putting the fragments in order from scratch based on overlapping sequences of multiple RNAs.

▼ Figure 20.11 Research Method

RT-PCR Analysis of the Expression of Single Genes

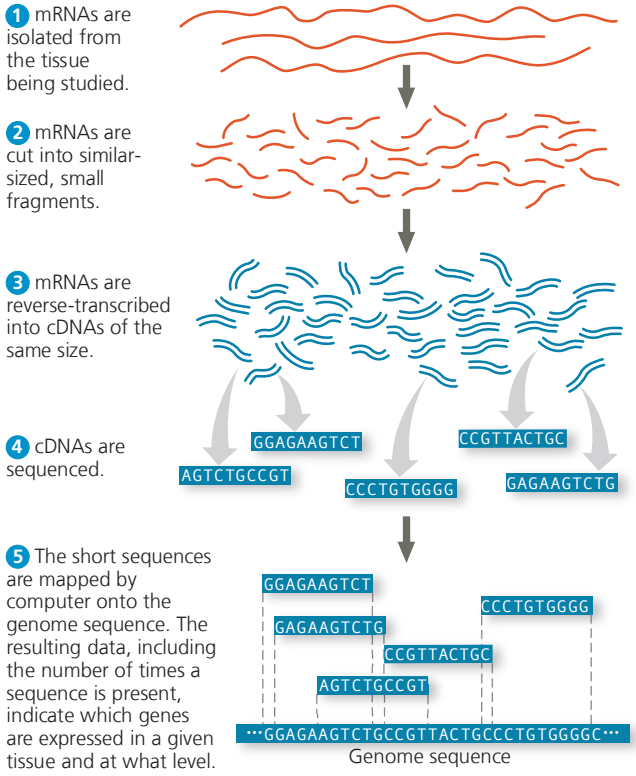
Application RT-PCR uses the enzyme reverse transcriptase (RT) in combination with PCR and gel electrophoresis. RT-PCR can be used to compare gene expression between samples—for instance, in different embryonic stages, in different tissues, or in the same type of cell under different conditions.

Technique In this example, samples containing mRNAs from six embryonic stages of *Drosophila* were analyzed for a specific mRNA as shown below. (In steps 1 and 2, the mRNA from only one stage is shown.)



Results The mRNA for this gene is expressed from stage 2 through stage 6. The size of the amplified fragment (shown by its position on the gel) depends on the distance between the primers that were used (not on the size of the mRNA).

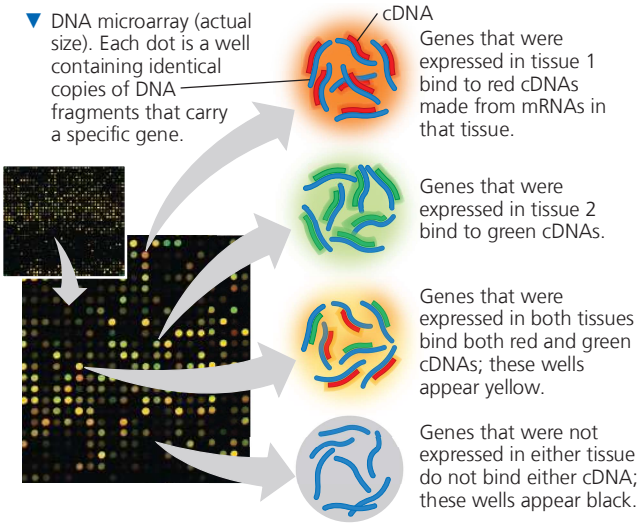
▼ Figure 20.12 Use of RNA sequencing (RNA-seq) to analyze expression of many genes. RNA-seq yields a wide range of information about expression of genes, including their level of expression.



Several characteristics make RNA-seq a powerful technique. First, it is not based on hybridization with a labeled probe, so it doesn't depend on knowing genomic sequences. Second, it can measure levels of expression over a very wide range. Third, a careful analysis provides a wealth of information about expression of a particular gene, such as relative levels of alternatively spliced mRNAs. In most cases, however, expression of individual genes still needs to be confirmed by RT-PCR.

An older method of genome-wide expression studies, called **DNA microarray assays**, is less powerful than RNA-seq but is still used for some applications, such as fetal testing (see Figure 14.19). A DNA microarray consists of tiny amounts of a large number of single-stranded DNA fragments representing different genes fixed to a glass slide in a tightly spaced array, or grid, of dots. (The microarray is also called a *DNA chip* by analogy to a computer chip.) Ideally, these fragments represent all the genes of an organism. The mRNAs from the cells being studied are reverse-transcribed into cDNAs (see Figure 20.10), and a fluorescent label is added so the cDNAs can be used as probes on the microarray. Different fluorescent labels are used for different cell samples so that multiple samples can be tested in the same experiment. The resulting pattern of colored dots, shown in an actual-size microarray in **Figure 20.13**, reveals the dots to which each probe was bound and thus the genes that are expressed in the cell samples being tested.

▼ Figure 20.13 Use of microarrays to analyze expression of many genes. Researchers extracted mRNAs from two different human tissues and synthesized two sets of cDNAs, fluorescently labeled red (tissue 1) or green (tissue 2). Labeled cDNAs were hybridized with a microarray containing 5,760 human genes (about 25% of human genes), part of which is shown in the enlargement. Red indicates that the gene in that well was expressed in tissue 1, green in tissue 2, yellow in both, and black in neither. The fluorescence intensity at each spot indicates the relative expression of the gene.



Scientists can now measure the expression of thousands of genes at one time. DNA technology makes such studies possible; with automation, they are easily performed on a large scale. By uncovering gene interactions and providing clues to gene function, DNA microarray assays and RNA-seq may contribute to a better understanding of diseases and suggest new diagnostic techniques or therapies. For instance, comparing patterns of gene expression in breast cancer tumors and noncancerous breast tissue has already resulted in more informed and effective treatment protocols (see Figure 18.27). Ultimately, information from these methods should provide a grander view of how ensembles of genes interact to form an organism and maintain its vital systems.

Determining Gene Function

Once they identify a gene of interest, how do scientists determine its function? A gene's sequence can be compared with sequences in other species. If the function of a similar gene in another species is known, one might suspect that the gene product in question performs a comparable task. Data about the location and timing of gene expression may reinforce the suggested function. To obtain stronger evidence, one approach is to disable the gene and then observe the consequences in the cell or organism.

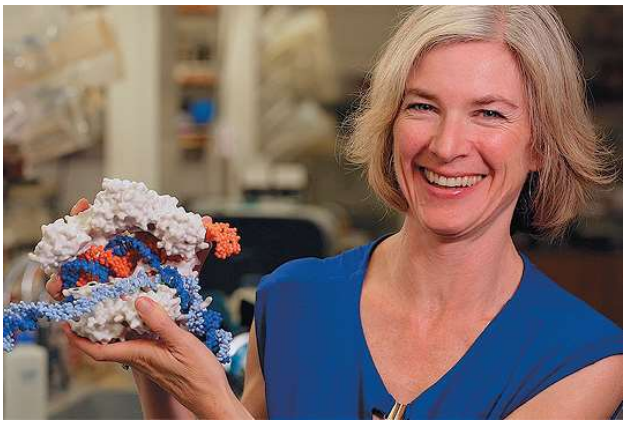
Editing Genes and Genomes

Molecular biologists have long sought techniques for altering, or editing, the genetic material of cells or organisms in a predictable way. In one such technique, called ***in vitro* mutagenesis**, specific mutations are introduced into a cloned gene, and the mutated gene is returned to a cell in such a way that it disables ("knocks out") the normal cellular copies of the same gene. If the introduced mutations alter or destroy the function of the gene product, the phenotype of the mutant cell may help reveal the function of the missing normal protein.

In Concept 17.5, you read about the CRISPR-Cas9 system, the powerful new technique for gene editing in living cells and organisms that is taking the field of genetic engineering by storm (see Figure 17.28). This system, worked out by Jennifer Doudna **(Figure 20.14)** and Emmanuelle Charpentier, is a highly effective way for researchers to knock out a given gene in order to study what that gene does. It has already been used in many organisms, including bacteria, fish, mice, insects, human cells, and various crop plants. Modifications of the technique allow researchers to repair a gene that has a mutation. This approach is used for gene therapy, which will be discussed later in the chapter.

In another application of the CRISPR-Cas9 system, scientists are attempting to address the global problem of insect-borne diseases by altering genes in insects so that,

▼ Figure 20.14 Jennifer Doudna holding a model of CRISPR-Cas9.



Mastering Biology **HHMI Video: Jennifer Doudna: CRISPR-Cas Proteins are the Swiss Army Knife of the Genome**

for example, they cannot transmit disease. An extra twist to this approach is to engineer the new allele so that it is much more highly favored for inheritance than the wild-type allele. This technology is called a **gene drive** because the biased inheritance of the engineered gene during reproduction rapidly "drives" the new allele through the population.

Other Methods for Studying Gene Function

Another method for silencing expression of selected genes doesn't alter the genome; instead, it exploits the phenomenon of **RNA interference (RNAi)**, described in Concept 18.3. This experimental approach uses synthetic double-stranded RNA molecules matching the sequence of a particular gene to trigger breakdown of the gene's messenger RNA or to block its translation. In organisms such as the nematode and the fruit fly, RNAi has already proved valuable for analyzing the functions of genes on a large scale. This method is quicker than using the CRISPR-Cas9 system, but it leads to only a temporary reduction of gene expression rather than a permanent gene knockout or alteration.

In humans, ethical considerations prohibit knocking out genes to determine their functions. An alternative approach is to analyze the genomes of large numbers of people with a certain phenotypic condition or disease, such as heart disease or diabetes, to try to find differences they all share compared with people without that condition. The logic is that these differences may be associated with one or more malfunctioning genes and thus in a sense are naturally occurring gene knockouts. In these large-scale analyses, called **genome-wide association studies**, researchers look for *genetic markers*, DNA sequences that vary in the population. In a gene, such sequence variation is the basis of different alleles, as we have seen for sickle-cell disease (see Figure 17.26). And just like the coding sequences of genes, noncoding DNA at a specific locus on a chromosome may exhibit small nucleotide differences among individuals. Variations in coding or noncoding DNA sequences among a population are called polymorphisms (from the Greek for "many forms").

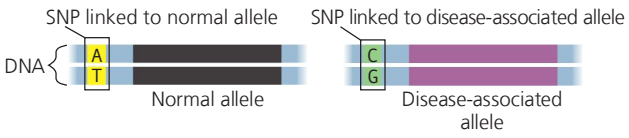
Mastering Biology **Interview with Charles Rotimi: Using genomics to study health-related conditions in African-Americans**



Among the most useful genetic markers in tracking down genes that contribute to diseases and disorders are single base-pair variations in the genomes of the human population. For more than 99% of the nucleotides in the human genome, virtually all people have the same nucleotide in each position. However, once in every 100–300 base pairs of both coding and noncoding DNA sequences are positions where the sequence varies between individuals. A single base-pair site where variation is found in at least 1% of the population is called a **single nucleotide polymorphism** (**SNP**, pronounced "snip"); a few million SNPs occur in the human genome. To find SNPs in large numbers of people, it isn't necessary to sequence their DNA; SNPs can be detected by very sensitive microarray assays, RNA-seq, or PCR.

Once a SNP is identified that is found in all people affected by the disease being studied, researchers focus on that region and sequence it. In nearly all cases, the SNP itself does not contribute directly to the disease in question by altering the encoded protein; in fact, most SNPs are in noncoding regions. Instead, having a particular SNP associated with a disease suggests that the gene whose mutation causes the disease is located very close to that SNP on that chromosome. This closeness (*genetic linkage*; see Concept 15.3) means that crossing over between the SNP and the gene is very unlikely during gamete formation. Therefore, the SNP and gene are almost always inherited together, so the SNP acts as a genetic marker for the disease-causing allele **(Figure 20.15)**. SNPs have been

▼ Figure 20.15 Single nucleotide polymorphisms (SNPs) as genetic markers for disease-associated alleles. This diagram depicts the same region of the genome from two groups of individuals, one group having a particular disease or condition with a genetic basis. Unaffected people have an AT pair at a given SNP locus, while affected people have a CG pair there. Once the allele is confirmed as being associated with the disease in question, the SNP that varies in this way can be used as a marker for the disease-associated allele.



found that correlate with diabetes, heart disease, and several types of cancer, and the search is on for genes that might be involved in these and other conditions that are based on genetics.

The experimental approaches you have learned about thus far focus on working with molecules, mainly DNA and proteins. In a parallel line of inquiry, biologists have been developing powerful techniques for cloning whole multicellular organisms. One aim of this work is to obtain special types of cells, called stem cells, that can give rise to all types of tissues. Being able to manipulate stem cells would allow scientists to use the DNA-based methods previously discussed to alter stem cells for the treatment of diseases. Methods involving the cloning of organisms and production of stem cells are the subject of the next section.

CONCEPT CHECK 20.2

1. Describe the role of complementary base pairing during RT-PCR, RNA sequencing, and DNA microarray analysis.
2. **VISUAL SKILLS** Consider the microarray in Figure 20.13. If a sample from normal tissue is labeled with a green fluorescent dye and a sample from cancerous tissue is labeled red, what color spots would represent genes you would be interested in if you were studying cancer? Explain.

For suggested answers, see Appendix A.

CONCEPT 20.3

Cloned organisms and stem cells are useful for basic research and other applications

Along with advances in DNA technology, scientists have been developing and refining methods for cloning whole multicellular organisms from single cells. In this context, cloning produces one or more organisms that are genetically identical to the "parent" that donated the single cell. This is often called *organismal cloning* to differentiate it from gene cloning and, more significantly, from cell cloning—the division of an asexually reproducing cell such as a bacterium into a group of genetically identical cells. (The common theme is that the product is genetically identical to the parent.) The current interest in organismal cloning is primarily due to its ability to generate stem cells. A **stem cell** is a relatively unspecialized cell that can both reproduce itself indefinitely and, under appropriate conditions, differentiate into specialized cells of one or more types. Stem cells have great potential for regenerating damaged tissues.

The cloning of plants and animals was first attempted over 50 years ago in experiments designed to answer basic questions about the genetic potential of single cells. For example, researchers wondered if all the cells of an organism have the same genes or whether cells lose genes during the process of differentiation (see Concept 18.4). One way to answer this question is to see whether a differentiated cell can generate a whole organism—in other words, whether cloning an organism is possible. Let's discuss these early experiments before we consider more recent progress in organismal cloning and procedures for producing stem cells.

Cloning Plants: Single-Cell Cultures

The successful cloning of whole plants from single differentiated cells was accomplished at Cornell University during the 1950s by F. C. Steward and his students, who worked with carrot plants. They found that differentiated cells taken from the root (the carrot) and incubated in culture medium could grow into normal adult plants, each genetically identical to the parent plant. These results showed that differentiation does not necessarily involve irreversible changes in the DNA. In plants, mature cells can "dedifferentiate" and then give rise to all the specialized cell types of the organism; any cell with this potential is said to be **totipotent**.

Plant cloning is used extensively in agriculture. For plants such as orchids, cloning is the only commercially practical means of producing new plants. In other cases, cloning has been used to reproduce a plant with valuable characteristics, such as resistance to plant pathogens. In fact, you yourself may be a plant cloner: If you have ever grown a new plant from a cutting, you have practiced cloning!

Cloning Animals: Nuclear Transplantation

Differentiated cells from animals generally do not divide in culture, much less develop into the multiple cell types of a new organism. Therefore, early researchers had to use a different approach to answer the question of whether differentiated animal cells are totipotent. Their approach was to remove the nucleus of an egg (creating an *enucleated* egg) and replace it with the nucleus of a differentiated cell, a procedure called *nuclear transplantation*, now more commonly called *somatic cell nuclear transfer*. If the nucleus from the differentiated donor cell retains its full genetic potential, then it should be able to direct development of the recipient cell into all the tissues and organs of an organism. Such experiments were conducted on one species of frog (*Rana pipiens*) by Robert Briggs and Thomas King in the 1950s and on another frog species (*Xenopus laevis*) by John Gurdon in the 1970s **(Figure 20.16)**. These researchers transplanted a nucleus from an embryonic or tadpole cell into an enucleated egg of the same species. In Gurdon's experiments, the transplanted nucleus was often able to support normal development of the egg into a tadpole. However, he found that the potential of a transplanted nucleus to direct normal development was inversely related

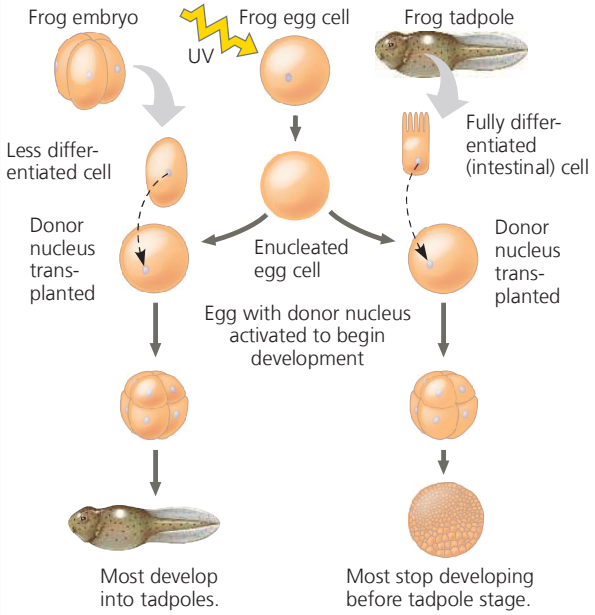
to the age of the donor: The older the donor nucleus, the lower the percentage of normal tadpoles.

From these results, Gurdon concluded that something in the nucleus *does* change as animal cells differentiate. In frogs and most other animals, nuclear potential tends to be restricted more and more as embryonic development and cell differentiation progress. These were foundational experiments that ultimately led to stem cell technology, and Gurdon received the 2012 Nobel Prize in Medicine for this work.

▼ Figure 20.16 Inquiry

Can the nucleus from a differentiated animal cell direct development of an organism?

Experiment John Gurdon and colleagues at Oxford University, in England, destroyed the nuclei of frog (*Xenopus laevis*) eggs by exposing the eggs to ultraviolet light. They then transplanted nuclei from cells of frog embryos and tadpoles into the enucleated eggs.



Results When the transplanted nuclei came from an early embryo, the cells of which are relatively undifferentiated, most of the recipient eggs developed into tadpoles. But when the nuclei came from the fully differentiated intestinal cells of a tadpole, fewer than 2% of the eggs developed into normal tadpoles, and most of the embryos stopped developing at a much earlier stage.

Conclusion The nucleus from a differentiated frog cell can direct development of a tadpole. However, its ability to do so decreases as the donor cell becomes more differentiated, presumably because of changes in the nucleus.

Data from J. B. Gurdon et al., The developmental capacity of nuclei transplanted from keratinized cells of adult frogs, *Journal of Embryology and Experimental Morphology* 34:93–112 (1975).

WHAT IF? *If each cell in a four-cell embryo was already so specialized that it was not totipotent, what results would you predict for the experiment on the left side of the figure?*

Mastering Biology **HHMI Video: Somatic Cell Nuclear Transfer**

Reproductive Cloning of Mammals

In addition to cloning frogs, researchers were able to clone mammals using early embryonic cells as a source of donor nuclei. Until about 25 years ago, though, it was not known whether a nucleus from a fully differentiated cell could be reprogrammed successfully to act as a donor nucleus. In 1997, researchers in Scotland announced the birth of Dolly, a lamb cloned from an adult sheep by nuclear transfer from a differentiated mammary gland cell **(Figure 20.17)**. Using a technique related to that in Figure 20.16, the researchers implanted early embryos into surrogate mothers. Out of several hundred embryos, one successfully completed normal development, and Dolly was born, a genetic clone of the nucleus donor. By the age of 6, Dolly had developed premature arthritis, and complications from a lung infection led to her euthanization. This led to speculation that this sheep's cells were in some way not quite as healthy as those of a normal sheep, possibly reflecting incomplete reprogramming of the original transplanted nucleus. Reprogramming involves epigenetic changes that lead to changes in chromatin structure (see Concept 18.2), to be discussed shortly.

▶ Figure 20.17
Reproductive cloning of a mammal by nuclear transfer. Dolly, shown here as a lamb, has a very different appearance from her surrogate mother, standing beside her.



Since that time, researchers have cloned many other mammals, including mice, cats, cows, horses, pigs, and dogs. In 2018, Chinese biologists reported the first cloning of a primate, the long-tailed macaque. In most cases, the research goal has been the production of new individuals, known as *reproductive cloning*. We have learned from such experiments that cloned animals of the same species are not always identical. For example, in 2016, scientists examined four 7- to 9-year-old clones genetically identical to Dolly and concluded that, unlike Dolly, they were healthy and aging normally. Another example of nonidentity in clones is the

▼ Figure 20.18 CC ("Carbon Copy"), the first cloned cat (right), and her single parent. Rainbow (left) donated the nucleus in a cloning procedure that resulted in CC. However, the two cats are not identical: Rainbow is a classic calico cat with orange patches on her fur and has a "reserved personality," while CC has a gray and white coat and is more playful.



first cloned cat, named CC for Carbon Copy **(Figure 20.18)**. The color and pattern of her calico coat differed from that of her single female parent because of random X chromosome inactivation, which is a normal occurrence during embryonic development (see Figure 15.8). And identical human twins, which are naturally occurring "clones," are always slightly different. Clearly, environmental influences and random phenomena play a significant role during development.

Epigenetic Differences in Cloned Animals

In most nuclear transplantation studies thus far, only a small percentage of cloned embryos develop normally to birth. And like Dolly, many cloned animals exhibit defects. Cloned mice, for instance, are prone to obesity, pneumonia, liver failure, and premature death. Scientists assert that even cloned animals that appear normal are likely to have subtle defects.

Researchers have uncovered some reasons for the low efficiency of cloning and the high incidence of abnormalities. In the nuclei of fully differentiated cells, a small subset of genes is turned on and expression of the rest of the genes is repressed. This regulation often is the result of epigenetic changes in chromatin, such as acetylation of histones or methylation of DNA (see Figure 18.7). During the nuclear transfer procedure, many of these changes must be reversed in the later-stage nucleus from a donor animal for genes to be expressed or repressed appropriately in earlier stages of development. Researchers have found that the DNA in cells from cloned embryos, like that of differentiated cells, often has abnormally high numbers of methyl groups. This finding suggests that the reprogramming of donor nuclei requires more accurate and complete chromatin restructuring than occurs during cloning procedures. Because DNA methylation helps regulate gene expression, misplaced or extra methyl groups in the DNA of donor nuclei may interfere with the pattern of gene expression necessary for normal embryonic development. In fact, the success of a cloning attempt may depend in large part on whether or not the chromatin in the donor nucleus can be artificially modified to resemble that of a newly fertilized egg.

Stem Cells of Animals

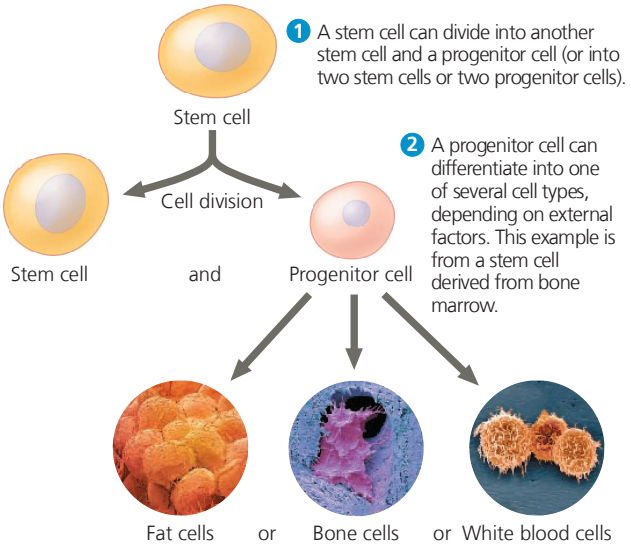
Progress in cloning mammalian embryos, including primates, has heightened speculation about the cloning of humans, which has not yet been achieved past very early embryonic stages. The main reason researchers have been trying to clone human embryos is not for reproduction, but for the production of stem cells to treat human diseases. Recall that a stem cell is a relatively unspecialized cell that can both reproduce itself indefinitely and, under appropriate conditions, differentiate into specialized cells of one or more types **(Figure 20.19)**.

Mastering Biology HHMI Animation: Somatic Cell Nuclear Transfer

Embryonic and Adult Stem Cells

Many early animal embryos contain stem cells capable of giving rise to differentiated cells of any type. Stem cells can be isolated from early embryos at a stage called the blastula stage or its human equivalent, the blastocyst stage. In culture, these

▼ Figure 20.19 How stem cells maintain their own population and generate differentiated cells.



Mastering Biology HHMI Video: Cultured Human Embryonic Stem Cells

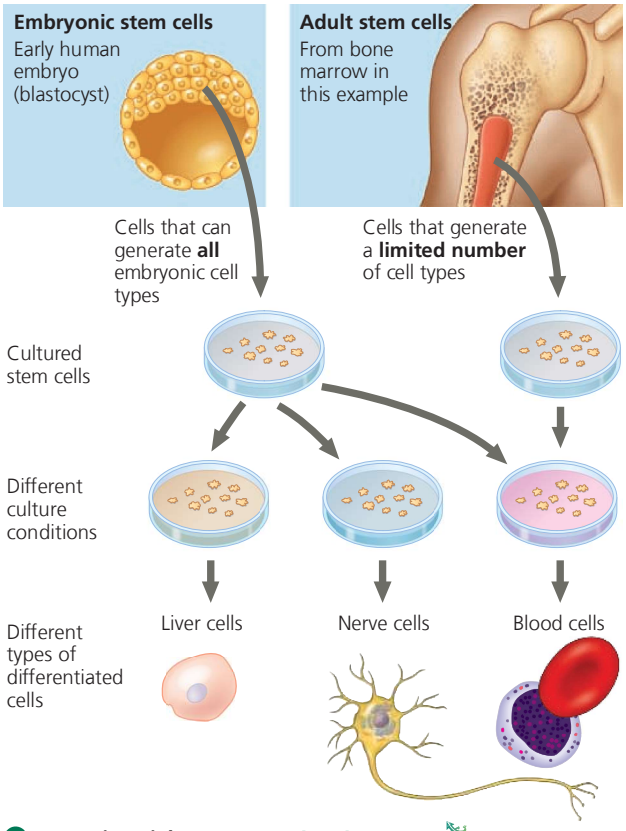
embryonic stem (ES) cells reproduce indefinitely, and depending on culture conditions, they can be made to differentiate into a wide variety of specialized cells **(Figure 20.20)**, including even eggs and sperm.

The adult body also has stem cells, which serve to replace nonreproducing specialized cells as needed. In contrast to ES cells, *adult stem cells* are not able to give rise to all cell types in the organism, though they can generate several defined types. For example, one of the several types of stem cells in bone marrow can generate all the different kinds of blood cells (see Figure 20.20), and another type of bone marrow stem cell can differentiate into bone, cartilage, fat, muscle, and the linings of blood vessels. To the surprise of many, the adult brain has been found to contain stem cells that continue to produce certain kinds of nerve cells there. Researchers have also reported finding stem cells in skin, hair, eyes, and dental pulp. Although adult animals have only tiny numbers of stem cells, scientists are learning to identify and isolate these cells from various tissues and, in some cases, to grow them in culture. With the right culture conditions (for instance, the addition of specific growth factors), cultured stem cells from adult animals have been made to differentiate into various defined types of specialized cells, although none are as versatile as ES cells.

Research with embryonic or adult stem cells is a source of valuable data about differentiation and has enormous potential for medical applications. The ultimate aim is to supply cells for the repair of damaged or diseased organs: for example, insulin-producing pancreatic cells for people with type 1 diabetes or certain kinds of brain cells for people with Parkinson's disease or Huntington's disease. Adult stem cells from bone marrow have long been used in bone marrow transplants as a source of immune system cells in patients whose own immune systems are nonfunctional because of genetic disorders or radiation treatments for cancer.

The developmental potential of adult stem cells is limited to certain tissues. ES cells hold more promise than adult stem cells for most medical applications because ES cells are **pluripotent**, capable of differentiating into many different cell types. One source of ES cells, first reported in 2013, is cloned human blastocysts produced by transferring a nucleus from a differentiated cell into an enucleated egg (similar to cloning Dolly the sheep). Prior to that report, cells were obtained only from embryos donated by patients undergoing infertility treatments or from long-term cell cultures originally established with cells isolated from donated embryos, an issue that prompts ethical and political discussions. Although the techniques for cloning early human embryos are still being optimized and the ethics discussed by researchers, they represent a potential new source for ES cells that may be less controversial. Furthermore, with a donor nucleus from a person with a particular disease, researchers should be able to produce ES cells that match the patient and are thus not rejected by his or her immune system when used for treatment. When the main aim of cloning is to produce ES cells to treat disease, the process is called *therapeutic cloning*. Although most people believe that reproductive cloning of humans is unethical, opinions vary about the morality of therapeutic cloning.

▼ Figure 20.20 Working with stem cells. Animal stem cells, which can be isolated from early embryos or adult tissues and grown in culture, are self-perpetuating, relatively undifferentiated cells. Embryonic stem cells are easier to grow than adult stem cells and can theoretically give rise to *all* types of cells in an organism. The range of cell types that can arise from adult stem cells is not yet fully understood.



Mastering Biology **HHMI Animation: Creating Embryonic Stem Cell Lines**

Induced Pluripotent Stem (iPS) Cells

In another approach to make stem cells for research and therapy, researchers succeeded in 2007 in learning how to turn back the clock in fully differentiated cells, reprogramming them to act like ES cells. Differentiated cells can be transformed into a type of ES cell by using a modified retrovirus to introduce extra, cloned copies of four "stem cell" master regulatory genes. The "deprogrammed" cells are known as *induced pluripotent stem* (*iPS*) cells because, in using this fairly simple laboratory technique to return them to their

undifferentiated state, pluripotency has been restored. The experiments that first transformed human differentiated cells into iPS cells are described in **Figure 20.21**. Shinya Yamanaka received the 2012 Nobel Prize in Medicine for this work, shared with John Gurdon, whose work you read about in Figure 20.16.

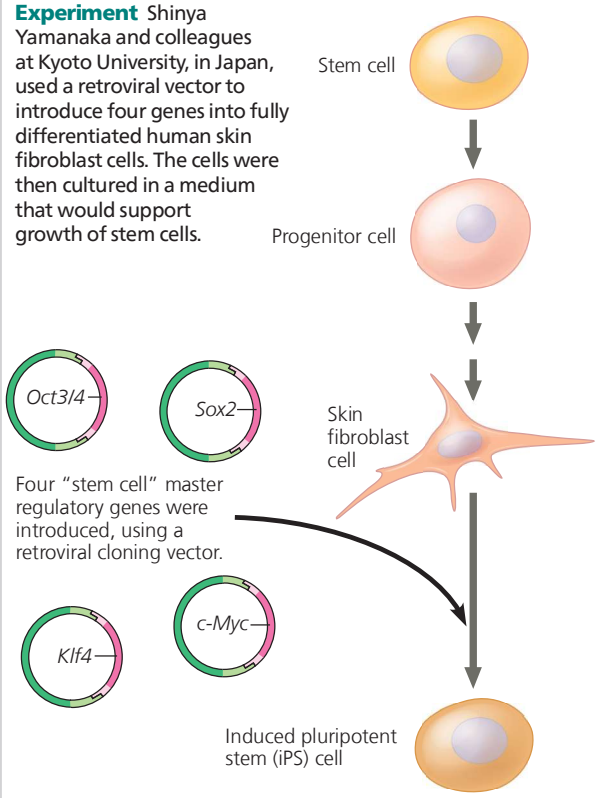
By many criteria, iPS cells can perform most of the functions of ES cells, but there are some differences in gene expression and other cellular functions, such as cell division. At least until these differences are fully understood, the study of ES cells will continue to make important contributions to the development of stem cell therapies. (In fact, it is likely that ES cells will always be a focus of basic research as well.) In the meantime, work is proceeding using the iPS cells that have been experimentally produced.

There are two major potential uses for human iPS cells. First, cells from patients with diseases have been reprogrammed to become iPS cells, which act as model cells for studying the disease and potential treatments. Human iPS cell lines have already been developed from individuals with type 1 diabetes, Parkinson's disease, Huntington's disease, Down syndrome, and many other diseases. Second, in the field of regenerative medicine, a patient's own cells could be reprogrammed into iPS cells and then used to replace nonfunctional tissues, such as cells of the retina of the eye that have been damaged by a condition called age-related macular degeneration (AMD). In fact, in 2014, Japanese researchers made iPS cells out of an AMD patient's skin cells, caused them to differentiate into retinal cells, and implanted them into her retina. Although her vision was not markedly improved, further deterioration was halted. Unfortunately, the cost of reprogramming cells into iPS cells is so high that it can't be used as a standard treatment. Over time, the procedure may become less expensive, and some researchers have suggested creating a bank for storage of donor iPS cells that could be matched to patients' tissues. Development of techniques that direct iPS cells to become specific cell types for regenerative medicine is an area of intense research. If the challenges are met, iPS cells created in this way could eventually provide tailor-made "replacement" cells for patients without using any human eggs or embryos, thus circumventing most ethical objections.

▼ Figure 20.21 Inquiry

Can a fully differentiated human cell be "deprogrammed" to become a stem cell?

Experiment Shinya Yamanaka and colleagues at Kyoto University, in Japan, used a retroviral vector to introduce four genes into fully differentiated human skin fibroblast cells. The cells were then cultured in a medium that would support growth of stem cells.



Results Two weeks later, the cells resembled embryonic stem cells in appearance and were actively dividing. Their gene expression patterns, gene methylation patterns, and other characteristics were also consistent with those of embryonic stem cells. The iPS cells were able to differentiate into heart muscle cells, as well as other cell types.

Conclusion The four genes induced differentiated skin cells to become pluripotent stem cells, with characteristics of embryonic stem cells.

Data from K. Takahashi et al., Induction of pluripotent stem cells from adult human fibroblasts by defined factors, *Cell* 131:861–872 (2007).

WHAT IF? *Patients with diseases such as heart disease or Alzheimer's could have their own skin cells reprogrammed to become iPS cells. Once affordable procedures have been developed for this and for converting iPS cells into heart or nervous system cells, the patients' own iPS cells might be used to treat their disease. When organs are transplanted from a donor to a recipient, the recipient's immune system may reject the transplant, a dangerous condition. Would using iPS cells be expected to carry the same risk? Why or why not? Given that these cells are actively dividing, undifferentiated cells, what risks might this procedure carry?*

CONCEPT CHECK 20.3

1. Based on current knowledge, how would you explain the difference in the percentage of tadpoles that developed from the two kinds of donor nuclei in Figure 20.16?
2. A few companies in China and South Korea provide the service of cloning dogs, using cells from their clients' pets to provide nuclei in procedures like that in Figure 20.17. Should their clients expect the clone to look identical to their original pet? Why or why not? What ethical questions does this bring up?
3. **MAKE CONNECTIONS** Based on what you know about muscle differentiation (see Figure 18.18) and genetic engineering, propose the first experiment you might try if you wanted to direct an embryonic stem cell or iPS cell to develop into a muscle cell.

For suggested answers, see Appendix A.

CONCEPT 20.4

The practical applications of DNA-based biotechnology affect our lives in many ways

In the last section, we'll survey the practical applications of DNA-based **biotechnology**, the manipulation of organisms or their components to make useful products. Today, major applications of DNA technology and genetic engineering include medicine, forensic evidence and genetic profiles, environmental cleanup, and agriculture.

Medical Applications

One important use of DNA technology is the identification of human genes whose mutation plays a role in genetic diseases. These discoveries may lead to ways of diagnosing, treating, and even preventing such conditions. DNA technology has also identified genes that play a role in a number of "nongenetic" diseases, from arthritis to AIDS, by influencing susceptibility to these diseases. Furthermore, a wide variety of diseases involve changes in gene expression within the affected cells and often within the patient's immune system. By using RNA-seq and DNA microarray assays (see Figures 20.12 and 20.13) or other techniques to compare gene expression in healthy and diseased tissues, researchers are finding genes that are turned on or off in particular diseases. These genes and their products are potential targets for prevention or therapy.

Diagnosis and Treatment of Diseases

A new chapter in the diagnosis of infectious diseases has been opened by DNA technology, in particular the use of PCR and labeled nucleic acid probes to track down pathogens. For example, because the sequence of the RNA genome of HIV is known, RT-PCR can be used to amplify and therefore detect and quantify even a small amount of HIV RNA in blood or tissue samples (see Figure 20.11). RT-PCR is often the best way to detect an otherwise elusive infectious agent.

Medical scientists can now diagnose hundreds of human genetic disorders by using PCR with primers that target the genes associated with these disorders. The amplified DNA product is then sequenced to reveal the presence or absence of the disease-causing mutation. Among the genes for human diseases targeted in this way are those for sickle-cell disease, hemophilia, cystic fibrosis, Huntington's disease, and Duchenne muscular dystrophy. Individuals with such diseases can often be identified before the onset of symptoms, even before birth (see Figure 14.19). PCR can also be used to identify symptomless carriers of potentially harmful recessive alleles as part of genetic counseling (see Concept 14.4).

Personal Genome Analysis

As you learned earlier, genome-wide association studies have pinpointed SNPs (single nucleotide polymorphisms) that are linked to disease-associated alleles (see Figure 20.15). Individuals can be tested by PCR and sequencing for a SNP that is correlated with the abnormal allele. The presence of particular SNPs is correlated with increased risk for particular adverse health conditions such as heart disease, Alzheimer's disease, and some types of cancer.

Direct-to-consumer genome analysis companies offer kits allowing individuals to send in a swab containing cheek cells that the company will analyze genetically. Genetic testing for risk factors like heart disease, Alzheimer's disease, and some types of cancer is carried out by looking for linked SNPs that have been previously identified (see Figure 20.15). It may be helpful for individuals to learn about their health risks, with the understanding that such genetic tests merely reflect correlations and do not make predictions.

In addition to health-related genetic information, these companies compare a person's DNA segments with those from reference populations around the world, established from thousands of individuals of known ancestry. Based on how closely the segments match up, the report can tell an individual about their likely ancestral breakdown. Females can also learn about their maternal lines of descent, based on comparisons of mitochondrial DNA, which is contributed to the egg only by the mother. For males, an analysis of the sequence of the Y chromosome can trace their paternal lines of descent. As the size of the database used by these companies increases, the results become more refined and more accurate.

Personalized Medicine

The techniques described in this chapter have also prompted improvements in disease treatments—for example, by knowing which specific cancer-related genes have been mutated in a particular patient's tumor. By analyzing the expression of many genes in large numbers of breast cancer patients, researchers can refine their understanding of the different subtypes of breast cancer (see Figure 18.27). Knowing the expression levels of particular genes in a given individual can help physicians determine the likelihood that the cancer will recur, thus helping them design an appropriate treatment. Given that some low-risk patients have a 96% survival rate over a ten-year period with no treatment, gene expression analysis allows doctors and patients access to valuable information when they are considering treatment options.

Many researchers envision a future of **personalized medicine**, a type of medical care in which each person's specific genetic profile can provide information about diseases or conditions for which the person is especially at risk

and help make health-care decisions. As we will see later in the chapter, a *genetic profile* is currently taken to mean a set of genetic markers such as SNPs. Ultimately, however, it will likely mean the complete DNA sequence of an individual—after sequencing becomes inexpensive enough. In fact, a 2016 study suggested that for difficult-to-diagnose patients it might be faster and more economical to sequence just the expressed regions (the exons, together called the *exome*) or even all of the genome right away, rather than running through a standard series of diagnostic tests. The ability to sequence a person's genome quickly and inexpensively is advancing rapidly, perhaps faster than development of appropriate treatments for the conditions. Still, the identification of genes involved in these conditions provides targets for therapeutic interventions.

An individual's genomic information can be used to predict the benefits and risks of particular medications, an approach called *pharmacogenetics*. There are over 300 medications for which the Food and Drug Administration recommends genetic testing for patients prior to their use.

Human Gene Therapy and Gene Editing

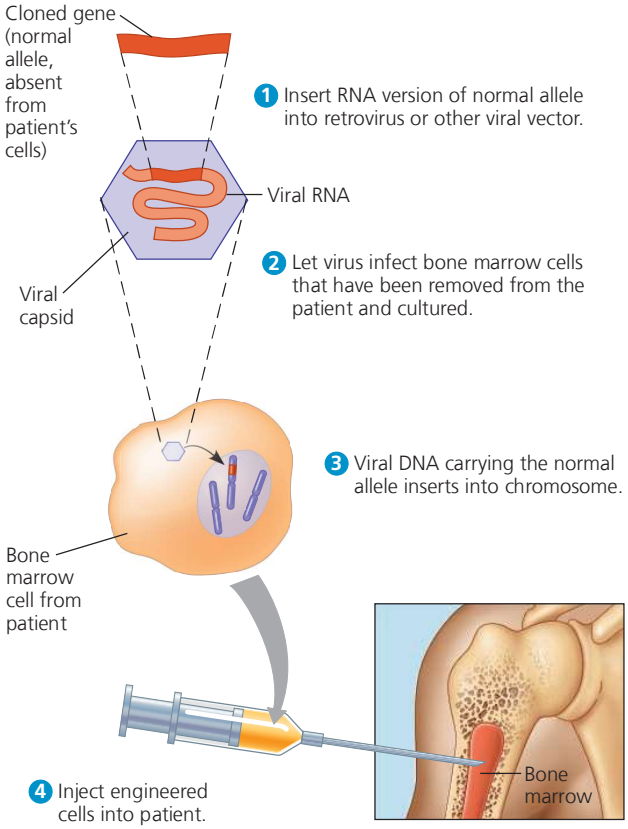
Gene therapy—the introduction of genes into an afflicted individual for therapeutic purposes—holds great potential for treating the relatively small number of disorders traceable to a single defective gene. The aim of this approach is to insert a normal allele of the defective gene into the somatic cells of the tissue affected by the disorder.

Mastering Biology **BBC Video: Cystic Fibrosis and the Promise of Gene Therapy**

For gene therapy of somatic cells to be permanent, the cells that receive the normal allele must be cells that multiply throughout the patient's life. Bone marrow cells, which include the stem cells that give rise to all the cells of the blood and immune system, are prime candidates. **Figure 20.22** outlines one procedure for gene therapy in an individual whose bone marrow cells do not produce a vital enzyme because of a single defective gene. One type of severe combined immunodeficiency (SCID) is caused by this kind of defect. If the treatment is successful, the patient's bone marrow cells will begin producing the missing protein, and the patient may be cured.

The procedure shown in Figure 20.22 was used in gene therapy trials for SCID in France in 2000, resulting in the first indisputable success of gene therapy. However, three patients subsequently developed leukemia, a type of blood cell cancer, and one of them died. Researchers have concluded it is likely that the insertion of the retroviral vector occurred near a gene that triggers the proliferation of blood cells. Using a viral vector that does not come from a retrovirus, clinical researchers have treated several other genetic diseases somewhat successfully with gene therapy, including

▼ Figure 20.22 Gene therapy using a retroviral vector.
A retrovirus that has been rendered harmless is used as a vector in this procedure, which exploits the ability of a retrovirus to insert a DNA transcript of its RNA genome into the chromosomal DNA of its host cell (see Figure 19.9). If the foreign gene carried by the retroviral vector is expressed, the cell and its descendants will possess the gene product. Cells that reproduce throughout life, such as bone marrow cells, are ideal candidates for gene therapy.



Mastering Biology **HHMI Video: Gene Therapy**
HHMI Video: The Search for a Mutated Gene
HHMI Video: Genes as Medicine

a type of progressive blindness (see Concept 50.3), a degenerative disease of the nervous system, and a blood disorder involving the β-globin gene.

Gene therapy raises many technical issues. For example, how can the activity of the transferred gene be controlled so that cells make appropriate amounts of the gene product at the right time and in the right place? How can we be sure that the insertion of the therapeutic gene does not harm some other necessary cell function? As more is learned about DNA control elements and gene interactions, researchers may be able to answer such questions.

A more direct approach that avoids the complications of using a viral vector in gene therapy is made possible by gene editing, especially using the CRISPR-Cas9 system

(see Figure 17.28). In this approach, the existing defective gene is edited to correct the mutation.

Mastering Biology **HHMI Video: Jennifer Doudna: CRISPR Provides a Tool for Changing Genes**

In 2018, researchers reported promising results using the CRISPR-Cas9 system in an attempt to correct the genetic defect that causes sickle-cell disease. They edited cells from patients with sickle-cell disease and injected them into bone marrow in mice. After 19 weeks, the gene remained corrected in 20–40% of the injected cells. Researchers, physicians, and patients alike are excited about the potential of CRISPR technology to treat or even cure human diseases that have a genetic basis, such as sickle-cell, Alzheimer's, and Parkinson's diseases, as well as some types of cancer. However, there are still concerns about using CRISPR-treated cells in humans because of possible effects on genes that are not being targeted; this is an active area of research.

In addition to technical challenges, gene therapy and gene editing provoke ethical questions. Some critics believe that tampering with human genes in any way is unethical. Other observers see no fundamental difference between transplanting genes into somatic cells and transplanting organs from one person to another.

An even more pressing question is the issue of engineering human germ-line cells to try to correct a defect in future generations. Such genetic engineering is now routinely done in laboratory mice, and some methods for genetic engineering of human embryos have been developed.

The development of the CRISPR-Cas9 system has engendered much debate about the ethics of gene editing. Jennifer Doudna (Figure 20.14), a co-discoverer of CRISPR-Cas9, and other biologists have agreed to exercise extreme caution as the field moves forward. Together, they called for the research community to "strongly discourage" any experimental work on human eggs or embryos.

In spite of this general consensus among biologists, a Chinese researcher reported in 2018 that he had used the CRISPR-Cas9 system to edit genes in embryos that completed fetal development and were born as twins and a third individual. He claimed to have edited the *CCR5* gene, which codes for a co-receptor for HIV (see Figure 7.8), so that HIV would be unable to bind and infect the cells. The twins' father was HIV-positive, which was the rationale the scientist used for carrying out this genetic engineering. However, this act was roundly condemned as highly unethical by the biological community; the researcher lost his job and may face a criminal investigation. In 2019, an advisory panel to the World Health Organization proposed establishing a registry to monitor any gene-editing research in humans.

There are many potential technical problems with the CRISPR-Cas9 technique, like the off-target effects mentioned previously. Beyond those technical concerns, however, are underlying ethical considerations: Under what circumstances, if any, should genomes of human germ lines be altered? Would alterations lead to the practice of eugenics, a deliberate effort to control the genetic makeup of human populations? It is imperative to consider these questions.

Pharmaceutical Products

The pharmaceutical industry derives significant benefit from advances in DNA technology and genetic research, applying them to the development of useful drugs to treat diseases. Pharmaceutical products are synthesized using methods of either organic chemistry or biotechnology, depending on the nature of the product.

Synthesis of Small Molecules for Use as Drugs

Determining the sequence and structure of proteins crucial for tumor cell survival has led to the identification of small molecules that combat certain cancers by blocking the function of these proteins. One drug, imatinib (trade name Gleevec), is a small molecule that inhibits one tyrosine kinase (see Figure 11.8). The overexpression of this kinase, resulting from a chromosomal translocation, is instrumental in causing chronic myelogenous leukemia (CML; see Figure 15.16). Patients in the early stages of CML who are treated with imatinib have exhibited nearly complete, sustained remission from the cancer. Drugs that work in a similar way have also been used with success to treat a few types of lung and breast cancers. This approach is feasible only for cancers whose molecular basis is fairly well understood.

Mastering Biology **HHMI Video: Gleevec**

In many cases of such drug-treated tumors, though, cells later arise that are resistant to the new drug. In one study, the whole genome of the tumor cells was sequenced both before and after the appearance of drug resistance. Comparison of the sequences showed genetic changes that allowed the tumor cells to "get around" the drug-inhibited protein. Here, we can see that cancer cells demonstrate the principles of evolution: Certain tumor cells have a random mutation that allows them to survive in the presence of a particular drug, and as a consequence of natural selection in the presence of the drug, these are the cells that survive and reproduce.

Protein Production in Cell Cultures Pharmaceutical products that are proteins are commonly synthesized on a large scale using cell cultures. You learned earlier in the chapter about DNA cloning and gene expression systems for producing large quantities of a chosen protein that is present naturally in only minute amounts. The host cells used in such expression systems can even be engineered to secrete a protein as it is made, thereby simplifying the task of purifying it by traditional biochemical methods.

Among the first pharmaceutical products manufactured in this way were human insulin and human growth hormone (HGH). Some 2 million people with diabetes in the United States depend on insulin treatment to control their disease. Human growth hormone has been a boon to children born with a form of dwarfism caused by inadequate amounts of HGH, as well as helping patients with AIDS gain weight. Another important pharmaceutical product produced by genetic engineering is tissue plasminogen activator (TPA). If administered shortly after a heart attack, TPA helps dissolve blood clots and reduces the risk of subsequent heart attacks.

Protein Production by "Pharm" Animals In some cases, instead of using cell systems to produce large quantities of protein products, pharmaceutical scientists can use whole organisms. To do this, they use a **transgene**, a gene that has been transferred into one organism from another. They first remove eggs from a female of the recipient species and fertilize them *in vitro*. Meanwhile, they clone the desired gene from a donor organism. They then inject the cloned DNA directly into the nuclei of the fertilized eggs. Some of the cells integrate the foreign DNA, the transgene, into their genome and are able to express the foreign gene. The engineered embryos that arise from these zygotes are then surgically implanted in a surrogate female. If an embryo develops successfully, the result is a **transgenic** organism that expresses its new, "foreign" gene.

Assuming that the introduced gene encodes a protein desired in large quantities, transgenic animals can act as pharmaceutical "factories." For example, a transgene for a human blood protein such as antithrombin, which prevents blood clots, can be inserted into the genome of a goat in such a way that the transgene's product is secreted in the animal's milk **(Figure 20.23)**. The protein is then purified from the milk (which is easier than purification from a cell culture). Such proteins must be tested to ensure that they (or contaminants from the farm animals) will not cause allergic reactions or other adverse effects in patients who receive them.

▼ Figure 20.23 Goats as "pharm" animals. This transgenic goat carries a gene for a human blood protein, antithrombin, which she secretes in her milk. Patients with a rare hereditary disorder in which this protein is lacking suffer from formation of blood clots in their blood vessels. Easily purified from the goat's milk, the protein is used to prevent blood clots in these patients during surgery or childbirth.





Forensic Evidence and Genetic Profiles

In violent crimes, body fluids or small pieces of tissue may be left at the scene or on the clothes or other possessions of the victim or assailant. If enough blood, semen, or tissue is available, forensic laboratories can determine the blood type or tissue type by using antibodies to detect specific cell-surface proteins. However, such tests require fairly fresh samples in relatively large amounts. Also, because many people have the same blood or tissue type, this approach can only exclude a suspect; it cannot provide strong evidence of guilt.

DNA testing, on the other hand, can identify the guilty individual with a high degree of certainty because the DNA sequence of every person is unique (except for identical twins). Genetic markers that vary in the population can be analyzed for a given person to determine that individual's unique set of genetic markers, or **genetic profile**. (This term is preferred over "DNA fingerprint" by forensic scientists, who want to emphasize the heritable aspect of these markers rather than the fact that they produce a pattern on a gel that, like a fingerprint, is visually recognizable.) The FBI started applying DNA technology to forensics in 1988, using a method involving gel electrophoresis and nucleic acid hybridization to detect similarities and differences in DNA samples. This method required much smaller samples of blood or tissue than earlier methods—only about 1,000 cells.

Mastering Biology Animation: Genetic Profiles

Today, forensic scientists use an even more sensitive method that takes advantage of variations in length of genetic markers called **short tandem repeats (STRs)**. These are tandemly repeated units of two- to five-nucleotide sequences in specific regions of the genome. The number of repeats present in these regions is highly variable from person to person (polymorphic); even for a single individual, the two alleles of an STR may differ from each other. For example, one individual may have the sequence ACAT repeated 30 times at one genome locus and 15 times at the same locus on the other homolog, whereas another individual may have 18 repeats at this locus on each homolog. (These two genotypes can be expressed by the two repeat numbers: 30,15 and 18,18.) PCR is used to amplify particular STRs, using sets of primers that are labeled with colored fluorescent tags; the length of the region, and thus the number of repeats, can then be determined by electrophoresis. The PCR step allows use of this method even when the DNA is in poor condition or available only in minute quantities: A tissue sample containing as few as 20 cells can be sufficient.

In a murder case, for example, this method can be used to compare DNA samples from the suspect, the victim, and a small amount of blood found at the crime scene. The forensic scientist tests only a few selected portions of the DNA—usually 13 STR markers. However, even this small set of markers can provide a forensically useful genetic profile because the probability that two people (who are not identical twins) would have exactly the same set of STR markers is vanishingly small. The Innocence Project, a nonprofit organization dedicated to overturning wrongful convictions, uses STR analysis of archived samples from crime scenes to revisit old cases. As of 2019, according to that group, 362 innocent people have been released from prison as a result of DNA-based forensic and legal work **(Figure 20.24)**.

Genetic profiles can also be useful for other purposes. A comparison of the DNA of a mother, her child, and the purported father can conclusively settle a question of paternity. Sometimes paternity is of historical interest: Genetic profiles provided strong evidence that Thomas Jefferson or one of his close male relatives fathered at least one of the children of Sally Hemings, one of his slaves. Genetic profiles can also identify victims of mass casualties. The largest such effort occurred after the attack on the World Trade Center in 2001; more than 10,000 samples of victims' remains were compared with DNA samples from personal items, such as toothbrushes, provided by families. Ultimately, forensic scientists succeeded in identifying almost 3,000 victims using these methods.

Just how reliable is a genetic profile? The greater the number of markers examined in a DNA sample, the more likely it is that the profile is unique to one individual. In forensic cases using STR analysis with 13 markers, the probability of two people having identical DNA profiles is somewhere between one in 10 billion and one in several trillion. (For comparison, the world's population is between 7 and 8 billion.) The exact probability depends on the frequency of those markers in the general population. Information on how common various markers are in different ethnic groups is critical because these marker frequencies may vary considerably among ethnic groups and between a particular ethnic group and the population as a whole. With the increasing availability of frequency data, forensic scientists can make extremely accurate statistical calculations. Thus, despite problems that can arise from insufficient data, human error, or flawed evidence, genetic profiles are now accepted as compelling evidence by legal experts and scientists alike.

▼ Figure 20.24 STR analysis used to release an innocent man from prison.

(a) In 1984, Earl Washington was convicted and sentenced to death for the 1982 rape and murder of Rebecca Williams. His sentence was commuted to life in prison in 1993 due to new doubts about the evidence. In 2000, STR analysis by forensic scientists associated with the Innocence Project showed conclusively that he was innocent. This photo shows Washington just before his release in 2001, after 17 years in prison.



Source of sample	STR marker 1	STR marker 2	STR marker 3
Semen on victim	17,19	13,16	12,12
Earl Washington	16,18	14,15	11,12
Kenneth Tinsley	17,19	13,16	12,12

(b) In STR analysis, selected STR markers in a DNA sample are amplified by PCR, and the PCR products are separated by electrophoresis. The procedure reveals how many repeats are present for each STR locus in the sample. An individual has two alleles per STR locus, each with a certain number of repeats. This table shows the number of repeats for three STR markers in three samples: from semen found on the victim, from Washington, and from another man (Kenneth Tinsley), who was in prison because of an unrelated conviction. These and other STR data (not shown) exonerated Washington and led Tinsley to plead guilty to the murder.

Environmental Cleanup

Increasingly, the diverse abilities of certain microorganisms to transform chemicals are being exploited for environmental cleanup. If the growth needs of such microorganisms make them unsuitable for direct use, scientists can now transfer the genes for their valuable metabolic capabilities into other microorganisms, which can then be used to treat environmental problems. For example, many bacteria can extract heavy metals, such as copper, lead, and nickel, from their environments and incorporate the metals into compounds such as copper sulfate or lead sulfate, which are readily recoverable. Genetically engineered microorganisms may become important in both mining (especially as ore reserves are depleted) and cleaning up highly toxic mining wastes. Biotechnologists are also trying to engineer microorganisms that can degrade chlorinated hydrocarbons and other harmful compounds. These microorganisms could be used in wastewater treatment plants or by manufacturers before the compounds are ever released into the environment.

Agricultural Applications

Scientists are working to learn more about the genomes of agriculturally important plants and animals. For a number of years, they have been using DNA technology in an effort to improve agricultural productivity. The selective breeding of

both livestock (animal husbandry) and crops has exploited naturally occurring mutations and genetic recombination for thousands of years.

As we described earlier, DNA technology enables scientists to produce transgenic animals, which speeds up the selective breeding process. The goals of creating a transgenic animal are often the same as the goals of traditional breeding—for instance, to make a sheep with better quality wool, a pig with leaner meat, or a cow that will mature in a shorter time. Scientists might, for example, identify and clone a gene that causes the development of larger muscles (muscles make up most of the meat we eat) in one breed of cattle and transfer it to other cattle or even to sheep. However, health problems are not uncommon among farm animals carrying genes from other species, and modification of the animal's own genes using the CRISPR-Cas9 system will likely emerge as a more useful technique. Animal health and welfare are important issues to consider when genetically altering animals.

Agricultural scientists have already endowed a number of crop plants with genes for desirable traits, such as delayed ripening and resistance to spoilage, disease, and drought. Modifications can also add value to food crops, giving them a longer shelf life or improved flavor or nutritional value. For many plant species, a single tissue cell grown in culture can give rise to an adult plant. Thus, genetic manipulations can be performed on an ordinary somatic cell and the cell then can be used to generate a plant with new traits.

Genetic engineering is rapidly replacing traditional plant-breeding programs, especially for useful traits, such as herbicide or pest resistance, determined by one or a few genes. Crops engineered with a bacterial gene making the plants resistant to an herbicide can grow while weeds are destroyed, and genetically engineered crops that can resist destructive insects reduce the need for chemical insecticides.

The Food and Agriculture Organization of the United Nations has predicted that we will need 70% more food by the year 2050 than the planet is currently producing. One novel approach is being taken by researchers working on the international C_4 Rice Project. The common form of rice, a global food staple, uses the C_3 form of photosynthesis (see Concept 10.5). The aim of researchers engaged in this project is to genetically engineer a strain of rice that can use C_4 photosynthesis, which is more efficient. In 2017, the first step was accomplished when they were able to genetically engineer a corn gene into rice plants.

Safety and Ethical Questions Raised by DNA Technology

Early concerns about potential dangers associated with recombinant DNA technology and genetic engineering focused on the possibility that hazardous new pathogens might be created. What might happen, for instance, if in a research study cancer cell genes were transferred into bacteria or viruses? To guard against such rogue microorganisms, scientists developed a set of guidelines that were adopted as formal government regulations in the United States and some other countries. One safety measure is a set of strict laboratory procedures designed to prevent engineered microorganisms from either infecting researchers or accidentally leaving the laboratory. In addition, strains of microorganisms to be used in recombinant DNA experiments are genetically crippled to ensure that they cannot survive outside the laboratory. Finally, certain obviously dangerous experiments have been banned.

Today, most public concern about possible hazards centers not on recombinant microorganisms but on **genetically modified organisms (GMOs)** used as food. A GMO is a transgenic organism that has acquired one or more genes from another species or from another variety of the same species. Some salmon, for example, have been genetically modified by addition of a more active salmon growth hormone gene. However, the majority of the GMOs that contribute to our food supply are not animals, but crop plants.

GM crops are widespread in the United States, Argentina, and Brazil; together, these countries account for over 80% of the world's acreage devoted to such crops. In the United States, most corn, soybean, and canola crops are genetically modified, and a recent law requires labeling of GM products. The same foods are an ongoing subject of controversy in Europe, where the GM revolution has been met with strong opposition. Many Europeans are concerned about the safety of GM foods and the possible environmental consequences of growing GM plants. Although a small number of GM crops have been grown on European soil, the European Union established a comprehensive legal framework regarding GMOs in 2015. Among other regulations, individual member states may ban either the growing or importing of GM crops, which must be clearly labeled. The high degree of consumer distrust in Europe makes the future of GM crops there uncertain.

Advocates of a cautious approach toward GM crops fear that transgenic plants might pass their new genes to close relatives in nearby wild areas. We know that lawn and crop grasses, for example, commonly exchange genes with wild relatives via pollen transfer. If crop plants carrying genes for resistance to herbicides, diseases, or insect pests pollinated wild ones, the offspring might become "super weeds" that are very difficult to control. Another worry involves possible risks to human health from GM foods. Some people fear that the protein products of transgenes might lead to allergic reactions. Although there is some evidence that this could happen, advocates claim that these proteins could be tested in advance to avoid producing ones that cause allergic reactions. (For further discussion of plant biotechnology and GM crops, see Concept 38.3.)

Today, governments and regulatory agencies throughout the world are grappling with how to facilitate the use of biotechnology in agriculture, industry, and medicine while ensuring that new products and procedures are safe. In the United States, such applications of biotechnology must be evaluated for potential risks by various regulatory agencies, including the Food and Drug Administration, the Environmental Protection Agency, the National Institutes of Health, and the Department of Agriculture. Meanwhile, these same agencies and the public must consider the ethical implications of biotechnology.

Advances in biotechnology have allowed us to obtain complete genome sequences for humans and many other species, providing a vast treasure trove of information about genes. We can ask how certain genes differ from species to species, as well as how genes and, ultimately, entire genomes have evolved. (These are the subjects of Chapter 21.) At the same time, the increasing speed and falling cost of sequencing the genomes of individuals are raising significant ethical questions. Who should have the right to examine someone else's genetic information? How should that information be used? Should a person's genome be a factor in determining eligibility for a job or insurance? Ethical considerations, as well as concerns about potential environmental and health hazards, will likely slow some applications of biotechnology. There is always a danger that too much regulation will stifle basic research and its potential benefits. On the other hand, genetic engineering—especially gene editing with the CRISPR-Cas9 system—enables us to profoundly and rapidly alter species that have been evolving for millennia. A good example is the potential use of a gene drive that would eliminate the ability of mosquito species to carry diseases or even eradicate certain mosquito species. There would probably be health benefits to this approach, at least initially, but unforeseen problems could easily arise. Given the tremendous power of DNA technology, we must proceed with humility and caution.

Mastering Biology **Interview with David Suzuki: Exploring ethical issues related to DNA technology**

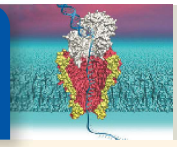


CONCEPT CHECK 20.4

1. What is the advantage of using stem cells for gene therapy or gene editing?
2. List at least three different properties that have been acquired by crop plants via genetic engineering.
3. **WHAT IF?** As a physician, you have a patient with symptoms that suggest a hepatitis A infection, but you have not been able to detect viral proteins in the blood. Knowing that hepatitis A is an RNA virus, what lab tests could you perform to support your diagnosis? Explain the results that would support your hypothesis.

For suggested answers, see Appendix A.

20 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

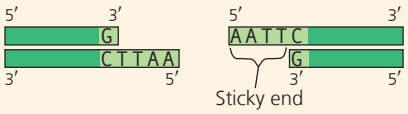
SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 20.1

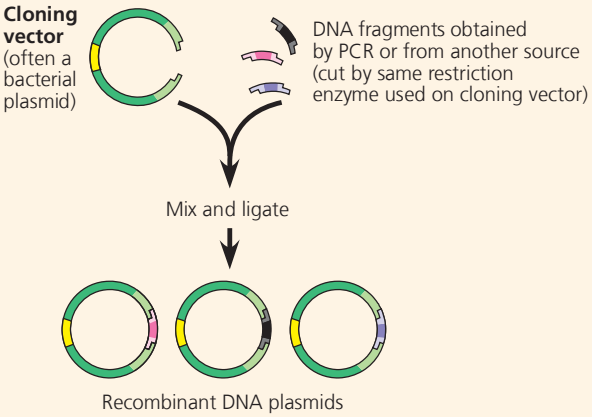
DNA sequencing and DNA cloning are valuable tools for genetic engineering and biological inquiry *(pp. 416–423)*

- **Nucleic acid hybridization**, the base pairing of one strand of a nucleic acid to the complementary sequence on a strand from another nucleic acid molecule, is widely used in **DNA technology**.
- **DNA sequencing** can be carried out using the dideoxy sequencing method in automated sequencing machines.
- Fast and inexpensive next-generation (high-throughput) techniques for sequencing DNA are based on sequencing by synthesis: DNA polymerase is used to synthesize a stretch of DNA using a single-stranded template, and the order in which nucleotides are added reveals the sequence. Third-generation sequencing methods, including nanopore technology, sequence long DNA molecules one at a time.
- **Gene cloning** (or **DNA cloning**) produces multiple copies of a gene (or DNA segment) that can be used to manipulate and analyze DNA and to produce useful new products or organisms with beneficial traits.
- In **genetic engineering**, bacterial **restriction enzymes** are used to cut DNA molecules within short, specific nucleotide sequences (**restriction sites**), yielding a set of double-stranded **restriction fragments** with single-stranded **sticky ends**:



- The sticky ends on restriction fragments from one DNA source can base-pair with complementary sticky ends on fragments from other DNA molecules. Sealing the base-paired fragments with DNA ligase produces **recombinant DNA molecules**.
- DNA restriction fragments of different lengths can be separated by **gel electrophoresis**.
- The **polymerase chain reaction (PCR)** can amplify (produce many copies of) a specific target segment of DNA, using primers that bracket the desired sequence and a heat-resistant DNA polymerase.

- To clone a eukaryotic gene:



  Recombinant **plasmids** are returned to host cells, each of which divides to form a clone of cells.
- Expressing cloned eukaryotic genes in bacterial host cells poses several technical difficulties. The use of cultured eukaryotic cells as host cells, coupled with appropriate **expression vectors**, helps avoid these problems.

? *Describe how the process of gene cloning results in a cell clone containing a recombinant plasmid.*

CONCEPT 20.2

Biologists use DNA technology to study gene expression and function *(pp. 423–428)*

- Several techniques use hybridization of a **nucleic acid probe** to detect the presence of specific mRNAs.
- ***In situ* hybridization** and **RT-PCR** can detect the presence of a given mRNA in a tissue or an RNA sample, respectively.
- Sets of genes co-expressed by a group of cells can be detected by **RNA sequencing (RNA-seq)**—sequencing the **cDNAs** corresponding to mRNAs from the cells. **DNA microarrays** are also used for this purpose.
- For a gene of unknown function, experimental inactivation of the gene (a gene knockout) and observation of the resulting phenotypic effects can provide clues to its function. The CRISPR-Cas9 system allows researchers to edit genes in living cells in a specific, desired way. The new alleles can be altered so that they are inherited in a biased way through a population (**gene drive**).
- In humans, **genome-wide association studies** identify and use **single nucleotide polymorphisms (SNPs)** as genetic markers for alleles that are associated with particular conditions.

? *What useful information is obtained by detecting expression of specific genes?*

CONCEPT 20.3

Cloned organisms and stem cells are useful for basic research and other applications *(pp. 428–432)*

- The question of whether all the cells in an organism have the same genome prompted the first attempts at organismal cloning.
- Single differentiated cells from plants are often **totipotent**: capable of generating all the tissues of a complete new plant.
- Transplantation of the nucleus from a differentiated animal cell into an enucleated egg can sometimes give rise to a new animal.
- Certain embryonic **stem cells** (ES cells) from animal embryos and particular adult stem cells from adult tissues can reproduce and differentiate both in the lab and in the organism, offering the potential for medical use. ES cells are **pluripotent** but difficult to acquire. Induced pluripotent stem (iPS) cells resemble ES cells in their capacity to differentiate; they can be generated by reprogramming differentiated cells. iPS cells hold promise for medical research and regenerative medicine.

? *Describe how, using mice, a researcher could carry out (1) organismal cloning, (2) production of ES cells, and (3) generation of iPS cells, focusing on how the cells are reprogrammed. (The procedures are basically the same in humans and mice.)*

CONCEPT 20.4

The practical applications of DNA-based biotechnology affect our lives in many ways *(pp. 433–439)*

- DNA technology, including the analysis of genetic markers such as SNPs, is increasingly being used in the diagnosis of genetic and other diseases and in personal genome analysis. Personalized medicine offers potential for an individual minimizing their known risk for a disease, as well as better treatment of genetic disorders or cancers. **Gene therapy** or gene editing with the CRISPR-Cas9 system may also lead to permanent cures. DNA technology is used with cell cultures in the large-scale production of protein hormones and other proteins with therapeutic uses. Some therapeutic proteins are being produced in **transgenic** "pharm" animals.
- Analysis of genetic markers such as **short tandem repeats (STRs)** in DNA isolated from tissue or body fluids found at crime scenes leads to a **genetic profile**. Use of genetic profiles can provide definitive evidence that a suspect is innocent or strong evidence of guilt. Such analysis is also useful in parenthood disputes and in identifying the remains of crime victims.
- Genetically engineered microorganisms can be used to extract minerals from the environment or degrade various types of toxic waste materials.
- The aims of developing transgenic plants and animals are to improve agricultural productivity and food quality.
- The potential benefits of genetic engineering must be carefully weighed against the potential for harm to humans or the environment.

? *What factors affect whether a given genetic disease would be a good candidate for successful gene therapy?*

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. In DNA technology, the term *vector* can refer to
   (A) the enzyme that cuts DNA into restriction fragments.
   (B) the sticky end of a DNA fragment.
   (C) a SNP marker.
   (D) a plasmid used to transfer DNA into a living cell.
2. Which of the following tools of DNA technology is incorrectly paired with its use?
   (A) electrophoresis—separation of DNA fragments
   (B) DNA ligase—cutting DNA, creating sticky ends of restriction fragments
   (C) DNA polymerase—polymerase chain reaction to amplify sections of DNA
   (D) reverse transcriptase—production of cDNA from mRNA

3. Plants are more readily manipulated by genetic engineering than are animals because
   (A) plant genes do not contain introns.
   (B) more vectors are available for transferring recombinant DNA into plant cells.
   (C) a somatic plant cell can often give rise to a complete plant.
   (D) plant cells have larger nuclei.

4. A paleontologist has recovered a bit of tissue from the 400-year-old preserved skin of an extinct dodo (a bird). To compare a specific region of the DNA from a sample with DNA from living birds, which of the following would be most useful for increasing the amount of dodo DNA available for testing?
   (A) SNP analysis
   (B) polymerase chain reaction (PCR)
   (C) electroporation
   (D) gel electrophoresis

Levels 3-4: Applying/Analyzing

5. Which of the following is true of cDNA produced using human brain tissue as the starting material?
   (A) The procedure to make it requires amplification by the polymerase chain reaction.
   (B) It is produced from pre-mRNA using reverse transcriptase.
   (C) It can be labeled and used as a probe to detect genes expressed in the brain.
   (D) It includes the introns of the pre-mRNA.

6. Expression of a cloned eukaryotic gene in a bacterial cell involves many challenges. The use of mRNA and reverse transcriptase is part of a strategy to solve the problem of
   (A) post-transcriptional processing.
   (B) post-translational processing.
   (C) nucleic acid hybridization.
   (D) restriction fragment ligation.

7. Which of the following sequences in double-stranded DNA is most likely to be recognized as a cutting site for a restriction enzyme?
   (A) AAGG
       TTCC
   (B) GGCC
       CCGG
   (C) ACCA
       TGGT
   (D) AAAA
       TTTT

Levels 5-6: Evaluating/Creating

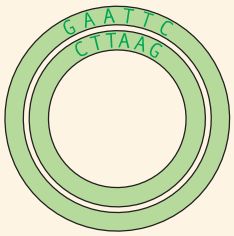
8. **MAKE CONNECTIONS** Imagine you want to study one of the human crystallins, proteins present in the lens of the eye (see Figure 1.8). To obtain a sufficient amount of the protein of interest, you decide to clone the gene that codes for it. Assume you know the sequence of this gene. Explain how you would go about this.

9. **MAKE CONNECTIONS** Looking at Figure 20.15, what does it mean for a SNP to be "linked" to a disease-associated allele? How does this allow the SNP to be used as a genetic marker? (See Concept 15.3.)

10. **DRAW IT** You are cloning an aardvark gene, using a bacterial plasmid as a vector. The green diagram shows the plasmid, which contains the restriction site for the enzyme used in Figure 20.5. Above the plasmid is a segment of linear aardvark DNA that was synthesized using PCR. Diagram your cloning procedure, and show what would happen to these two molecules during each step. Use one color for the aardvark DNA and its bases and another color for those of the plasmid. Label each step and all 5′ and 3′ ends.

5′ GAATTCTAAAGCGCTTATGAATTC 3′
3′ CTTAAGATTTCGCGAATACTTAAG 5′

Aardvark DNA



Plasmid

11. **EVOLUTION CONNECTION** Ethical considerations aside, if DNA-based technologies became widely used, discuss how they might change the way evolution proceeds, as compared with the natural evolutionary mechanisms that have operated for the past 4 billion years.

12. **SCIENTIFIC INQUIRY** You hope to study a gene that codes for a neurotransmitter protein produced in human brain cells. You know the amino acid sequence of the protein. Explain how you might (a) identify what genes are expressed in a specific type of brain cell, (b) identify (and isolate) the neurotransmitter gene, (c) produce multiple copies of the gene for study, and (d) produce large quantities of the neurotransmitter for evaluation as a potential medication.

13. **WRITE ABOUT A THEME: INFORMATION** In a short essay (100–150 words), discuss how the genetic basis of life plays a central role in biotechnology.

14. **SYNTHESIZE YOUR KNOWLEDGE**



The water in the Yellowstone National Park hot springs shown here is around 160°F (70°C). Biologists assumed that no species of organisms could live in water above about 130°F (55°C), so they were surprised to find several species of bacteria there, now called *thermophiles* ("heat-lovers"). You've learned in this chapter how an enzyme from one species, *Thermus aquaticus*, made feasible one of the most important DNA-based techniques used in labs today. Identify the enzyme, and indicate the value of its being isolated from a thermophile. Suggest other reasons why enzymes from this bacterium (or other thermophiles) might also be valuable.

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
How can CRISPR-Cas9 be used to create a gene drive?
Go to "Can We Handle the Power of CRISPR?"
at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.

21 Genomes and Their Evolution

KEY CONCEPTS

21.1 The Human Genome Project fostered development of faster, less expensive sequencing techniques *p. 443*

21.2 Scientists use bioinformatics to analyze genomes and their functions *p. 444*

21.3 Genomes vary in size, number of genes, and gene density *p. 448*

21.4 Multicellular eukaryotes have a lot of noncoding DNA and many multigene families *p. 450*

21.5 Duplication, rearrangement, and mutation of DNA contribute to genome evolution *p. 454*

21.6 Comparing genome sequences provides clues to evolution and development *p. 459*

Study Tip

Write genomic analysis questions: As you go through the chapter, make a table of genomic analyses that are discussed and write down some questions they can address. You can include questions not covered in the text.

Genomic projects and websites	Questions that can be addressed
Human Genome Project	How much DNA is in the human genome? How much of the human genome codes for proteins? What functions are carried out by other regions of the genome?
ENCODE project	

Go to Mastering Biology

For Students (in eText and Study Area)
- Get Ready for Chapter 21
- HHMI Animation: Shotgun Sequencing
- Animation: The Human Genome Project: The Genes on Human Chromosome 17

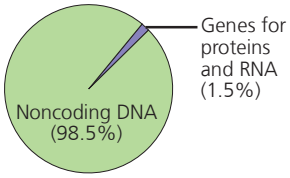
For Instructors to Assign (in Item Library)
- Tutorial: Scientific Skills Exercise: Reading an Amino Acid Sequence Identity Table
- Tutorial: Using BLAST: What Can a Protein Sequence Reveal About Cancer?



Figure 21.1 The elephant shark (*Callorhinchus milii*) looks vaguely prehistoric and has been called a "living fossil." In fact, it has the slowest evolving genome of any vertebrate sequenced so far. Comparing rates of genome change in different species provides insights into the evolutionary past.

What are some questions that can be explored by sequencing and comparing genomes?

What are the **functions** of the human genome?



How do genomes differ in **number of genes?**



Escherichia coli 4,400 genes

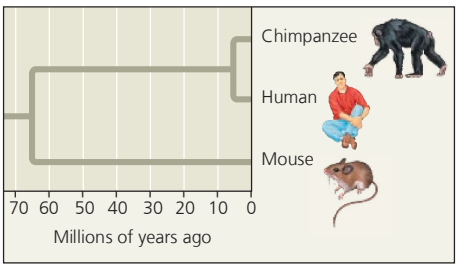


Homo sapiens 21,300 genes



Zea mays (corn) 32,000 genes

What do gene sequences tell us about **evolutionary relationships** between species?



How do genomes **evolve** over time?



Elephant shark: slower-evolving genome



Tiger tail sea horse: faster-evolving genome

CONCEPT 21.1

The Human Genome Project fostered development of faster, less expensive sequencing techniques

Upon sequencing the genome of any species, scientists can study whole sets of genes and their interactions, an approach called **genomics**. The sequencing efforts that feed this approach generate enormous volumes of data. The necessity of dealing with this ever-increasing flood of information has energized the field of **bioinformatics**, the application of computational methods to store and analyze biological data.

The sequencing of the human genome, an ambitious undertaking, officially began as the **Human Genome Project** in 1990. Organized by an international, publicly funded consortium of scientists at universities and research institutes, the project involved 20 large sequencing centers in six countries plus a host of other labs working on smaller parts of the project.

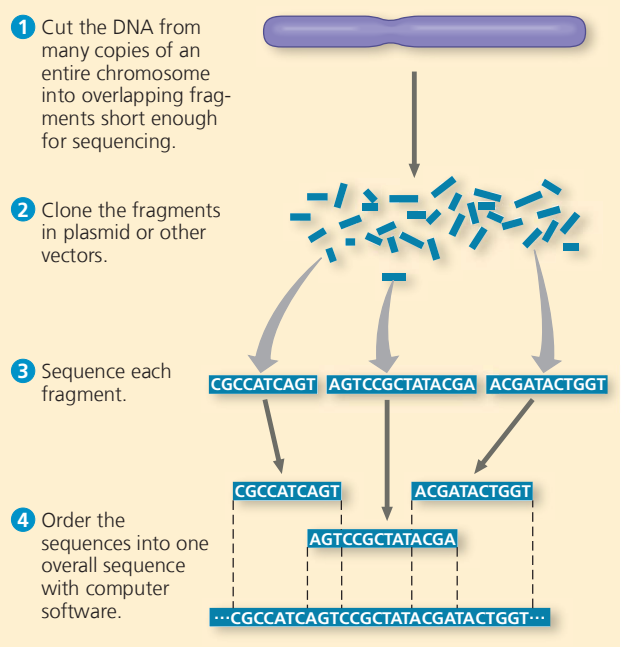
After the human genome sequence was largely completed in 2003, the sequence of each chromosome was analyzed and described in a series of papers, the last of which covered chromosome 1 and was published in 2006. The sequence is over 99% complete—the figure is not 100% because numerous gaps remain, usually in regions with much repetitive DNA that are difficult to sequence. The sequenced DNA was pooled from a few individuals; scientists reviewed the results and agreed on a **reference genome**, a full sequence that best represents the genome of a species. The human reference genome is continuously being revised and re-released.

The ultimate goal in mapping any genome is to determine the complete nucleotide sequence of each chromosome. For the human genome, this was accomplished by scientists using sequencing machines and the dideoxy chain termination method mentioned in Concept 20.1. Two approaches complemented each other in this endeavor. The initial approach was a methodical one that ordered each fragment based on earlier genetic mapping of the human genome. In 1998, however, molecular biologist J. Craig Venter set up a company (Celera Genomics) and declared his intention to sequence the entire human genome using an alternative strategy. The **whole-genome shotgun approach** starts with the cloning and sequencing of DNA fragments from randomly cut DNA. Powerful computer programs then assemble the resulting very large number of overlapping short sequences into a single continuous sequence **(Figure 21.2)**. The whole-genome shotgun approach is widely used today.

A major thrust of the Human Genome Project was the development of technology for faster sequencing (see Concept 20.1). Improvements over the years chipped away at each time-consuming step, enabling the rate of sequencing to accelerate impressively: Whereas a productive lab could typically sequence 1,000 base pairs a day in the 1980s, by the year 2000 each research center working on the Human Genome Project was sequencing 1,000 base pairs *per second*. Currently, the most widely used automated "next-generation" sequencing machines can sequence nearly 35 million base pairs per second (see Figure 20.3). Also, because of the sensitivity of these techniques, the DNA can be sequenced directly, either in fragments or as a single whole molecule; the cloning step (2 in Figure 21.2) is unnecessary. Methods that can analyze biological materials very rapidly and produce enormous volumes of data are said to be "high-throughput." Sequencing machines using rapid sequencing techniques are an example of high-throughput devices.

Along with massive increases in sequencing speed, the cost of sequencing entire genomes has plummeted. Whereas sequencing the first human genome took 13 years and cost between $500 million and $1 billion, by 2007 the same task took four months at a cost of $1 million. In 2019, the fastest machines are able to sequence the complete genomes of 48 individuals in 44 hours for under $1,000 per genome.

▼ Figure 21.2 Whole-genome shotgun approach to sequencing. In this approach, random DNA fragments are cloned (see Figure 20.4), sequenced, and then ordered relative to each other.



VISUAL SKILLS *The fragments in step 2 of this figure are depicted as scattered, rather than arranged in an ordered array. How does this depiction reflect the approach?*

Mastering Biology **HHMI Animation: Shotgun Sequencing**

These technological advances have also facilitated an approach called **metagenomics** (from the Greek *meta*, beyond), in which DNA from an entire community of species (a *metagenome*) is collected from an environmental sample and sequenced. Again, computer software sorts out the partial sequences and assembles them into the individual species' partial of complete genomes. An advantage of this technique is the ability to sequence the DNA of mixed microbial populations, which eliminates the need to culture each species separately in the lab, a challenge that has limited the study of microorganisms. So far, this approach has been applied to communities found in environments as diverse as the human intestine and extreme habitats like thermal springs where the temperature exceeds 80°C.

Mastering Biology
BBC Video: Collecting New Genes to Create New Life

At first glance, genome sequences of humans and other organisms are simply dry lists of nucleotide bases—millions of A's, T's, C's, and G's in mind-numbing succession. Making sense of this massive amount of data has called for new analytical approaches, which we discuss next.

CONCEPT CHECK 21.1

1. Describe the whole-genome shotgun approach.

For suggested answers, see Appendix A.

CONCEPT 21.2

Scientists use bioinformatics to analyze genomes and their functions

Each of the 20 or so sequencing centers around the world working on the Human Genome Project churned out voluminous amounts of DNA sequence day after day. As the data began to accumulate, the need to coordinate efforts to keep track of all the sequences became clear. Thanks to the foresight of research scientists and government officials involved in the Human Genome Project, its goals included establishing centralized databases and refining analytical software, all made readily accessible on the Internet.

Mastering Biology
HHMI Video: Leading Edge Bioinformatics

Centralized Resources for Analyzing Genome Sequences

Making bioinformatics resources available to researchers worldwide and speeding up the dissemination of information served to accelerate progress in DNA sequence analysis. The National Library of Medicine (NLM) and the National Institutes of Health (NIH) maintain the National Center for Biotechnology Information (NCBI), which today maintains a website (www.ncbi.nlm.nih.gov), which has extensive bioinformatics resources. On this site are links to databases, software, and a wealth of information about genomics and related topics. Similar websites have also been established by three genome centers with which the NCBI collaborates: the European Molecular Biology Laboratory, the DNA Data Bank of Japan, and BGI (formerly known as the Beijing Genomics Institute) in Shenzhen, China. These large, comprehensive websites are complemented by others maintained by individual or small groups of laboratories. Smaller websites often provide databases and software designed for a narrower purpose, such as studying genetic and genomic changes in one particular type of cancer.

The NCBI database of sequences is called GenBank. As of August 2019, it included the sequences of 214 million fragments of genomic DNA, totaling 366 billion base pairs! GenBank is constantly updated, and the amount of data it contains increases rapidly. Any sequence in the database can be retrieved and analyzed using software from the NCBI website or elsewhere.

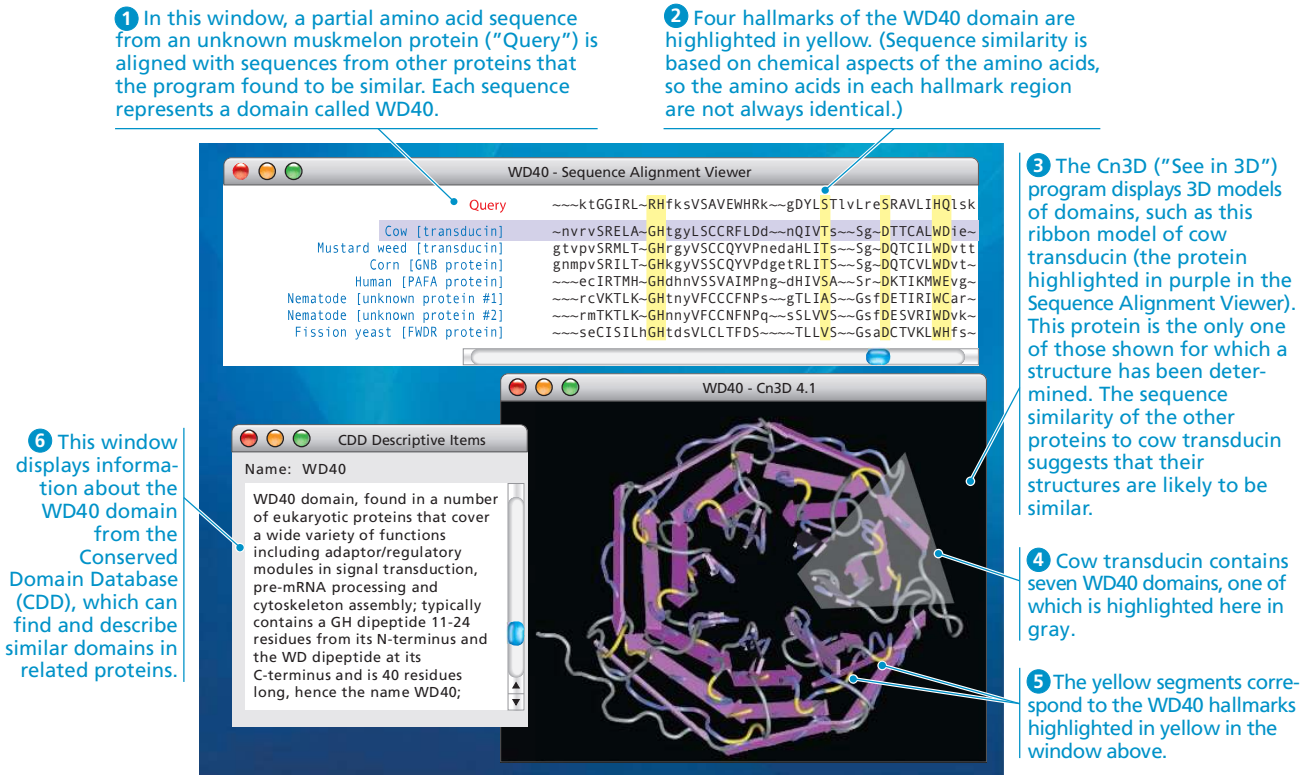
One very widely used software program available on the NCBI website, called BLAST (Basic Local Alignment Search Tool), allows the user to compare a DNA sequence with every sequence in GenBank, base by base. A researcher might search for similar regions in other genes of the same species or among the genes of other species. Another program allows comparison of protein sequences. A third program can search any protein sequence for *conserved* (common) stretches of amino acids (domains) for which a function is known or suspected, and it can show a three-dimensional model of the domain alongside other relevant information **(Figure 21.3)**. There are even various software programs that can align and compare a collection of sequences, either nucleic acids or polypeptides, and diagram them in the form of an evolutionary tree based on the sequence relationships. (One such diagram is shown in Figure 21.17.)

Two research institutions, Rutgers University and the University of California, San Diego, also maintain a worldwide database of all three-dimensional protein structures that have been experimentally determined, called the Protein Data Bank (www.wwpdb.org). These structures can be rotated by the viewer to show all sides of the protein. Throughout this book, you'll find images of protein structures that have been obtained from the Protein Data Bank.

There is a vast array of resources available for researchers anywhere in the world to use free of charge. Let us now consider the types of questions scientists can address using these resources.

Instructors: BLAST Data Analysis Tutorials, which teach students how to work with real data from the BLAST database, can be assigned in **Mastering Biology**.

▼ Figure 21.3 National Center for Biotechnology Information (NCBI) website. Some results are shown from a search for regions of proteins similar to an amino acid sequence in a muskmelon protein.



Identifying Protein-Coding Genes and Understanding Their Functions

Using available DNA sequences, geneticists can study genes directly, rather than taking the classical genetic approach, which requires determining the function of an unknown gene from the phenotype. But this more recent approach poses a new challenge: What does the gene actually do? Given a long DNA sequence from a database like GenBank, scientists aim to identify all protein-coding genes in the sequence and ultimately their functions. This process, called **gene annotation**, uses three lines of evidence to identify a gene.

First, computers are utilized in a search for patterns that indicate the presence of genes. The usual approach is to use software to scan the stored sequences for those that represent transcriptional and translational start and stop signals, RNA-splicing sites, and other telltale signs of protein-coding genes, such as promoter sequences. The software also looks for certain short sequences that specify known mRNAs. Thousands of such sequences, called *expressed sequence tags*, or *ESTs*, have been collected from cDNA sequences and are cataloged in computer databases. This type of analysis identifies sequences that may turn out to be previously unknown protein-coding genes.

Although the identities of about half of the human genes were known before the Human Genome Project began, the other genes, previously unknown, were revealed by DNA sequence analysis. Once such suspected genes are identified, the second step is to obtain clues about their identities and functions. Software is used to compare their sequences with those of known genes from other organisms. Due to redundancy in the genetic code, the DNA sequence itself may vary more among species than the protein sequence does. Thus, scientists interested in proteins often compare the predicted amino acid sequence of a protein to that of other proteins. The final step is to confirm the identities of these genes using RNA-seq (see Figure 20.12) or some other method to show that the relevant RNA is actually expressed from the proposed gene.

Mastering Biology **Animation: The Human Genome Project: Genes on Human Chromosome 17**

Sometimes a newly identified sequence will match, at least partially, the sequence of a gene or protein in another species whose function is well known. For example, a plant researcher working on signaling pathways in the muskmelon would be excited to see that a partial amino acid sequence from a gene she had identified matched sequences in other

species encoding a functional part of a protein called a WD40 domain (see Figure 21.3). WD40 domains are present in many eukaryotic proteins and are known to function in signal transduction pathways. Alternatively, a new gene sequence might be similar to a previously encountered sequence whose function is still unknown. Another possibility is that the sequence is entirely unlike anything ever seen before. This was true for about a third of the genes of *Escherichia coli* when its genome was sequenced. In such cases, protein function is usually deduced through a combination of biochemical and functional studies. The biochemical approach aims to determine the three-dimensional structure of the protein as well as other attributes, such as potential binding sites for other molecules. Functional studies usually involve *knocking out* (blocking or disabling) the gene in an organism to see how the phenotype is affected. The CRISPR-Cas 9 system, described in Figure 17.28, is an example of an experimental technique used to block gene function.

Understanding Genes and Gene Expression at the Systems Level

The impressive computational power provided by the tools of bioinformatics allows the study of whole sets of genes and their interactions, as well as the comparison of genomes from different species. Genomics is a rich source of new insights into fundamental questions about genome organization, regulation of gene expression, embryonic development, and evolution.

One informative approach was taken by a long-term research project called ENCODE (Encyclopedia of DNA Elements), which began in 2003. The aim of the project is to learn everything possible about the functionally important elements in the human genome, at first using multiple experimental techniques on different types of cultured cells. Investigators sought to identify protein-coding genes and genes for noncoding RNAs, along with sequences that regulate gene expression, such as enhancers and promoters. In addition, they extensively characterized DNA and histone modifications and chromatin structure—features termed "epigenetic" since they affect gene expression without changing the sequence of nucleotide bases (see Concept 18.3). The second phase of the project, involving more than 440 scientists in 32 research groups, culminated in 2012 with the simultaneous publication of 30 papers describing over 1,600 large data sets. Now in its fourth phase, the ENCODE project is expanding its analysis of the human genome and that of the mouse (a mammalian model organism), seeking to identify regulatory and other important sequences. The considerable power of ENCODE has been that it provides the opportunity to compare results from specific projects with each other, yielding a much richer picture of the human and mouse genomes.

Perhaps the most striking finding was that about 75% of the human genome is transcribed at some point in at least one of the cell types studied, even though less than 2% codes for proteins. Furthermore, biochemical functions have been assigned to DNA elements making up at least 80% of the genome. To learn more about the different types of functional elements, parallel projects analyzed in a similar way the genomes of two model organisms, the soil nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. Because genetic and biochemical experiments using DNA technology can be performed on these species, testing the activities of potentially functional DNA elements in their genomes is expected to illuminate the workings of the human genome.

Because the ENCODE project analyzed cells in culture, its potential for clinical applications was limited. A related project called the Roadmap Epigenomics Project set out to characterize the *epigenome*—the epigenetic features of the genome of hundreds of human cell types and tissues. The aim was to focus on the epigenomes of stem cells, normal tissues from mature adults, and relevant tissues from individuals with diseases such as cancer and neurodegenerative and autoimmune disorders. In 2015, a series of papers reported on the results from 111 tissues. One of the most useful findings was that the original tissue in which a cancer arose can be identified in cells of a secondary tumor based on characterization of their epigenomes.

Systems Biology

The scientific progress resulting from sequencing genomes and studying large sets of genes has encouraged scientists to attempt similar systematic studies of sets of proteins and their properties (such as their abundance, chemical modifications, and interactions), an approach called **proteomics**. (A **proteome** is the entire set of proteins expressed by a cell or group of cells.) Proteins, not the genes that encode them, carry out most of the activities of the cell. Therefore, if we are to understand the functioning of cells and organisms, we must study when and where proteins are produced in an organism, as well as how they interact in networks.

Genomics and proteomics enable molecular biologists to approach the study of life from an increasingly integrated perspective. Using the tools we have described, biologists have begun to compile catalogs of genes and proteins, listing all the "parts" that contribute to the operation of cells, tissues, and organisms. With such catalogs in hand, researchers have shifted their attention from the individual parts—genes and proteins—to their functional integration in biological systems. As you may recall, Concept 1.1 discussed this approach, called **systems biology**, which aims to model the dynamic behavior of whole biological systems based on the study of

the interactions among the system's parts. Because of the vast amounts of data generated in these types of studies, advances in computer technology and bioinformatics are crucial to studying systems biology.

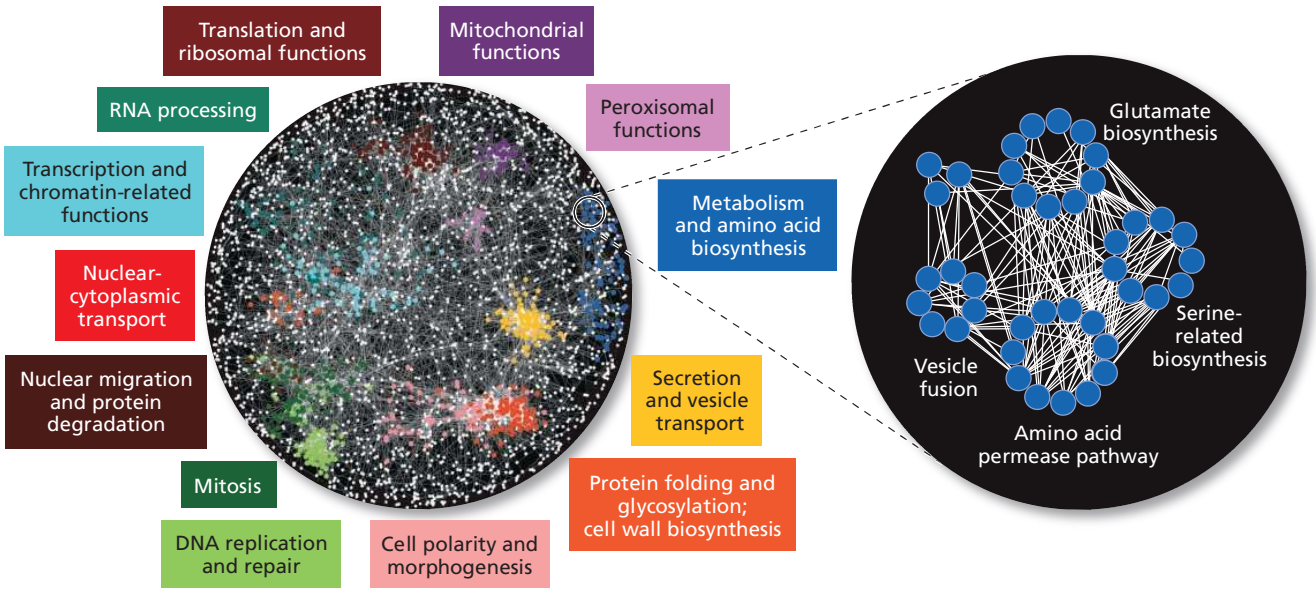
One important use of the systems biology approach is to define gene and protein interaction networks. To map the protein interaction network in the yeast *Saccharomyces cerevisiae*, for instance, researchers used sophisticated techniques to knock out pairs of genes, one pair at a time, creating doubly mutant cells. They then compared the fitness of each double mutant (based in part on the size of the cell colony it formed) to that predicted from the fitness of each of the two single mutants. The researchers reasoned that if the observed fitness matched the prediction, then the products of the two genes didn't interact with each other, but if the observed fitness was greater or less than predicted, then the gene products interacted in the cell. They then used computer software to build a graphic model by "mapping" the gene products to certain locations in the model, based on the similarity of their interactions. This resulted in the network-like "functional map" of protein interactions shown in **Figure 21.4**. Processing the vast number of protein-protein interactions generated by this experiment and integrating them into the completed map required powerful computers, mathematical tools, and newly developed software.

Application of Systems Biology to Medicine

The Cancer Genome Atlas began in 2007 and culminated in 2018 with publications called the Pan-Cancer Atlas. This project is another example of systems biology in which many interacting genes and gene products are analyzed together as a group. Under the joint leadership of the National Cancer Institute and the NIH, the project aimed to determine how changes in biological systems lead to cancer. A pilot project set out to find all the common mutations in three types of cancer—lung cancer, ovarian cancer, and glioblastoma of the brain—by comparing gene sequences and patterns of gene expression in cancer cells with those in normal cells. Work on glioblastoma confirmed the role of several suspected genes and identified a few previously unknown ones, suggesting possible new targets for therapies.

As high-throughput techniques become more rapid and less expensive, they are increasingly being applied to the problem of cancer. The approach described previously proved so fruitful for those three types of cancer that it was extended to ten other types, chosen because they are common and often lethal in humans, as well as to *metastatic* tumors (see Figure 12.20), those that have dispersed from primary tumors and invaded organs far away in the body. Ninety percent of cancer deaths are primarily caused by metastasis. Results from the study of metastatic tumors, published in 2017, highlighted several key genes whose mutations were frequently found in

▼ Figure 21.4 The systems biology approach to protein interactions. This global protein interaction map shows the likely interactions (lines) among about 4,500 gene products (dots) in *Saccharomyces cerevisiae,* the budding yeast. Dots of the same color represent gene products involved in one of the 13 similarly colored cellular functions listed around the map. The white dots represent proteins that haven't been assigned to any color-coded function. The expanded area shows additional details of one map region where the gene products (blue dots) carry out amino acid biosynthesis, uptake, and related functions.





◀ Figure 21.5 A human gene microarray chip. Tiny spots of DNA arranged in a grid on this silicon wafer represent almost all of the genes in the human genome. Using this chip, researchers can analyze expression patterns for all these genes at the same time (see Figure 20.13).

Mastering Biology
HHMI Animation: Gene Chip Manufacturing

metastases and could be targets for chemotherapy. Overall, the Pan-Cancer Atlas contributed significantly to understanding how, where, and why tumors arise, underscoring the value of an integrative systems biology approach to treating cancer.

In addition to whole-genome sequencing, DNA microarrays on glass or silicon chips **(Figure 21.5)** and, increasingly, RNA-seq (see Figure 20.12) are used to analyze gene expression patterns in patients who have various cancers and other diseases. Analyzing which genes are overexpressed or underexpressed in a particular cancer allows physicians to tailor patients' treatment to their unique genetic makeup and the specifics of their cancers. This approach has been used to characterize subsets of particular cancers, enabling more refined treatments. Breast cancer is one example (see Figure 18.27).

Eventually, medical records may include an individual's DNA sequence, a sort of genetic bar code, with regions highlighted that predispose the person to specific diseases. The use of such sequences for personalized medicine—disease prevention and treatment—has great potential.

Systems biology is a very efficient way to study emergent properties at the molecular level. Novel properties arise at each successive level of biological complexity as a result of the arrangement of building blocks at the underlying level (see Concept 1.1). The more we can learn about the arrangement and interactions of the components of genetic systems, the deeper will be our understanding of whole organisms. The rest of this chapter surveys what we've learned from genomic studies.

CONCEPT CHECK 21.2

1. What role does the Internet play in current genomics and proteomics research?
2. Explain the advantage of the systems biology approach to studying cancer versus the approach of studying a single gene at a time.
3. **MAKE CONNECTIONS** The ENCODE pilot project found that at least 75% of the genome is transcribed into RNAs, far more than could be accounted for by protein-coding genes. Review Concepts 17.3 and 18.3 and suggest some roles that these RNAs might play.
4. **MAKE CONNECTIONS** In Concept 20.2, you learned about genome-wide association studies. Explain how these studies use the systems biology approach.

For suggested answers, see Appendix A.

CONCEPT 21.3

Genomes vary in size, number of genes, and gene density

The sequences of thousands of genomes have been completed, with tens of thousands of genomes either in progress or considered permanent drafts (because they require more work than it would be worth to complete them). Among the sequences in progress are roughly 22,000 metagenomes. In the completely sequenced group, about 137,500 are genomes of bacteria, and 1,200 are archaeal genomes. There are 430 completed eukaryotic species, along with 4,500 permanent drafts. Among these are vertebrates, invertebrates, protists, fungi, and plants. Next, we'll discuss what we've learned about genome size, number of genes, and gene density, focusing on general trends.

Genome Size

Comparing the three domains (Bacteria, Archaea, and Eukarya), we find a general difference in genome size between prokaryotes and eukaryotes **(Table 21.1)**. While

Table 21.1 Genome Sizes and Estimated Numbers of Genes*

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
Bacteria			
Haemophilus influenzae	1.8	1,700	940
Escherichia coli	4.6	4,400	950
Archaea			
Archaeoglobus fulgidus	2.2	2,500	1,130
Methanosarcina barkeri	4.8	3,600	750
Eukaryotes			
Saccharomyces cerevisiae (yeast, a fungus)	12	6,300	525
Utricularia gibba (floating bladderwort)	82	28,500	348
Caenorhabditis elegans (nematode)	100	20,100	200
Arabidopsis thaliana (mustard family plant)	120	27,000	225
Drosophila melanogaster (fruit fly)	165	14,000	85
Daphnia pulex (water flea)	200	31,000	155
Zea mays (corn)	2,300	32,000	14
Ailuropoda melanoleuca (giant panda)	2,400	21,000	9
Homo sapiens (human)	3,000	21,300	7
Paris japonica (Japanese canopy plant)	149,000	ND	ND

*Some values given here are likely to be revised as genome analysis continues. Mb = million base pairs; the haploid number is used because it represents a complete set of genetic information. ND = not determined.

there are some exceptions, most bacterial genomes have between 1 and 6 million base pairs (Mb); for example, the genome of *E. coli* has 4.6 Mb. Genomes of archaea are mostly within the size range of bacterial genomes. (Keep in mind, however, that many fewer archaeal genomes have been completely sequenced, so this picture may change.) Eukaryotic genomes tend to be larger: The genome of the single-celled yeast *Saccharomyces cerevisiae* (a fungus) has about 12 Mb, while most animals and plants, which are multicellular, have genomes of at least 100 Mb. There are 165 Mb in the fruit fly genome, while humans have 3,000 Mb, about 500 to 3,000 times as many as a typical bacterium.

Aside from this general difference between prokaryotes and eukaryotes, a comparison of genome sizes among eukaryotes fails to reveal any systematic relationship between genome size and the organism's phenotype. For instance, the range among plants is enormous: The genome of *Paris japonica*, the Japanese canopy plant, contains 149 billion base pairs (149,000 Mb), while that of another plant, *Utricularia gibba*, a bladderwort, contains only 82 Mb. Even more striking, there is a single-celled amoeba, *Polychaos dubium*, whose genome size has been estimated at 670 billion base pairs (670,000 Mb). (This genome has not yet been sequenced.) On a finer scale, comparing two insect species, the cricket (*Anabrus simplex*) genome turns out to have 11 times as many base pairs as the fruit fly (*Drosophila melanogaster*) genome. There is a wide range of genome sizes within the groups of protists, insects, amphibians, and plants and less of a range within mammals and reptiles.

Number of Genes

The number of genes also varies between prokaryotes and eukaryotes: Bacteria and archaea, in general, have fewer genes than eukaryotes. Free-living bacteria and archaea have from 1,500 to 7,500 genes, while the number of genes in eukaryotes ranges from about 5,000 for unicellular fungi (yeasts) to at least 40,000 for some multicellular eukaryotes.

Within the eukaryotes, the number of genes in a species is often lower than expected from considering simply the size of its genome. Looking at Table 21.1, you can see that the genome of the nematode *C. elegans* is 100 Mb in size and contains roughly 20,100 genes. In comparison, the genome of *Drosophila melanogaster* is much bigger (165 Mb) but has only about two-thirds the number of genes—14,000 genes.

Considering an example closer to home, we noted that the human genome contains 3,000 Mb, well over ten times the size of either the *D. melanogaster* or *C. elegans* genome. At the outset of the Human Genome Project, biologists expected somewhere between 50,000 and 100,000 genes to be identified in the completed sequence, based on the number of known human proteins. As the project progressed, the estimate was revised downward several times. There is still debate about the number, but it is much lower, probably around 21,300. This estimate, similar to the number of genes in the nematode *C. elegans*, surprised biologists, who had been expecting many more human genes.

What genetic attributes allow humans (and other vertebrates) to get by with no more genes than nematodes? An important factor is that vertebrate genomes "get more bang for the buck" from their coding sequences because of extensive alternative splicing of RNA transcripts. Recall that this process generates more than one polypeptide from a single gene (see Figure 18.14). A typical human gene contains about ten exons, and an estimated 90% or more of these multi-exon genes are spliced in at least two different ways. Some genes are expressed in hundreds of alternatively spliced forms, others in just two. Scientists have not yet catalogued all of the different forms, but it is clear that the number of different proteins encoded in the human genome far exceeds the proposed number of genes.

Additional polypeptide diversity could result from post-translational modifications such as cleavage or the addition of carbohydrate groups in different cell types or at different developmental stages. Finally, the discovery of miRNAs and other RNAs that play regulatory roles has added a new variable to the mix (see Concept 18.3). Some scientists think that this added level of regulation of some genes, when present, may contribute to greater organismal complexity.

Gene Density and Noncoding DNA

We can take both genome size and number of genes into account by comparing gene density in different species. In other words, we can ask how many genes are in a given length of DNA. When we compare the genomes of bacteria, archaea, and eukaryotes, we see that eukaryotes generally have larger genomes but fewer genes in a given number of base pairs. Humans have hundreds or thousands of times as many base pairs in their genome as most bacteria, as we already noted, but only 5 to 15 times as many genes; thus, gene density is lower in humans (see Table 21.1). Even unicellular eukaryotes, such as yeasts, have fewer genes per million base pairs than bacteria and archaea. Among the genomes that have been sequenced completely, humans and other mammals have the lowest gene density.

In all bacterial genomes studied so far, most of the DNA consists of genes for protein, tRNA, or rRNA; the small amount remaining consists mainly of nontranscribed regulatory sequences, such as promoters. The sequence of nucleotides along a bacterial protein-coding gene is not interrupted by noncoding sequences (introns). In eukaryotic genomes, by contrast, most of the DNA neither encodes protein nor is transcribed into RNA molecules of known

function, and the DNA includes more complex regulatory sequences. In fact, humans have 10,000 times as much noncoding DNA as bacteria. Some of this DNA in multicellular eukaryotes is present as introns within genes. Indeed, introns account for most of the difference in average length between human genes (27,000 base pairs) and bacterial genes (1,000 base pairs).

In addition to introns, multicellular eukaryotes have a vast amount of non-protein-coding DNA between genes. In the next section, we will describe the composition and arrangement of these great stretches of DNA in the human genome.

CONCEPT CHECK 21.3

1. The current best estimate is that the human genome contains around 21,300 genes. However, there is evidence that human cells produce many more than 21,300 different polypeptides. What processes might account for this discrepancy?
2. The Genomes Online Database (GOLD) website of the Joint Genome Institute has information about genome sequencing projects. Scroll through the page at https://gold.jgi.doe.gov/statistics and describe the information you find. What percent of bacterial genome projects have medical relevance?
3. **WHAT IF?** What evolutionary processes might account for prokaryotes having smaller genomes than eukaryotes?

For suggested answers, see Appendix A.

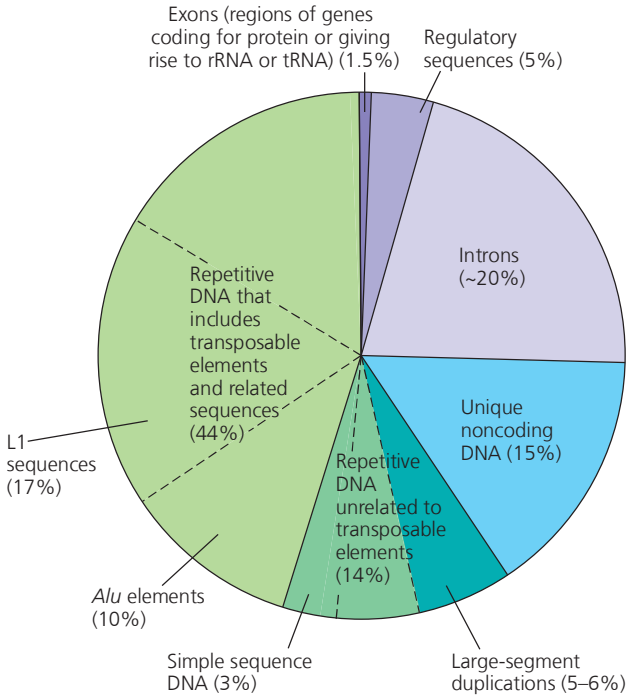
CONCEPT 21.4

Multicellular eukaryotes have a lot of noncoding DNA and many multigene families

We have spent most of our time focusing on genes that code for proteins. Yet the coding regions of these genes and the genes for small RNAs like tRNAs make up a small portion of most multicellular eukaryotic genomes. For example, only a tiny part—about 1.5%—codes for proteins or is transcribed into rRNAs or tRNAs. **Figure 21.6** shows what is known about the makeup of the remaining 98.5% of the genome.

Gene-related regulatory sequences and introns account, respectively, for 5% and about 20%, respectively, of the human genome. The rest, located between functional genes, includes some unique (single-copy) noncoding DNA, such as gene fragments and **pseudogenes**, former genes that have accumulated mutations over a long time and no longer produce functional proteins. (The genes that produce small noncoding RNAs are a tiny percentage of the genome, distributed between the 20% introns and the 15% unique noncoding DNA.) Most of the DNA between functional genes, however, is **repetitive DNA**, which consists of sequences that are present in multiple copies in the genome.

▼ Figure 21.6 Types of DNA sequences in the human genome. The gene sequences that code for proteins or are transcribed into rRNA or tRNA molecules make up only about 1.5% of the human genome (dark purple in the pie chart), while introns and regulatory sequences associated with genes (lighter purple) make up about a quarter. The vast majority of the human genome does not code for proteins (although much of it gives rise to RNAs), and a large amount is repetitive DNA (dark and light green and teal).



Mastering Biology HHMI Animation: Coding Sequences in DNA

The bulk of many eukaryotic genomes consists of DNA sequences that neither code for proteins nor are transcribed to produce RNAs with known functions; this noncoding DNA was often described in the past as "junk DNA." However, genome comparisons over the past ten years have revealed the persistence of this DNA in diverse genomes over many hundreds of generations. For example, the genomes of humans, rats, and mice contain almost 500 regions of noncoding DNA that are *identical* in sequence in all three species. This is a higher level of sequence conservation than is seen for protein-coding regions in these species, strongly suggesting that the noncoding regions have important functions. The results of the ENCODE project discussed earlier have underscored the key roles played by much of this noncoding DNA. Next, we'll examine how genes and noncoding DNA sequences are organized within genomes of multicellular eukaryotes, using the human genome as our main example. Genome organization tells us a lot about how genomes have evolved and continue to evolve, as we'll see in Concept 21.5.

Transposable Elements and Related Sequences

Both prokaryotes and eukaryotes have stretches of DNA that can move from one location to another within the genome. These stretches are known as *transposable genetic elements*, or simply **transposable elements**. During the process called *transposition*, a transposable element moves from one site in a cell's DNA to a different target site by a type of recombination process. Transposable elements are sometimes called "jumping genes," but actually they never entirely detach from the cell's DNA. Instead, the original and new DNA sites are brought very close together by enzymes and other proteins that bend the DNA. Surprisingly, about 75% of human repetitive DNA (44% of the entire human genome) is made up of transposable elements and sequences related to them.

The first evidence for wandering DNA segments came from American geneticist Barbara McClintock's breeding experiments with calico corn (maize) in the 1940s and 1950s **(Figure 21.7)**. Tracking corn plants through many generations, McClintock analyzed changes in the color of corn kernels. The patterns she saw led her to propose that there were genetic elements capable of moving from other locations in the genome into the genes for kernel color, disrupting the genes and changing the kernel color. McClintock's hypothesis provoked great interest among her maize colleagues, but most other scientists thought the phenomenon she had observed might occur only in maize. Her careful work and insightful ideas were finally validated many years later when transposable elements were found in bacteria. In 1983, at the age of 81, McClintock received the Nobel Prize for her pioneering research.

Movement of Transposons and Retrotransposons

Eukaryotic transposable elements are of two types. The first type, **transposons**, move within a genome by means of a DNA intermediate. Transposons can move by a "cut-and-paste" mechanism, which removes the element from the original site, or by a "copy-and-paste" mechanism, which leaves a copy behind **(Figure 21.8)**. Both mechanisms require an enzyme called *transposase*, which is generally encoded by the transposon.

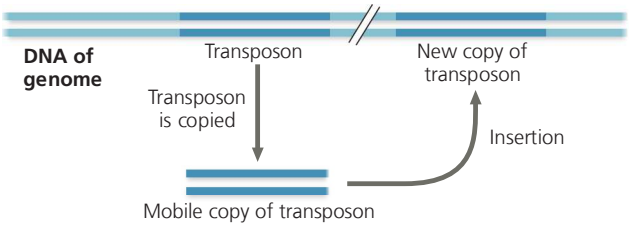
Most transposable elements in eukaryotic genomes are of the second type, **retrotransposons**, which move by means of an RNA intermediate that is a transcript of the retrotransposon DNA. Thus, retrotransposons always leave a copy at the original site during transposition **(Figure 21.9)**. To insert at another site, the RNA intermediate is first converted back to DNA by reverse transcriptase, an enzyme encoded by the retrotransposon. (Reverse transcriptase is also encoded by retroviruses, as you learned in Concept 19.2. In fact, retroviruses may have evolved from retrotransposons, or vice versa.) Another cellular enzyme catalyzes insertion of the reverse-transcribed DNA at a new site.





◀ Figure 21.7 The effect of transposable elements on corn kernel color. Barbara McClintock first proposed the idea of mobile genetic elements after observing variegations in the color of the kernels on a corn cob (top right).

▼ Figure 21.8 Transposon movement. Movement of transposons by either the copy-and-paste mechanism (shown here) or the cut-and-paste mechanism involves a double-stranded DNA intermediate that is inserted into the genome.

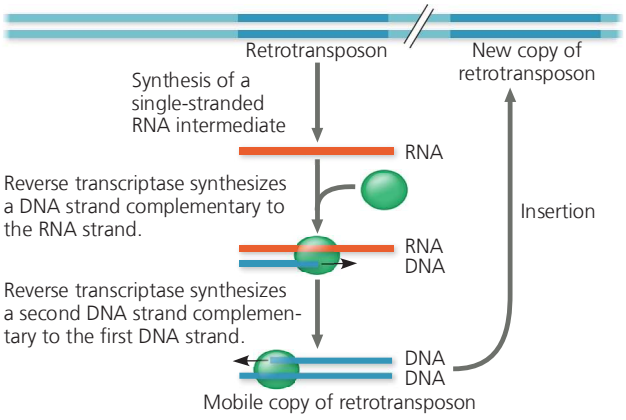


VISUAL SKILLS *How would this figure differ if it showed the cut-and-paste mechanism?*

Mastering Biology
Interview with Virginia Walbot: Plant genetics and development



▼ Figure 21.9 Retrotransposon movement. Movement begins with synthesis of a single-stranded RNA intermediate. The remaining steps are essentially identical to part of the retrovirus replicative cycle (see Figure 19.8).



Sequences Related to Transposable Elements

Multiple copies of transposable elements and sequences related to them are scattered throughout eukaryotic genomes. A single unit is usually hundreds to thousands of base pairs long, and the dispersed copies are similar but usually not identical to each other. Some of these are transposable elements that can move; the enzymes required for this movement may be encoded by any transposable element, including the one that is moving. Others are related sequences that have lost the ability to move altogether. Transposable elements and related sequences make up 25–50% of most mammalian genomes (see Figure 21.6) and even higher percentages in amphibians and many plants. In fact, the very large size of some plant genomes is accounted for by extra transposable elements rather than by extra genes. For example, transposable elements make up 85% of the corn genome!

In humans and other primates, a large portion of transposable element–related DNA consists of a family of similar sequences called *Alu elements*. These sequences alone account for approximately 10% of the human genome. *Alu* elements are about 300 nucleotides long, much shorter than most functional transposable elements, and they do not code for any protein. However, many *Alu* elements are transcribed into RNA, and at least some of these RNAs are thought to help regulate gene expression.

An even larger percentage (17%) of the human genome is made up of a type of retrotransposon called *LINE-1*, or *L1*. These sequences are much longer than *Alu* elements—about 6,500 base pairs—and typically have a very low rate of transposition. However, researchers working with mice have discovered that transcription of L1 retrotransposons is crucial for the development of early (one- and two-cell stage) embryos. They have proposed that transcription of L1 retrotransposons may affect the chromatin structure in ways important for embryonic development.

Although some transposable elements encode proteins, these proteins do not carry out normal cellular functions. Therefore, transposable elements are usually included in the "noncoding" DNA category, along with other repetitive sequences.

Other Repetitive DNA, Including Simple Sequence DNA

Repetitive DNA that is not related to transposable elements has probably arisen from mistakes during DNA replication or recombination. Such DNA accounts for about 14% of the human genome (see Figure 21.6). About a third of this (5–6% of the human genome) consists of duplications of long stretches of DNA, with each unit ranging from 10,000 to 300,000 base pairs. These long segments seem to have been copied from one chromosomal location to another site on the same or a different chromosome and probably include some functional genes.

In contrast to scattered copies of long sequences, stretches of DNA known as **simple sequence DNA** contain many copies of tandemly repeated short sequences, as in the following example (showing one DNA strand only):

. . . GTTACGTTACGTTACGTTACGTTACGTTAC. . .

In this case, the repeated unit (GTTAC) consists of 5 nucleotides, but the number can range from 2 to 500. When the unit contains 2–5 nucleotides, the series of repeats is called a **short tandem repeat**, or **STR**; we discussed the use of STR analysis in preparing genetic profiles by use of PCR in Concept 20.4 (see Figure 20.24). The number of copies of the repeated unit can vary from site to site within a given genome. There could be as many as several hundred thousand repetitions of the GTTAC unit at one site, but only half that number at another. STR analysis is performed on sites selected because they have relatively few repeats. The repeat number varies from person to person, and since humans are diploid, each person has two alleles per repeat site; these can differ in repeat number. This diversity produces the variation represented in the genetic profiles that result from STR analysis.

Simple sequence DNA makes up 3% of the human genome, much of it located at chromosomal telomeres and centromeres, where it may play a structural role. The DNA at centromeres is essential for the separation of chromatids in cell division (see Concept 12.2) and, along with simple sequence DNA located elsewhere, may also help organize the chromatin within the interphase nucleus. The simple sequence DNA located at telomeres binds proteins that protect chromosomal ends from degradation and from joining to other chromosomes.

Short repetitive sequences like those described here provide a challenge for whole-genome shotgun sequencing because the presence of many short repeats hinders accurate reassembly of fragment sequences by computers. Regions of simple sequence DNA account for much of the uncertainty present in estimates of whole-genome sizes and are the reason some sequences are considered "permanent drafts."

Genes and Multigene Families

Now, let's take a look at genes. Recall that DNA sequences that code for proteins or give rise to tRNA or rRNA make up only 1.5% of the human genome (see Figure 21.6). If we include introns and regulatory sequences, the total amount of DNA that is gene-related—coding and noncoding—constitutes about 25% of the human genome. Put another way, only about 6% (1.5% out of 25%) of the length of the average gene is represented in the final gene product.

Many eukaryotic genes are present as unique sequences, with only one copy per haploid set of chromosomes. But in the human genome and the genomes of many other animals and plants, these unique genes make up less than half of the total gene-related DNA. The rest occur in **multigene families**, collections of two or more identical or very similar genes.

In multigene families that consist of *identical* DNA sequences, those sequences are usually clustered tandemly and, with the notable exception of the genes for histone proteins, have RNAs as their final products. An example is the family of identical DNA sequences that each include the genes for the three largest rRNA molecules **(Figure 21.10a)**. These rRNA molecules are transcribed from a single transcription unit that is repeated tandemly hundreds to thousands of times in one or several clusters in the genome of a multicellular eukaryote. The many copies of this rRNA transcription unit help cells to quickly make the millions of ribosomes needed for active protein synthesis. The primary transcript is cleaved to yield three rRNA molecules, which combine with proteins and one other kind of rRNA (5*S* rRNA) to form ribosomal subunits.

The classic examples of multigene families of *nonidentical* genes are two related families of genes that encode globins, a group of proteins that include the α and β polypeptide subunits of hemoglobin. One family, located on chromosome 16 in humans, encodes various forms of α-globin; the other, on chromosome 11, encodes forms of β-globin **(Figure 21.10b)**. The different forms of each globin subunit are expressed at different times in development, allowing hemoglobin to function effectively in the changing environment of the developing animal. In humans, for example, the embryonic and fetal forms of hemoglobin have a higher affinity for oxygen than the adult forms, ensuring the efficient transfer of oxygen from mother to fetus. Also found in the globin gene family clusters are several pseudogenes.

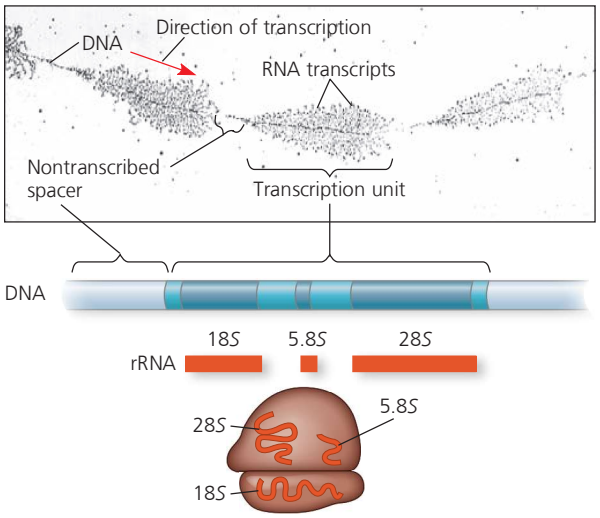
In Concept 21.5, we'll consider the evolution of these two globin gene families as we explore how arrangements of genes provide insight into the evolution of genomes. We'll also examine some processes that have shaped the genomes of different species over evolutionary time.

CONCEPT CHECK 21.4

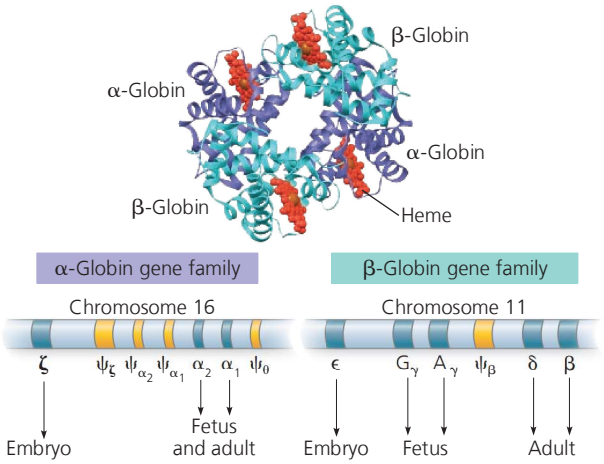
1. Discuss the characteristics of mammalian genomes that make them larger than prokaryotic genomes.
2. **VISUAL SKILLS** Which of the three mechanisms described in Figures 21.8 and 21.9 result(s) in a copy remaining at the original site as well as a copy appearing in a new location?
3. Contrast the organizations of the rRNA gene family and the globin gene families. For each, explain how the existence of a family of genes benefits the organism.
4. **MAKE CONNECTIONS** Assign each DNA segment at the top of Figure 18.9 to a sector in the pie chart in Figure 21.6.

For suggested answers, see Appendix A.

▼ Figure 21.10 Gene families.



(a) Part of the ribosomal RNA gene family. The TEM at the top shows three of the hundreds of copies of rRNA transcription units in the rRNA gene family of a salamander genome. Each "feather" corresponds to a single unit being transcribed by about 100 molecules of RNA polymerase (dark dots along the DNA), moving left to right (red arrow). The growing RNA transcripts extend from the DNA, accounting for the feather-like appearance. In the diagram of a transcription unit below the TEM, the genes for three types of rRNA (darker blue) are adjacent to regions that are transcribed but later removed (medium blue). A single transcript is processed to yield one of each of the three rRNAs (red), key components of the ribosome.



(b) The human α-globin and β-globin gene families. Adult hemoglobin is composed of two α-globin and two β-globin polypeptide subunits, as shown in the molecular model. The genes (darker blue) encoding α- and β-globins are found in two families, organized as shown here. The noncoding DNA (light blue) separating the functional genes within each family includes pseudogenes (ψ; gold), versions of the functional genes that no longer encode functional polypeptides. Genes and pseudogenes are named with Greek letters, as you have seen previously for the α- and β-globins. Some genes are expressed only in the embryo or fetus.

VISUAL SKILLS *In the TEM at the top of part (a), how could you determine the direction of transcription if it weren't indicated by the red arrow?*

CONCEPT 21.5

Duplication, rearrangement, and mutation of DNA contribute to genome evolution

EVOLUTION Now that we have explored the makeup of the human genome, let's see what its composition reveals about how the genome evolved. The basis of change at the genomic level is mutation, which underlies much of genome evolution. It seems likely that the earliest forms of life had a minimal number of genes—those necessary for survival and reproduction. If this were indeed the case, one aspect of evolution must have been an increase in the size of the genome, with the extra genetic material providing the raw material for gene diversification. In this section, we'll first look at how extra copies of all or part of a genome can arise and then consider subsequent processes that can lead to the evolution of proteins (or RNA products) with slightly different or entirely new functions.

Duplication of Entire Chromosome Sets

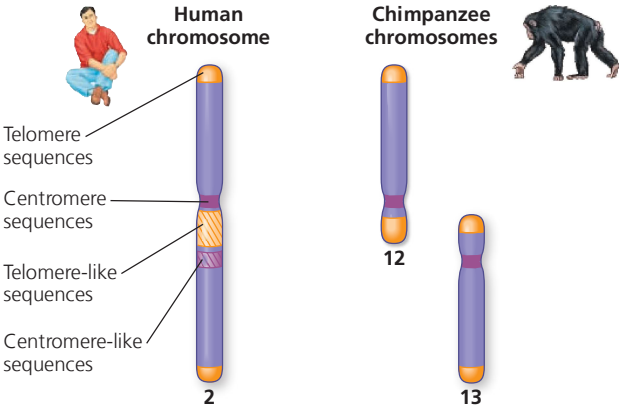
An accident in meiosis, such as failure to separate homologs during meiosis I, can result in one or more extra sets of chromosomes, a condition known as polyploidy (see Concept 15.4). Although such accidents would most often be lethal, in rare cases they could facilitate the evolution of genes. In a polyploid organism, one set of genes can provide essential functions for the organism. The genes in the one or more extra sets can diverge by accumulating mutations; these variations may persist if the organism carrying them survives and reproduces. In this way, genes with novel functions can evolve. As long as one copy of an essential gene is expressed, the divergence of another copy can lead to its encoded protein acting in a novel way, thereby changing the organism's phenotype.

The outcome of this accumulation of mutations may eventually be the branching off of a new species. While polyploidy is rare among animals, it is relatively common among plants, especially flowering plants. Some botanists estimate that as many as 80% of the plant species that are alive today show evidence of polyploidy having occurred among their ancestral species. You'll learn more about the details of how polyploidy leads to plant speciation in Concept 24.2.

Alterations of Chromosome Structure

With the recent explosion in genomic sequence information, we can now compare the chromosomal organizations of many different species in detail. This information allows us to make inferences about the evolutionary processes that shape chromosomes and may drive speciation. For example, scientists have long known that sometime in the last 6 million years, when the ancestors of humans and chimpanzees diverged as species, the fusion of two ancestral chromosomes

▼ **Figure 21.11 Human and chimpanzee chromosomes.** The positions of telomere-like and centromere-like sequences on human chromosome 2 (left) match those of telomeres on chimpanzee chromosomes 12 and 13 and the centromere on chimpanzee chromosome 13 (right). This suggests that chromosomes 12 and 13 in a human ancestor fused end to end to form human chromosome 2. The centromere from ancestral chromosome 12 remained functional on human chromosome 2, while the one from ancestral chromosome 13 did not.



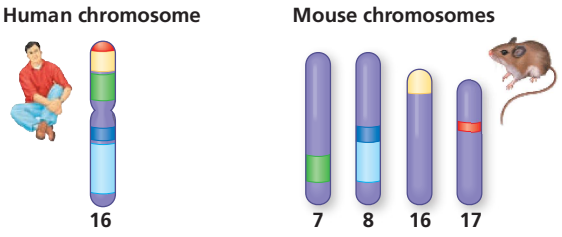
in the human line led to different haploid numbers for humans ($n = 23$) and chimpanzees ($n = 24$). The banding patterns in stained chromosomes suggested that the ancestral versions of current chimpanzee chromosomes 12 and 13 fused end to end, forming chromosome 2 in an ancestor of the human lineage **(Figure 21.11)**.

How do we know a chromosome didn't just split into two in an ancestor of chimps? A comparison of chromosomes in other great apes—gorillas, chimps, bonobos, and orangutans, our closest relatives—shows that those species have 24 chromosomes, so the simplest conclusion is that the last common ancestor of all great apes had 24 chromosomes. (You'll learn more about this type of reasoning in Concept 26.3.) Sequencing and analysis of human chromosome 2 during the Human Genome Project revealed sequences for telomeres and an extra, unused centromere in the middle of it, among other very strong supporting evidence for the model described previously (see Figure 21.11).

In another study of broader scope, researchers compared the DNA sequence of each human chromosome with the whole-genome sequence of the mouse **(Figure 21.12)**. One part of their study showed that large blocks of genes on human chromosome 16 are found on four mouse chromosomes, indicating that the genes in each block stayed together in both the mouse and the human lineages during their divergent evolution from a common ancestor.

Performing the same comparison of chromosomes of humans and six other mammalian species allowed the researchers to reconstruct the evolutionary history of chromosomal rearrangements in these eight species. They found many duplications and inversions of large portions of chromosomes, the result of errors during meiotic recombination in which the DNA was broken and rejoined incorrectly. The rate of

▼ Figure 21.12 Human and mouse chromosomes. Here, we can see that DNA sequences very similar to large blocks of human chromosome 16 (colored areas in this diagram) are found on mouse chromosomes 7, 8, 16, and 17. This finding suggests that the DNA sequence in each block has stayed together in the mouse and human lineages since the time they diverged from a common ancestor.



Mastering Biology
Interview with Eric Lander: Exploring the human genome



these events seems to have begun accelerating about 100 million years ago, around 35 million years before large dinosaurs became extinct and the number of mammalian species began rapidly increasing. The apparent coincidence is interesting because chromosomal rearrangements are thought to contribute to the generation of new species. Although two individuals with different arrangements could still mate and produce offspring, the offspring would have two nonequivalent sets of chromosomes, making meiosis inefficient or even impossible. Thus, chromosomal rearrangements would lead to two populations that could not successfully mate with each other, a step on the way to their becoming two separate species. (You'll learn more about this in Concept 24.2.)

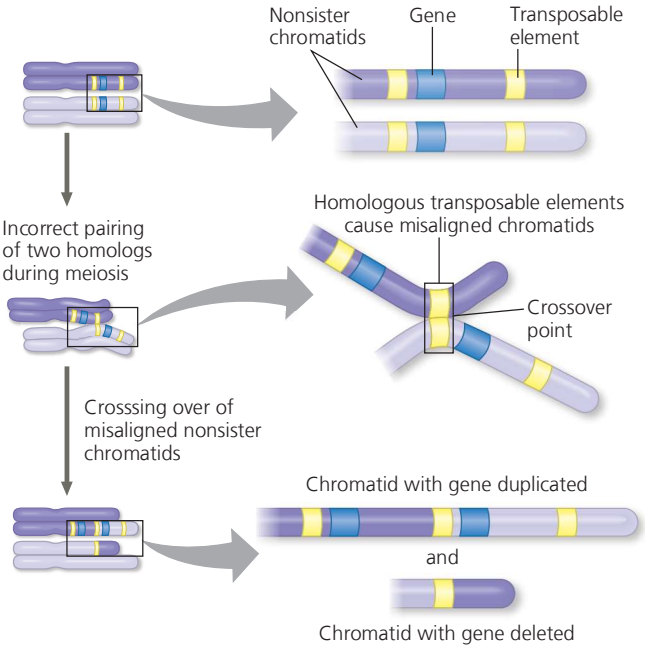
The same study also unearthed a pattern with medical relevance. Analysis of the chromosomal breakage points associated with the rearrangements showed that specific sites were used over and over again. A number of these recombination "hot spots" correspond to locations of chromosomal rearrangements within the human genome that are associated with congenital diseases (see Concept 15.4).

Duplication and Divergence of Gene-Sized Regions of DNA

Errors during meiosis can also lead to the duplication of chromosomal regions that are smaller than the ones we've just discussed, including segments the length of individual genes. Unequal crossing over during prophase I of meiosis, for instance, can result in one chromosome with a deletion and another with a duplication of a particular gene or genes. Transposable elements can provide homologous sites where nonsister chromatids can cross over, even when other chromatid regions are not correctly aligned **(Figure 21.13)**.

Also, slippage can occur during DNA replication, such that the template shifts with respect to the new complementary strand, and a part of the template strand is either skipped by the replication machinery or used twice as a template. As a result, a segment of DNA is deleted or duplicated. It is easy to imagine how such errors could occur in regions of repeats. (See the question in Figure 21.13.) The variable number of repeated units of simple sequence DNA at a given site, used for STR analysis, is probably due to errors like these. Evidence that unequal crossing over and template slippage during DNA replication lead to duplication of genes is found in the existence of multigene families, such as the globin family.

▼ Figure 21.13 Gene duplication due to unequal crossing over. One mechanism by which a gene (or other DNA segment) can be duplicated is recombination during meiosis between copies of a transposable element (yellow) flanking the gene (blue). Such recombination between misaligned nonsister chromatids of homologous chromosomes produces one chromatid with two copies of the gene and one chromatid with no copy. (Genes and transposable elements are shown only in the region of interest.)

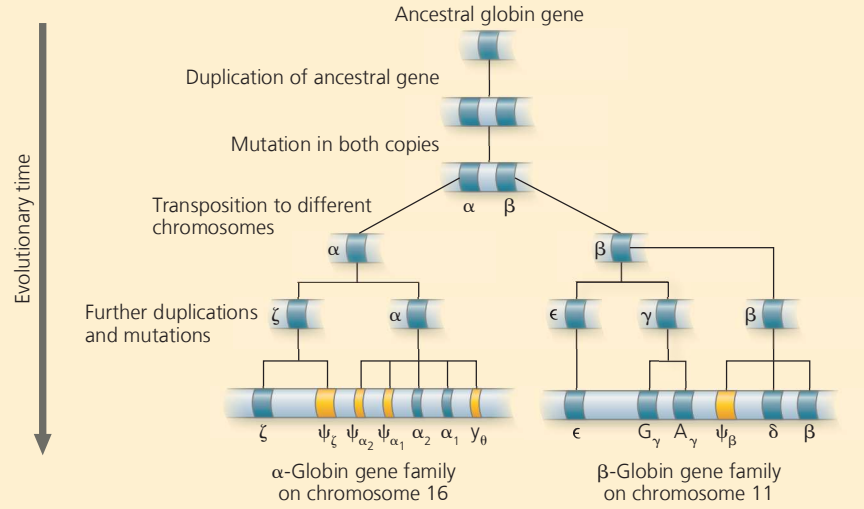


MAKE CONNECTIONS *Examine how crossing over occurs in Figure 13.9. In the middle panel above, draw a line through the portions that result in the upper chromatid in the bottom panel. Use a different color to do the same for the other chromatid.*

Evolution of Genes with Related Functions: The Human Globin Genes

In Figure 21.10b, you saw the organization of the α-globin and β-globin gene families as they exist in the human genome today. Now, let's consider how events such as duplications can lead to the evolution of genes with related functions like the globin genes. A comparison of gene sequences within a multigene family can suggest the order in which the genes arose. Re-creating the evolutionary history of the globin genes using this approach indicates that they all evolved from one common ancestral globin gene that underwent duplication and divergence into the α-globin and β-globin ancestral genes about 450–500 million years ago

▼ Figure 21.14 A proposed model for the sequence of events in the evolution of the human α-globin and β-globin gene families from a single ancestral globin gene.



? *The gold elements are pseudogenes. Explain how they could have arisen after gene duplication.*

(Figure 21.14). Each of these genes was later duplicated several times, and the copies then diverged from each other in sequence, yielding the current family members. In fact, the common ancestral globin gene also gave rise to the oxygen-binding muscle protein myoglobin and to the plant protein leghemoglobin. The latter two proteins function as monomers, and their genes are included in a "globin superfamily."

After the duplication events, the differences between the genes in the globin families undoubtedly arose from mutations that accumulated in the gene copies over many generations. The current model is that the necessary function provided by an α-globin protein, for example, was fulfilled by one gene, while other copies of the α-globin gene accumulated random mutations. Many mutations may have had an adverse effect on the organism, and others may have had no effect. However, a few mutations must have altered the function of the protein product in a way that benefitted the organism at a particular life stage without substantially changing the protein's oxygen-carrying function. Presumably, natural selection acted on these altered genes, maintaining them in the population.

In the **Scientific Skills Exercise**, you can compare amino acid sequences of the globin family proteins and see how such comparisons were used to generate the model for globin gene evolution shown in Figure 21.14. The existence of several pseudogenes among the functional globin genes provides additional evidence for this model: Random mutations in these "genes" over evolutionary time have destroyed their function.

Evolution of Genes with Novel Functions

In the evolution of the globin gene families, gene duplication and subsequent divergence produced family members whose protein products performed similar functions (oxygen transport). Alternatively, one copy of a duplicated gene can undergo alterations that lead to a completely new function for the protein product. The genes for lysozyme and α-lactalbumin are a good example.

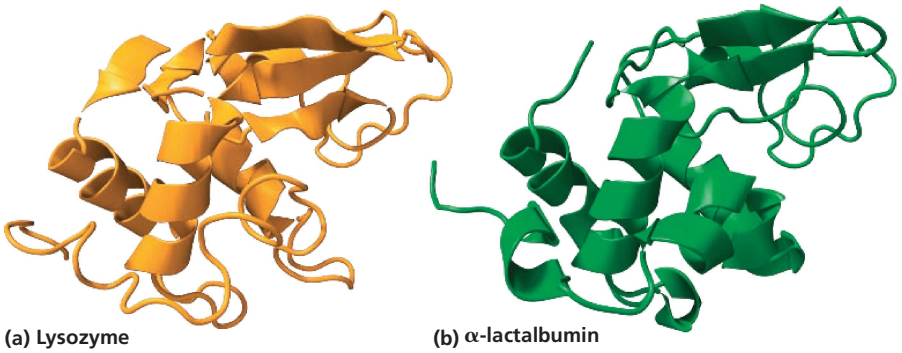
Lysozyme is an enzyme that helps protect animals against bacterial infection by hydrolyzing bacterial cell walls (see Visualizing Figure 5.16); α-lactalbumin is a nonenzymatic protein that plays a role in milk production in mammals. The two proteins are quite similar in their amino acid sequences and three-dimensional structures **(Figure 21.15)**. Both genes are found in mammals, but only the lysozyme gene is present in birds. These findings suggest that at some time after the lineages leading to mammals and birds had separated, the lysozyme gene was duplicated in the mammalian lineage but not in the avian lineage. Subsequently, one copy of the duplicated lysozyme gene evolved into a gene encoding α-lactalbumin, a protein with a completely new function associated with a key characteristic of mammals—milk production. In one study, evolutionary biologists searched vertebrate genomes for genes with similar sequences. There appear to be at least eight members of the lysozyme family, distributed widely among mammalian species. The functions of all the encoded gene products are not yet known, but it will be exciting to discover whether they are as different as the functions of lysozyme and α-lactalbumin.

Besides the duplication and divergence of whole genes, rearrangement of existing DNA sequences within genes has also contributed to genome evolution. The presence of introns may have promoted the evolution of new proteins by facilitating the duplication or shuffling of exons, as we'll see next.

Rearrangements of Parts of Genes: Exon Duplication and Exon Shuffling

Recall from Concept 17.3 that an exon often codes for a protein domain, a distinct structural and functional region of a protein molecule, such as the WD40 domain in Figure 21.3. We've already seen that unequal crossing over during meiosis can lead to duplication of a gene on one chromosome and its loss from the homologous chromosome (see Figure 21.13). By a similar process, a particular exon within a gene could be duplicated on one chromosome and deleted from the other. The gene with the duplicated exon would code for a protein containing a second copy of the encoded domain. This change in the protein's structure might

▼ Figure 21.15 Comparison of lysozyme and α-lactalbumin proteins. Computer-generated ribbon models of the similar structures of **(a)** lysozyme and **(b)** α-lactalbumin are shown, along with **(c)** a comparison of the amino acid sequences of the two proteins. Single-letter amino acid codes are used (see Figure 5.14). Identical amino acids are highlighted in yellow, and dashes indicate gaps in one sequence that have been introduced by the software to optimize the alignment.



(a) Lysozyme (b) α-lactalbumin

Lysozyme	1	KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATN
α-lactalbumin	1	KQFTKCELSQLLK--DIDGYGGIALPELICTMFHTSGYDTQAIV
Lysozyme	51	STDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVA
α-lactalbumin	51	STEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDITDDIM
Lysozyme	101	RDPQGIRAWVAWRNRCQ-NRDVRQYVQGCGV
α-lactalbumin	101	D-IKGIDYWLAHKALCT--EKLEQWLCEKL-

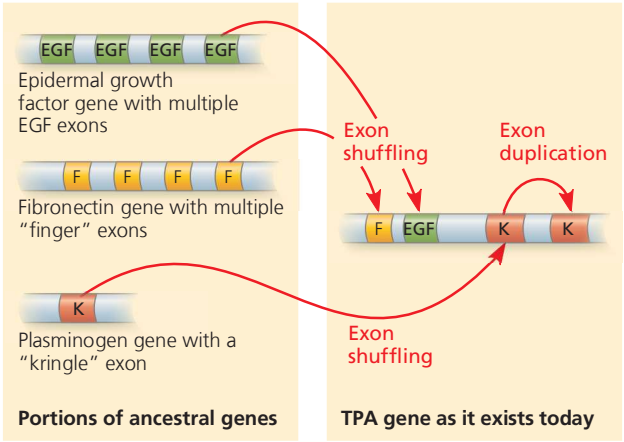
(c) Amino acid sequence alignments of lysozyme and α-lactalbumin

MAKE CONNECTIONS *Even though two amino acids are not identical, they may be structurally and chemically similar and therefore behave similarly. Using Figure 5.14 as a reference, examine the nonidentical amino acids in positions 1–30 and note cases where the amino acids in the two sequences are similarly acidic or basic.*

augment its function by increasing its stability, enhancing its ability to bind a particular ligand, or altering some other property. Quite a few protein-coding genes have multiple copies of related exons, which presumably arose by duplication and then diverged. The gene encoding the extracellular matrix protein collagen is a good example. Collagen is a structural protein (see Figure 5.18) with a highly repetitive amino acid sequence, which reflects the repetitive pattern of exons in the collagen gene.

As an alternative possibility, we can imagine the occasional mixing and matching of different exons either within a gene or between two different (nonallelic) genes owing to errors in meiotic recombination. This process, termed *exon shuffling*, could lead to new proteins with novel combinations of functions. As an example, let's consider the gene for tissue plasminogen activator (TPA). The TPA protein is an extracellular protein that helps control blood clotting. It has four domains of three types, each encoded by an exon, and one of those exons is present in two copies. Because each type of exon is also found in other proteins, the current version of the gene for TPA is thought to have arisen by several instances of exon shuffling during errors in meiotic recombination and subsequent duplication **(Figure 21.16)**.

▼ Figure 21.16 Evolution of a new gene by exon shuffling. Meiotic errors could have moved exons, each encoding a particular domain, from ancestral forms of the genes for epidermal growth factor, fibronectin, and plasminogen (left) into the evolving gene for tissue plasminogen activator, TPA (right). Subsequent duplication of the "kringle" exon (K) from the plasminogen gene after its movement into the TPA gene could account for the two copies of this exon in the TPA gene existing today.



VISUAL SKILLS *Looking at Figure 21.13, describe the steps by which transposable elements within introns might have facilitated the exon shuffling shown here.*

Scientific Skills Exercise

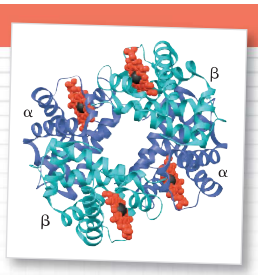
Reading an Amino Acid Sequence Identity Table

How Have Amino Acid Sequences of Human Globin Genes Diverged During Their Evolution? To build a model of the evolutionary history of the globin genes (see Figure 21.14), researchers compared the amino acid sequences of the polypeptides they encode. In this exercise, you will analyze comparisons of the amino acid sequences of the globin polypeptides to shed light on their evolutionary relationships.

How the Experiment Was Done Scientists obtained the DNA sequences for each of the eight globin genes and "translated" them into amino acid sequences. They then used a computer program to align the sequences (with dashes indicating gaps in one sequence) and calculate a percent identity value for each pair of globins. The percent identity reflects the number of positions with identical amino acids relative to the total number of amino acids in a globin polypeptide. The data were displayed in a table to show the pairwise comparisons.

Data from the Experiment The following table shows an example of a pairwise alignment—that of the α_1-globin (alpha-1 globin) and ζ-globin (zeta globin) amino acid sequences—using the standard single-letter symbols for amino acids. To the left of each line of amino acid sequence is the number of the first amino acid in that line. The percent identity value for the α_1- and ζ-globin amino acid sequences was calculated by counting the number of matching amino acids (86, highlighted in yellow), dividing by the total number of amino acid positions (143), and then multiplying by 100. This resulted in a 60% identity value for the α_1-ζ pair, as shown in the amino acid identity table at the bottom of the page. The values for other globin pairs were calculated in the same way.

Globin	Alignment of Globin Amino Acid Sequences
α_1	1 MVLSPADKTNVKAAWGKVGAHAGEYGAEAL
ζ	1 MSLTKTERTIIVSMWAKISTQADTIGTETL
α_1	31 ERMFLSFPTTKTYFPHFDLSH-GSAQVKGH
ζ	31 ERLFLSHPQTKTYFPHFDL-HPGSAQLRAH
α_1	61 GKKVADALTNAVAHVDDMPNALSALSDLHA
ζ	61 GSKVVAAVGDAVKSIDDIGGALSKLSELHA
α_1	91 HKLRVDPVNFKLLSHCLLVTLAAHLPAEFT
ζ	91 YILRVDPVNFKLLSHCLLVTLAARFPADFT
α_1	121 PAVHASLDKFLASVSTVLTSKYR
ζ	121 AEAHAAWDKFLSVVSSVLTEKYR



▲ Hemoglobin

INTERPRET THE DATA

1. Note that in the amino acid identity table, the data are arranged so each globin pair can be compared. **(a)** Some cells in the table have dashed lines. Given the pairs that are being compared for these cells, what percent identity value is implied by the dashed lines? **(b)** Notice that the cells in the lower left half of the table are blank. Using the information already provided in the table, fill in the missing values. Why does it make sense that these cells were left blank in the table?
2. The earlier that two genes arose from a duplicated gene, the more their nucleotide sequences can have diverged, which may result in amino acid differences in the protein products. **(a)** Based on that premise, identify which two genes are most divergent from each other. What is the percent amino acid identity between their polypeptides? **(b)** Using the same approach, identify which two globin genes are the most recently duplicated. What is the percent identity between them?
3. The model of globin gene evolution shown in Figure 21.14 suggests that an ancestral gene duplicated and mutated to become α- and β-globin genes, and then each one was further duplicated and mutated. What features of the data set support the model?
4. Make an ordered list of all the percent identity values from the table, starting with 100% at the top. Next to each number write the globin pair(s) with that percent identity value. Use one color for the globins from the α family and a different color for the globins from the β family. **(a)** Compare the order of pairs on your list with their positions in the model shown in Figure 21.14. Does the order of pairs describe the same relative "closeness" of globin family members seen in the model? **(b)** Compare the percent identity values for pairs within the α or β family to the values for between-family pairs.

Instructors: A version of this Scientific Skills Exercise can be assigned in **Mastering Biology**.

Further Reading R. C. Hardison, Globin genes on the move, *Journal of Biology* 7:35.1–35.5 (2008).

Amino Acid Identity Table

		α Family			β Family				
		α_1 (alpha 1)	α_2 (alpha 2)	ζ (zeta)	β (beta)	δ (delta)	ϵ (epsilon)	A_γ (gamma A)	G_γ (gamma G)
α Family	α_1	-----	100	60	45	44	39	42	42
	α_2		-----	60	45	44	39	42	42
	ζ			-----	38	40	41	41	41
β Family	β				-----	93	76	73	73
	δ					-----	73	71	72
	ϵ						-----	80	80
	A_γ							-----	99
	G_γ								-----

Compiled using data from the National Center for Biotechnology Information (NCBI).

How Transposable Elements Contribute to Genome Evolution

The persistence of transposable elements as a large fraction of some eukaryotic genomes is consistent with the idea that they play an important role in shaping a genome over evolutionary time. These elements can contribute to the evolution of the genome in several ways. They can promote recombination, disrupt cellular genes or control elements, and carry entire genes or individual exons to new locations.

Transposable elements of similar sequence scattered throughout the genome facilitate recombination between different (nonhomologous) chromosomes by providing homologous regions for crossing over (see Figure 21.13). Most such recombination events are probably detrimental, causing chromosomal translocations and other changes in the genome that may be lethal to the organism. But over the course of evolutionary time, an occasional recombination event of this sort may be advantageous to the organism. (For the change to be heritable, of course, it must happen in a cell that will give rise to a gamete.)

The movement of a transposable element can have a variety of consequences. For instance, a transposable element that "jumps" into a protein-coding sequence will prevent the production of a normal transcript of the gene. (Introns provide a sort of "safety zone" that does not affect the transcript because the transposable element will be spliced out—unless it affects the splicing process.) If a transposable element inserts within a regulatory sequence, the transposition may lead to increased or decreased production of one or more proteins. Transposition caused both types of effects on the genes coding for pigment-synthesizing enzymes in McClintock's corn kernels. Again, while such changes are usually harmful, in the long run some may provide a survival advantage. A possible example was mentioned earlier: At least some of the *Alu* transposable elements in the human genome are known to produce RNAs that regulate expression of human genes.

During transposition, a transposable element may carry along a gene or even a group of genes to a new position in the genome. This occurrence probably accounts for the location of the α-globin and β-globin gene families on different human chromosomes, as well as the dispersion of the genes of certain other gene families. By a similar tag-along process, an exon from one gene may be inserted into another gene in a mechanism similar to that of exon shuffling during recombination. For example, an exon may be inserted by transposition into the intron of a protein-coding gene. If the inserted exon is retained in the RNA transcript during RNA splicing, the protein that is synthesized will have an additional domain, which may confer a new function on the protein.

Most often, the processes discussed in this section produce harmful effects, which may be lethal, or have no effect at all. In a few cases, however, small heritable changes that are beneficial may occur. Over many generations, the resulting genetic diversity provides valuable raw material for natural selection. Diversification of genes and their products is an important factor in the evolution of new species. Thus, the accumulation of changes in the genome of each species provides a record of its evolutionary history. To read this record, we must be able to identify genomic changes. Comparing the genomes of different species allows us to do that, increasing our understanding of how genomes evolve. You will learn more about these topics next.

CONCEPT CHECK 21.5

1. Describe three examples of errors in cellular processes that lead to DNA duplications.
2. Explain how multiple exons might have arisen in the ancestral EGF and fibronectin genes shown in Figure 21.16 (left).
3. What are three ways that transposable elements are thought to contribute to genome evolution?
4. **WHAT IF?** In 2005, Icelandic scientists reported finding a large chromosomal inversion present in 20% of northern Europeans, and they noted that Icelandic women with this inversion had significantly more children than women without it. What would you expect to happen to the frequency of this inversion in the Icelandic population in future generations?

For suggested answers, see Appendix A.

CONCEPT 21.6

Comparing genome sequences provides clues to evolution and development

EVOLUTION In the last 30 years, we have seen rapid advances in genome sequencing and data collection, new techniques for assessing gene activity across the whole genome and for editing a gene sequence in a specific way in living cells, and refined approaches for understanding how genes and their products work together in complex systems. In the field of biology, we are truly poised on the brink of a new world.

Comparing genome sequences from different species reveals a lot about the evolutionary history of life, from very ancient to more recent. Similarly, comparative studies of the genetic programs that direct embryonic development in different species are uncovering the mechanisms that generated the great diversity of life-forms present today. We'll now look at what has been learned from these two approaches.

Comparing Genomes

The more similar in sequence the genes and genomes of two species are, the less time has passed for mutations and other changes to accumulate, and therefore the more closely related those species are in their evolutionary history. Comparing genomes of closely related species sheds light on more recent

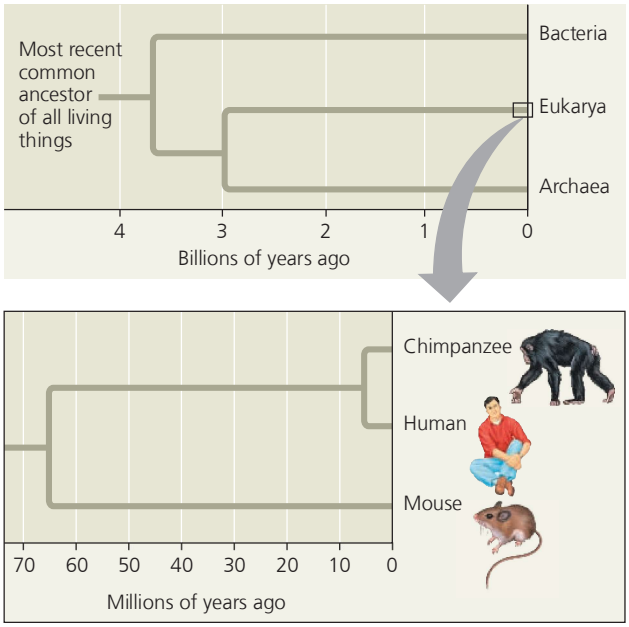
evolutionary events, whereas comparing genomes of very distantly related species helps us understand ancient evolutionary history. In either case, learning about characteristics that are shared or divergent between groups enhances our picture of the evolution of organisms and biological processes. Evolutionary relationships between species can be represented by a diagram in the form of a tree (often turned sideways), where each branch point marks the divergence of two lineages (see Figure 1.20). **Figure 21.17** shows the evolutionary relationships of some groups and species we'll now examine.

Comparing Distantly Related Species

Determining which genes have remained similar—that is, are *highly conserved*—in distantly related species can help clarify evolutionary relationships among species that diverged from each other long ago. Indeed, comparisons of the specific gene sequences of bacteria, archaea, and eukaryotes indicate that these three groups diverged between 2 and 4 billion years ago and strongly support the theory that they are the fundamental domains of life (see Figure 21.17).

In addition to their value in evolutionary biology, comparative genomic studies confirm the relevance of research on model organisms to our understanding of biology in general and human biology in particular. Very ancient genes can still be surprisingly similar in disparate species. One experimental study tested the ability of the human version of each of 414 important yeast genes to function equivalently in yeast cells. Remarkably, the researchers concluded that 47% of these yeast genes could be replaced by the corresponding human gene. This striking result underscores the common origin of yeasts and humans—two distantly related species.

▼ Figure 21.17 Evolutionary relationships of the three domains of life. The tree diagram at the top shows the ancient divergence of bacteria, archaea, and eukaryotes. A portion of the eukaryote lineage is expanded in the inset to show the more recent divergence of three mammalian species discussed in this chapter.



Comparing Closely Related Species

The genomes of two closely related species are likely to be organized similarly because of their relatively recent divergence. Their long shared history also means that only a small number of gene differences are found when their genomes are compared. These genetic differences can thus be more easily correlated with phenotypic differences between the two species. An exciting application of this type of analysis is seen as researchers compare the human genome with the genomes of the chimpanzee, mouse, rat, and other mammals. Identifying the genes shared by all of these species but not by nonmammals gives us clues about what it takes to make a mammal, while finding the genes shared by chimpanzees and humans but not by rodents tells us something about primates. And, of course, comparing the human genome with that of the chimpanzee helps us answer a tantalizing question: What genomic information defines a human or a chimpanzee?

An analysis of the overall composition of the human and chimpanzee genomes, which are thought to have diverged only about 6 million years ago (see Figure 21.17), reveals some general differences. Considering single nucleotide substitutions, the two genomes differ by only 1.2%. When researchers looked at longer stretches of DNA, however, they were surprised to find a further 2.7% difference due to insertions or deletions of larger regions in the genome of one or the other species; many of the insertions were duplications or other repetitive DNA. In fact, a third of the human duplications are not present in the chimpanzee genome, and some of these duplications contain regions associated with human diseases. More *Alu* elements have undergone transposition in the human genome, leading to 7,000 human-specific elements compared with 2,300 chimpanzee-specific elements. The chimpanzee genome, in turn, contains many copies of a retroviral provirus not present in humans. All of these observations provide clues to the forces that might have swept the two genomes along different paths, but we don't have a complete picture yet.

Along with chimpanzees, bonobos are the other African ape species that are the closest living relatives to humans. The sequencing of the bonobo genome revealed that in some regions, human sequences were more closely related to either chimpanzee or bonobo sequences than chimpanzee or bonobo sequences were to each other. Such a fine-grained comparison of three closely related species allows even more detail to be worked out in reconstructing their related evolutionary history.

We also don't know how the genetic differences revealed by genome sequencing might account for the distinct characteristics of each species. To discover the basis for the phenotypic differences between chimpanzees and humans, biologists are studying specific genes and types of genes that differ between the two species and comparing them with their

counterparts in other mammals. This approach has revealed a number of genes that are apparently changing (evolving) faster in the human than in either the chimpanzee or the mouse. Among them are genes involved in defense against malaria and tuberculosis as well as at least one gene that regulates brain size. When genes are classified by function, the genes that seem to be evolving the fastest are those that code for transcription factors. This discovery makes sense because transcription factors regulate gene expression and thus play a key role in orchestrating the overall genetic program.

One transcription factor of interest, whose gene is called *FOXP2*, might be involved with the acquisition of speech in humans. Attention was first drawn to *FOXP2* in 2002 when a research publication reported that this gene was evolving faster in humans (*Homo sapiens*) than in other primates, based in part on two amino acids found only in the human protein sequence. The researchers proposed that this was a human-specific version of the gene that resulted from strong natural selection and might have conferred on humans the ability to speak.

In 2014, a high-quality sequence of the Neanderthal (*Homo neanderthalensis*) genome was achieved from a small amount of preserved genomic DNA, and shortly thereafter the DNA of another member of the species *Homo*, the Denisovan, was sequenced as well (see Concept 34.7). It turns out that some DNA sequences from both species encode the same two amino acids, so the two amino acids are not human-specific. In a paper published in 2018, researchers analyzed genomic sequences from a larger and more diverse population of modern humans. In this more robust study, the researchers found no evidence for strong natural selection for those two amino acids in the human lineage during the time frame relevant to language acquisition.

However, the *FOXP2* gene is still interesting in its own right. Several lines of evidence suggest that the *FOXP2* gene product regulates genes that function in vocalization in vertebrates. First, mutations in this gene can produce severe speech and language impairment in humans. Second, the *FOXP2* gene is expressed in the brains of zebra finches and canaries at the time when these songbirds are learning their songs. And third, perhaps the strongest evidence comes from a "knockout" experiment in which researchers disrupted the *FOXP2* gene in mice and analyzed the resulting phenotype **(Figure 21.18)**. The homozygous mutant mice had malformed brains and failed to emit normal ultrasonic vocalizations; mice with one faulty copy of the gene also showed significant problems with vocalization. These results support the idea that the *FOXP2* gene product turns on genes involved in vocalization.

Expanding on this analysis, another research group replaced the *FOXP2* gene in mice with a "humanized" copy coding for the human versions of the FOXP2 protein, with the two amino acids that differ between human and chimpanzee sequences. The researchers reported that although the mice with the human *FOXP2* gene were generally healthy, they had subtly different vocalizations and showed changes in brain cells in circuits associated with speech in human brains. While the *FOXP2* gene clearly affects vocalization and speech, the precise mechanism by which it does this is still to be determined.

The *FOXP2* story is an excellent example of how different approaches can complement each other in uncovering biological phenomena of widespread importance. The *FOXP2* experiments used mice as a model for humans because it would be unethical (as well as impractical) to carry out such experiments in humans. Mice and humans, which diverged about 65.5 million years ago (see Figure 21.17), share about 85% of their genes. This genetic similarity can be exploited in studying human genetic disorders. If researchers know the organ or tissue that is affected by a particular genetic disorder, they can look for genes that are expressed in these locations in mice.

Even though more distantly related to humans, fruit flies have also been a useful model species for study of such human disorders as Parkinson's disease and alcoholism, while nematodes (soil worms) have yielded a wealth of information about aging. Further research efforts are under way to extend genomic studies to many more species, including neglected species from diverse branches of the tree of life. These studies will advance our understanding of evolution, of course, as well as all aspects of biology, from human health to ecology.

Comparing Genomes Within a Species

Another exciting consequence of our ability to analyze genomes is our growing understanding of the spectrum of genetic variation in humans. Because the history of the human species is so short—probably about 200,000 years—the amount of DNA variation among humans is small compared to that of many other species. Much of our diversity seems to be in the form of single nucleotide polymorphisms (SNPs). SNPs are single base-pair sites where genetic variation is found in at least 1% of the population (see Concept 20.2); they are usually detected by DNA sequencing. In the human genome, SNPs occur on average about once in 100–300 base pairs. Scientists have already identified the locations of several million SNP sites in the human genome and continue to find additional SNPs. These are stored in databases around the world, one of which is run by the National Center for Biotechnology Information (NCBI) and can be accessed at http://www.ncbi.nlm.nih.gov/SNP/.

In the course of this search, they have also found other variations—including chromosomal regions with inversions, deletions, and duplications. The most surprising discovery has been the widespread occurrence of *copy-number variants* (*CNVs*), loci where some individuals have one or multiple copies of a particular gene or genetic region rather than the standard two copies (one on each homolog). CNVs result from regions of the genome being duplicated or deleted inconsistently within the population. One study of 40 people found more than 8,000 CNVs involving 13% of the genes in

▼ Figure 21.18 Inquiry

What is the function of a gene (*FOXP2*) that may be involved in language acquisition?

Experiment Several lines of evidence support a role for the *FOXP2* gene in the development of speech and language in humans and of vocalization in other vertebrates. In 2005, Joseph Buxbaum and collaborators at the Mount Sinai School of Medicine and several other institutions tested the function of *FOXP2*. They used the mouse, a model organism in which genes can be easily knocked out, as a representative vertebrate that vocalizes: Mice produce ultrasonic squeaks (whistles) to communicate stress. The researchers used genetic engineering to produce mice in which one or both copies of *FOXP2* were disrupted.

Wild type: two normal copies of *FOXP2*	Heterozygote: one copy of *FOXP2* disrupted	Homozygote: both copies of *FOXP2* disrupted

They then compared the phenotypes of these mice. Two of the characters they examined are included here: brain anatomy and vocalization.

Experiment 1: Researchers cut thin sections of brain and stained them with reagents that allow visualization of brain anatomy in a UV fluorescence microscope.



Experiment 2: To induce stress, researchers separated each newborn mouse pup from its mother and recorded the number of ultrasonic whistles produced by the pup.

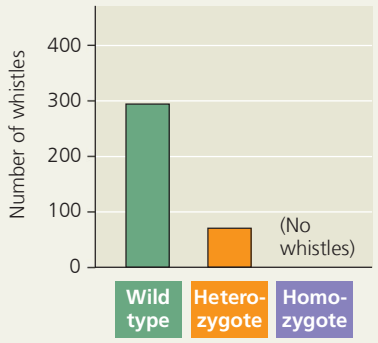
Results

Experiment 1 Results: Disruption of both copies of *FOXP2* led to brain abnormalities in which the cells were disorganized. Phenotypic effects on the brain of heterozygotes, with one disrupted copy, were less severe. (Each color in the micrographs below reveals a different cell or tissue type.)



Wild type	Heterozygote	Homozygote

Experiment 2 Results: Disruption of both copies of *FOXP2* led to an absence of ultrasonic vocalization in response to stress. The effect on vocalization in the heterozygote was also extreme.



Conclusion *FOXP2* plays a significant role in the development of functional communication systems in mice. The results augment evidence from studies of birds and humans, supporting the hypothesis that *FOXP2* may act similarly in diverse organisms.

Data from W. Shu et al., Altered ultrasonic vocalization in mice with a disruption in the *Foxp2* gene, *Proceedings of the National Academy of Sciences USA* 102:9643–9648 (2005).

WHAT IF? *Since the results support a role for mouse FOXP2 in vocalization, you might wonder whether the human FOXP2 protein is a key regulator of speech. If you were given the amino acid sequences of wild-type and mutant human FOXP2 proteins and the wild-type chimpanzee FOXP2 protein, how would you investigate this question? What further clues could you obtain by comparing these sequences to that of the mouse FOXP2 protein?*

the genome, and these CNVs probably represent just a small subset of the total. Since these variants encompass much longer stretches of DNA than the single nucleotides of SNPs, CNVs are more likely to have phenotypic consequences and to play a role in complex diseases and disorders. At the very least, the high incidence of copy-number variation blurs the meaning of the phrase "a normal human genome."

Copy-number variants, SNPs, and variations in repetitive DNA such as short tandem repeats (STRs) are useful genetic markers for studying human evolution. In one study, the genomes of two Africans from different communities were sequenced: Archbishop Desmond Tutu, the South African civil rights advocate of Xhosa and Tswana tribal heritage and a speaker of a Bantu language, and !Gubi, a hunter-gatherer from a Khoisan language-speaking community in the Kalahari Desert of Namibia. This community is believed to be the human group with the oldest known lineage. The comparison revealed many differences, as you might expect. The analysis was then broadened to compare the protein-coding regions of !Gubi's genome with those of three other Khoisan-speaking men also living in the Kalahari. Remarkably, the four genomes of these men living in the same area differed

more from each other than a European would from an Asian. These data highlight the extensive diversity among African genomes, consistent with our understanding that humans and their ancestors arose in Africa and that native Africans have therefore been evolving for longer than any migrant (non-African) population. Extending this approach will help answer questions about the differences between human populations and the migratory routes of human populations throughout history.

Mastering Biology
Interview with Charles Rotimi: Using genomics to study health-related conditions in African-Americans



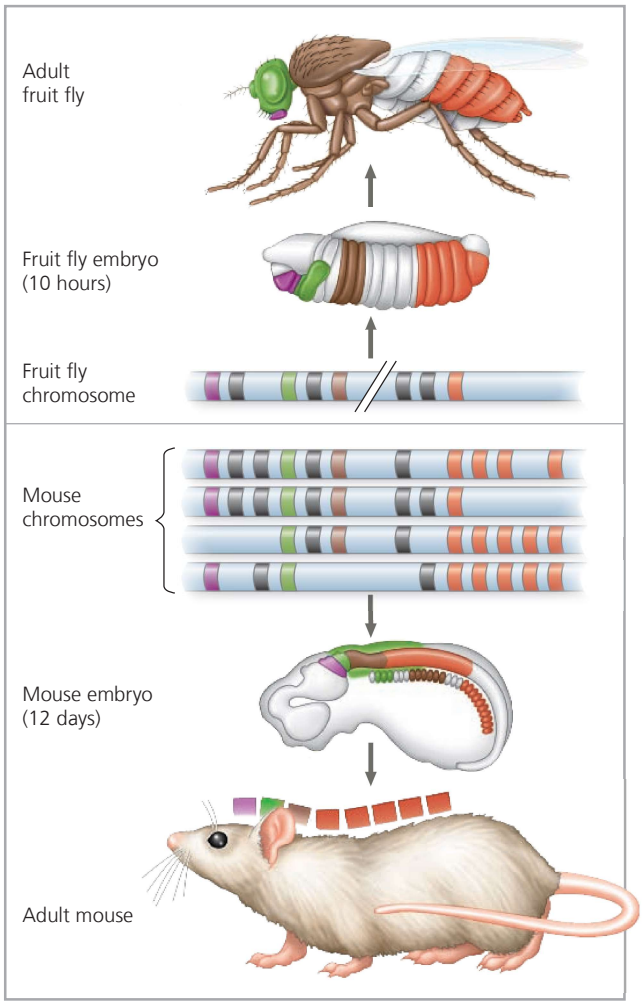
Widespread Conservation of Developmental Genes Among Animals

Biologists in the field of evolutionary developmental biology, or **evo-devo**, as it is often called, compare developmental processes of different multicellular organisms. Their aim is to understand how these processes have evolved and how changes in them can modify existing organismal features or lead to new ones. With the advent of molecular techniques and the recent flood of genomic information, we are beginning to realize that the genomes of related species with strikingly different forms may have only minor differences in gene sequence or, perhaps more importantly, in gene regulation. Discovering the molecular basis of these differences in turn helps us understand the origins of the myriad diverse forms that cohabit this planet, thus informing our study of the evolution of life.

In Concept 18.4, you learned about the homeotic genes in *Drosophila melanogaster*, which encode transcription factors that regulate gene expression and specify the identity of body segments in the fruit fly (see Figure 18.20). Molecular analysis of the homeotic genes in *Drosophila* has shown that they all include a 180-nucleotide sequence called a **homeobox**, which codes for a 60-amino-acid *homeodomain* in the encoded proteins. An identical or very similar nucleotide sequence has been discovered in the homeotic genes of many invertebrates and vertebrates. The resemblance even extends to the organization of these genes: The vertebrate genes homologous to the homeotic genes of fruit flies have kept the same chromosomal arrangement **(Figure 21.19)**. The similarities in sequence and organization are so striking that one researcher has whimsically referred to flies as "little people with wings." Homeobox-containing sequences have also been found in regulatory genes of much more distantly related eukaryotes, including plants and yeasts. From these similarities, we can deduce that the homeobox DNA sequence evolved very early and was sufficiently valuable to organisms to have been conserved in animals and plants virtually unchanged for hundreds of millions of years.

Homeotic genes in animals were named *Hox* genes, short for homeobox-containing genes, because homeotic genes were the first genes found to have this sequence. Other homeobox-containing genes were later found that do not act as homeotic genes; that is, they do not directly control the identity of body parts. However, most of these genes, in animals at least, are associated with development, suggesting their ancient and fundamental importance in that process. In *Drosophila*, for example, homeoboxes are present not only in the homeotic genes but also in the egg-polarity gene *bicoid* (see Figures 18.21 and 18.22), in several of the segmentation genes, and in a master regulatory gene for eye development.

▼ Figure 21.19 Conservation of homeotic genes in a fruit fly and a mouse. Homeotic genes that control the form of anterior and posterior structures of the body occur in the same linear sequence on chromosomes in Drosophila and mice. Each colored band on the chromosomes shown here represents a homeotic gene. In fruit flies, all homeotic genes are found on one chromosome. The mouse and other mammals have the same or similar sets of genes on four chromosomes. The color code indicates the parts of the embryos in which these genes are expressed and the adult body regions that result. All of these genes are essentially identical in flies and mice, except for those represented by black bands, which are less similar in the two animals.



Researchers have discovered that the homeobox-encoded homeodomain is the part of the protein that binds to DNA when the protein functions as a transcription factor. Elsewhere in the protein, domains that are more variable interact with other transcription factors, allowing the homeodomain-containing protein to recognize specific enhancers and regulate the associated genes. Proteins with homeodomains probably regulate development by coordinating the transcription of batteries of developmental genes, switching them on or off. In embryos of *Drosophila* and other animal species, different combinations of homeobox genes are active in different parts of the embryo. This selective expression of regulatory genes, varying over time and space, is central to pattern formation.

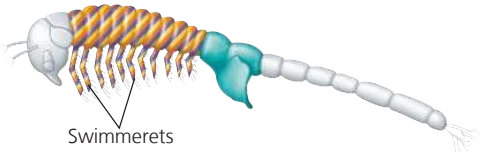
Developmental biologists have found that in addition to homeotic genes, many other genes involved in development are highly conserved from species to species. These include numerous genes encoding components of signaling pathways. The extraordinary similarity among some developmental genes in different animal species raises a question: How can the same genes be involved in the development of animals whose forms are so very different from each other?

In some cases, small changes in regulatory sequences of particular genes cause changes in gene expression patterns that can lead to major changes in body form. For example, the differing patterns of expression of the *Hox* genes along the body axis in a crustacean and an insect can explain the variation in number of leg-bearing segments among these closely related animals **(Figure 21.20)**. In other cases, similar genes direct different developmental processes in various organisms, resulting in diverse body shapes. Several *Hox* genes, for instance, are expressed in the embryonic and larval stages of the sea urchin, a nonsegmented animal that has a body plan quite different from those of insects and mice. Sea urchin adults make the pincushion-shaped shells you may have seen on the beach; two species of live sea urchins are shown in the photo. Sea urchins are among the organisms long used in classical embryological studies (see Concept 47.2).

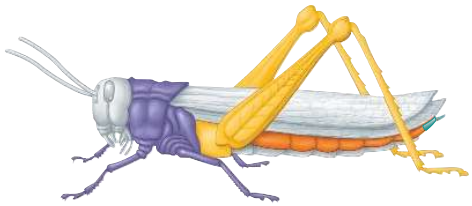
▼ **Two species of sea urchins**



▼ **Figure 21.20 Effect of differences in Hox gene expression in a crustacean and an insect.** Changes in the expression patterns of Hox genes have occurred over evolutionary time since insects diverged from a crustacean ancestor. These changes account in part for the different body plans of **(a)** the brine shrimp Artemia, a crustacean, and **(b)** the grasshopper, an insect. Shown here are regions of the adult body color-coded for expression of four Hox genes that determine the formation of particular body parts during embryonic development. Each color represents a specific Hox gene.



(a) Expression of four *Hox* genes in the brine shrimp *Artemia*. Three of the *Hox* genes are expressed together in one region (indicated by stripes), specifying the identity of the segments that have swimmerets. The fourth (teal) specifies the identity of the genital segments.



(b) Expression of the grasshopper versions of the same four *Hox* genes. In the grasshopper, each *Hox* gene is expressed in a discrete region and specifies the identity of that region.

In this final chapter of the genetics unit, you have learned how studying genomic composition and comparing the genomes of different species can illuminate the process by which genomes evolve. Furthermore, comparing developmental programs, we can see that the unity of life is reflected in the similarity of molecular and cellular mechanisms used to establish body pattern, although the genes directing development may differ among organisms. The similarities between genomes reflect the common ancestry of life on Earth. But the differences are also crucial, for they have created the huge diversity of organisms that have evolved. In the remaining chapters, we expand our perspective beyond the level of molecules, cells, and genes to explore this diversity on the organismal level.

CONCEPT CHECK 21.6

1. Would you expect the genome of the macaque (a monkey) to be more like that of a mouse or that of a human? Explain.
2. DNA sequences called homeoboxes help homeotic genes in animals direct development. Given that they are common to flies and mice, explain why these animals are so different.
3. **WHAT IF?** There are three times as many human-specific insertions of *Alu* elements in the human genome as in the chimpanzee genome. How did these extra *Alu* elements arise in the human genome? Propose a role they might have played in the divergence of these two species.

For suggested answers, see Appendix A.

21 Chapter Review



Go to **Mastering Biology** for Assignments, the eText, the Study Area, and Dynamic Study Modules.

SUMMARY OF KEY CONCEPTS

To review key terms, go to the **Vocabulary Self-Quiz** in the **Mastering Biology** eText or Study Area, or go to goo.gl/zkjz9t.

CONCEPT 21.1

The Human Genome Project fostered development of faster, less expensive sequencing techniques *(pp. 443–444)*

- The **Human Genome Project** was largely completed in 2003, aided by major advances in sequencing technology.
- In the **whole-genome shotgun** approach, the whole genome is cut into many small, overlapping fragments that are sequenced; computer software then assembles the genome sequence.

? *How did the Human Genome Project result in more rapid, less expensive DNA-sequencing technology?*

CONCEPT 21.2

Scientists use bioinformatics to analyze genomes and their functions *(pp. 444–448)*

- Computer analysis of genome sequences aids **gene annotation**, the identification of protein-coding sequences. Methods to determine gene function include comparing sequences of newly discovered genes with those of known genes in other species and observing the effects of experimentally inactivating the genes.
- In **systems biology**, scientists use the computer-based tools of **bioinformatics** to compare genomes and study sets of genes and proteins as whole systems (**genomics** and **proteomics**). Studies include large-scale analyses of protein interactions, functional DNA elements, and genes contributing to medical conditions.

? *What has been the most significant finding of the ENCODE project? Why was the project expanded to include nonhuman species?*

CONCEPT 21.3

Genomes vary in size, number of genes, and gene density *(pp. 448–450)*

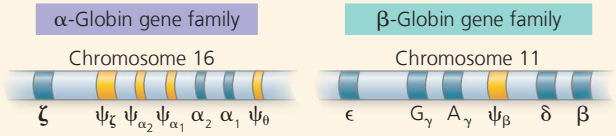
	Bacteria	Archaea	Eukarya
Genome size	Most are 1–6 Mb		Most are 10–4,000 Mb, but a few are much larger
Number of genes	1,500–7,500		Most are 5,000–45,000
Gene density	Higher than in eukaryotes		Lower than in prokaryotes (Within eukaryotes, lower density is correlated with larger genomes.)
Introns	None in protein-coding genes	Present in some genes	Present in most genes of multicellular eukaryotes, but only in some genes of unicellular eukaryotes
Other noncoding DNA	Very little		Can exist in large amounts; generally more repetitive noncoding DNA in multicellular eukaryotes

? *Compare genome size, gene number, and gene density (a) in the three domains and (b) among eukaryotes.*

CONCEPT 21.4

Multicellular eukaryotes have a lot of noncoding DNA and many multigene families *(pp. 450–453)*

- Only 1.5% of the human genome codes for proteins or gives rise to rRNAs or tRNAs; the rest is noncoding DNA, including **pseudogenes** and **repetitive DNA** of unknown function.
- The most abundant type of repetitive DNA in multicellular eukaryotes consists of **transposable elements** and related sequences. In eukaryotes, there are two types of transposable elements: **transposons**, which move via a DNA intermediate, and **retrotransposons**, which are more prevalent and move via an RNA intermediate.
- Other repetitive DNA includes short, noncoding sequences that are tandemly repeated thousands of times (**simple sequence DNA**, which includes **STRs**); these sequences are especially prominent in centromeres and telomeres, where they probably play structural roles in the chromosome.
- Though many eukaryotic genes are present in one copy per haploid chromosome set, others (most, in some species) are members of a gene family, such as the human globin gene families:



? *Explain how the function of transposable elements might account for their prevalence in human noncoding DNA.*

CONCEPT 21.5

Duplication, rearrangement, and mutation of DNA contribute to genome evolution *(pp. 454–459)*

- Errors in cell division can lead to extra copies of all or part of entire chromosome sets, which may then diverge if one set accumulates sequence changes. Polyploidy occurs more often among plants than animals and contributes to speciation.
- The chromosomal organization of genomes can be compared among species, providing information about evolutionary relationships. Within a given species, rearrangements of chromosomes are thought to contribute to the emergence of new species.
- The genes encoding the various related but different globin proteins evolved from one common ancestral globin gene, which duplicated and diverged into the α-globin and β-globin ancestral genes. Subsequent duplication and random mutation gave rise to the present globin genes, all of which code for oxygen-binding proteins. The copies of some duplicated genes have diverged so much that the functions of their encoded proteins (such as lysozyme and α-lactalbumin) are now substantially different.
- Rearrangement of exons within and between genes during evolution has led to genes containing multiple copies of similar exons and/or several different exons derived from other genes.
- Movement of transposable elements or recombination between copies of the same element can generate new sequence combinations that are beneficial to the organism. These may alter the functions of genes or their patterns of expression and regulation.

? *How could chromosomal rearrangements lead to the emergence of new species?*

CONCEPT 21.6

Comparing genome sequences provides clues to evolution and development *(pp. 459–464)*

- Comparisons of genomes from widely divergent and closely related species provide valuable information about ancient and more recent evolutionary history, respectively. Analysis of single nucleotide polymorphisms (SNPs) and copy-number variants (CNVs) among individuals in a species can also shed light on the evolution of that species.
- Evolutionary developmental (**evo-devo**) biologists have shown that homeotic genes and some other genes associated with animal development contain a **homeobox** region whose sequence is highly conserved among diverse species.

? *What type of information can be obtained by comparing the genomes of closely related species? Of very distantly related species?*

TEST YOUR UNDERSTANDING

For more multiple-choice questions, go to the **Practice Test** in the **Mastering Biology** eText or Study Area, or go to goo.gl/GruWRg.

Levels 1-2: Remembering/Understanding

1. Bioinformatics includes
   (A) using DNA technology to clone genes.
   (B) using computer programs to align DNA sequences.
   (C) using a person's genomic sequence to inform decisions about medical treatment.
   (D) amplifying DNA segments from a species' genome.
2. Homeotic genes
   (A) encode transcription factors that control the expression of genes responsible for specific anatomical structures.
   (B) are found only in *Drosophila* and other arthropods.
   (C) are the only genes that contain the homeobox domain.
   (D) encode proteins that form anatomical structures in the fly.

Levels 3-4: Applying/Analyzing

3. Two eukaryotic proteins have one domain in common but are otherwise very different. Which of the following processes is most likely to have contributed to this similarity?
   (A) gene duplication (C) exon shuffling
   (B) alternative splicing (D) random point mutations
4. **DRAW IT** Below are the amino acid sequences (using single letters; see Figure 5.14) of three short segments of the FOXP2 protein from five species. These segments contain all amino acid differences between the FOXP2 proteins of these species. Compare the amino acid sequences by answering parts (a)–(d).

Chimpanzee	PKSSD … TSSTT … NARRD
Mouse	PKSSE … TSSTT … NARRD
Gorilla	PKSSD … TSSTT … NARRD
Human	PKSSD … TSSNT … SARRD
Rhesus monkey	PKSSD … TSSTT … NARRD

   (a) Circle the names of any species that have identical amino acid sequences for the FOXP2 protein.
   (b) In the sequence for the mouse, circle any amino acid that differs from the sequence for the chimpanzee, gorilla, and rhesus monkey. Then draw a box around any amino acid that differs from the human sequence.
   (c) In the human sequence, underline any amino acid that differs from the sequence for the chimpanzee, gorilla, and rhesus monkey.
   (d) Primates and rodents diverged about 65 million years ago, and chimpanzees and humans diverged about 6 million years ago (see Figure 21.17). How many amino acid differences are there between the sequence for the mouse and the sequence for the chimpanzee, gorilla, and rhesus monkey? How many amino acid differences are there between the human sequence and the sequence for the chimpanzee, gorilla, and rhesus monkey? Based solely on the numbers of amino acid differences occurring over these time periods, what might you hypothesize about the rate of evolution of the *FOXP2* gene? Based on the information in the chapter regarding the *FOXP2* gene, is your hypothesis correct?

Levels 5-6: Evaluating/Creating

5. **EVOLUTION CONNECTION** Genes important in the embryonic development of animals, such as homeobox-containing genes, have been relatively well conserved during evolution; that is, they are more similar among different species than are many other genes. Explain why this is.
6. **SCIENTIFIC INQUIRY** The scientists mapping the SNPs in the human genome noticed that groups of SNPs tended to be inherited together, in blocks known as haplotypes, ranging in length from 5,000 to 200,000 base pairs. There are as few as four or five commonly occurring combinations of SNPs per haplotype. Integrating what you've learned throughout this chapter and this unit, propose an explanation for this observation.
7. **WRITE ABOUT A THEME: INFORMATION** The continuity of life is based on heritable information in the form of DNA. In a short essay (100–150 words), explain how mutations in protein-coding genes and regulatory DNA contribute to evolution.
8. **SYNTHESIZE YOUR KNOWLEDGE**



Insects have three thoracic (trunk) segments. While researchers have found insect fossils with wings on all segments, modern insects have wings or related structures on only the second and third segment. In modern insects, *Hox* gene products act to inhibit wing formation on the first segment. The treehopper insect (above) is an exception. In addition to the pair of prominent wings on its second segment, its first segment has an ornate helmet resembling a set of thorns, which is a modified, fused pair of "wings." (This provides camouflage in tree branches, reducing the risk of predation.) Explain how changes in gene regulation could have led to the evolution of such a structure.

For selected answers, see Appendix A.

Explore Scientific Papers with Science in the Classroom AAAS
How do transposable elements drive evolution?
Go to "Jumping Genes!" at www.scienceintheclassroom.org.
Instructors: Questions can be assigned in Mastering Biology.