10 Meiosis and Sexual Life Cycles

KEY CONCEPTS

10.1 Offspring acquire genes from parents by inheriting chromosomes
10.2 Fertilization and meiosis alternate in sexual life cycles
10.3 Meiosis reduces the number of chromosome sets from diploid to haploid
10.4 Genetic variation produced in sexual life cycles contributes to evolution

OVERVIEW

Variations on a Theme

Most people who send out birth announcements mention the sex of the baby, but they don't feel the need to specify that their offspring is a human being! One of the characteristics of life is the ability of organisms to reproduce their own kind—elephants produce little elephants, and oak trees generate oak saplings. Exceptions to this rule show up only as sensational but highly suspect stories in tabloid newspapers.

Another rule often taken for granted is that offspring resemble their parents more than they do unrelated individuals. If you examine the family members shown in **Figure 10.1**, you can pick out some similar features among them. The transmission of traits from one generation to the next is called inheritance, or **heredity** (from the Latin *heres*, heir). However, sons and daughters are not identical copies of either parent or of their siblings. Along with inherited similarity, there is also **variation**. Farmers have exploited the principles of heredity and variation for thousands of years, breeding plants and animals for desired traits. But what are the biological mechanisms leading to the hereditary similarity and variation that we call a "family resemblance"? The answer to this question eluded biologists until the advance of genetics in the 20th century.

▼ **Figure 10.1** What accounts for family resemblance?

Genetics is the scientific study of heredity and hereditary variation. In this unit, you'll learn about genetics at multiple levels, from organisms to cells to molecules. On the practical side, you'll see how genetics continues to revolutionize medicine, and you'll be asked to consider some social and ethical questions raised by our ability to manipulate DNA, the genetic material. At the end of the unit, you'll be able to stand back and consider the whole genome, an organism's entire complement of DNA. Rapid acquisition and analysis of the genome sequences of many species, including our own, have taught us a great deal about evolution on the molecular level—in other words, evolution of the genome itself. In fact, genetic methods and discoveries are catalyzing progress in all areas of biology, from cell biology to physiology, developmental biology, behavior, and even ecology.

We begin our study of genetics in this chapter by examining how chromosomes pass from parents to offspring in sexually reproducing organisms. The processes of meiosis (a special type of cell division) and fertilization (the fusion of sperm and egg) maintain a species' chromosome count during the sexual life cycle. We'll describe the cellular mechanics of meiosis and explain how this process differs from mitosis. Finally, we'll consider how both meiosis and fertilization contribute to genetic variation, such as the variation obvious in the family shown in Figure 10.1.

CONCEPT 10.1

Offspring acquire genes from parents by inheriting chromosomes

Family friends may tell you that you have your mother's freckles or your father's eyes. Of course, parents do not, in any literal sense, give their children freckles, eyes, hair, or any other traits. What, then, *is* actually inherited?

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 the 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 (see Chapter 3). 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.

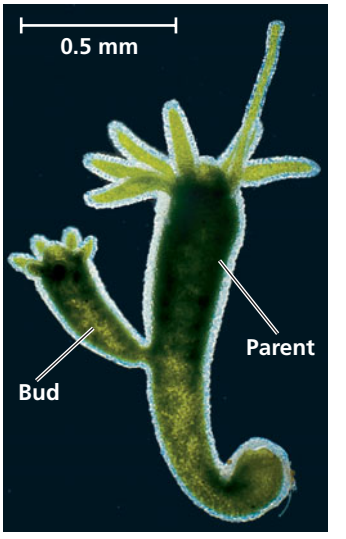
The transmission of hereditary traits has its molecular basis in the precise replication of DNA, which produces copies of genes that can be passed from parents to offspring. In animals and plants, reproductive cells called **gametes** are the vehicles that transmit genes from one generation to the next.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

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 the 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 of which is a specific sequence of nucleotides within 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 consists of the genes that are part of 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 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 10.2)**. Because the cells of the offspring are derived by mitosis in the parent, the "chip off the old block" is usually genetically identical to its parent. An individual that reproduces asexually gives rise to a **clone**, a group of genetically identical individuals. Genetic differences occasionally arise in asexually reproducing organisms as a result of changes in the DNA called mutations, which we will discuss in Chapter 14.

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 10.1 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.



(a) Hydra

(b) Redwoods

▲ **Figure 10.2 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.

CONCEPT CHECK 10.1

1. Explain what causes the traits of parents (such as hair color) to show up in their offspring.
2. How do asexually reproducing organisms produce offspring that are genetically identical to each other and to their parent?
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. To produce more plants like this one, should she cross-breed it with another plant or clone it? Why?

For suggested answers, see Appendix A.

CONCEPT 10.2

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, usually found in a diffused state throughout the nucleus. During mitosis, however, the chromosomes become condensed enough to be distinguished microscopically from each other. They differ in size, centromere position, and the pattern of 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 10.3)**. 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 homolog will also have a version of the same gene specifying eye color at the equivalent locus.

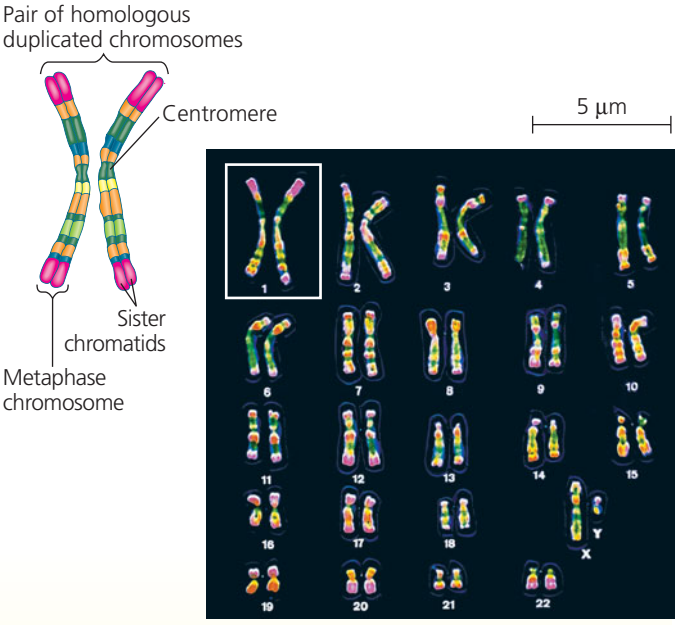
The two chromosomes referred to as X and Y are an important exception to the general pattern of homologous chromosomes in human somatic cells. Human females have a homologous pair of X chromosomes (XX), but males have one X and one Y chromosome (XY). Only small parts of the X and Y are homologous. Most of the genes carried on the X chromosome do not have counterparts on the tiny Y, and the Y chromosome has genes not present on the X. Because they determine an

▼ Figure 10.3 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 in metaphase, when chromosomes are most highly condensed, are stained and then viewed with a microscope equipped with a digital camera. A photograph of the chromosomes is displayed on a computer monitor, and the images of the chromosomes are arranged into pairs according to their appearance.



Results This karyotype shows the chromosomes from a normal human male. 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 a pair of homologous duplicated chromosomes).

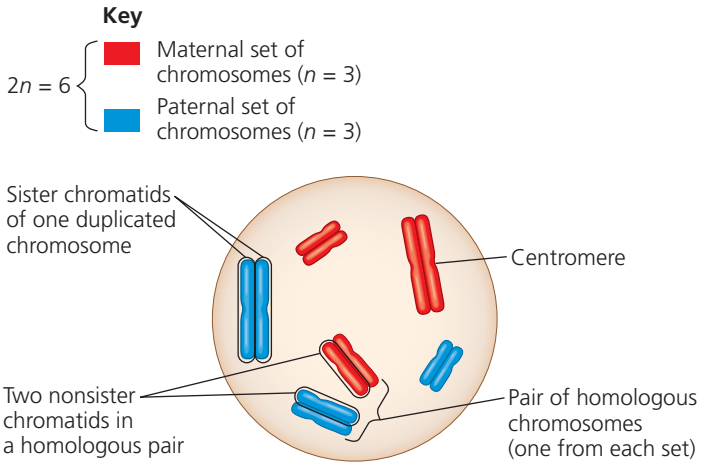
individual's sex, the X and Y chromosomes are called **sex chromosomes**. The other chromosomes are called **autosomes**.

The occurrence of pairs of homologous chromosomes in each human somatic cell is a consequence of our sexual origins. We inherit one chromosome of each 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. **Figure 10.4** helps clarify the various terms that we use to describe duplicated chromosomes in a diploid cell. Study this figure so that you understand the differences between homologous chromosomes, sister chromatids, nonsister chromatids, and chromosome sets.

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, but a sperm may contain an X or a Y chromosome.

Note that each sexually reproducing species has a characteristic diploid number 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 dogs have a diploid number of 78 and a haploid number of 39.

Now that you have learned the concepts of diploid and haploid numbers of chromosomes, let's consider chromosome behavior during sexual life cycles. We'll use the human life cycle as an example.



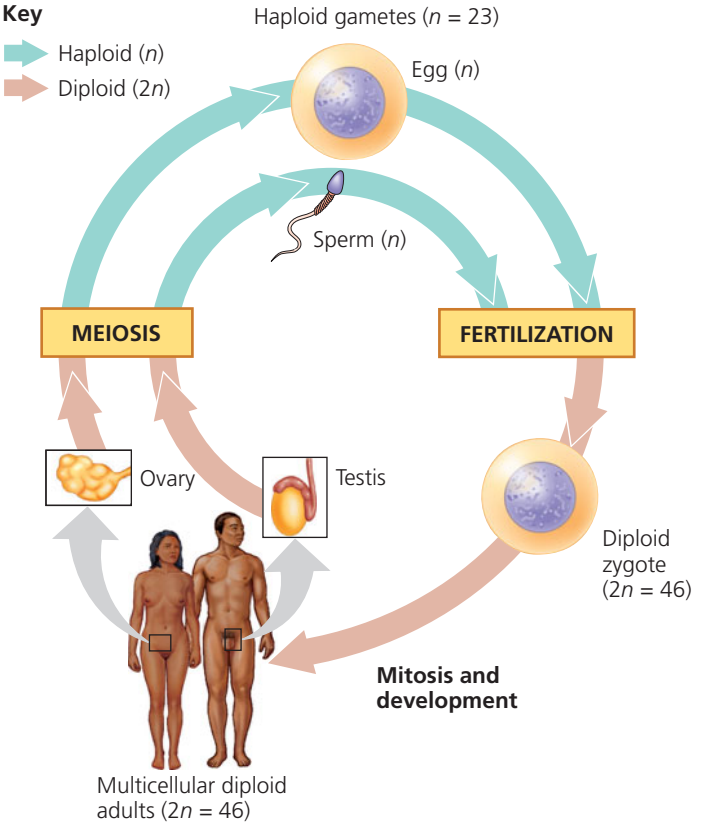
▲ Figure 10.4 Describing chromosomes. A cell from an organism with a diploid number of 6 ($2n = 6$) is depicted here following chromosome duplication and condensation. Each of the six duplicated chromosomes consists of two sister chromatids associated closely along their lengths. Each homologous pair 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. Nonsister chromatids are any two chromatids in a pair of homologous chromosomes that are not sister chromatids—in other words, one maternal and one paternal chromatid.

? *What is the haploid number of this cell? Is a "set" of chromosomes haploid or diploid? How many sets are present in this cell? In the karyotype in Figure 10.3?*

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. This union of gametes, culminating in fusion of their nuclei, is 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 **(Figure 10.5)**. Imagine what would happen if human gametes



▲ Figure 10.5 The human life cycle. In each generation, the number of chromosome sets doubles at fertilization but is halved during meiosis. 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.

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 sort of cell division called **meiosis**. This type of cell division reduces the number of sets of chromosomes from two to one in the gametes, counterbalancing the doubling that occurs at fertilization. In animals, meiosis occurs only in germ cells, which are in the ovaries or testes. As a result of meiosis, each human sperm and egg is haploid ($n = 23$). Fertilization restores the diploid condition by combining two haploid sets of chromosomes, and the human life cycle is repeated, generation after generation (see Figure 10.5). You will learn more about the production of sperm and eggs in Chapter 36.

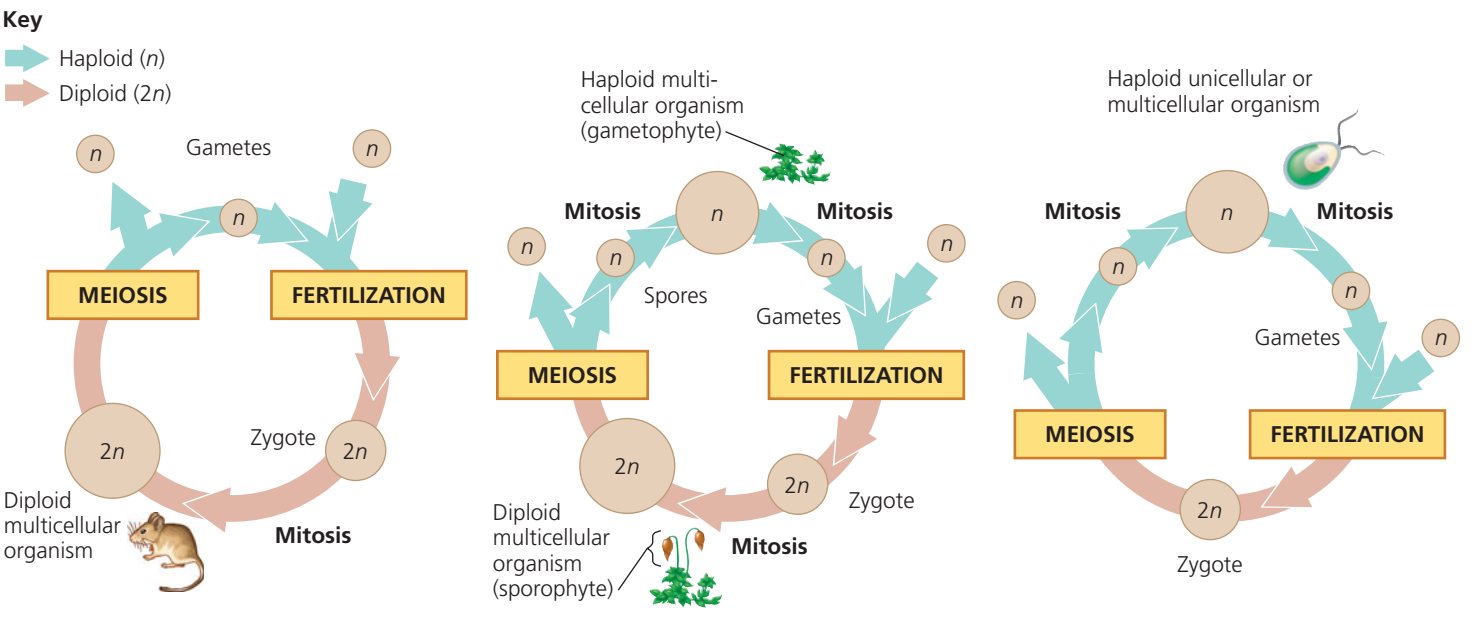
In general, the steps of the human life cycle are typical of many sexually reproducing animals. Indeed, the processes of fertilization and meiosis are the hallmarks of sexual reproduction in plants, fungi, and protists as well as in animals. Fertilization and meiosis alternate in sexual life cycles, maintaining a constant number of chromosomes in each 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. 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 **(Figure 10.6a)**.

Plants and some species of algae exhibit a second type of life cycle called **alternation of generations**. 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 the next sporophyte generation **(Figure 10.6b)**. The term *alternation of generations* fits well as a name for this type of life cycle.

A third type of life cycle occurs in most fungi and some protists, including some algae. After gametes fuse and form a diploid zygote, meiosis occurs without a multicellular diploid offspring developing. Meiosis produces not gametes but 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 **(Figure 10.6c)**.



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

▲ Figure 10.6 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.

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 because haploid cells 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. A closer look at meiosis will reveal the sources of this variation.

CONCEPT CHECK 10.2

1. **MAKE CONNECTIONS** In Figure 10.4, how many DNA molecules (double helices) are present (see Figure 9.5)?
2. How does the alternation of meiosis and fertilization in the life cycles of sexually reproducing organisms maintain the normal chromosome count for each species?
3. Each sperm of a pea plant contains seven chromosomes. What are the haploid and diploid numbers for this species?
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.

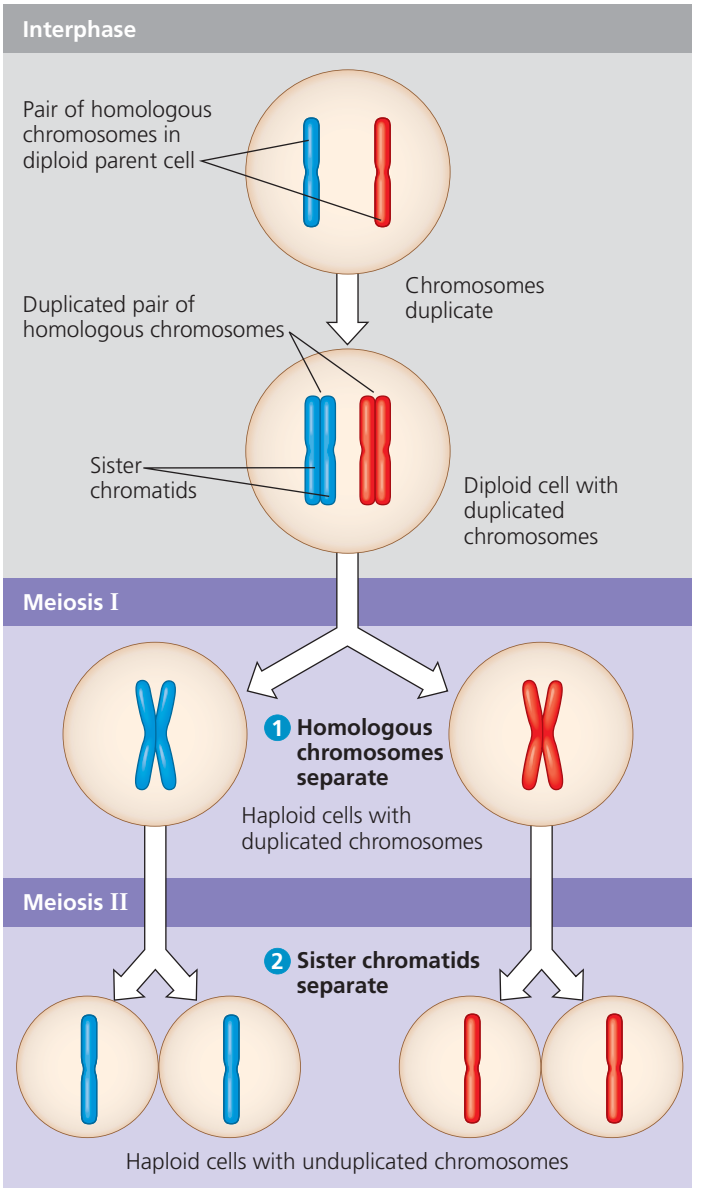
CONCEPT 10.3

Meiosis reduces the number of chromosome sets from diploid to haploid

Many of the steps of meiosis closely resemble corresponding steps in mitosis. Meiosis, like mitosis, is preceded by the duplication of chromosomes. However, this single duplication is followed not by one but by 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.

The Stages of Meiosis

The overview of meiosis in **Figure 10.7** 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 10.4). In contrast, the two chromosomes of a homologous pair are individual chromosomes that were inherited from different parents. Homologs appear alike in the microscope, but they may have different versions of genes, each called an *allele*, at corresponding loci (for example, an allele for freckles on one chromosome and an allele for the absence of freckles at the same locus on the homolog).



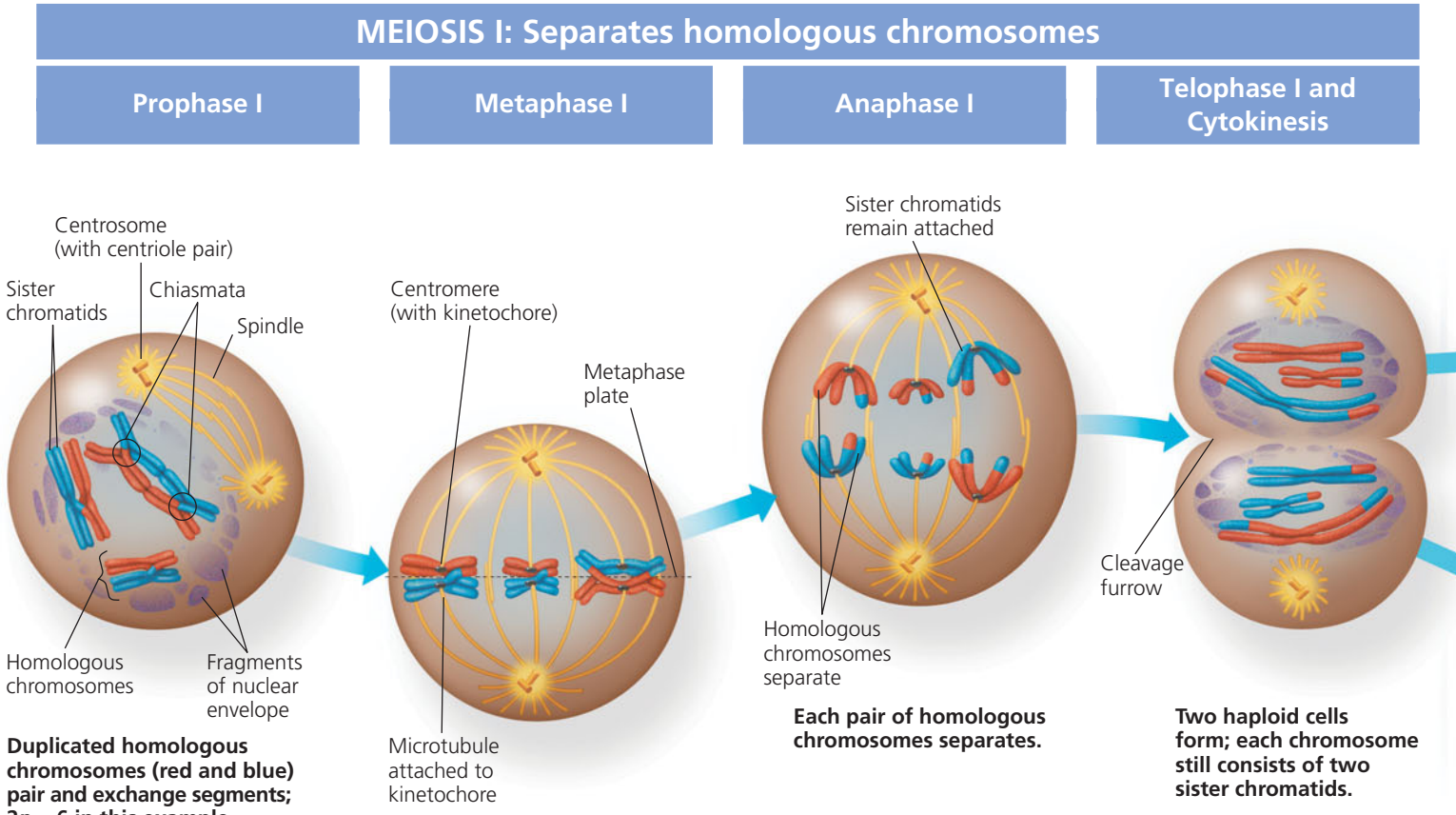
▲ Figure 10.7 Overview of meiosis: how meiosis reduces chromosome number. After the chromosomes duplicate in interphase, the diploid cell divides *twice*, yielding four haploid daughter cells. This overview tracks just one pair of homologous chromosomes, which for the sake of simplicity are drawn in the condensed state throughout. (They would not normally be condensed during interphase.) The red chromosome was inherited from the female parent, the blue chromosome from the male parent.

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

Homologs are not associated with each other in any obvious way except during meiosis, as you will soon see.

Figure 10.8, on the next two pages, describes in detail the stages of the two divisions of meiosis for an animal cell whose diploid number is 6. Meiosis halves the total number of chromosomes in a very specific way, reducing the number of sets from two to one, with each daughter cell receiving one set of chromosomes. Study Figure 10.8 thoroughly before going on.

▼ Figure 10.8 Exploring Meiosis in an Animal Cell



Prophase I

During early prophase I, before the stage shown above:

- Chromosomes begin to condense, and homologs loosely pair along their lengths, aligned gene by gene.
- Paired homologs become physically connected to each other along their lengths by a zipper-like protein structure, the *synaptonemal complex*; this state is called **synapsis**.
- **Crossing over**, a genetic rearrangement between nonsister chromatids involving the exchange of corresponding segments of DNA molecules, begins during pairing and synaptonemal complex formation and is completed while homologs are in synapsis.

At the stage shown above:

- Synapsis has ended with the disassembly of the synaptonemal complex in midprophase, and the chromosomes in each pair have moved apart slightly.
- Each homologous pair has one or more X-shaped regions called **chiasmata** (singular, *chiasma*). A chiasma exists at the point where a crossover has occurred. It appears as a cross because sister chromatid cohesion still holds the two original sister chromatids together, even in regions beyond the crossover point, where one chromatid is now part of the other homolog.
- Centrosome movement, spindle formation, and nuclear envelope breakdown occur as in mitosis.

Later in prophase I, after the stage shown above:

- Microtubules from one pole or the other attach to the two kinetochores, protein structures at the centromeres of the two homologs. The homologous pairs then move toward the metaphase plate.

Metaphase I

- Pairs of homologous chromosomes are now arranged at the metaphase plate, with one chromosome in each pair facing each pole.
- Both chromatids of one homolog are attached to kinetochore microtubules from one pole; those of the other homolog are attached to microtubules from the opposite pole.

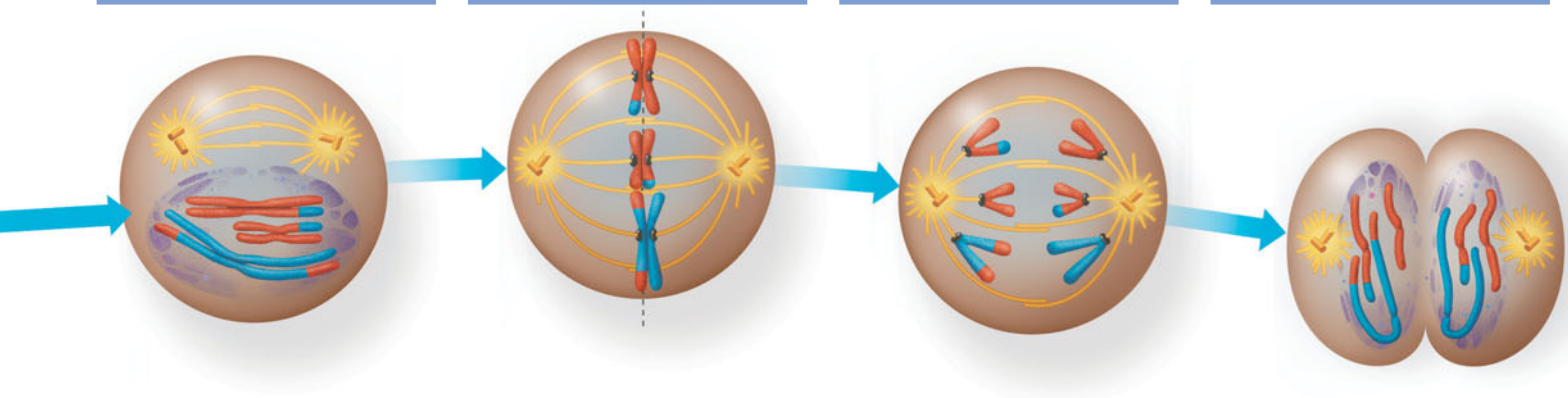
Anaphase I

- Breakdown of proteins 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 chromatids to move as a unit toward the same pole.

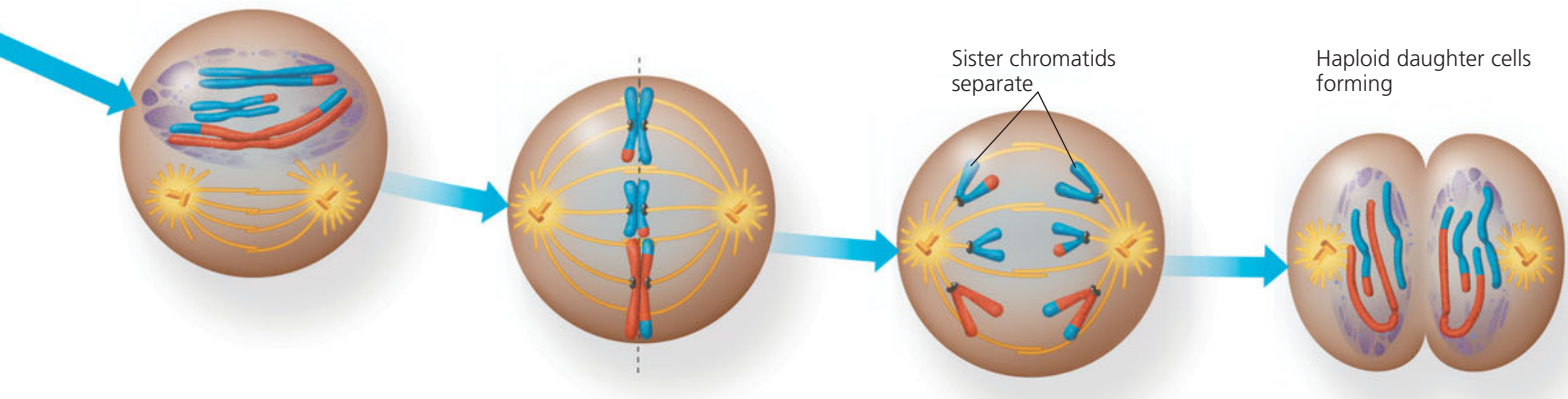
Telophase I and Cytokinesis

- At the beginning of telophase I, 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.

MEIOSIS II: Separates sister chromatids



During another round of cell division, the sister chromatids finally separate; four haploid daughter cells result, containing unduplicated chromosomes.



Prophase II

- A spindle apparatus forms.
- In late prophase II (not shown here), chromosomes, each still composed of two chromatids associated at the centromere, move 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.

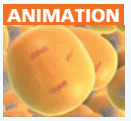
Anaphase II

- Breakdown of proteins holding the sister chromatids together at the centromere allows the chromatids to separate. The chromatids move toward opposite poles as individual chromosomes.

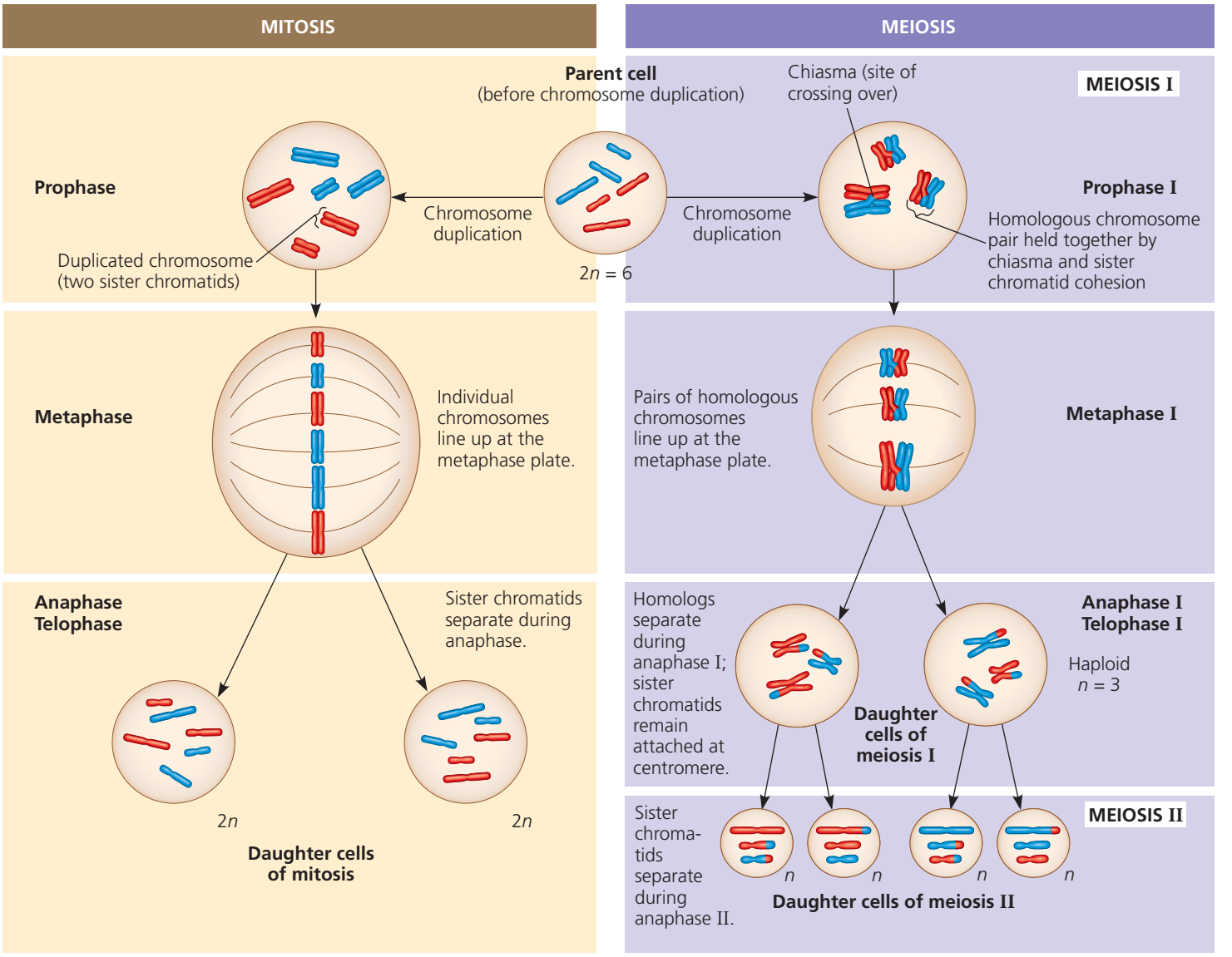
Telophase II and Cytokinesis

- 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 *Imagine the two daughter cells in Figure 9.7 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 10.8, after meiosis. What is it about the process of meiosis that accounts for this difference, even though meiosis also includes two cell divisions?*



BioFlix Visit the Study Area in **MasteringBiology** for the BioFlix® 3-D Animation on Meiosis.



SUMMARY

Property	Mitosis	Meiosis
DNA replication	Occurs during interphase before mitosis begins	Occurs during interphase before meiosis I begins
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 diploid ($2n$) and genetically identical to the parent cell	Four, each haploid (n), containing half as many chromosomes as the parent cell; genetically different from the parent cell and from each other
Role in the animal body	Enables multicellular adult to arise from zygote; produces cells for growth, repair, and, in some species, asexual reproduction	Produces gametes; reduces number of chromosome sets by half and introduces genetic variability among the gametes

▲ Figure 10.9 A comparison of mitosis and meiosis in diploid cells.

DRAW IT *Could any other combinations of chromosomes be generated during meiosis II from the specific cells shown in telophase I? Explain. (Hint: Draw all possible arrangements of chromosomes at metaphase II.)*

A Comparison of Mitosis and Meiosis

Figure 10.9 summarizes the key differences between meiosis and mitosis in diploid cells. Basically, meiosis reduces the number of chromosome sets from two to one, whereas mitosis conserves the number. Meiosis produces cells that differ genetically from their 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 the formation of the synaptonemal complex between them holds them in synapsis. Crossing over also occurs during prophase I. Synapsis and crossing over normally do not occur during prophase of mitosis.
2. **Homologous pairs at the metaphase plate.** At metaphase I of meiosis, chromosomes are positioned at the metaphase plate as pairs of homologs, rather than individual chromosomes, as in metaphase of mitosis.
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.

How do sister chromatids stay together through meiosis I but separate from each other in meiosis II and mitosis? Sister chromatids are attached along their lengths by protein complexes called *cohesins* and are said to exhibit *sister chromatid cohesion*. 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, one at the start of anaphase I and one at anaphase II. In metaphase I, homologs are held together by cohesion between sister chromatid arms in regions beyond points of crossing over, where stretches of sister chromatids now belong to different chromosomes. As shown in Figure 10.8, 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 is called the *reductional division* because it halves the number of chromosome sets per cell—a reduction from two sets (the diploid state) to one set (the haploid state). During the second meiotic division, meiosis II (sometimes called the *equational division*), the sister chromatids separate, producing haploid daughter cells. The mechanism for separating sister chromatids is virtually identical in meiosis II and mitosis. 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 from an experiment that tracked the amount of DNA in cells as they proceeded through the stages of meiosis.

CONCEPT CHECK 10.3

1. **MAKE CONNECTIONS** How are the chromosomes in a cell at metaphase of mitosis similar to and different from the chromosomes in a cell at metaphase of meiosis II? (Compare Figures 9.7 and 10.8.)
2. **WHAT IF?** Given that the synaptonemal complex has disappeared by the end of prophase, how would the two homologs be associated if crossing over did not occur? What effect might this ultimately have on gamete formation?

For suggested answers, see Appendix A.

CONCEPT 10.4

Genetic variation produced in sexual life cycles contributes to evolution

How do we account for the genetic variation among the family members illustrated in Figure 10.1? As you'll learn later in more detail, mutations are the original source of all 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. Let's examine three mechanisms that 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 the mother and 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 given daughter cell of meiosis I will get the maternal

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 them 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 = 1×10^{-15} gram).

Data from the Experiment

Time after Induction (hours)	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

Interpret the Data

1. First, set up your graph. (a) Place the labels for the independent variable and the dependent variable on the appropriate axes, followed by units of measurement in parentheses. Explain your choices. (b) Add tick marks and values for each axis in your graph. Note that while the timed samples were not all taken at equal intervals, the tick marks signifying the elapsed times along the *x*-axis should be regularly spaced and labeled. Explain your choices. (For additional information about graphs, see the Scientific Skills Review in Appendix F and in the Study Area in MasteringBiology.)
2. Because the variable on the *x*-axis varies continuously, it makes sense to plot the data on a line graph. (a) Plot each data point from the table onto the graph by placing a dot at the appropriate (*x, y*) coordinate. (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 9.2 and Figure 9.6.) At the end of meiosis I (MI)? At the end of meiosis II (MII)? (See Figure 10.7.) (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 10.7). (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 to the downward sloping line?
4. Given the fact that 1 fg of DNA = 9.78×10^5 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?

Further Reading G. Simchen, Commitment to meiosis: what determines the mode of division in budding yeast? *BioEssays* 31:169–177 (2009). doi 10.1002/bies.200800124

A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

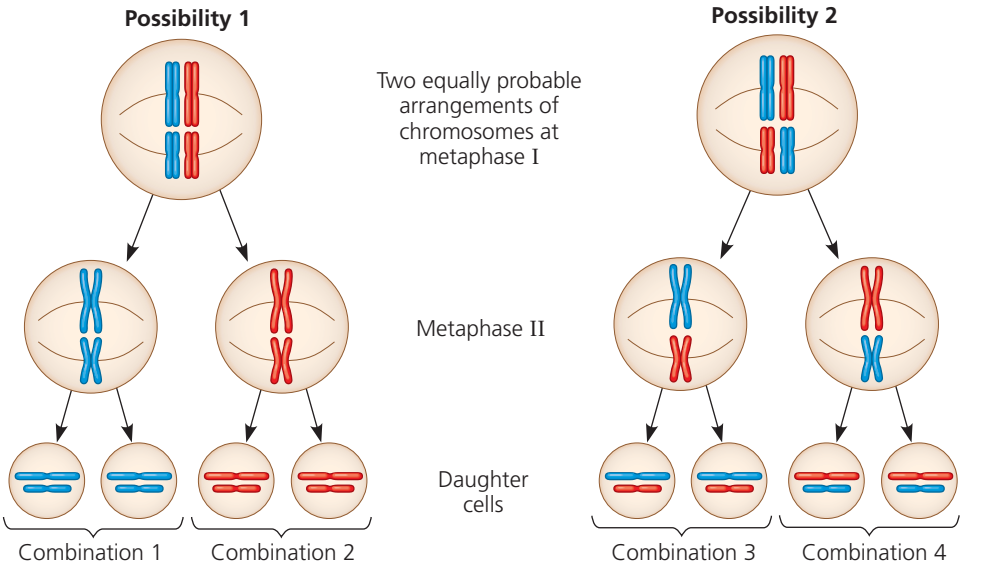
chromosome of a certain homologous pair and a 50% chance that it will get the paternal chromosome.

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 10.10**, the number of combinations possible for daughter cells formed by meiosis of a diploid cell with $n = 2$ (two pairs of homologous chromosomes) 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 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 organism.

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.

Crossing Over

As a consequence of the independent assortment of chromosomes during meiosis, each of us produces a collection of gametes differing greatly in their combinations of the chromosomes we inherited from our two parents. Figure 10.10 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) derived from two different parents **(Figure 10.11)**. In meiosis in humans, an average of one to three crossover events occur per chromosome pair, depending on the size of the chromosomes and the position of their centromeres.



▲ **Figure 10.10 The independent assortment of homologous chromosomes in meiosis.**

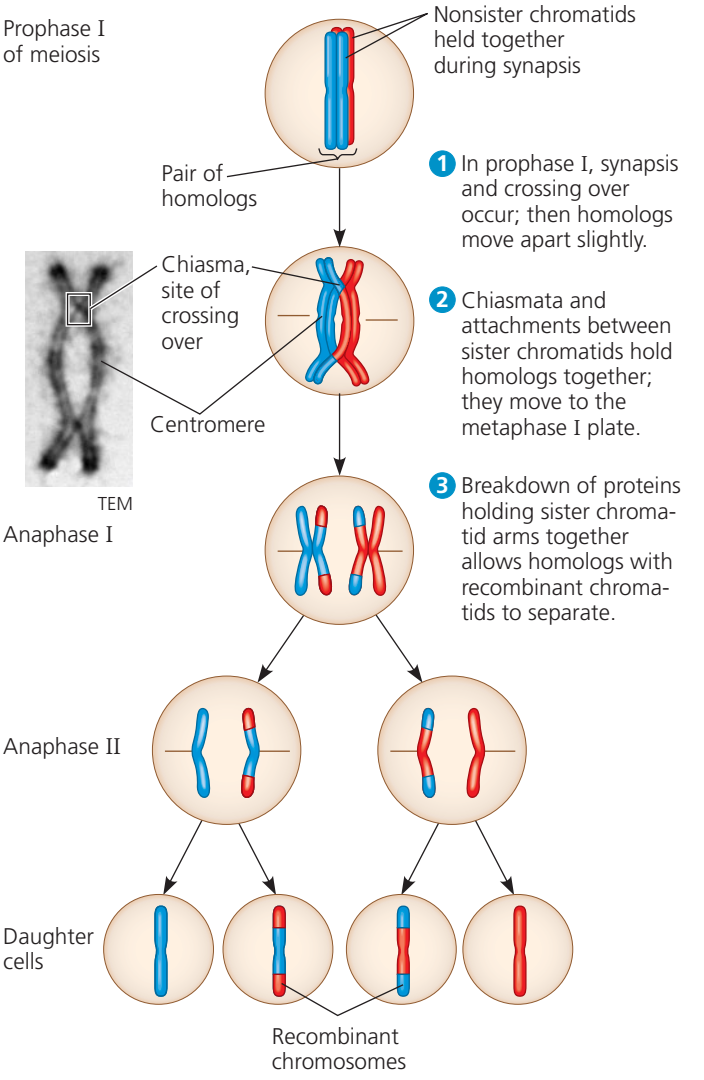
Crossing over begins very early in prophase I as homologous chromosomes pair loosely along their lengths. Each gene on one homolog is aligned precisely with the corresponding gene on the other homolog. In a single crossover event, the DNA of two *nonsister* chromatids—one maternal and one paternal chromatid of a homologous pair—is broken by specific proteins at precisely corresponding points, and the two segments beyond the crossover point are each joined to the other chromatid. Thus, a paternal (blue) chromatid is joined to a piece of maternal (red) chromatid beyond the crossover point, and vice versa. In this way, crossing over produces chromosomes with new combinations of maternal and paternal alleles (see Figure 10.11).

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. 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 Chapter 12. 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. It may sound trite, but you really *are* unique.



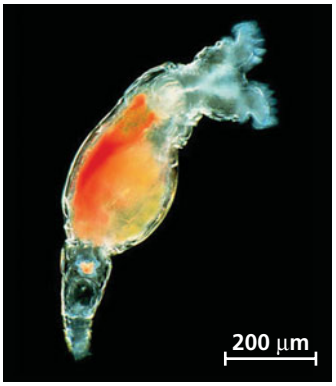
▲ **Figure 10.11 The results of crossing over during meiosis.**

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, let's see how the genetic variation in a population relates 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 by the 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 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. Consider the rare case of the bdelloid rotifer **(Figure 10.12)**. This group has apparently not reproduced sexually throughout the 40 million years of its evolutionary history. Does this mean that genetic diversity is not advantageous in this species? It turns out that bdelloid rotifers are an exception that proves the rule: This group has 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 rotifers and even other species. Evidence suggests that this DNA can become incorporated into the genome of the rotifer, leading to increased genetic diversity. This supports the idea that genetic diversity is advantageous, and that sexual reproduction has persisted because it generates such diversity.



▲ Figure 10.12 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 10.4

1. What is the original source of variation among the different alleles of a gene?
2. **WHAT IF?** Under what circumstances would crossing over during meiosis *not* contribute to genetic variation among daughter cells?

For suggested answers, see Appendix A.

10 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 10.1

Offspring acquire genes from parents by inheriting chromosomes (pp. 193–194)

- 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 10.2

Fertilization and meiosis alternate in sexual life cycles (pp. 194–197)

- 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 **homologous** pairs of **autosomes** and one pair of **sex chromosomes**; the latter 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.

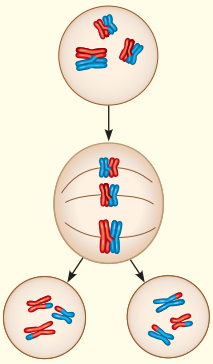
- Sexual **life cycles** differ in the timing of meiosis relative to fertilization and in the point(s) of the cycle at which a multicellular organism is produced by mitosis.

? *Compare the life cycles of animals and plants, mentioning their similarities and differences.*

CONCEPT 10.3

Meiosis reduces the number of chromosome sets from diploid to haploid (pp. 197–201)

- **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, the reductional division.
- Meiosis is distinguished from mitosis by three events of meiosis I:



Prophase I: Each homologous pair 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 then 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.

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

CONCEPT 10.4

Genetic variation produced in sexual life cycles contributes to evolution (pp. 201–204)

- Three events in sexual reproduction contribute to genetic variation in a population: independent assortment of chromosomes during meiosis, 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 diversity.

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

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

1. A human cell containing 22 autosomes and a Y chromosome is
a. a sperm. **b.** an egg. **c.** a zygote. **d.** a somatic cell of a male. **e.** a somatic cell of a female.

2. Homologous chromosomes move toward opposite poles of a dividing cell during
a. mitosis. **b.** meiosis I. **c.** meiosis II. **d.** fertilization. **e.** binary fission.

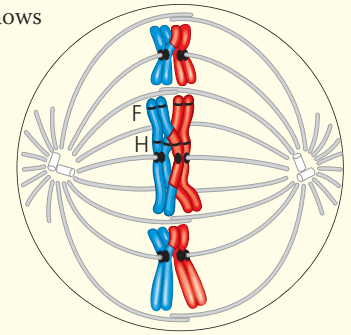
Level 2: Application/Analysis

3. 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 would be
a. $0.25x$. **b.** $0.5x$. **c.** x. **d.** $2x$. **e.** $4x$.

4. If we continued to follow the cell lineage from question 3, then the DNA content of a single cell at metaphase of meiosis II would be
a. $0.25x$. **b.** $0.5x$. **c.** x. **d.** $2x$. **e.** $4x$.

5. How many different combinations of maternal and paternal chromosomes can be packaged in gametes made by an organism with a diploid number of 8 ($2n = 8$)?
a. 2 **b.** 4 **c.** 8 **d.** 16 **e.** 32

6. DRAW IT The diagram at right 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, homologs, chiasma, sister chromatid cohesion, gene loci.
(b) Identify the stage of meiosis shown.
(c) Describe the makeup of a haploid set and a diploid set.

Level 3: Synthesis/Evaluation

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

8. SCIENTIFIC INQUIRY
The diagram above represents a meiotic cell. A previous study has shown 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. List other possible combinations of these alleles in this individual's gametes.

9. FOCUS ON EVOLUTION
Many species can reproduce either asexually or sexually. What might be the evolutionary significance of the switch from asexual to sexual reproduction that occurs in some organisms when the environment becomes unfavorable?

10. FOCUS ON 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.

For selected answers, see Appendix A.

11 Mendel and the Gene Idea

KEY CONCEPTS

11.1 Mendel used the scientific approach to identify two laws of inheritance

11.2 The laws of probability govern Mendelian inheritance

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

11.4 Many human traits follow Mendelian patterns of inheritance

OVERVIEW

Drawing from the Deck of Genes

Scanning the crowd at a soccer match attests to the marvelous variety and diversity of humankind. Brown, blue, green, or gray eyes; black, brown, blond, or red hair—these are just a few examples of heritable variations that we may observe among individuals in a population. What are the genetic principles that account for the transmission of such traits from parents to offspring in humans and other organisms?

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 in a manner analogous to the way 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. However, our everyday observations and the results of breeding experiments with animals and plants contradict that prediction. The blending hypothesis also fails to explain other phenomena of inheritance, such as traits reappearing after skipping a generation.

An alternative to the blending model is a "particulate" hypothesis of inheritance: the gene idea. According to this model, parents pass on discrete heritable units—genes—that retain their separate identities in offspring. An organism's collection of genes is more like a deck of cards than a bucket of paint. Like playing cards, genes can be shuffled and passed along, generation after generation, in undiluted form.

Modern genetics had its genesis in an abbey garden, where a monk named Gregor Mendel documented a particulate mechanism for inheritance. **Figure 11.1** shows Mendel (back row, holding a sprig of fuchsia) with his fellow monks. Mendel developed his theory of inheritance several decades before chromosomes were observed under the microscope and well before the significance of their behavior was understood. In this chapter, we will step into Mendel's garden to re-create his experiments and explain how he arrived at his theory of inheritance. We'll also explore inheritance patterns more complex than those observed by Mendel in garden peas. Finally, we'll see how the Mendelian model applies to the inheritance of human variations, including hereditary disorders such as sickle-cell disease.

▼ **Figure 11.1** What principles of inheritance did Gregor Mendel discover by breeding garden pea plants?

CONCEPT 11.1

Mendel used the scientific approach to identify two laws of inheritance

Mendel discovered the basic principles of heredity by breeding garden peas in carefully planned experiments. As we retrace his work, you'll recognize the key elements of the scientific process that were introduced in Chapter 1.

Mendel's Experimental, Quantitative Approach

One reason Mendel probably chose to work with peas is that they are available in 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**.

Mendel could strictly control mating between plants. 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 cross-pollination (fertilization between different plants), Mendel removed the immature stamens of a plant before they produced pollen and then dusted pollen from another plant onto the altered flowers **(Figure 11.2)**. Each resulting zygote then developed into a plant embryo encased in a seed (a pea). Mendel could thus 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, over many generations of self-pollination, had produced only the same variety as the parent plant. Such plants are said to be **true-breeding**. 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.

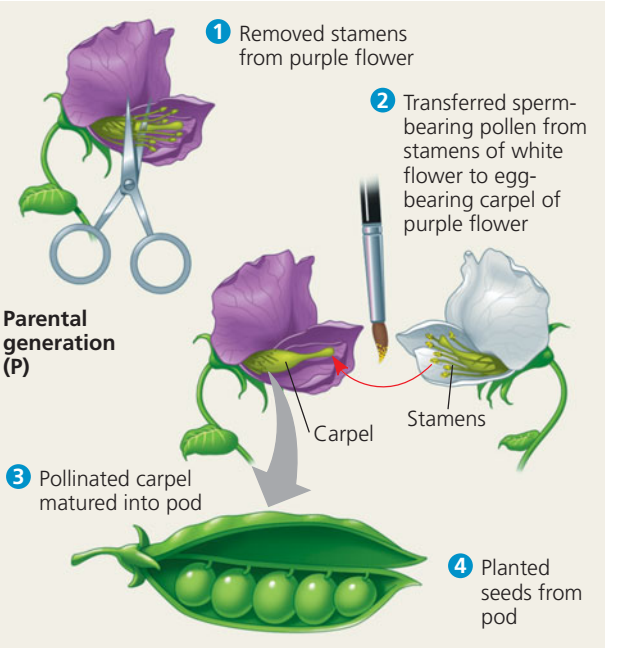
In a typical breeding experiment, Mendel cross-pollinated two contrasting, true-breeding pea varieties—for example, purple-flowered plants and white-flowered plants (see Figure 11.2). This mating, or *crossing*, of two true-breeding varieties is called **hybridization**. The true-breeding parents are 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 Mendel stopped his experiments with the F_1 generation, the basic patterns of inheritance would have escaped him.

▼ Figure 11.2 Research Method

Crossing Pea Plants

Application By crossing (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.

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, which have come to be called the law of segregation and the law of independent assortment.

The Law of Segregation

If the blending model of inheritance were correct, the F_1 hybrids from a cross between purple-flowered and white-flowered pea plants would have pale purple flowers, a trait intermediate between those of the P generation. Notice in Figure 11.2 that the experiment produced a very different result: All the F_1 offspring had flowers just as purple 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-pollinate and planted their seeds, the white-flower trait reappeared in the F_2 generation.

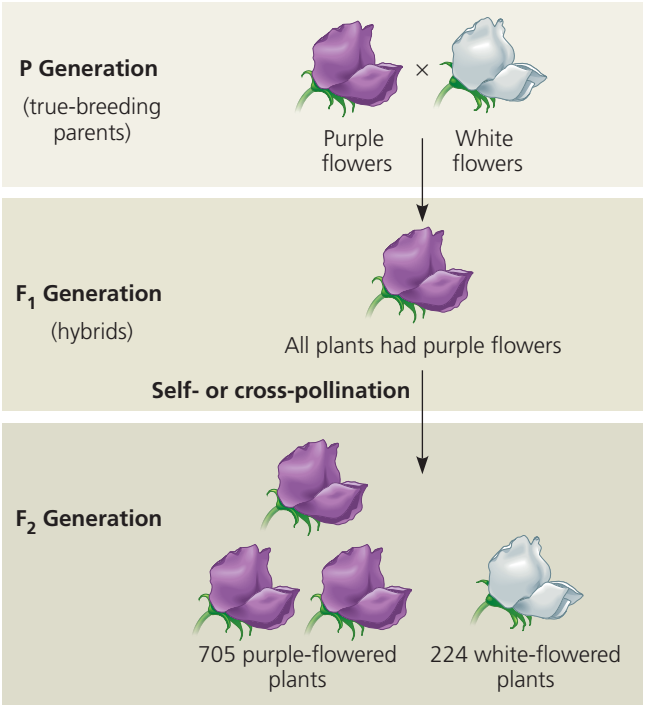
Mendel used very large sample sizes and kept accurate records of his results: 705 of the F_2 plants had purple flowers, and 224 had white flowers. These data fit a ratio of approximately three purple to one white **(Figure 11.3)**. Mendel reasoned that the heritable factor for white flowers did not disappear in the F_1 plants, but was somehow hidden, or masked, 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 11.1)**. 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 11.3. 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.")

▼ Figure 11.3 Inquiry

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

Experiment Around 1860, in a monastery garden in Brünn, Austria, Gregor Mendel used the character of flower color in pea plants to follow traits through two generations. He 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.

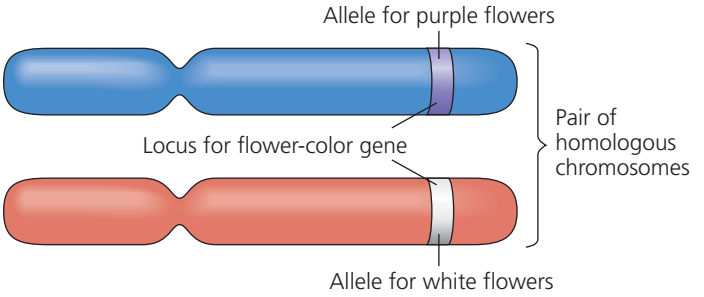
Conclusion The "heritable factor" for the recessive trait (white flowers) had not been destroyed, deleted, or "blended" in the F_1 generation but was merely masked by the presence of the factor for purple flowers, which is the dominant trait.

Source 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.

Table 11.1 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 shape	Inflated	×	Constricted	882:299	2.95:1
Pod color	Green	×	Yellow	428:152	2.82:1
Flower position	Axial	×	Terminal	651:207	3.14:1
Stem length	Tall	×	Dwarf	787:277	2.84:1



▲ Figure 11.4 Alleles, alternative versions of a gene. A somatic cell has two copies of each chromosome (forming a homologous pair) and thus two versions of each gene; the alleles may be identical or different. This figure depicts a pair of homologous chromosomes in an F_1 hybrid pea plant. The paternally inherited chromosome (blue), which was present in the sperm within a pollen grain, has an allele for purple flowers, and the maternally inherited chromosome (red), which was present in an egg within a carpel, has an allele for white flowers.

Mendel's Model

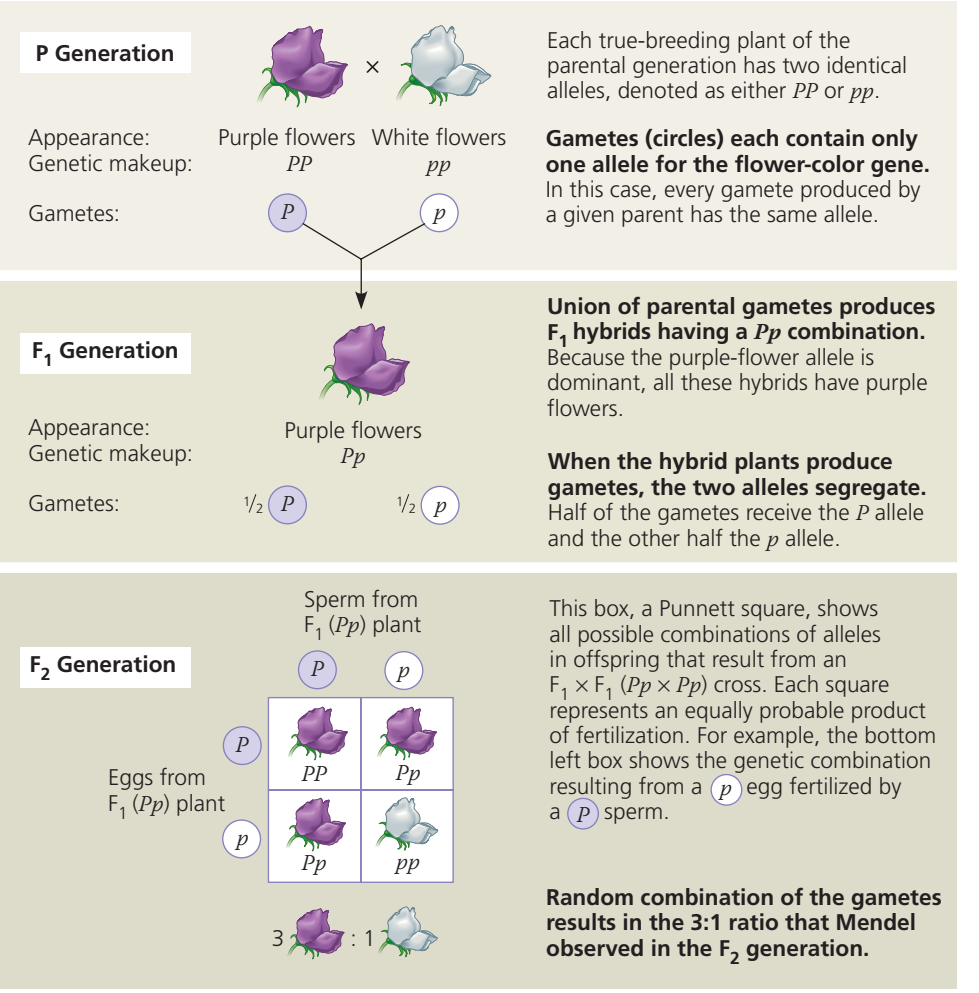
Mendel developed a model to explain the 3:1 inheritance pattern that he consistently observed among the F_2 offspring in his pea experiments. We describe four related concepts making up this model, the fourth of which is the law of segregation.

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 (Figure 11.4)**. Today, we can relate this concept to chromosomes and DNA: 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 and hence in its information content. The purple-flower allele and the white-flower allele are two DNA sequence variations possible at the flower-color locus on one of a pea plant's chromosomes.

Second, *for each character, an organism inherits two copies (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 Chapter 10). 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 11.4).

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 (separate from each other) during gamete formation and end up in different gametes* **(Figure 11.5)**. Thus, an egg or a sperm gets only one of the two alleles that are present in the somatic cells of the organism making the gamete. In terms of



▶ Figure 11.5 Mendel's law of segregation. This diagram shows the genetic makeup of the generations in Figure 11.3. 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.

chromosomes, this segregation corresponds to the distribution of the two members of a pair of homologous chromosomes to different gametes in meiosis (see Figure 10.7). Note that if an organism has identical alleles for a particular character—that is, the organism is true-breeding for that character—then that allele is present in all gametes. But if different alleles are 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 one 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 11.5 illustrates these combinations using a **Punnett square**, a handy diagrammatic device for predicting the allele composition of all offspring resulting 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; the gene itself is sometimes referred to as the *P*/*p* gene.

In the F_2 offspring, what color will the flowers be? One-fourth of the plants have inherited two purple-flower alleles; 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.

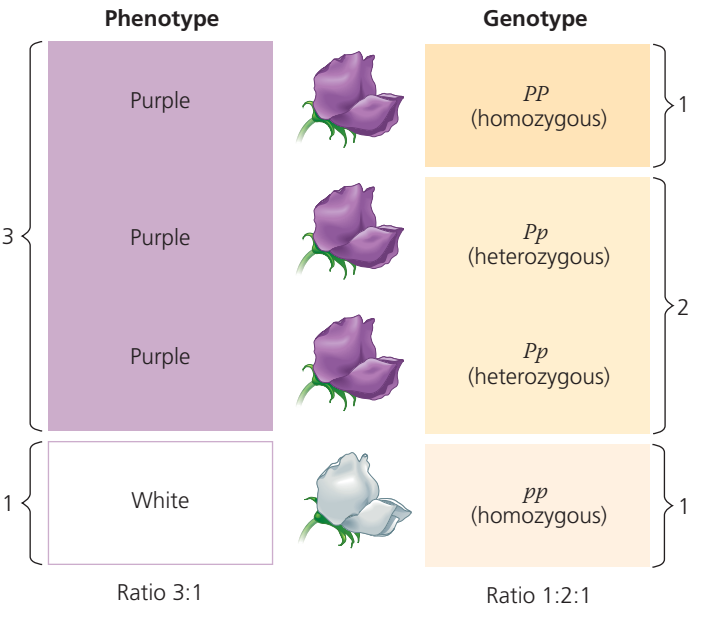
Useful Genetic Vocabulary

An organism that has a pair of identical alleles for a character is said to be **homozygous** for the gene controlling that character. In the parental generation in Figure 11.5, the purple 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 11.5). An organism that has two different alleles for a gene 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**. In the case of flower color in pea plants, *PP* and *Pp* plants have the same phenotype (purple) but different genotypes. **Figure 11.6** reviews these terms. Note that "phenotype" refers to physiological traits as well as traits that relate directly to appearance. For example, there is a pea variety that lacks the normal ability to self-pollinate. This physiological variation (non-self-pollination) is a phenotypic trait.

The Testcross

Suppose we have a "mystery" pea plant that has purple flowers. We cannot tell from its flower color if this plant 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 mystery plant will therefore determine the appearance of the offspring **(Figure 11.7)**. 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 appear among the offspring, then the purple-flowered parent must be heterozygous. The offspring of a *Pp* × *pp* cross will be expected to



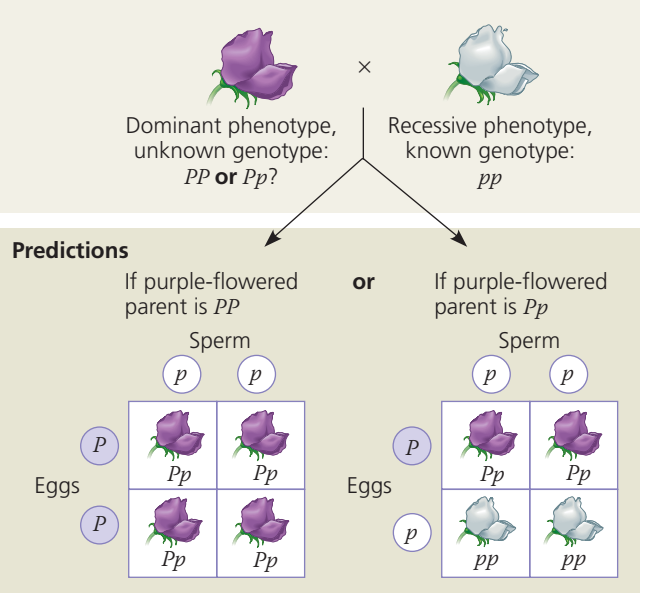
▲ **Figure 11.6 Phenotype versus genotype.** Grouping F_2 offspring from a cross for flower color according to phenotype results in the typical 3:1 phenotypic ratio. In terms of genotype, however, there are actually two categories of purple-flowered plants, *PP* (homozygous) and *Pp* (heterozygous), giving a 1:2:1 genotypic ratio.

▼ Figure 11.7 Research Method

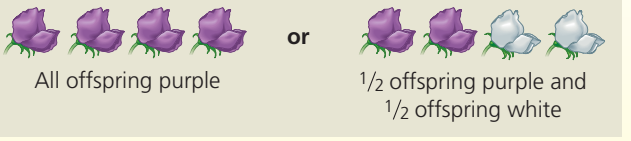
The Testcross

Application An organism that exhibits a dominant trait, 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.



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 an important tool of 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 identified his 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 (*Y*) is dominant and the allele for green seeds (*y*) is recessive. For the seed-shape character, the allele for round (*R*) is dominant, and the allele for wrinkled (*r*) is recessive.

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 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 11.8** 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*. This "dependent assortment" hypothesis predicts that the phenotypic ratio of the F_2 generation will be 3:1, just as in a monohybrid cross (see Figure 11.8, left side).

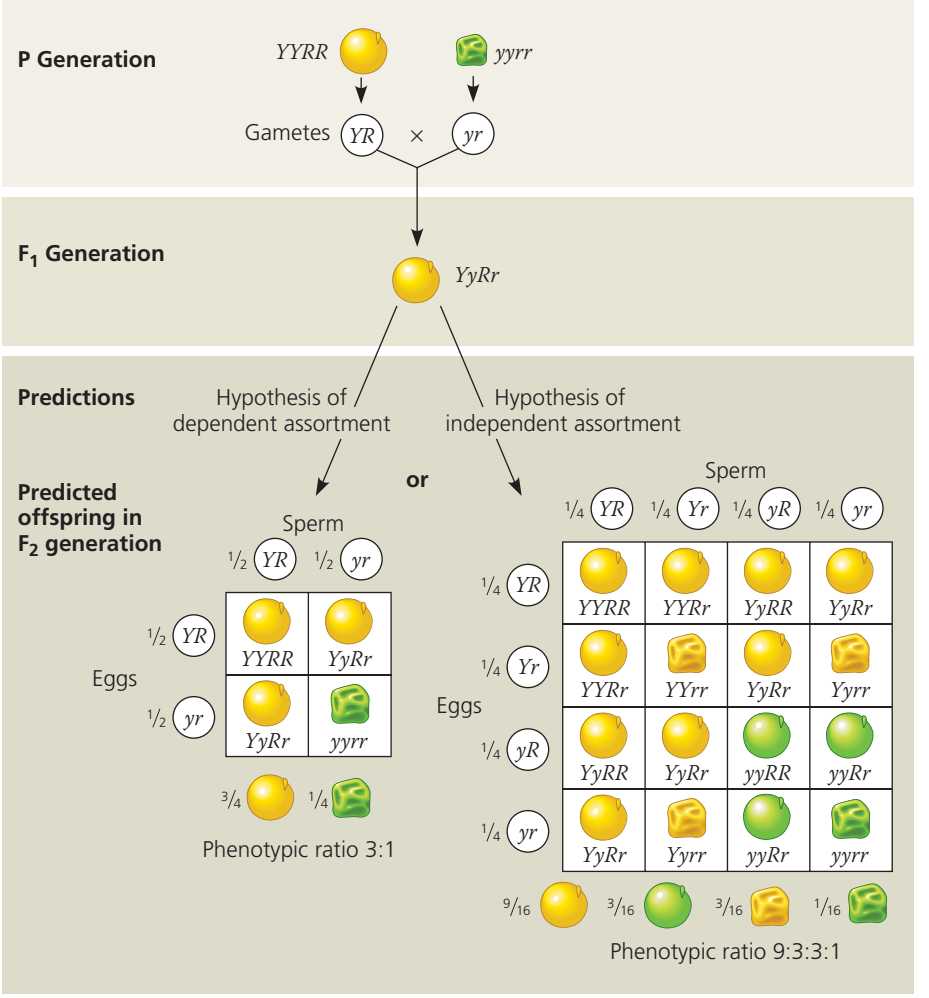
The alternative hypothesis is that the two pairs of alleles segregate independently of each other. In other words, genes are packaged into gametes in all possible allelic combinations, as long as each gamete has one allele for each gene. In our example, an F_1 plant will produce four classes of gametes 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 in Figure 11.8, right side. These combinations result in four phenotypic categories with a ratio of 9:3:3:1 (nine yellow-round to three green-round to three yellow-wrinkled to one 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—controlling seed color or seed shape, in this example—segregate into gametes independently of the alleles of other genes.

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. Does this override the 3:1 phenotypic ratio seen for the monohybrid cross shown in Figure 11.5? To investigate this question, let's consider one of the two dihybrid characters by itself: Looking only at pea color, we see that there are 416 yellow and 140 green peas—a 2.97:1 ratio, or roughly

▼ Figure 11.8 Inquiry

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

Experiment Gregor Mendel followed the characters of seed color and seed shape through the F_2 generation. He crossed two true-breeding plants, one with yellow-round seeds and one 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 Only the hypothesis of independent assortment predicts two of the observed phenotypes: green-round seeds and yellow-wrinkled seeds (see the right-hand Punnett square). The alleles for each gene segregate independently, and the two genes are said to assort independently.

Source 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?

3:1. In the dihybrid cross, the pea color alleles segregate as if this were a monohybrid cross. 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 each other pair during gamete formation.*

This law applies only to genes (allele pairs) located on different chromosomes—that is, on chromosomes that are not homologous—or very far apart on the same chromosome. (The latter case will be explained in Chapter 12, along with the more complex inheritance patterns of genes located near each other, which tend to be inherited together.) All the pea characters Mendel chose for analysis were controlled by genes on different chromosomes or far apart on one 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.

CONCEPT CHECK 11.1

1. In the dihybrid cross shown in Figure 11.8, calculate the phenotypic ratio in the F_2 generation, considering only the character of pea shape.
2. **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 11.1.)
3. List all gametes that could be made by a pea plant heterozygous for seed color, seed shape, and pod shape (*YyRrIi;* see Table 11.1). How large a Punnett square would you need to draw to predict the offspring of a self-pollination of this "trihybrid"?
4. **MAKE CONNECTIONS** In some pea plant crosses, the plants are self-pollinated. Explain whether self-pollination is considered asexual or sexual reproduction (refer back to Concept 10.1).

For suggested answers, see Appendix A.

CONCEPT 11.2

The laws of probability govern Mendelian inheritance

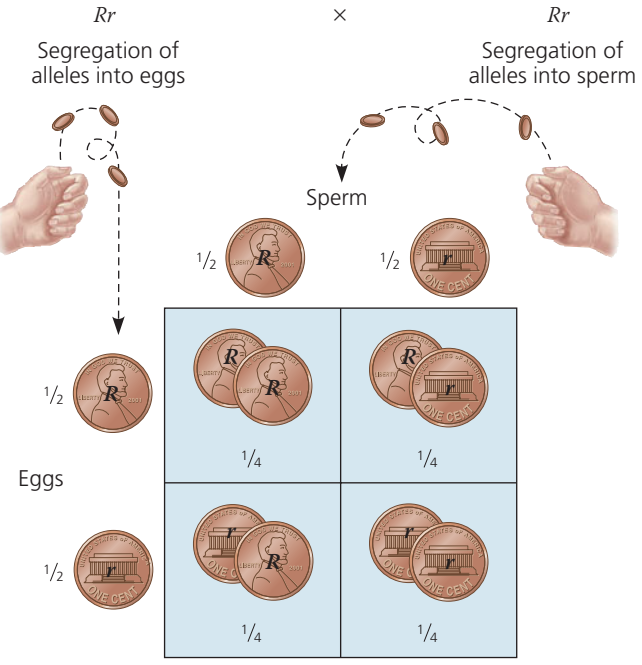
Mendel's laws of segregation and independent assortment reflect the same rules of probability that apply to tossing coins, rolling dice, and drawing cards from a deck. The probability scale ranges from 0 to 1. An event that is certain to occur has a probability of 1, while an event that is certain *not* 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 1/2, and the chance of tossing tails is 1/2. The probability of drawing the ace of spades from a 52-card deck is 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 51/52.

Tossing a coin illustrates an important lesson about probability. For every toss, the probability of heads is 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). Two basic rules of probability, described below, can help us predict the outcome of the fusion of such gametes in simple monohybrid crosses and more complicated crosses.

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 this probability, 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 $1/2 \times 1/2 = 1/4$.

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 1/2 chance of carrying the dominant allele (*R*) and a 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 both gametes at fertilization is found by multiplying 1/2 (the probability that the egg will have an *r*) × 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 1/4 **(Figure 11.9)**. Likewise, the probability of an F_2 plant carrying both dominant alleles for seed shape (*RR*) is 1/4.



▲ Figure 11.9 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).

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 11.9 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 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 1/4. The probability for the other possible way—the recessive allele from the egg and the dominant allele from the sperm—is also 1/4 (see Figure 11.9). Using the rule of addition, then, we can calculate the probability of an F_2 heterozygote as $1/4 + 1/4 = 1/2$.

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 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 11.8. 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 1/4 for *YY*, 1/2 for *Yy*, and 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: 1/4 *RR*, 1/2 *Rr*, and 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* = ¼ (probability of *YY*) × ¼ (*RR*) = 1/16

Probability of *YyRR* = ½ (*Yy*) × ¼ (*RR*) = ⅛

The *YYRR* genotype corresponds to the upper left box in the larger Punnett square in Figure 11.8 (one box = 1/16). Looking closely at the larger Punnett square in Figure 11.8, 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 is 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 offspring is 1/2 and the probability of homozygous 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, as shown below:

ppyyRr	¼ (probability of *pp*) × ½ (*yy*) × ½ (*Rr*)	= 1/16
ppYyrr	¼ × ½ × ½	= 1/16
Ppyyrr	½ × ½ × ½	= 2/16
PPyyrr	¼ × ½ × ½	= 1/16
ppyyrr	¼ × ½ × ½	= 1/16
Chance of *at least two* recessive traits		= 6/16 or 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 *chance* of various outcomes. Usually, the larger the sample size, the closer the results will conform to our predictions. The reason Mendel counted so many offspring from his crosses is that he understood this statistical feature of inheritance and had a keen sense of the rules of chance.

CONCEPT CHECK 11.2

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*). What fraction of offspring are predicted to be homozygous recessive for at least two of the three characters?

For suggested answers, see Appendix A.

CONCEPT 11.3

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

In the 20th century, 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'll extend Mendelian genetics to hereditary patterns that were not reported by Mendel.

Extending Mendelian Genetics for a Single Gene

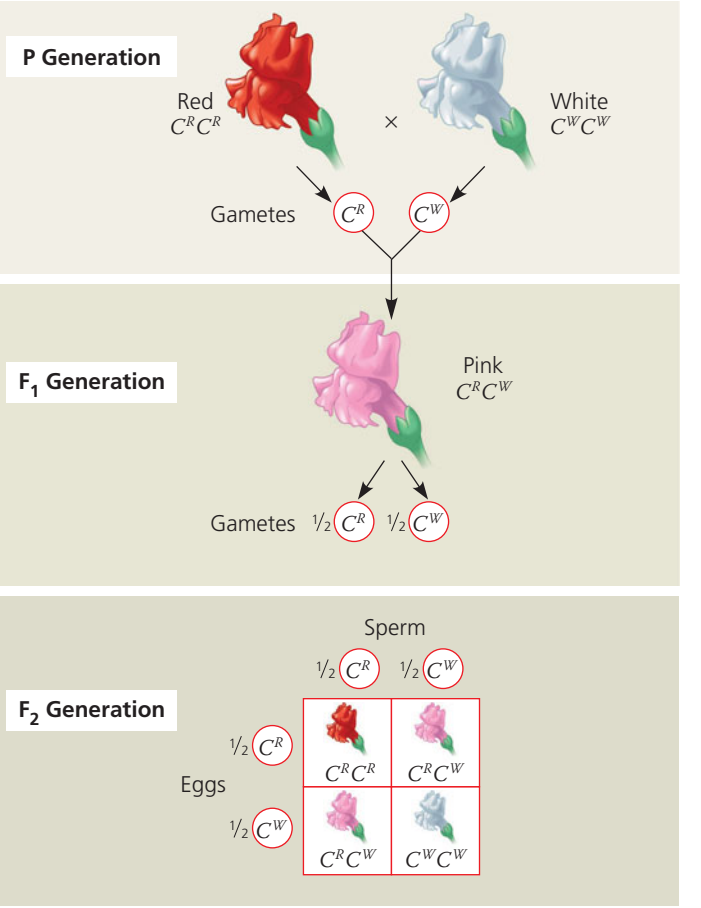
The inheritance of characters determined by a single gene deviates from simple Mendelian patterns when alleles are not completely dominant or recessive, when a particular gene has more than two alleles, or when a single gene produces multiple phenotypes. We'll describe examples of each of these situations 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.

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 11.10)**. 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 be retrieved from 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, 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, inheritance is particulate.



▲ Figure 11.10 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?*

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 locus, at which two allelic variations are possible, determines the phenotype of this blood group. Individuals homozygous for the *M* allele (*MM*) have red blood cells with only M molecules; individuals homozygous for the *N* allele (*NN*) 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 (*MN*). 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.

The Relationship Between Dominance and Phenotype

We've now seen that the relative effects of two alleles range from complete dominance of one allele, through 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. 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 the phenotype. **Tay-Sachs disease**, an inherited disorder in humans, provides 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 that in individuals homozygous for the normal allele and that in individuals with Tay-Sachs disease. 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 in a population than the recessive one, this is not always so. For example, about one baby out of 400 in the United States is born with extra digits (fingers or toes), a condition known as polydactyly. Some cases are caused by the presence of a dominant allele. The low frequency of polydactyly indicates that the recessive allele, which results in five digits per appendage when homozygous, is far more prevalent than the dominant allele. In Chapter 21, you'll learn how relative frequencies of alleles in a population are affected by natural selection.

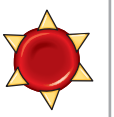
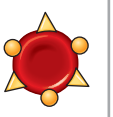
Multiple Alleles

Only two alleles exist for each of the seven pea characters that Mendel studied, but most genes exist in more than two allelic forms. The ABO blood groups in humans, for instance, are determined by three alleles of a single gene: I^A, I^B, and i. Each person has two alleles of the three for the blood group gene, which determines his or her blood group (phenotype): A, B, AB, or O. These letters refer to two carbohydrates—A and B—that may be found on the surface of red blood cells. A person's blood cells may have carbohydrate A (type A blood), carbohydrate B (type B), both (type AB), or neither (type O), as shown schematically in **Figure 11.11**. Matching compatible blood groups is critical for safe blood transfusions (see Chapter 35).

(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 appearance				
Phenotype (blood group)	A	B	AB	O

▲ Figure 11.11 Multiple alleles for the ABO blood groups. The four blood groups result from different combinations of three alleles.

? *Based on the surface carbohydrate phenotypes in (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 characteristics in an organism.

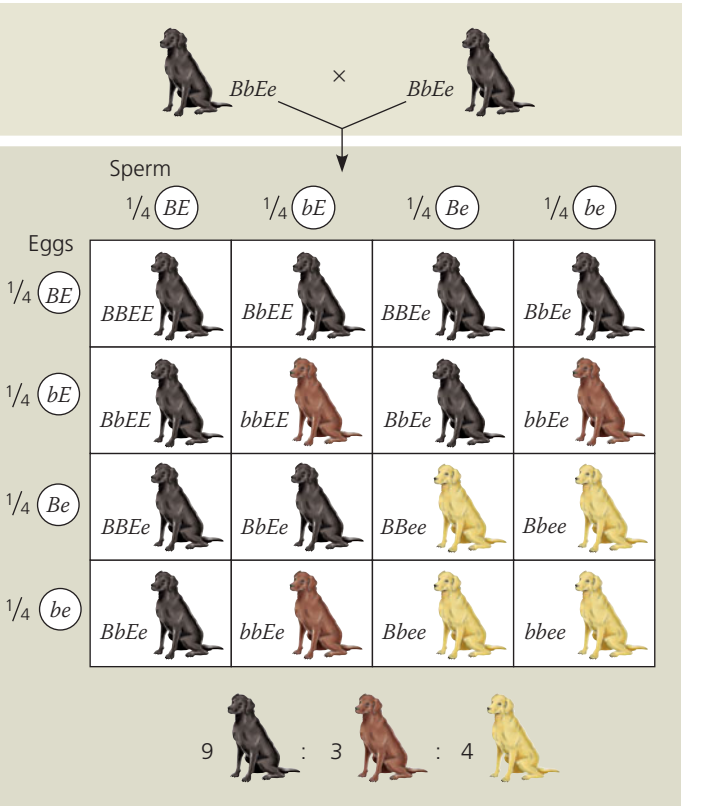
Extending Mendelian Genetics for Two or More Genes

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, one gene affects the phenotype of another because the two gene products interact, whereas in the second, 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. 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 11.12)**. As a result of epistasis, the phenotypic ratio among the F_2 offspring is nine black to three chocolate (brown) to four yellow. Other types of epistatic interactions produce different ratios, but all are modified versions of 9:3:3:1.

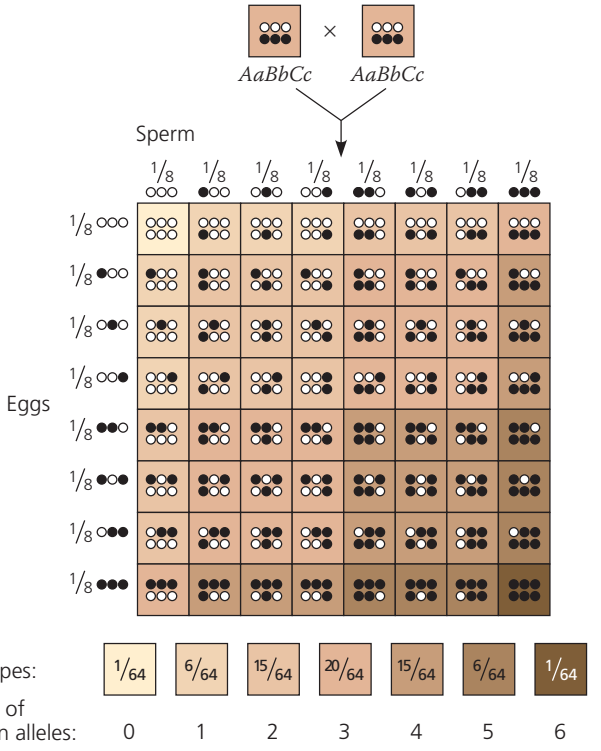


▲ **Figure 11.12 An example of epistasis.** This Punnett square illustrates the genotypes and phenotypes predicted for offspring of matings between two black Labrador retrievers of genotype *BbEe*. The *E*/*e* gene, which is epistatic to the *B*/*b* gene coding for hair pigment, controls whether or not pigment of any color will be deposited in the hair.

Polygenic Inheritance

Mendel studied characters that could be classified on an either-or basis, such as purple versus white flower color. But for many characters, such as human skin color and height, an either-or classification is impossible because the characters vary in the population in gradations along a continuum. These are called **quantitative characters**. Quantitative variation usually indicates **polygenic inheritance**, the 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.)

There is evidence, for instance, that skin pigmentation in humans is controlled by at least three separately inherited genes (probably more, but we will simplify). 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*). An *AABBCC* person would be very dark, while an *aabbcc* individual would be very light. 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. In a large



▲ Figure 11.13 A simplified model for polygenic inheritance of skin color. In this model, three separately inherited genes affect skin color. The heterozygous individuals (*AaBbCc*) represented by the two rectangles at the top of this figure each carry three dark-skin alleles (black circles, representing *A*, *B*, or *C*) and three light-skin alleles (white circles, representing *a*, *b*, or *c*). The Punnett square shows all the possible genetic combinations in gametes and offspring of many hypothetical matings between these heterozygotes. The results are summarized by the phenotypic frequencies (fractions) under the Punnett square. (The phenotypic ratio of the skin colors shown in the boxes is 1:6:15:20:15:6:1.)

number of such matings, the majority of offspring would be expected to have intermediate phenotypes (skin color in the middle range), as shown in **Figure 11.13**. 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. In 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 characteristics are more influenced by genes or the environment— in everyday terms, nature versus nurture—is a very old and hotly contested 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. For some characters, such as the ABO blood group system, the range is extremely narrow; that is, a given genotype mandates a very specific phenotype. Other characteristics, 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.

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.

Integrating a Mendelian View of Heredity and Variation

We have now broadened our view of Mendelian inheritance by exploring the degrees of dominance as well as multiple alleles, pleiotropy, epistasis, polygenic inheritance, and the phenotypic impact of the environment. Stepping back to see the big picture, 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 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 two laws, those of segregation and 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

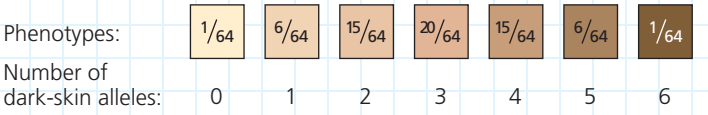
Scientific Skills Exercise

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 several different genes. In this exercise, you will work with a simplified model of skin-color genetics where three genes are assumed to affect the darkness of skin color and where each gene has two alleles—dark or light. In this model, each dark allele contributes equally to the darkness of skin color, and each pair of alleles segregates independently of each 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 F and in the Study Area in MasteringBiology.)

How This Model Is Analyzed To predict the phenotypes of the offspring of heterozygous parents, the ratios of the genes for this trait must be calculated. Figure 11.13 shows a simplified model for polygenic inheritance of skin color that includes three of the known genes. According to this model, three separately inherited genes affect the darkness of skin. 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 the possible genetic combinations in gametes and in offspring of a large number of hypothetical matings between these heterozygotes. The possible phenotypes are shown under the Punnett square.

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 phenotypic frequencies of individuals with the same number of dark-skin alleles.



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 number of offspring (out of 64) with each phenotype on the *y*-axis. There are no gaps in our allele data, so draw the bars side-to-side 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 dotted 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 11.13.)
3. If one of the three genes were lethal when homozygous recessive, what would happen to the distribution of phenotype frequencies? To determine this, use *bb* as an example of a lethal genotype. Using Figure 11.13, identify offspring where the center circle (the *B*/*b* gene) in both the top and bottom rows of the square is white, representing the homozygous state *bb*. Because *bb* individuals would not survive, cross out those squares, then count the phenotype frequencies of the surviving offspring according to the number of dark-skin alleles (0–6) and graph the new data. What happens to the shape of the curve compared with the curve in question 2? What does this indicate about the distribution of phenotype frequencies?

Further Reading R.A. Sturm, A golden age of human pigmentation genetics, *Trends in Genetics* 22: 464–468 (2006). doi:10.1016/j.tig.2006.06.010

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

garden came a theory of particulate inheritance that anchors modern genetics. In the last section of this chapter, we'll apply Mendelian genetics to human inheritance, with emphasis on the transmission of hereditary diseases.

CONCEPT CHECK 11.3

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 11.4

Many human traits follow Mendelian patterns of inheritance

Peas are convenient subjects for genetic research, but humans are not. The human generation span is long—about 20 years—and human parents produce many fewer offspring than peas and most other species. Even more important, it wouldn't be ethical to ask pairs of humans to breed so that the phenotypes of their offspring could be analyzed! In spite of these constraints, the study of human genetics continues, spurred on by our desire to understand our own inheritance. New molecular biological techniques have led to many breakthrough discoveries, but basic Mendelian genetics endures as the foundation of human genetics.

Unable to manipulate the matings of people, geneticists instead analyze results that have already occurred by collecting information about a family's history for a particular trait.

Pedigree Analysis

Geneticists assemble information about members of a family into a tree diagram that describes the traits of parents and children across the generations, called a **pedigree**.

Figure 11.14a shows a three-generation pedigree that traces the occurrence of a pointed contour of the hairline on the forehead. This trait, called a widow's peak, is due to a dominant allele, *W*. 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* × *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*).

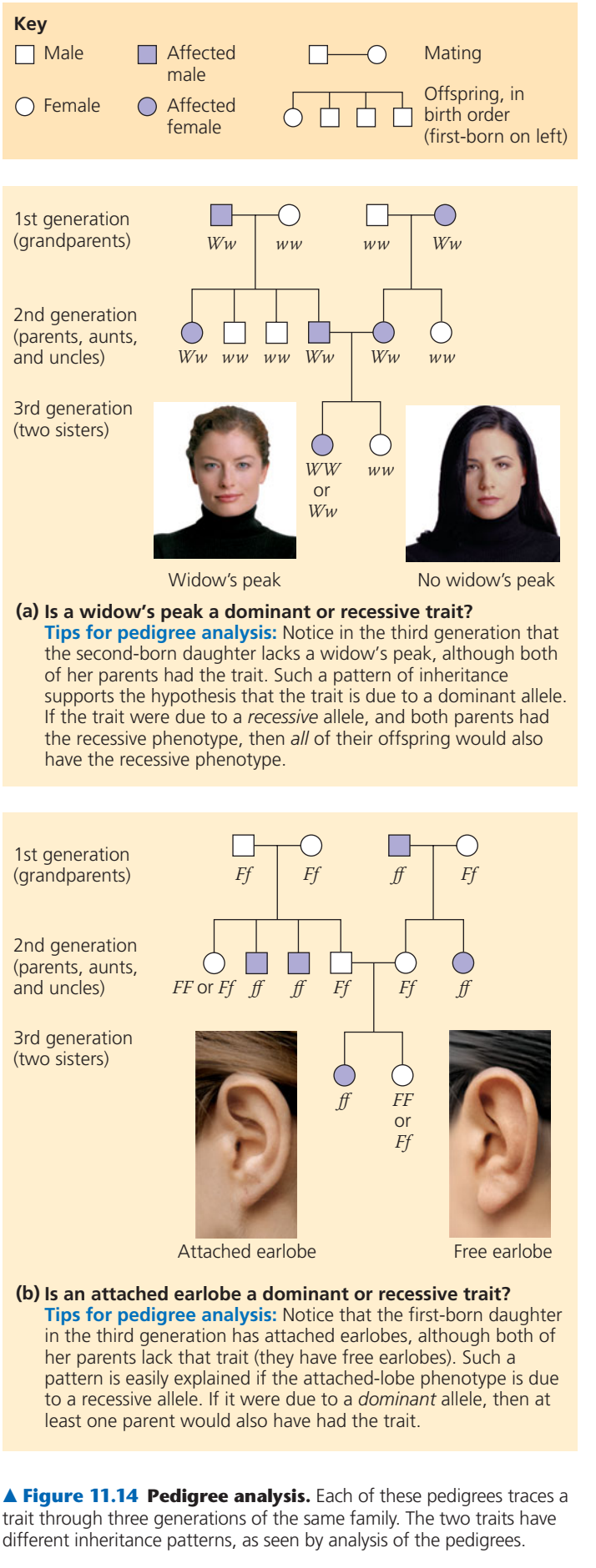
Figure 11.14b is a pedigree of the same family, but this time we focus on a recessive trait, attached earlobes. We'll use *f* for the recessive allele and *F* for the dominant allele, which results in free earlobes. As you work your way through the pedigree, notice once again that you can apply what 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 11.14 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 monohybrid cross (*Ww* × *Ww*), and thus the probability that a child will inherit a dominant allele and have a widow's peak is 3/4 (1/4 *WW* + 1/2 *Ww*). What is the probability that the child will have attached earlobes? Again, we can treat this as a monohybrid cross (*Ff* × *Ff*), but this time we want to know the chance that the offspring will be homozygous recessive (*ff*). That probability is 1/4. Finally, what is the chance that the child will have a widow's peak *and* attached earlobes? Assuming that the genes for these two characters are on different chromosomes, the two pairs of alleles will assort independently in this dihybrid cross (*WwFf* × *WwFf*). Thus, we can use the multiplication rule: 3/4 (chance of widow's peak) × 1/4 (chance of attached earlobes) = 3/16 (chance of widow's peak and attached earlobes).

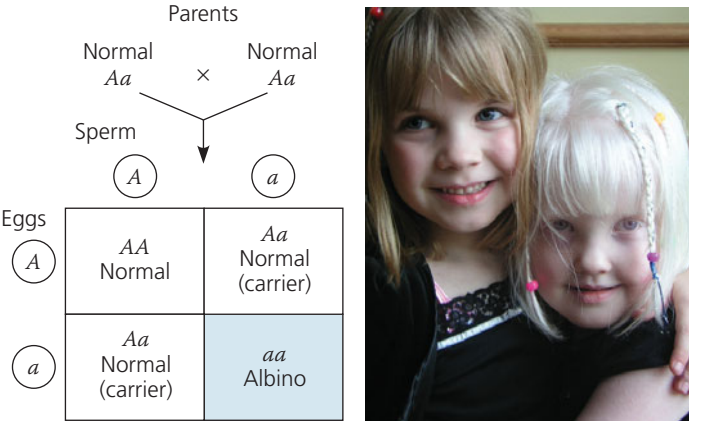
Pedigrees are a more serious matter when the alleles in question cause disabling or deadly diseases instead of innocuous human variations such as hairline or earlobe configuration. However, for disorders inherited as simple Mendelian traits, the same techniques of pedigree analysis apply.

Recessively Inherited Disorders

Thousands of genetic disorders are known to be inherited as simple recessive traits. These disorders range in severity from



▲ **Figure 11.14 Pedigree analysis.** Each of these pedigrees traces a trait through three generations of the same family. The two traits have different inheritance patterns, as seen by analysis of the pedigrees.



▲ Figure 11.15 Albinism: a recessive trait. One of the two sisters shown here has normal coloration; the other is albino. 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 with normal coloration is a carrier of the albinism allele?*

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*) are typically normal in 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 one 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 11.15** 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 11.15. 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 1/4 chance of inheriting a double dose of the recessive allele; in the case of albinism, such a child will be albino. From the genotypic ratio, we also can see that out of three offspring with the *normal* phenotype (one *AA* plus two *Aa*), two are predicted to be heterozygous carriers, a 2/3 chance. Recessive homozygotes could also result from *Aa* × *aa* and *aa* × *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, such individuals will still account for a much smaller percentage of the population than heterozygous carriers (for reasons we will examine in Chapter 21).

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. However, if the man and woman are close relatives (for example, siblings or first cousins), the probability of passing on recessive traits increases greatly. These are called consanguineous ("same blood") matings, and they are indicated in pedigrees by double lines. Because people with recent common ancestors are more likely to carry the same recessive alleles than are unrelated people, it is more likely that a mating of close relatives will 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.

There is debate among geneticists about the extent to which human consanguinity increases the risk of inherited diseases. Many deleterious alleles have such severe effects that a homozygous embryo spontaneously aborts long before birth. Still, most societies and cultures have laws or taboos forbidding marriages between close relatives. These rules may have evolved out of empirical observation that in most populations, stillbirths and birth defects are more common when parents are closely related. Social and economic factors have also influenced the development of customs and laws against consanguineous marriages.

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 extracellular chloride, which causes the mucus that coats certain cells to become thicker and stickier than normal. The mucus builds up in the pancreas, lungs, digestive tract, and other organs, leading to multiple (pleiotropic) effects, including poor absorption of nutrients from the intestines, chronic bronchitis, and recurrent bacterial infections.

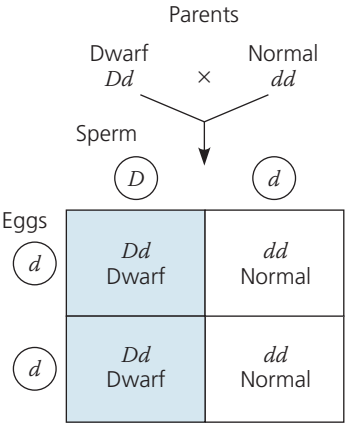
Untreated, cystic fibrosis can cause death by the age of 5. Daily doses of antibiotics to stop infection, gentle pounding on the chest to clear mucus from clogged airways, and other therapies can prolong life. In the U.S., more than half of those with cystic fibrosis now survive into their 30s and beyond. Recent research on gene-based treatments also shows much promise.

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 molecules aggregate into long rods that deform the red cells into a sickle shape (see Figure 3.22). 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 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, but there is no cure.

Although two sickle-cell alleles are necessary for an individual to manifest full-blown sickle-cell disease, 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. Heterozygotes (carriers), said to have *sickle-cell trait*, are usually healthy, but they may suffer some sickle-cell symptoms during prolonged periods of reduced blood oxygen. At the molecular level, the two alleles are codominant; both normal and abnormal (sickle-cell) hemoglobins are made in heterozygotes.

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 this allele from 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 25.26), and the presence of even heterozygous amounts of sickle-cell hemoglobin results in lower parasite densities and hence reduced malaria symptoms. 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 Chapter 21.) The relatively high frequency of African-Americans with sickle-cell trait is a vestige of their African roots.



▲ Figure 11.16 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.

Dominantly Inherited Disorders

Although many harmful alleles are recessive, a number of human 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 11.16)**. Therefore, all people who are not achondroplastic dwarfs—99.99% of the population—are homozygous for the recessive allele. 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.

Dominant alleles that cause a lethal disease are much less common than recessive alleles that have lethal effects. All lethal alleles arise by mutations (changes to the DNA) in cells that produce sperm or eggs; presumably, such mutations are equally likely to be recessive or dominant. A lethal recessive allele 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, so the allele is not passed on to future generations.

In cases of late-onset diseases, however, a lethal dominant allele may be passed on. If symptoms first appear after reproductive age, the individual may already have transmitted the allele to his or her children. For example, **Huntington's disease**, a degenerative disease of the nervous system, 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 11.16). In the United States, this devastating 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 allele for Huntington's disease to a locus near the tip of chromosome 4, and the gene was sequenced in 1993. This information led to the development of a genetic test that could detect the presence of the Huntington's allele in an individual's genome. 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.

Multifactorial Disorders

The hereditary diseases we have discussed so far are sometimes described as simple Mendelian disorders because they result from an 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 many 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 Counseling Based on Mendelian Genetics

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.

Consider the case of a hypothetical couple, John and Carol. Each had a brother who died from the same recessively inherited lethal disease. Before conceiving their first child, John and Carol 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 John and both parents of Carol must have been carriers of the recessive allele. Thus, John and Carol are both products of $Aa \times Aa$ crosses, where *a* symbolizes the allele that causes this particular disease. We also know that John and Carol 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, John and Carol each have a 2/3 chance of being carriers (*Aa*). According to the rule of multiplication, the overall probability of their firstborn having the disorder is 2/3 (the chance that John is a carrier) times 2/3 (the chance that Carol is a carrier) times 1/4 (the chance of two carriers having a child with the disease), which equals 1/9. Suppose that Carol and John decide to have a child—after all, there is an 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* John and Carol are, in fact, carriers (*Aa* genotype). If both John and Carol are carriers, there is a 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 John and Carol have three more children, and *all three* have the hypothetical hereditary disease. There is only one chance in 64 ($1/4 \times 1/4 \times 1/4$) that such an outcome will occur. Despite this run of misfortune, the chance that still another child of this couple will have the disease remains 1/4.

Genetic counseling like this relies 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 early in the 20th century, decades after he reported his findings. In the next chapter, you'll 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 11.4

1. Beth and Tom each have a sibling with cystic fibrosis, but neither Beth nor Tom 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 Tom is a carrier but Beth is not? Explain your answers.
2. **MAKE CONNECTIONS** In Table 11.1, 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 11.14b. What accounts for the difference in the two ratios?

For suggested answers, see Appendix A.

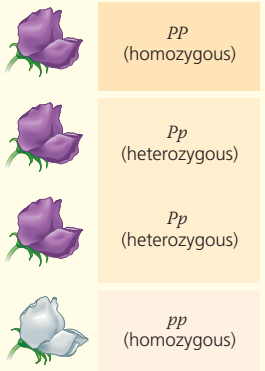
11 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 11.1

Mendel used the scientific approach to identify two laws of inheritance (pp. 207–212)

- 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** self-pollinate. Each organism inherits one allele for each gene from each parent. In **heterozygotes**, the two alleles are different, and expression of one (the **dominant allele**) masks the phenotypic effect of the other (the **recessive allele**). **Homozygotes** have identical alleles of a given gene and are **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 11.2

The laws of probability govern Mendelian inheritance (pp. 213–214)

- The **multiplication rule** states that the probability of two or more events occurring together 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 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 11.8 as two smaller monohybrid Punnett squares, one for each gene. Below each square, list the fraction of each phenotype produced. Use the rule of multiplication to compute the overall fraction of each possible dihybrid phenotype. Write the phenotypic ratio.*

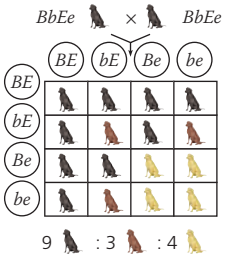
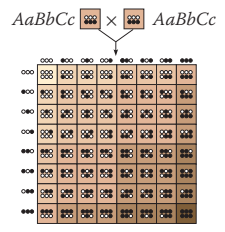
CONCEPT 11.3

Inheritance patterns are often more complex than predicted by simple Mendelian genetics (pp. 214–219)

- 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	
Incomplete dominance of either allele	Heterozygous phenotype intermediate between the two homozygous phenotypes	
Codominance	Both phenotypes expressed in heterozygotes	
Multiple alleles	In the whole population, some genes have more than two alleles	ABO blood group alleles I^A, I^B, i
Pleiotropy	One gene is able to affect 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	
Polygenic inheritance	A single phenotypic character is affected by two or more genes	

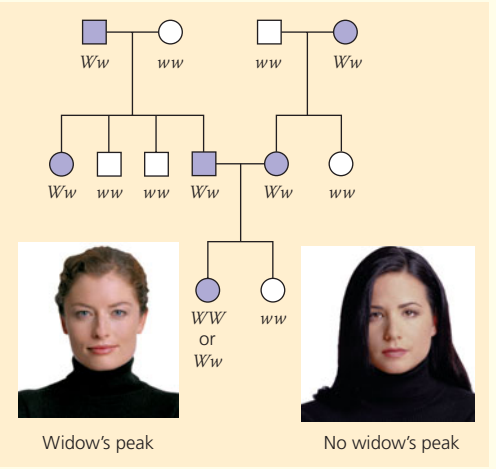
- The expression of a genotype can be affected by environmental influences. Polygenic characters that are also influenced by the environment are called **multifactorial** characters.
- An organism's overall phenotype reflects its complete genotype and unique environmental history. Even in more complex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

? *Which relationships (in the first column of the two tables above) are demonstrated by the inheritance patterns of the ABO blood group alleles? Explain why or why not, for each genetic relationship.*

CONCEPT 11.4

Many human traits follow Mendelian patterns of inheritance (pp. 219–223)

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



- Many genetic disorders are inherited as simple recessive traits, ranging from relatively mild disorders (albinism, for example) to life-threatening ones such as sickle-cell disease and cystic fibrosis. Most affected (homozygous recessive) individuals are children of phenotypically normal, heterozygous **carriers**.
- The sickle-cell allele has probably persisted for evolutionary reasons: 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 alleles that are expressed relatively late in life are 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 of their children having genetic disorders.

? *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 more uncertainty in their prediction if they could find out from genetic tests 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 allele 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. 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. Instead, see tip 5.
5. You can use the rules of probability if the Punnett square would be too big. (For example, see the question at the end of the summary for Concept 11.2 and question 7.) You can consider each gene separately (see the section Solving Complex Genetics Problems with the Rules of Probability in Concept 11.2).
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 1/2 of the offspring have the recessive phenotype and 1/2 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 11.14 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 both of the parents must be 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 individuals in the pedigree as well as you can. If an individual has the dominant phenotype, the genotype must be *AA* or *Aa*; you can write this as *A*-; the recessive phenotype means the genotype must be *aa.* Try different possibilities to see how well each fits the results. Use Mendel's laws and the rules of probability to calculate the probability of each possible genotype being the correct one.

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

1. Match each term on the left with a statement on the right.

Term	Statement
— Gene	a. Has no effect on phenotype in a heterozygote
— Allele	b. A variant for a character
— Character	c. Having two identical alleles for a gene
— Trait	d. A cross between individuals heterozygous for a single character
— Dominant allele	e. An alternative version of a gene
— Recessive allele	f. Having two different alleles for a gene
— Genotype	g. A heritable feature that varies among individuals
— Phenotype	h. An organism's appearance or observable traits
— Homozygous	i. A cross between an individual with an unknown genotype and a homozygous recessive individual
— Heterozygous	j. Determines phenotype in a heterozygote
— Testcross	k. The genetic makeup of an individual
— Monohybrid cross	l. A heritable unit that determines a character; can exist in different forms

2. **DRAW IT** Two pea plants heterozygous for the characters of pod color and pod shape are crossed. Draw a Punnett square to determine the phenotypic ratios of the offspring.

3. 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?

4. A man has six fingers on each hand and six toes on each foot. His wife and their daughter have the normal 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?

5. **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. Assume that pollen comes from the *ii* plant.

Level 2: Application/Analysis

6. 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 follows:

Character	Dominant	Recessive
Flower position	Axial (*A*)	Terminal (*a*)
Stem length	Tall (*T*)	Dwarf (*t*)
Seed shape	Round (*R*)	Wrinkled (*r*)

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 as follows? (*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, while heterozygous for seed shape

7. A black guinea pig crossed with an albino guinea pig produces 12 black offspring. When the albino is crossed with a second black one, 7 blacks and 5 albinos are obtained. What is the best explanation for this genetic outcome? Write genotypes for the parents, gametes, and offspring.

8. In some plants, a true-breeding, red-flowered strain gives all pink flowers when crossed with a white-flowered strain: C^RC^R (red) × C^WC^W (white) → C^RC^W (pink). If flower position (axial or terminal) is inherited as it is in peas (see Table 11.1), what will be the ratios of genotypes and phenotypes of the F_1 generation resulting from the following cross: axial-red (true-breeding) × terminal-white? What will be the ratios in the F_2 generation?

9. In sesame plants, the one-pod condition (*P*) is dominant to the three-pod condition (*p*), and normal leaf (*L*) is dominant to wrinkled leaf (*l*). Pod type and leaf type are inherited independently. Determine the genotypes for the two parents for all possible matings producing the following offspring:

(a) 318 one-pod, normal leaf and 98 one-pod, wrinkled leaf
(b) 323 three-pod, normal leaf and 106 three-pod, wrinkled leaf
(c) 401 one-pod, normal leaf
(d) 150 one-pod, normal leaf, 147 one-pod, wrinkled leaf, 51 three-pod, normal leaf, and 48 three-pod, wrinkled leaf
(e) 223 one-pod, normal leaf, 72 one-pod, wrinkled leaf, 76 three-pod, normal leaf, and 27 three-pod, wrinkled leaf

10. Phenylketonuria (PKU) 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.)

11. 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*

12. 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*

13. Karen and Steve each have a sibling with sickle-cell disease. Neither Karen nor Steve nor any of their parents have the disease, and none of them have been tested to see if they have the sickle-cell trait. Based on this incomplete information, calculate the probability that if this couple has a child, the child will have sickle-cell disease.

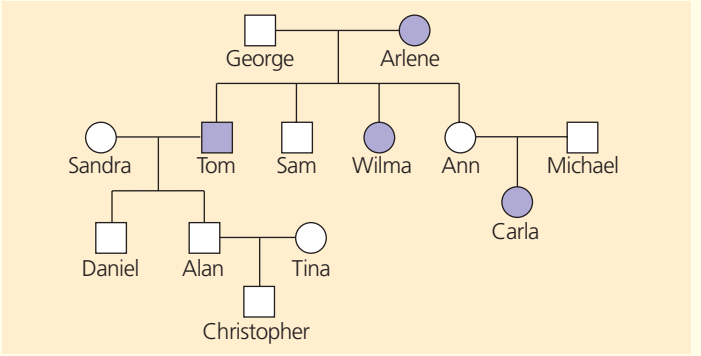
14. In tigers, a recessive allele that is pleiotropic causes an absence of fur pigmentation (a white tiger) and a cross-eyed condition. If two phenotypically normal tigers that are heterozygous at this locus are mated, what percentage of their offspring will be cross-eyed? What percentage of cross-eyed tigers will be white?

15. In 1981, a stray black cat with unusual rounded, curled-back ears was adopted by a family in California. Hundreds of descendants of the cat have since been born, and cat fanciers hope to develop 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?

16. Imagine that a newly discovered, recessively inherited disease is expressed only in individuals with type O blood, although the disease and blood group are independently inherited. A normal man with type A blood and a normal woman with type B blood have already had one child with the disease. The woman is now pregnant for a second time. What is the probability that the second child will also have the disease? Assume that both parents are heterozygous for the gene that causes the disease.

17. 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 genotypic and phenotypic ratios of the offspring?

18. 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?



19. 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.)

20. In mice, black fur (*B*) is dominant to white (*b*). At a different locus, a dominant allele (*A*) produces a band of yellow just below the tip of each hair in mice with black fur. This gives a frosted appearance known as agouti. Expression of the recessive allele (*a*) results in a solid coat color. If mice that are heterozygous at both loci are crossed, what are the expected genotypic and phenotypic ratios of their offspring?

Level 3: Synthesis/Evaluation

21. 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) What are *all* the possible genotypes for your mystery plant?
(b) Describe the *one* cross you would do, out 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. How do you do this? Why is this 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.

22. SCIENCE, TECHNOLOGY, AND SOCIETY

Imagine that one of your parents has Huntington's disease. What is the probability that you, too, will someday manifest the disease? There is no cure for Huntington's. Would you want to be tested for the Huntington's allele? Why or why not?

23. FOCUS ON EVOLUTION

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?

24. FOCUS ON 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.

For selected answers, see Appendix A.

12 The Chromosomal Basis of Inheritance

KEY CONCEPTS

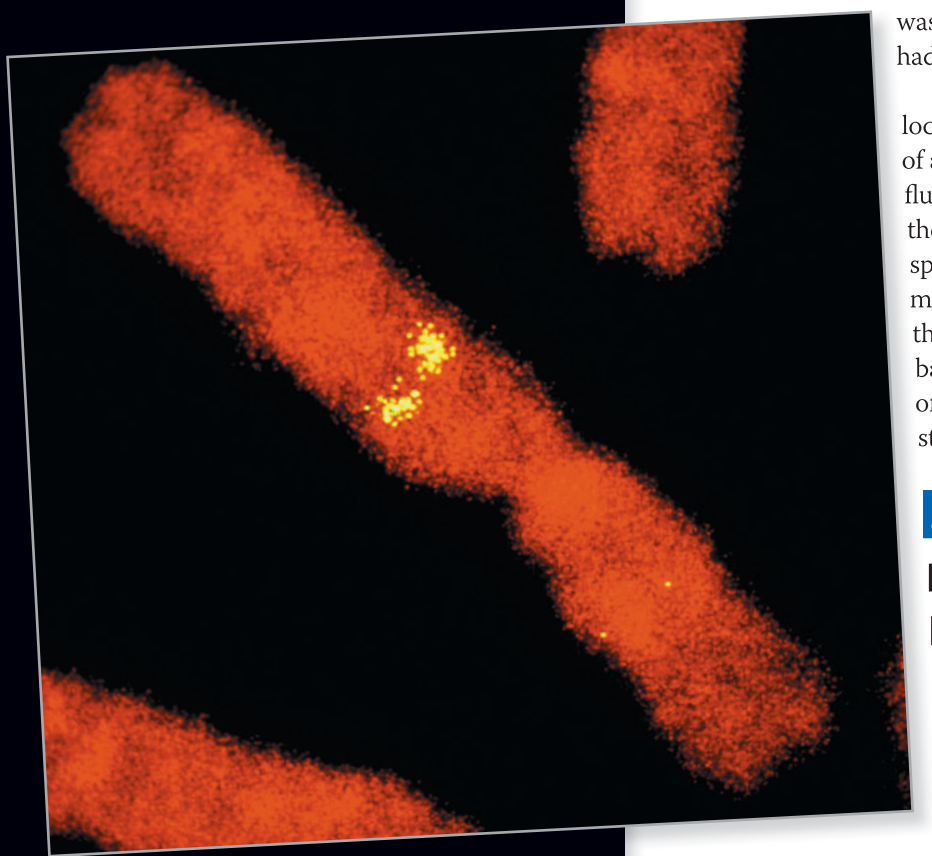
12.1 Mendelian inheritance has its physical basis in the behavior of chromosomes

12.2 Sex-linked genes exhibit unique patterns of inheritance

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

12.4 Alterations of chromosome number or structure cause some genetic disorders

▼ **Figure 12.1** Where are Mendel's hereditary factors located in the cell?



OVERVIEW

Locating Genes Along Chromosomes

Gregor Mendel's "hereditary factors" were purely an abstract concept when he proposed their existence in 1860. At that time, no cellular structures were known that could house these imaginary units. Even after chromosomes were first observed, many biologists remained skeptical about Mendel's laws of segregation and independent assortment until there was sufficient evidence that these principles of heredity had a physical basis in chromosomal behavior.

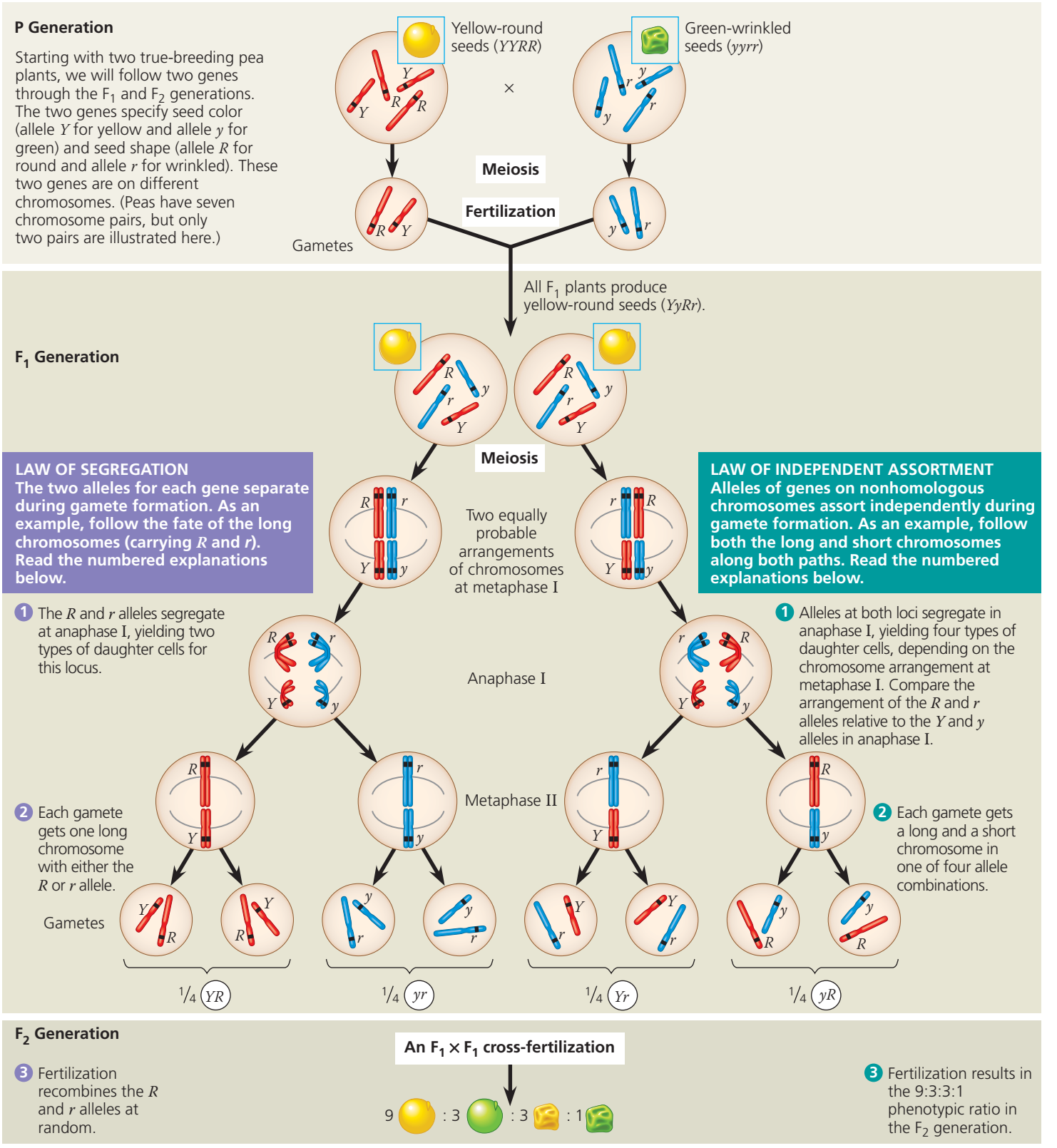
Today, we know that genes—Mendel's "factors"—are located along chromosomes. We can see the location of a particular gene by tagging chromosomes with a fluorescent dye that highlights that gene. For example, the two yellow spots in **Figure 12.1** mark the locus of a specific gene on the sister chromatids of human chromosome 6. This chapter will extend what you learned in the past two chapters. We'll describe the chromosomal basis for the transmission of genes from parents to offspring, along with some important exceptions to the standard mode of inheritance.

CONCEPT 12.1

Mendelian inheritance has its physical basis in the behavior of chromosomes

Using improved techniques of microscopy, cytologists worked out the process of mitosis in 1875 and meiosis in the 1890s. Cytology and genetics converged when biologists began to see parallels between the behavior of chromosomes and the behavior of Mendel's proposed hereditary factors during sexual life cycles: Chromosomes and genes are both present in pairs in diploid cells; homologous chromosomes separate and alleles segregate during the process of meiosis; and fertilization restores the paired condition for both chromosomes and genes. Around 1902, Walter S. Sutton, Theodor Boveri, and others independently noted these parallels, and the **chromosome theory of inheritance** began to take form. According to this theory, Mendelian genes have specific loci (positions) along chromosomes, and it is the chromosomes that undergo segregation and independent assortment.

Figure 12.2 shows that the behavior of homologous chromosomes during meiosis can account for the segregation of the alleles at each genetic locus to



▲ Figure 12.2 The chromosomal basis of Mendel's laws. Here we correlate a dihybrid cross that Mendel performed (see Figure 11.8) with the behavior of chromosomes during meiosis (see Figure 10.8). The arrangement of chromosomes at metaphase I of meiosis and their movement during anaphase I account, respectively, for 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.

? *If you crossed an F_1 plant 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?*

different gametes. The figure also shows that the behavior of nonhomologous chromosomes can account for the independent assortment of the alleles for two or more genes located on different chromosomes. By carefully studying this figure, which traces the same dihybrid pea cross you learned about in Figure 11.8, 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.

Morgan's Experimental Evidence: *Scientific Inquiry*

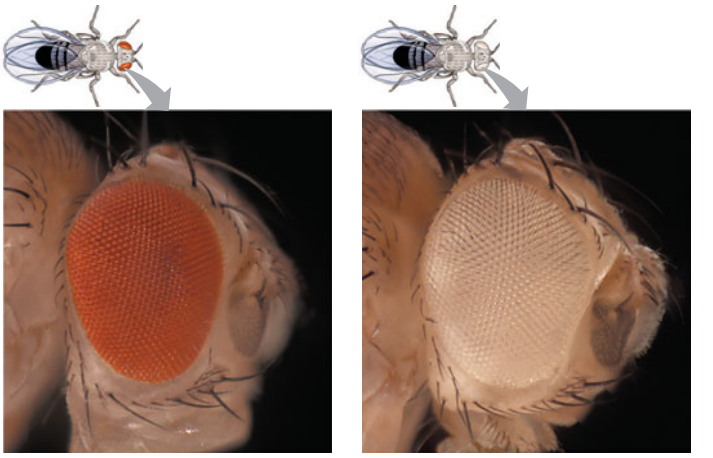
The first solid evidence associating a specific gene with a specific chromosome came early in the 20th century from the work of Thomas Hunt Morgan, an experimental embryologist at Columbia University. Although Morgan was initially skeptical about both Mendelian genetics and the chromosome theory, his 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 lamented, "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 12.3). 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 12.3 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 (LMs).

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*. The 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 *HD* for the allele for Huntington's disease.

Correlating Behavior of a Gene's Alleles with Behavior of a Chromosome Pair

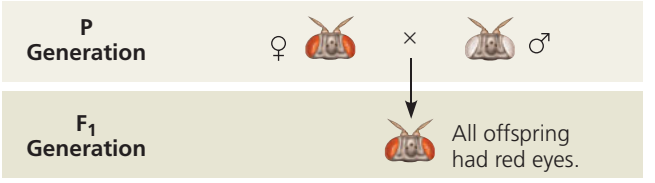
Morgan mated his white-eyed male fly with a red-eyed female. All the F_1 offspring had red eyes, suggesting that the wild-type allele is dominant. When Morgan bred the F_1 flies to each other, he observed the classical 3:1 phenotypic ratio among the F_2 offspring. However, there was a surprising additional 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 male and half 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 12.4**. For a male, a single

▼ Figure 12.4 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?

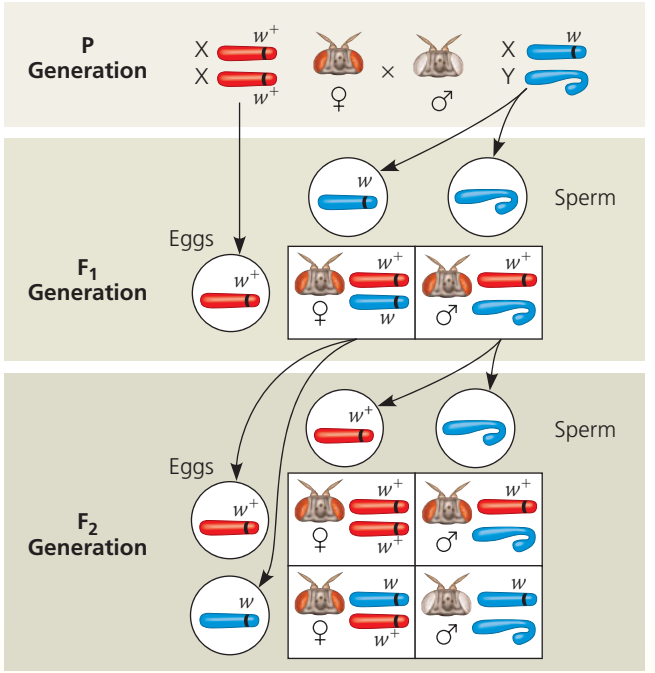
Experiment Thomas Hunt Morgan wanted to analyze the behavior of two alleles of a fruit fly eye-color gene. In crosses similar to those done by Mendel with pea plants, Morgan and his colleagues mated a wild-type (red-eyed) female with a mutant white-eyed male.



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 three red-eyed flies to one white-eyed fly. However, no females displayed the white-eye trait; all white-eyed flies were males.

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.



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

A related Experimental Inquiry Tutorial can be assigned in MasteringBiology.

WHAT IF? Suppose this eye-color gene were located on an autosome. Predict the phenotypes (including gender) of the F_2 flies in this hypothetical cross. (*Hint*: Draw a Punnett square.)

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. On the other hand, 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.

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

CONCEPT CHECK 12.1

1. Which one of Mendel's laws relates to the inheritance of alleles for a single character? Which law relates to the inheritance of alleles for two characters in a dihybrid cross?
2. **MAKE CONNECTIONS** Review the description of meiosis (see Figure 10.8) and Mendel's laws of segregation and independent assortment (see Concept 11.1). What is the physical basis for each of Mendel's laws?
3. **WHAT IF?** Propose a possible reason that the first naturally occurring mutant fruit fly Morgan saw involved a gene on a sex chromosome.

For suggested answers, see Appendix A.

CONCEPT 12.2

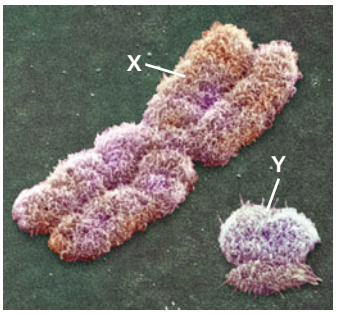
Sex-linked genes exhibit unique patterns of inheritance

As you just learned, Morgan's discovery of a trait (white eyes) that correlated with a fly's sex was a key episode in the development of the chromosome theory of inheritance. Because the identity of a fly's sex chromosomes 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. We'll begin by reviewing the chromosomal basis of sex determination in humans and some other animals.

The Chromosomal Basis of Sex

Whether we are male or female is one of our more obvious phenotypic characters. Although the anatomical and physiological differences between women and men are numerous, the chromosomal basis for determining sex is rather simple. In humans and other mammals, there are two varieties of sex

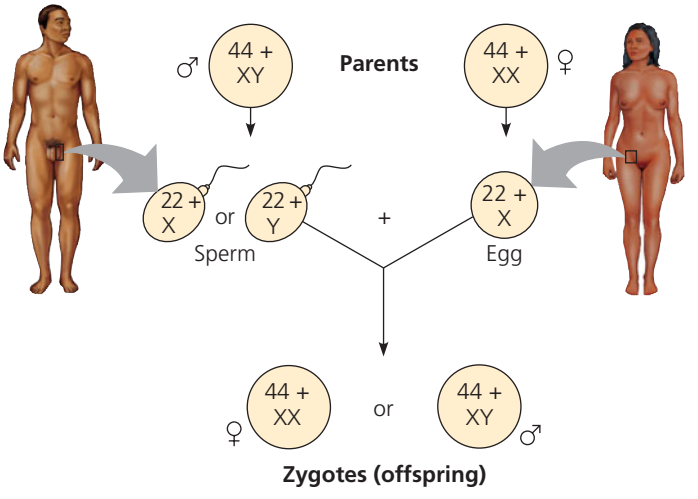
chromosomes, designated X and Y. The Y chromosome is much smaller than the X chromosome **(Figure 12.5)**. A person who inherits two X chromosomes, one from each parent, usually develops as a female. A male develops from a zygote containing one X chromosome and one Y chromosome **(Figure 12.6)**. Short segments at either end of the Y chromosome are the only regions that are homologous with corresponding regions of the X. These homologous regions allow the X and Y chromosomes in males to pair and behave like homologous chromosomes during meiosis in the testes.



▲ Figure 12.5 Human sex chromosomes.

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 happens to fertilize 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 12.6). Thus, mammalian sex determination is a matter of chance—a fifty-fifty chance. In *Drosophila*, males are XY, but sex depends on the ratio between the number of X chromosomes and the number of autosome sets, not simply on the presence of the Y. There are other chromosomal systems as well, besides the X-Y system, for determining sex.

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. In 1990, a British research team identified a



▲ Figure 12.6 The mammalian X-Y chromosomal system of sex determination. In mammals, the sex of an offspring depends on whether the sperm cell contains an X chromosome or a Y. Numerals indicate the number of autosomes.

gene on the Y chromosome required for the development of testes. They named the gene *SRY*, for sex-determining region of Y. In the absence of *SRY*, the gonads develop into ovaries. The biochemical, physiological, and anatomical features that distinguish males and females are complex, and many genes are involved in their development. In fact, *SRY* codes for a protein that regulates other genes.

Researchers have sequenced the human Y chromosome and 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. A gene located on either sex chromosome is called a **sex-linked gene**; those located on the Y chromosome are called *Y-linked genes*. 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. A rare example is that in the absence of certain Y-linked genes, an XY individual is male but does not produce normal sperm.

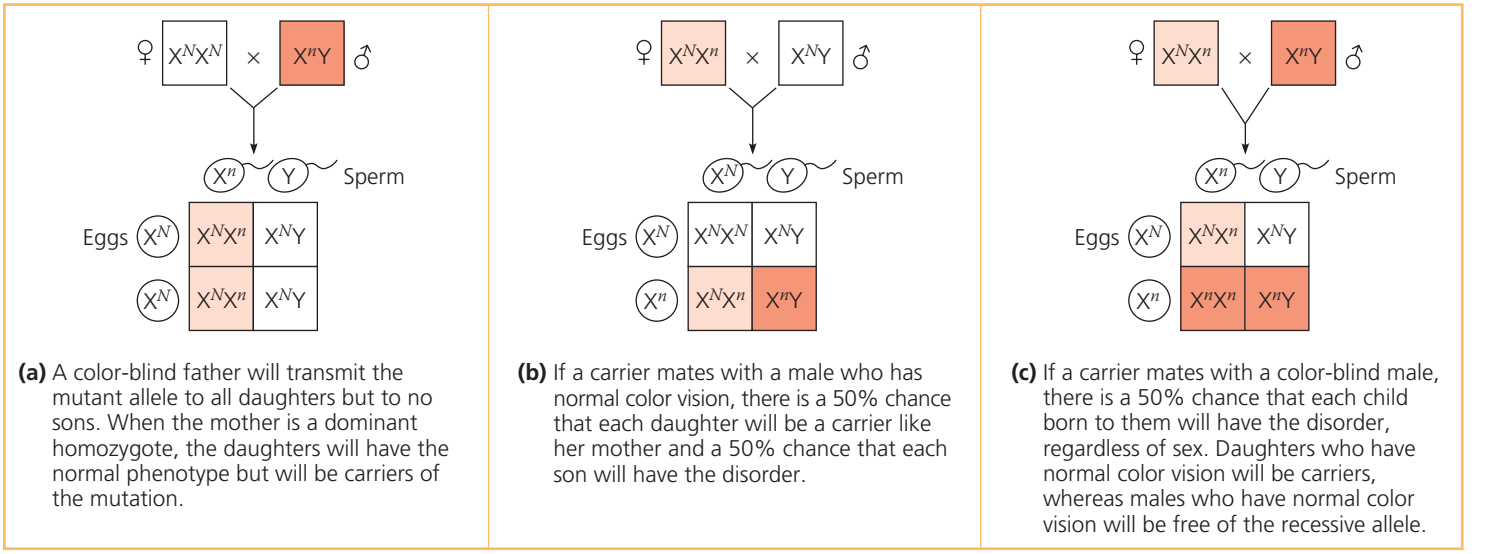
The human X chromosome contains approximately 1,100 genes, which are called **X-linked genes**. The fact that males and females inherit a different number of X chromosomes leads to a pattern of inheritance different from that produced by genes located on autosomes.

Inheritance of X-Linked Genes

While most Y-linked genes help determine sex, the X chromosomes have genes for many 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 12.4). 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 12.7** for the inheritance of a mild X-linked disorder, 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* are meaningless when 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 12.7c). Because the X-linked allele for color blindness is relatively rare, however, 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. An example is **Duchenne muscular dystrophy**, which affects about one out of every 3,500 males



(a) A color-blind father 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.

▲ Figure 12.7 The transmission of X-linked recessive traits. In this diagram, color blindness is used as an example. The superscript *N* represents the dominant allele for normal color vision carried on the X chromosome, and the superscript *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?*

born in the United States. The disease is characterized by a progressive weakening of the muscles and loss of coordination. Affected individuals rarely live past their early 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.

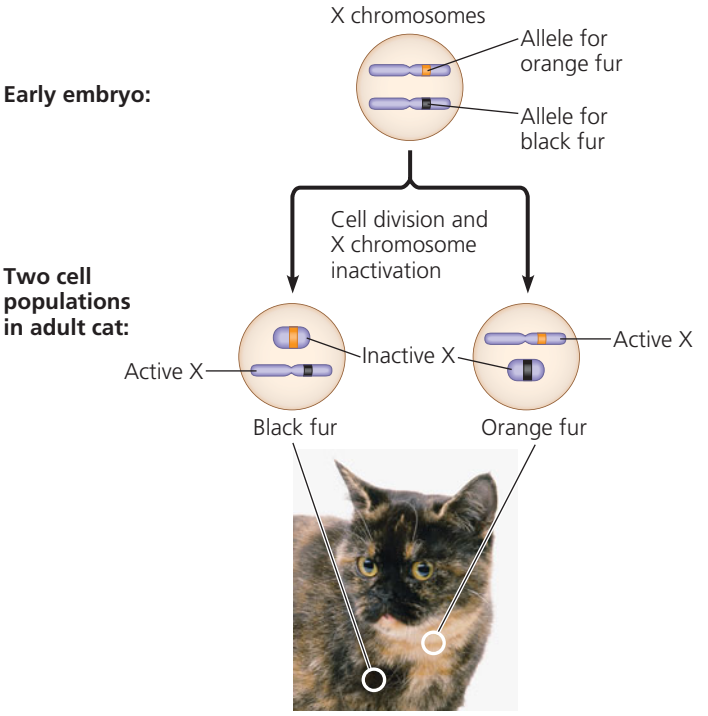
Hemophilia is an X-linked recessive disorder defined by the absence of one or more of the proteins required for blood clotting. When a person with hemophilia is injured, bleeding is prolonged because a firm clot is slow to form. Small cuts in the skin are usually not a problem, but bleeding in the muscles 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. 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 humans, inherit two X chromosomes—twice the number inherited by males—so you may wonder if females make twice as much of the proteins encoded by X-linked genes. In fact, most 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 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, Barr-body chromosomes are reactivated in the cells that give rise to eggs, so every female gamete has an active X.

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 her cells will express one allele, while the others will express the alternate allele. **Figure 12.8** shows how this mosaicism results in the mottled 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 the histone proteins bound to it, including attachment of methyl groups (—CH_3) to one of the nitrogenous bases of DNA nucleotides. (The regulatory role of DNA methylation is discussed further in Chapter 15.) 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



▲ Figure 12.8 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 yet another gene.)

genes, called *XIST* (for X-inactive specific transcript), becomes active *only* on the chromosome that will become the Barr body. 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 seems to initiate X inactivation, and the RNA products of other genes nearby on the X chromosome help to regulate the process.

CONCEPT CHECK 12.2

1. A white-eyed *Drosophila* female is mated with a red-eyed (wild-type) male, the reciprocal of the cross shown in Figure 12.4. What phenotypes and genotypes do you predict for the offspring?
2. Neither Tim nor Rhoda has Duchenne muscular dystrophy, but their firstborn son does have it. What is the probability that a second child of this couple 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 11.1. 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 12.3

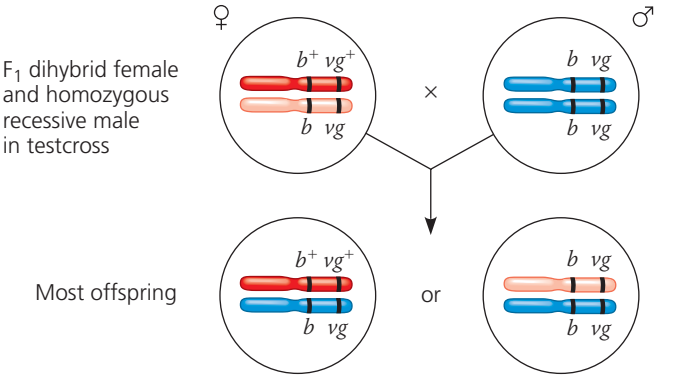
Linked genes tend to be inherited together because they are located near each other on the same chromosome

The number of genes in a cell is far greater than the number of chromosomes; in fact, each chromosome has hundreds or thousands of genes. (The small Y chromosome is an exception.) 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**. (Note the distinction between the terms *sex-linked gene*, referring to a single gene on a sex chromosome, and *linked genes*, referring to two or more genes on the same chromosome that tend to be inherited together.) 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 12.9**. The first was a P generation cross to generate F_1 dihybrid flies, and the second was a testcross.

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 for these characters are near each other on the same chromosome:

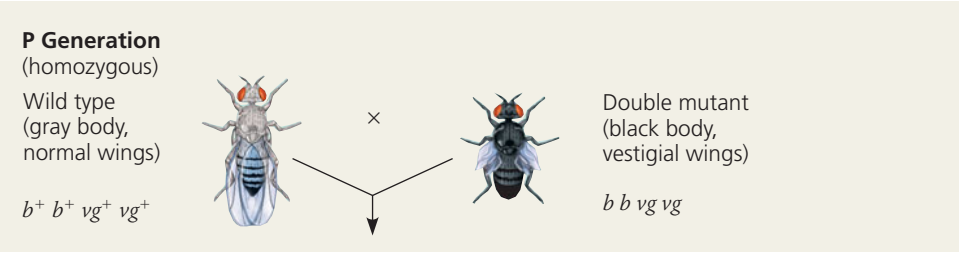


▼ Figure 12.9 Inquiry

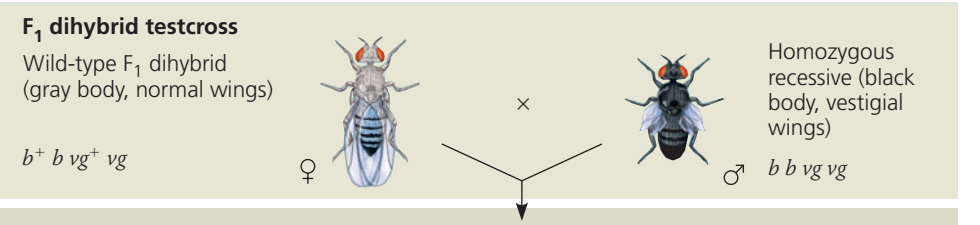
How does linkage between two genes affect inheritance of characters?

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.

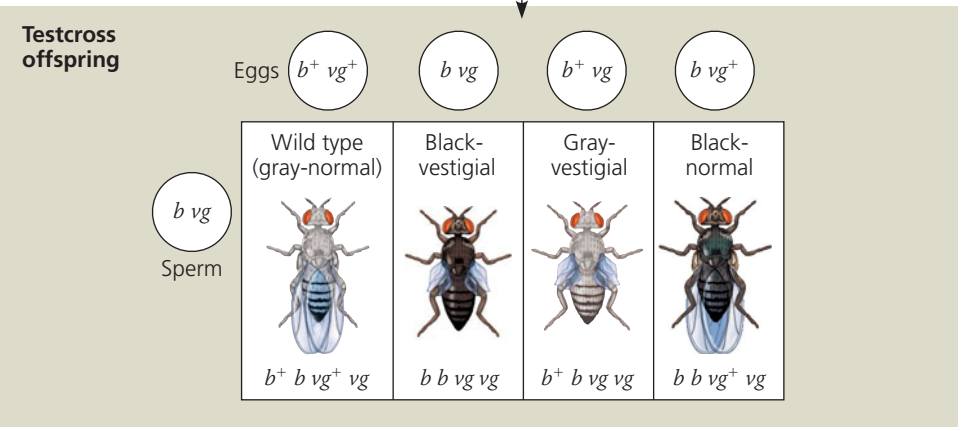


He then mated wild-type F_1 dihybrid females with homozygous recessive males. This testcross will reveal the genotype of the eggs made by the dihybrid female.



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

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	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.

Source 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 true-breeding for black body with normal wings, which phenotypic class(es) would be largest among the testcross offspring?*

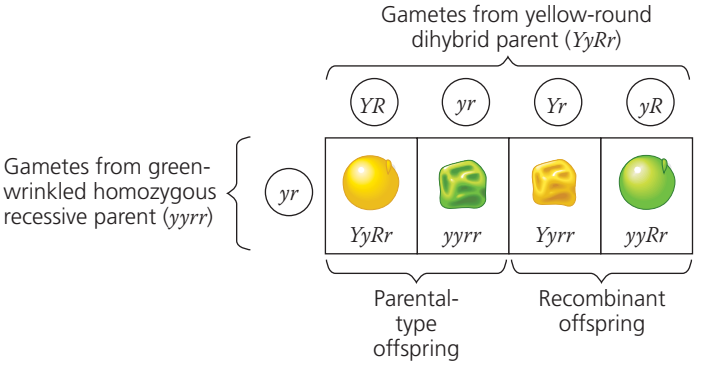
However, as Figure 12.9 shows, both of the combinations of traits not seen in the P generation (called nonparental phenotypes) were also produced in Morgan's experiments, suggesting that the body-color and wing-size alleles are not always linked genetically. To understand this conclusion, we need to 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 and crossing over in meiosis I, and the possibility of any sperm fertilizing any egg (see Chapter 10). Here we'll examine the chromosomal basis of recombination 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, we can represent the cross between a pea plant with yellow-round seeds that is heterozygous for both seed color and seed shape (a dihybrid, *YyRr*) and a plant with green-wrinkled seeds (homozygous for both recessive alleles, *yyrr*) 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 parental (P generation) phenotypes. These offspring are called **parental types**. 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 *YyRr* × *yyrr* crosses (a type of testcross because it reveals the genotype of the gametes made by the dihybrid *YyRr* plant).

A 50% frequency of recombination in such testcrosses is observed for any two genes that are located on different chromosomes and thus unlinked. 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 10.10 and the question in the Figure 12.2 legend).

Recombination of Linked Genes: Crossing Over

Now let's return to Morgan's fly room to see how we can explain the results of the *Drosophila* testcross illustrated in Figure 12.9. Recall that most 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.

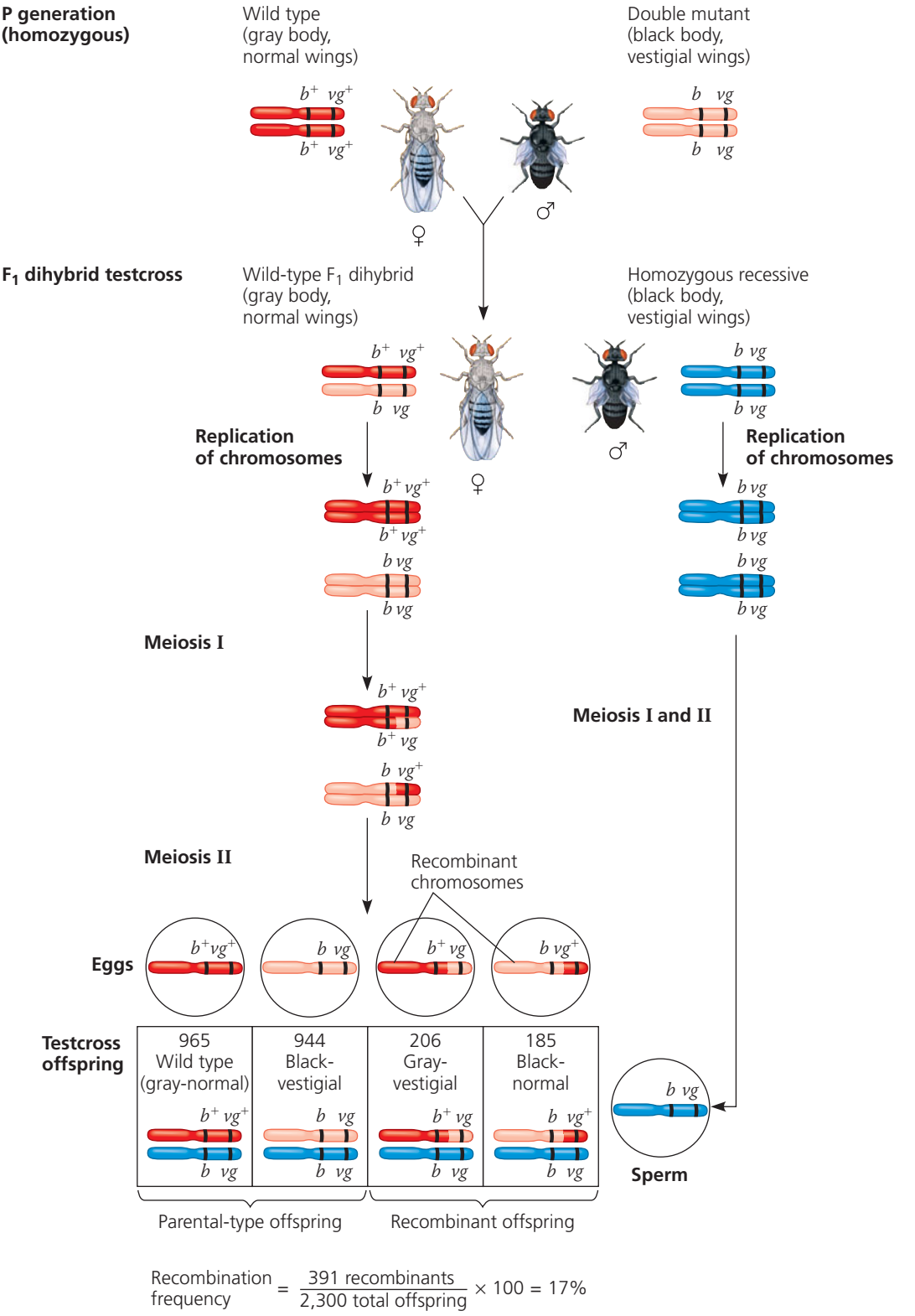
Faced with these results, Morgan proposed that some process must occasionally break the physical connection between specific genes on the same chromosome. Subsequent experiments demonstrated that this process, now called **crossing over**, accounts for the recombination of linked genes. In crossing over, which occurs while replicated homologous chromosomes are paired during prophase of meiosis I, a set of proteins orchestrates an exchange of corresponding segments of one maternal and one paternal chromatid (see Figure 10.11). In effect, end portions of two nonsister chromatids trade places each time a crossover occurs.

Figure 12.10 shows how crossing over in a dihybrid female fly resulted in recombinant eggs and ultimately recombinant offspring in Morgan's testcross. Most of the eggs had a chromosome with either the b^+ vg^+ or *b vg* parental genotype for body color and wing size, but some eggs had a recombinant chromosome (b^+ *vg* or *b* vg^+). Fertilization of these various 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 another F_1 dihybrid testcross to see whether the two genes are assorting independently or are linked.

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 Chapter 10). 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 Chapter 11). 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 12.9 and 12.10. 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 genotype will determine which genetic combinations persist over time.



▶ Figure 12.10 Chromosomal basis for recombination of linked genes. In these diagrams re-creating the testcross in Figure 12.9, we track chromosomes as well as genes. The maternal chromosomes are color-coded red and pink to distinguish one homolog from the other before any meiotic crossing over has taken place. Because crossing over between the b^+/b and vg^+/vg loci occurs in some, but not all, 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 12.9, 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."*

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

The discovery of linked genes and recombination due to crossing over led one of Morgan's students, Alfred H. Sturtevant, to 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 that in Figures 12.9 and 12.10, 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.

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? In this exercise, you will use a simple statistical test, the chi-square (χ^2) test, to analyze phenotypes of F_1 testcross progeny to see whether two genes are linked or unlinked.

How These Experiments Are Done If genes are unlinked and therefore assort independently, the phenotypic ratio of offspring from an F_1 testcross is expected to be 1:1:1:1 (see Figure 12.9). If the two genes are linked, however, the observed phenotypic ratio of the offspring will not match the expected ratio. Given random fluctuations in the data, 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 called a chi-square (χ^2) test. This test compares an observed data set to 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 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 test crossed (*AaBb* × *aabb*). 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

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 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 the 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 o = observed and e = expected. Calculate the χ^2 value for the data using the table below. Enter the data into the table, and carry 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*)**	**$(o-e)^2$**	**$(o-e)^2/e$**
Purple stem/ short petals (*A–B–*)		220			
Green stem/ short petals (*aaB–*)		210			
Purple stem/ long petals (*A–bb*)		231			
Green stem/ long petals (*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 the observed data is not consistent with the hypothesis, and thus the hypothesis should be rejected. A standard cut-off point biologists use 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 is consistent with the hypothesis. To find the probability, locate your χ^2 value in the χ^2 Distribution Table in Appendix F. 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?

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

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 to their locations on the chromosomes.

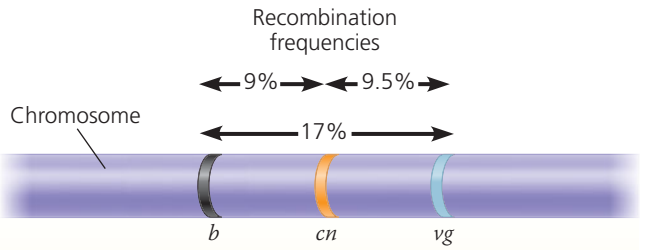
▼ Figure 12.11 Research Method

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 12.9 and 12.10. 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*–*cn* 9%, *cn*–*vg* 9.5%, and *b*–*vg* 17%) best fit a linear order in which *cn* is positioned about halfway between the other two genes:

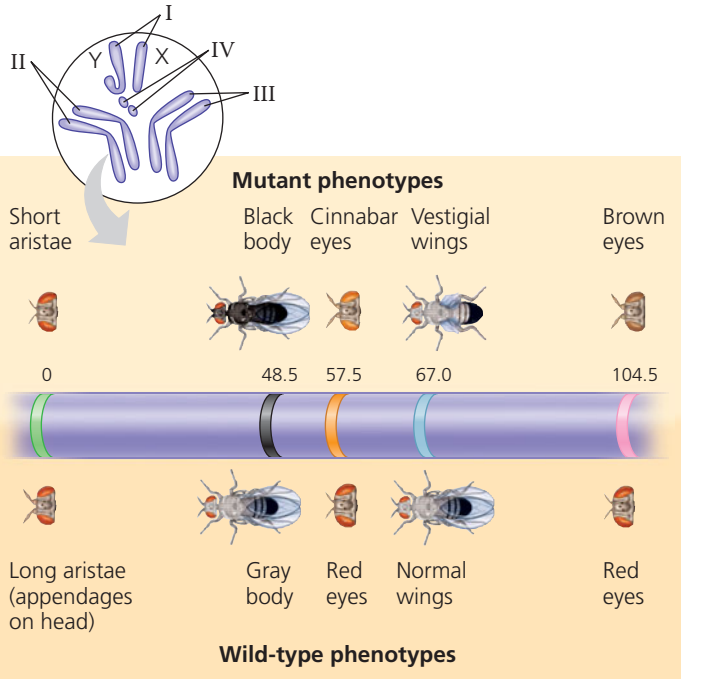


The *b*–*vg* recombination frequency (17%) is slightly less than the sum of the *b*–*cn* and *cn*–*vg* frequencies (9 + 9.5 = 18.5%) because of the few times that one crossover occurs between *b* and *cn* and another crossover occurs between *cn* and *vg*. The second crossover would "cancel out" the 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, so a geneticist would add the smaller distances in constructing a map.

A genetic map based on recombination frequencies is called a **linkage map**. **Figure 12.11** shows Sturtevant's linkage map of three genes: the body-color (*b*) and wing-size (*vg*) genes depicted in Figure 12.10 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 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 12.12)**.



▲ Figure 12.12 A partial genetic (linkage) map of a *Drosophila* chromosome. This simplified map shows just a few of the many genes that have been mapped on *Drosophila* chromosome II. The number at each gene locus indicates the number of map units between that locus and the locus for arista length (left). Notice that more than one gene can affect a given phenotypic characteristic, such as eye color. Also, note that in contrast to the homologous autosomes (II–IV), the X and Y sex chromosomes (I) have distinct shapes.

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. The ultimate maps display the physical distances between gene loci in DNA nucleotides (see Chapter 18). 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.

CONCEPT CHECK 12.3

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. For each type of offspring of the testcross in Figure 12.9, 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 between *A* and *C* is 12%. Can you determine the linear order of these genes? Explain.

For suggested answers, see Appendix A.

CONCEPT 12.4

Alterations of chromosome number or structure cause some genetic disorders

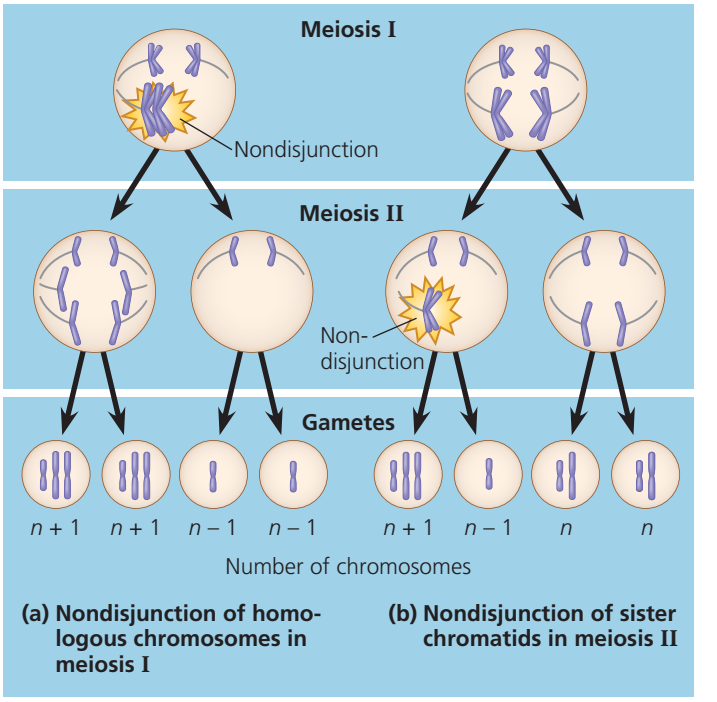
As you have learned so far in this chapter, the phenotype of an organism can be affected by small-scale changes involving individual genes. Random mutations are the source of all new alleles, which can lead to new phenotypic traits.

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 may tolerate such genetic defects better than animals do.

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 12.13)**. In these cases, 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**. (Aneuploidy may involve more than one chromosome.) 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. 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 shortly. Nondisjunction can also



▲ **Figure 12.13 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 10.6.)

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* (3*n*) and *tetraploidy* (4*n*) indicate three or 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 2*n* zygote to divide after replicating its chromosomes. Subsequent normal mitotic divisions would then produce a 4*n* embryo.

Polyploidy is fairly common in plants; the spontaneous origin of polyploid individuals plays an important role in the evolution of plants (see Chapter 22). Many of the plant species we eat are polyploid; for example, bananas are triploid, wheat is hexaploid (6*n*), and strawberries are octoploid (8*n*).

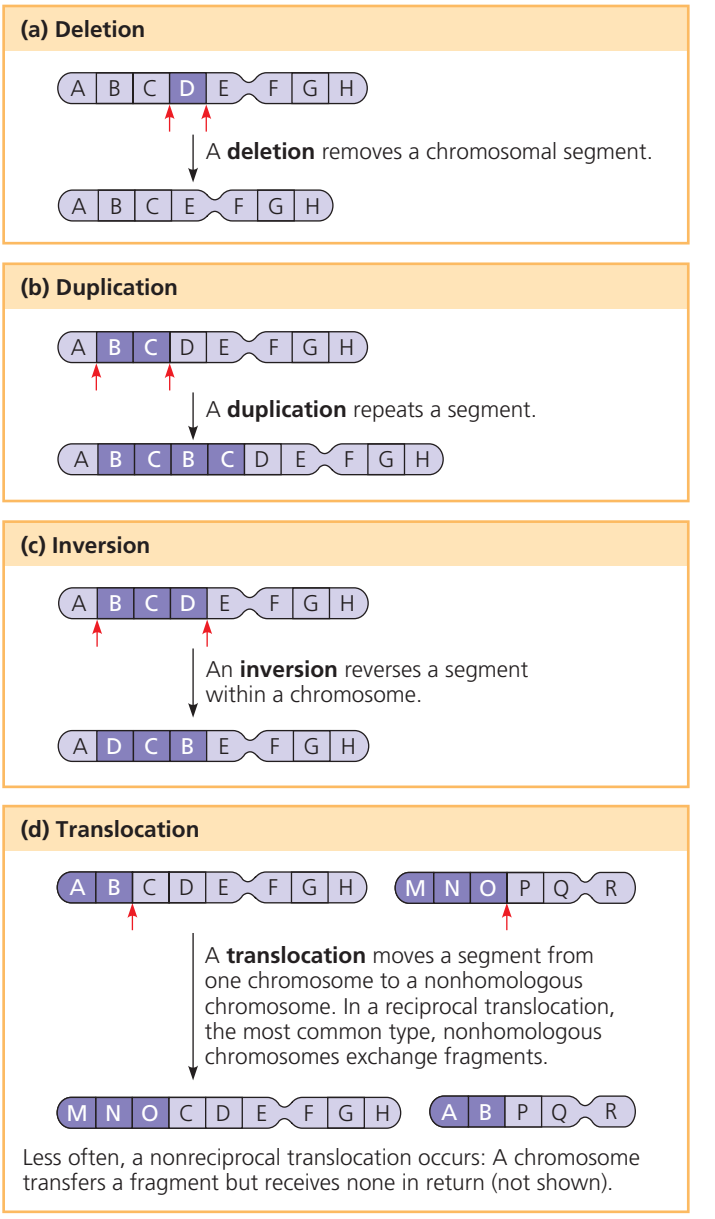
Alterations of Chromosome Structure

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 12.14)**. A **deletion** occurs when a chromosomal fragment is lost. The affected chromosome is then missing certain genes. The "deleted" fragment may become attached as an extra segment to a sister chromatid, producing a **duplication**. Alternatively, a detached fragment could attach to a nonsister chromatid of a homologous chromosome. In that case, though, the "duplicated" segments might not be identical because the homologs could carry different alleles of certain genes. 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. 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 (or has a single X chromosome with a large deletion, in a male) is usually missing a number of essential genes, a condition that is ordinarily lethal. Duplications and translocations also tend to be harmful. In reciprocal translocations, in which segments are exchanged between nonhomologous chromosomes, and in 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; such events sometimes have devastating effects.

▼ Figure 12.14 Alterations of chromosome structure. Red arrows indicate breakage points. Dark purple highlights the chromosomal parts affected by the rearrangements.



Human Disorders Due to Chromosomal Alterations

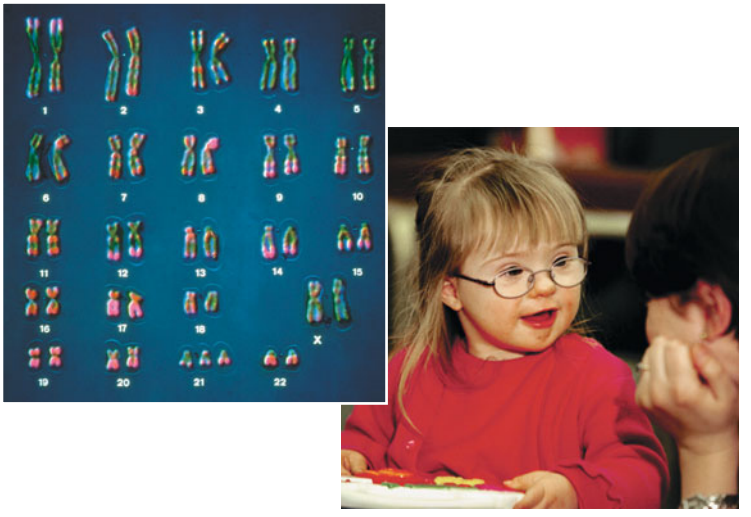
Alterations of chromosome number and structure are associated with a number of serious human disorders. As described earlier, nondisjunction in meiosis results in aneuploidy in gametes and subsequently in any resulting zygotes. Although the frequency of aneuploid zygotes may be quite high in humans, 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, with the result that 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 genetic testing of the fetus.

Down Syndrome (Trisomy 21)

One aneuploid condition, **Down syndrome**, affects approximately one out of every 700 children born in the United States **(Figure 12.15)**. 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. Although people with Down syndrome, on average, have a life span shorter than normal, most, with proper medical treatment, live to middle age and beyond. 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 disorder occurs in just 0.04% of children born to women under age 30, the risk climbs to 0.92% 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. Due to its low risk and its potential for providing useful information, prenatal screening for trisomies in the embryo is now offered to all pregnant women. Passed in 2008, the Prenatally and Postnatally Diagnosed Conditions Awareness Act stipulates 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.



▲ **Figure 12.15 Down syndrome.** The karyotype shows trisomy 21, the most common cause of Down syndrome. The child exhibits the facial features characteristic of this disorder.

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 in somatic cells.

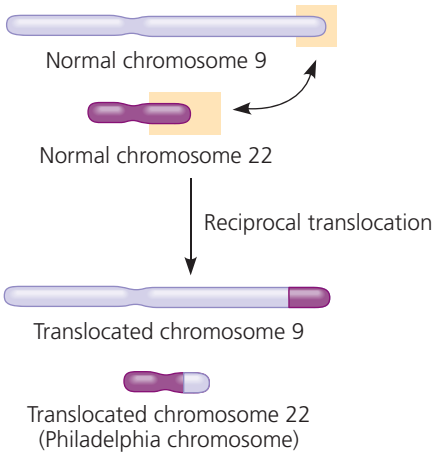
An extra X chromosome in a male, producing XXY, occurs approximately once in every 500 to 1,000 live male births. People with this disorder, called *Klinefelter syndrome*, have male sex organs, but the testes are abnormally small and the man is sterile. Even though the extra X is inactivated, some breast enlargement and other female body characteristics are common. Affected individuals may have subnormal intelligence. About one of every 1,000 males is born with an extra Y chromosome (XYY). These males undergo normal sexual development and do not exhibit any well-defined syndrome.

Females with trisomy X (XXX), which occurs once in approximately 1,000 live female births, are healthy and have no unusual physical features other than being slightly taller than average. Triple-X females are at risk for learning disabilities but are fertile. Monosomy X, 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 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 normal 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 is severely intellectually disabled, has a small head with unusual facial features, and has a cry that sounds like the mewing of a distressed cat. Such individuals usually die in infancy or early childhood.

Chromosomal translocations have been implicated in certain cancers, including *chronic myelogenous leukemia* (*CML*). This disease occurs when a reciprocal translocation happens during mitosis of cells that will become white blood cells. In these 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 12.16)**. Such an exchange causes cancer by activating a gene that leads to uncontrolled cell cycle progression. (The mechanism of gene activation will be discussed in Chapter 16.)



▲ Figure 12.16 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 12.4

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. **WHAT IF?** 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.
3. **MAKE CONNECTIONS** The gene that is activated on the Philadelphia chromosome codes for an intracellular kinase. Review the discussion of cell cycle control and cancer in Concept 9.3, and explain how the activation of this gene could contribute to the development of cancer.
4. Women born with an extra X chromosome (XXX) are generally healthy and indistinguishable in appearance from normal XX women. What is a likely explanation for this finding? How could you test this explanation?

For suggested answers, see Appendix A.

12 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 12.1

Mendelian inheritance has its physical basis in the behavior of chromosomes (pp. 228–231)

- The **chromosome theory of inheritance** states that genes are located on chromosomes and that the behavior of chromosomes during meiosis accounts for Mendel's laws of segregation and independent assortment.
- Morgan's discovery that transmission of the X chromosome in *Drosophila* correlates with inheritance of an eye-color trait was the first solid evidence indicating that a specific gene is associated with a specific chromosome.

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

CONCEPT 12.2

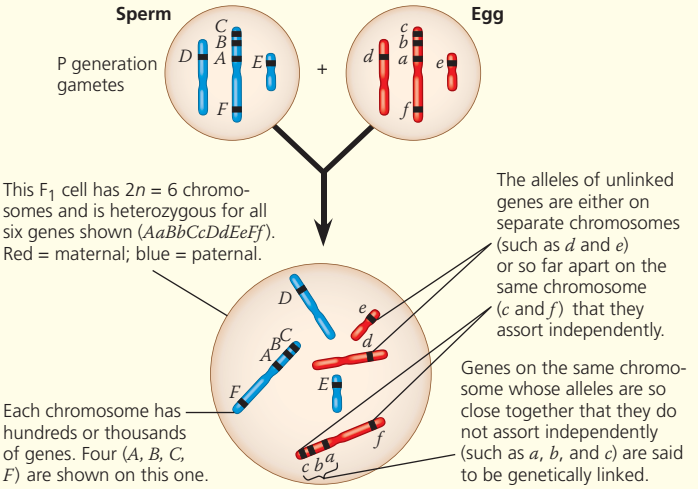
Sex-linked genes exhibit unique patterns of inheritance (pp. 231–234)

- Sex is often chromosomally based. Humans and other mammals have an X-Y system in which sex is determined by whether a Y chromosome is present.
- 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 much more often than females by X-linked disorders?*

CONCEPT 12.3

Linked genes tend to be inherited together because they are located near each other on the same chromosome (pp. 234–240)



- An F_1 testcross yields **parental types** with the same combination of traits as those in the P generation parents and **recombinant types** with new combinations of traits. Unlinked genes exhibit a 50% frequency of recombination in the gametes. For genetically **linked genes**, **crossing over** accounts for the observed recombinants, always less than 50%.
- Recombination frequencies observed in genetic crosses allow construction of a **linkage map** (a type of **genetic map**).

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

CONCEPT 12.4

Alterations of chromosome number or structure cause some genetic disorders (pp. 240–243)

- **Aneuploidy**, an abnormal chromosome number, results 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 complete nondisjunction.
- Chromosome breakage can result in alterations of chromosome structure: **deletions**, **duplications**, **inversions**, and **translocations**.

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

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

1. A man with hemophilia (a recessive, sex-linked condition) has a normal daughter, who marries a normal man. What is the probability that a daughter will be a hemophiliac? A son? If the couple has 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 apparently normal 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 space probe discovers a planet 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 morphology (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.

Level 2: Application/Analysis

4. Using the information from problem 3, scientists do a further testcross using a heterozygote for height and nose morphology. The offspring are tall–upturned snout, 40; dwarf–upturned snout, 9; dwarf–downturned snout, 42; tall–downturned snout, 9. Calculate the recombination frequency from these data; then use your answer from problem 3 to determine the correct sequence of the three linked genes.
5. A man with red-green color blindness (a recessive, sex-linked condition) marries a woman with normal vision whose father was color-blind. What is the probability that they will have a color-blind daughter? That their first son will be color-blind? (Note the different wording in the two questions.)
6. You design *Drosophila* crosses to provide recombination data for gene *a*, which is located on the chromosome shown in Figure 12.12. Gene *a* has recombination frequencies of 14% with the vestigial-wing locus and 26% with the brown-eye locus. Approximately where is gene *a* located along the chromosome?
7. 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 Figure 12.9, what fruit flies (genotypes and phenotypes) would you mate to determine the sequence of the body-color, wing-size, and eye-color genes on the chromosome?
8. 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? Without knowing that these genes are on the same chromosome, how would you interpret the results of this cross?
9. 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 F_2 progeny are obtained. How many F_2 plants of each of the four phenotypes do you expect?

Level 3: Synthesis/Evaluation

10. **SCIENTIFIC INQUIRY**
Butterflies have an X-Y sex determination system that is different from that of flies or humans. Female butterflies may be either XY or XO, while butterflies with two or more X chromosomes are males. This photograph shows a tiger swallowtail *gynandromorph*, an individual that 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.
11. **FOCUS ON EVOLUTION**
Crossing over, or recombination, is thought to be evolutionarily advantageous because it continually shuffles genetic alleles into novel combinations. Until recently, it was thought that Y-linked genes might degenerate because they have no homologous genes on the X chromosome with which to recombine. However, when the Y chromosome was sequenced, eight large regions were found to be internally homologous to each other, and quite a few of the 78 genes represent duplicates. How might this be beneficial?
12. **FOCUS ON 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.

For selected answers, see Appendix A.

13 The Molecular Basis of Inheritance

KEY CONCEPTS

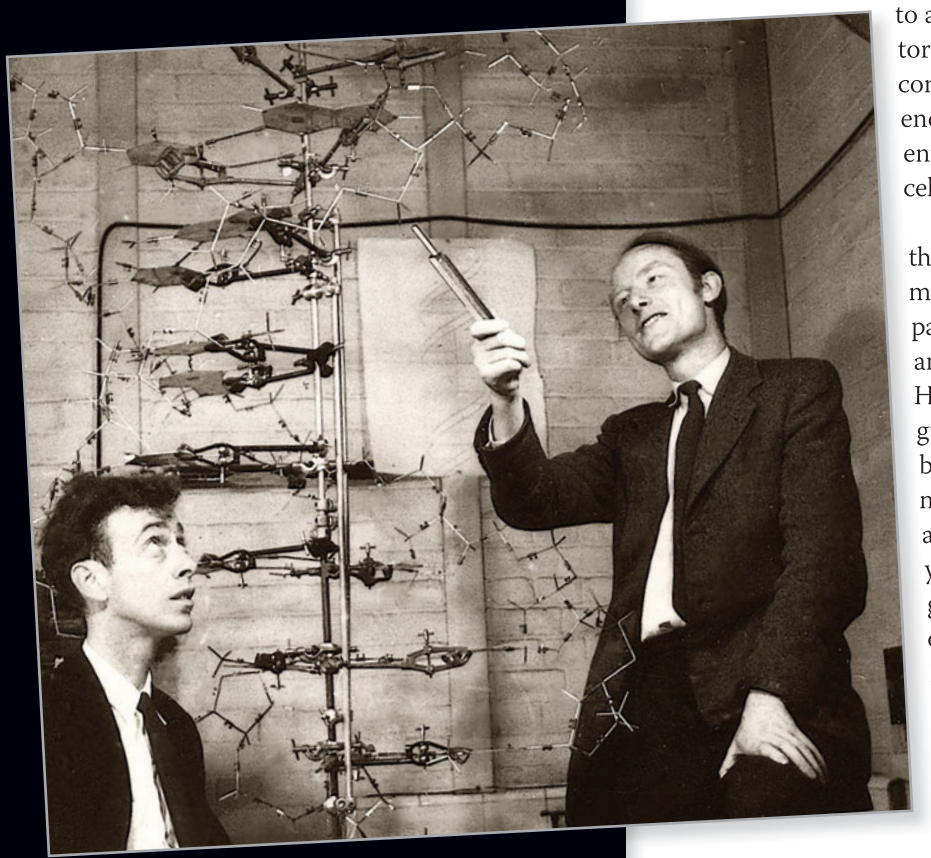
13.1 DNA is the genetic material

13.2 Many proteins work together in DNA replication and repair

13.3 A chromosome consists of a DNA molecule packed together with proteins

13.4 Understanding DNA structure and replication makes genetic engineering possible

▼ **Figure 13.1** How was the structure of DNA determined?



OVERVIEW

Life's Operating Instructions

In April 1953, James Watson and Francis Crick shook the scientific world with an elegant double-helical model for the three-dimensional structure of deoxyribonucleic acid, or DNA. **Figure 13.1** shows Watson (left) and Crick admiring their DNA model, which they built from tin and wire. Over the past 60 years or so, their model has evolved from a novel proposition to an icon of modern biology. Mendel's heritable factors and Morgan's genes on chromosomes are, in fact, composed of DNA. Chemically speaking, your genetic endowment is the DNA you inherited from your parents. DNA, the substance of inheritance, is the most celebrated molecule of our time.

Of all nature's molecules, nucleic acids are unique in their ability to direct their own replication from monomers. Indeed, the resemblance of offspring to their parents has its basis in the precise replication of DNA and its transmission from one generation to the next. Hereditary information is encoded in the chemical language of DNA and reproduced in all the cells of your body. It is this DNA program that directs the development of your biochemical, anatomical, physiological, and, to some extent, behavioral traits. In this chapter, you'll discover how biologists deduced that DNA is the genetic material and how Watson and Crick worked out its structure. You'll also learn how a molecule of DNA is copied during **DNA replication** and how cells repair their DNA. Next, you'll see how DNA is packaged with proteins in a chromosome. Finally, you'll explore how an understanding of DNA-related processes has allowed scientists to directly manipulate genes for practical purposes.

CONCEPT 13.1

DNA is the genetic material

Today, even schoolchildren have heard of DNA, and scientists routinely manipulate DNA in the laboratory, often to change the heritable traits of cells in their experiments. Early in the 20th century, however, identifying the molecules of inheritance loomed as 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 Chapter 12), 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, especially since biochemists had identified them 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 experiments with microorganisms yielded unexpected results. As with the work of Mendel and Morgan, a key factor in determining the identity of the genetic material was the choice of appropriate experimental organisms. The role of DNA in heredity was first worked out while studying bacteria and the viruses that infect them, which are far simpler than pea plants, fruit flies, or humans. In this section, we'll trace the search for the genetic material in some detail 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 13.2)**. Furthermore, this newly acquired trait of pathogenicity was inherited by all the descendants of the transformed bacteria. Clearly, 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 and others identified the transforming substance as DNA.

Scientists remained skeptical, however, many viewing proteins as better candidates for the genetic material. Moreover, many biologists were not convinced that the genes of bacteria 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.

Evidence That Viral DNA Can Program Cells

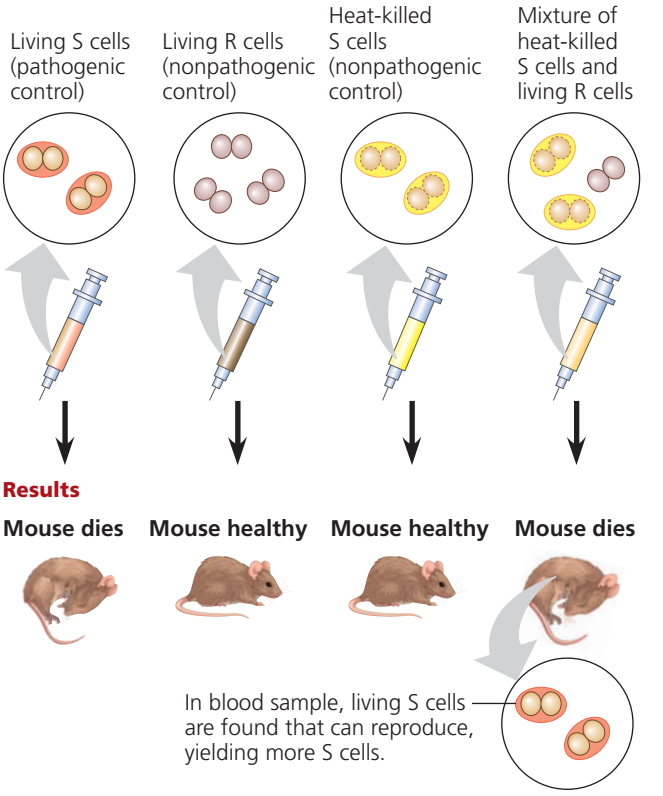
Additional evidence for DNA as the genetic material came from studies of viruses that infect bacteria. These viruses are called **bacteriophages** (meaning "bacteria-eaters"), or **phages** for short. Viruses are much simpler than cells. A **virus** is little more than DNA (or sometimes RNA) enclosed by a protective coat, which is often simply protein **(Figure 13.3)**. To produce more viruses, a virus must infect a cell and take over the cell's metabolic machinery.

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

▼ Figure 13.2 Inquiry

Can a genetic trait be transferred between different bacterial strains?

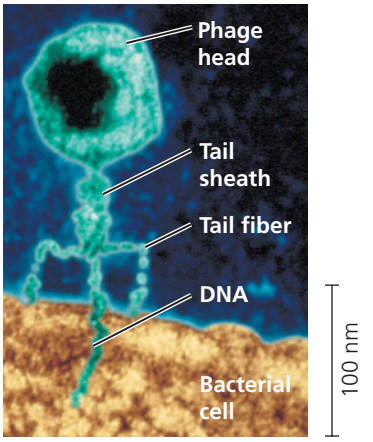
Experiment Frederick Griffith studied two strains of the bacterium *Streptococcus pneumoniae*. Bacteria of the S (smooth) strain can cause pneumonia in mice; they are pathogenic because they have an outer capsule that protects them from an animal's immune system. Bacteria of the R (rough) strain lack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



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

Source 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 could have simply used the capsules of the dead S cells to become pathogenic?



▶ Figure 13.3 Viruses infecting a bacterial cell. Phages called T2 attach to the host cell and inject their genetic material through the plasma membrane while the head and tail parts remain on the outer bacterial surface (colorized TEM).

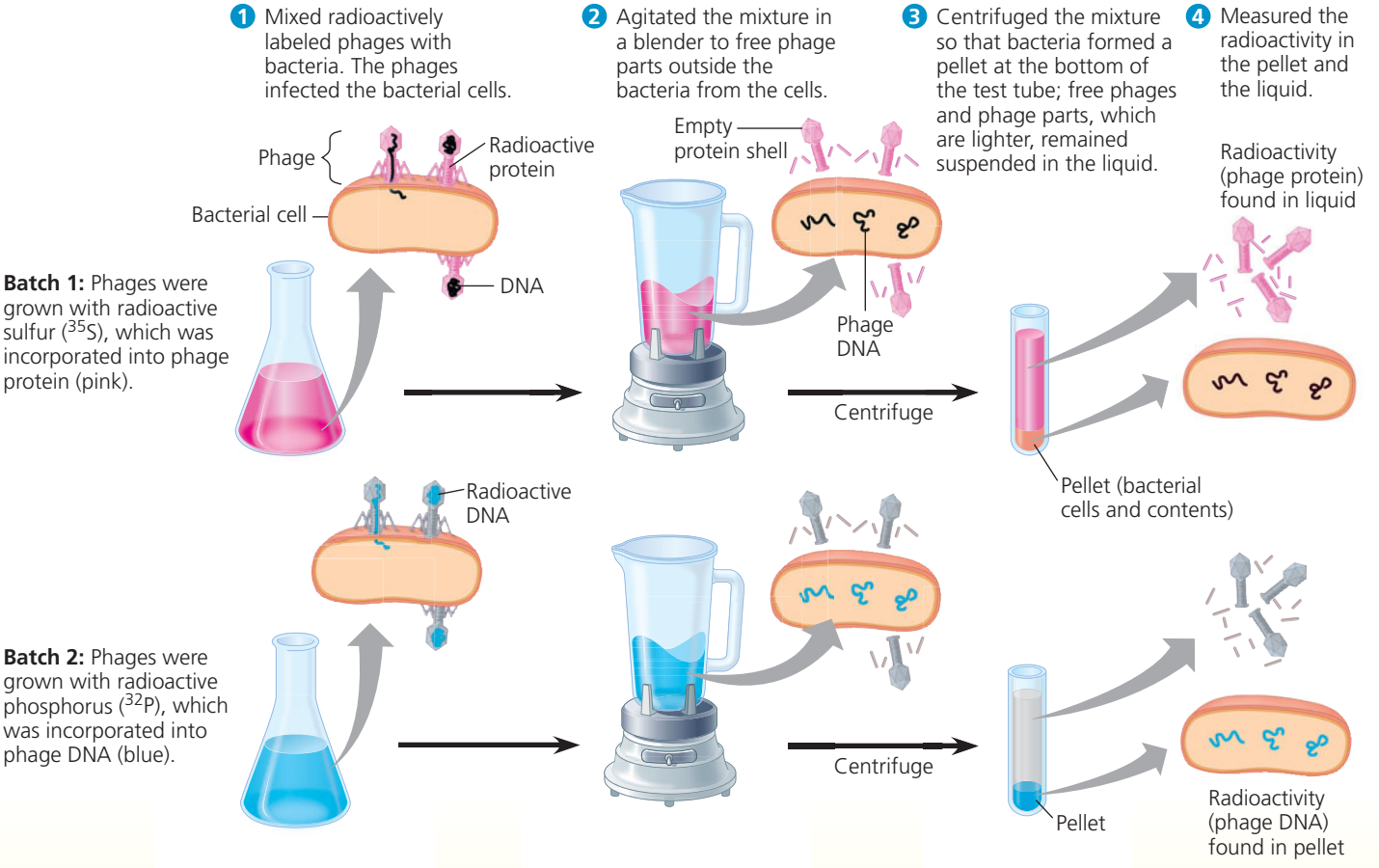
infect *Escherichia coli* (*E. coli*), a bacterium that normally lives in the intestines of mammals and is a model organism for molecular biologists. At that time, biologists already knew that T2, like many other phages, was composed almost entirely of DNA and protein. They also knew that the T2 phage could quickly turn an *E. coli* cell into a T2-producing factory that released many copies 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 13.4)**.

▼ Figure 13.4 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. Bacterial cells with 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.

Source 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?*

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 allowed to be infected by the protein-labeled and DNA-labeled batches of T2. The researchers then tested the two samples shortly after the onset of infection to see which type of molecule—protein or DNA—had entered the bacterial cells and would therefore have been 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, the infection ran its course, and the *E. coli* released phages that contained some radioactive phosphorus, further showing that the DNA inside the cell played an ongoing role during the infection process.

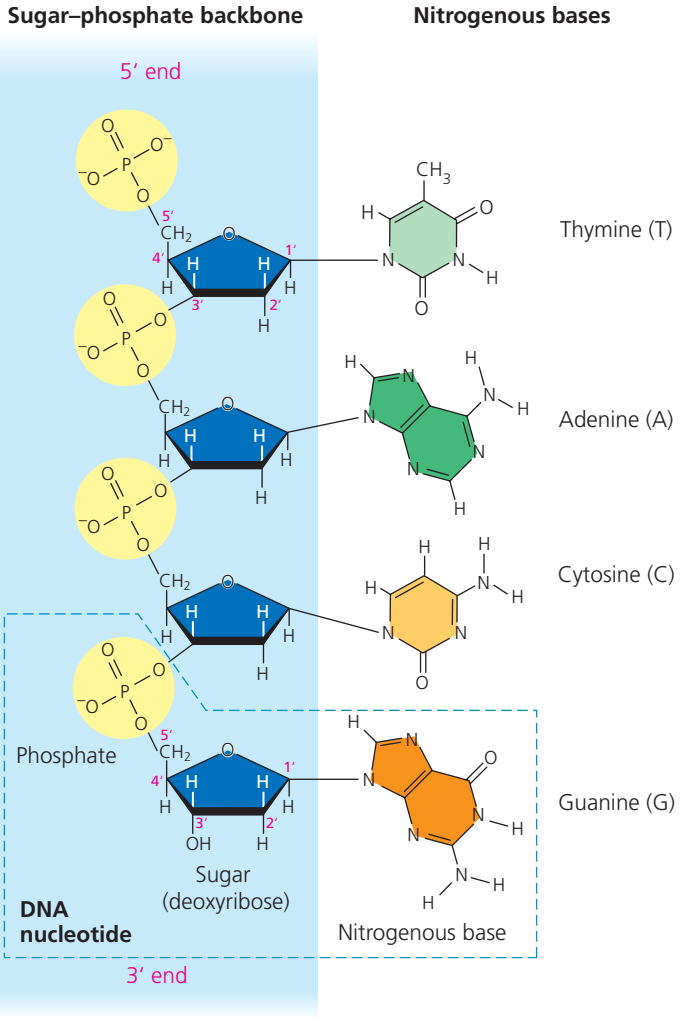
Hershey and Chase concluded that the DNA injected by the phage must be the molecule carrying the genetic information that makes the cells produce new viral DNA and proteins. The Hershey-Chase experiment was a landmark study because it provided powerful evidence that nucleic acids, rather than proteins, are the hereditary material, at least for 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. It was already known that DNA is a polymer of nucleotides, each consisting of three components: a nitrogenous (nitrogen-containing) base, a pentose sugar called deoxyribose, and a phosphate group **(Figure 13.5)**. 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, 32.8% of sea urchin DNA nucleotides have the base A, whereas only 24.7% of the DNA nucleotides from the bacterium *E. coli* have an A. This evidence of molecular diversity among species, which had been presumed 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%.

These two findings became known as *Chargaff's rules*: (1) the base composition varies between species, and



▲ Figure 13.5 The structure of a DNA strand. Each DNA nucleotide monomer consists of a nitrogenous base (T, A, C, or G), the sugar deoxyribose (blue), and a phosphate group (yellow). The phosphate group of one nucleotide is attached to the sugar of the next, forming a "backbone" of alternating phosphates and sugars from which the bases project. The 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.

(2) within a species, the number of A and T bases are roughly equal and the number of G and C bases are roughly equal. In the **Scientific Skills Exercise,** you can use Chargaff's rules to predict unknown percentages of nucleotide bases. The rationale for these rules remained unexplained until the discovery of the double helical structure of DNA.

Building a Structural Model of DNA: *Scientific Inquiry*

Once most biologists were convinced that DNA was the genetic material, the challenge was to determine how the structure of DNA could account for its role in inheritance. By the early 1950s, the arrangement of covalent bonds in a single nucleic acid polymer was well established (see Figure 13.5), and researchers focused on discovering the three-dimensional

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 coworkers noticed a pattern in the base composition of nucleotides from different organisms: the number of adenine (A) bases roughly equaled the number of thymine (T) bases, and the number of cytosine (C) bases roughly equaled the number of guanine (G) bases. Further, each species they studied had a different distribution of A/T and C/G bases. We now know that these consistent ratios are due to complementary base pairing between A and T and between C and G in the DNA double helix, and interspecies differences 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 nucleotide bases in a genome.

How the Experiments Were Done In Chargaff's experiments, DNA was extracted from the given organism, denatured, and hydrolyzed to break apart the individual nucleotides before analyzing them chemically. These experiments provided approximate values for each type of nucleotide. Today, the availability of whole-genome sequencing has allowed 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, species). You can apply the patterns that you see in the known data to predict unknown values. In the table in the upper right, 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	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			

Interpret the Data

1. Explain how the sea urchin and salmon data demonstrate both of Chargaff's rules.
2. Based on 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 is valid, that the amount of A equals the amount of T and the amount of C equals the amount of G, 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?

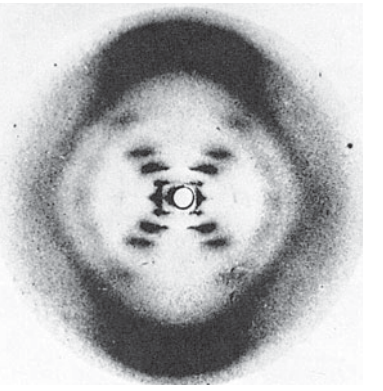
Data from several papers by Chargaff: for example, E. Chargaff et al., Composition of the desoxypentose nucleic acids of four genera of sea-urchin, *Journal of Biological Chemistry* 195:155–160 (1952).

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

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 correct answer, however, were two scientists who were relatively unknown at the time—the American James Watson and the Englishman Francis Crick.

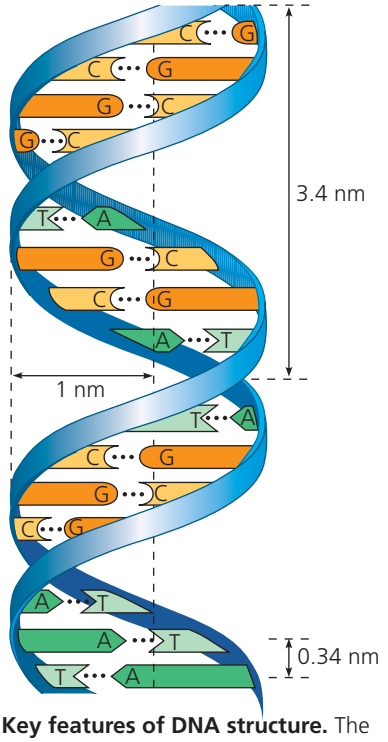
The brief but celebrated partnership that solved the puzzle of DNA structure began soon after Watson journeyed to Cambridge University, where Crick was studying protein structure with a technique called X-ray crystallography (see Figure 3.24). While visiting the laboratory of Maurice Wilkins, Watson saw an X-ray diffraction image of DNA produced by Wilkins's accomplished colleague Rosalind Franklin **(Figure 13.6a)**. Images produced by X-ray crystallography are not actually pictures of molecules. The spots and smudges in **Figure 13.6b** 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 that helical molecules produce, and an examination of the photo that Wilkins showed him confirmed that DNA was helical in shape. It also augmented 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 that Linus Pauling had proposed a short time earlier. The

(a) Rosalind Franklin

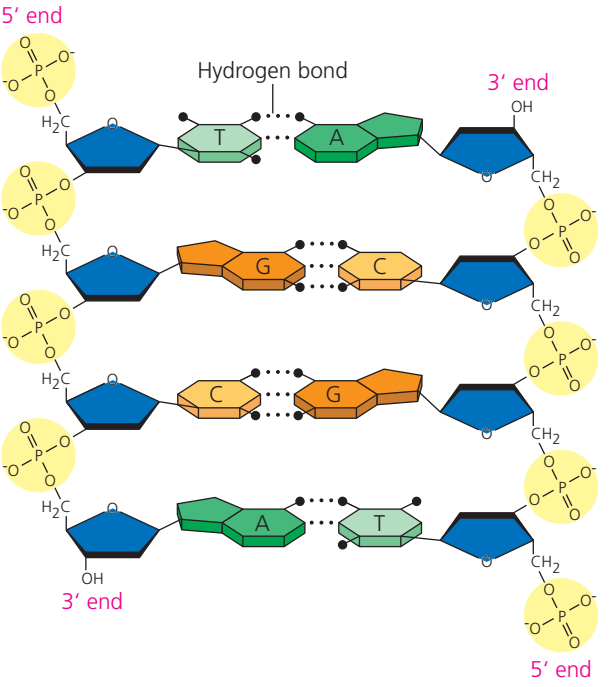


(b) Franklin's X-ray diffraction photograph of DNA

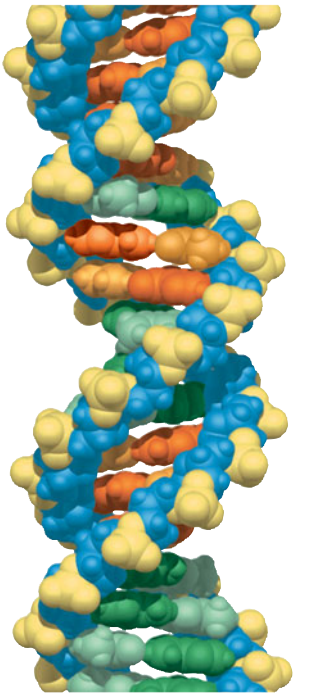
▲ Figure 13.6 Rosalind Franklin and her X-ray diffraction photo of DNA. Franklin, a very accomplished X-ray crystallographer, conducted critical experiments resulting in the photograph that allowed Watson and Crick to deduce the double-helical structure of DNA.



(a) Key features of DNA structure. The "ribbons" in this diagram represent the sugar-phosphate backbones of the two DNA strands. The helix is "right-handed," curving up to the right. The two strands are held together by hydrogen bonds (dotted lines) between the nitrogenous bases, which are paired in the interior of the double helix.



(b) Partial chemical structure. For clarity, the two DNA strands are shown untwisted in this partial chemical structure. Strong covalent bonds link the units of each strand, while weaker hydrogen bonds between the bases hold one strand to the other. Notice that the strands are antiparallel, meaning that they are oriented in opposite directions.



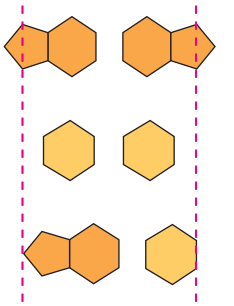
(c) Space-filling model. The tight stacking of the base pairs is clear in this computer model. Van der Waals interactions between the stacked pairs play a major role in holding the molecule together.

▲ Figure 13.7 The double helix.

presence of two strands accounts for the now-familiar term **double helix** **(Figure 13.7)**.

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 rules. They knew that Franklin had concluded that the sugar-phosphate backbones were on the outside of the DNA molecule. This 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 (see Figure 13.1). In this model, the two sugar-phosphate backbones are **antiparallel**—that is, their subunits run in opposite directions (see Figure 13.7b). You can imagine the overall arrangement as a rope ladder with rigid rungs. The side ropes represent 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 "rungs" of base pairs in each full turn of the helix.

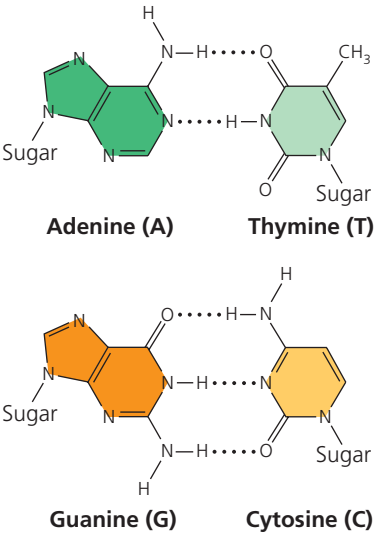
The nitrogenous bases of the double helix are paired in specific combinations: adenine (A) with thymine (T), and guanine (G) with cytosine (C). It was mainly by trial and error that Watson and Crick arrived at this key feature of DNA. At first, Watson imagined that the bases paired like with like—for 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. Thus, purines (A and G) are about twice as wide as pyrimidines (C and T). A purine-purine pair is too wide and a pyrimidine-pyrimidine pair too narrow to account for the 2-nm diameter of the double helix. Always pairing a purine with a pyrimidine, however, results in a uniform diameter:



Purine + purine: too wide

Pyrimidine + pyrimidine: too narrow

Purine + pyrimidine: width consistent with X-ray data



▲ **Figure 13.8 Base pairing in DNA.** The pairs of nitrogenous bases in a DNA double helix are held together by hydrogen bonds, shown here as black dotted lines.

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 forms 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 13.8)**.

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. 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 order, or 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 the symbol 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 38 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.

CONCEPT CHECK 13.1

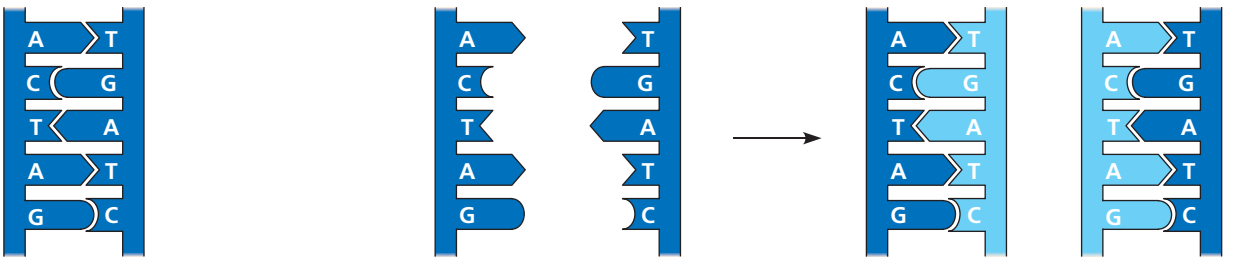
1. Given a polynucleotide sequence such as GAATTC, can you tell which is the 5′ end? If not, what further information do you need to identify the ends? (See Figure 13.5.)
2. **WHAT IF?** Griffith did not expect transformation to occur in his experiment. What results was he expecting? Explain.

For suggested answers, see Appendix A.

CONCEPT 13.2

Many proteins work together in DNA replication and repair

The relationship between structure and function is manifest in the double helix. The idea that there is specific pairing of nitrogenous bases in DNA was the flash of inspiration that led Watson and Crick to the double helix. At the same time, they saw the functional significance of the base-pairing rules. They ended their classic paper with this wry statement: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." In this section, you'll learn about the basic principle of DNA replication (**Figure 13.9**), as well as some important details of the process.



(a) The parental molecule (dark blue) 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 strands are connected to form the sugar-phosphate backbones of the new (light blue) strands.

▲ **Figure 13.9 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, here represented as ladder rungs. Dark blue represents DNA strands present in the parental molecule; light blue represents newly synthesized DNA.

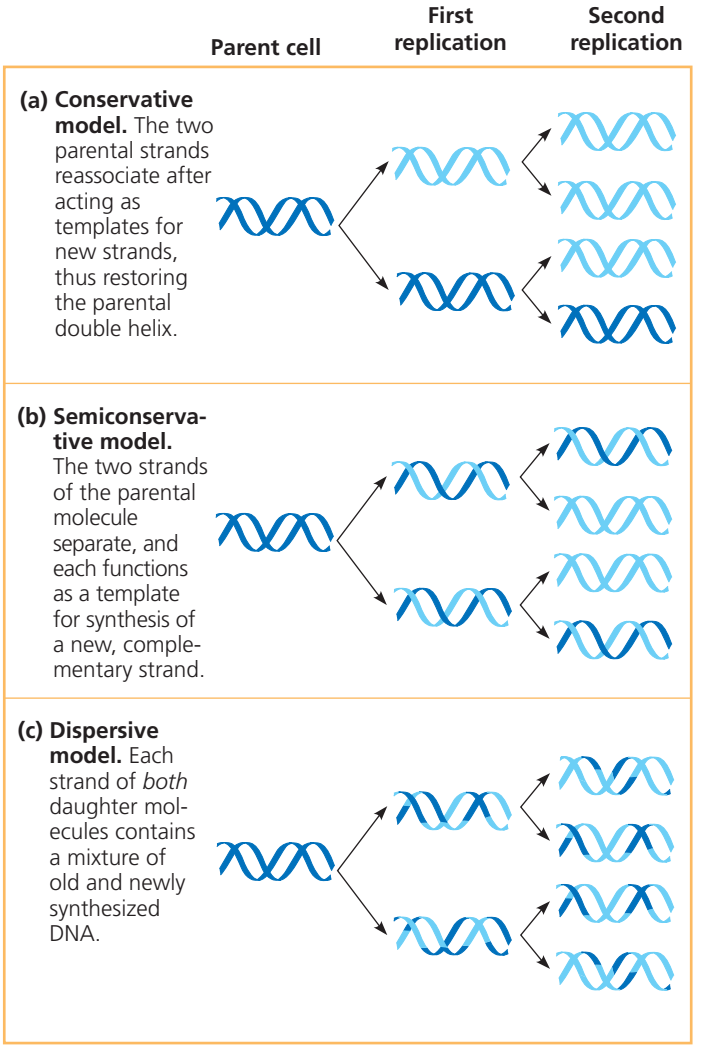
The Basic Principle: Base Pairing to a Template Strand

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 onto 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.*

Figure 13.9 illustrates Watson and Crick's basic idea. To make it easier to follow, only a short section of double helix is shown, in untwisted form. Notice that if you cover one of the two DNA strands of Figure 13.9a, you can still determine its linear sequence of nucleotides by referring to the uncovered strand and applying the base-pairing rules. 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. Where there was one double-stranded DNA molecule at the beginning of the process, there are soon two, each an exact replica of the "parental" molecule. The copying mechanism is analogous to using a photographic negative to make a positive image, which can in turn be used to make another negative, and so on.

This model of DNA replication remained untested for several years following publication of the DNA structure. The requisite 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. These three models are shown in **Figure 13.10**. Although mechanisms for conservative or dispersive DNA replication are not easy to come up with, these models remained possibilities until they could be ruled out. After two years of preliminary work in the late 1950s, Matthew Meselson and Franklin Stahl devised a clever experiment that distinguished between the three models, described in detail in **Figure 13.11**. Their experiment supported the semiconservative model of DNA replication, as predicted by Watson and Crick,



▲ **Figure 13.10 Three alternative models of DNA replication.** 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. Newly made DNA is light blue.

and is widely acknowledged among biologists to be a classic example of elegant experimental 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*

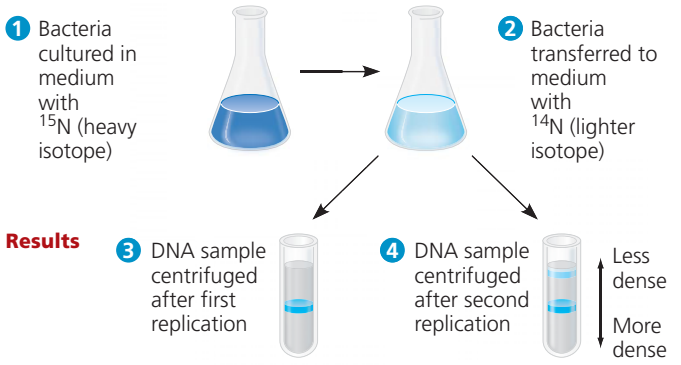
The bacterium *E. coli* has a single chromosome of about 4.6 million nucleotide pairs. In a favorable environment, an *E. coli* cell can copy all this DNA and divide to form two genetically identical daughter cells in less than an hour. Each of *your* cells has 46 DNA molecules in its nucleus, one long double-helical molecule per chromosome. In all, that represents about 6 billion nucleotide pairs, or over a thousand times more DNA than is found in a bacterial cell. If we were to print the one-letter symbols for these bases (A, G, C, and T) the size of the

*F. H. C. Crick and J. D. Watson, The complementary structure of deoxyribonucleic acid, *Proceedings of the Royal Society of London A* 223:80 (1954).

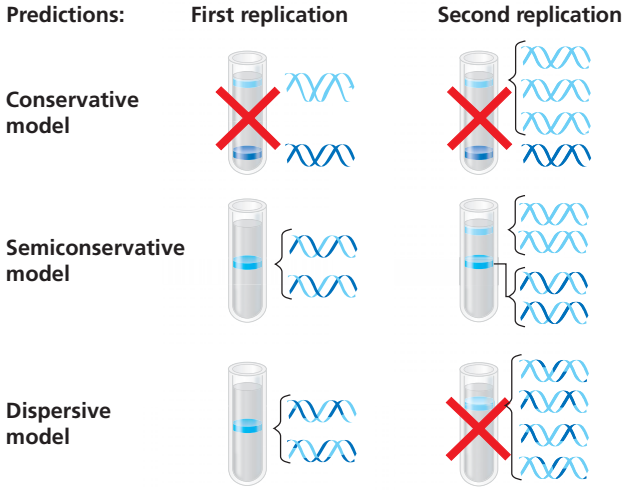
▼ Figure 13.11 Inquiry

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

Experiment At the California Institute of Technology, 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. A sample was taken after DNA replicated once; another sample was taken after DNA replicated again. 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 with those predicted by each of the three models in Figure 13.10, as shown below. The first replication in the ^{14}N medium produced a band 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.



Source 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*.

MB A related Experimental Inquiry Tutorial can be assigned in MasteringBiology.

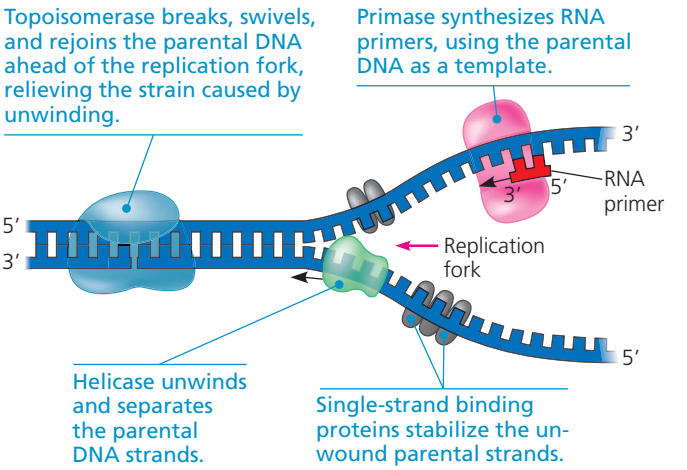
WHAT IF? If Meselson and Stahl had first grown the cells in ^{14}N-containing medium and then moved them into ^{15}N-containing medium before taking samples, what would have been the result?

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. 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 a DNA molecule begins at particular sites called **origins of replication**, short stretches of DNA having a specific sequence of nucleotides. Proteins that initiate DNA replication recognize this sequence and attach to the DNA, separating the two strands and opening up a replication "bubble." At each end of a 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 13.12)**. **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** helps relieve this strain by breaking, swiveling, and rejoining DNA strands.



▲ Figure 13.12 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.

The *E. coli* chromosome, like many other bacterial chromosomes, is circular and has a single origin of replication, forming one replication bubble **(Figure 13.13a)**. 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 long DNA molecules **(Figure 13.13b)**. As in bacteria, eukaryotic DNA replication proceeds in both directions from each origin.

Synthesizing a New DNA Strand

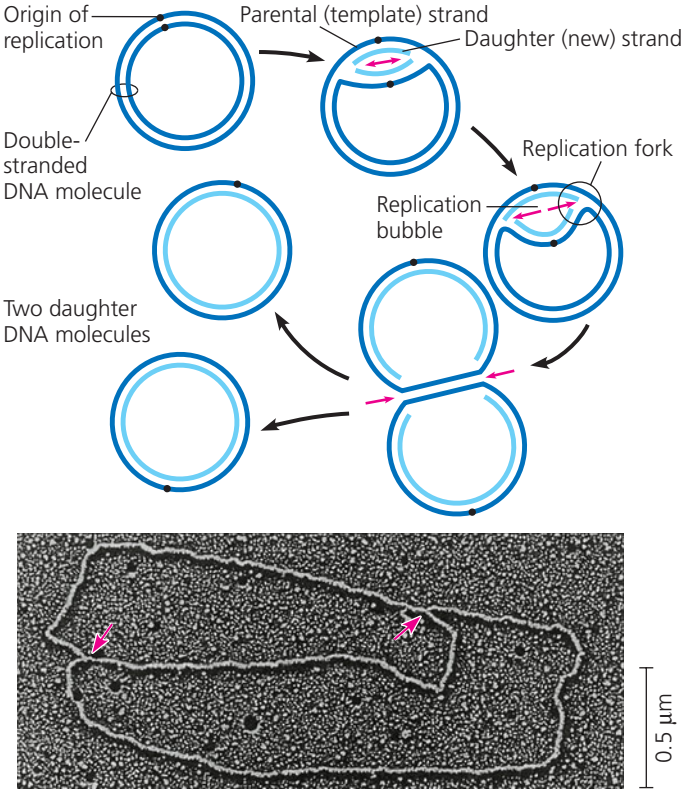
Within a bubble, the unwound sections of parental DNA strands are available to serve as templates for the synthesis of new complementary DNA strands. However, the enzymes that synthesize DNA cannot *initiate* the synthesis of a polynucleotide; they can only add nucleotides to the end of an already existing chain that is base-paired with the template strand. The initial nucleotide chain that is produced during DNA synthesis is actually a short stretch of RNA, not DNA. This RNA chain is called a **primer** and is synthesized by the enzyme **primase** (see Figure 13.12). Primase starts a complementary RNA chain from a single RNA nucleotide, adding RNA nucleotides one at a time, using the parental DNA strand as a template. The completed primer, generally 5–10 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 a preexisting chain. In *E. coli*, there are several different DNA polymerases, but two 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; however, the general principles are the same.

Most DNA polymerases require a primer and a DNA template strand along which complementary DNA nucleotides line up. In *E. coli*, DNA polymerase III (abbreviated DNA pol

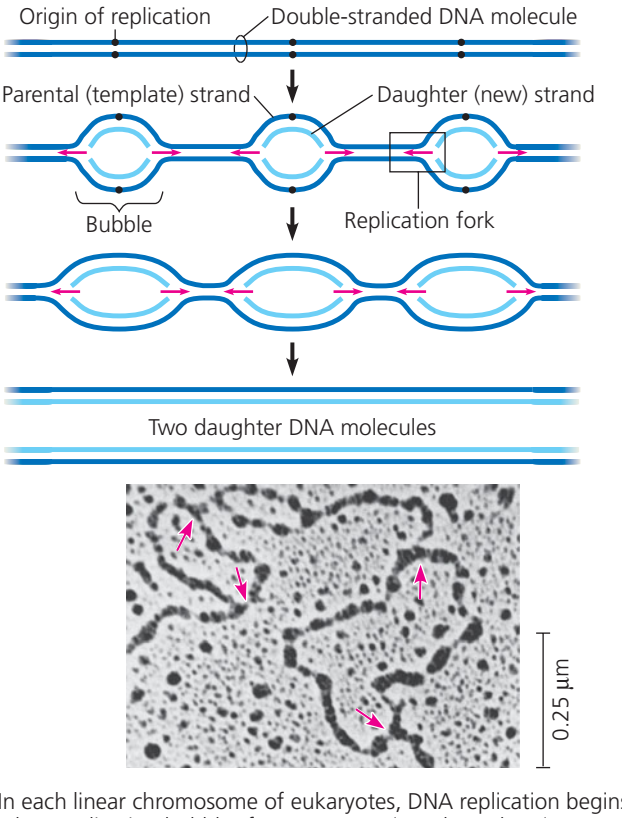
▼ Figure 13.13 Origins of replication in *E. coli* and eukaryotes. The red arrows indicate the movement of the replication forks and thus the overall directions of DNA replication within each bubble.

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



In the circular chromosome of *E. coli* and many other bacteria, only one origin of replication is present. The parental strands separate at the origin, forming a replication bubble with two forks. 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 each linear chromosome of eukaryotes, DNA replication begins when replication bubbles form at many sites along the giant DNA molecule. The bubbles expand as replication proceeds in both directions. 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.

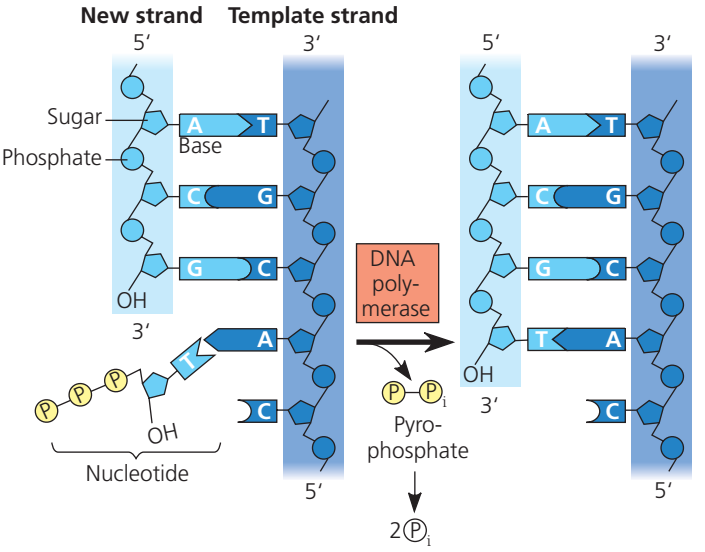
DRAW IT *In the TEM above, add arrows for the third bubble.*

III) adds a DNA nucleotide to the RNA primer and then continues adding DNA nucleotides, 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 three phosphate groups. You have already encountered such a molecule—ATP (adenosine triphosphate; see Figure 6.8). 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. As each monomer joins the growing end of a DNA strand, 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 13.14)**.

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 13.5). 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 a divided highway (see Figure 13.14). Therefore, the two new strands formed during DNA replication must also end up antiparallel to their template strands.

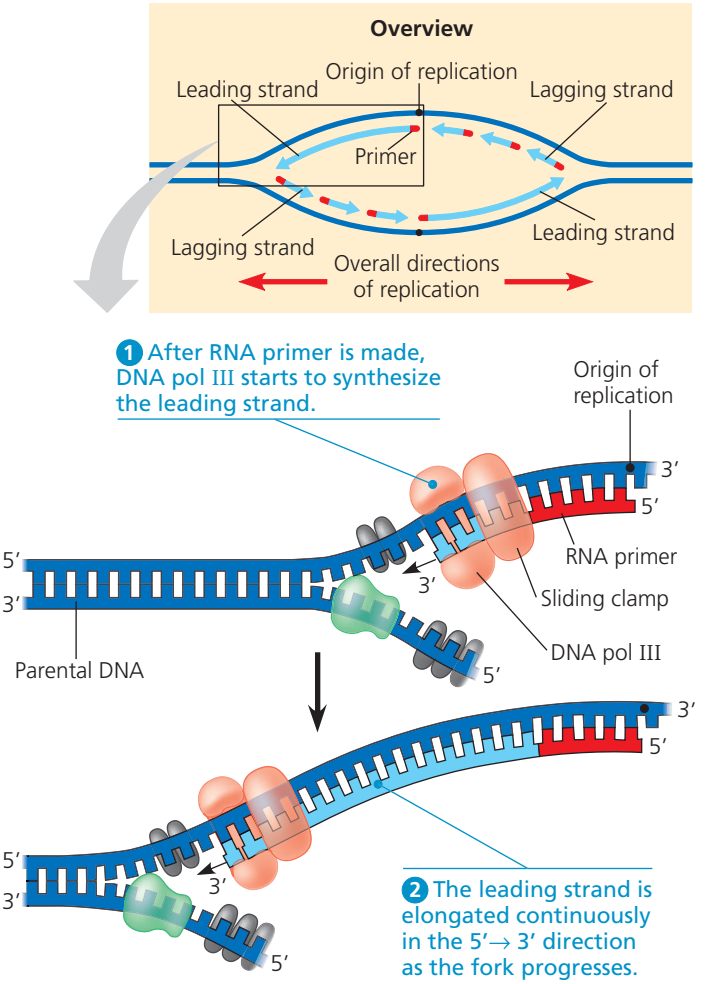


▲ Figure 13.14 Addition of a nucleotide to a DNA strand. DNA polymerase catalyzes the addition of a nucleotide to the 3′ end of a growing DNA strand, with the release of two phosphates.

? *Use this diagram to explain what we mean when we say that each DNA strand has directionality.*

How does the antiparallel arrangement of the double helix affect replication? 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 13.14). 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 13.15)**. 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 leading strand.

To elongate the other new strand of DNA in the mandatory 5′ → 3′ direction, DNA pol III must work along the other



▲ Figure 13.15 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.

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 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.

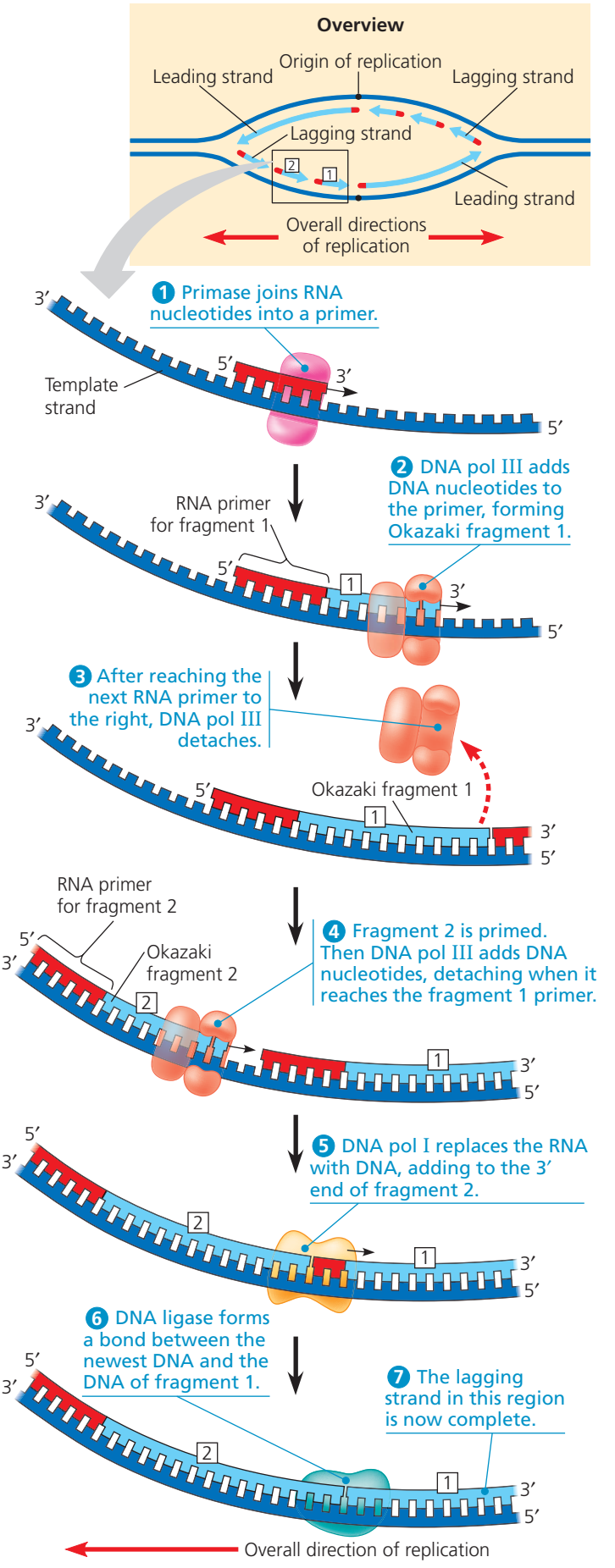
Figure 13.16 illustrates the steps in the synthesis of the lagging strand at one fork. Whereas only one primer is required on the leading strand, each Okazaki fragment on the lagging strand must be primed separately (1 and 4). After DNA pol III forms an Okazaki fragment (2–4), another DNA polymerase, DNA polymerase I (DNA pol I), replaces the RNA nucleotides of the adjacent primer with DNA nucleotides (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 (6).

Figure 13.17 summarizes DNA replication. Study it carefully before proceeding.

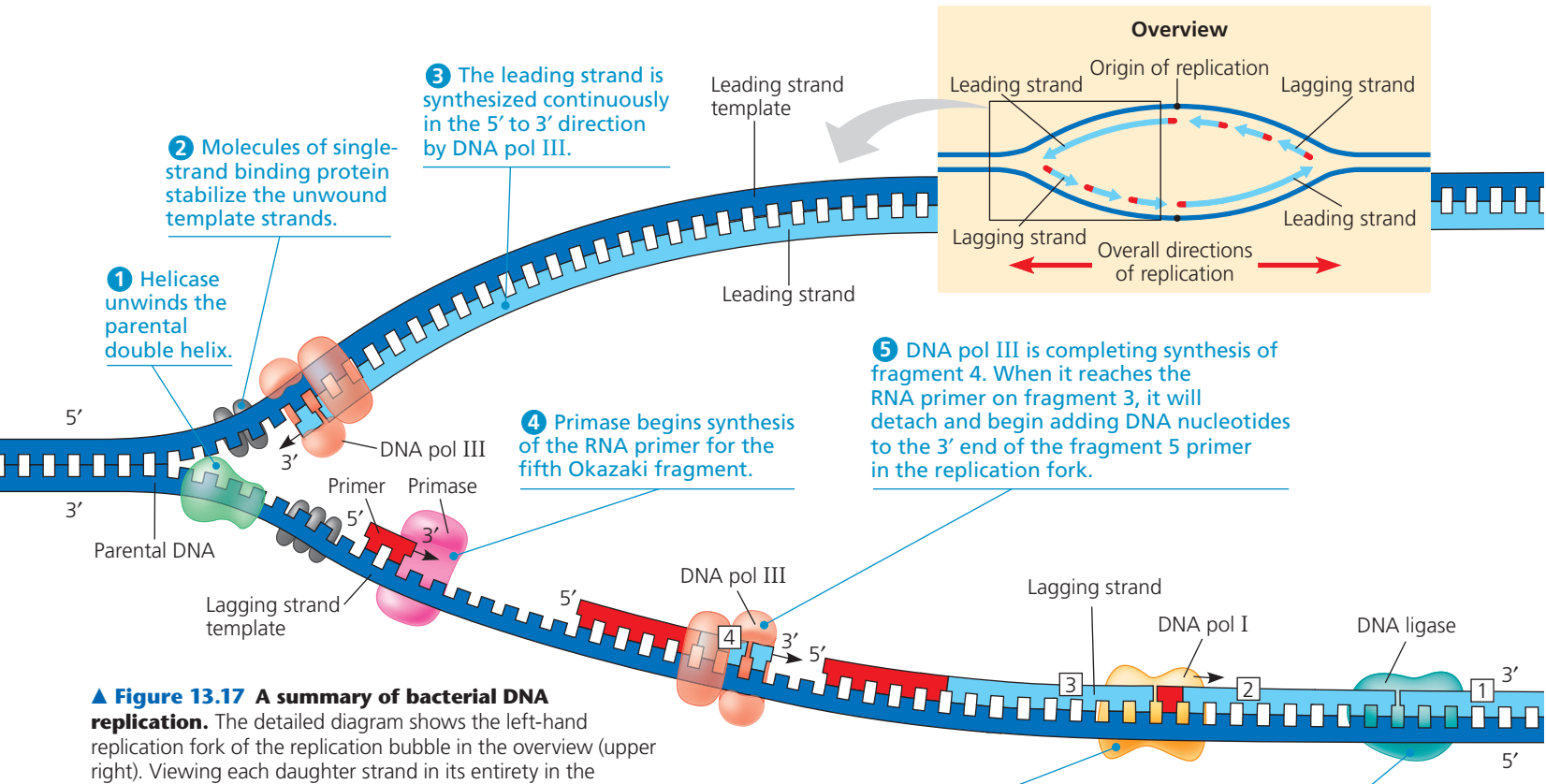
The DNA Replication Complex

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. Recent studies support 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. Additional evidence suggests that the lagging strand is looped back through the complex **(Figure 13.18)**.

*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.

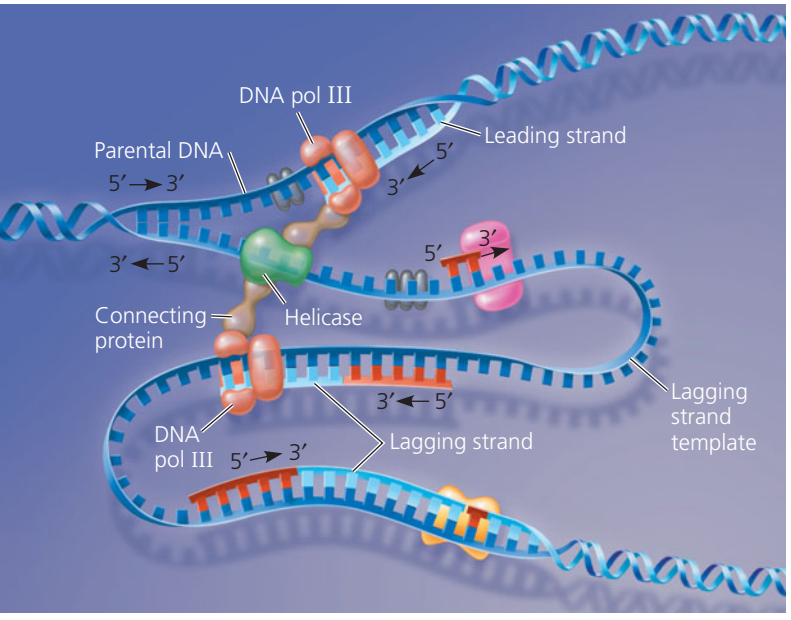


▲ **Figure 13.16 Synthesis of the lagging strand.**

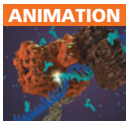


▲ Figure 13.17 A summary of bacterial DNA replication. The detailed diagram shows the left-hand replication fork of the replication bubble in the overview (upper right). Viewing each daughter strand in its entirety in the overview, you can see that half of it is made continuously as the leading strand, while the other half (on the other side of the origin) is synthesized in fragments as the lagging strand.

DRAW IT *Draw a similar diagram showing the right-hand fork of this bubble, numbering the Okazaki fragments appropriately. Label all 5′ and 3′ ends.*



▲ Figure 13.18 A current model of the DNA replication complex. Two DNA polymerase III molecules work together in a complex, one on each template strand. The lagging strand template DNA loops through the complex.



BioFlix® Visit the Study Area in **MasteringBiology** for the BioFlix® 3-D Animation on DNA Replication.

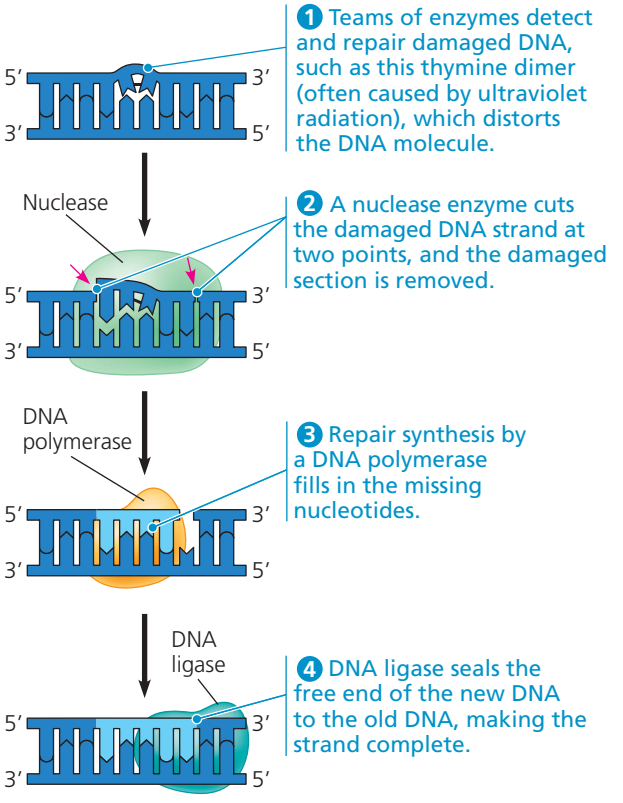
Proofreading and Repairing DNA

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, DNA polymerases proofread each nucleotide against its template as soon as it is added 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 typing error by deleting the wrong letter and then entering the correct letter.)

Mismatched nucleotides sometimes do evade proofreading by a DNA polymerase. In **mismatch repair**, other enzymes remove and replace incorrectly paired nucleotides resulting from replication errors. Researchers spotlighted 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 cigarette smoke and X-rays (as we'll discuss in Chapter 14). In addition, DNA bases often 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 130 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. One such DNA repair system, shown in **Figure 13.19**, is called **nucleotide excision repair**.



▲ **Figure 13.19** **Nucleotide excision repair of DNA damage.**

An important function of the DNA repair enzymes in our skin cells is to repair genetic damage caused by the ultraviolet rays of sunlight. One example of this damage is when adjacent thymine bases on a DNA strand become covalently linked. Such *thymine dimers* cause the DNA to buckle (see Figure 13.19) and interfere with DNA replication. The importance of repairing this kind of damage is underscored by the disorder xeroderma pigmentosum, which in most cases is caused by an inherited defect in a nucleotide excision repair enzyme. Individuals with this disorder are hypersensitive to sunlight; mutations in their skin cells caused by ultraviolet light are left uncorrected, resulting in skin cancer.

Evolutionary Significance of Altered DNA Nucleotides

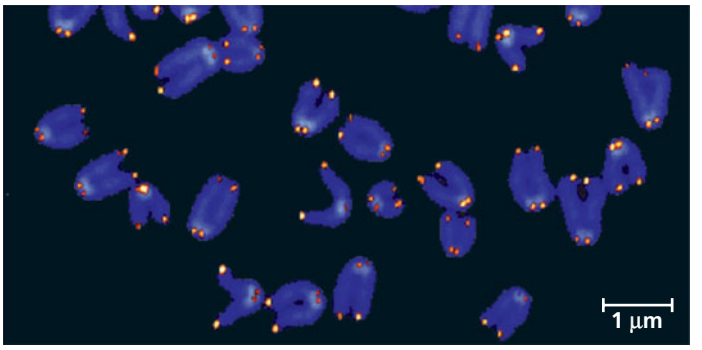
EVOLUTION Faithful replication of the genome and repair of DNA damage are important for the functioning of the organism and for passing on a complete, accurate genome to the next generation. The error rate after proofreading and repair is extremely low, but rare mistakes do slip through. Once a mismatched 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 you know, a permanent change in the DNA sequence is called a mutation.

Mutations can change the phenotype of an organism (as you'll learn in Chapter 14). 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 are harmful, but a very small percentage can be beneficial. In either case, mutations are the 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 Three.) The balance between complete fidelity of DNA replication or repair and a low mutation rate has, over long periods of time, allowed the evolution of the rich diversity of species we see 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. (This is a consequence of the fact that a DNA polymerase can add nucleotides only to the 3′ ends.) As a result, repeated rounds of replication produce shorter and shorter DNA molecules with uneven ends.

What protects the genes near the ends of eukaryotic chromosomes from being eroded away during successive replications? Eukaryotic chromosomal DNA molecules have special nucleotide sequences called telomeres at their ends **(Figure 13.20)**. 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 sequence



▲ **Figure 13.20 Telomeres.** Eukaryotes have repetitive, noncoding sequences called telomeres at the ends of their DNA. Telomeres are stained orange in these mouse chromosomes (LM).

TTAGGG is repeated 100 to 1,000 times. Telomeric DNA acts as a buffer zone that protects the organism's genes.

Telomeres do not prevent the erosion of genes near ends of chromosomes; they merely postpone it. As you would expect, telomeres tend to be shorter in cultured cells that have divided many times and in dividing somatic cells of older individuals. Shortening of telomeres is proposed to play a role in the aging process of some tissues and even of the organism as a whole.

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. Telomerase is not active in most human somatic cells, but shows inappropriate activity in some cancer cells that may remove limits to a cell's normal life span. Thus, telomerase is under study as a target for cancer therapies.

CONCEPT CHECK 13.2

1. What role does base pairing play in the replication of DNA?
2. Make a table listing the functions of seven proteins involved in DNA replication in *E. coli*.
3. **MAKE CONNECTIONS** What is the relationship between DNA replication and the S phase of the cell cycle? See Figure 9.6.

For suggested answers, see Appendix A.

CONCEPT 13.3

A chromosome consists of a DNA molecule packed together with proteins

Now that you have learned about the structure and replication of DNA, let's take a step back and examine how DNA is packaged into chromosomes, the structures that carry genetic information. The main component of the genome in most bacteria is one double-stranded, circular DNA molecule that is associated with a small amount of protein. Although we refer to this structure as a bacterial chromosome, it is very different from a eukaryotic chromosome, which consists of one linear DNA molecule associated with a large amount of protein. 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. Still, that is a lot 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, 500 times longer than 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. Unlike the nucleus of a eukaryotic cell, this dense region of DNA in a bacterium, called the **nucleoid**, is not surrounded by membrane (see Figure 4.5).

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!

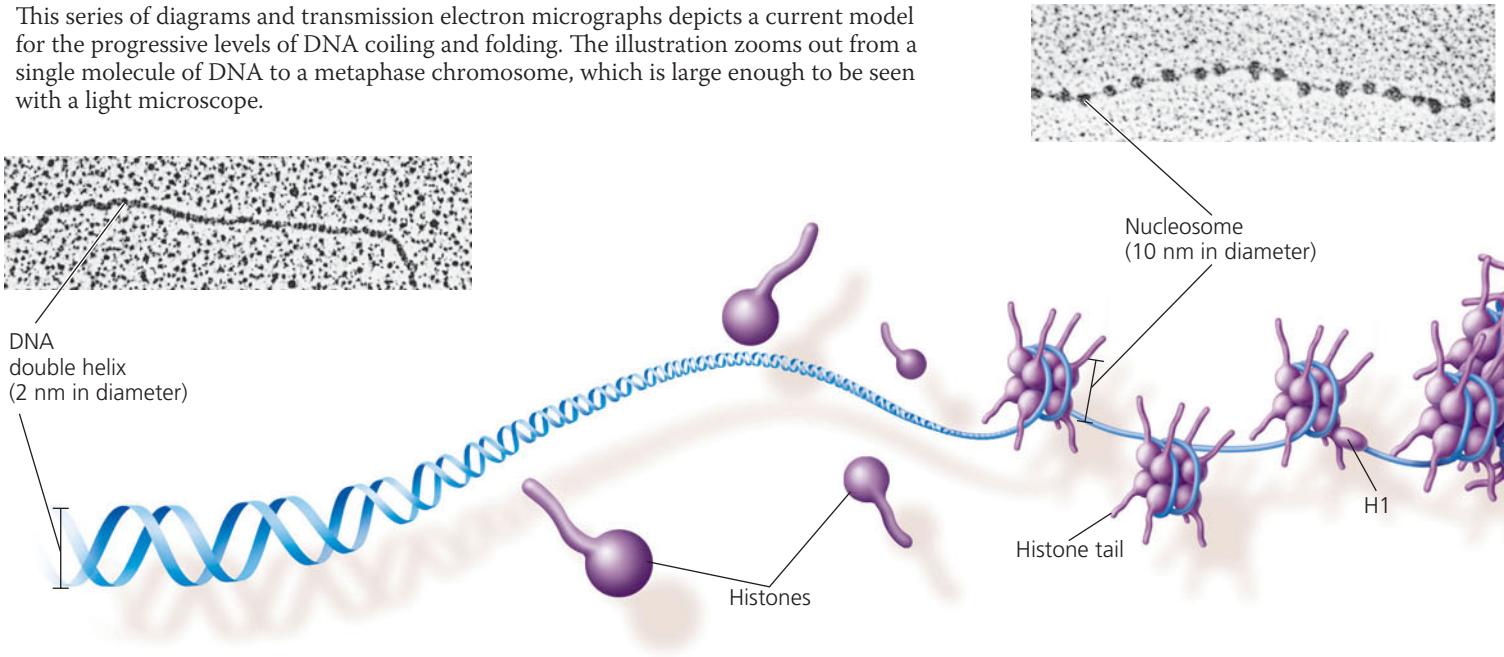
In the cell, eukaryotic DNA is precisely combined with a large amount of protein. Together, this complex of DNA and protein, called **chromatin**, fits into the nucleus through an elaborate, multilevel system of packing.

Chromatin undergoes striking changes in its degree of packing during the course of the cell cycle (see Figure 9.7). In interphase cells stained for light microscopy, the chromatin usually appears as a diffuse mass within the nucleus, suggesting that the chromatin is highly extended. As a cell prepares for mitosis, its chromatin coils and folds up (condenses), eventually forming a characteristic number of short, thick metaphase chromosomes that are distinguishable from each other with the light microscope. Our current view of the successive levels of DNA packing in a chromosome is outlined in **Figure 13.21**. Study this figure carefully before reading further.

Though interphase chromatin is generally much less condensed than the chromatin of mitotic chromosomes, it shows several of the same levels of higher-order packing. Some of the chromatin comprising a chromosome seems to be present as a 10-nm fiber, but much is compacted into a 30-nm fiber, which in some regions is further folded into looped domains. Even during interphase, the centromeres of chromosomes, as well as other chromosomal regions in some cells, exist in a highly condensed state similar to that seen in a metaphase chromosome. This type of interphase chromatin, visible as irregular clumps with a light microscope, is called **heterochromatin**, to distinguish it from the less compacted, more dispersed **euchromatin** ("true chromatin"). Because of its compaction, heterochromatic DNA is largely inaccessible to the machinery

▼ Figure 13.21 Exploring Chromatin Packing in a Eukaryotic Chromosome

This series of diagrams and transmission electron micrographs depicts a current model for the progressive levels of DNA coiling and folding. The illustration zooms out from a single molecule of DNA to a metaphase chromosome, which is large enough to be seen with a light microscope.



DNA, the double helix

Shown here is a ribbon model of DNA, with each ribbon representing one of the sugar-phosphate backbones. As you will recall from Figure 13.7, the phosphate groups along the backbone contribute a negative charge along the outside of each strand. The TEM shows a molecule of naked DNA; the double helix alone is 2 nm across.

Histones

Proteins called **histones** are responsible for the first level of DNA packing in chromatin. Although each histone is small—containing only about 100 amino acids—the total mass of histone in chromatin approximately equals the mass of DNA. More than a fifth of a histone's amino acids are positively charged (lysine or arginine) and therefore bind tightly to the negatively charged DNA.

Four types of histones are most common in chromatin: H2A, H2B, H3, and H4. The histones are very similar among eukaryotes; for example, all but two of the amino acids in cow H4 are identical to those in pea H4. The apparent conservation of histone genes during evolution probably reflects the important role of histones in organizing DNA within cells.

The four main types of histones are critical to the next level of DNA packing. (A fifth type of histone, called H1, is involved in a further stage of packing.)

Nucleosomes, or "beads on a string" (10-nm fiber)

In electron micrographs, unfolded chromatin is 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 composed of two molecules each of the four main histone types. The amino end (N-terminus) of each histone (the *histone tail*) extends outward from the nucleosome.

In the cell cycle, the histones leave the DNA only briefly during DNA replication. Generally, they do the same during transcription, another process that requires access to the DNA by the cell's molecular machinery. Chapter 18 will discuss some recent findings about the role of histone tails and nucleosomes in the regulation of gene expression.

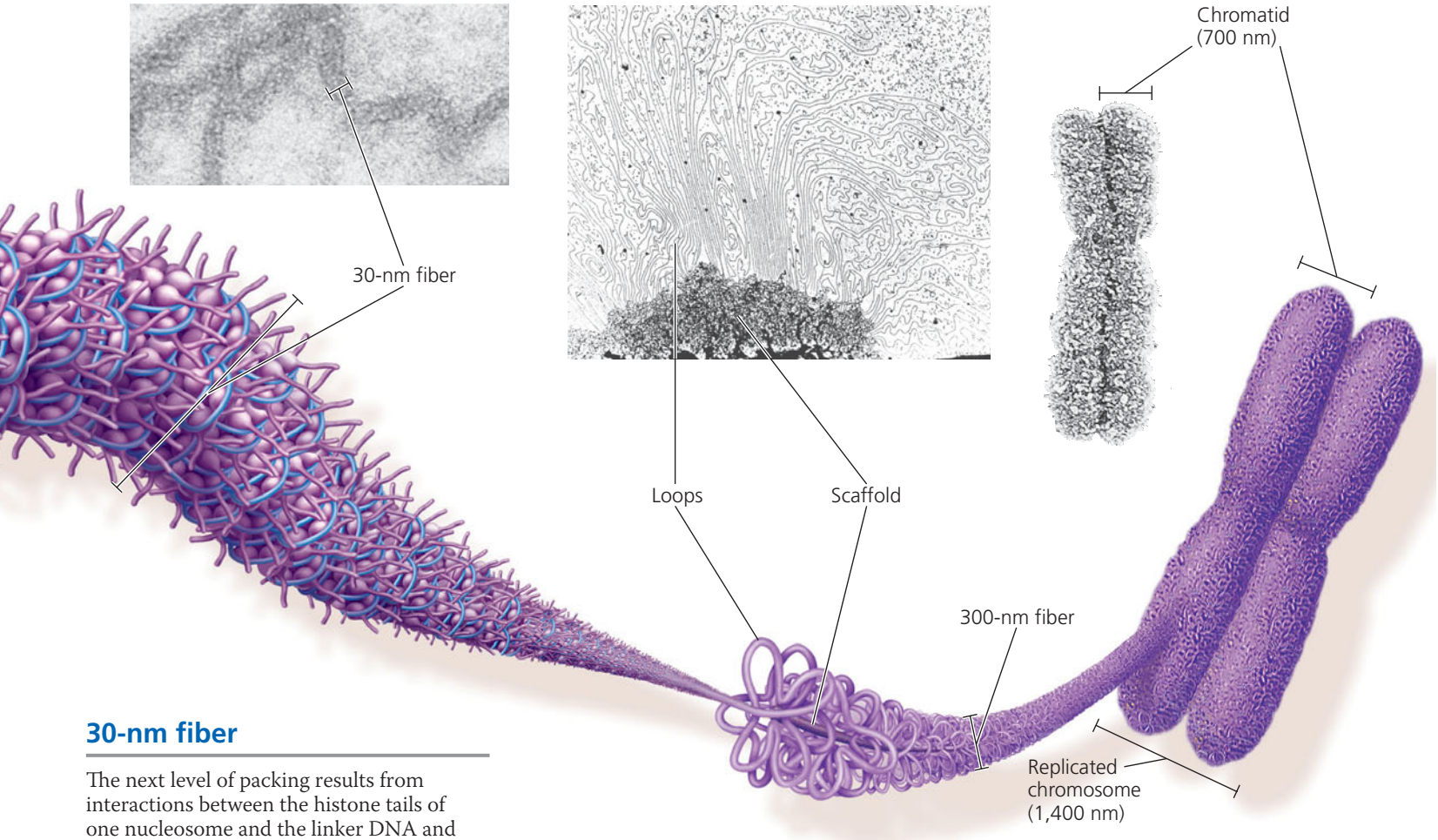
in the cell responsible for transcribing the genetic information coded in the DNA, a crucial early step in gene expression. In contrast, the looser packing of euchromatin makes its DNA accessible to this machinery, so the genes present in euchromatin can be transcribed.

The chromosome is a dynamic structure that is condensed, loosened, modified, and remodeled as necessary for various cell processes, including mitosis, meiosis, and gene activity. Certain chemical modifications of histones affect the state of chromatin condensation and also have multiple effects on gene activity (as you'll see in Chapter 15).

CONCEPT CHECK 13.3

1. Describe the structure of a nucleosome, the basic unit of DNA packing in eukaryotic cells.
2. What two properties, one structural and one functional, distinguish heterochromatin from euchromatin?

For suggested answers, see Appendix A.



30-nm fiber

The next level of packing results from interactions between the histone tails of one nucleosome and the linker DNA and nucleosomes on either side. A fifth histone, H1, is involved at this level. These interactions cause the extended 10-nm fiber to coil or fold, forming a chromatin fiber roughly 30 nm in thickness, the *30-nm fiber*. Although the 30-nm fiber is quite prevalent in the interphase nucleus, the packing arrangement of nucleosomes in this form of chromatin is still a matter of some debate.

Looped domains (300-nm fiber)

The 30-nm fiber, in turn, forms loops called *looped domains* attached to a chromosome scaffold composed of proteins, thus making up a *300-nm fiber*. The scaffold is rich in one type of topoisomerase, and H1 molecules also appear to be present.

Metaphase chromosome

In a mitotic chromosome, the looped domains themselves coil and fold in a manner not yet fully understood, further compacting all the chromatin to produce the characteristic metaphase chromosome shown in the micrograph above. The width of one chromatid is 700 nm. Particular genes always end up located at the same places in metaphase chromosomes, indicating that the packing steps are highly specific and precise.

CONCEPT 13.4

Understanding DNA structure and replication makes genetic engineering possible

The discovery of the structure of DNA marked a milestone in biology and changed the course of biological research. Most notable was the realization that the two strands of a DNA molecule are complementary to each other. This fundamental structural property of DNA is the basis for **nucleic acid hybridization**, the base pairing of one strand of a nucleic acid to a complementary sequence on another strand. 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 ranging from agriculture to criminal law to medical and basic biological research. In this section, we'll describe several of the most important techniques and their uses.

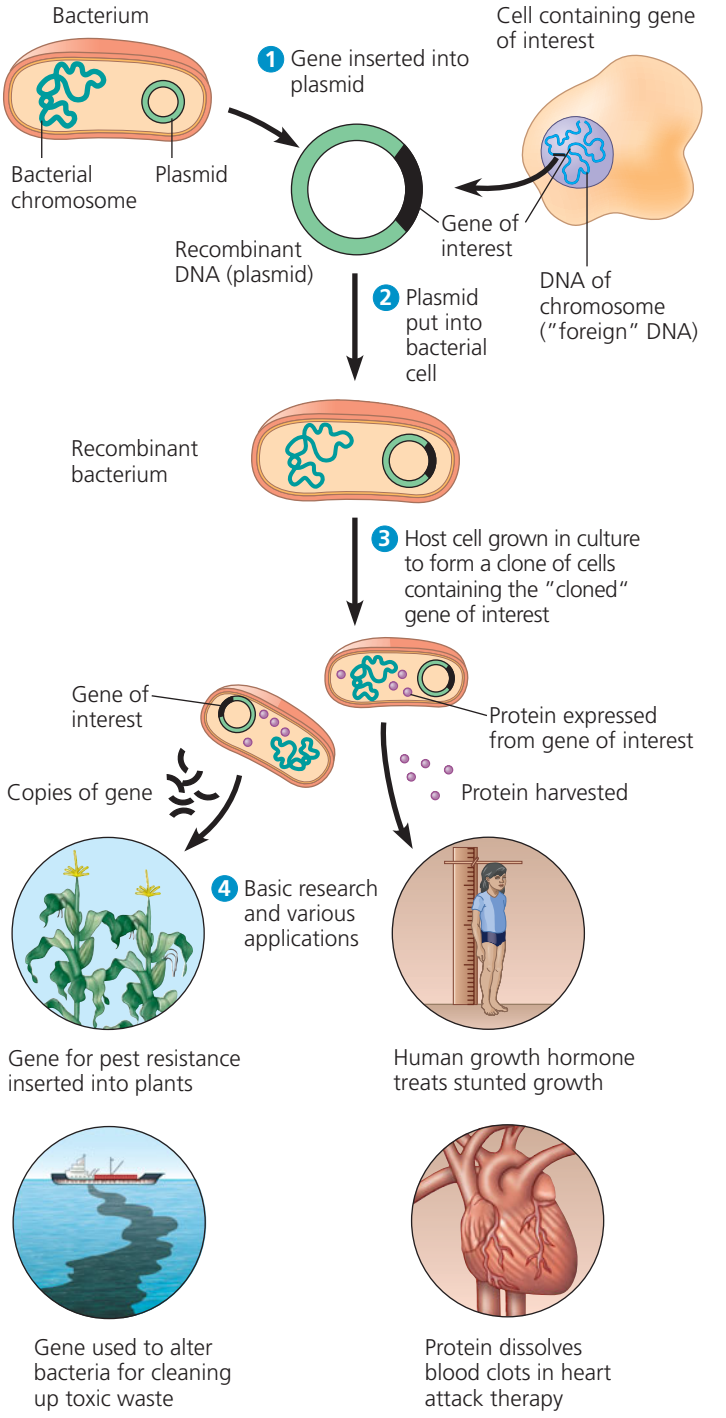
DNA Cloning: Making Multiple Copies of a Gene or Other DNA Segment

The molecular biologist studying a particular gene faces a challenge. Naturally occurring DNA molecules are very long, and a single molecule usually carries many genes. Moreover, in many eukaryotic genomes, 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, the distinctions between a gene and the surrounding DNA are subtle, consisting only of differences in nucleotide sequence. To work directly with specific genes, scientists have developed methods for preparing well-defined segments of DNA in multiple identical copies, 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 *E. coli*. Recall from Figure 13.13 that the *E. coli* chromosome is a large circular molecule of DNA. In addition, *E. coli* and many other bacteria have **plasmids**, small circular DNA molecules that replicate separately from the bacterial chromosome. 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 in the laboratory, researchers first obtain a plasmid (originally isolated from a bacterial cell and genetically engineered for efficient cloning) and insert DNA from another source ("foreign" DNA) into it **(Figure 13.22)**. The resulting plasmid is now **recombinant DNA**, a DNA molecule formed when segments of DNA from two different sources—often different species—are combined *in vitro* (in a test tube). 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 called **gene cloning**.

Gene cloning is useful for two basic purposes: to make many copies of, or *amplify*, a particular gene and to produce a protein product. Researchers can isolate copies of a cloned gene from bacteria for use in basic research or to endow an organism with a new metabolic trait, such as pest resistance. For example, a resistance gene present in one crop species might be cloned and transferred into plants of another species. Alternatively, a protein with medical uses, such as human growth hormone, can be harvested in large quantities from cultures of bacteria carrying the cloned gene for the protein. Since a single gene is usually a very small part of the total DNA in a cell, the ability to amplify such rare DNA fragments is therefore crucial for any application involving a single gene.



▲ Figure 13.22 An overview of 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.

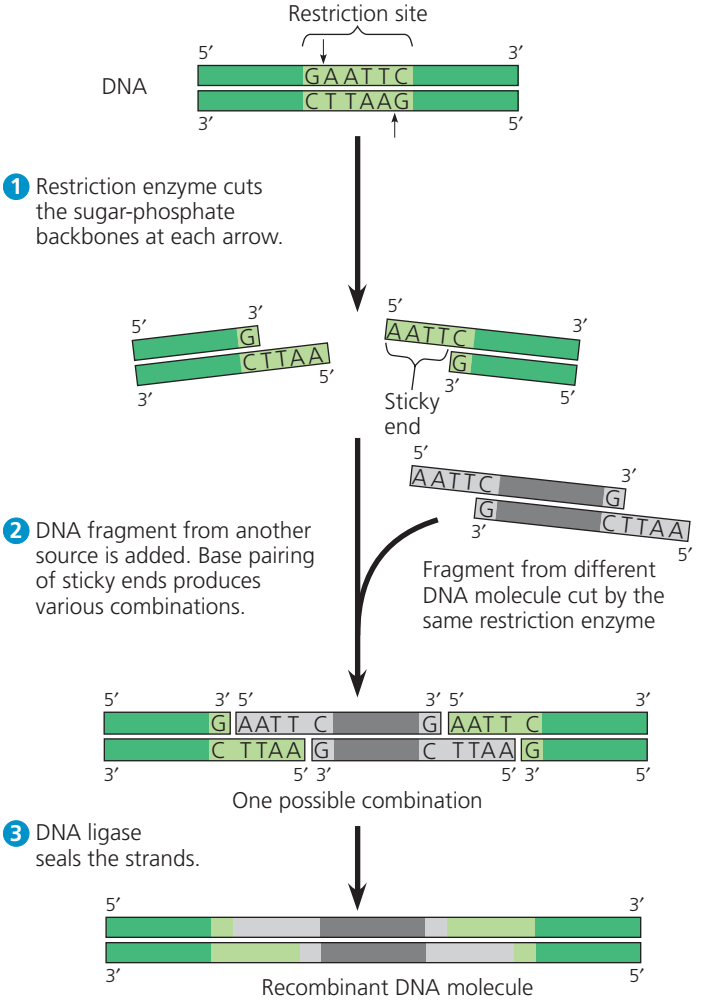
Using Restriction Enzymes to Make Recombinant DNA

Gene cloning and genetic engineering 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.

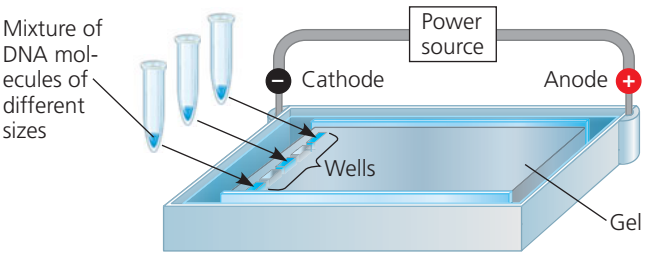
Hundreds of different restriction enzymes have been identified and isolated. Each restriction enzyme is very specific, recognizing a particular short DNA sequence, or **restriction site**, and cutting both DNA strands at precise points within this restriction site. The DNA of a bacterial cell is protected 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.

The top of **Figure 13.23** illustrates a restriction site recognized by a particular restriction enzyme from *E. coli*. As shown in this example, most restriction sites are symmetric. That is, 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 4–8 nucleotides. Because any sequence this short usually occurs (by chance) many times in a long DNA molecule, a restriction enzyme will make many cuts in a DNA molecule, yielding a set of **restriction fragments**. All copies of a particular DNA molecule always yield the same set of restriction fragments when exposed to the same restriction enzyme. To see the fragments, researchers carry out a technique called **gel electrophoresis**, which can separate a mixture of nucleic acid fragments by length **(Figure 13.24)**.

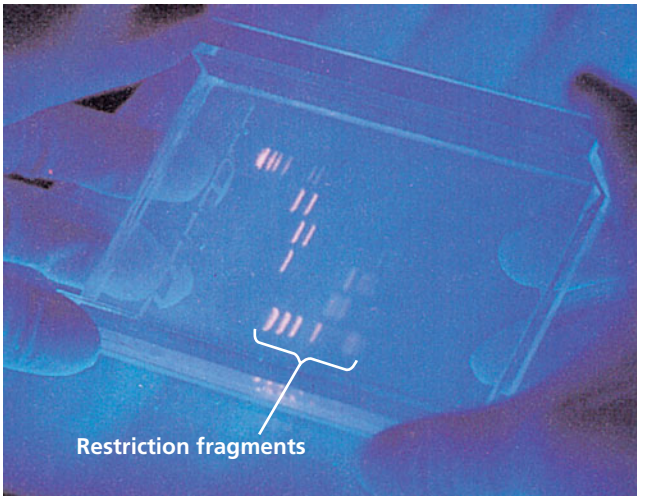


▲ Figure 13.23 Using a restriction enzyme and DNA ligase to make recombinant DNA. The restriction enzyme in this example (called *Eco*RI) recognizes a specific six-base-pair sequence, the restriction site, and makes staggered cuts in the sugar-phosphate backbones within this sequence, producing fragments with sticky ends. Any fragments with complementary sticky ends can base-pair, including the two original fragments. If the fragments come from different DNA molecules, the ligated product is recombinant DNA.

DRAW IT *The restriction enzyme HindIII recognizes the sequence 5′-AAGCTT-3′, cutting between the two A's. Draw the double-stranded sequence before and after the enzyme cuts.*



(a) Each sample, a mixture of 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 impeded less than longer ones, so they move faster through the gel. After the current is turned off, a DNA-binding dye is added that fluoresces pink in ultraviolet light. Each pink band corresponds to many thousands of DNA molecules of the same length. The horizontal ladder of bands at the bottom of the gel is a set of restriction fragments used as size standards.

▲ Figure 13.24 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 the polysaccharide agarose.

The most useful restriction enzymes cleave the sugar-phosphate backbones in the two DNA strands in a staggered manner, as indicated in Figure 13.23. 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 (hybridize) with complementary sticky ends on any other DNA molecules cut with the same enzyme. The associations formed in this way are only temporary but can be made permanent by DNA ligase. As you saw in Figure 13.16, this enzyme catalyzes the formation of covalent bonds that close up the sugar-phosphate backbones of DNA strands; for example, it joins Okazaki fragments during replication.

You can see at the bottom of Figure 13.23 that the ligase-catalyzed joining of DNA from two different sources produces a stable recombinant DNA molecule. In gene cloning, the two DNA molecules to be joined are a **cloning vector**—a DNA molecule that can carry foreign DNA into a host cell and replicate there—and the gene to be cloned (see Figure 13.22). The cloning vector is often a bacterial plasmid that has one copy of a restriction site recognized by a particular restriction enzyme, selected by the researcher and purchased from a commercial source. The most common way to obtain many copies of the gene to be cloned is described next.

Amplifying DNA *in Vitro*: The Polymerase Chain Reaction (PCR) and Its Use in Cloning

Today, most researchers have some information about the sequence of the gene or DNA fragment they want to clone. Using this information, they can start with the entire collection of genomic DNA from the particular species of interest and obtain enough copies of the desired gene by using a technique called the **polymerase chain reaction**, or **PCR**. **Figure 13.25** 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.

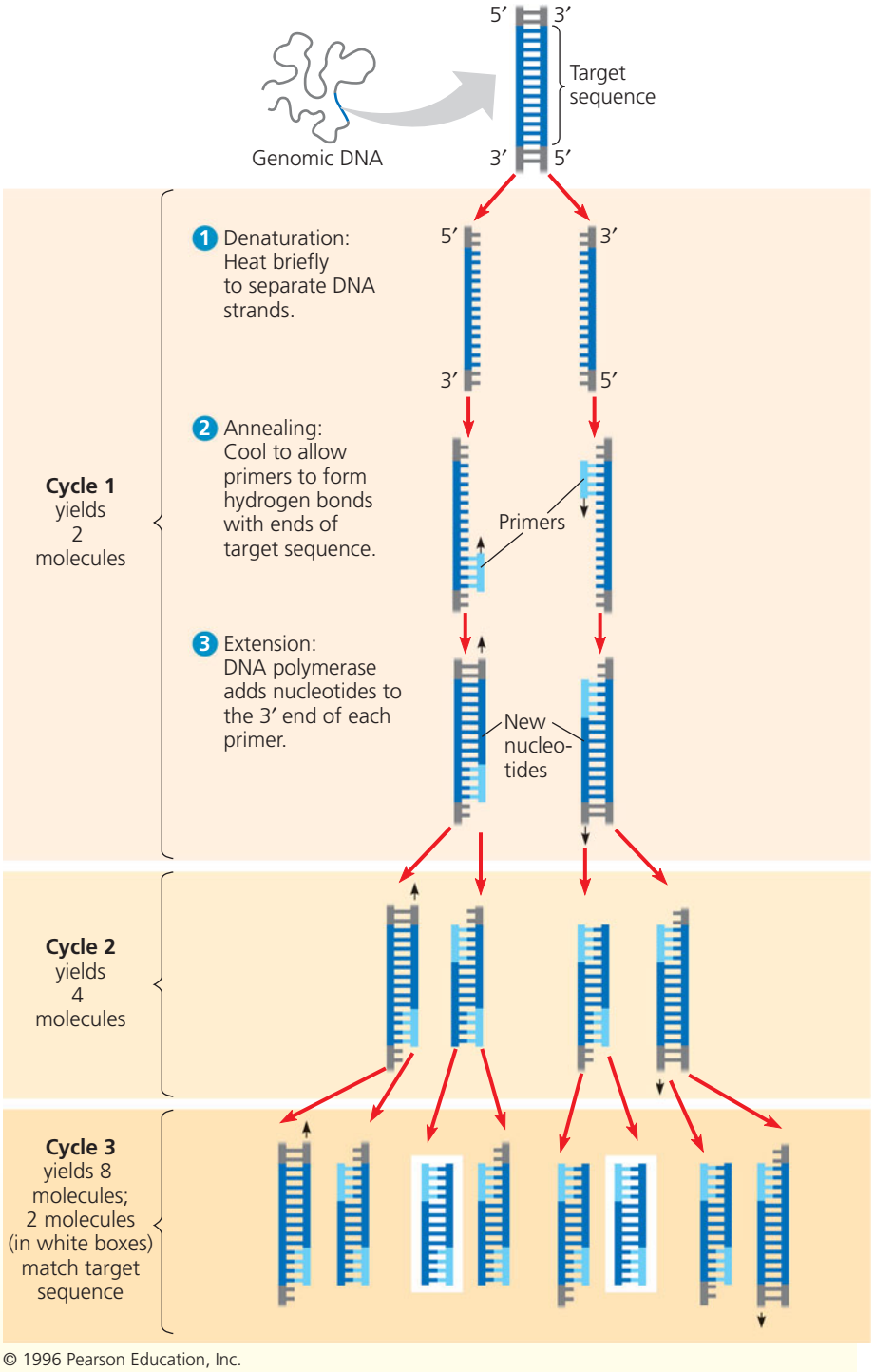
In the PCR procedure, a three-step cycle brings about a chain reaction that produces

▼ Figure 13.25 Research Method

The Polymerase Chain Reaction (PCR)

Application With PCR, any specific segment—the target sequence—within a DNA sample can be copied many times (amplified), completely *in vitro*.

Technique PCR requires double-stranded DNA containing the target sequence, a heat-resistant DNA polymerase, all four nucleotides, and two 15- to 20-nucleotide 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 3 cycles, two molecules match the target sequence exactly. After 30 more cycles, over 1 billion (10^9) molecules match the target sequence.

an exponentially growing population of identical DNA molecules. During each cycle, the reaction mixture is heated to denature (separate) the DNA strands and then cooled to allow annealing (hybridization) of short, single-stranded DNA primers complementary to sequences on opposite strands at each end of the target segment; finally, a DNA polymerase extends the primers in the 5′ → 3′ direction. If a standard DNA polymerase were used, the protein 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 unusually heat-stable DNA polymerase 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 bacterium to survive at temperatures up to 95°C.

PCR is speedy and very specific. Only minuscule amounts of DNA need be present in the starting material, and this DNA can be partially degraded, as long as a few molecules contain the complete target segment. The key to this high specificity is the primers, the sequences of which are chosen so they hybridize *only* with complementary sequences at opposite ends of the target segment. (For high specificity, the primers must be at least 15 or so nucleotides long.) By the end of the third cycle, one-fourth of the molecules are identical to the target segment, with both strands the appropriate length. 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 more cycles, about a billion copies of the target sequence are present!

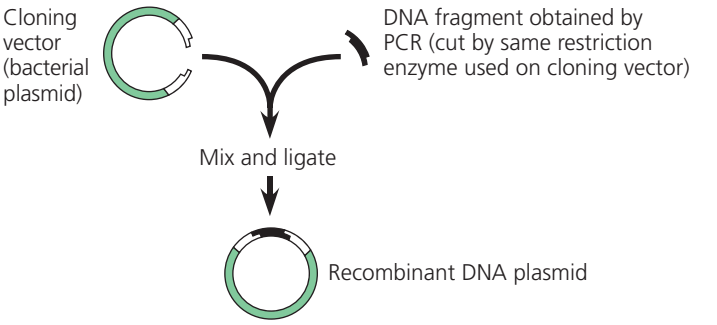
Despite its speed and specificity, PCR amplification alone cannot substitute for gene cloning in cells to make large amounts of a gene. This is because occasional errors during PCR replication limit the number of good copies and the length of DNA fragments that can be copied. Instead, PCR is used to 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 13.26)**. The resulting clones are sequenced so that clones 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; and cells infected with viruses that are difficult to detect, such as HIV (in the latter case, viral genes are amplified).

DNA Sequencing

Once a gene is cloned, researchers can exploit the principle of complementary base pairing to determine the gene's complete nucleotide sequence, a process called **DNA sequencing**. In the last ten years, "next-generation" sequencing techniques have been developed that are rapid and inexpensive. A single template strand is immobilized, and DNA polymerase and other reagents are added that allow so-called *sequencing by synthesis* of the complementary strand, one nucleotide at a time. A chemical trick enables electronic monitors to identify which of the four nucleotides is being added, allowing determination of the sequence. Technical advances continue to produce "third-generation" sequencing techniques, with each new technique being faster and less expensive than the previous. In Chapter 18, you'll learn more about how this rapid acceleration of sequencing technology has enhanced our study of genes and whole genomes.

In this chapter, you've 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 few chapters, we'll examine how the cell expresses the genetic information encoded in DNA. We'll also return to the subject of genetic engineering by exploring a few techniques for analyzing gene expression.



▲ Figure 13.26 Use of restriction enzymes and PCR in gene cloning. In a closer look at the process shown at the top of Figure 13.22, 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. Both the plasmid and the DNA fragments are cut with the same restriction enzyme, combined so the sticky ends can hybridize, ligated together, and introduced into bacterial cells.

CONCEPT CHECK 13.4

1. The restriction site for an enzyme called *Pvu*I is the following sequence:

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

   Staggered cuts are made between the T and C on each strand. What type of bonds are being cleaved?
2. **DRAW IT** One strand of a DNA molecule has the following sequence: 5′-CCTTGACGATCGTTACCG-3′. Draw the other strand. Will *Pvu*I cut this molecule? If so, draw the products.
3. Describe the role of complementary base pairing during cloning, DNA sequencing, and PCR.

For suggested answers, see Appendix A.

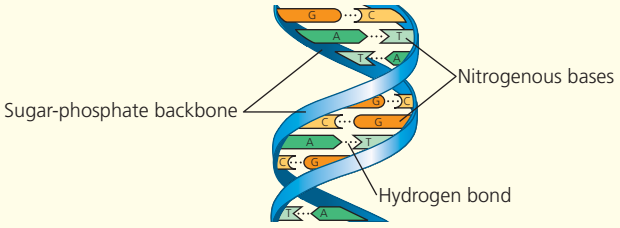
13 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 13.1

DNA is the genetic material (pp. 245–251)

- Experiments with bacteria and **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.

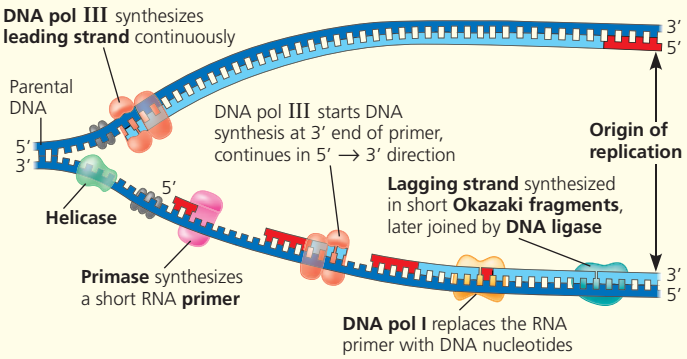


? *What does it mean when we say that the two DNA strands in the double helix are antiparallel? What would an end of the double helix look like if the strands were parallel?*

CONCEPT 13.2

Many proteins work together in DNA replication and repair (pp. 251–259)

- 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 general process by which **nucleases** cut out and replace damaged stretches of DNA.

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

CONCEPT 13.3

A chromosome consists of a DNA molecule packed together with proteins (pp. 259–261)

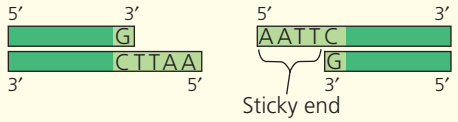
- The chromosome of most bacterial species is a circular DNA molecule with some associated proteins, making up the **nucleoid** of the cell. The **chromatin** making up a eukaryotic chromosome is composed of DNA, **histones**, and other proteins. The histones bind to each other and to the DNA to form **nucleosomes**, the most basic units of DNA packing. Additional coiling and folding lead ultimately to the highly condensed chromatin of the metaphase chromosome. In interphase cells, most chromatin is less compacted (**euchromatin**), but some remains highly condensed (**heterochromatin**). Euchromatin, but not heterochromatin, is generally accessible for transcription of genes.

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

CONCEPT 13.4

Understanding DNA structure and replication makes genetic engineering possible (pp. 261–265)

- **Gene cloning** (or DNA cloning) produces multiple copies of a gene (or DNA fragment) 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**.



- DNA fragments of different lengths can be separated and their lengths assessed by **gel electrophoresis**.
- The sticky ends on restriction fragments from one DNA source—such as a bacterial **plasmid** or other **cloning vector**—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.
- The **polymerase chain reaction (PCR)** can produce many copies of (amplify) a specific target segment of DNA *in vitro* for use as a DNA fragment for cloning. PCR uses primers that bracket the desired segment and requires a heat-resistant DNA polymerase.
- The rapid development of fast, inexpensive techniques for **DNA sequencing** is based on *sequencing by synthesis*: DNA polymerase is used to replicate a stretch of DNA from a single-stranded template, and the order in which nucleotides are added reveals the sequence.

? *Describe how the process of gene cloning results in a cell clone containing a recombinant plasmid.*

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

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.
   - **e.** bacteriophages injected DNA into bacteria.

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 growing strand.
   d. DNA ligase works only in the 3′ → 5′ direction.
   e. Polymerase can work on only one strand at a time.

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 + T
   d. A = C
   e. G = T

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.
   e. does not require a template strand.

5. In a nucleosome, the DNA is wrapped around
   a. polymerase molecules.
   b. ribosomes.
   c. histones.
   d. a thymine dimer.
   e. satellite DNA.

6. Which of the following sequences in double-stranded DNA is most likely to be recognized as a cutting site for a restriction enzyme?

a.	**b.**	**c.**	**d.**	**e.**
AAGG TTCC	AGTC TCAG	GGCC CCGG	ACCA TGGT	AAAA TTTT

Level 2: Application/Analysis

7. *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
   e. one low-density band

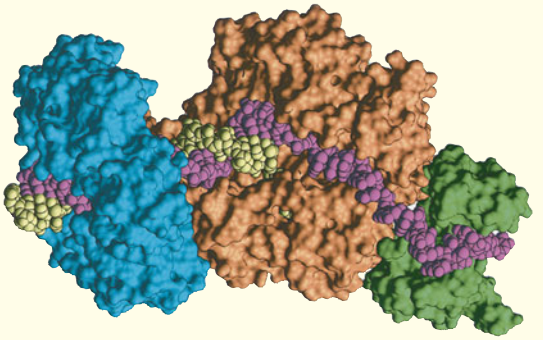
8. A biochemist isolates, purifies, and combines in a test tube a variety of molecules needed for DNA replication. When she adds some DNA to the mixture, replication occurs, but each DNA molecule consists of a normal strand paired with numerous segments of DNA a few hundred nucleotides long. What has she probably left out of the mixture?
   a. DNA polymerase
   b. DNA ligase
   c. nucleotides
   d. Okazaki fragments
   e. primase

9. 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. topoisomerase, primase, DNA polymerase
   c. topoisomerase, helicase, single-strand binding protein
   d. DNA ligase, replication fork proteins, adenylyl cyclase
   e. nuclease, topoisomerase, primase

10. **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 3.17)?

Level 3: Synthesis/Evaluation

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. Use what you've learned in this chapter to clarify this model by labeling each DNA strand and each protein and indicating the overall direction of DNA replication.

12. **FOCUS ON EVOLUTION**
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 of this ability? Explain.

13. **FOCUS ON 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.

For selected answers, see Appendix A.

14 Gene Expression: From Gene to Protein

KEY CONCEPTS

OVERVIEW

The Flow of Genetic Information

In 2006, a young albino deer seen frolicking with several brown deer in the mountains of eastern Germany elicited a public outcry **(Figure 14.1)**. A local hunting organization announced that the albino deer suffered from a "genetic disorder" and should be shot. Some argued that the deer should merely be prevented from mating with other deer to safeguard the population's gene pool. Others favored relocating the albino deer to a nature reserve because they worried that it might be more noticeable to predators if left in the wild. A German rock star even held a benefit concert to raise funds for the relocation. What led to the striking phenotype of this deer, the cause of this lively debate?

▼ **Figure 14.1** How does a single faulty gene result in the dramatic appearance of an albino deer?

Inherited traits are determined by genes, and the trait of albinism is caused by a recessive allele of a pigmentation gene (see Chapter 11). The information content of genes is in the form of specific sequences of nucleotides along strands of DNA, the genetic material. But how does this information determine an organism's traits? Put another way, what does a gene actually say? And how is its message translated by cells into a specific trait, such as brown hair, type A blood, or, in the case of an albino deer, a total lack of pigment? The albino deer has a faulty version of a key protein, an enzyme required for pigment synthesis, and this protein is faulty because the gene that codes for it contains incorrect information.

This example illustrates the main point of this chapter: 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 other words, 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. This chapter describes the flow of information from gene to protein in detail and explains how genetic mutations affect organisms through their proteins. Understanding the processes of gene expression, which are similar in all three domains of life, will allow us to revisit the concept of the gene in more detail at the end of the chapter.

CONCEPT 14.1

Genes specify proteins via 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 the Study of Metabolic Defects

In 1902, British physician Archibald Garrod was the first to suggest that genes dictate phenotypes through enzymes that catalyze specific chemical reactions in the cell. Garrod postulated that the symptoms of an inherited disease reflect a person's inability to make a particular enzyme. He later referred to such diseases as "inborn errors of metabolism." Garrod gave as one example the hereditary condition called alkaptonuria. In this disorder, the urine is black because it contains the chemical alkapton, which darkens upon exposure to air. Garrod reasoned that most people have an enzyme that metabolizes alkapton, whereas people with alkaptonuria have inherited an inability to make that enzyme.

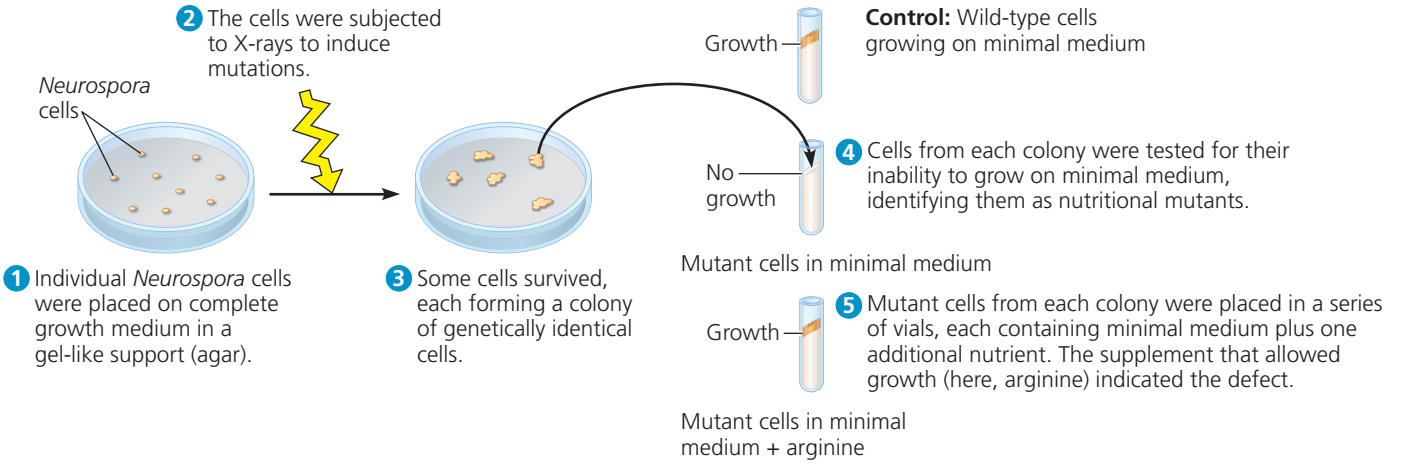
Garrod may have been the first to recognize that Mendel's principles of heredity apply to humans as well as peas. Garrod's realization was ahead of its time, but research several decades later supported his hypothesis that a gene dictates the production of a specific enzyme. Biochemists accumulated much evidence 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 6.1). Such metabolic pathways lead, for instance, to the synthesis of the pigments that give the brown deer in Figure 14.1 their fur color or fruit flies (*Drosophila*) their eye color (see Figure 12.3). In the 1930s, the American geneticist George Beadle and his French colleague Boris Ephrussi speculated that in *Drosophila*, each of the mutations 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 the bread mold *Neurospora crassa* to investigate the role of genes in this organism's metabolic pathways. Their experimental approach still plays a central role in genetic research today: They disabled genes one by one and looked for changes in each mutant's phenotype, thereby revealing the normal function of the gene.

Like Mendel and T. H. Morgan before them, Beadle and Tatum chose their experimental organism carefully. They elected to work with *Neurospora*, a haploid species. They realized that it would be easier to detect a disabled gene in a haploid species than in a diploid species like *Drosophila*. In a diploid species, two copies of each gene are present, and both would need to be disabled for an effect to be seen on the organism's phenotype. In *Neurospora*, though, disabling a single gene would allow them to see the consequences and thus to deduce what the function of the wild-type gene might be. (In other words, haploidy makes it easier to detect recessive mutations.)

What was known about metabolism in *Neurospora* also made it a good choice. Wild-type *Neurospora* has modest food requirements. It can grow in the laboratory on a simple solution of inorganic salts, glucose, and the vitamin biotin **(Figure 14.2)**. From this *minimal medium*, the mold cells use their metabolic



▲ **Figure 14.2 The experimental approach of Beadle and Tatum.** To obtain nutritional mutants, Beadle and Tatum exposed *Neurospora* cells to X-rays to induce mutations. They then screened mutants with new nutritional requirements, such as arginine, as shown here.

WHAT IF? *What do you predict would happen if the same mutant were tested in a tube of minimal medium plus the amino acid glycine?*

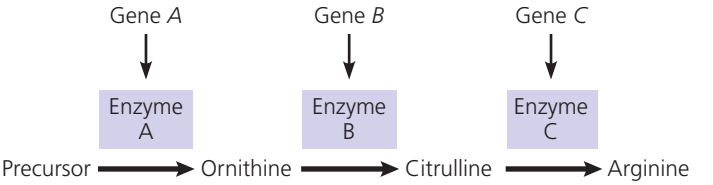
pathways to produce all the other molecules they need. Wild-type cells grow and divide repeatedly on this medium.

Neurospora cells can be plated individually on a petri dish containing minimal medium embedded in a gel-like substance called agar. Although single cells are microscopic, after many divisions the resulting daughter cells can be seen by eye on the surface of the agar. Thus their ability to grow and divide on a particular medium can easily be monitored. Any change in their ability to grow on minimal medium would be easy to recognize in the lab simply as a new food requirement for growth (cell division).

Because any mutant that could not synthesize an essential nutrient would be unable to grow on minimal medium, Beadle and Tatum placed single *Neurospora* cells on a *complete growth medium*, which consisted of minimal medium supplemented with all 20 amino acids and a few other nutrients. As diagrammed in Figure 14.2 they bombarded the cells with X-rays, shown in the 1920s to cause mutations. Each of the surviving cells formed a visible colony of genetically identical cells. Next, the researchers screened the surviving colonies for "nutritional mutants" that grew well on complete medium but not at all on minimal medium. Apparently, each nutritional mutant was unable to synthesize a certain essential molecule from the minimal ingredients. In the final step of this experimental approach, Beadle and Tatum took cells from each mutant colony growing on complete medium and distributed them to a number of different vials. Each vial contained minimal medium plus a single additional nutrient. The particular supplement that allowed growth indicated the nutrient that the mutant could not synthesize.

Thus, the researchers amassed a valuable collection of mutant strains of *Neurospora*, catalogued by their defects. The collection would prove useful for focusing in on particular metabolic pathways in which the individual steps were either known or strongly suspected. For example, a series of experiments on mutants requiring the amino acid arginine revealed that they could be grouped into classes, each corresponding to a particular step in the biochemical pathway for arginine synthesis. These results, along with the results of similar experiments with other nutritional mutants, suggested that each class was blocked at a different step in the pathway because mutants in that class lacked the enzyme that catalyzes the blocked step due to a faulty gene **(Figure 14.3)**.

Because each mutant was defective in a single gene, Beadle and Tatum saw that, taken together, the collected results provided strong support for a working hypothesis they had proposed earlier. The *one gene–one enzyme hypothesis*, as they dubbed it, states 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.



▲ Figure 14.3 The one gene–one protein hypothesis. Based on results from work in their lab on nutritional mutants, Beadle and Tatum proposed that the function of a specific gene is to dictate production of a specific enzyme that catalyzes a particular reaction. The model shown here for the arginine-synthesizing pathway illustrates their hypothesis.

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 (see Figure 3.21). Beadle and Tatum's idea was therefore restated as the *one gene–one polypeptide hypothesis*. Even this description is not entirely accurate, though. First, many eukaryotic genes can each 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'll encounter in this text—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 Chapter 3). 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.

It is customary to describe the flow of information from gene to protein in linguistic terms because both nucleic acids and proteins are polymers with specific sequences of monomers that convey information, much as specific sequences of letters communicate information in a language like English. In

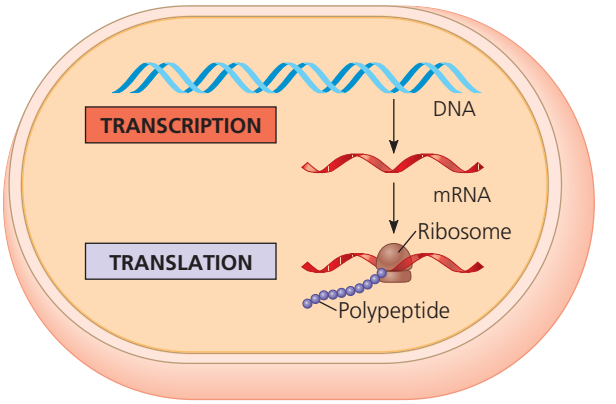
DNA or RNA, the monomers are the four types of nucleotides, which differ in their nitrogenous bases. Genes are typically hundreds or thousands of nucleotides long, each gene having a specific sequence of nucleotides. Each polypeptide of a protein also has monomers arranged in a particular linear order (the protein's primary structure), but its monomers are amino acids. Thus, nucleic acids and proteins contain information written in two different chemical languages. Getting from DNA to protein requires two major stages: transcription and translation.

Transcription is the synthesis 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, 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 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'll learn about some other types of RNA produced by transcription.)

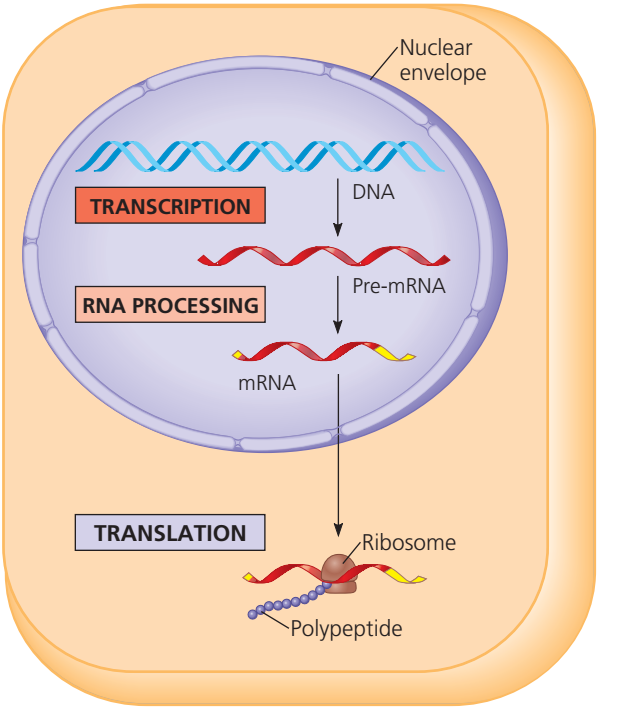
Translation is the synthesis of a polypeptide using the information in the mRNA. During this stage, there is a change in language: The cell must translate the nucleotide sequence of an mRNA molecule into the amino acid sequence of a polypeptide. The sites of translation are **ribosomes**, complex particles that facilitate the orderly linking of amino acids into polypeptide chains.

Transcription and translation occur in all organisms—those that lack a membrane-enclosed nucleus (bacteria and archaea) and those that have one (eukaryotes). Because most studies of transcription and translation have used bacteria and eukaryotic cells, they are our main focus in this chapter. While our understanding of transcription and translation in archaea lags behind, 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. Because bacteria do not have nuclei, their DNA is not separated by nuclear membranes from ribosomes and the other protein-synthesizing equipment **(Figure 14.4a)**. As you will see later, this lack of compartmentalization allows translation of an mRNA to begin while its transcription is still in progress. In a eukaryotic cell, by contrast, the nuclear envelope separates transcription from translation in space and time **(Figure 14.4b)**. Transcription occurs in the nucleus, and mRNA is then transported to the cytoplasm, where translation occurs. But before eukaryotic RNA transcripts from protein-coding genes can leave the nucleus, they are modified in various ways to produce the final, functional mRNA.



(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.

▲ Figure 14.4 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.

The transcription of a protein-coding eukaryotic gene results in *pre-mRNA*, and further processing yields the finished mRNA. The initial RNA transcript from any gene, including those specifying RNA that is not translated into protein, 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, shown here by arrows:

DNA → RNA → Protein

This concept was dubbed the *central dogma* by Francis Crick in 1956. How has the concept held up over time? In the 1970s, scientists were surprised to discover that some RNA molecules can act as templates for DNA synthesis (a process you'll read about in Chapter 17). However, these exceptions do not invalidate the idea that, in general, genetic information flows from DNA to RNA to protein. Now let's discuss how the instructions for assembling amino acids into a specific order are encoded in nucleic acids.

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, correspond to an amino acid?

Codons: Triplets of Nucleotides

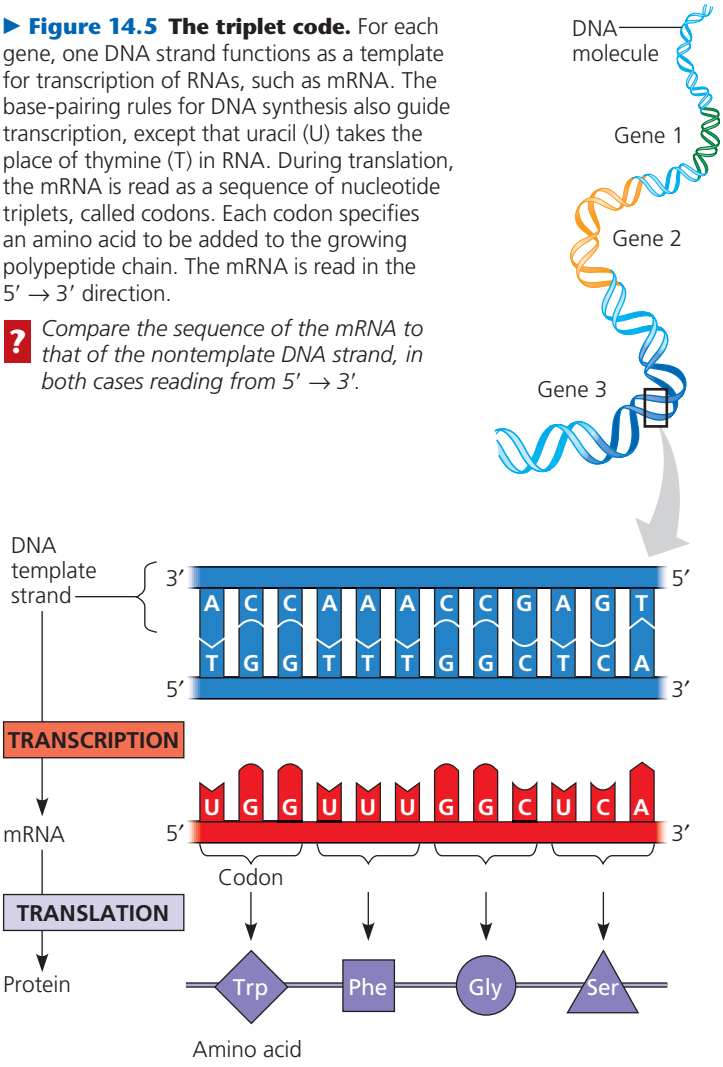
If each kind of nucleotide base were translated into an amino acid, only four amino acids could be specified, one per nucleotide base. Would a language of two-letter code words suffice? The two-nucleotide sequence AG, for example, could specify one amino acid, and GT could specify another. Since there are four possible nucleotide bases in each position, this would give us 16 (that is, 4^2) 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 14.5)**.

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 the gene is transcribed. For other genes on the same DNA molecule, however, the opposite strand may be the one that always functions as the template.

▶ Figure 14.5 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 uracil (U) takes the place of thymine (T) in RNA. 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.

? *Compare the sequence of the mRNA to that of the nontemplate DNA strand, in both cases reading from 5′ → 3′.*



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 14.5). 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 13.7.) In the example in Figure 14.5, the nucleotide triplet ACC along the DNA (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). 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 that they have T instead of U. (For this reason, the nontemplate DNA strand is sometimes called the "coding strand.")

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′ 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, it takes 300 nucleotides along an mRNA strand to code for the amino acids in a polypeptide that is 100 amino acids long.

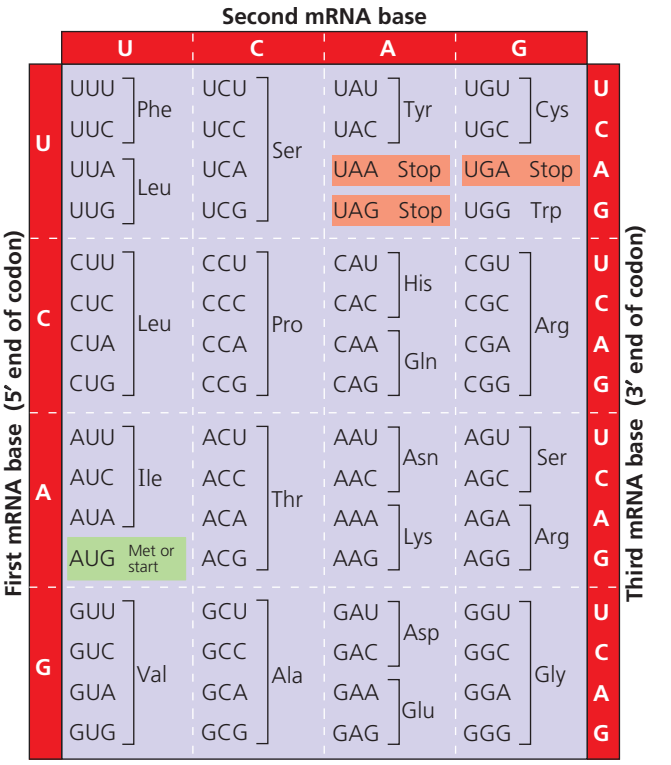
Cracking the Code

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 RNA codons. The first codon was deciphered in 1961 by Marshall Nirenberg, of the National Institutes of Health, and his colleagues. Nirenberg synthesized an artificial mRNA by linking identical RNA nucleotides containing uracil as their base. No matter where this message started or stopped, it could contain only one codon in repetition: UUU. Nirenberg added this "poly-U" to a test-tube mixture containing amino acids, ribosomes, and the other components required for protein synthesis. His artificial system translated the poly-U into a polypeptide containing many units of the amino acid phenylalanine (Phe), strung together as a long 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 determined in the same way.

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 14.6** shows, 61 of the 64 triplets code for amino acids. The three codons that do not designate amino acids are "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) 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 begin with methionine when they are synthesized. However, an enzyme may subsequently remove this starter amino acid from the chain.)

Notice in Figure 14.6 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'll consider a possible benefit of this redundancy later in the chapter.

Our ability to extract the intended message from a written language depends on reading the symbols in the correct



▲ Figure 14.6 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. (Practice using this table by finding the codons in Figure 14.5.) The codon AUG not only stands for the amino acid methionine (Met) 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. See Figure 3.17 for a list of the full names of all the amino acids.

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 14.5, 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.

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 RNA codon CCG, for instance, is translated as the amino acid proline in all organisms whose 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

(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) Pig expressing a jellyfish gene. Researchers injected the gene for a fluorescent protein into fertilized pig eggs. One of the eggs developed into this fluorescent pig.

▲ Figure 14.7 Expression of genes from different species. Because diverse forms of life share a common genetic code, one species can be programmed to produce proteins characteristic of a second species by introducing DNA from the second species into the first.

striking results, as shown in **Figure 14.7**. 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 genetic engineering (see Chapter 13).

Despite a small number of exceptions in which a few codons differ from the standard ones, 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 that bonds all life on Earth.

CONCEPT CHECK 14.1

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 the disorder. Today, would this disorder be called dominant or recessive? (See Concept 11.4.)
2. What polypeptide product would you 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′. Imagine that the nontemplate sequence was transcribed instead of the template sequence. Draw the mRNA sequence and translate it using Figure 14.6. (Be sure to pay attention to the 5′ and 3′ ends.) Predict how well the protein synthesized from the nontemplate strand would function, if at all.

For suggested answers, see Appendix A.

CONCEPT 14.2

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 more 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 14.8)**. Like the DNA polymerases that function in DNA replication, RNA polymerases can assemble a polynucleotide only in its 5′ → 3′ direction. Unlike DNA polymerases, however, RNA polymerases are able to start a chain from scratch; they don't need a 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 is different in eukaryotes; we'll describe it 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 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 protein synthesis, such as ribosomal RNA. In contrast, eukaryotes have at least three types of RNA polymerase in their nuclei; the one used for mRNA synthesis is called RNA polymerase II. In the discussion of transcription 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

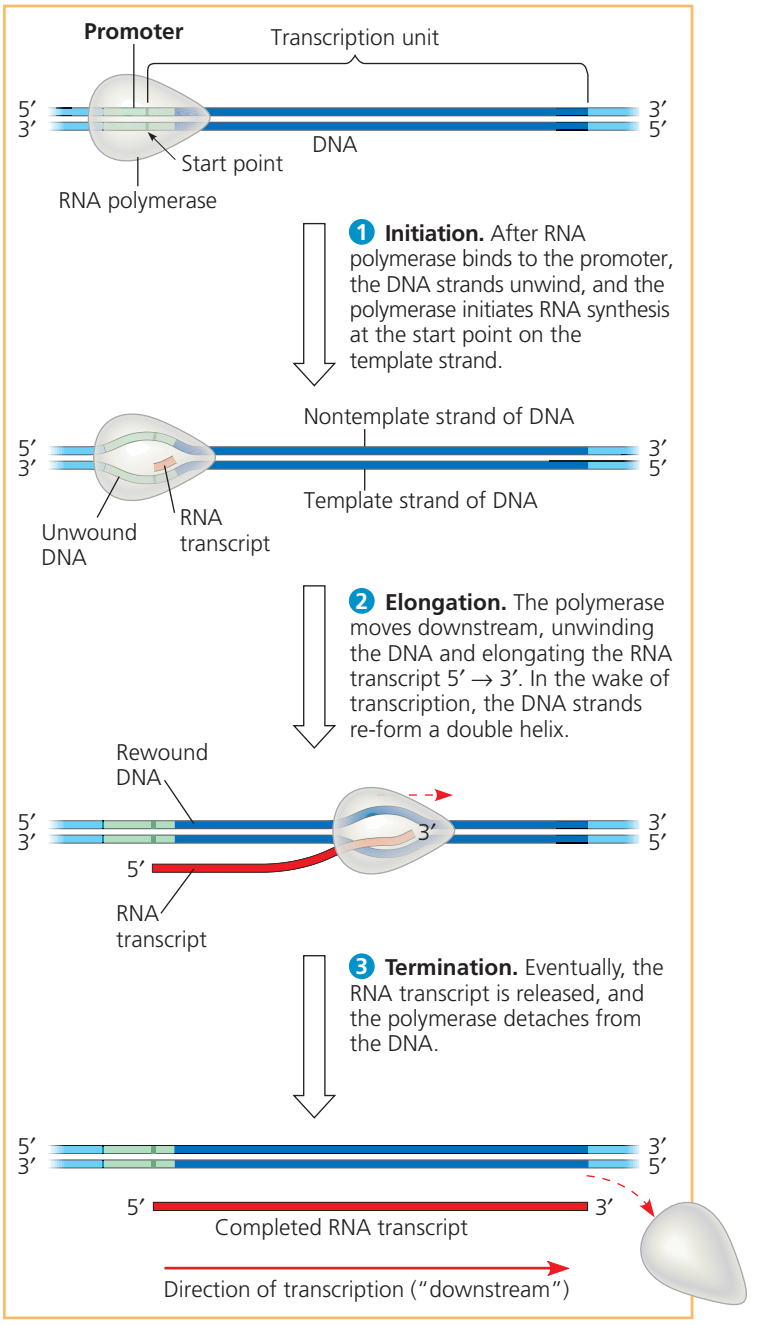
The three stages of transcription, as shown in Figure 14.8 and described next, are initiation, elongation, and termination of the RNA chain. Study Figure 14.8 to familiarize yourself with the stages and the terms used to describe them.

RNA Polymerase Binding and Initiation of Transcription

The promoter of a gene includes within it the transcription **start point** (the nucleotide where RNA synthesis actually begins) and typically extends several dozen or more nucleotide

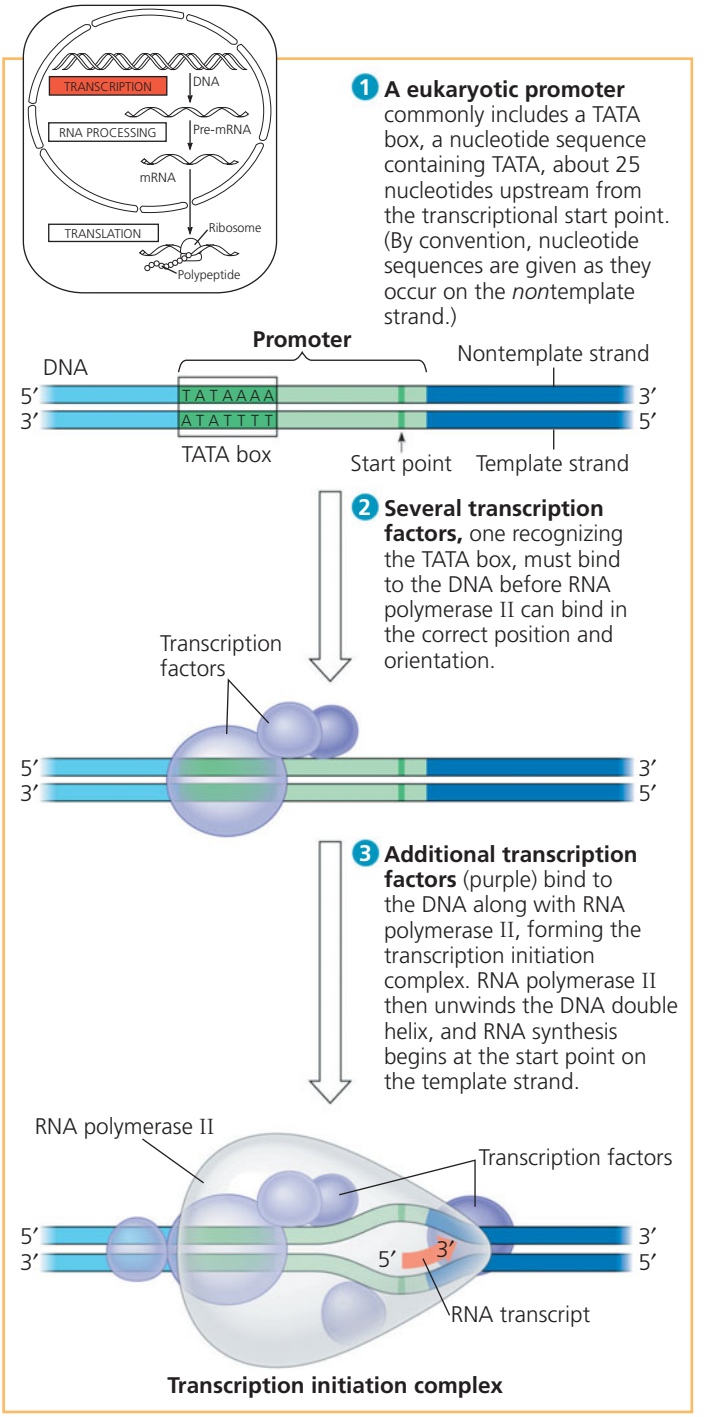
pairs upstream from the start point. RNA polymerase binds in a precise location and orientation on the promoter, thereby determining where transcription starts and which of the two strands of the DNA helix is used as the template.

Certain sections of a promoter are especially important for binding RNA polymerase. In bacteria, the RNA polymerase itself specifically recognizes and binds to the promoter. In eukaryotes, a collection of proteins called **transcription factors** mediate 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 14.9** shows the role of transcription factors and a crucial



▲ Figure 14.8 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 13.17.*



▲ Figure 14.9 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.*

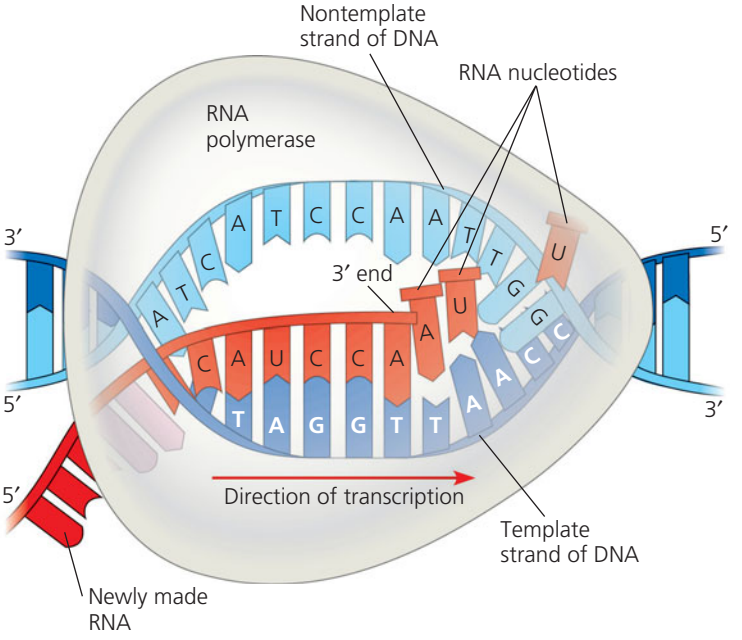
DNA sequence in the promoter—the so-called **TATA box**—in forming the initiation complex at a eukaryotic promoter.

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 in the correct orientation, the enzyme unwinds the two DNA strands and starts transcribing the template strand.

Elongation of the RNA Strand

As RNA polymerase moves along the DNA, it continues to untwist the double helix, exposing about 10–20 DNA nucleotides at a time for pairing with RNA nucleotides **(Figure 14.10)**. The enzyme adds nucleotides to the 3′ end of the growing RNA molecule as it continues along the double helix. In the wake of this advancing wave of RNA synthesis, the new RNA molecule 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 14.23). 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.



▲ Figure 14.10 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.

Termination of Transcription

The mechanism of termination differs between bacteria and eukaryotes. 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 transcribes a sequence on the DNA called the polyadenylation signal sequence, which codes for a polyadenylation signal (AAUAAA) in the pre-mRNA. Then, at a point about 10–35 nucleotides downstream from the AAUAAA signal, proteins associated with the growing RNA transcript cut it free from the polymerase, releasing the pre-mRNA. The pre-mRNA then undergoes processing, the topic of the next section.

CONCEPT CHECK 14.2

1. What is a promoter, and 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 14.9.)

For suggested answers, see Appendix A.

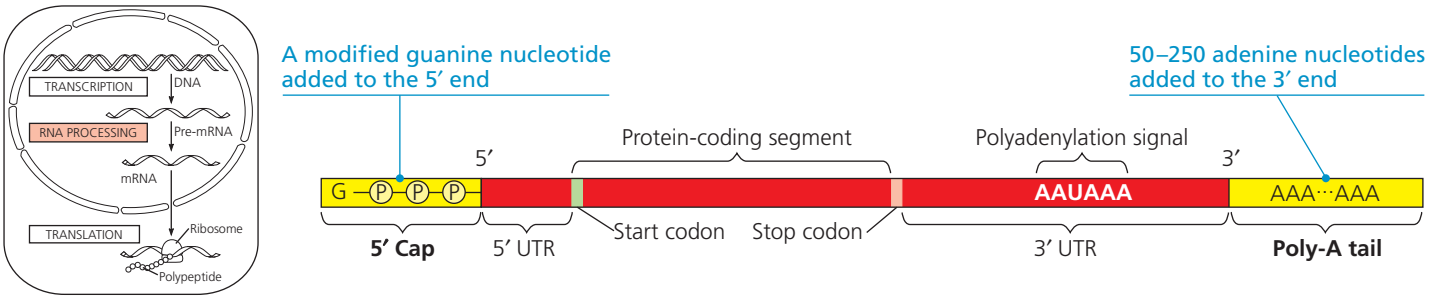
CONCEPT 14.3

Eukaryotic cells modify RNA after transcription

Enzymes in the eukaryotic nucleus modify pre-mRNA in specific ways before the genetic messages are 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 14.11)**. The 5′ end is synthesized first; it 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. The 3′ end of the pre-mRNA molecule is also modified before the mRNA exits the nucleus. Recall that the pre-mRNA is released soon after the polyadenylation signal, AAUAAA, is transcribed. At the 3′ end, an enzyme 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



▲ Figure 14.11 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, facilitate 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).

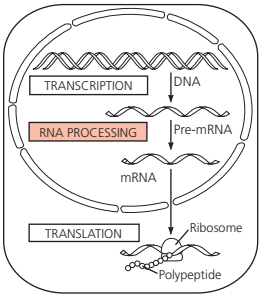
5′ end of the mRNA once the mRNA reaches the cytoplasm. Figure 14.11 shows a diagram of a eukaryotic mRNA molecule with cap and tail. The figure also shows the untranslated regions (UTRs) at the 5′ and 3′ ends of the mRNA (referred to as the 5′ UTR and 3′ UTR). The UTRs are parts of the mRNA that will not be translated into protein, but they have other functions, such as ribosome binding.

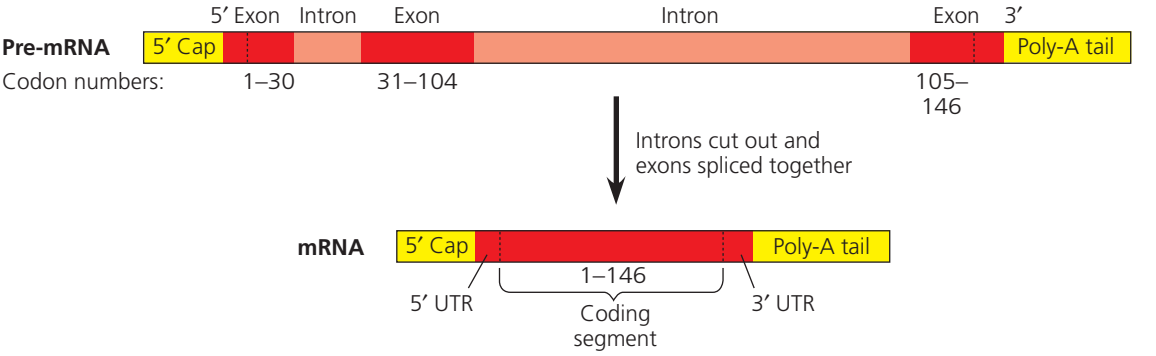
Split Genes and RNA Splicing

A remarkable stage of RNA processing in the eukaryotic nucleus is the removal of large portions of the RNA molecule that is initially synthesized—a cut-and-paste job called **RNA splicing**, similar to editing a video **(Figure 14.12)**. 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 means that 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 *in*tervening sequences, or **introns**. The other regions are called **exons**, because they are eventually *ex*pressed, 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 find it helpful to think of exons as sequences of RNA that *exit* the nucleus.) The terms *intron* and *exon* are used for both RNA sequences and the DNA sequences that encode 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. The introns are cut out from the molecule and the exons joined together, forming an mRNA molecule with a continuous coding sequence. This is the process of RNA splicing.

One important consequence of the presence of introns in genes is that a single gene can encode more than one kind of





▲ Figure 14.12 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 leave the nucleus as mRNA. (The 5′ UTR and 3′ UTR are parts of exons because they are included in the mRNA; however, they do not code for protein.) During RNA processing, the introns are cut out and the exons spliced together. In many genes, the introns are much larger than the exons.

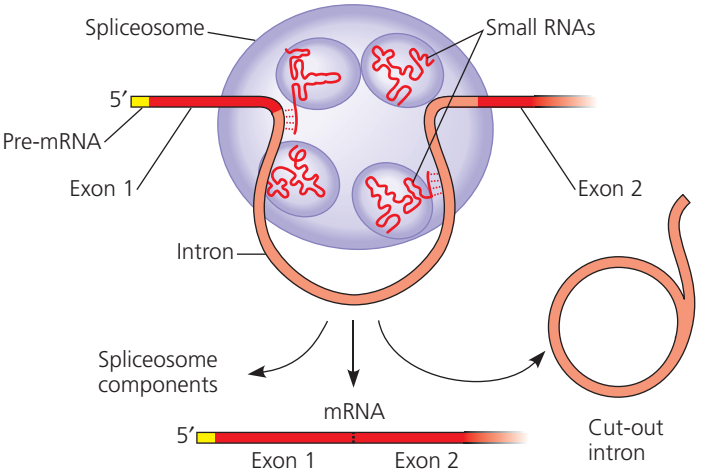
polypeptide. Many genes are known to give rise to two or more different polypeptides, depending on which segments are treated as exons during RNA processing; this is called **alternative RNA splicing** (see Figure 15.12). Because of alternative splicing, the number of different protein products an organism produces can be much greater than its number of genes.

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 the intron, including key sequences at each end **(Figure 14.13)**. The intron is then released (and rapidly degraded), and the spliceosome joins together the two exons that flanked the intron. It turns out that the small RNAs in the spliceosome catalyze these processes, as well as participating in spliceosome assembly and splice site recognition.

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 excision! For example, in the ciliate protist *Tetrahymena*, self-splicing occurs in the production of ribosomal RNA (rRNA), a component of the organism's ribosomes. The pre-rRNA actually removes its own introns. The discovery of ribozymes rendered obsolete 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 with a complementary region elsewhere in the same molecule, which 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 may 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'll 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.



▲ Figure 14.13 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, the spliceosome catalyzes cutting of the pre-mRNA and the splicing together of the exons, releasing the intron for rapid degradation.

CONCEPT CHECK 14.3

1. How can human cells make 75,000–100,000 different proteins, given that there are about 20,000 human genes?
2. How is RNA splicing similar to editing a video? 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 cap from mRNAs?

For suggested answers, see Appendix A.

CONCEPT 14.4

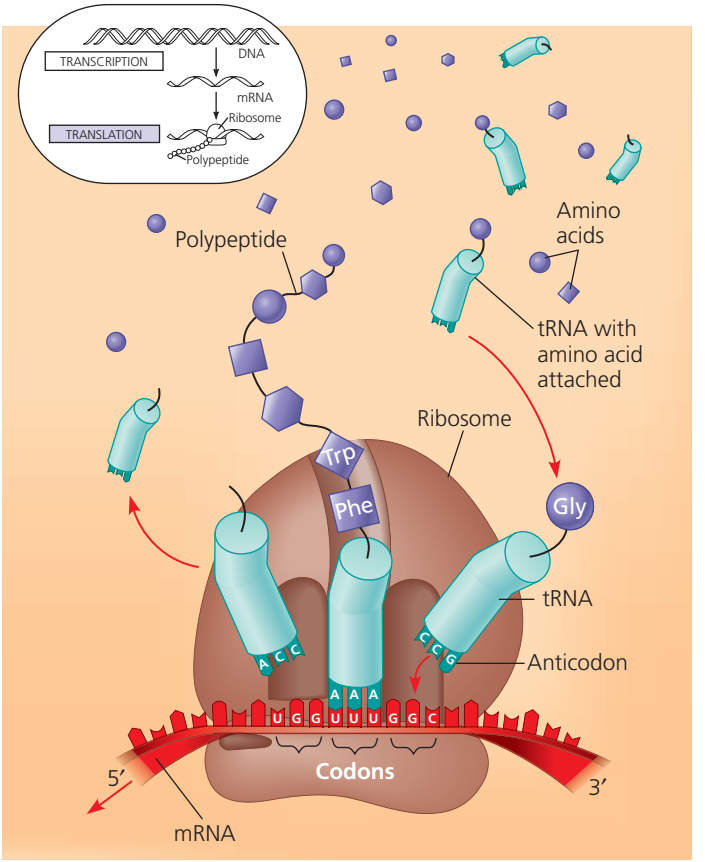
Translation is the RNA-directed synthesis of a polypeptide: *a closer look*

We will now examine in greater detail how genetic information flows from mRNA to protein—the process of translation. As we did for transcription, we'll concentrate on the basic steps of translation that occur in both bacteria and eukaryotes, while pointing out key differences.

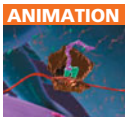
Molecular Components of Translation

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 **transfer RNA (tRNA)**. The function of tRNA is to transfer amino acids 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 tRNA to the growing end of a polypeptide chain **(Figure 14.14)**.

Translation is simple in principle but complex in its biochemistry and mechanics, especially in the eukaryotic cell. In dissecting translation, we'll concentrate on the slightly less complicated version of the process that occurs in bacteria. We'll begin by looking at the major players in this cellular process and then see how they act together in making a polypeptide.



▲ Figure 14.14 Translation: the basic concept. As a molecule of mRNA is moved through a ribosome, codons are translated into amino acids, one by one. The interpreters are tRNA molecules, each type with a specific nucleotide triplet called an 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 after the anticodon hydrogen-bonds to a complementary codon on the mRNA. The figures that follow show some of the details of translation in a bacterial cell.

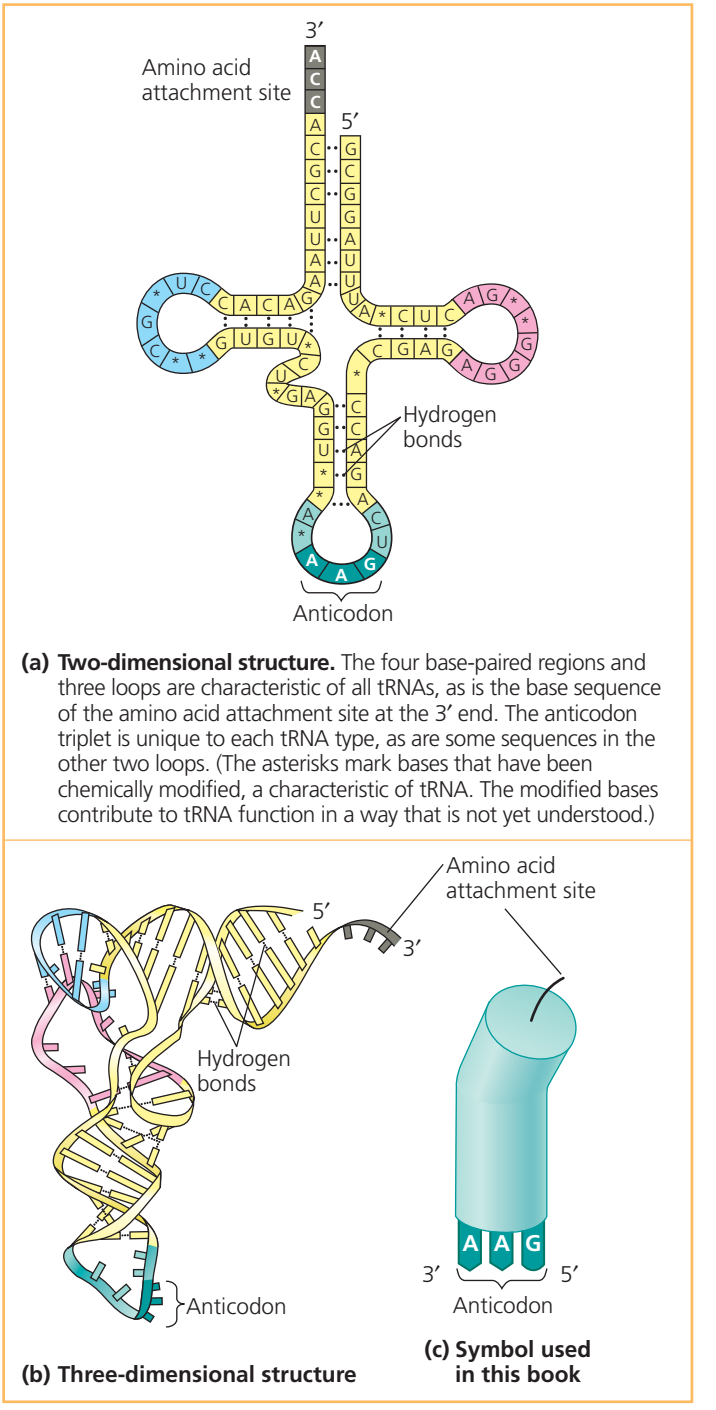


BioFlix Visit the Study Area in **MasteringBiology** for the BioFlix® 3-D Animation on Protein Synthesis.

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 can translate a particular mRNA codon into a given amino acid. This is possible because a tRNA bears a specific amino acid at one end, 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). Because of the presence of complementary stretches of nucleotide bases that can hydrogen-bond to each other, this single strand can fold back on itself and form a molecule with a three-dimensional structure. Flattened into one plane to clarify this base pairing, a tRNA molecule looks like a cloverleaf **(Figure 14.15a)**. The tRNA actually twists and folds into a compact three-dimensional structure that is roughly L-shaped **(Figure 14.15b)**. The loop extending from one end of the L includes the **anticodon**, the particular nucleotide triplet that base-pairs to a specific mRNA codon. From the other end of the L-shaped tRNA molecule protrudes its 3' end, which is the attachment site for an amino acid. Thus, the structure of a tRNA molecule fits its function.



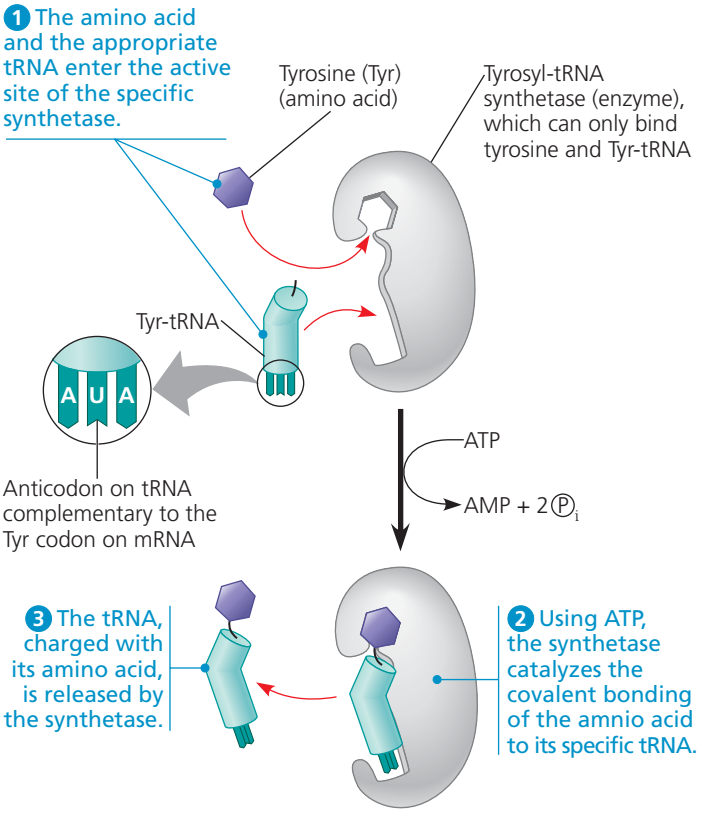
▲ Figure 14.15 The structure of transfer RNA (tRNA). Anticodons are conventionally written 3′ → 5′ to align properly with codons written 5′ → 3′ (see Figure 14.14). For base pairing, RNA strands must be antiparallel, like DNA. For example, anticodon 3′-AAG-5′ pairs with mRNA codon 5′-UUC-3′.

As an example, consider the mRNA codon GGC, which is translated as the amino acid glycine. The tRNA that base-pairs with this codon by hydrogen bonding has CCG as its anticodon and carries glycine at its other end (see the incoming tRNA approaching the ribosome in Figure 14.14). As an mRNA molecule is moved through a ribosome, glycine will be added to the polypeptide chain whenever the codon GGC is presented for translation. Codon by codon, the genetic message is translated as tRNAs deposit amino acids in the order prescribed, and the ribosome joins the amino acids into a chain. The tRNA molecule is a translator in the sense that 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 from the nucleus to the cytoplasm, where translation occurs. 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 called **aminoacyl-tRNA synthetases** **(Figure 14.16)**. 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 are instrumental in ensuring the specific fit.) There are 20 different synthetases, one for each amino acid; each synthetase is able to bind to all the different tRNAs that match the codons 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 variety existed for each mRNA codon specifying an amino acid, there would be 61 tRNAs (see Figure 14.6). In fact, there are only about 45, signifying that some tRNAs must be able to bind to more than one codon. Such versatility 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

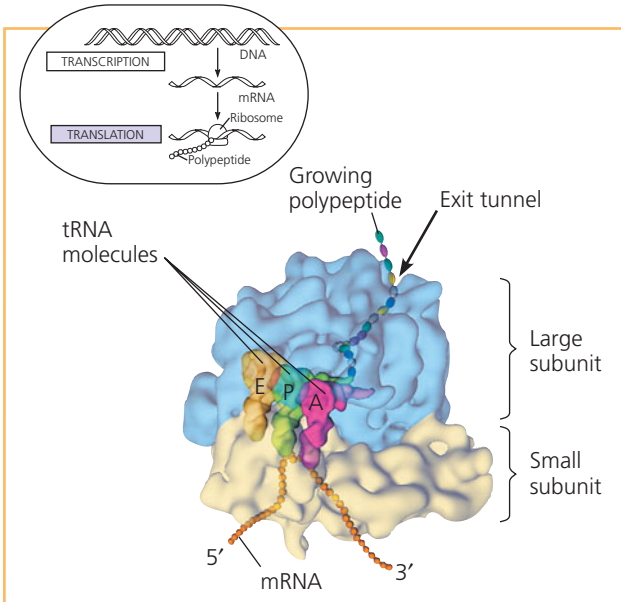


▲ **Figure 14.16 Aminoacyl-tRNA synthetases provide specificity in joining amino acids to their tRNAs.** Linkage of a tRNA to its amino acid is an endergonic process that occurs at the expense of ATP (which loses two phosphate groups, becoming AMP). Shown as an example is the joining of the amino acid tyrosine (Tyr) to the appropriate tRNA by the aminoacyl-tRNA synthetase specific for tyrosine.

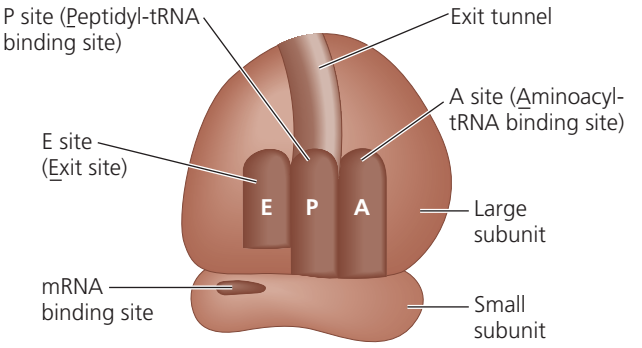
this codon position is called **wobble**. Wobble explains why the synonymous codons for a given amino acid most often differ in their third nucleotide base, but not in the other bases. For example, a tRNA with the anticodon 3′-UCU-5′ can base-pair with either the mRNA codon 5′-AGA-3′ or 5′-AGG-3′, both of which code for arginine (see Figure 14.6).

Ribosomes

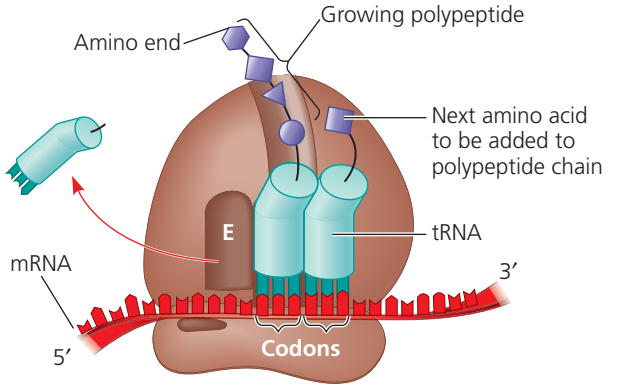
Ribosomes facilitate the specific coupling of tRNA anticodons with mRNA codons during protein synthesis. A ribosome consists of a large subunit and a small subunit, each made up of proteins and one or more **ribosomal RNAs (rRNAs)** **(Figure 14.17)**. In eukaryotes, the subunits are made in the nucleolus. Ribosomal RNA genes are transcribed, and the RNA is processed and assembled with proteins imported from the cytoplasm. The resulting ribosomal subunits are then exported via nuclear pores to the cytoplasm. In both bacteria and eukaryotes, large and small subunits join to form a functional ribosome only when they attach to an mRNA molecule. About one-third of the mass of a ribosome is made up of proteins; the rest consists of rRNAs, either three molecules (in bacteria) or four



(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 a binding 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.

▲ Figure 14.17 The anatomy of a functioning ribosome.

(in eukaryotes). Because most cells contain thousands of ribosomes, rRNA is the most abundant type of cellular RNA.

Although the ribosomes of bacteria and eukaryotes are very similar in structure and function, eukaryotic ribosomes are slightly larger, and they differ somewhat from bacterial ribosomes in their molecular composition. The differences are medically significant. Certain antibiotic drugs can inactivate bacterial ribosomes without inhibiting the ability of eukaryotic ribosomes to make proteins. These drugs, including tetracycline and streptomycin, are used to combat bacterial infections.

The structure of a ribosome reflects its function of bringing mRNA together with tRNAs carrying amino acids. In addition to a binding site for mRNA, each ribosome has three binding sites for tRNA, as described in Figure 14.17. 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 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.

There is strong evidence supporting the hypothesis that rRNA, not protein, is 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 ribosomal subunits; it also acts as the catalyst of peptide bond formation. Thus, a ribosome can be regarded as one colossal ribozyme!

Building a Polypeptide

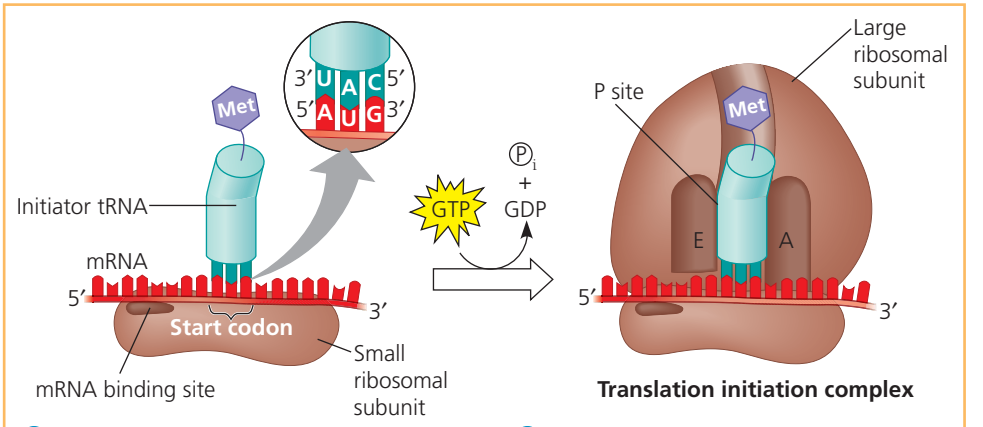
We can divide translation, the synthesis of a polypeptide chain, into three stages (analogous to those of transcription): initiation, elongation, and termination. All three stages require protein "factors" that aid in the translation process. For certain aspects of chain initiation and elongation, energy is also required. It is provided by the hydrolysis of guanosine triphosphate (GTP), a molecule closely related to ATP.

Ribosome Association and Initiation of Translation

The initiation stage of translation brings together 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 mRNA and a specific initiator tRNA, which carries the amino acid methionine. In bacteria, the small subunit can bind these two in either order; it binds the mRNA at a specific RNA sequence, just upstream of the start codon, AUG. In eukaryotes, the small subunit, with the initiator tRNA already

bound, binds to the 5′ cap of the mRNA and then moves, or *scans*, downstream along the mRNA until it reaches the start codon; the initiator tRNA then hydrogen-bonds to the AUG start codon **(Figure 14.18)**. In either case, the start codon signals the start of translation; this is important because it establishes the codon reading frame for the mRNA. In the **Scientific Skills Exercise**, you can work with DNA sequences encoding the ribosomal binding sites on the mRNAs of a group of *E. coli* genes.

The union of mRNA, initiator tRNA, and a small ribosomal subunit is followed by the attachment of a large ribosomal subunit, completing the *translation initiation complex*. 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 3.18).



▲ **Figure 14.18 The 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 the participation of several proteins called *elongation factors* and occurs in a three-step cycle described in **Figure 14.19**. 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 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 important 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 to the chain until the polypeptide is completed.

Termination of Translation

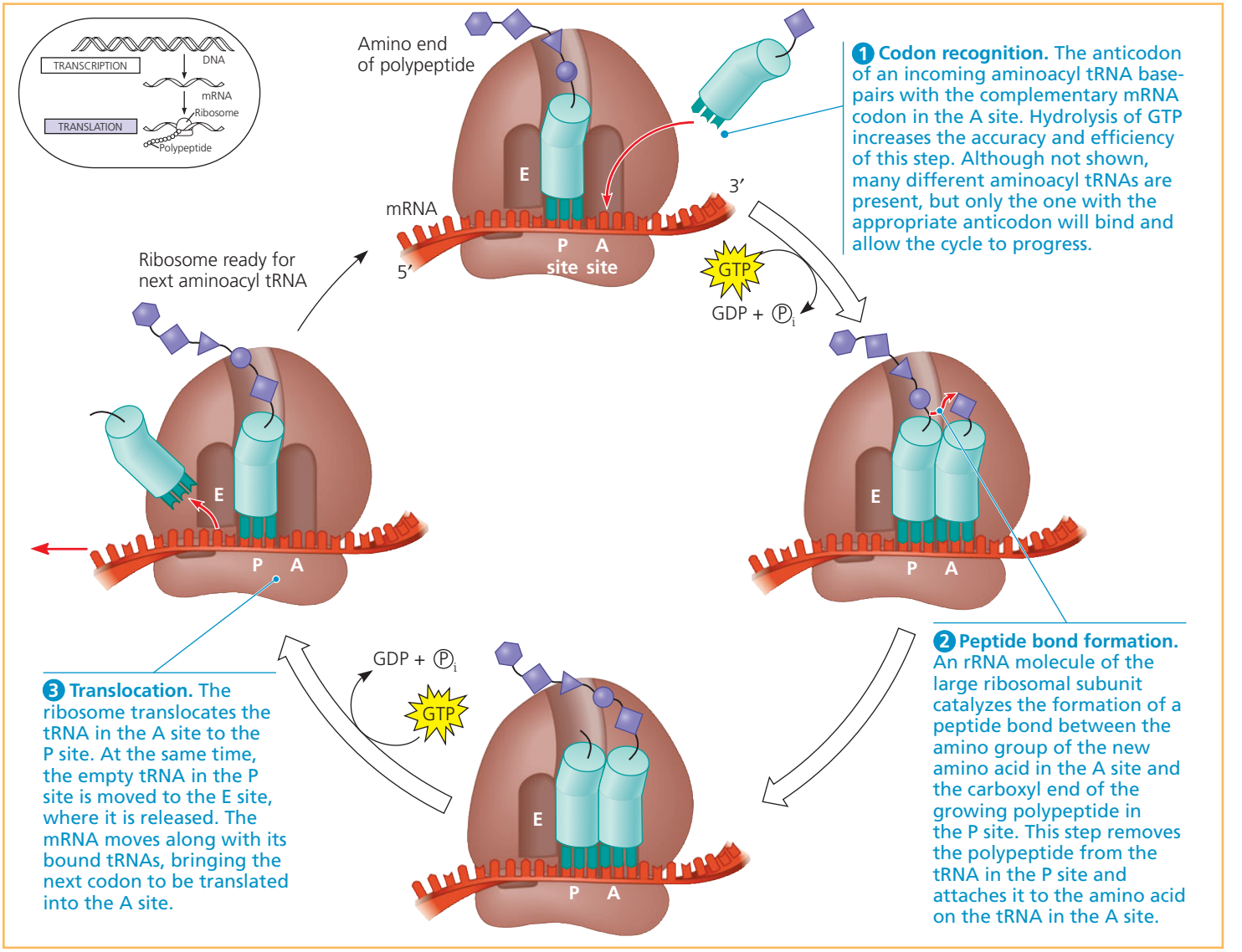
The final stage of translation is termination **(Figure 14.20)**. Elongation continues until a stop codon in the mRNA reaches the A site of the ribosome. The nucleotide base triplets UAG, UAA, and UGA do not code for amino acids but instead act as signals to stop translation. A *release factor*, a protein 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. (There are plenty of water molecules available in the aqueous cellular environment.) This reaction breaks (hydrolyzes) 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

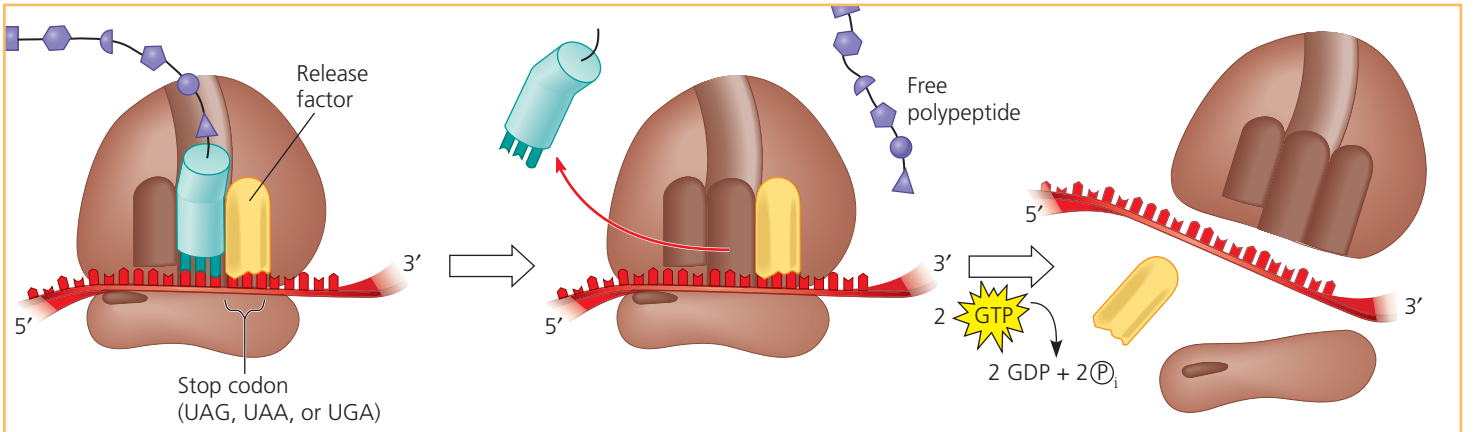
The process of translation is often not sufficient to make a functional protein. In this section, you'll learn about modifications that polypeptide chains undergo after the translation process as well as some of the mechanisms used to target completed proteins to specific sites in the cell.

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 3.21). Thus, a gene determines primary structure, and primary structure in turn determines shape. In many cases, a chaperone protein helps the polypeptide fold correctly.



▲ **Figure 14.19** **The elongation cycle of translation.** The hydrolysis of GTP plays an important role in the elongation process. Not shown are the proteins called elongation factors.



▲ **Figure 14.20** **The termination of translation.** Like elongation, termination requires GTP hydrolysis as well as additional protein factors, which are not shown here.

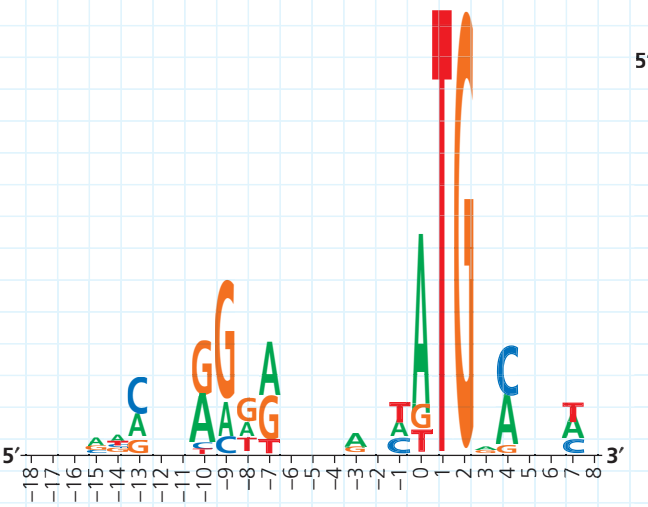
Interpreting a Sequence Logo

How Can a Sequence Logo Be Used to Identify Ribosome-Binding Sites? 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 and analyzed using computer software. The aim was to identify similar base sequences—at the appropriate location in each gene—that might be 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.

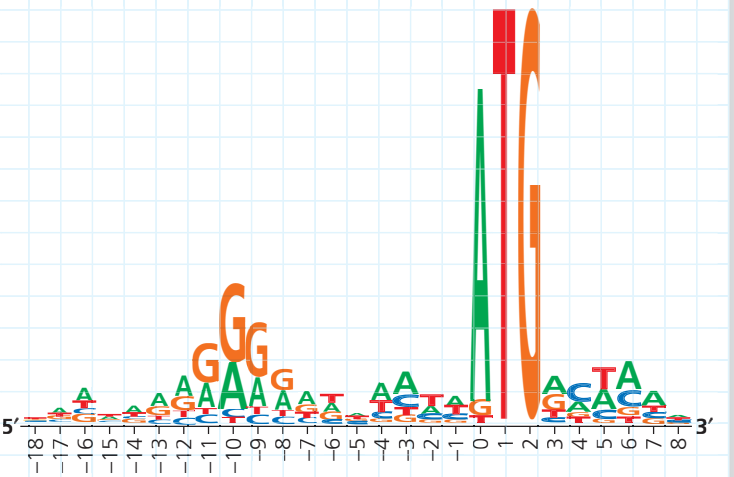
Data from the Experiment To show how sequence logos are made, the potential ribosome-binding regions from 10 of the *E. coli* genes are shown in a sequence alignment below, followed by the sequence logo derived from the aligned sequences. Note that the DNA shown is the nontemplate (coding) strand, which is how DNA sequences are typically given.

	−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
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



Interpret the Data

1. In the sequence logo (bottom, left), 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, all 10 sequences have a G; the probability of finding a G there in a new sequence is very high, as is the stack. For short stacks, the bases all have about the same frequency, and so it's hard to predict a base at those positions. (a) 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 in the 10 sequences? Answer only for the two left-most of the 12 positions.
3. In the actual experiment, the researchers used 149 sequences to build their sequence logo (shown below). There is a stack at each position, even if short, because the sequence logo includes more data. (a) Which three positions in the sequence logo have the most predictable bases? Name the most frequent base at each. (b) Which 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?

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).

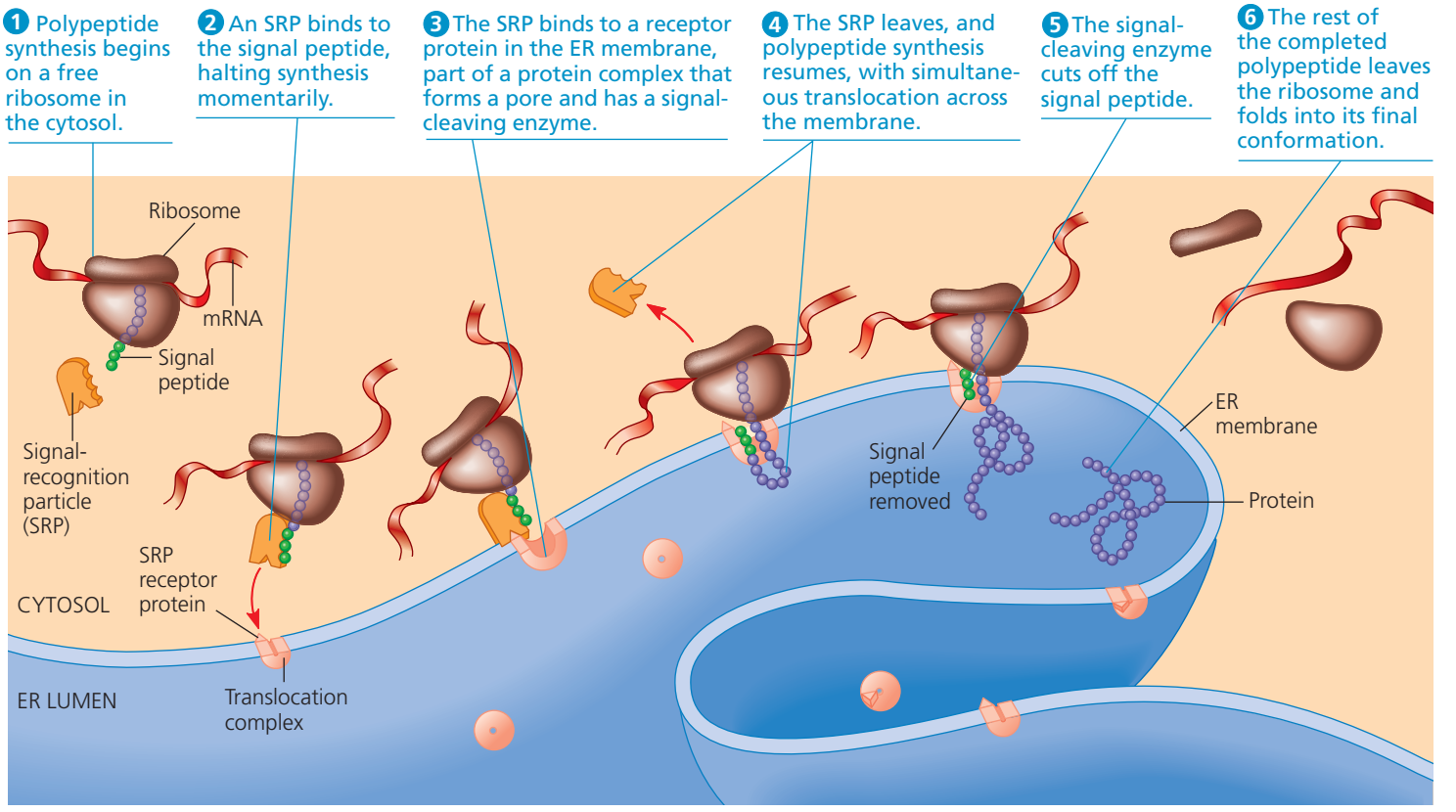
MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

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. For example, the protein insulin is first synthesized as a single polypeptide chain but becomes active only after an enzyme cuts out a central part of the chain, leaving a protein made up of two shorter polypeptide chains connected by disulfide bridges. In other cases, two or more polypeptides that are synthesized separately may come together, becoming the subunits of a protein that has quaternary structure. A familiar example is hemoglobin (see Figure 3.21).

Targeting Polypeptides to Specific Locations

In electron micrographs of eukaryotic cells active in protein synthesis, two populations of ribosomes (and polyribosomes) are evident: free and bound (see Figure 4.9). 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 (the nuclear envelope, ER, Golgi apparatus, lysosomes, vacuoles, and plasma membrane) 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 and bound.

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 14.21)**. 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 functions as an escort that brings 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 usually 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 (as in Figure 14.21). Alternatively, if the polypeptide is to be a membrane protein,



▲ **Figure 14.21** **The signal mechanism for targeting proteins to the ER.**

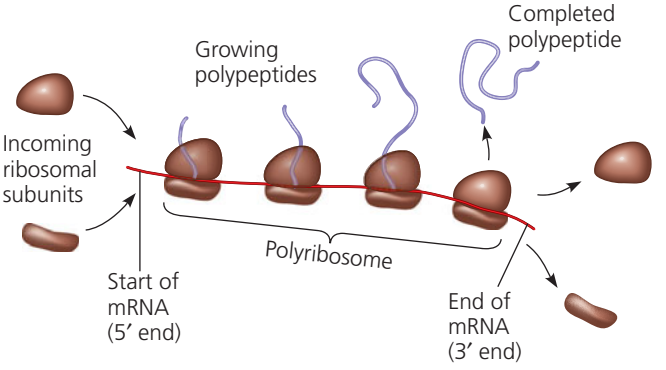
it remains partially embedded in the ER membrane. In either case, it travels in a transport vesicle to the plasma membrane (see Figure 5.8).

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 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 for secretion.

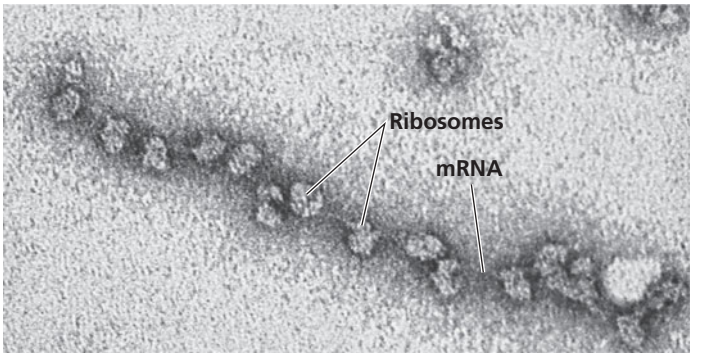
Making Multiple Polypeptides in Bacteria and Eukaryotes

In previous sections, you have 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 14.22)**; that is, a single mRNA is used to make many copies of a polypeptide



(a) An mRNA molecule is generally translated simultaneously by several ribosomes in clusters called polyribosomes.



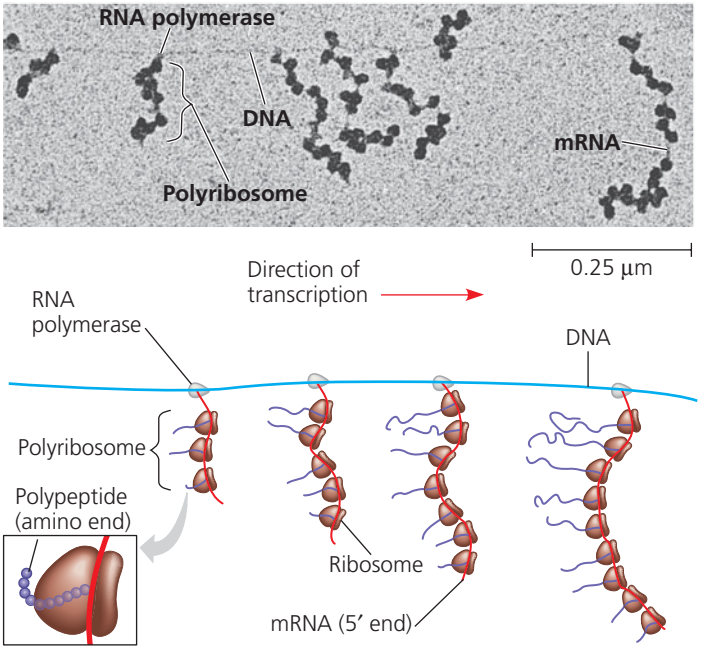
(b) This micrograph shows a large polyribosome in a bacterial cell. Growing polypeptides are not visible here (TEM).

▲ Figure 14.22 Polyribosomes.

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 (see Figure 14.22). They enable a cell to make many copies of a polypeptide very quickly.

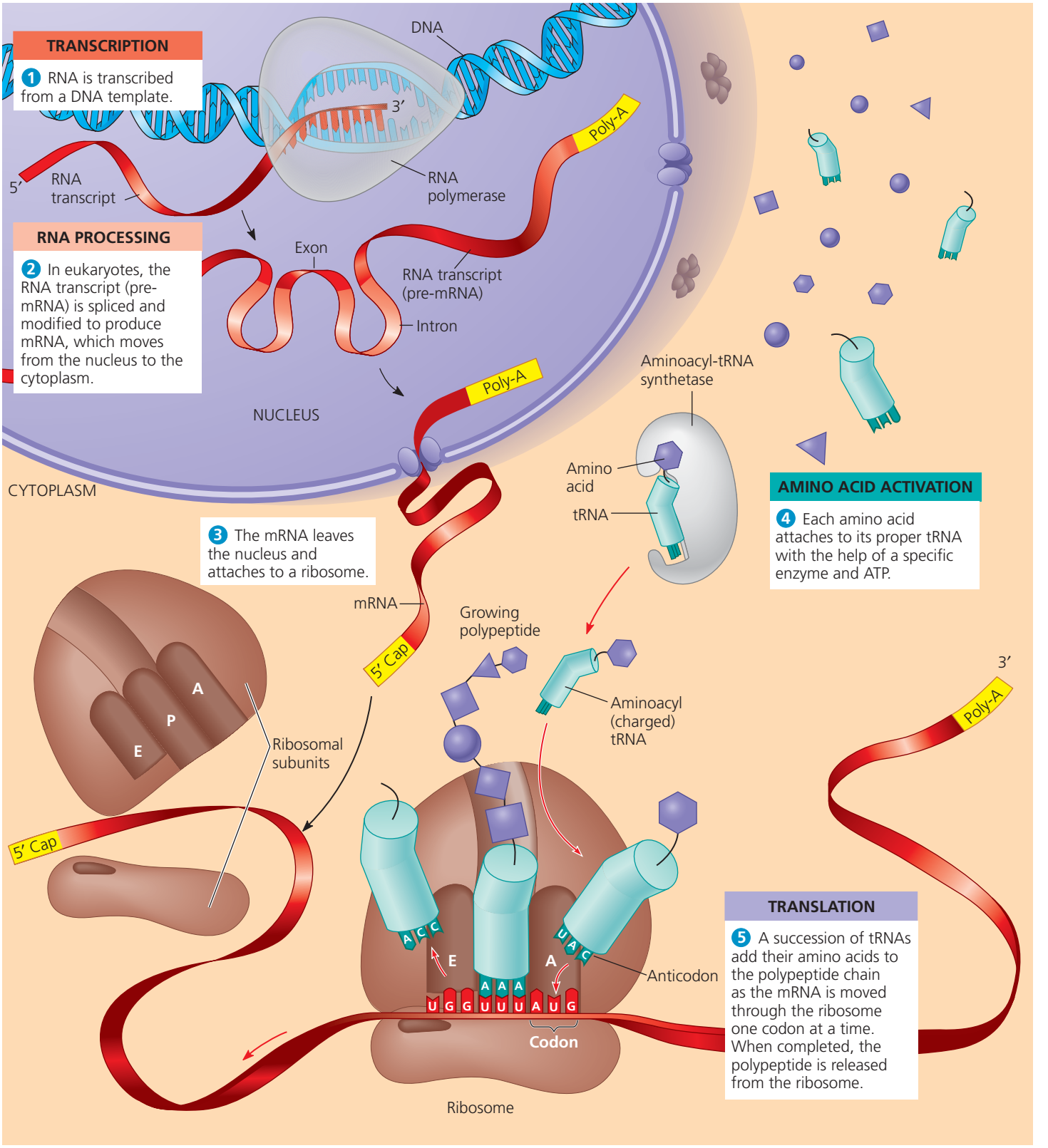
Another way both bacteria and eukaryotes augment the number of copies of a polypeptide is by transcribing multiple mRNAs from the same gene, as we mentioned earlier. However, the coordination of the two processes—transcription and translation—differ in the two groups. The most important differences between 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. In the absence of a nuclear envelope, it can simultaneously transcribe and translate the same gene **(Figure 14.23)**, and the newly made protein can then 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 whose regulation can help coordinate the eukaryotic cell's elaborate activities (see Chapter 15). **Figure 14.24** summarizes the path from gene to polypeptide in a eukaryotic cell.



▲ Figure 14.23 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 stretch 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.

? *Which one of the mRNA molecules started being transcribed first? On that mRNA, which ribosome started translating first?*



▲ Figure 14.24 A summary of transcription and translation in a eukaryotic cell. This diagram shows the path from one gene to one polypeptide. Keep in mind that each gene in the DNA can be transcribed repeatedly into many identical RNA molecules and that each mRNA can be translated repeatedly to yield many identical polypeptide molecules. (Also, remember that the final products of some genes are not polypeptides but RNA molecules, including tRNA and rRNA.) 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.

CONCEPT CHECK 14.4

1. What two processes ensure that the correct amino acid is added to a growing polypeptide chain?
2. Discuss the ways in which rRNA structure likely contributes to ribosomal function.
3. Describe how a polypeptide to be secreted is transported to the endomembrane system.
4. **DRAW IT** Draw a tRNA with the anticodon 3′-CGU-5′. 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.

For suggested answers, see Appendix A.

CONCEPT 14.5

Mutations of one or a few nucleotides can affect protein structure and function

Now that you have explored the process of gene expression, you are ready to understand the effects of changes to the genetic information of a cell (or virus). These changes, called **mutations**, are responsible for the huge diversity of genes found among organisms because mutations are the ultimate source of new genes. Chromosomal rearrangements that affect long segments of DNA are considered large-scale mutations (see Figure 12.14). Here we'll 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 a succession of future generations. If the mutation has an adverse effect on the phenotype of an organism, 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 the production of an abnormal protein **(Figure 14.25**; also see Figure 3.22). 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 Chapter 11). Another disorder caused by a point mutation is a heart condition, familial cardiomyopathy, that is responsible for some incidents of sudden death in young athletes. Point mutations in several genes have been identified, any of which can lead to this disorder.

Types of Small-Scale Mutations

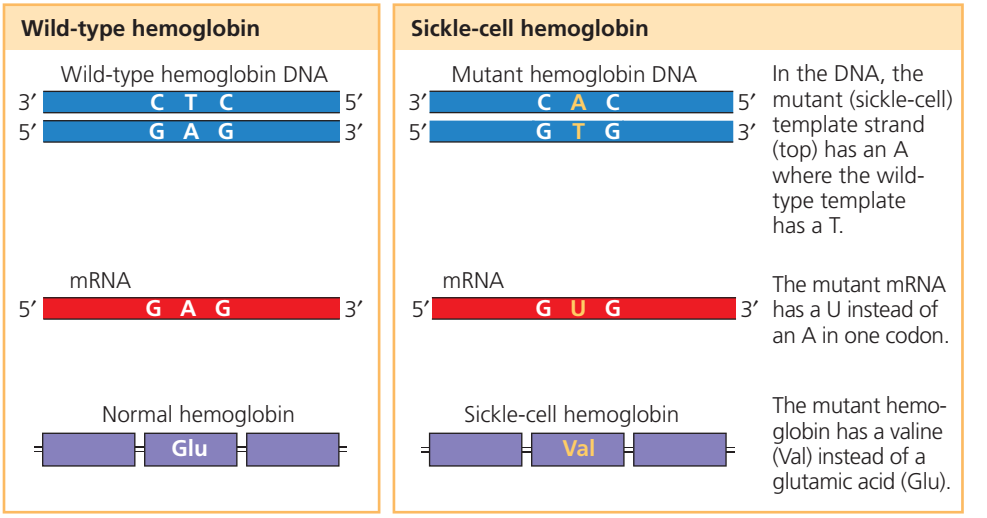
Let's now consider how small-scale mutations affect proteins. Small-scale mutations within a gene can be divided into two general categories: (1) single nucleotide-pair substitutions and (2) nucleotide-pair insertions or deletions. Insertions and deletions can involve one or more nucleotide pairs.

Substitutions

A **nucleotide-pair substitution** is the replacement of one nucleotide and its partner with another pair of nucleotides **(Figure 14.26a)**. 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 14.6). 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.) 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

▼ Figure 14.25 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.



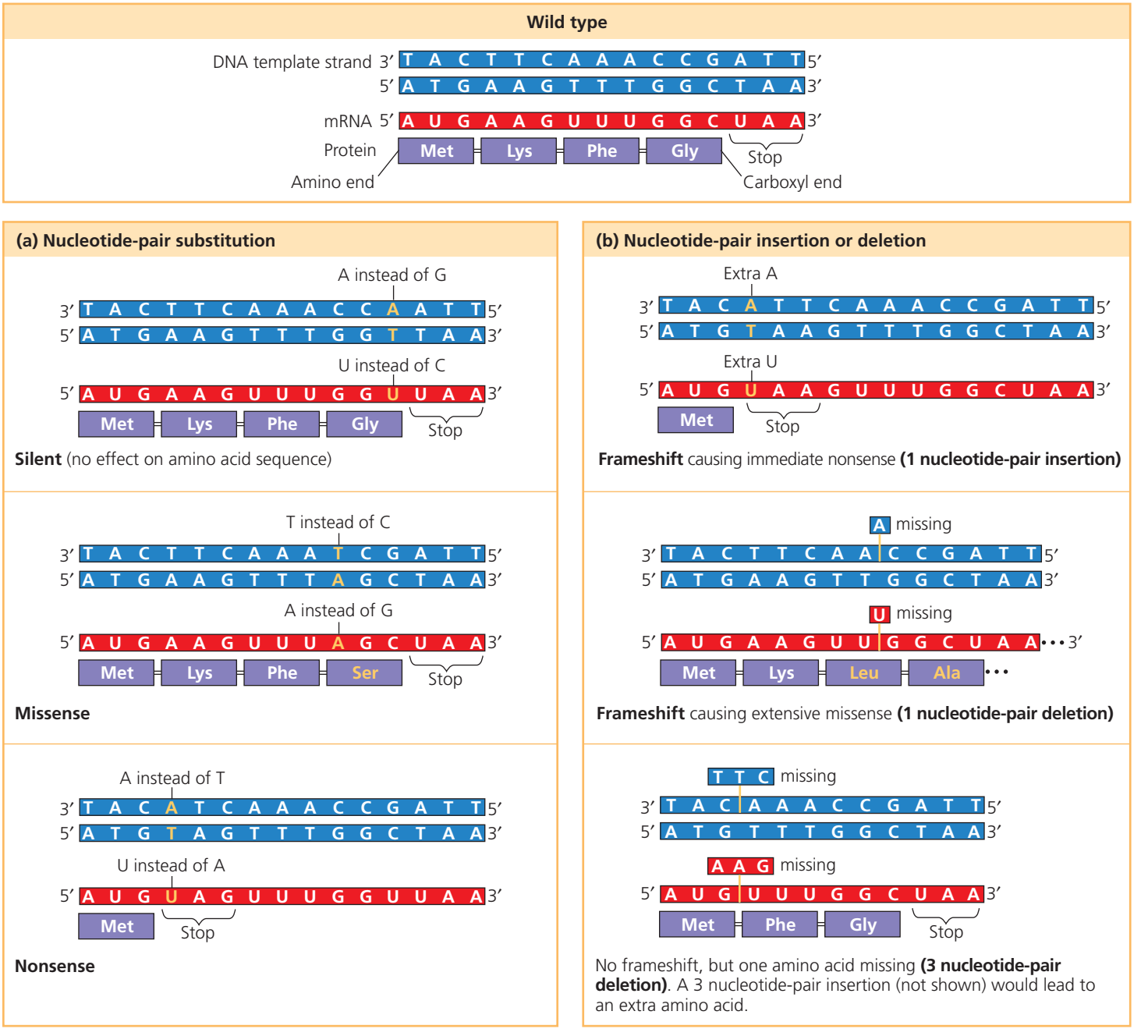
alteration of a single amino acid in a crucial area of a protein—such as in the part of hemoglobin shown in Figure 14.25 or in the active site of an enzyme—will significantly alter protein activity. Occasionally, such a mutation leads to an improved protein or one with novel capabilities, but much more often such mutations are detrimental, leading to a useless or less active protein that impairs cellular function.

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 resulting polypeptide will be shorter than the polypeptide encoded by the normal gene. Nearly all nonsense mutations lead to nonfunctional proteins.

Insertions and Deletions

Insertions and **deletions** are additions or losses of nucleotide pairs in a gene **(Figure 14.26b)**. These mutations have a disastrous effect on the resulting protein more often than substitutions do. Insertion or deletion of nucleotides

▼ Figure 14.26 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.



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**, will occur whenever the number of nucleotides inserted or deleted is not a multiple of three. All the nucleotides that are downstream of the deletion or insertion will be improperly grouped into codons, and the result will be extensive missense, usually ending sooner or later in nonsense and premature termination. Unless the frameshift is very near the end of the gene, the protein is almost certain to be nonfunctional.

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 proofreading and repair systems (see Chapter 13). 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. But he 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 13.19).

Chemical mutagens fall into several categories. Nucleotide analogs are chemicals that are similar to normal DNA nucleotides but that pair incorrectly during DNA replication. Some 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.

CONCEPT CHECK 14.5

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 show effects of the allele under some circumstances (see Concept 11.4). Explain in terms of gene expression.
3. **DRAW IT** The template strand of a gene includes this sequence: 3′-TACTTGTCCGATATC-5′. It is mutated to 3′-TACTTGTCCAATATC-5′. For both versions, draw the DNA, the mRNA, and the encoded amino acid sequence. What is the effect on the amino acid sequence?

For suggested answers, see Appendix A.

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 11). We saw that Morgan and his colleagues assigned such genes to specific loci on chromosomes (Chapter 12). 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 13). Finally, in this chapter, we have considered a functional definition of a gene as a DNA sequence that codes for a specific polypeptide chain. All these definitions are useful, depending on the context in which genes are being studied.

Clearly, the statement that a gene codes for a polypeptide is too simple. 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 useful to focus on genes that code for polypeptides. In this chapter, you have learned 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: Gene expression is precisely regulated. We'll explore gene regulation in the next chapter, beginning with the simpler case of bacteria and continuing with eukaryotes.

14 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 14.1

Genes specify proteins via transcription and translation (pp. 269–274)

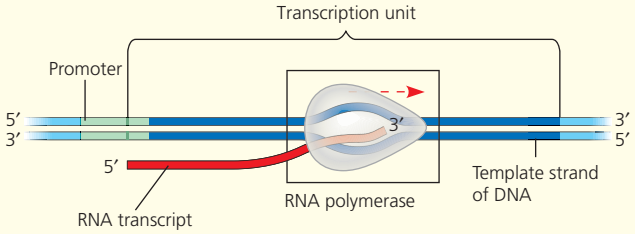
- DNA controls metabolism by directing cells to make specific enzymes and other proteins, via the process of **gene expression**. Beadle and Tatum's studies of mutant strains of *Neurospora* led to the one gene–one polypeptide hypothesis. Genes code for polypeptide chains or specify RNA molecules.
- **Transcription** is the synthesis of RNA complementary to a **template strand** of DNA, providing a nucleotide-to-nucleotide transfer of information. **Translation** is the synthesis of a polypeptide whose amino acid sequence is specified by the nucleotide sequence in **mRNA**; this informational transfer involves a change of language, from that of nucleotides to that of amino acids.
- Genetic information is encoded as a sequence of nonoverlapping nucleotide triplets, or **codons**. A codon in messenger RNA (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 14.2

Transcription is the DNA-directed synthesis of RNA: *a closer look* (pp. 274–276)

- RNA synthesis is catalyzed by **RNA polymerase**, which links together RNA nucleotides complementary to a DNA template strand. This process follows the same base-pairing rules as DNA replication, except that in RNA, uracil substitutes for thymine.

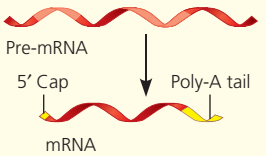


- The three stages of transcription are initiation, elongation, and termination. A **promoter**, often including a **TATA box** in eukaryotes, establishes where RNA synthesis is initiated. **Transcription factors** help eukaryotic RNA polymerase recognize promoter sequences, forming a **transcription initiation complex**. The mechanisms of termination are different in bacteria and eukaryotes.

? *What are the similarities and differences in the initiation of gene transcription in bacteria and eukaryotes?*

CONCEPT 14.3

Eukaryotic cells modify RNA after transcription (pp. 276–278)



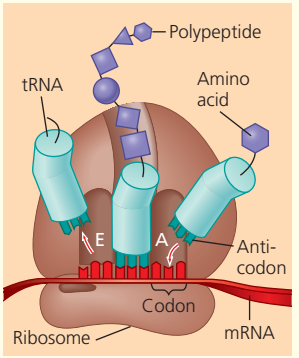
- Eukaryotic pre-mRNAs undergo **RNA processing**, which includes RNA splicing, the addition of a modified nucleotide **5′ cap** to the 5′ end, and the addition of a **poly-A tail** to the 3′ end.
- Most eukaryotic genes are split into segments: They have **introns** interspersed among the **exons** (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 catalytic ability of some RNA molecules, called **ribozymes**, derives from the properties of RNA. 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 14.4

Translation is the RNA-directed synthesis of a polypeptide: *a closer look* (pp. 278–288)

- 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 rRNA as tRNAs move through the **A** and **P sites** and exit through the **E site**.
- After translation, modifications to proteins can affect their shape. 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. 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 time and space by the nuclear membrane.



? *What function do tRNAs serve in the process of translation?*

CONCEPT 14.5

Mutations of one or a few nucleotides can affect protein structure and function (pp. 288–290)

- 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.

? *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

Level 1: Knowledge/Comprehension

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.
   e. DNA nucleases have isolated the transcription unit.
2. Which of the following is *not* true of a codon?
   a. It consists of three nucleotides.
   b. It may code for the same amino acid as another codon.
   c. It never codes for more than one amino acid.
   d. It extends from one end of a tRNA molecule.
   e. It is the basic unit of the genetic code.
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. changeable, depending on the amino acid that attaches to the tRNA.
   e. catalytic, making the tRNA a ribozyme.
4. Which of the following is *not* true of RNA processing?
   a. Exons are cut out before mRNA leaves the nucleus.
   b. Nucleotides may be added at both ends of the RNA.
   c. Ribozymes may function in RNA splicing.
   d. RNA splicing can be catalyzed by spliceosomes.
   e. A primary transcript is often much longer than the final RNA molecule that leaves the nucleus.
5. Which component is *not* directly involved in translation?
   a. mRNA
   b. DNA
   c. tRNA
   d. ribosomes
   e. GTP

Level 2: Application/Analysis

6. Using Figure 14.6, 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′-UUUGGGAAA-3′
   b. 5′-GAACCCCTT-3′
   c. 5′-AAAACCTTT-3′
   d. 5′-CTTCGGGAA-3′
   e. 5′-AAACCCUUU-3′
7. Which of the following mutations would be *most* likely to have a harmful effect on an organism?
   a. a nucleotide-pair substitution
   b. a deletion of three nucleotides near the middle of a gene
   c. a single nucleotide deletion in the middle of an intron
   d. a single nucleotide deletion near the end of the coding sequence
   e. a single nucleotide insertion downstream of, and close to, the start of the coding sequence
8. Fill in the following table:

Type of RNA	Functions
Messenger RNA (mRNA)	
Transfer RNA (tRNA)	
	Plays catalytic (ribozyme) roles and structural roles in ribosomes
Primary transcript	
Small RNAs in spliceosome	

Level 3: Synthesis/Evaluation

9. **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 14.12) 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.
10. **FOCUS ON EVOLUTION**
    Most amino acids are coded for by a set of similar codons (see Figure 14.6). What evolutionary explanations can you give for this pattern? (*Hint*: There is one explanation relating to ancestry, and some less obvious ones of a "form-fits-function" type.)
11. **FOCUS ON 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 13.2.)

For selected answers, see Appendix A.

15 Regulation of Gene Expression

KEY CONCEPTS

15.1 Bacteria often respond to environmental change by regulating transcription

15.2 Eukaryotic gene expression is regulated at many stages

15.3 Noncoding RNAs play multiple roles in controlling gene expression

15.4 Researchers can monitor expression of specific genes

OVERVIEW

Differential Expression of Genes

The fish shown in **Figure 15.1** is keeping an eye out for predators—or, more precisely, half of each eye! *Anableps anableps* is commonly known as "cuatro ojos" ("four eyes") where it lives in regions of southern Mexico and Central and South America. The fish glides through freshwater lakes and ponds with the upper half of each eye protruding from the water. The eye's upper half is particularly well suited for aerial vision and the lower half for aquatic vision. The molecular basis of this specialization has recently been revealed: The cells of the two parts of the eye express a slightly different set of genes involved in vision, even though these two groups of cells are quite similar and contain identical genomes.

▼ **Figure 15.1** How can this fish's eyes see equally well in both air and water?

A hallmark of prokaryotic and eukaryotic cells alike—from bacteria to the cells of a fish—is their intricate and precise regulation of gene expression. Both prokaryotes and eukaryotes must alter their patterns of gene expression in response to changes in environmental conditions. Multicellular eukaryotes must also develop and maintain multiple cell types, each expressing a different subset of genes. This is a significant challenge in gene regulation.

In this chapter, we'll first explore how bacteria regulate expression of their genes in response to different environmental conditions. We'll then examine how eukaryotes regulate gene expression to maintain different cell types. Gene expression in eukaryotes, as in bacteria, is often regulated at the stage of transcription, but control at other stages is also important. In recent years, researchers have been surprised to discover the many roles played by RNA molecules in regulating eukaryotic gene expression, a topic we'll touch on next. Finally, we'll describe a few techniques related to those in Chapter 13 that have been developed to investigate gene expression. Elucidating how gene expression is regulated in different cells is crucial to our understanding of living systems.

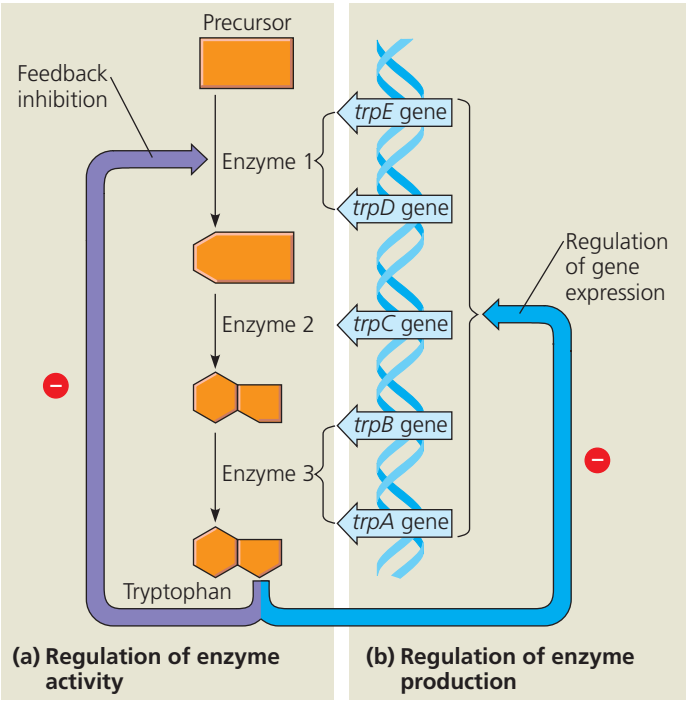
CONCEPT 15.1

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 *E. coli* cell living in the erratic environment of 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. Later, if the human host eats a tryptophan-rich meal, the bacterial cell stops producing tryptophan, thus avoiding wasting its resources to produce a substance that is available from the surrounding solution in prefabricated form. This is just one example of how bacteria tune their metabolism to changing environments.

Metabolic control occurs on two levels, as shown for the synthesis of tryptophan in **Figure 15.2**. First, cells can adjust the activity of enzymes already present. This is a fairly fast response, which relies on the sensitivity of many enzymes to chemical cues that increase or decrease their catalytic activity (see Chapter 6). The activity of the first enzyme in the tryptophan synthesis pathway is inhibited by the pathway's end product **(Figure 15.2a)**. 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.



▲ Figure 15.2 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. Genes *trpE* and *trpD* encode the two subunits of enzyme 1, and genes *trpB* and *trpA* encode the two subunits of enzyme 3. (The genes were named before the order in which they functioned in the pathway was determined.) The ⊖ symbol stands for inhibition.

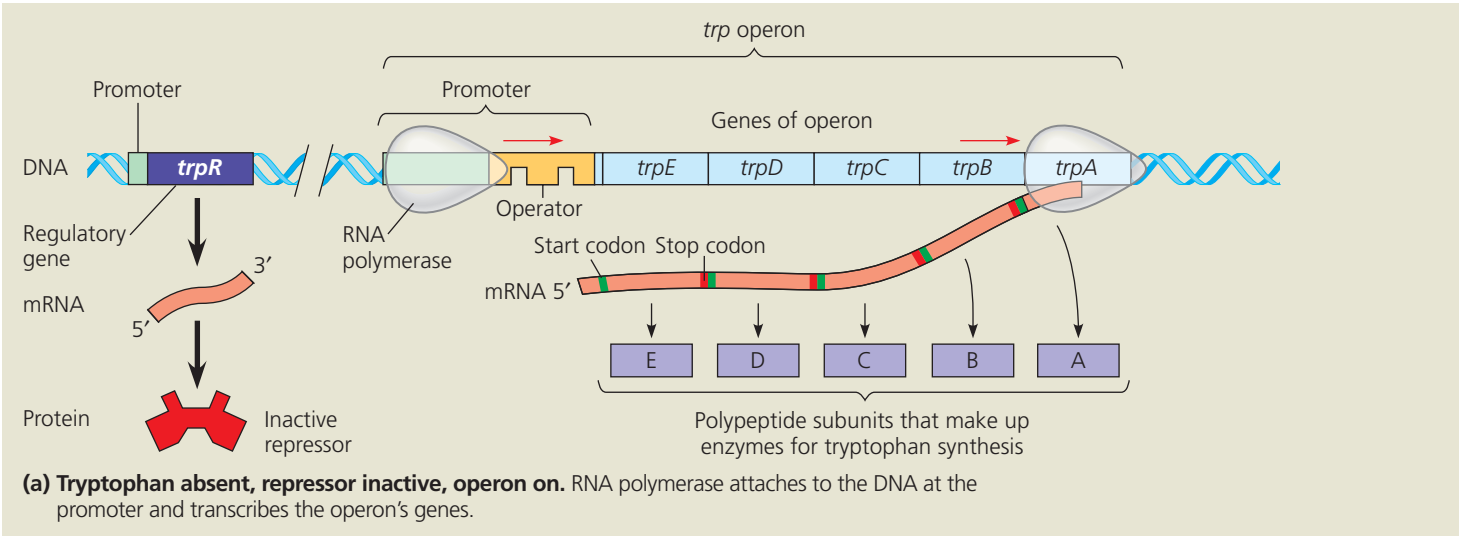
Second, cells can adjust the production level of certain enzymes; that is, they can regulate the expression of the genes encoding the enzymes. If, in our example, the environment provides all the tryptophan the cell needs, the cell stops making the enzymes that catalyze the synthesis of tryptophan **(Figure 15.2b)**. In this case, the control of enzyme production occurs at the level of transcription, the synthesis of messenger RNA coding for these enzymes. More generally, many genes of the bacterial genome are switched on or off by changes in the metabolic status of the cell. One basic mechanism for this control of gene expression 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, using the control of tryptophan synthesis as our first example.

Operons: The Basic Concept

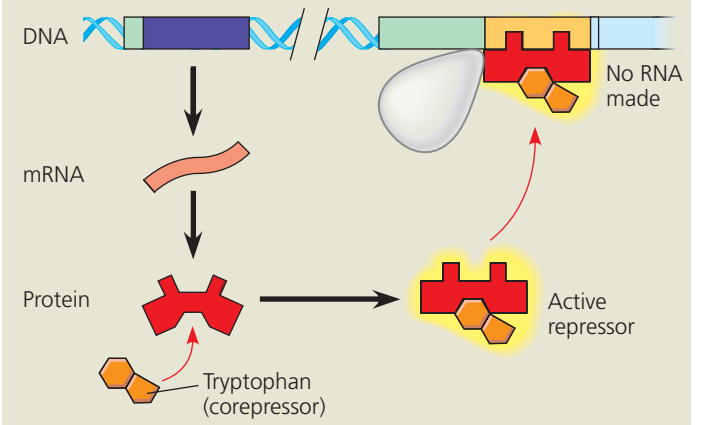
E. coli synthesizes the amino acid tryptophan from a precursor molecule in the multistep pathway shown in Figure 15.2. Each reaction in the pathway is catalyzed by a specific enzyme, and the five genes that code for the subunits of these enzymes are clustered together on the bacterial chromosome. A single promoter serves all five genes, which together constitute a transcription unit. (Recall from Chapter 14 that a promoter is a site where RNA polymerase can bind to DNA and begin transcription.) Thus, transcription gives rise to one long mRNA molecule that codes for the five polypeptides making up the enzymes in the tryptophan pathway. 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 the nutrient medium lacks this amino acid, all the enzymes for the metabolic pathway are synthesized at one time. The 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 RNA polymerase to the genes. All 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 **(Figure 15.3)**.

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 operon



(a) Tryptophan absent, repressor inactive, operon on. RNA polymerase attaches to the DNA at the promoter and transcribes the operon's genes.



(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.

▲ Figure 15.3 The *trp* operon in *E. coli*: regulated synthesis of repressible enzymes. Tryptophan is an amino acid produced by an anabolic pathway catalyzed by repressible enzymes. **(a)** The five genes encoding the polypeptide subunits of the enzymes in this pathway (see Figure 15.2) 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.

? *Describe what happens to the* trp *operon as the cell uses up its store of tryptophan.*

can be switched off by a protein called the *trp* **repressor**. The repressor binds to the operator and blocks attachment of RNA polymerase to the promoter, preventing transcription of the genes. A repressor protein is specific for the operator of a particular operon. For example, the repressor that switches off the *trp* operon by binding to the *trp* operator has no effect on other operons in the *E. coli* genome.

The *trp* repressor is the protein product of a **regulatory gene** called *trpR*, which is located some distance from the *trp* operon and has its own promoter. Regulatory genes 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 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 6.18). The *trp* repressor is synthesized in an inactive form with little affinity for the *trp* operator. Only if tryptophan 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, transcription of the operon's genes resumes. 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

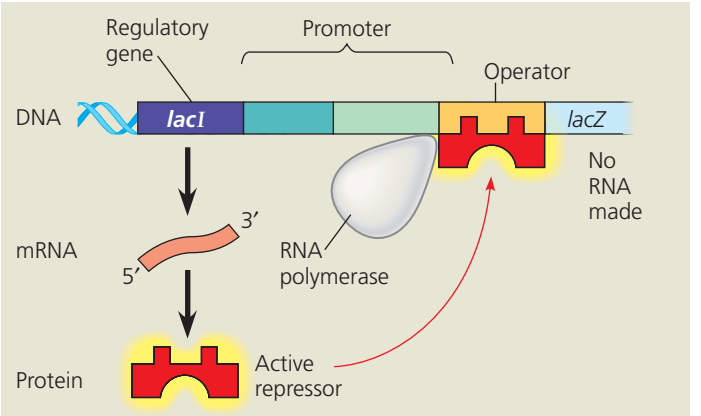
The *trp* operon is said to be a *repressible operon* because its transcription is usually on but can be inhibited (repressed) when a specific small molecule (in this case, tryptophan) binds allosterically to a regulatory protein. In contrast, an *inducible operon* is usually off but can be stimulated (induced) when a specific small molecule interacts with a regulatory protein. The classic example of an inducible operon is the *lac* operon (*lac* for lactose), which was the subject of Jacob and Monod's pioneering research.

The disaccharide lactose (milk sugar) is available to *E. coli* in the human colon if the host drinks milk. Lactose metabolism 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 a thousandfold within about 15 minutes.

The gene for β-galactosidase is part of the *lac* operon, which includes two other genes coding for enzymes that function in lactose utilization. The entire transcription unit is under the command of one main operator and promoter. The regulatory gene, *lacI*, located outside the operon, codes for an allosteric repressor protein that can switch off the *lac* operon by binding to the 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.

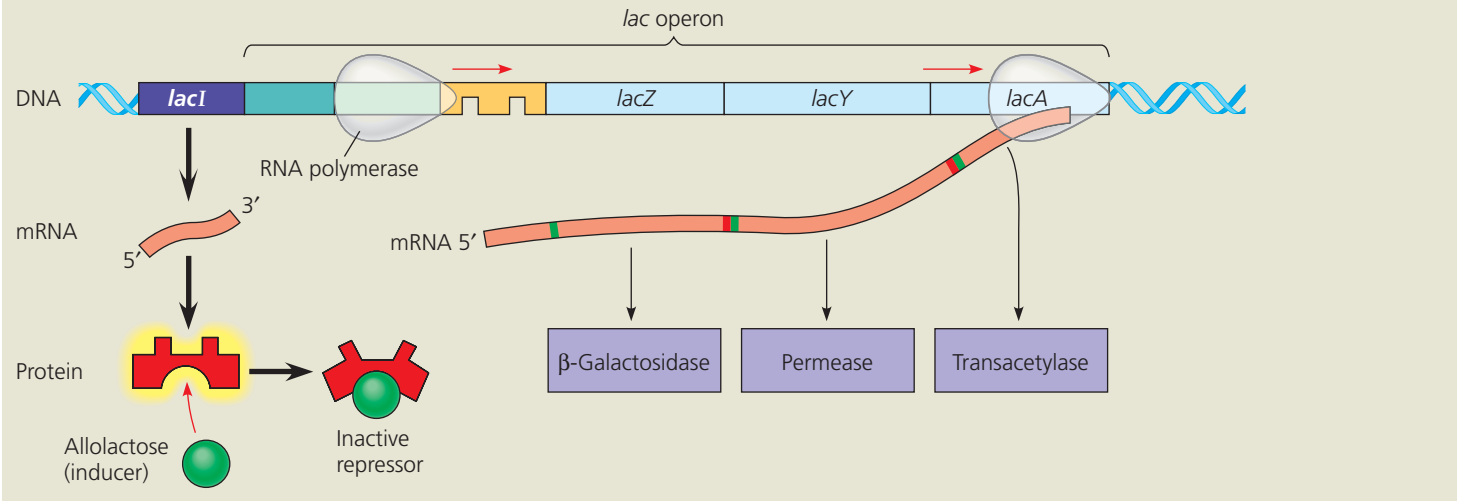
For the *lac* operon, the inducer is allolactose, an isomer of lactose formed in small amounts from lactose that enters the cell. In the absence of lactose (and hence allolactose), the *lac* repressor is in its active configuration, and the genes of the *lac* operon are silenced **(Figure 15.4a)**. If lactose is added to the cell's surroundings, allolactose binds to the *lac* repressor and alters its conformation, nullifying the repressor's ability to attach to the operator. Without bound repressor, the *lac* operon is transcribed into mRNA for the lactose-utilizing enzymes **(Figure 15.4b)**.

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 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



(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.

▼ Figure 15.4 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 these three enzymes are clustered in the *lac* operon. One gene, *lacZ*, codes for β-galactosidase, which hydrolyzes lactose to glucose and galactose. The second gene, *lacY*, codes for a permease, the membrane protein that transports lactose into the cell. The third gene, *lacA*, codes for an enzyme called transacetylase, whose function in lactose metabolism is still unclear. The gene for the *lac* repressor, *lacI*, happens to be adjacent to the *lac* operon, an unusual situation. The function of the teal region at the upstream end of the promoter (the left end in these diagrams) will be revealed in Figure 15.5.



(b) Lactose present, repressor inactive, operon on. Allolactose, an isomer of lactose, derepresses the operon by inactivating the repressor. In this way, the enzymes for lactose utilization are induced.

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.

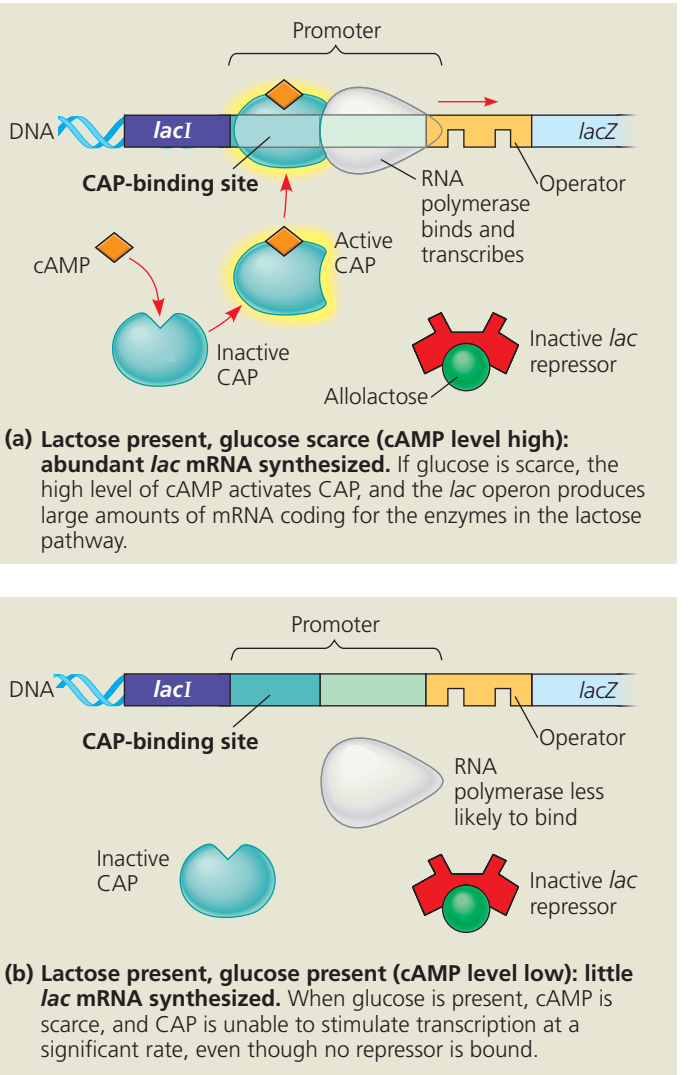
Regulation of both the *trp* and *lac* operons involves the *negative* control of genes, because the operons are switched off by the active form of the repressor protein. It may be easier to see this for the *trp* operon, but it is also true for the *lac* operon. Allolactose induces enzyme synthesis not by acting directly on the genome, but by freeing the *lac* operon from the negative effect of the repressor. Gene regulation is said to be *positive* only when a regulatory protein interacts directly with the genome to switch transcription on. Let's look at an example of the positive control of genes, again involving the *lac* operon.

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 7.9) 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 genome? Again, the mechanism depends on the interaction of an allosteric regulatory protein with a small organic molecule, in this case **cyclic AMP (cAMP)**, which accumulates when glucose is scarce. The regulatory protein, called *catabolite activator protein (CAP)*, is an **activator**, a protein that binds to DNA and stimulates transcription of a gene. When cAMP binds to this regulatory protein, CAP assumes its active shape and can attach to a specific site at the upstream end of the *lac* promoter **(Figure 15.5a)**. This attachment increases the affinity of RNA polymerase for the promoter, which is actually rather low even when no repressor is bound to the operator. By facilitating the binding of RNA polymerase to the promoter and thereby increasing the rate of transcription, the attachment of CAP 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, CAP detaches from the operon. Because CAP is inactive, RNA polymerase binds less efficiently to the promoter, and transcription of the *lac* operon proceeds at only a low level, even in the presence of lactose **(Figure 15.5b)**. Thus, the *lac* operon is under dual control: negative control by the *lac* repressor and positive control by CAP. The state of the *lac* repressor (with or without bound allolactose) determines whether or not transcription of the *lac* operon's genes occurs at all; the state of CAP (with or without bound cAMP) controls the *rate* of transcription if the operon



(a) Lactose present, glucose scarce (cAMP level high): abundant *lac* mRNA synthesized. If glucose is scarce, the high level of cAMP activates CAP, and the *lac* operon produces large amounts of mRNA coding for the enzymes in the lactose pathway.

(b) Lactose present, glucose present (cAMP level low): little *lac* mRNA synthesized. When glucose is present, cAMP is scarce, and CAP is unable to stimulate transcription at a significant rate, even though no repressor is bound.

▲ Figure 15.5 Positive control of the lac operon by catabolite activator protein (CAP). RNA polymerase has high affinity for the *lac* promoter only when catabolite activator protein (CAP) is bound to a DNA site at the upstream end of the promoter. CAP 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 also is available, the cell preferentially catabolizes glucose and makes very little of the lactose-utilizing enzymes.

is repressor-free. It is as though the operon has both an on-off switch and a volume control.

In addition to regulating the *lac* operon, CAP 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 CAP 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 glucose to survive. The compounds present in the cell at the 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 15.1

1. How does binding of the *trp* corepressor and the *lac* inducer to their respective repressor proteins alter repressor function and transcription in each case?
2. Describe the binding of RNA polymerase, 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 15.2

Eukaryotic gene expression is regulated at many stages

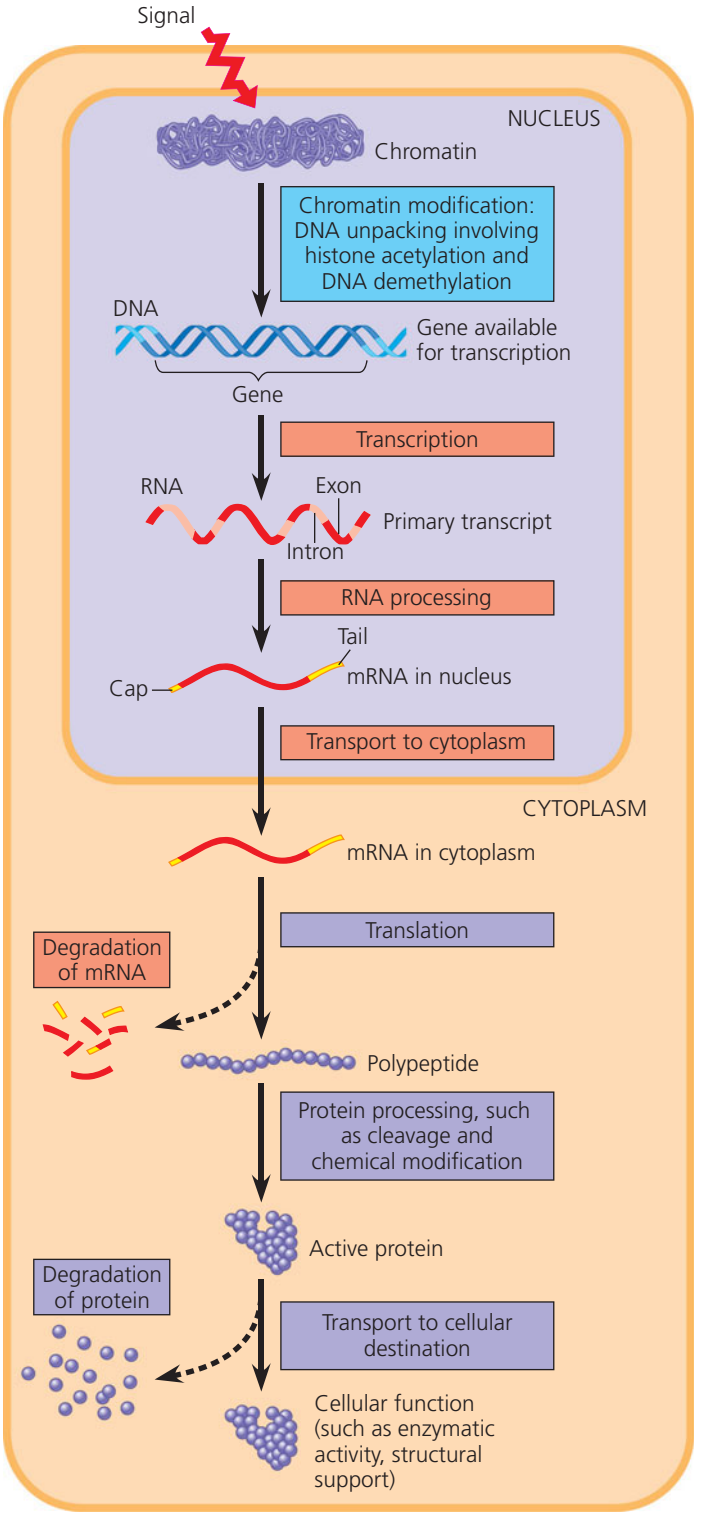
All organisms, whether prokaryotes or eukaryotes, must regulate which genes are expressed at any given time. Both unicellular organisms and the cells of multicellular organisms must continually turn genes on and off in response to signals from their external and internal environments. Regulation of gene expression is also essential for cell specialization in multicellular organisms, which are made up of different types of cells, each with a distinct role. To perform its role, each cell type must maintain a specific program of gene expression in which certain genes are expressed and others are not.

Differential Gene Expression

A typical human cell might express about 20% of its protein-coding genes at any given time. Highly differentiated cells, such as muscle or nerve cells, express an even smaller fraction of their genes. Almost all the cells in an organism contain an identical genome. (Cells of the immune system are one exception, as you will see in Chapter 35.) However, the subset of genes expressed in the cells of each type is unique, allowing 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, whether a single-celled eukaryote or a particular cell type in a multicellular organism, depends on the appropriate set of genes being expressed. The transcription factors of a cell must locate the right genes at the right time, a task on a par with finding a needle in a haystack. When gene expression proceeds abnormally, serious imbalances and diseases, including cancer, can arise.

Figure 15.6 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 15.6 is a potential control point at which gene expression can be turned on or off, accelerated, or slowed down.



▲ Figure 15.6 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, orange = RNA, purple = protein). The nuclear envelope separating transcription from translation in eukaryotic cells offers an opportunity for post-transcriptional control in the form of RNA processing that is absent in prokaryotes. In addition, eukaryotes have a greater variety of control mechanisms operating before transcription and after translation. The expression of any given gene, however, does not necessarily involve every stage shown; for example, not every polypeptide is cleaved.

Fifty years ago, an understanding of the mechanisms that control gene expression in eukaryotes seemed almost hopelessly out of reach. Since then, new research methods, notably advances in DNA technology (see Concept 13.4), have enabled molecular biologists to uncover many of the details of eukaryotic gene regulation. In all organisms, a common control point for gene expression is 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 is most often the case for bacteria, the greater complexity of eukaryotic cell structure and function provides opportunities for regulating gene expression at many additional stages (see Figure 15.6). In the remainder of this section, we'll examine some of the important control points of eukaryotic gene expression more closely.

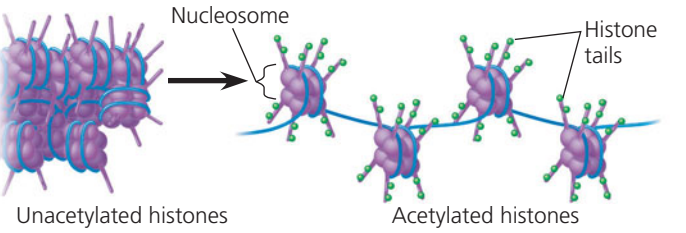
Regulation of Chromatin Structure

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 13.21). 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. The location of a gene's promoter relative to nucleosomes and to the sites where the DNA attaches to the chromosome scaffold or nuclear lamina can affect whether the gene is transcribed. In addition, genes within heterochromatin, which is highly condensed, are usually not expressed. Lastly, certain chemical modifications to the histone proteins and to the DNA of chromatin can influence both chromatin structure and gene expression. Here we examine the effects of these modifications, which are catalyzed by specific enzymes.

Histone Modifications and DNA Methylation

There is abundant evidence that chemical modifications to histones, the proteins around which the DNA is wrapped in nucleosomes, play a direct role in the regulation of gene transcription. The N-terminus of each histone molecule in a nucleosome protrudes outward from the nucleosome. These 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. Generally, **histone acetylation** appears to promote transcription by opening up the chromatin structure **(Figure 15.7)**, while addition of methyl groups can lead to condensation of chromatin and reduced transcription.

While some enzymes methylate the tails of histone proteins, a different set of enzymes can methylate certain bases in the DNA itself, 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 12.8), are generally more methylated than regions of actively transcribed DNA, although there are exceptions.



▲ Figure 15.7 A simple model of the effect of histone acetylation. The amino acids in the N-terminal tails of histones are accessible for chemical modification such as addition of acetyl groups (green balls). A region of chromatin in which nucleosomes are unacetylated forms a compact structure (left) in which the DNA is not transcribed. Highly acetylated nucleosomes (right) cause the chromatin to be less compact and the DNA accessible for transcription.

On a smaller scale, individual genes are usually more heavily methylated in cells in which they are not expressed. Removal of the extra methyl groups can turn on some of these genes. 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. In this way, methylation patterns can be inherited.

Epigenetic Inheritance

The chromatin modifications discussed above do not entail a change in the DNA sequence, yet they may be passed along to future generations of cells. Inheritance of traits transmitted by mechanisms not directly involving the nucleotide sequence is called **epigenetic inheritance**. Whereas mutations in DNA are permanent, modifications to the chromatin can be reversed.

Researchers are amassing more and more evidence for the importance of epigenetic information in the regulation of gene expression. Epigenetic variations might help explain why 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 also seen in some cancers, where the alterations 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 typical eukaryotic gene and its transcript.

Organization of a Typical Eukaryotic Gene

A eukaryotic gene and the DNA elements (segments) that control it are typically organized as shown in **Figure 15.8**, which extends what you learned about eukaryotic genes in Chapter 14. Recall that a cluster of proteins called a *transcription initiation complex* assembles on the promoter sequence at the "upstream" end of the gene. One of these proteins, RNA polymerase II, then proceeds to transcribe the gene, synthesizing a primary RNA transcript (more specifically, 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 having particular nucleotide sequences that serve as binding sites for the proteins called transcription factors, which in turn regulate transcription. Control elements on the DNA and the transcription factors they bind are critical to the precise regulation of gene expression seen in different cell types.

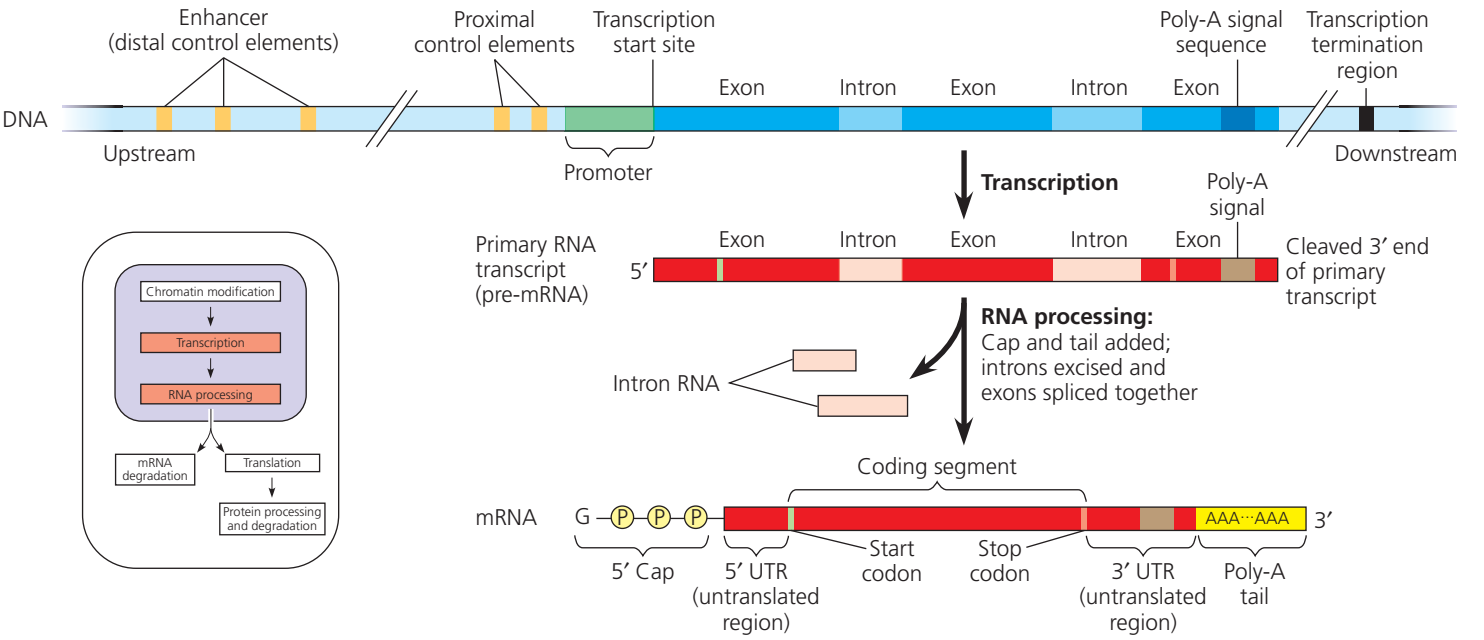
The Roles of Transcription Factors

To initiate transcription, eukaryotic RNA polymerase requires the assistance of transcription factors. Some transcription factors, such as those illustrated in Figure 14.9, are essential for the transcription of *all* protein-coding genes; therefore, they are often called *general transcription factors*. Only a few general transcription factors independently bind a DNA sequence, such as the TATA box within the promoter; the others primarily bind proteins, including each other and 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, producing a complementary strand of RNA.

The interaction of general transcription factors and RNA polymerase II with a promoter usually leads to only a low rate of initiation and production of few RNA transcripts. In eukaryotes, high levels of transcription of 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 15.8, some control elements, named *proximal control elements*, are located close to the promoter. (Although some biologists consider proximal control elements part of the promoter, in this book we do not.) 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 or in a different cell type or location in the organism. Each enhancer, however, is generally associated with only that gene and no other.

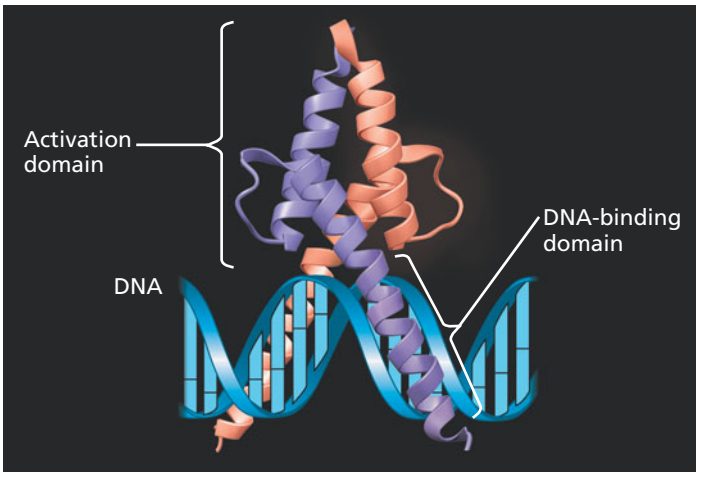
In eukaryotes, the rate of gene expression can be strongly increased or decreased by the binding of specific transcription



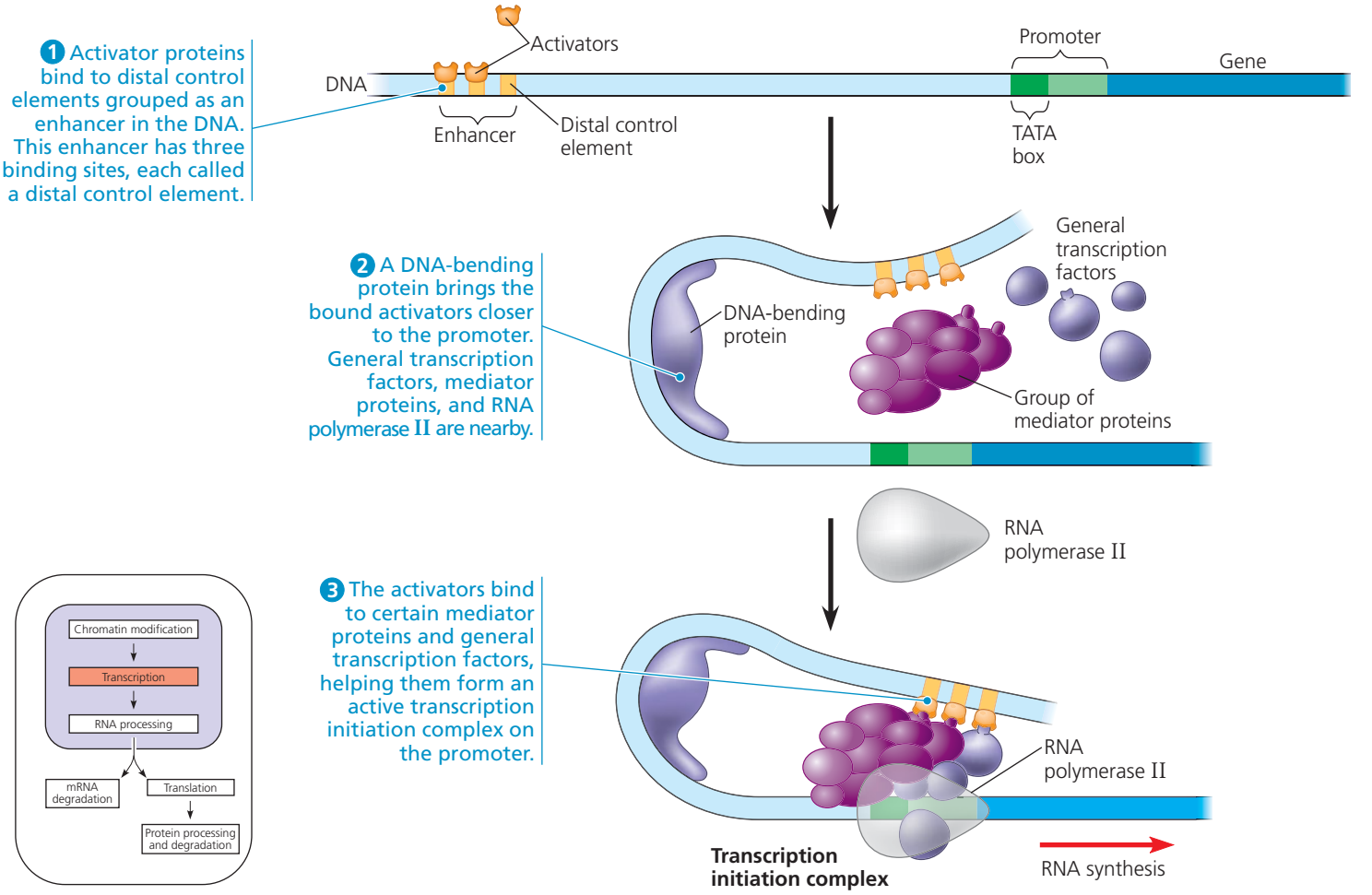
▲ Figure 15.8 A eukaryotic gene and its transcript. Each eukaryotic gene has a promoter, a DNA sequence where RNA polymerase binds and starts transcription, proceeding "downstream." A number of control elements (gold) are involved in regulating the initiation of transcription; these are DNA sequences 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. 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; splicing and poly-A tail addition may also occur while transcription is under way (see Figure 14.11).

factors, either activators or repressors, to the control elements of enhancers. Hundreds of transcription activators have been discovered in eukaryotes; the structure of one example is shown in **Figure 15.9**. In a large number of activator proteins, researchers have identified two common structural elements: 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 transcription of a given gene.

Figure 15.10 shows a current model for how binding of activators to an enhancer located far from the promoter can influence transcription. Protein-mediated bending of the DNA is thought to bring the bound activators into contact with a group of *mediator proteins*, which in turn interact with proteins at the promoter. These protein-protein interactions help assemble and position the initiation complex on the promoter.



▲ **Figure 15.9 MyoD, a transcription activator.** The MyoD protein is made up of two subunits (purple and salmon) with extensive regions of α helix. Each subunit has one DNA-binding domain and one activation domain. The latter includes binding sites for the other subunit and other proteins. MyoD is involved in muscle development in vertebrate embryos.



▲ **Figure 15.10 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 called activators bind to the enhancer DNA sequences and then to a group of mediator proteins, which in turn bind to general transcription factors, assembling the transcription initiation complex. These protein-protein interactions facilitate the 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.

Support for this model includes a study showing that the proteins regulating a mouse globin gene contact both the gene's promoter and an enhancer located about 50,000 nucleotides upstream. Evidently, these two regions in the DNA must come together in a very specific fashion for this interaction to occur.

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 or, in some cases, turning off transcription even when activators are bound. Other repressors block the binding of activators to proteins that allow the activators to bind to DNA. In the **Scientific Skills Exercise**, you can work with data from an experiment that identified the control elements in the enhancer of a particular human gene.

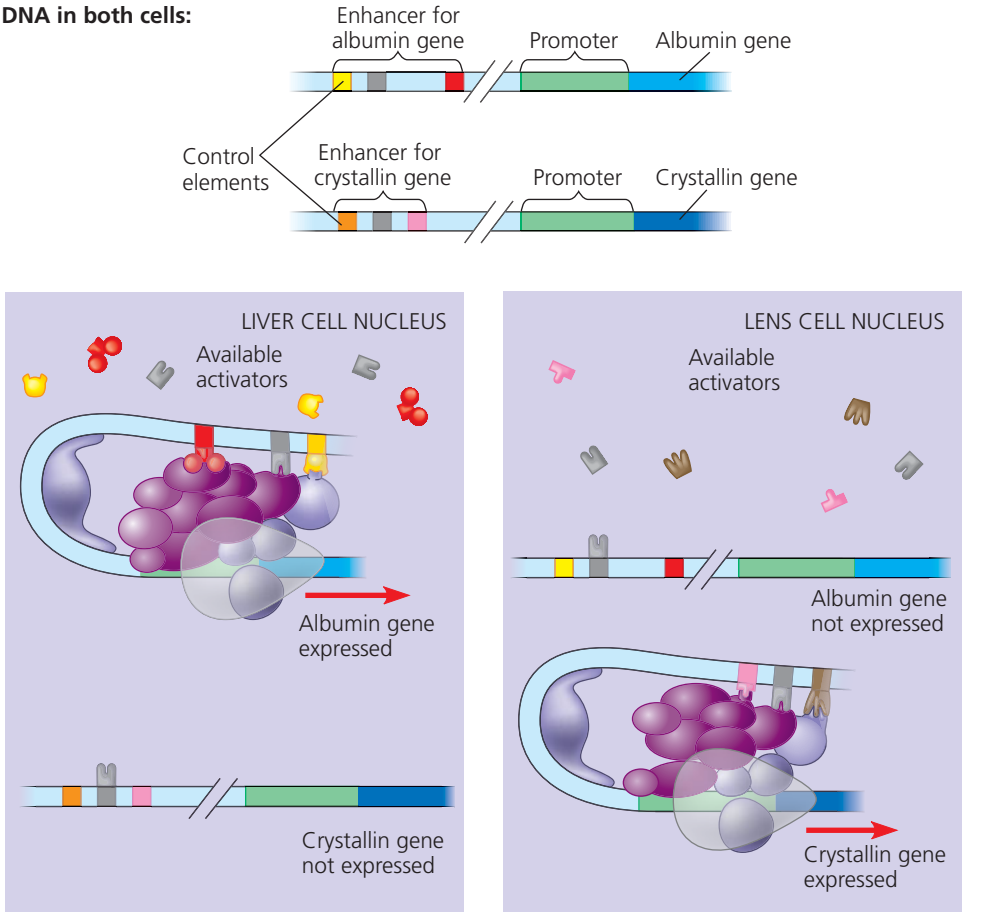
In addition to influencing transcription directly, some activators and repressors act indirectly by affecting chromatin structure. Studies using yeast and mammalian cells show that some activators recruit proteins that acetylate histones near the promoters of specific genes, thus promoting transcription (see Figure 15.7). Similarly, some repressors recruit proteins that remove acetyl groups from histones, leading to reduced transcription, a phenomenon called *silencing*. Indeed, the recruitment of proteins that modify chromatin seems to be the most common mechanism of repression in eukaryotes.

Combinatorial Control of Gene Activation In eukaryotes, the precise control of transcription depends largely on the binding of activators to DNA control elements. Considering the great number of genes that must be regulated in a typical animal or plant cell, the number of completely different nucleotide sequences found 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 of which can bind only one or two specific transcription factors. It is the particular *combination* of control elements in an enhancer associated with a gene, rather than the presence of a single unique control element, that is important in regulating transcription of the gene.

Even with only a dozen control element sequences available, a very large number of combinations are possible. A particular combination of control elements will be able to activate transcription only when the appropriate activator proteins are present, which may occur at a precise time during development or in a particular cell type. **Figure 15.11** illustrates how the use of different combinations of just a few control elements can allow differential regulation of transcription in two cell types. This can occur because each cell type contains a different group of activator proteins. How these groups came to differ during embryonic development will be explored in Chapter 16.

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



(a) Liver cell. The albumin gene is expressed, and the crystallin gene is not.

(b) Lens cell. The crystallin gene is expressed, and the albumin gene is not.

▲ **Figure 15.11 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 share one control element (gray), each enhancer has a unique combination of elements. All the activators required for high-level expression of the albumin gene are present only in liver cells **(a)**, whereas the activators needed for expression of the crystallin gene are present only in lens cells **(b)**. For simplicity, we consider only the role of activators here, although the presence or absence of repressors may also influence transcription in certain cell types.

? *Describe the enhancer for the albumin gene in each cell. How would the nucleotide sequence of this enhancer in the liver cell compare with that in the lens cell?*

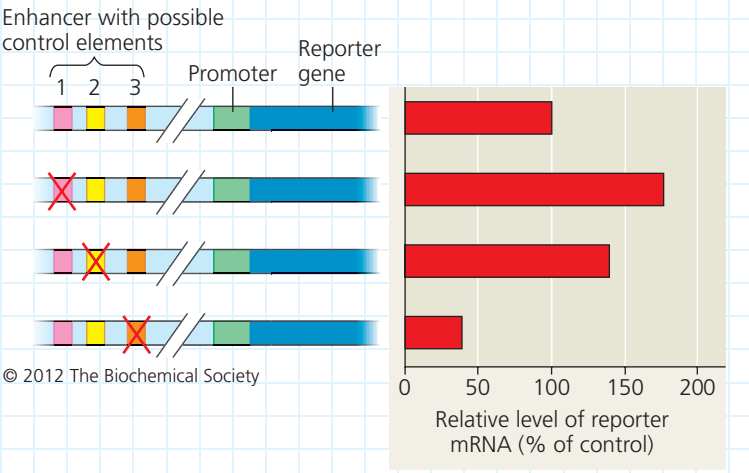
Scientific Skills Exercise

Analyzing DNA Deletion Experiments

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 expression of the gene can also be affected by control elements. These can be thousands of base pairs upstream of the promoter, grouped in an enhancer. Since the distance and spacing of these 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 inflammation.

How the Experiment Was Done The researchers hypothesized that there were three possible control elements in an enhancer region 8–9 kilobases 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 ("constructs") with the intact enhancer region upstream of a "reporter gene," a gene whose mRNA product could be easily measured experimentally. Next, they synthesized three more DNA constructs but deleted one of the three proposed control elements in each (see left side of figure). The researchers then introduced each DNA construct into a separate human cell culture, where the cells took up the artificial DNA molecules. 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 that deletions would have had 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, so expression of the reporter gene is used to mimic expression of the *mPGES-1* gene.)

Data from the Experiment The diagrams on the left side of the figure show the intact DNA sequence (top) and the three experimental DNA sequences. A red X indicates the possible control element (1, 2, or 3) that was deleted in each experimental DNA sequence. The area between the slashes represents the approximately 8 kilobases 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%).



Interpret the Data

1. (a) What is the independent variable in the graph (that is, what variable was manipulated by the scientists)? (b) What is the dependent variable (that is, what variable responded to the changes in the independent 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.

Data from J. N. Walters et al., Regulation of human microsomal prostaglandin E synthase-1 by IL-1β requires a distal enhancer element with a unique role for C/EBPβ, *Biochemical Journal* (2012). doi:10.1042/BJ20111801

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

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 concurrently. With a few minor exceptions, operons that work in this way have *not* been found in eukaryotic cells.

Co-expressed eukaryotic genes, such as genes coding for the enzymes of a metabolic pathway, are typically scattered over different chromosomes. In these cases, coordinate gene expression depends on the association of a specific combination of control elements with every gene of a dispersed group. The presence of these elements can be compared to the raised flags on a few mailboxes out of many, signaling to the mail carrier to check those boxes. Copies of the activators 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 5.23). Every gene whose transcription is stimulated by a particular 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 lead to activation of particular transcription activators or repressors (see Figure 5.26). Coordinate regulation in such pathways is the same as for steroid hormones: Genes with the same control elements are activated by the same chemical signals. Systems for coordinating gene regulation probably arose early in evolutionary history.

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. Researchers are discovering more and more regulatory mechanisms that operate at various stages after transcription (see Figure 15.6). These mechanisms allow a cell to fine-tune gene expression rapidly in response to environmental changes without altering its transcription patterns. Here we discuss how cells can regulate gene expression once a gene has been transcribed.

RNA Processing

RNA processing in the nucleus and the export of mature RNA to the cytoplasm provide opportunities for regulating gene expression that are 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 RNA sequences within the primary transcript.

A simple example of alternative RNA splicing is shown in **Figure 15.12** for the troponin T gene, which encodes two different (though related) proteins. Other genes offer possibilities for far greater numbers of products. For instance, researchers have found a gene in *Drosophila* with enough alternatively spliced exons to generate about 19,000 membrane proteins with 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 unique form of the protein, which acts as an identification badge on the cell surface.

It is clear that 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 about ten years ago. The number of human genes was found to be similar to that of a soil worm (nematode), mustard plant, or sea anemone. This discovery prompted questions about what, if not the number of genes, accounts for the more complex morphology (external form) of humans. It turns out that more than 90% of human protein-coding genes probably 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.

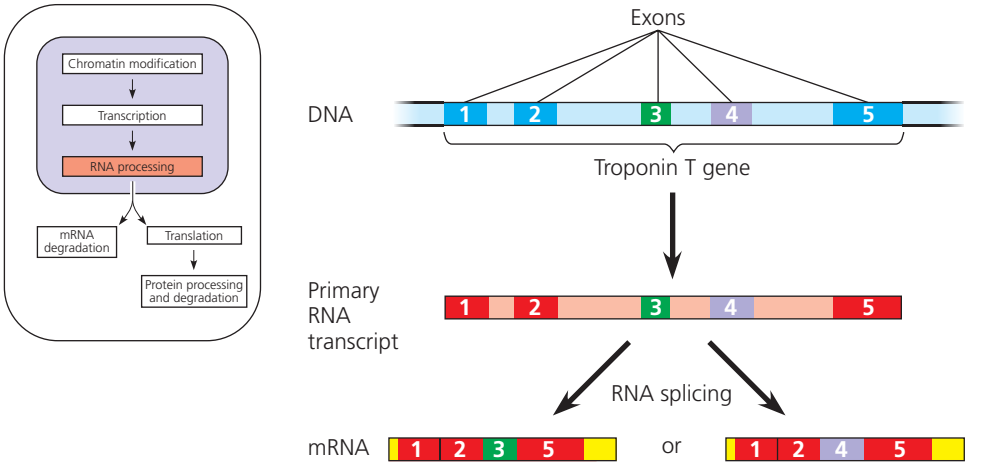
mRNA Degradation

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 of their synthesis. 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, 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 these cells. Nucleotide sequences that affect how long an mRNA remains intact are often found in the untranslated region (UTR) at the 3′ end of the molecule (see Figure 15.8).

During the past few years, other mechanisms that degrade or block expression of mRNA molecules have come to light. These mechanisms involve an important group of newly discovered RNA molecules that regulate gene expression at several levels, and we'll discuss them later in this chapter.

Initiation of Translation

Translation presents another opportunity for regulating gene expression; such regulation occurs most commonly at the initiation stage (see Figure 14.18). For some mRNAs, the initiation of translation can be blocked by regulatory proteins that bind



▲ **Figure 15.12 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.

to specific sequences or structures within the 5′ or 3′ UTR, preventing the attachment of ribosomes. (Recall from Chapter 14 that both the 5′ cap and the poly-A tail of an mRNA molecule are important for ribosome binding.) A different mechanism for blocking translation is seen in a variety of mRNAs present in the eggs of many organisms: Initially, these stored mRNAs lack poly-A tails of sufficient length to allow translation initiation. At the appropriate time during embryonic development, however, a cytoplasmic enzyme adds more adenine (A) nucleotides, prompting translation to begin.

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 of the protein factors required to initiate translation. This mechanism plays a role in starting translation of mRNAs that are stored in eggs. Just after 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.

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, and proteins destined for the surface of animal cells acquire sugars. Cell-surface proteins and many others must also be transported to target destinations in the cell in order to function. 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 means of 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. To mark a particular protein for destruction, the cell commonly attaches molecules of a small protein called ubiquitin to the protein, which triggers its destruction by protein complexes in the cell.

CONCEPT CHECK 15.2

1. In general, what is the effect of histone acetylation and DNA methylation on gene expression?
2. Compare the roles of general and specific transcription factors in regulating gene expression.
3. 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 15.3

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, most of the remaining DNA was assumed to be untranscribed. The idea was that since it didn't specify proteins or the few known types of RNA, such DNA didn't contain meaningful genetic information. However, a flood of recent data has contradicted this idea. For example, an in-depth study of a region comprising 1% of the human genome showed that more than 90% of that region was transcribed. Introns accounted for only a fraction of this transcribed, nontranslated RNA. These and other results suggest that a significant amount of the genome may be transcribed into non-protein-coding RNAs (also called *noncoding RNAs*, or *ncRNAs*), including a variety of small RNAs and longer RNA transcripts. While many questions about the functions of these RNAs remain unanswered, researchers are uncovering more evidence of their biological roles every day.

Biologists are excited about these recent discoveries, which hint at a large, diverse population of RNA molecules in the cell that play crucial roles in regulating gene expression—and have gone largely unnoticed until now. Clearly, we must revise our long-standing view that because mRNAs code for proteins, they are the most important RNAs functioning in the cell. This represents a major shift in the thinking of biologists, one that you are witnessing as students entering this field of study. It's as if our exclusive focus on a famous rock star has blinded us to the many backup musicians and songwriters working behind the scenes.

Regulation by both small and large ncRNAs is known to occur at several points in the pathway of gene expression, including mRNA translation and chromatin modification. We'll focus mainly on two types of small ncRNAs that have been extensively studied in the past few years; the importance of these RNAs was acknowledged when they were the focus of the 2006 Nobel Prize in Physiology or Medicine.

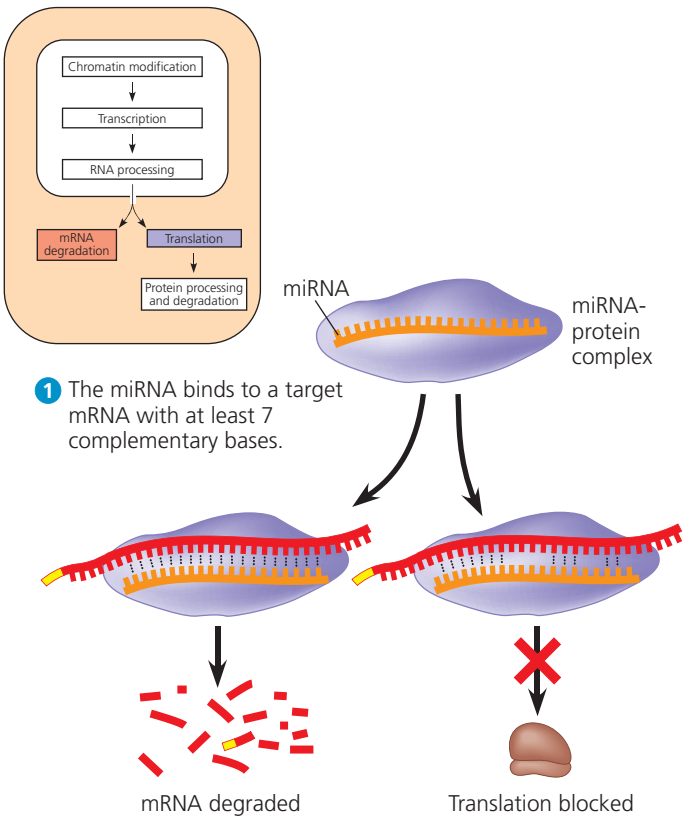
Effects on mRNAs by MicroRNAs and Small Interfering RNAs

Since 1993, a number of research studies have uncovered small single-stranded RNA molecules, called **microRNAs (miRNAs)**, that are 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. The miRNA allows the complex to bind to any mRNA molecule with 7–8 nucleotides of complementary sequence. The

miRNA-protein complex then either degrades the target mRNA or blocks its translation **(Figure 15.13)**. It has been estimated 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 a mere two decades ago.

Another class of small RNAs are called **small interfering RNAs (siRNAs)**. These are similar in size and function to miRNAs—both can associate with the same proteins, producing similar results. The distinction between miRNAs and siRNAs is based on subtle differences in the structure of their double-stranded RNA precursor molecules. If researchers inject siRNA precursor molecules into a cell, the cell's machinery can process them into siRNAs that turn off expression of genes with related sequences. The blocking of gene expression by siRNAs is called **RNA interference (RNAi)**; it 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 Chapter 17, some viruses have double-stranded RNA genomes. Because the cellular RNAi pathway can process double-stranded RNAs into homing devices that lead to the destruction of RNAs with complementary sequences, 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. Moreover, many species, including mammals, apparently produce their own long, double-stranded RNA precursors to small RNAs such as siRNAs. Once produced, these RNAs can interfere with gene expression at stages other than translation, as we'll discuss next.



▲ Figure 15.13 Regulation of gene expression by miRNAs. A 22-nucleotide miRNA, formed by enzymatic processing of an RNA precursor, associates with one or more proteins in a complex that can affect target mRNAs.

Chromatin Remodeling and Effects on Transcription by ncRNAs

In addition to affecting mRNAs, small RNAs can cause remodeling of chromatin structure. In some yeasts, siRNAs produced by the yeast cells themselves are required for the formation of heterochromatin at the centromeres of chromosomes. According to one model, an RNA transcript produced from DNA in the centromeric region of the chromosome is copied into double-stranded RNA by a yeast enzyme and then processed into siRNAs. These siRNAs associate with a complex of proteins (different from the one shown in Figure 15.13) and act as a homing device, targeting the complex back to RNA transcripts being made from the centromeric sequences of DNA. Once there, proteins in the complex recruit enzymes that modify the chromatin, turning it into the highly condensed heterochromatin found at the centromere.

A newly discovered class of small ncRNAs called *piwi-associated RNAs* (*piRNAs*) also induce formation of heterochromatin, blocking expression of some parasitic DNA elements in the genome known as transposons. (Transposons are discussed in Chapter 18.) Usually 24–31 nucleotides in length, piRNAs are probably processed from single-stranded RNA precursors. 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.

The role of ncRNAs in regulation of gene expression adds yet another layer to the complex and intricate process described in the previous section. As more is learned about the multiple, interacting ways a cell can fine-tune expression of its genes, the goal is to understand how a specific set of genes is expressed in a particular cell. In the next section, we'll describe a few methods that researchers use to monitor expression of specific genes.

CONCEPT CHECK 15.3

1. **WHAT IF?** If the mRNA being degraded in Figure 15.13 coded for a protein that promotes cell division in a multicellular organism, what would happen if a mutation disabled the gene encoding the miRNA that triggers this degradation?
2. **MAKE CONNECTIONS** Inactivation of one of the X chromosomes in female mammals results in a Barr body (see Concept 12.2). Suggest a model for how the *XIST* noncoding RNA functions to cause Barr body formation.

For suggested answers, see Appendix A.

CONCEPT 15.4

Researchers can monitor expression of specific genes

The diverse mechanisms of regulating gene expression discussed in this chapter underlie one basic generality: Cells of a given multicellular organism differ from each other because they express different genes from an identical genome. 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. Techniques related to those developed for genetic engineering (see Concept 13.4) are widely used to track expression of mRNAs. In this section we'll first examine techniques that look for patterns of expression of specific individual genes. Next, we'll explore techniques that characterize groups of genes being expressed by cells or tissues of interest. As you will see, all of these techniques depend in some way on base pairing between complementary nucleotide sequences.

Studying the Expression of Single Genes

Suppose we have cloned a gene that may play an important role in the embryonic development of *Drosophila* (the fruit fly). The first thing we might want to know is which embryonic cells express the gene—in other words, where in the embryo is the corresponding mRNA found? We can detect the mRNA using the technique of **nucleic acid hybridization**, the base pairing of one strand of a nucleic acid to the complementary sequence on another strand. The complementary molecule, a short single-stranded nucleic acid that can be either RNA or DNA, is called a **nucleic acid probe**. Using our cloned gene as a template, we can synthesize a probe complementary to the mRNA. For example, if part of the sequence on the mRNA were

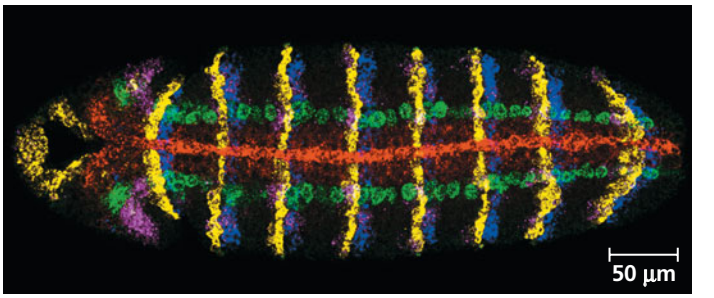
5′ ···CUCAUCACCGGC··· 3′

then we would synthesize this single-stranded DNA probe:

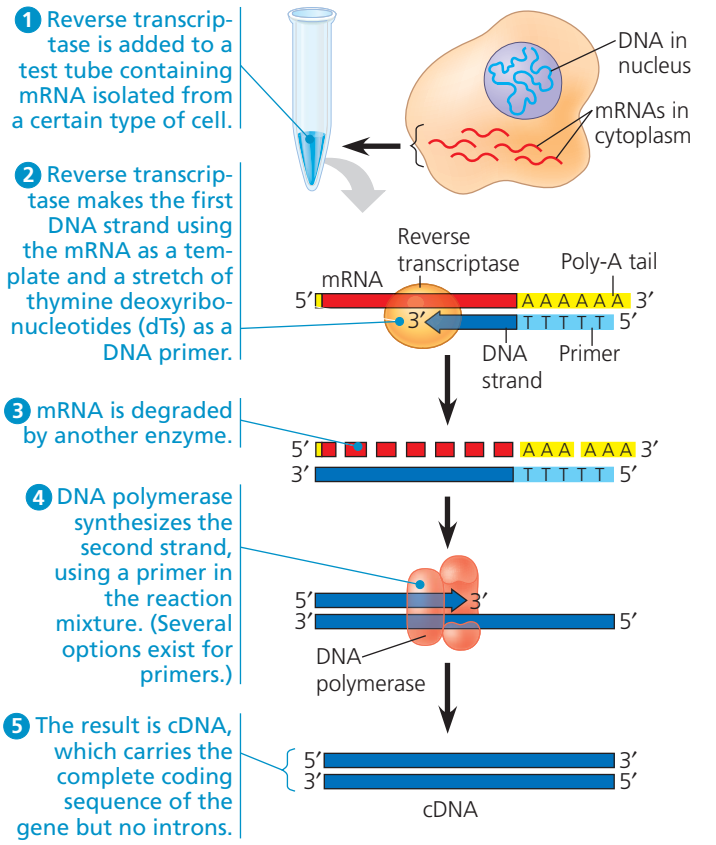
3′ GAGTAGTGGCCG 5′

Each probe molecule is labeled during synthesis with a fluorescent tag so we can follow it. A solution with the probe is applied to *Drosophila* embryos, allowing probe molecules to hybridize specifically to any complementary sequences on the many mRNAs in embryonic cells that are transcribing the gene. Because this technique allows us to see the mRNA in place (or *in situ*) in the intact organism, this technique is called ***in situ* hybridization**. Different probes can be labeled with different fluorescent dyes, sometimes with strikingly beautiful results **(Figure 15.14)**.

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 (RT-PCR)**. RT-PCR begins by turning sample sets of mRNAs into double-stranded DNAs with the corresponding sequences. This feat is accomplished by an enzyme called *reverse transcriptase*, isolated in the late 1980s from a type of virus called a retrovirus. (You'll learn more about retroviruses, including HIV, in Chapter 17.) Reverse transcriptase is able to synthesize a complementary DNA copy of an mRNA, thus making a *reverse transcript* **(Figure 15.15)**. Recall that the 3′



▲ Figure 15.14 Determining where genes are expressed by *in situ* hybridization analysis. This *Drosophila* embryo was incubated in a solution containing 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 cells in which a specific gene is expressed as mRNA.



▲ Figure 15.15 Making complementary DNA (cDNA) from eukaryotic genes. Complementary DNA is DNA made *in vitro* using mRNA as a template for the first strand. Although only one mRNA is shown here, the final collection of cDNAs would reflect all the mRNAs that were present in the cell.

end of the mRNA has a stretch of adenine (A) ribonucleotides called a poly-A tail. This feature allows use of a short strand of thymine deoxyribonucleotides (dT's) as a primer for the reverse transcriptase. Following enzymatic degradation of the mRNA, a second DNA strand, complementary to the first, is synthesized by DNA polymerase. The resulting double-stranded DNA is called **complementary DNA (cDNA)**, and the reverse transcription step accounts for the "RT" in the name RT-PCR. 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 then make cDNA from each stage **(Figure 15.16)**.

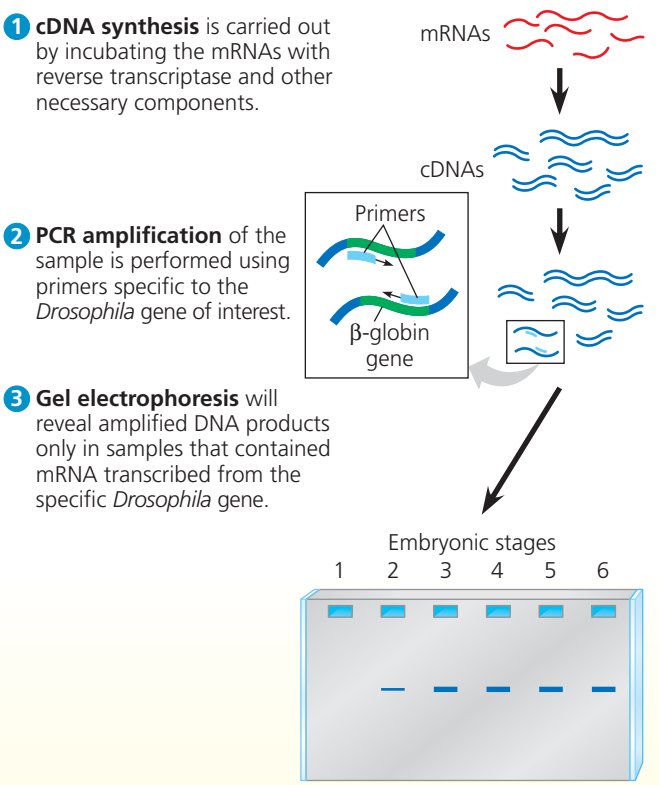
Next in RT-PCR is the PCR step (see Figure 13.25). As you may recall, PCR is a way of rapidly making many copies of one specific stretch of double-stranded DNA, using primers that hybridize to the opposite ends of the region of interest. In our case, we would add primers corresponding to a region of our *Drosophila* gene, using the cDNA from each sample as a template for PCR amplification. When the products are run on a gel, copies of the amplified region will be observed as bands only in samples that originally contained mRNA from the gene we're focusing on. RT-PCR can also be carried out with mRNAs collected from different tissues at one time to discover which tissue is producing a specific mRNA.

▼ Figure 15.16 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 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 processed as shown below. (The mRNAs from only one stage are shown here.)



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.

Studying the Expression of Groups of Genes

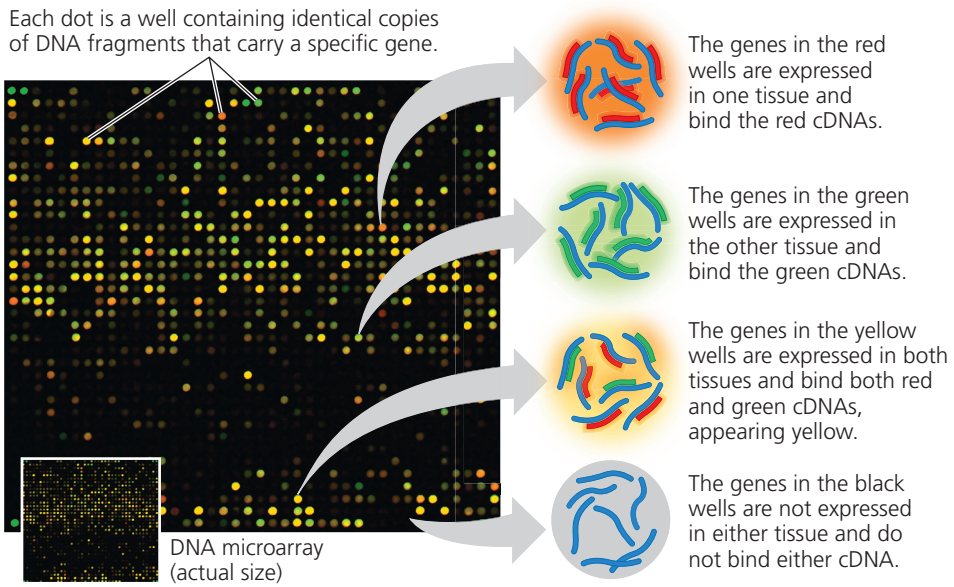
A major goal of biologists is to learn how genes act together to produce and maintain a functioning organism. Now that the entire genomes of a number of organisms have been sequenced, it is possible to study the expression of large groups of genes—the so-called systems approach. Researchers use what is known about the whole genome to investigate which groups of genes are transcribed in different tissues or at different stages of development. One of their aims is to identify networks of gene expression across an entire genome.

Genome-wide expression studies can be carried out using **DNA microarray assays**. 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. (The microarray is also called a *DNA chip* by analogy to a computer chip.) Ideally, these fragments represent all the genes in the genome of an organism.

The basic strategy in such studies is to isolate the mRNAs made in a cell of interest and use these mRNAs as templates for making the corresponding cDNAs by reverse transcription. In microarray assays, these cDNAs are labeled with fluorescent molecules and then allowed to hybridize to a microarray slide. Most often, the cDNAs from two samples are labeled with molecules that emit different colors and tested on the same microarray. **Figure 15.17** shows the result of such an experiment, identifying the subsets of genes in the genome that are being expressed in one tissue compared with another. DNA technology makes such studies possible; with automation, they are easily performed on a large scale. Scientists can now measure the expression of thousands of genes at one time.

Alternatively, with the advent of rapid, inexpensive DNA sequencing methods (see Chapter 13), researchers can now afford to simply sequence the cDNA samples from different tissues or different embryonic stages in order to discover which genes are expressed. This straightforward method is called *RNA sequencing* or *RNA-seq*, even though it is the cDNA that is actually sequenced. As the price of sequencing plummets, this method is growing more widespread.

By characterizing sets of genes that are expressed together in some tissues but not others, genome-wide gene expression studies 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



▲ Figure 15.17 DNA microarray assay of gene expression levels. Researchers synthesized two sets of cDNAs, fluorescently labeled red or green, from mRNAs from two different human tissues. These cDNAs were hybridized with a microarray containing 5,760 human genes (about 25% of human genes), resulting in the pattern shown here. The intensity of fluorescence at each spot measures the relative expression in the two samples of the gene represented by that spot: Red indicates expression in one sample, green in the other, yellow in both, and black in neither.

more informed and effective treatment protocols. Ultimately, information from genome-wide studies should provide a grander view of how ensembles of genes interact to form an organism and maintain its vital systems. The genetic basis of embryonic development and disease will be considered in the next chapter.

CONCEPT CHECK 15.4

1. Describe the role of complementary base pairing during RT-PCR and microarray analysis.
2. **WHAT IF?** Consider the microarray in Figure 15.17. 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.

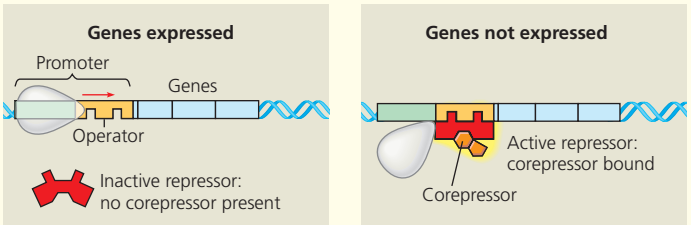
15 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 15.1

Bacteria often respond to environmental change by regulating transcription (pp. 293–298)

- In bacteria, certain groups of genes are clustered into an operon with a single promoter. 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. 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, the repressor is active when bound to a **corepressor**.



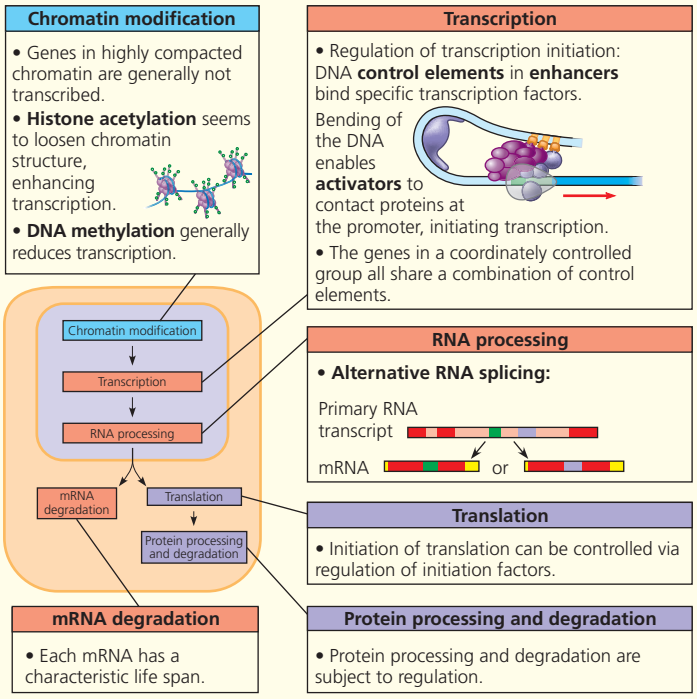
In an inducible operon, binding of an **inducer** to an innately active repressor inactivates the repressor and turns on transcription. Inducible enzymes usually function in catabolic pathways.

- Some operons have positive gene regulation. A stimulatory **activator** protein (such as CAP, when activated by **cyclic AMP**), binds to a site within the promoter and stimulates transcription.

? *Compare and contrast the roles of the corepressor and the inducer in negative regulation of an operon.*

CONCEPT 15.2

Eukaryotic gene expression is regulated at many stages (pp. 298–305)



? *Describe what must happen for a cell type–specific gene to be transcribed in a cell of that type.*

CONCEPT 15.3

Noncoding RNAs play multiple roles in controlling gene expression (pp. 305–306)

- Noncoding RNAs (e.g., **miRNAs** and **siRNAs**) can block translation or cause degradation of mRNAs.

? *Why are miRNAs called noncoding RNAs? Explain how they participate in gene regulation.*

CONCEPT 15.4

Researchers can monitor expression of specific genes (pp. 307–309)

- In **nucleic acid hybridization**, a **nucleic acid probe** is used to detect the presence of a specific mRNA.
- ***In situ*** **hybridization** and **RT-PCR** can detect the presence of a given mRNA in a tissue or an RNA sample, respectively.
- **DNA microarrays** are used to identify sets of genes co-expressed by a group of cells. Their cDNAs can also be sequenced.

? *What useful information is obtained by detecting expression of specific genes?*

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

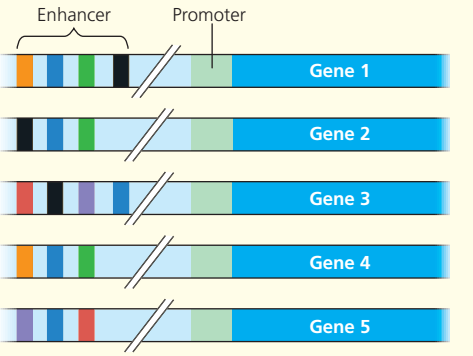
1. If a particular operon encodes enzymes for making an essential amino acid and is regulated like the *trp* operon, then
   a. the amino acid inactivates the repressor.
   b. the enzymes produced are called inducible enzymes.
   c. the repressor is active in the absence of the amino acid.
   d. the amino acid acts as a corepressor.
   e. the amino acid turns on transcription of the operon.
2. The functioning of enhancers is an example of
   a. transcriptional control of gene expression.
   b. a post-transcriptional mechanism to regulate mRNA.
   c. the stimulation of translation by initiation factors.
   d. post-translational control that activates certain proteins.
   e. a eukaryotic equivalent of prokaryotic promoter functioning.
3. 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. the binding of RNA polymerase to transcription factors
   e. the folding of DNA to form heterochromatin

Level 2: Application/Analysis

4. 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
   e. overproduction of catabolite activator protein (CAP)
5. 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. Each gene lies immediately adjacent to an enhancer.
   d. Many genes are grouped into operon-like clusters.
   e. It is the same as the DNA in one of your kidney cells.
6. Which of the following would *not* be true of cDNA produced using human brain tissue as the starting material?
   a. It could be amplified by the polymerase chain reaction.
   b. It would contain sequences representing all the genes in the genome.
   c. It was produced from mRNA using reverse transcriptase.
   d. It could be used as a probe to detect genes expressed in the brain.
   e. It lacks the introns of the human genes.

Level 3: Synthesis/Evaluation

7. **DRAW IT** The diagram to the right shows five genes, including their enhancers, from the genome of a certain species. Imagine that orange, 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. 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 orange activators are present. 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. Which activators would have to be present in each cell type to ensure transcription of the appropriate genes?
8. **SCIENTIFIC INQUIRY**
   Imagine you want to study one of the mouse crystallins, proteins present in the lens of the eye. Assuming the gene has been cloned, describe two ways you could investigate expression of this gene in the developing mouse embryo.
9. **FOCUS ON EVOLUTION**
   DNA sequences can act as "tape measures of evolution" (see Concept 3.6). 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.
10. **FOCUS ON INTERACTIONS**
    In a short essay (100–150 words), discuss how the processes shown in Figure 15.2 are examples of feedback mechanisms regulating biological systems in bacterial cells.

For selected answers, see Appendix A.

16 Development, Stem Cells, and Cancer

KEY CONCEPTS

16.1 A program of differential gene expression leads to the different cell types in a multicellular organism

16.2 Cloning of organisms showed that differentiated cells could be "reprogrammed" and ultimately led to the production of stem cells

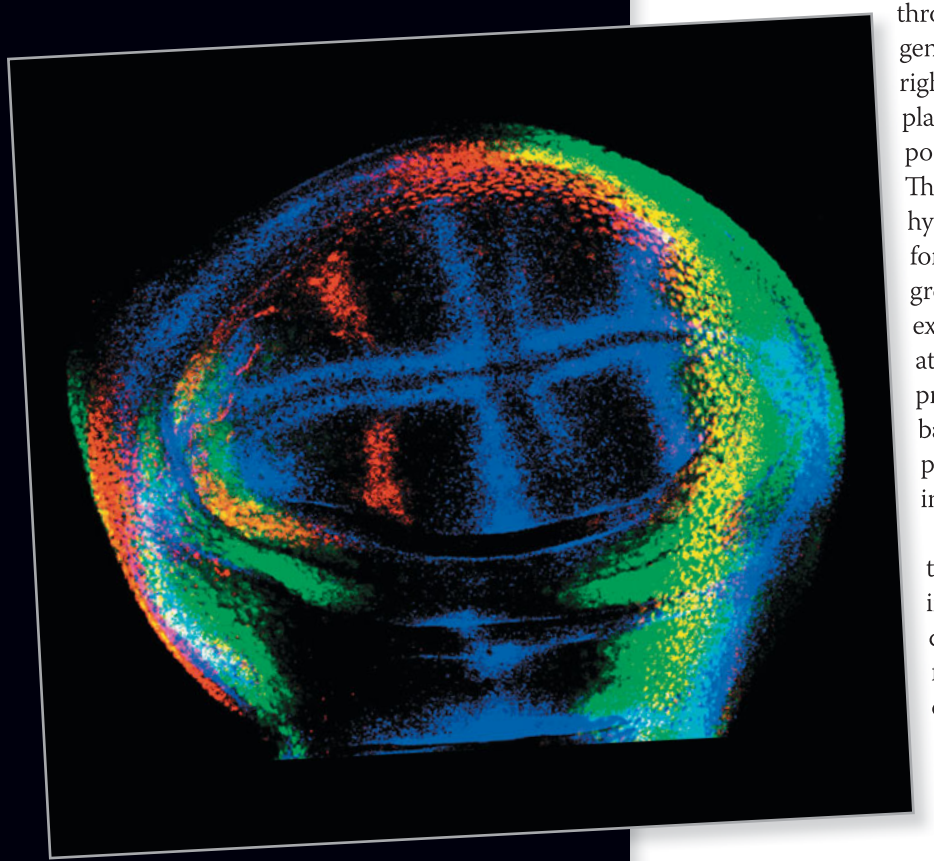
16.3 Abnormal regulation of genes that affect the cell cycle can lead to cancer

OVERVIEW

Orchestrating Life's Processes

The development of the fertilized egg, a single cell, into an embryo and later an adult is an astounding transformation that requires a precisely regulated program of gene expression. All of the levels of eukaryotic gene regulation you learned about in the previous chapter come into play during embryonic development. The elaborate sequence of genes being turned on and off in different cells is the ultimate example of regulation of gene expression.

Understanding the genetic underpinnings of development has progressed mainly by studying the process in **model organisms**, species that are easy to raise in the lab and use in experiments. A prime example is the fruit fly *Drosophila melanogaster*. An adult fruit fly develops from a fertilized egg, passing through a wormlike stage called a larva. At every stage, gene expression is carefully regulated, ensuring that the right genes are expressed only at the correct time and place. In the larva, the adult wing forms in a disk-shaped pocket of several thousand cells, shown in **Figure 16.1**. The tissue in this image has been analyzed by *in situ* hybridization (see Figure 15.14) to reveal the mRNA for three genes—labeled red, blue, and green. (Red and green together appear yellow.) The intricate pattern of expression for each gene is the same from larva to larva at this stage, and it provides a graphic display of the precision of gene regulation. But what is the molecular basis for this pattern? Why is one particular gene expressed only in the few hundred cells that appear blue in this image and not in the other cells?



▼ **Figure 16.1** What regulates the precise pattern of gene expression in the developing wing of a fly embryo?

Part of the answer involves the transcription factors and other regulatory molecules you learned about in the previous chapter. But how do they come to be different in distinct cell types? In this chapter, we'll first explain the mechanisms that send cells down diverging genetic pathways to adopt different fates. Then we'll take a closer look at *Drosophila* development. Next, we'll describe the discovery of stem cells, a powerful cell type that is key to the developmental process. These cells offer hope for medical treatments as well. Finally, after having explored embryonic development and stem cells, we will underscore the crucial role played by regulation of gene expression by investigating how cancer can result when this regulation goes awry. Orchestrating proper gene expression by all cells is crucial to the functions of life.

CONCEPT 16.1

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. Here, we'll focus on the program of regulation of gene expression that orchestrates development using a few animal species as examples.

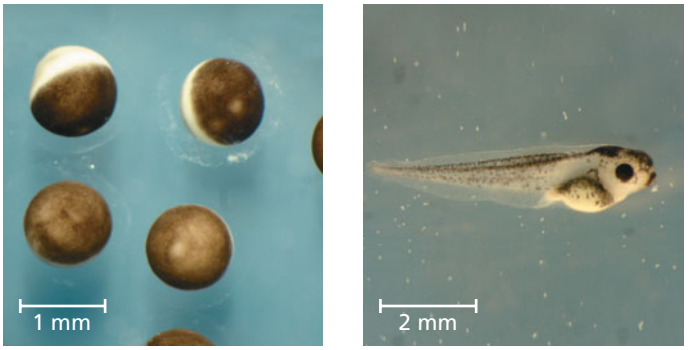
A Genetic Program for Embryonic Development

The photos in **Figure 16.2** illustrate the dramatic difference between a zygote and the organism 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 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 have their basis in cellular behavior. Even morphogenesis, the shaping of the organism, can be traced back to changes in the motility, shape, 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 15.11, 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 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 the mother set up a sequential program of gene regulation that is carried out as cells divide, and this program makes the cells become different from each other in a coordinated fashion. To understand how this works, we'll 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.



(a) Fertilized eggs of a frog (b) Newly hatched tadpole

▲ **Figure 16.2 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).

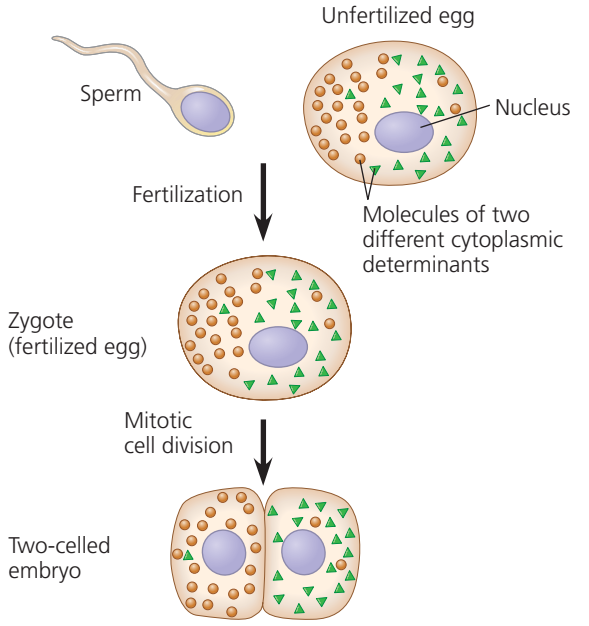
Cytoplasmic Determinants and Inductive Signals

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 RNA, 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 16.3a)**. 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.

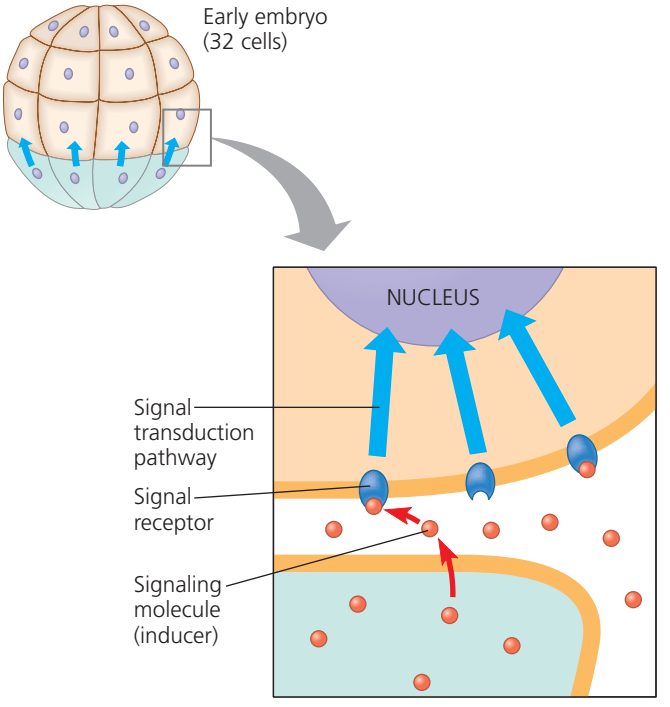
▼ Figure 16.3 Sources of developmental information for the early embryo.

(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 sets of cytoplasmic determinants and, as a result, express different genes.

(b) Induction by nearby cells



The cells at the bottom of the early embryo depicted here are releasing chemicals that signal 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 communicated 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 5.6). Such signals cause changes in the target cells, a process called **induction** **(Figure 16.3b)**. The molecules conveying these signals within the target cell are cell-surface receptors and other proteins expressed by the embryo's own genes. In general, the signaling molecules send a cell down a specific developmental path by causing changes in its gene expression that eventually result in observable cellular changes. Thus, interactions between embryonic cells help induce differentiation of the many specialized cell types making up a new organism.

Sequential Regulation of Gene Expression during Cellular Differentiation

As the tissues and organs of an embryo develop and their cells differentiate, the cells become noticeably different in structure and function. These observable changes are actually the outcome of a cell's developmental history, which begins at the first mitotic division of the zygote, as we have just seen. The earliest changes that set a cell on a 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 unseen events that lead to the observable differentiation of a cell. Once it has undergone determination, an embryonic cell is irreversibly committed to its final fate. If a committed cell is experimentally placed in another location in the embryo, it will still differentiate into the cell type that is its normal fate.

Differentiation of Cell Types

Today we understand determination in terms of molecular changes. The outcome of determination, observable cell differentiation, is marked by the expression of genes for *tissue-specific proteins*. These proteins are found only in a specific cell type and give the cell its characteristic structure and function. The first evidence of differentiation is the appearance of mRNAs for these proteins. Eventually, 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 from division of their precursors. A number of the steps in gene expression may be regulated during differentiation, with transcription among the most important. At the end of the process, in the fully differentiated cell, transcription remains the principal regulatory point for maintaining appropriate gene expression.

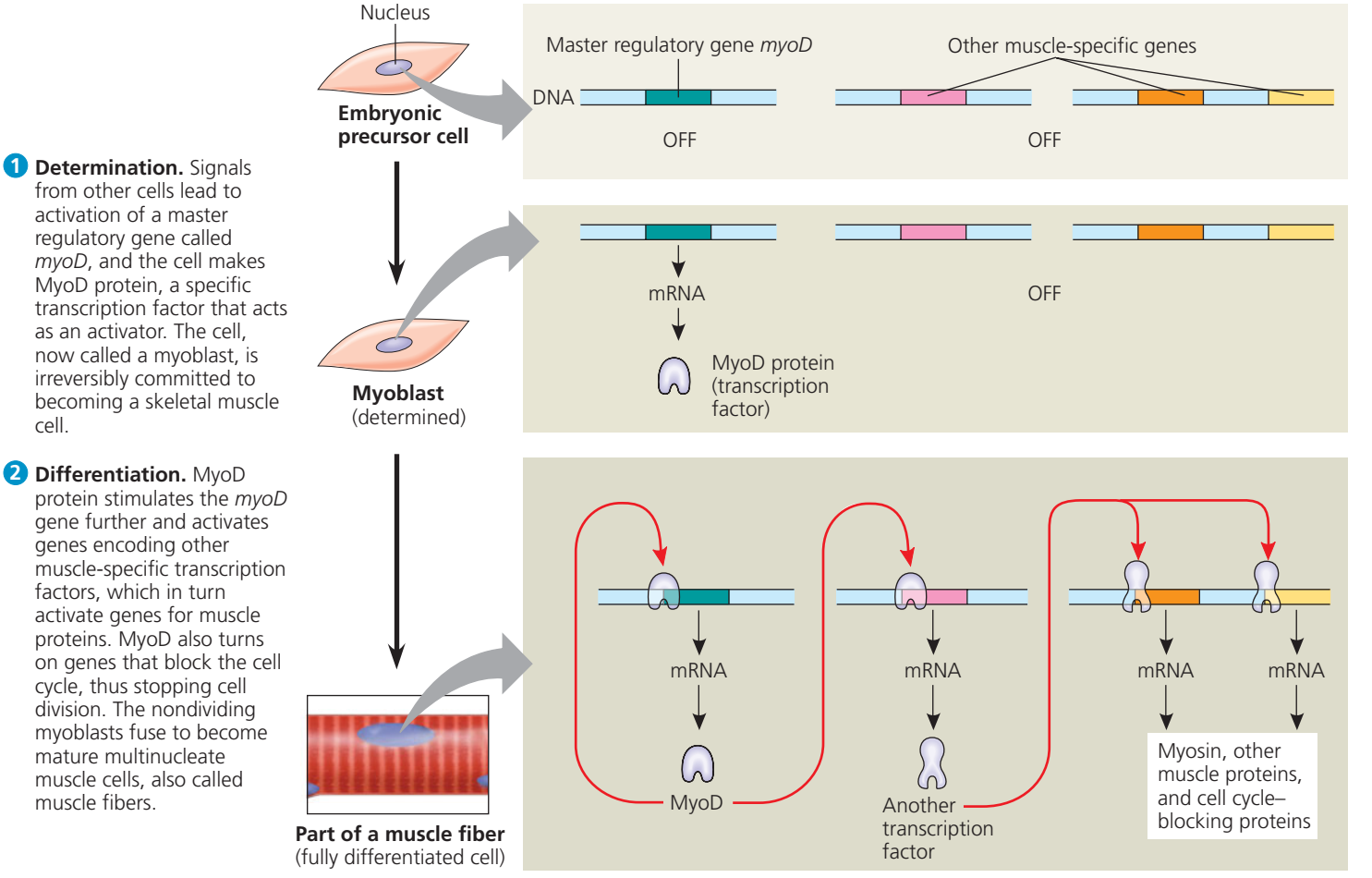
Differentiated cells are specialists at making tissue-specific proteins. For example, as a result of transcriptional regulation, liver cells specialize in making albumin, and lens cells

specialize in making crystallin (see Figure 15.11). 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 *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 **(Figure 16.4)**. To do so, they grew embryonic precursor cells in culture and analyzed them using molecular biological techniques like those described in Chapters 13 and 15. They isolated different genes one by one, caused each to be expressed in a separate 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. 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 commitment occurs in muscle cell differentiation, let's focus on the master regulatory gene called *myoD* (see Figure 16.4). This gene encodes MyoD protein, a transcription factor that binds to specific control elements in the enhancers of various target genes and stimulates their expression (see Figure 15.9). Some target genes for MyoD encode still other muscle-specific transcription factors. MyoD also stimulates expression of the *myoD* gene itself, thus perpetuating its effect in maintaining the cell's differentiated state. Since all the genes activated by MyoD have enhancer control elements recognized by MyoD, they are coordinately controlled. Finally, the secondary transcription factors activate the genes



▲ **Figure 16.4 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 a MyoD protein that could not activate the* myoD *gene?*

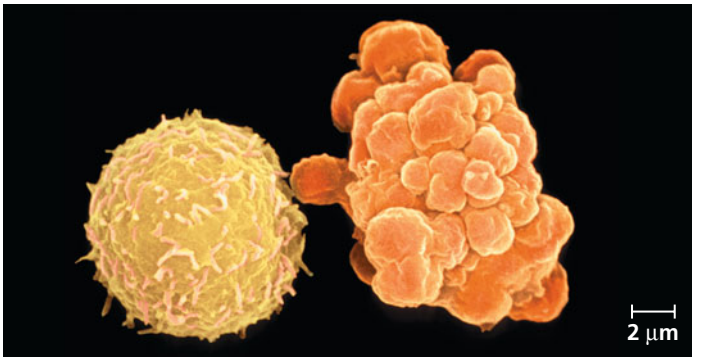
for proteins such as myosin and actin that confer the unique properties of skeletal muscle cells. The MyoD protein deserves its designation as a master regulatory gene.

The regulation of genes that play important roles in development of embryonic tissues and structures is often complex. In the **Scientific Skills Exercise**, you'll work with data from an experiment that tested how different regulatory regions in the DNA affect expression of a gene that helps establish the pattern of the different digits in a mouse's paw.

Apoptosis: A Type of Programmed Cell Death

During the time when most cells are differentiating, some cells in the developing organism are genetically programmed to die. The best-understood type of "programmed cell death" is **apoptosis** (from the Greek, meaning "falling off," and used in a classic Greek poem to refer to leaves falling from a tree). Apoptosis also occurs in cells of the mature organism that are infected, damaged, or have reached the end of their functional life span. During this process, cellular agents chop up the DNA and fragment the organelles and other cytoplasmic components. The cell becomes multilobed, a change called "blebbing" **(Figure 16.5)**, and the cell's parts are packaged up in vesicles. These "blebs" are then engulfed by scavenger cells, leaving no trace. Apoptosis protects neighboring cells from damage that they would otherwise suffer if a dying cell merely leaked out all its contents, including its many digestive enzymes.

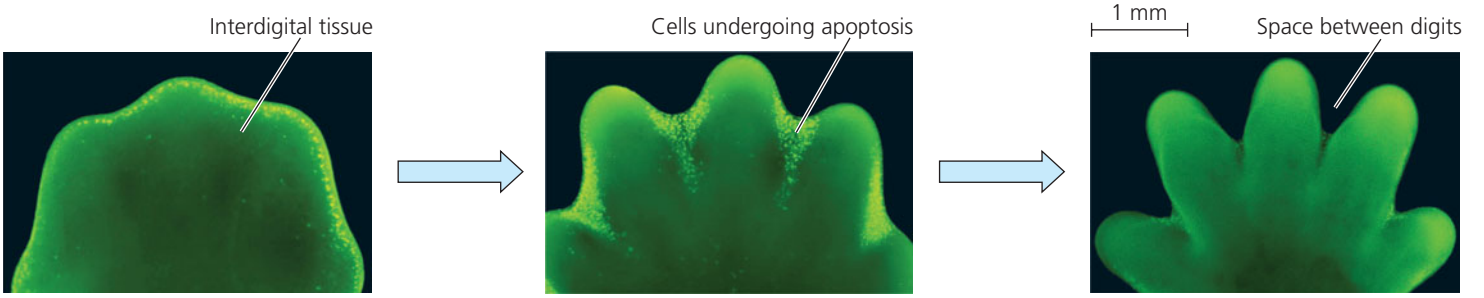
Apoptosis plays a crucial role in the developing embryo. The molecular mechanisms underlying apoptosis were worked out in detail by researchers studying embryonic development of a small soil worm, a nematode called *Caenorhabditis elegans* that has now become a popular model organism for genetic studies. Because the adult worm has only about a thousand cells, the researchers were able to work out the complete ancestry of each cell. The timely suicide of cells occurs exactly 131 times during normal development of *C. elegans*, at precisely the same points in the cell lineage of each worm. In worms and other species, apoptosis is triggered by signal transduction pathways (see Figure 5.20). These activate a cascade of apoptotic "suicide" proteins in the cells destined to die, including the enzymes that break down and package cellular molecules in the "blebs."



▲ **Figure 16.5 Apoptosis of a human white blood cell.** We can compare a normal white blood cell (left) with a white blood cell undergoing apoptosis (right). The apoptotic cell is shrinking and forming lobes ("blebs"), which eventually are shed as membrane-enclosed cell fragments (colorized SEMs).

Apoptosis is essential to development and maintenance in all animals. There are similarities in genes encoding apoptotic proteins in nematodes and mammals, and apoptosis is known to occur as well in multicellular fungi and single-celled yeasts, evidence that the basic mechanism evolved early among eukaryotes. In vertebrates, apoptosis is essential for normal development of the nervous system and for normal morphogenesis of hands and feet in humans and paws in other mammals **(Figure 16.6)**. The level of apoptosis between the developing digits is lower in the webbed feet of ducks and other water birds than in the nonwebbed feet of land birds, such as chickens. In the case of humans, the failure of appropriate apoptosis can result in webbed fingers and toes.

We have seen how different programs of gene expression that are activated in the fertilized egg can result in differentiated cells and tissues as well as the death of some cells. But for tissues to function properly 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. Next we'll look at the molecular basis for establishing the body plan, using the well-studied *Drosophila* as an example.

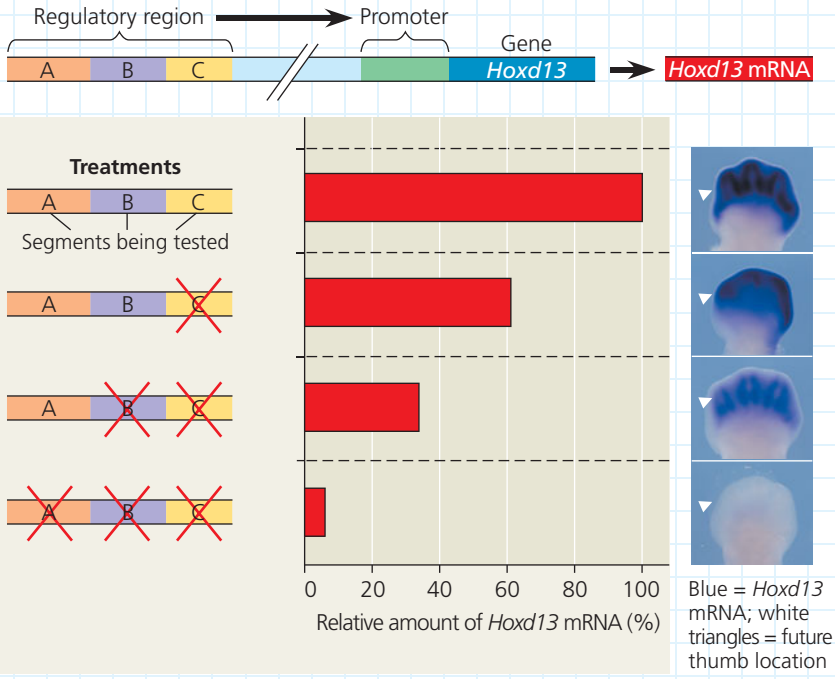


▲ **Figure 16.6 Effect of apoptosis during paw development in the mouse.** In mice, humans, other mammals, and land birds, the embryonic region that develops into feet or hands initially has a solid, platelike structure. Apoptosis eliminates the cells in the interdigital regions, thus forming the digits. The embryonic mouse paws shown in these fluorescence light micrographs are stained so that cells undergoing apoptosis appear a bright yellowish green. Apoptosis of cells begins at the margin of each interdigital region (left), peaks as the tissue in these regions is reduced (middle), and is no longer visible when the interdigital tissue has been eliminated (right). (Note that the Scientific Skills Exercise shows a different genetic process involved in mouse paw development.)

Analyzing Quantitative and Spatial Gene Expression Data

How Is a Particular *Hox* Gene Regulated During Paw Development? *Hox* genes code for transcription factor proteins, which in turn control sets of genes important for animal development (see Concept 18.6 for more information on *Hox* genes). One group of *Hox* genes, the *Hoxd* genes, plays a role in establishing the pattern of the different digits (fingers and toes) at the end of a limb. Unlike the *mPGES-1* gene mentioned in the last chapter, *Hox* genes have very large, complicated regulatory regions, including control elements that may be hundreds of kilobases (kb; thousands of nucleotides) away from the gene.

In cases like this, how do biologists narrow down the segments that contain important elements? They begin by removing (deleting) large segments of DNA and studying the effect on gene expression. In this exercise, you'll compare data from two different but complementary approaches that look at the expression of a specific *Hoxd* gene (*Hoxd13*). One approach quantifies overall expression; the other approach is less quantitative but gives important spatial localization information.



How the Experiment Was Done Researchers interested in the regulation of *Hoxd13* gene expression genetically engineered a set of mice (*transgenic* mice) that had different segments of DNA deleted upstream of the gene. They then compared levels and patterns of *Hoxd13* gene expression in the developing paws of 12.5-day-old transgenic mouse embryos with those seen in wild-type mouse embryos of the same age.

They used two different approaches: In some mice, they extracted the mRNA from the embryonic paws and quantified the overall level of *Hoxd13* mRNA in the whole paw. In another set of the same transgenic mice, they used *in situ* hybridization (see Concept 15.4) to pinpoint exactly where in the paws the *Hoxd13* gene was expressed as mRNA. The particular technique that was used causes the *Hoxd13* mRNA to appear blue.

Data from the Experiment The top diagram (upper right) depicts the very large regulatory region upstream of the *Hoxd13* gene. The area between the slashes represents the DNA located between the promoter and the regulatory region.

The diagrams to the left of the bar graph show, first, the intact DNA (830 kb) and, next, the three altered DNA sequences. (Each is called a "deletion," since a particular section of DNA has been deleted from it.) A red X indicates the segment (A, B, and/or C) that was deleted in each experimental treatment.

The horizontal bar graph shows the amount of *Hoxd13* mRNA that was present in the digit-formation zone of each mutant 12.5-day-old embryo paw relative to the amount that was in the digit-formation zone of the mouse that had the intact regulatory region (top bar = 100%).

The images on the right are the embryo paws showing the location of the *Hoxd13* mRNA (blue stain). The white triangles show the location where the thumb will form.

Interpret the Data

1. The researchers hypothesized that all three regulatory segments (A, B, and C) were required for full expression of the *Hoxd13* gene. By measuring the amount of *Hoxd13* mRNA in the embryo paw zones where digits will develop, they could measure the effect of the regulatory segments singly and in combination. Refer to the graph to answer these questions, noting that the segments being tested are shown on the vertical axis and the relative amount of *Hoxd13* mRNA is shown on the horizontal axis. (a) Which of the four treatments was used as a control for the experiment? (b) Their hypothesis is that all three segments are required for highest expression of the *Hoxd13* gene. Is this supported by their results? Explain your answer.
2. (a) What is the effect on the amount of *Hoxd13* mRNA when segments B and C are both deleted, compared with the control? (b) Is this effect visible in the blue-stained regions of the *in situ* hybridizations? How would you describe the spatial pattern of gene expression in the embryo paws that lack segments B and C?
3. (a) What is the effect on the amount of *Hoxd13* mRNA when just segment C is deleted, compared with the control? (b) Is this effect visible in the *in situ* hybridizations? How would you describe the spatial pattern of gene expression in embryo paws that lack just segment C, compared with the control and with the paws that lack segments B and C?
4. If the researchers had only measured the amount of *Hoxd13* mRNA and not done the *in situ* hybridizations, what important information about the role of the regulatory segments in *Hoxd13* gene expression during paw development would have been missed? Conversely, if the researchers had only done the *in situ* hybridizations, what information would have been inaccessible?

Data from T. Montavon et al., A regulatory archipelago controls *Hox* genes transcription in digits, *Cell* 147:1132–1145 (2011). doi 10.1016/j.cell.2011.10.023

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Pattern Formation: Setting Up the Body Plan

Cytoplasmic determinants and inductive signals both contribute to the development of a spatial organization in which the tissues and organs of an organism are all in their characteristic places. This process is called **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 symmetric 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 tissues and organs appear. The molecular cues that control pattern formation, collectively called **positional information**, are provided by cytoplasmic determinants and inductive signals (see Figure 16.3). These cues tell a cell its location relative to the body axes and to neighboring cells and determine how the cell and its progeny will respond to future molecular signals.

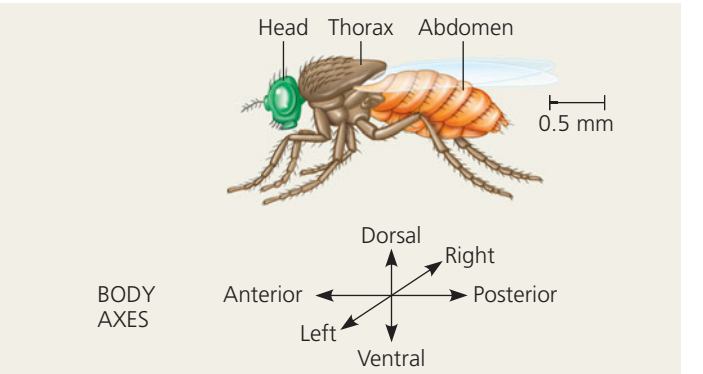
During the first half of the 20th century, classical embryologists made detailed anatomical observations of embryonic development in a number of species and performed experiments in which they manipulated embryonic tissues. This research laid the groundwork for understanding the mechanisms of development, but 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 and continues today. Genetic 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.

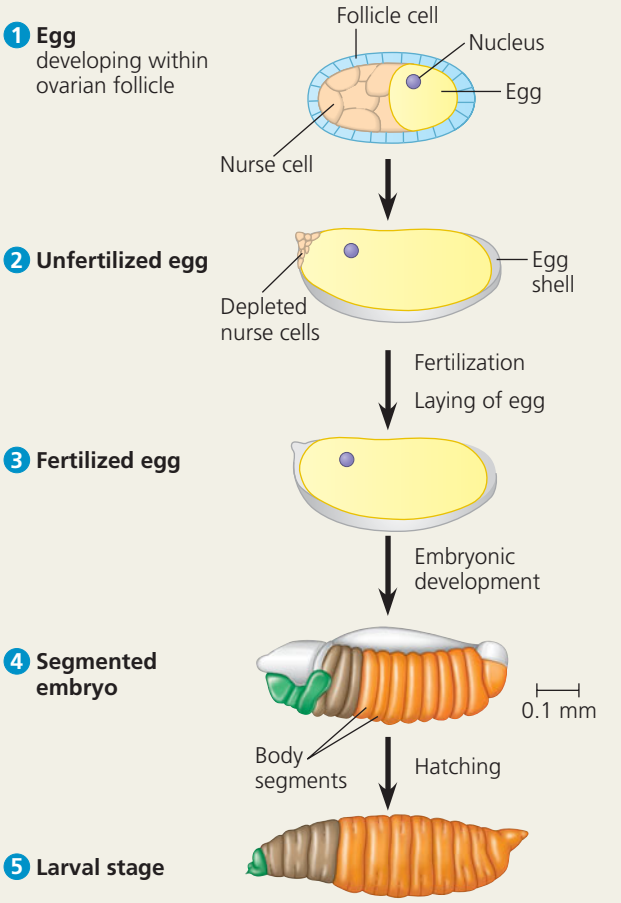
The Life Cycle of Drosophila

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 midbody, from which the wings and legs extend), and the abdomen **(Figure 16.7a)**. Like other bilaterally symmetric 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 the female's ovary, surrounded by ovarian cells called nurse cells and follicle cells **(Figure 16.7b**, top). These support cells supply the egg with



(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. 1 The egg (yellow) is surrounded by other cells that form a structure called the follicle within one of the mother's ovaries. 2 The nurse cells shrink as they supply nutrients and mRNAs to the developing egg, which grows larger. Eventually, the mature egg fills the egg shell that is secreted by the follicle cells. 3 The egg is fertilized within the mother and then laid. 4 Embryonic development forms 5 a larva, which goes through three stages. The third stage forms a cocoon (not shown), within which the larva metamorphoses into the adult shown in (a).

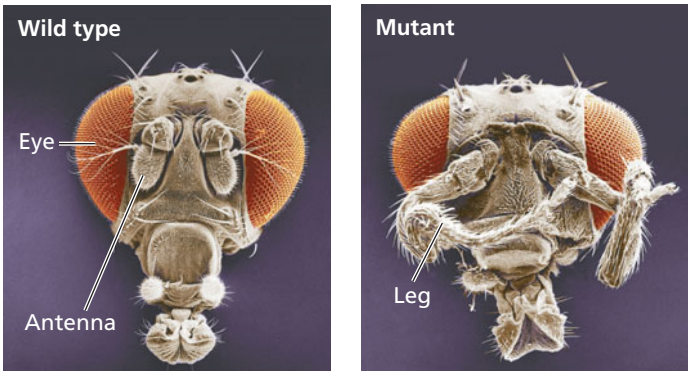
▲ Figure 16.7 Key developmental events in the life cycle of *Drosophila*.

nutrients, mRNAs, and other substances needed for development, and make the egg shell. 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 cocoon in which it metamorphoses into the adult fly pictured in Figure 16.7a.

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 16.8)**. 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**, control pattern formation in the late embryo, larva, and adult.

Insight into pattern formation during early embryonic development did not come for another 30 years, 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* protein-coding genes, now known to total about 13,900. 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. Since 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.



▲ Figure 16.8 Abnormal pattern formation in *Drosophila*. Mutations in certain regulatory genes, called homeotic genes, cause a misplacement of structures in an animal. These scanning electron micrographs contrast the head of a wild-type fly, bearing a pair of small antennae, with that of a homeotic mutant (a fly with a mutation in a single gene), bearing a pair of legs in place of antennae.

Nüsslein-Volhard and Wieschaus began their search for segmentation genes by exposing flies to a mutagenic chemical that affected the flies' gametes. They mated the mutagenized flies and then scanned 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. Over several years, the researchers were able to group these segmentation genes by general function, to map them, and to clone many of them for further study in the lab. 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 further the genes that Nüsslein-Volhard, Wieschaus, and co-workers found for cytoplasmic determinants deposited in the egg by the mother. These genes set up the initial pattern of the embryo by regulating gene expression in broad regions of the early embryo.

Axis Establishment

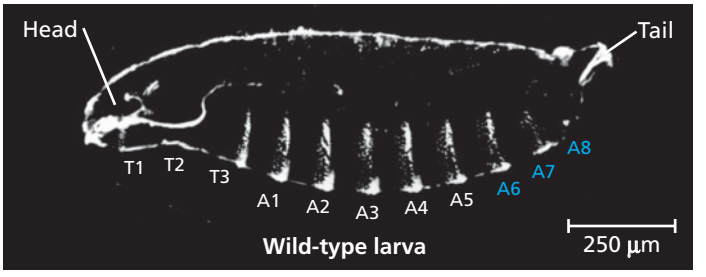
As we mentioned earlier, cytoplasmic determinants in the egg are the substances that initially establish the axes of the *Drosophila* body. These substances are encoded by genes of the mother, fittingly called maternal effect genes. A **maternal effect gene** is a gene that, when mutant in the mother, 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 defective; when these eggs are fertilized, they fail to develop properly.

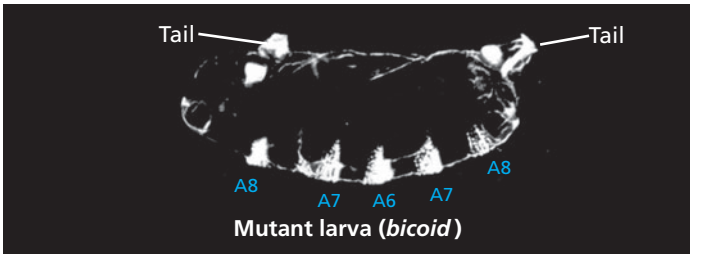
Because they control the orientation (polarity) of the egg and consequently of the fly, maternal effect genes are also called **egg-polarity genes**. One group of these genes sets up the anterior-posterior axis of the embryo, while a second group establishes the dorsal-ventral axis. Like mutations in

segmentation genes, mutations in maternal effect genes are generally embryonic lethals.

Bicoid: A Morphogen Determining Head Structures To see how maternal effect genes determine the body axes of the offspring, we'll focus on one such gene, called ***bicoid***, a term meaning "two-tailed." An embryo whose mother has two mutant alleles of the *bicoid* gene lacks the front half of its body and has posterior structures at both ends **(Figure 16.9)**. 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 this hypothesis, 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. The first question they asked was whether the mRNA and protein products of these genes are located in the egg in a position consistent with the hypothesis. They found that *bicoid* mRNA is highly concentrated at the extreme anterior end of the mature egg, as predicted by the hypothesis **(Figure 16.10)**. 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, with the highest concentration at the anterior end. These results are consistent with the hypothesis that Bicoid protein specifies the fly's anterior end. To test the hypothesis 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.





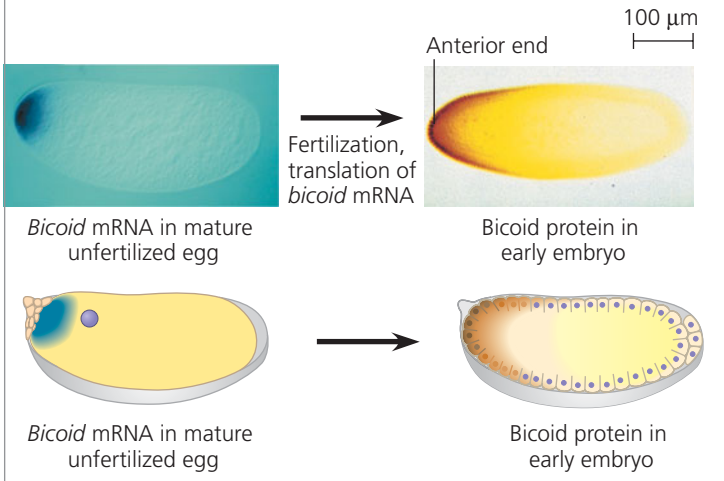
▲ Figure 16.9 Effect of the *bicoid* gene on *Drosophila* development. A wild-type fruit fly larva has a head, three thoracic (T) segments, eight abdominal (A) segments, and a tail. A larva whose mother has two mutant alleles of the *bicoid* gene has two tails and lacks all anterior structures (LMs).

▼ Figure 16.10 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 the European Molecular Biology Laboratory in Heidelberg, 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) 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.

Source 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, predict what happened.*

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. Finally, 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 determine the posterior and anterior ends and 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. (This process involves miRNAs in *Drosophila* and other species.) 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. When the genes operating in this final step are abnormal, the pattern of the adult is abnormal, as you saw in Figure 16.8.

EVOLUTION The fly with legs emerging from its head in Figure 16.8 is the result of a single mutation in one gene. The gene does not encode an antenna protein, however. Instead, it encodes a transcription factor that regulates other genes, and its malfunction leads to misplaced structures like 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 discussed further in Chapter 18.

CONCEPT CHECK 16.1

1. **MAKE CONNECTIONS** As you learned in Chapter 9, 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 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 Figure 5.26.)
3. Why are fruit fly maternal effect genes also called egg-polarity genes?

For suggested answers, see Appendix A.

CONCEPT 16.2

Cloning of organisms showed that differentiated cells could be "reprogrammed" and ultimately led to the production of stem cells

When the field of developmental biology (then called embryology) was first taking shape at the beginning of the 20th century, a major question was whether all the cells of an organism have the same genes (a concept called *genomic equivalence*) or whether cells lose genes during the process of differentiation. Today, we know that genes are not lost—but the question that remains is whether each cell is able to express all of its genes.

One way to answer this question is to see whether a differentiated cell has the potential to generate a whole organism. Because the organism develops from a single cell without either meiosis or fertilization, this is called "cloning." In this context, cloning produces one or more organisms 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 into a collection of genetically identical cells. (The common theme for all types of cloning is that the product is genetically identical to the parent. In fact, the word *clone* comes from the Greek *klon*, meaning "twig.")

The current interest in organismal cloning arises primarily from its potential to generate stem cells, which can in turn generate many different tissues. Conceptually, though, the series of experiments discussed here provides a context for thinking about how regulation of gene expression genetically programs the overall potential of a cell—what genes it can express. Let's discuss early organismal cloning experiments before we consider more recent progress in cloning and procedures for producing stem cells.

Cloning Plants and Animals

The successful cloning of whole plants from single differentiated cells was accomplished during the 1950s by F. C. Steward and his students at Cornell University, who worked with carrot plants. They found that single 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, at least, 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**.

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 the question of whether differentiated animal cells can be totipotent. Their approach was to remove the nucleus of an unfertilized or fertilized egg and replace it with the nucleus of a differentiated cell, a procedure called *nuclear transplantation*. If the nucleus from the differentiated donor cell retains its full genetic capability, 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 (*Xenopus laevis*) by John Gurdon in the 1970s. These researchers transplanted a nucleus from an embryonic or tadpole cell into an enucleated (nucleus-lacking) egg of the same species. In Gurdon's experiments, the transplanted nucleus was often able to support normal development of the

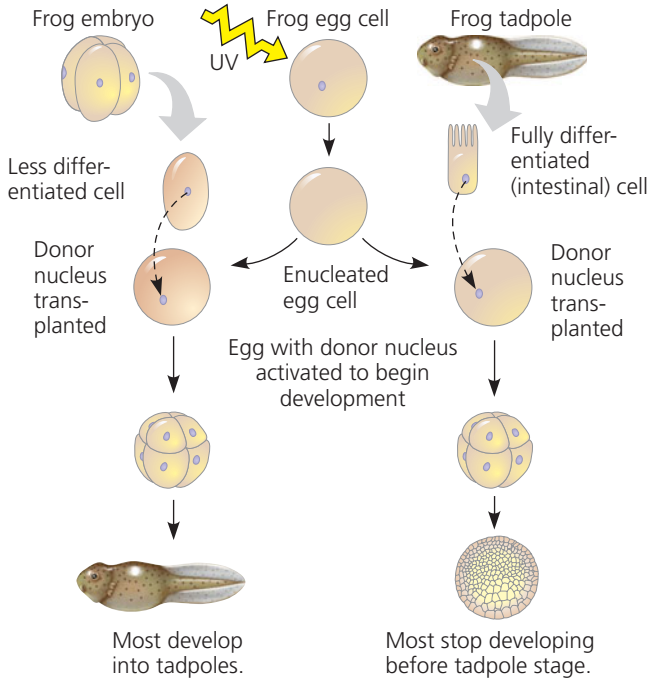
egg into a tadpole **(Figure 16.11)**. However, he found that the potential of a transplanted nucleus to direct normal development was inversely related to the age of the donor: the older the donor nucleus, the lower the percentage of normally developing tadpoles.

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.

▼ Figure 16.11 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, whose cells 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.

Source 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?*

Reproductive Cloning of Mammals

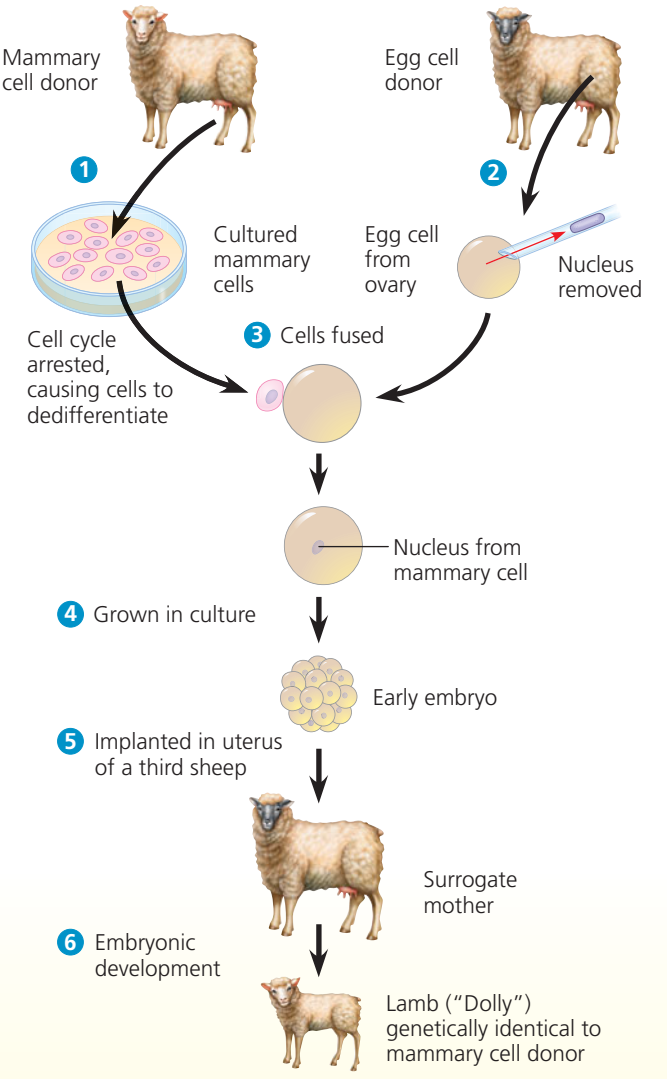
In addition to cloning frogs, researchers have long been able to clone mammals by transplanting nuclei or cells from a variety of early embryos. But until about 15 years ago, it was not known whether a nucleus from a fully differentiated cell could be reprogrammed to successfully act as a donor nucleus. In 1997, however, researchers at the Roslin Institute in Scotland captured newspaper headlines when they announced the birth of Dolly, a lamb cloned from an adult sheep by nuclear transplantation from a differentiated cell **(Figure 16.12)**. These researchers achieved

▼ Figure 16.12 Research Method

Reproductive Cloning of a Mammal by Nuclear Transplantation

Application This method produces cloned animals with nuclear genes identical to those of the animal supplying the nucleus.

Technique The procedure below produced Dolly, the first case of a mammal cloned using the nucleus of a differentiated cell.



Results The cloned animal is genetically identical to the animal supplying the nucleus but differs from the egg donor and surrogate mother. (The latter two are "Scottish blackface" sheep.)

the necessary dedifferentiation of donor nuclei by culturing mammary cells in nutrient-poor medium. They then fused these cells with enucleated sheep eggs. The resulting diploid cells divided to form early embryos, which were implanted into surrogate mothers. Out of several hundred embryos, one successfully completed normal development, and Dolly was born.

Later analyses showed that Dolly's chromosomal DNA was indeed identical to that of the nucleus donor. (Her mitochondrial DNA came from the egg donor, as expected.) At the age of 6, Dolly suffered complications from a lung disease usually seen only in much older sheep and was euthanized. Dolly's premature death, as well as an arthritic condition, led to speculation that her cells were in some way not quite as healthy as those of a normal sheep, possibly reflecting incomplete reprogramming of the original transplanted nucleus.

Since that time, researchers have cloned numerous other mammals, including mice, cats, cows, horses, pigs, dogs, and monkeys. In most cases, their goal has been the production of new individuals; this is known as *reproductive cloning*. We have already learned a lot from such experiments. For example, cloned animals of the same species do *not* always look or behave identically. In a herd of cows cloned from the same line of cultured cells, certain cows are dominant in behavior and others are more submissive. Another example of nonidentity in clones is the first cloned cat, named CC for Carbon Copy **(Figure 16.13)**. She has a calico coat, like her single female parent, but the color and pattern are different because of random X chromosome inactivation, which is a normal occurrence during embryonic development (see Figure 12.8). And identical human twins, which are naturally occurring "clones," are always slightly different. Clearly, environmental influences and random phenomena can play a significant role during development.

Faulty Gene Regulation in Cloned Animals

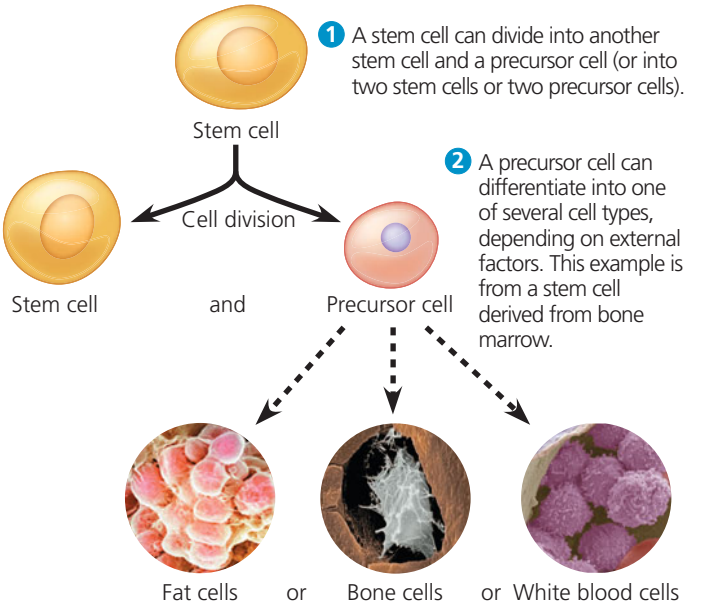
In most nuclear transplantation studies thus far, only a small percentage of cloned embryos develop normally to birth. And like Dolly, many cloned animals exhibit defects. Cloned mice, for instance, are prone to obesity, pneumonia, liver failure, and premature death. Scientists assert that even cloned animals that appear normal are likely to have subtle defects.

In recent years, we have begun to uncover 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 is repressed. This regulation often is the result of epigenetic changes in chromatin, such as acetylation of histones or methylation of DNA (see Figure 15.7). 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 early stages of development. Researchers have found that the DNA in cells from cloned embryos, like that of differentiated cells, often has more methyl groups than does the DNA in equivalent cells from normal embryos of the same species. 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 "rejuvenated" to resemble that of a newly fertilized egg.

▶ Figure 16.13 CC, 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 has orange patches on her fur, but CC does not.

Stem Cells of Animals

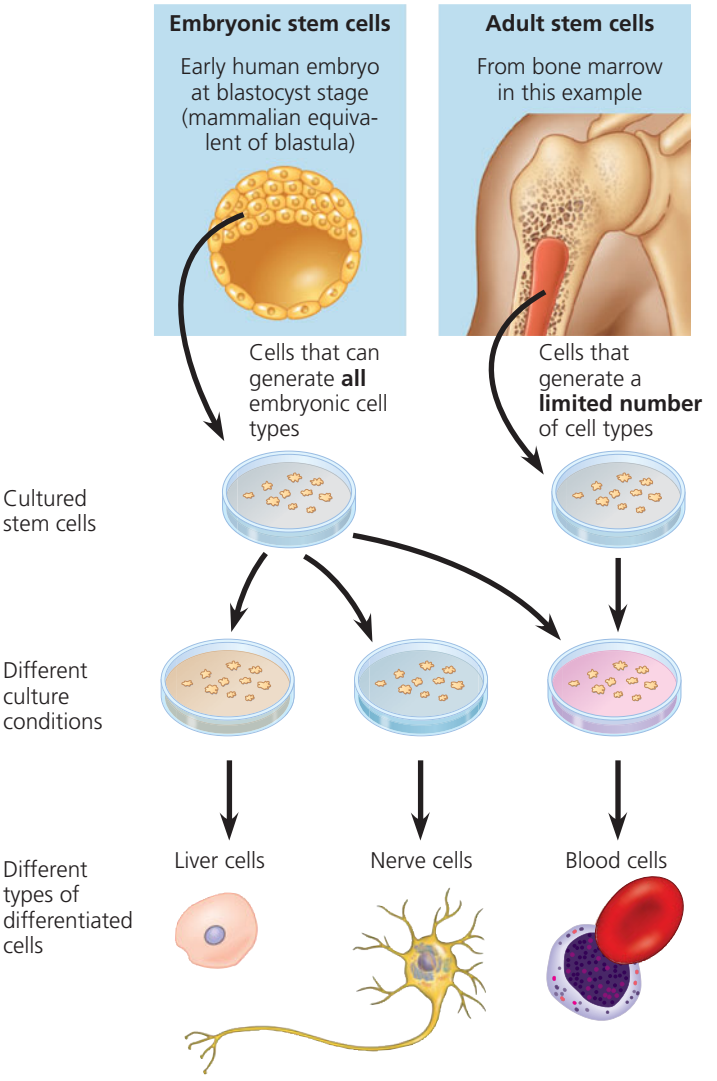
The successful cloning of many mammals, including primates, has heightened speculation about the cloning of humans, which has not yet been achieved. The main reason researchers are trying to clone human embryos is not for reproduction, but for the production of stem cells to treat human diseases. A **stem 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 16.14)**. Thus, stem cells can both replenish their own undifferentiated population and generate cells that travel down specific differentiation pathways.



▲ Figure 16.14 How stem cells maintain their own population and generate differentiated cells.

Many early animal embryos contain stem cells capable of giving rise to differentiated embryonic 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 **(Figure 16.15)**. In culture, these *embryonic stem (ES) cells* reproduce indefinitely; and depending on culture conditions, they can be made to differentiate into a wide variety of specialized cells, 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 in many cases they can generate multiple 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 16.15), and another 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 (often including the addition of specific growth factors), cultured stem cells from adult animals have been made to differentiate into multiple types of specialized cells, although none are as versatile as ES cells.



▲ Figure 16.15 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.

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 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. The only way to obtain ES cells thus far, however, has been to harvest them from human embryos, which raises ethical and political issues.

ES cells are currently obtained from embryos donated (with informed consent) by patients undergoing infertility treatment or from long-term cell cultures originally established with cells isolated from donated embryos. If scientists were able to clone human embryos to the blastocyst stage, they might be able to use such clones as the source of ES cells in the future. Furthermore, with a donor nucleus from a person with a particular disease, they might be able to produce ES cells for treatment that match the patient and are thus not rejected by his or her immune system. 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.

Resolving the debate now seems less imperative because researchers have been able to turn back the clock in fully differentiated cells, reprogramming them to act like ES cells. The accomplishment of this feat, which posed formidable obstacles, was announced in 2007, first by labs using mouse skin cells and then by additional groups using cells from human skin and other organs or tissues. In all these cases, researchers transformed the differentiated cells into ES cells by using types of viruses called retroviruses to introduce extra cloned copies of four "stem cell" master regulatory genes. All the tests that were carried out at the time indicated that the transformed cells, known as *induced pluripotent stem (iPS) cells*, could do everything ES cells can do. More recently, however, several research groups have uncovered differences between iPS and ES cells 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, ES cells will likely always be a focus of basic research as well.) In the meantime, work is proceeding using the iPS cells in hand.

There are two major potential uses for human iPS cells. First, cells from patients suffering from diseases can be reprogrammed to become iPS cells, which can 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, and at least a dozen 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. Developing techniques that direct iPS cells to become specific cell types for this purpose is an area of intense research, one that has already seen some success. The 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.

The research described in this and the preceding section on stem cells and cell differentiation has underscored the key role of gene regulation in embryonic development. The genetic 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 one or a few genes can lead to the development of cancer.

CONCEPT CHECK 16.2

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 16.11?
2. If you were to clone a sheep using the technique shown in Figure 16.12, would all the progeny sheep ("clones") look identical? Why or why not?
3. **WHAT IF?** If you were a doctor who wanted to use iPS cells to treat a patient with severe type 1 diabetes, what new technique would have to be developed?

For suggested answers, see Appendix A.

CONCEPT 16.3

Abnormal regulation of genes that affect the cell cycle can lead to cancer

In Chapter 9, we considered cancer as a set of diseases in which cells escape from the control mechanisms that normally limit their growth. Now that we have discussed the molecular basis of gene expression and its regulation, we are ready to 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 maintenance of stem cell populations, and many other biological processes. Thus, research into the molecular basis of cancer has both benefited from and informed many other fields of biology.

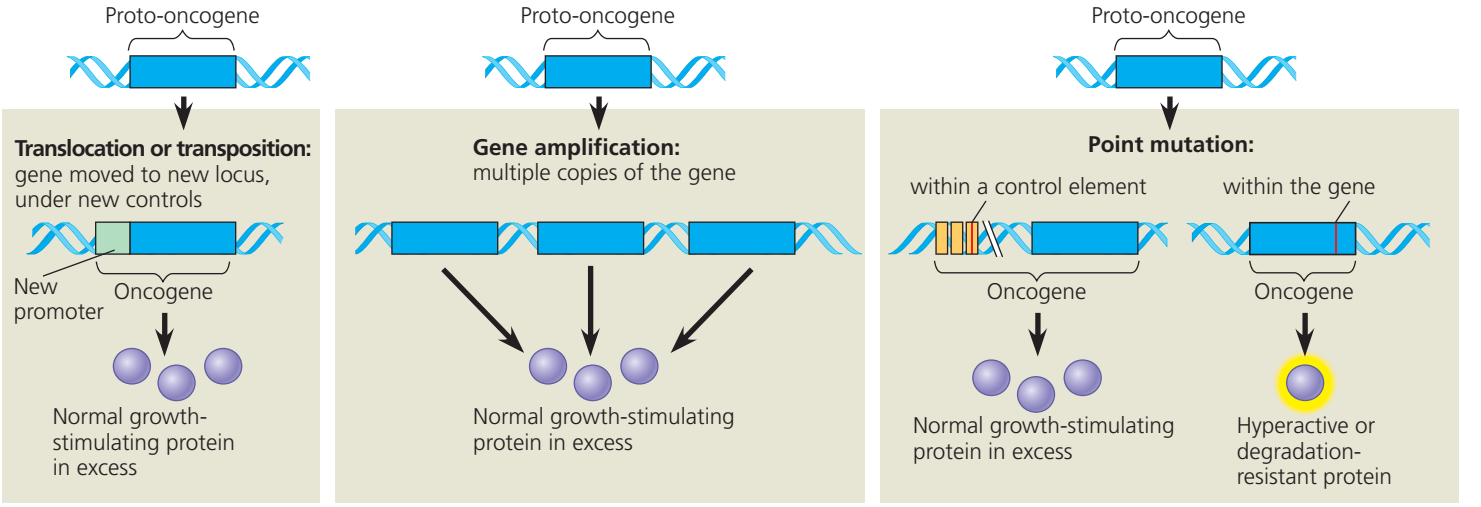
Types of Genes Associated with Cancer

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 the cell cycle, see Chapter 9; for cell signaling, see Concept 5.6.) Mutations that alter any of these genes in somatic cells can lead to cancer. The agent of such change can be random spontaneous mutation. However, it is likely that many cancer-causing mutations result from environmental influences, such as chemical carcinogens, X-rays and other high-energy radiation, and some viruses.

Cancer research led to the discovery of cancer-causing genes called **oncogenes** (from the Greek *onco*, tumor) in certain types of viruses (see Chapter 17). Subsequently, close counterparts 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 three main categories: movement of DNA within the genome, amplification of a proto-oncogene, and point mutations in a control element or in the proto-oncogene itself **(Figure 16.16)**.

Cancer cells are frequently found to contain chromosomes that have broken and rejoined incorrectly, translocating fragments from one chromosome to another (see Figure 12.14). Now that you have learned how gene expression is regulated, you can understand the possible consequences of such translocations. If a translocated proto-oncogene ends up near an especially active promoter (or other control element), its transcription may increase, making it an oncogene. The second main type of genetic change, amplification, increases the number of copies of the proto-oncogene in the cell through repeated gene duplication (discussed in Chapter 18). The third possibility is a point mutation either (1) in the promoter or an enhancer that controls a proto-oncogene, causing an increase in its expression, or (2) in the coding sequence of the proto-oncogene, changing the gene's product to a protein that is more active or more resistant to degradation than the normal protein. All these mechanisms can lead to abnormal stimulation of the cell cycle and put the cell on the path to becoming malignant.



▲ Figure 16.16 Genetic changes that can turn proto-oncogenes into oncogenes.

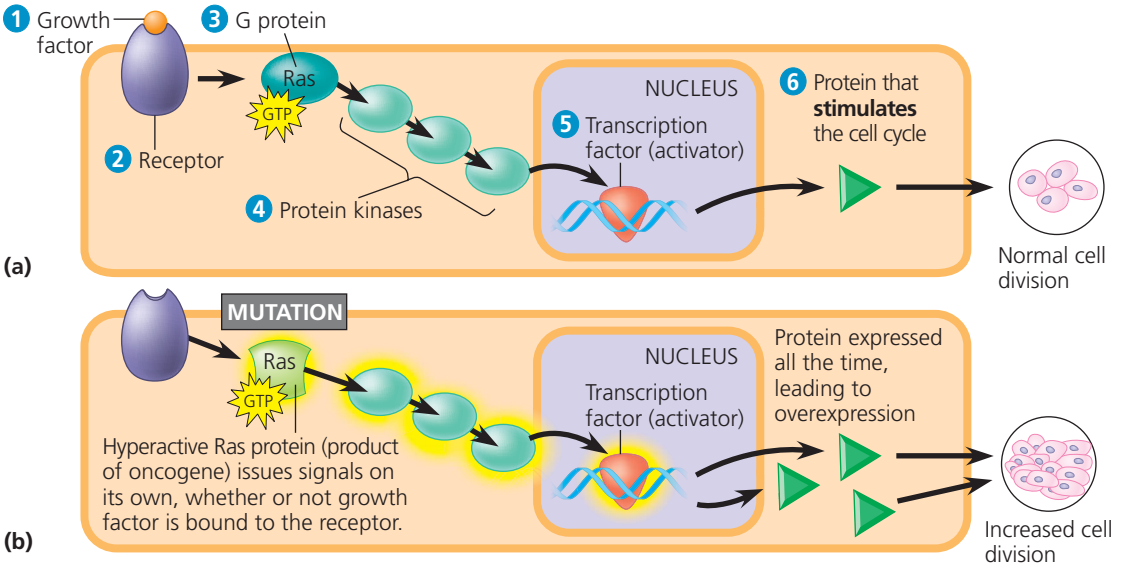
In addition to genes whose products normally promote cell division, cells contain genes whose normal products *inhibit* cell division. Such **tumor-suppressor genes** encode proteins that help prevent uncontrolled cell growth. Any mutation that decreases the normal activity of a tumor-suppressor protein may contribute to the onset of cancer, in effect stimulating growth through the absence of suppression.

Tumor-suppressor gene products have various functions. Some tumor-suppressor proteins repair damaged DNA, a function that prevents the cell from accumulating cancer-causing mutations. Other tumor-suppressor proteins control the 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.

Interference with Cell-Signaling Pathways

The proteins encoded by many proto-oncogenes and tumor-suppressor genes are components of cell-signaling pathways. Let's take a closer look at how such proteins function in normal cells and what goes wrong with their function in cancer cells. We'll 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 Figure 5.21). The cellular response at the end of the pathway is the synthesis of a protein that stimulates the cell cycle **(Figure 16.17)**. 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. In fact, hyperactive versions or excess amounts of any of the pathway's components can have the same outcome: excessive cell division.



▶ Figure 16.17 Normal and mutant cell cycle–stimulating pathway. (a) 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)** If a mutation makes Ras or any other pathway component abnormally active, excessive cell division and cancer may result.

Figure 16.18 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 53,000-dalton 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, like a mutation that leads to a hyperactive Ras protein, can lead to excessive cell growth and cancer.

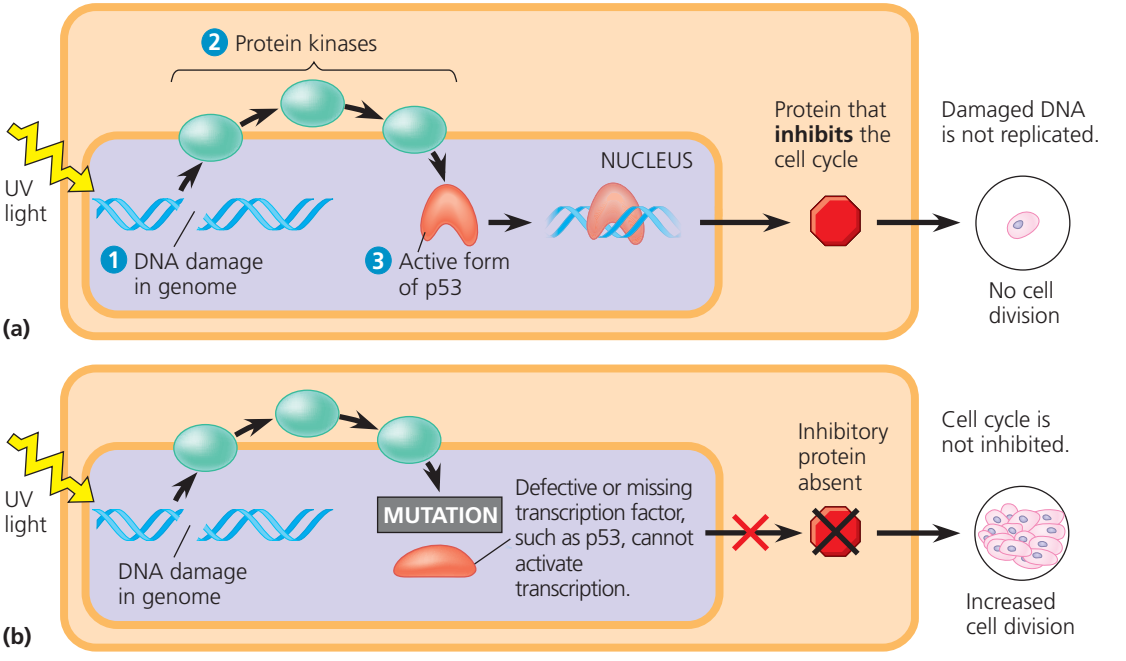
The *p53* gene has been called the "guardian angel of the genome." Once the gene is activated—for example, by DNA damage—the p53 protein functions as an activator for several other genes. Often it activates a gene called *p21*, whose product halts the cell cycle by binding to cyclin-dependent kinases, allowing time for the cell to repair the DNA. Researchers recently showed that p53 also activates expression of a group of miRNAs, which in turn inhibit the cell cycle. In addition, the p53 protein can turn on genes directly involved in DNA repair. Finally, when DNA damage is irreparable, p53 activates "suicide" genes, whose protein products bring about apoptosis, as described in the first section of this chapter. 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.

For the present, the diagrams in Figures 16.17 and 16.18 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. As we discover previously unknown aspects of gene regulation, it is informative to study their role in the onset of cancer. Such studies have shown, for instance, that DNA methylation and histone modification patterns differ in normal and cancer cells and that miRNAs probably participate in cancer development. While we've learned a lot about cancer by studying cell-signaling pathways, there are still a lot of outstanding questions that need to be answered.

The Multistep Model of Cancer Development

More than one somatic mutation 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 and if mutations 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. About 140,000 new cases of colorectal cancer are diagnosed each year in the United States, and the disease causes 50,000 deaths each year. Like most cancers, colorectal cancer develops gradually **(Figure 16.19)**. 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.



▶ **Figure 16.18 Normal and mutant cell cycle–inhibiting pathway. (a)** In the normal pathway, 1 DNA damage is an intracellular signal that is passed via 2 protein kinases and leads to activation of 3 p53. Activated p53 promotes transcription of the gene for 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, the p53 signal leads to programmed cell death (apoptosis). **(b)** Mutations causing deficiencies in any pathway component can contribute to the development of cancer.

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.) The order in which these changes must occur is still under investigation, as is the relative importance of different mutations.

Since we understand the progression of this type of cancer, routine screenings are recommended to identify and remove any suspicious polyps. The colorectal cancer rate has been declining for the past 20 years, due in part to increased screening and in part to improved treatments. Treatments for other cancers have improved as well. Technical advances in the sequencing of DNA and mRNA have allowed medical researchers to compare the genes expressed by different types of tumors and by the same type in different individuals. These comparisons have led to personalized cancer treatments based on the molecular characteristics of an individual's tumor.

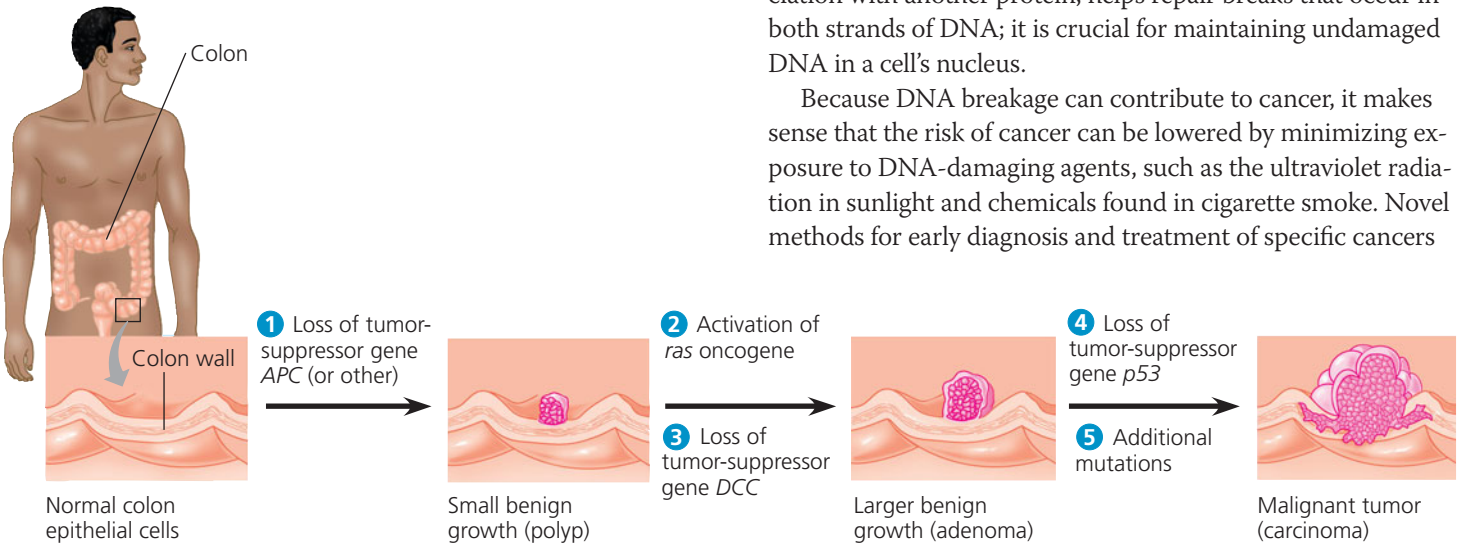
Inherited Predisposition and Other Factors Contributing to Cancer

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 devoting much effort to identifying 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. Many of these affect the tumor-suppressor gene called *adenomatous polyposis coli*, or *APC* (see Figure 16.19). 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 occur in both *APC* alleles before the gene's function is lost. Since only 15% of colorectal cancers are associated with known inherited mutations, researchers continue in their efforts to identify "markers" that could predict the risk of developing this type of cancer.

There is evidence of a strong inherited predisposition in 5–10% of patients with breast cancer. This is the second most common type of cancer in the United States, striking over 230,000 women (and some men) annually and killing 40,000 each year. In 1990, after 16 years of research, geneticist Mary-Claire King 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 the related *BRCA2* gene 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. Both *BRCA1* and *BRCA2* are considered tumor-suppressor genes because their wild-type alleles protect against breast cancer and their mutant alleles are recessive. Apparently, the BRCA1 and BRCA2 proteins both function in the cell's DNA damage repair pathway. More is known about BRCA2, which, in association with another protein, helps repair breaks that occur in both strands of DNA; it is crucial for maintaining undamaged DNA in a cell's nucleus.

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 methods for early diagnosis and treatment of specific cancers



▲ Figure 16.19 A multistep model for the development of colorectal cancer. Affecting the colon and/or rectum, 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 *DCC* stands for "deleted in colorectal cancer." Other mutation sequences can also lead to colorectal cancer.

are being developed that rely on new techniques for analyzing, and perhaps interfering with, gene expression in tumors. Ultimately, such approaches may lower the death rate from 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. The Epstein-Barr virus, which causes infectious mononucleosis, has been linked to several types of cancer in humans, notably Burkitt's lymphoma. Papillomaviruses are associated with cancer of the cervix, and a virus called HTLV-1 causes a type of adult leukemia. Worldwide, viruses seem to 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. In addition, 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, and you'll learn more about their function in Chapter 17.

CONCEPT CHECK 16.3

1. The p53 protein can activate genes involved in apoptosis, or programmed cell death. Review Concept 16.1 and discuss how mutations in genes coding for proteins that function in apoptosis could contribute to cancer.
2. Under what circumstances is cancer considered to have a hereditary component?
3. **WHAT IF?** Explain how the types of mutations that lead to cancer are different for a proto-oncogene and a tumor-suppressor gene in terms of the effect of the mutation on the activity of the gene product.

For suggested answers, see Appendix A.

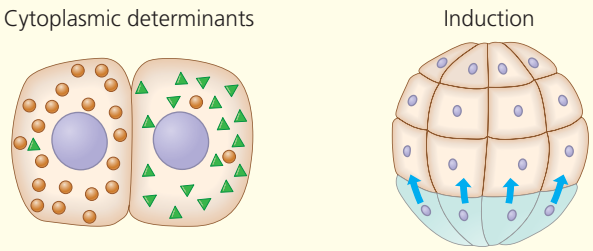
16 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 16.1

A program of differential gene expression leads to the different cell types in a multicellular organism (pp. 312–320)

- Embryonic cells undergo **differentiation**, becoming specialized in structure and function. **Morphogenesis** encompasses the processes that give shape to the organism and its various structures. Cells differ in structure and function not because they contain different genes but because they express different portions of a common genome.
- Localized **cytoplasmic determinants** in the unfertilized egg are distributed differentially to daughter cells, where they regulate the expression of genes that affect those cells' developmental fates. In the process called **induction**, signaling molecules from embryonic cells cause transcriptional changes in nearby target cells.



- Differentiation is heralded by the appearance of tissue-specific proteins, which enable differentiated cells to carry out their specialized roles.
- **Apoptosis** is a type of programmed cell death in which cell components are disposed of in an orderly fashion, without damage to neighboring cells. Studies of the soil worm *Caenorhabditis elegans* showed that apoptosis occurs at defined times during embryonic development. Related apoptotic signaling pathways exist in the cells of humans and other mammals, as well as yeasts.
- 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 16.2

Cloning of organisms showed that differentiated cells could be "reprogrammed" and ultimately led to the production of stem cells (pp. 320–324)

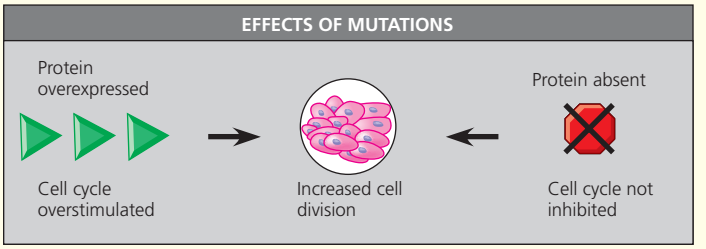
- Studies showing genomic equivalence (that an organism's cells all have the same genome) provided the first examples of organismal cloning.
- Single differentiated cells from plants are often **totipotent**: capable of generating all the tissues of a complete new plant.
- Transplantation of the nucleus 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 or adult stem cells from adult tissues can reproduce and differentiate *in vitro* as well as *in vivo*, 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 a researcher could carry out organismal cloning, production of ES cells, and generation of iPS cells, focusing on how the cells are reprogrammed and using mice as an example. (The procedures are basically the same in humans and mice.)*

CONCEPT 16.3

Abnormal regulation of genes that affect the cell cycle can lead to cancer (pp. 324–328)

- The products of **proto-oncogenes** and **tumor-suppressor genes** control cell division. A DNA change that makes a proto-oncogene excessively active converts it to an **oncogene**, which may promote excessive cell division and cancer. A tumor-suppressor gene encodes a protein that inhibits abnormal cell division. A mutation in such a gene that reduces the activity of its protein product may also lead to excessive cell division and possibly to cancer.
- Many proto-oncogenes and tumor-suppressor genes encode components of growth-stimulating and growth-inhibiting signaling pathways, respectively, and mutations in these genes 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 function 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.
- Individuals who inherit a mutant oncogene or tumor-suppressor allele 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 the functions of proteins encoded by tumor-suppressor genes.*

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

1. 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.
   e. have different chromosomes.
2. Cell differentiation always involves
   a. the production of tissue-specific proteins, such as muscle actin.
   b. the movement of cells.
   c. the transcription of the *myoD* gene.
   d. the selective loss of certain genes from the genome.
   e. the cell's sensitivity to environmental cues, such as light or heat.

Level 2: Application/Analysis

3. Apoptosis involves all but which of the following?
   a. fragmentation of the DNA
   b. cell-signaling pathways
   c. activation of cellular enzymes
   d. lysis of the cell
   e. digestion of cellular contents by scavenger cells
4. 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. is transcribed in the early embryo.
   b. normally leads to formation of tail structures.
   c. normally leads to formation of head structures.
   d. is a protein present in all head structures.
   e. leads to programmed cell death.
5. 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 normally help regulate cell division.
   c. Proto-oncogenes are genetic "junk."
   d. Proto-oncogenes are mutant versions of normal genes.
   e. Cells produce proto-oncogenes as they age.

Level 3: Synthesis/Evaluation

6. **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 5.23 to review the action of these steroid hormones.)
7. **FOCUS ON EVOLUTION**
   Cancer cells can be considered a population that undergoes evolutionary processes such as random mutation and natural selection. Apply what you learned about evolution in Chapter 1 and about cancer in this chapter to discuss this concept.
8. **FOCUS ON ORGANIZATION**
   The property of life emerges at the biological level of the cell. The highly regulated process of apoptosis is not simply the destruction of a cell; it is also an emergent property. In a short essay (about 100–150 words), briefly explain the role of apoptosis in the development and proper functioning of an animal and describe how this form of programmed cell death is a process that emerges from the orderly integration of signaling pathways.

For selected answers, see Appendix A.

17 Viruses

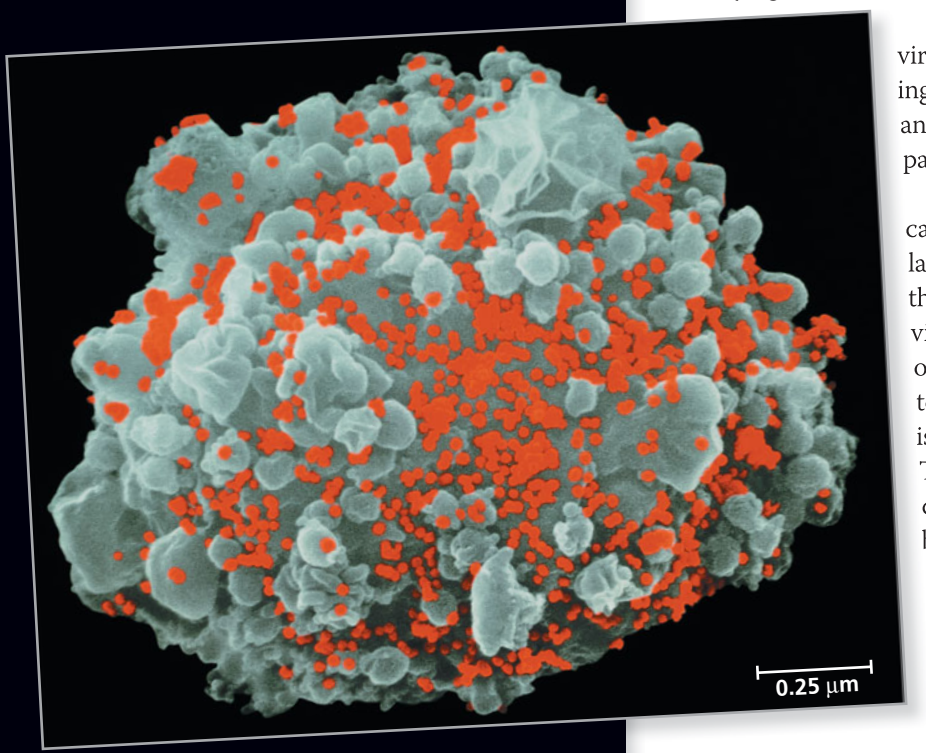
KEY CONCEPTS

17.1 A virus consists of a nucleic acid surrounded by a protein coat

17.2 Viruses replicate only in host cells

17.3 Viruses are formidable pathogens in animals and plants

▼ **Figure 17.1** Are the tiny viruses (red) budding from this cell alive?



OVERVIEW

A Borrowed Life

The photo in **Figure 17.1** shows a remarkable event: a cell under siege, releasing thousands more of its attackers, each capable of infecting another cell. The attackers (red) are human immunodeficiency viruses (HIV) emerging from a human immune cell. By injecting its genetic information into the infected cell, a single virus hijacks the cell, recruiting cellular machinery to manufacture many new viruses and promote further infection. Left untreated, HIV causes acquired immunodeficiency syndrome (AIDS) by destroying the immune system.

Compared to eukaryotic and even prokaryotic cells, viruses are much smaller and simpler in structure. Lacking the 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? Because viruses are capable of causing many diseases, 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 metabolism outside of a host cell. Most biologists studying viruses today would likely agree that they are not alive but exist in a shady area between life-forms and chemicals. The simple phrase used recently by two researchers describes them aptly enough: Viruses lead "a kind of borrowed life."

In this chapter, we'll explore the biology of viruses, beginning with their structure and then describing how they replicate. We'll end the chapter with a look at the role of viruses as disease-causing agents, or pathogens, of plants and animals.

CONCEPT 17.1

A virus consists of a nucleic acid surrounded by a protein coat

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 several hundred nanometers, is barely visible under the light microscope. An early 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 nucleic acid enclosed in a protein coat and, for some viruses, surrounded by a membranous envelope.

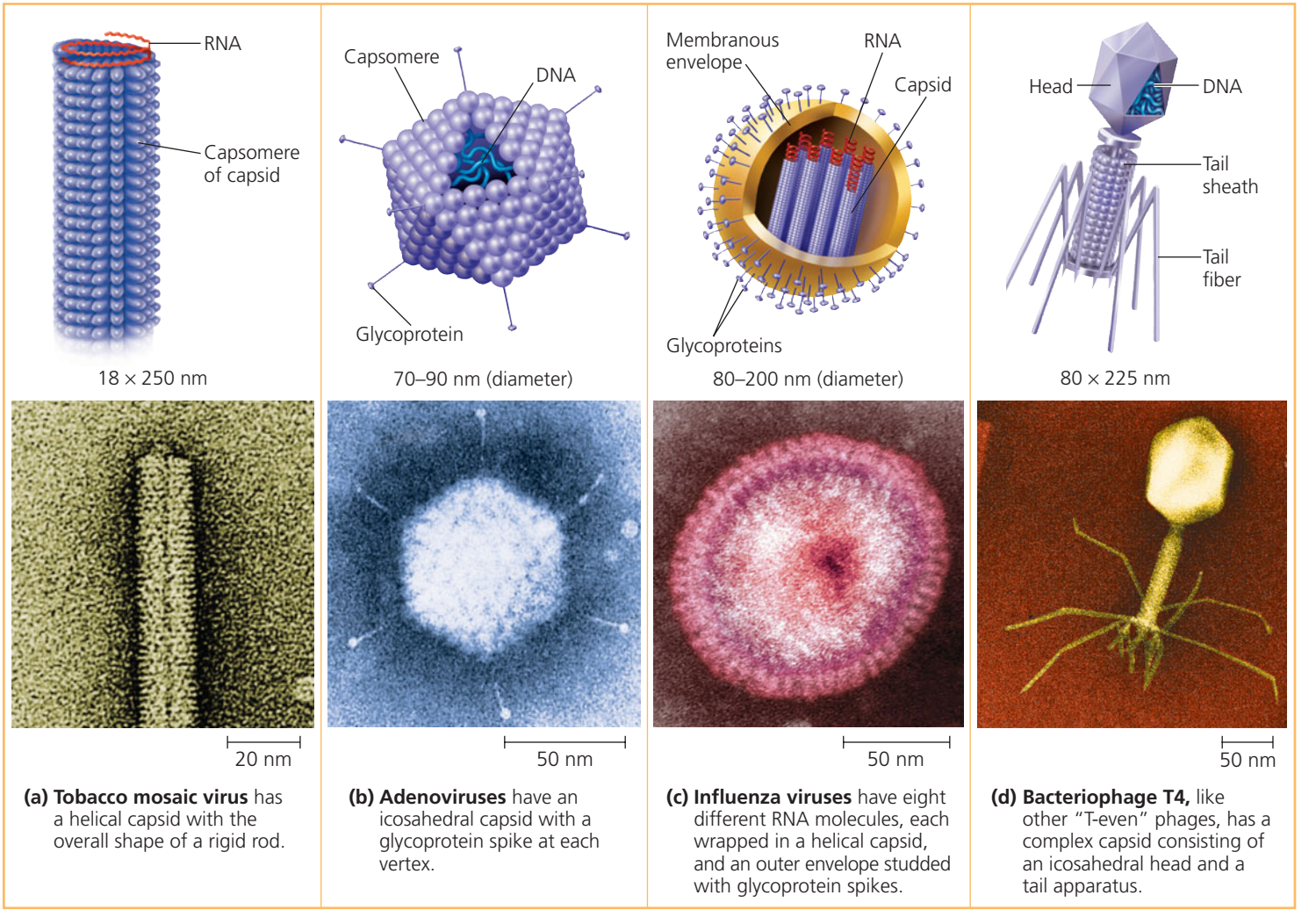
Viral Genomes

We usually think of genes as being made of double-stranded DNA—the conventional double helix—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 four genes in their genome, while the largest have several hundred to a thousand. 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 (TMV), for example, has a rigid, rod-shaped capsid made from over a thousand molecules of a single type of protein arranged in a helix; rod-shaped viruses are commonly called *helical viruses* for this reason **(Figure 17.2a)**. 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 17.2b)**.

Some viruses have accessory structures that help them infect their hosts. For instance, a membranous envelope surrounds



(a) Tobacco mosaic virus has a helical capsid with the overall shape of a rigid rod.

(b) Adenoviruses have an icosahedral capsid with a glycoprotein spike at each vertex.

(c) Influenza viruses have eight different RNA molecules, each wrapped in a helical capsid, and an outer envelope studded with glycoprotein spikes.

(d) Bacteriophage T4, like other "T-even" phages, has a complex capsid consisting of an icosahedral head and a tail apparatus.

▲ **Figure 17.2 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.)

the capsids of influenza viruses and many other viruses found in animals **(Figure 17.2c)**. 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 carbohydrates covalently attached.) Some viruses carry a few viral enzyme molecules 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 *E. coli*. These seven phages were named type 1 (T1), type 2 (T2), and so forth, in the order of their discovery. 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 bacterium **(Figure 17.2d)**. 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 17.1

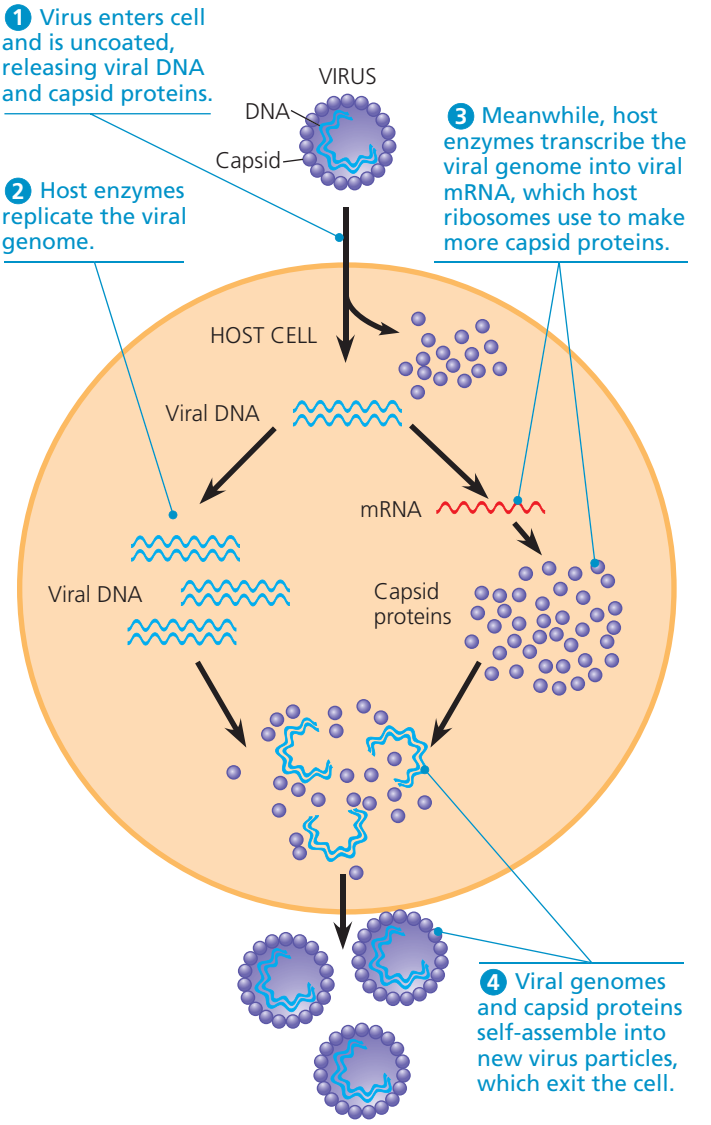
1. Compare the structures of tobacco mosaic virus and influenza virus (see Figure 17.2).
2. **MAKE CONNECTIONS** Bacteriophages were used to provide evidence that DNA carries genetic information (see Figure 13.4). Briefly describe the experiment carried out by Hershey and Chase, including in your description why the researchers chose to use phages.

For suggested answers, see Appendix A.

CONCEPT 17.2
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 "lock-and-key" fit between viral surface proteins and specific receptor molecules on the outside of cells. 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 virus binds to receptors present only on certain types of white blood cells (see Figure 17.1).



▲ Figure 17.3 A simplified viral replicative cycle. A virus is an obligate 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 black arrows with one word representing the name of the process that is occurring. (Review Figure 14.24.)*

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 17.3)**. 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 17.2d). 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 nucleic acid and manufacture viral proteins. The host provides the nucleotides

for making viral nucleic acids, as well as enzymes, ribosomes, 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'll now take a look at some of these variations in bacterial viruses (phages) and animal viruses; later in the chapter, we'll consider plant viruses.

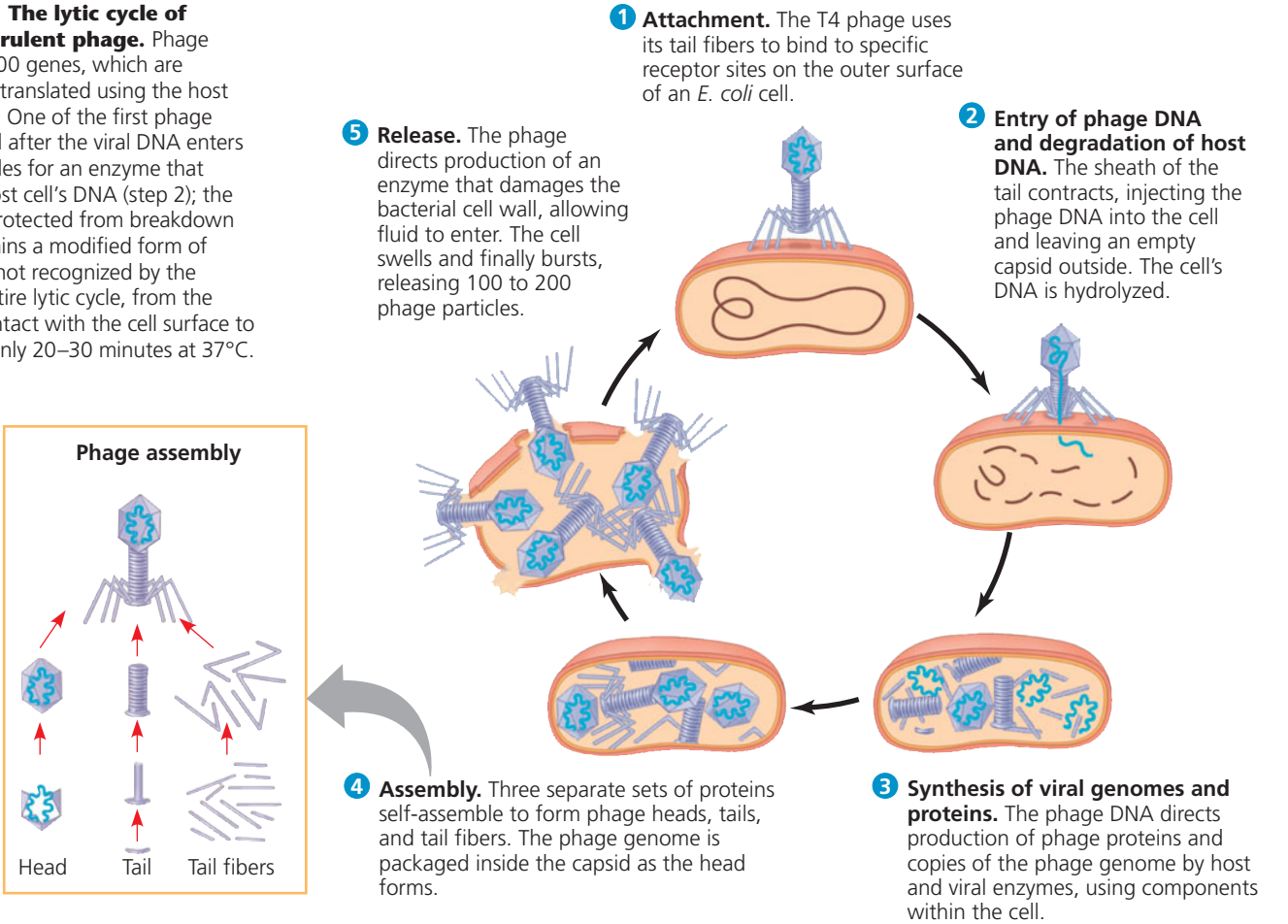
Replicative Cycles of Phages

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 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 called a **virulent phage**. **Figure 17.4** illustrates the major steps in the lytic cycle of T4, a typical virulent phage. Study this figure before proceeding.

After reading about the lytic cycle, you may wonder why phages haven't exterminated all bacteria. The reason is that bacteria have their own defenses. First, natural selection favors bacterial mutants with receptors that are no longer recognized by a particular type of phage. Second, when phage DNA does



▶ Figure 17.4 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 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.

enter a bacterium, the DNA often is identified as foreign and cut up by cellular enzymes called **restriction enzymes**, which are so named because their activity *restricts* the ability of the phage to infect the bacterium. (These enzymes are used in molecular biology and DNA cloning techniques; see Concept 13.4.) The bacterial cell's own DNA is methylated in a way that prevents attack by its own restriction enzymes. But just as natural selection favors bacteria with mutant receptors or effective restriction enzymes, it also favors phage mutants that can bind the altered receptors or are resistant to particular restriction enzymes. Thus, the parasite-host relationship is in constant evolutionary flux.

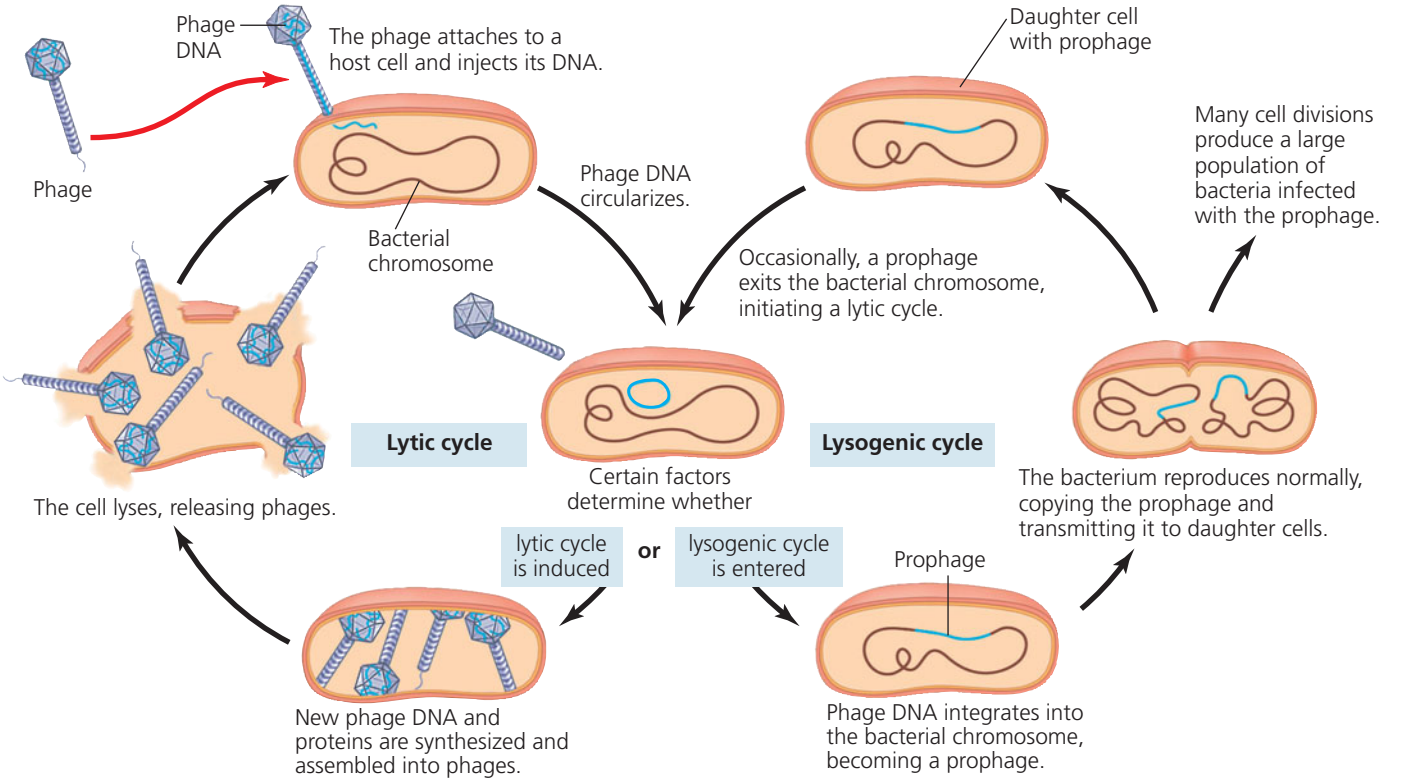
There is yet a third important reason bacteria have been spared from extinction as a result of phage activity. Instead of lysing their host cells, many phages coexist with them in a state called lysogeny, which we'll now discuss.

The Lysogenic Cycle

In contrast to the lytic cycle, which kills the host cell, the **lysogenic cycle** allows 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, is 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 17.5)**. 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 viral products. 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 and passes the copies on to daughter cells. 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* implies that prophages are capable of generating active phages that lyse their host cells. This occurs when the λ genome is induced to exit the bacterial chromosome and initiate a lytic cycle. An environmental signal, such



▲ Figure 17.5 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 17.4. 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.

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 transcription-preventing protein, 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 that resides in our intestines and the O157:H7 strain that has caused several deaths by food poisoning appears to be the presence of prophages in the O157:H7 strain.

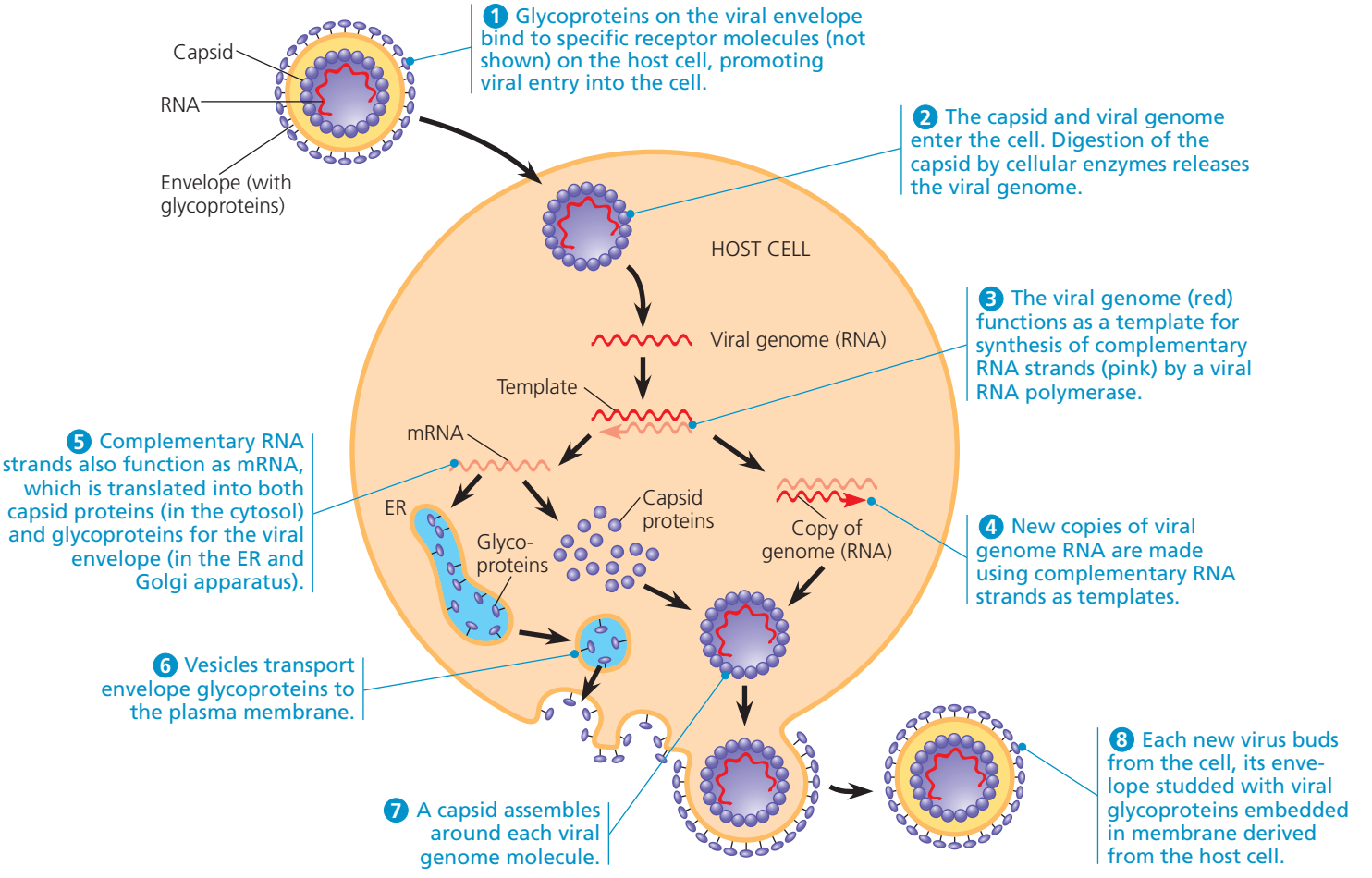
Replicative Cycles of Animal Viruses

Everyone has suffered from viral infections, whether cold sores, influenza, or the common cold. Like all viruses, those 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 few bacteriophages have an RNA genome or envelope, 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'll focus on the roles of viral envelopes and on the functioning of RNA as the genetic material of many animal viruses.

Viral Envelopes

An animal virus equipped with an envelope—that is, an outer membrane—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 17.6** 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



▲ **Figure 17.6 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. Some enveloped viruses enter the host cell by fusion of the envelope with the cell's plasma membrane; others enter by endocytosis. For all enveloped RNA viruses, the formation of new envelopes for progeny viruses occurs by the mechanism depicted in this figure.

host cell, are transported to the cell surface. In a process much like exocytosis, new viral capsids are wrapped in membrane as they bud from the cell. In other words, the viral envelope is derived from the host cell's plasma membrane, although some 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.

RNA as Viral Genetic Material

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. In the first type, the viral genome can directly serve as mRNA and thus can be translated into viral protein immediately after infection. In a second type, the RNA genome serves 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 require RNA → RNA synthesis to make mRNA use a viral enzyme capable of carrying out this process; there are no such enzymes in most cells. The viral enzyme is packaged with the genome inside the viral capsid.

The RNA animal viruses with the most complicated replicative cycles are the third type, the **retroviruses**. These viruses are equipped with an enzyme called **reverse transcriptase**, which transcribes an RNA template into DNA, providing an RNA → DNA information flow, the opposite of the usual direction. (Reverse transcriptase is the enzyme used in the technique called RT-PCR, described in Concept 15.4.) This unusual phenomenon is the source of the name retroviruses (*retro* means "backward"). Of particular medical importance is **HIV (human immunodeficiency virus)**, the retrovirus 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 (traced in **Figure 17.7**) 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 host's RNA polymerase 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 Chapter 35, we'll describe how HIV causes the deterioration of the immune system that occurs in AIDS.

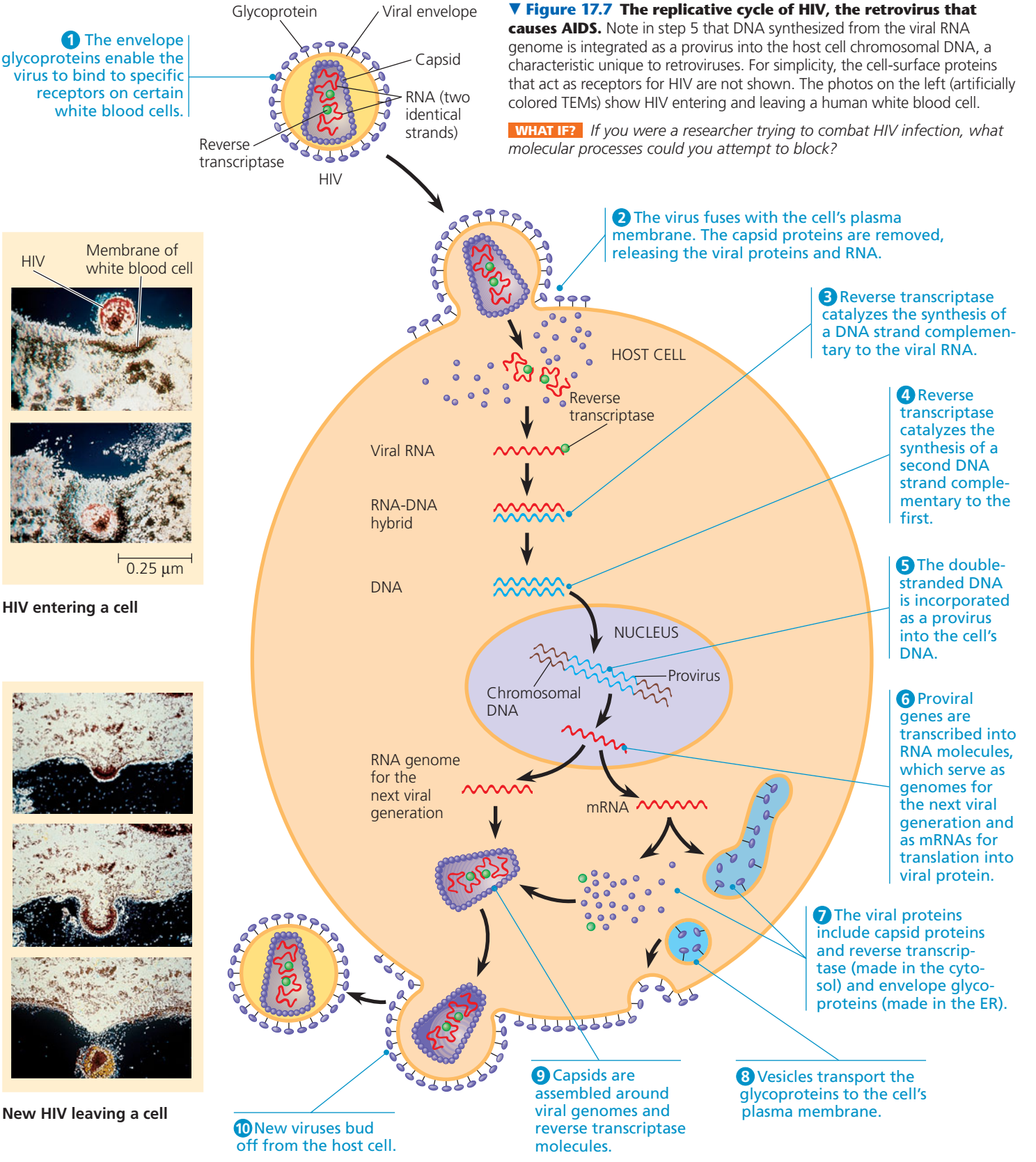
Evolution of Viruses

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 supply of 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 just bacteria, animals, and plants, but also archaea, fungi, and algae and other protists. Because they 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 facilitated 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 eukaryotes called yeasts. Plasmids exist apart from the cell's genome, can replicate independently of the genome, and are occasionally transferred between cells. (Use of plasmids in gene cloning was discussed in Concept 13.4.) *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 Chapter 24 and transposons in Chapter 18.)

Consistent with this vision of pieces of DNA shuttling from cell to cell is the observation that a viral genome can have more



▼ Figure 17.7 The replicative cycle of HIV, the retrovirus that causes AIDS. Note in step 5 that DNA synthesized from the viral RNA genome is integrated as a provirus into the host cell chromosomal DNA, a characteristic unique to retroviruses. For simplicity, the cell-surface proteins that act as receptors for HIV are not shown. The photos on the left (artificially colored TEMs) show HIV entering and leaving a human white blood cell.

WHAT IF? *If you were a researcher trying to combat HIV infection, what molecular processes could you attempt to block?*

in common with the genome of its host than with the genomes of viruses that infect other hosts. The ongoing evolutionary relationship between viruses and the genomes of their host cells is an association that makes 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 17.2

1. Compare the effect on the host cell of a lytic (virulent) phage and a lysogenic (temperate) phage.
2. **MAKE CONNECTIONS** The RNA virus in Figure 17.6 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 14.10).
3. Why is HIV called a retrovirus?

For suggested answers, see Appendix A.

CONCEPT 17.3

Viruses are formidable pathogens in animals and plants

Diseases caused by viral infections afflict humans, crops, and livestock worldwide. We'll first discuss animal viruses.

Viral Diseases in Animals

A viral infection can produce symptoms by a number of different routes. 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 aches, actually result from the body's own efforts at defending itself against infection rather than from cell death caused by the virus.

The immune system is a complex and critical part of the body's natural defenses (see Chapter 35). It is also the basis for the major medical tool for preventing viral infections—vaccines. A **vaccine** is a harmless variant or derivative of a pathogen that stimulates the immune system to mount defenses against the harmful pathogen. Smallpox, a viral disease that was at one time a devastating scourge in many parts of the world, was eradicated by 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 currently under way to eradicate polio and measles. Effective vaccines are also available to protect against rubella, mumps, hepatitis A and B, and a number of other viral diseases.

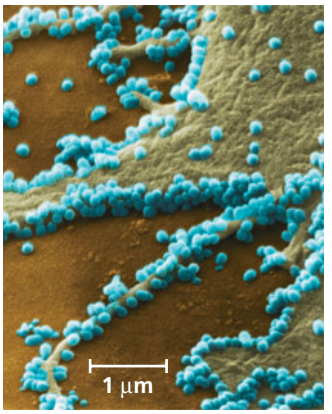
Although vaccines can prevent certain viral illnesses, medical technology can do little, at present, to cure most viral infections once they occur. The antibiotics that help us recover from bacterial infections are powerless against viruses. Antibiotics kill bacteria by inhibiting enzymes specific to bacteria but have no effect on eukaryotic or virally encoded enzymes. However, the few enzymes that are encoded by viruses have provided targets for other drugs. Most antiviral drugs resemble nucleosides and as a result interfere with viral nucleic acid synthesis. One such drug is acyclovir, which impedes herpesvirus replication by inhibiting the viral polymerase that synthesizes viral DNA. Similarly, azidothymidine (AZT) curbs HIV replication by interfering with the synthesis of DNA by reverse transcriptase. In the past two decades, much effort has gone into developing drugs against HIV. Currently, multidrug treatments, sometimes called "cocktails," have been found 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.

Emerging Viruses

Viruses that suddenly become apparent are often referred to as *emerging viruses*. HIV, the AIDS virus, is a classic example: This virus appeared in San Francisco in the early 1980s, seemingly out of nowhere, although later studies uncovered a case in the Belgian Congo in 1959. The deadly Ebola virus, recognized initially in 1976 in central Africa, is one of several emerging viruses that cause *hemorrhagic fever*, an often fatal syndrome (set of symptoms) characterized by fever, vomiting, massive bleeding, and circulatory system collapse. A number of other dangerous emerging viruses cause encephalitis, inflammation of the brain. One example is the West Nile virus, which appeared in North America for the first time in 1999 and has spread to all 48 contiguous states in the United States.

In 2009, a general outbreak, or **epidemic**, of a flu-like illness appeared in Mexico and the United States. The infectious agent was quickly identified as an influenza virus related to viruses that cause the seasonal flu **(Figure 17.8a)**. This particular virus was named H1N1 for reasons that will be explained shortly. The viral disease spread rapidly, prompting WHO to declare a global epidemic, or **pandemic**, shortly thereafter. Half a year later, the disease had reached 207 countries, infecting over 600,000 people and killing almost 8,000. Public health agencies responded rapidly with guidelines for shutting down schools and other public places, and vaccine development and screening efforts were accelerated **(Figure 17.8b)**.

How do such viruses burst on the human scene, giving rise to harmful diseases that were previously rare or even unknown? Three processes contribute to the emergence of viral diseases. The first, and perhaps most important, is the mutation of existing viruses. RNA viruses tend to have an unusually high rate of mutation because errors in replicating their RNA genomes are not corrected by proofreading. Some mutations change existing viruses into new genetic varieties (strains) that can cause disease, even in individuals who are immune to the ancestral virus. For instance, seasonal flu epidemics are caused by new strains of influenza virus



(a) 2009 pandemic H1N1 influenza A virus. Viruses (blue) are seen on an infected cell (green) in this colorized SEM.

(b) 2009 pandemic screening. At a South Korean airport, thermal scans were used to detect passengers with a fever who might have the H1N1 flu.

▲ Figure 17.8 Influenza in humans.

genetically different enough from earlier strains that people have little immunity to them. You'll see an example of this process in the **Scientific Skills Exercise**, where you'll analyze genetic changes in variants of the 2009 flu virus and correlate them with spread of the disease.

A second process that can lead to the emergence of viral diseases is the dissemination of a viral disease from a small, isolated human population. For instance, AIDS went unnamed and virtually unnoticed for decades before it began to spread around the world. In this case, technological and social factors, including affordable international travel, blood transfusions, sexual promiscuity, and the abuse of intravenous drugs, allowed a previously rare human disease to become a global scourge.

A third source 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 that harbor and can transmit a particular virus but are generally unaffected by it are said to act as a natural reservoir for that virus. For example, the 2009 flu pandemic mentioned earlier was likely passed to humans from pigs; for this reason, it was originally called "swine flu."

In general, flu epidemics provide an instructive example of the effects of viruses moving between species. 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. Influenza A strains have caused four major flu epidemics among humans in the last 100 years. The worst was the first one, the "Spanish flu" pandemic of 1918–1919, which killed between 10 and 20% of those infected—about 40 million people, including many World War I soldiers.

Different strains of influenza A are given standardized names; for example, both the strain that caused the 1918 flu and the one that caused the 2009 pandemic flu are called H1N1. The name identifies which forms of two viral surface proteins are present: hemagglutinin (H) and neuraminidase (N). There are 16 different types of hemagglutinin, a protein that helps the flu virus attach to host cells, and 9 types of neuraminidase, an enzyme that helps release new virus particles from infected cells. Waterbirds have been found that carry viruses with all possible combinations of H and N.

A likely scenario for the 1918 pandemic is that the virus mutated as it passed from one host species to another. When an animal like a pig or a bird is infected with more than one strain of flu virus, the different strains can undergo genetic recombination if the RNA molecules of their genomes mix and match during viral assembly. Pigs were probably the breeding ground for the 2009 flu virus, which contains sequences from bird, pig, and human flu viruses. Coupled with mutation, these reassortments can lead to the emergence of a viral strain capable of infecting human cells. People who have never been exposed to that particular strain before will lack immunity, and the recombinant virus has the potential to be highly pathogenic. If such a flu virus 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.

One potential long-term threat is the avian flu caused by an H5N1 virus carried by wild and domestic birds. The first documented transmission from birds to humans occurred in Hong Kong in 1997. Since then, the overall mortality rate of the H5N1 virus has been greater than 50% of those infected, an alarming number. Also, the host range of H5N1 is expanding, which provides increasing chances for reassortment between different strains. If the H5N1 avian flu virus evolves so that it can spread easily from person to person, it could represent a major global health threat akin to that of the 1918 pandemic.

How easily could this happen? Recently, scientists working with ferrets, small mammals that are animal models for human flu, found that only a few mutations of the avian flu virus would 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 at a scientific conference in 2011 ignited a firestorm of debate about whether to publish the results. Ultimately, the scientific community decided the benefits of potentially understanding how to prevent pandemics would outweigh the risks of the information being used for harmful purposes, and the work was published in 2012.

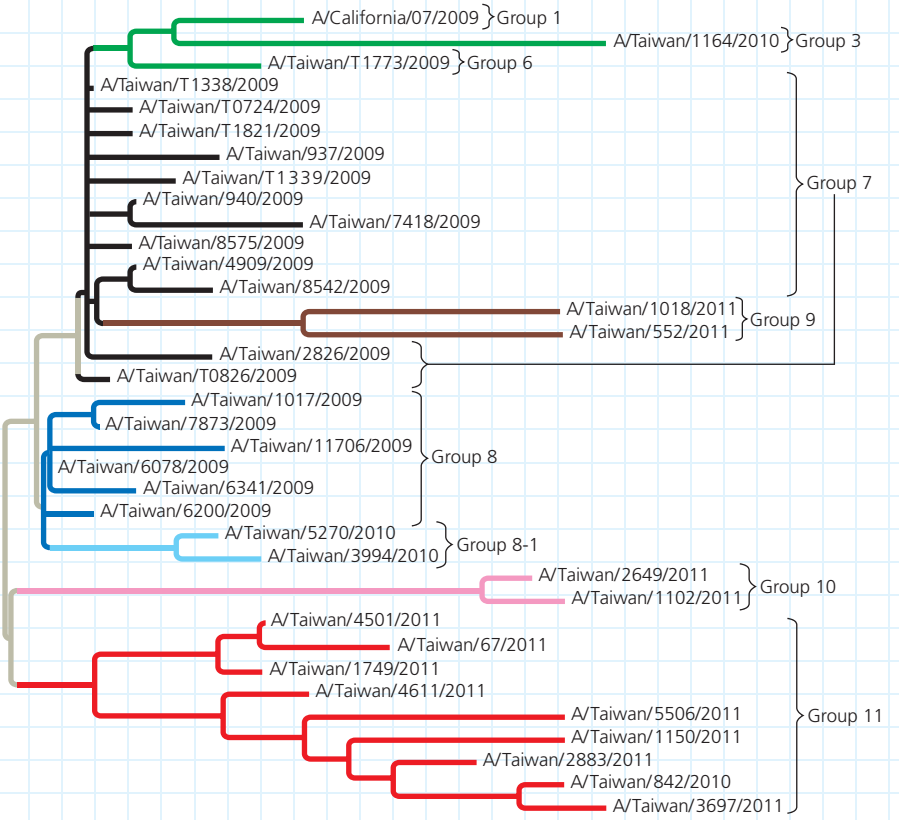
As we have seen, emerging viruses are generally not new; rather, they are existing viruses that mutate, disseminate more widely in the current host species, or spread to new host species. Changes in host behavior or environmental changes can increase the viral traffic responsible for emerging diseases. 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 other animals that may host viruses capable of infecting humans.

Analyzing a DNA Sequence-Based Phylogenetic Tree to Understand Viral Evolution

How Can DNA Sequence Data Be Used to Track Flu Virus Evolution During Pandemic Waves? 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 variants of the H1N1 virus were able to evade human immune system defenses. To test this hypothesis, they needed to determine if each wave of the flu outbreak was caused by a different H1N1 variant.

How the Experiment Was Done Scientists obtained the genome sequences for 4,703 virus isolates collected from patients with H1N1 flu in Taiwan. 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 20.5 for information on how to read phylogenetic trees).

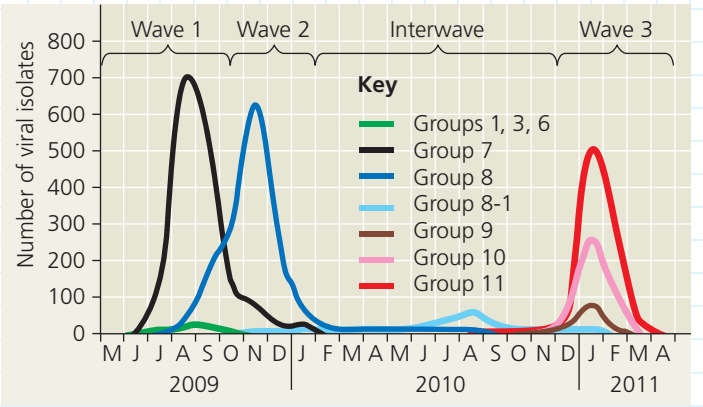
Data from the Experiment The figure below, left, shows a phylogenetic tree; each branch tip is one variant of the H1N1 virus with a unique HA gene sequence. The tree is a way to visualize a working hypothesis about the evolutionary relationships between H1N1 variants.



Interpret the Data

1. The phylogenetic tree shows the hypothesized evolutionary relationship between the variant strains of H1N1 virus. The more closely connected two variants are, 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 DNA sequence differences there are between the variants, thus how distantly related they are. Referring to the phylogenetic tree, which variants are more closely related to each other: A/Taiwan1018/2011 and A/Taiwan/552/2011 or A/Taiwan1018/2011 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 from ill patients 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 H1N1 flu in over 100 patients in Taiwan? (b) Once a group of variants had a peak number of infections, did members of that same group cause another wave of infection? (c) One variant in Group 1 was used to make a vaccine 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 also graphed the isolates by the month and year of isolate collection, which reflects the time period in which each viral variant was actively causing illness in people.

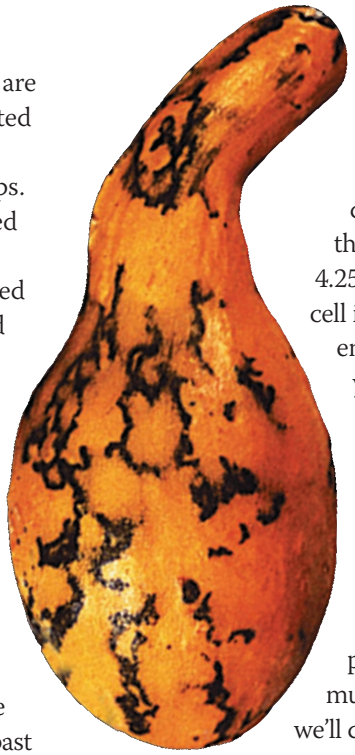


Data from J-R. Yang et al., New variants and age shift to high fatality groups contribute to severe successive waves in the 2009 influenza pandemic in Taiwan, *PLoS ONE* 6(11): e28288 (2011). doi:10.1371/journal.pone.0028288.

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Viral Diseases in Plants

More than 2,000 types of viral diseases of plants are known, and together they account for an estimated annual loss of $15 billion worldwide due to their destruction of agricultural and horticultural crops. Common signs of viral infection include bleached or brown spots on leaves and fruits (as on the squash to the right), stunted growth, and damaged flowers or roots, all tending to diminish the yield and quality of crops.



Plant viruses have the same basic structure and mode of replication as animal viruses. Most plant viruses discovered thus far, including tobacco mosaic virus (TMV), have an RNA genome. Many have a helical capsid, like TMV, while others have an icosahedral capsid (see Figure 17.2).

Viral diseases of plants spread by two major routes. In the first route, called *horizontal transmission*, a plant is infected from an external source of the virus. 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 act as carriers of viruses, transmitting disease from plant to plant. Moreover, farmers and 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 by means of plasmodesmata, the cytoplasmic connections that penetrate the walls between adjacent plant cells (see Figure 4.25). 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. Consequently, research efforts are focused largely on reducing the transmission of such diseases and on breeding resistant varieties of crop plants.

Earlier in this chapter, we mentioned the ongoing evolutionary relationship between viruses and the genomes of their host cells. In fact, the original source of viral genetic material may have been transposons, mobile genetic elements that are present in multiple copies in many genomes. In the next chapter, we'll discuss the structure of genomes and how they evolve.

CONCEPT CHECK 17.3

1. Describe two ways 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.

17 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 17.1

A virus consists of a nucleic acid surrounded by a protein coat (pp. 330–332)

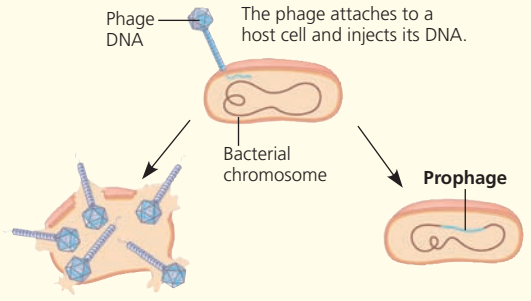
- A **virus** is a small nucleic acid genome enclosed in a protein **capsid** and sometimes a membranous **viral envelope** containing viral proteins that help the virus enter a cell. The genome may be single- or double-stranded DNA or RNA.

? *Are viruses generally considered living or nonliving? Explain.*

CONCEPT 17.2

Viruses replicate only in host cells (pp. 332–338)

- 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**.
- **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

- 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. The origin of viruses is still being debated.

? *Describe enzymes that are not found in most cells but are necessary for the replication of certain viruses.*

CONCEPT 17.3

Viruses are formidable pathogens in animals and plants (pp. 338–341)

- Symptoms of viral diseases in animals 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.
- Outbreaks of "new" viral diseases in humans are usually 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.
- Viruses enter plant cells through damaged cell walls (horizontal transmission) or are inherited from a parent (vertical transmission).

? *What aspect of an RNA virus makes it more likely than a DNA virus to become an emerging virus?*

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

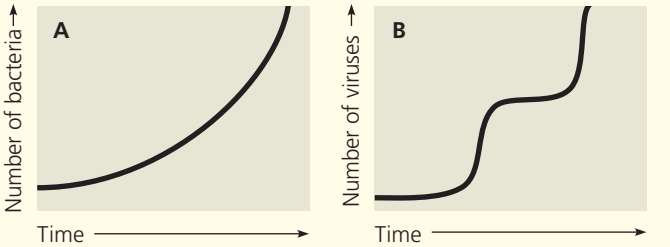
1. Which of the following characteristics, structures, or processes is common to both bacteria and viruses?
   a. metabolism
   b. ribosomes
   c. genetic material composed of nucleic acid
   d. cell division
   e. independent existence
2. Emerging viruses arise by
   a. mutation of existing viruses.
   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
   e. none of the above
3. A human pandemic is
   a. a viral disease that infects all humans.
   b. a flu that kills more than 1 million people.
   c. an epidemic that extends around the world.
   d. a viral disease that can infect multiple species.
   e. a virus that increases in mortality rate as it spreads.

Level 2: Application/Analysis

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 will have
   a. T2 protein and T4 DNA.
   b. T2 protein and T2 DNA.
   c. a mixture of the DNA and proteins of both phages.
   d. T4 protein and T4 DNA.
   e. T4 protein and T2 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.
   e. these enzymes cannot be made in host cells.
6. **DRAW IT** Redraw Figure 17.6 to show the replicative cycle of a virus with a single-stranded genome that can function as mRNA.

Level 3: Synthesis/Evaluation

7. **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. Then the number of viruses rises suddenly and subsequently increases in a series of steps (graph B). Explain the difference in the curves.



8. **FOCUS ON EVOLUTION**
   The success of some viruses lies in their ability to evolve rapidly within the host. Such a virus evades 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?
9. **FOCUS ON 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.

For selected answers, see Appendix A.

18 Genomes and Their Evolution

KEY CONCEPTS

18.1 The Human Genome Project fostered development of faster, less expensive sequencing techniques

18.2 Scientists use bioinformatics to analyze genomes and their functions

18.3 Genomes vary in size, number of genes, and gene density

18.4 Multicellular eukaryotes have much noncoding DNA and many multigene families

18.5 Duplication, rearrangement, and mutation of DNA contribute to genome evolution

18.6 Comparing genome sequences provides clues to evolution and development

OVERVIEW

Reading the Leaves from the Tree of Life

The chimpanzee (*Pan troglodytes*) is our closest living relative on the evolutionary tree of life. The boy in **Figure 18.1** and his chimpanzee companion are intently studying the same leaf, but only one of them is able to talk about what he sees. What accounts for this difference between two primates that share so much of their evolutionary history? With the advent of recent techniques for rapidly sequencing complete genomes, we can now start to address the genetic basis of intriguing questions like this.

▼ **Figure 18.1** What genomic information distinguishes a human from a chimpanzee?

The chimpanzee genome was sequenced two years after sequencing of the human genome was largely completed. Now that we can compare our genome, base by base, with that of the chimpanzee, we can tackle the more general issue of what differences in genetic information account for the distinct characteristics of these two species of primates.

In addition to determining the sequences of the human and chimpanzee genomes, researchers have obtained complete genome sequences for *E. coli* and numerous other prokaryotes, as well as many eukaryotes, including *Zea mays* (corn), *Drosophila melanogaster* (fruit fly), *Mus musculus* (house mouse), and *Pongo pigmaeus* (orangutan). In 2010, a draft sequence was announced for the genome of *Homo neanderthalensis,* an extinct species closely related to present-day humans. These whole and partial genomes are of great interest in their own right and are also providing important insights into evolution and other biological processes. Broadening the human-chimpanzee comparison to the genomes of other primates and more distantly related animals should reveal the sets of genes that control group-defining characteristics. Beyond that, comparisons with the genomes of bacteria, archaea, fungi, protists, and plants will enlighten us about the long evolutionary history of shared ancient genes and their products.

With the genomes of many species fully sequenced, scientists can study whole sets of genes and their interactions, an approach called **genomics**. The sequencing efforts that feed this approach have generated, and continue to generate, enormous volumes of data. The need to deal with this

ever-increasing flood of information has spawned the field of **bioinformatics**, the application of computational methods to the storage and analysis of biological data.

We'll begin this chapter by discussing genome sequencing and some of the advances in bioinformatics and its applications. We'll then summarize what has been learned from the genomes that have been sequenced thus far. Next, we'll describe the composition of the human genome as a representative genome of a complex multicellular eukaryote. Finally, we'll explore current ideas about how genomes evolve and about how the evolution of developmental mechanisms could have generated the great diversity of life on Earth today.

CONCEPT 18.1

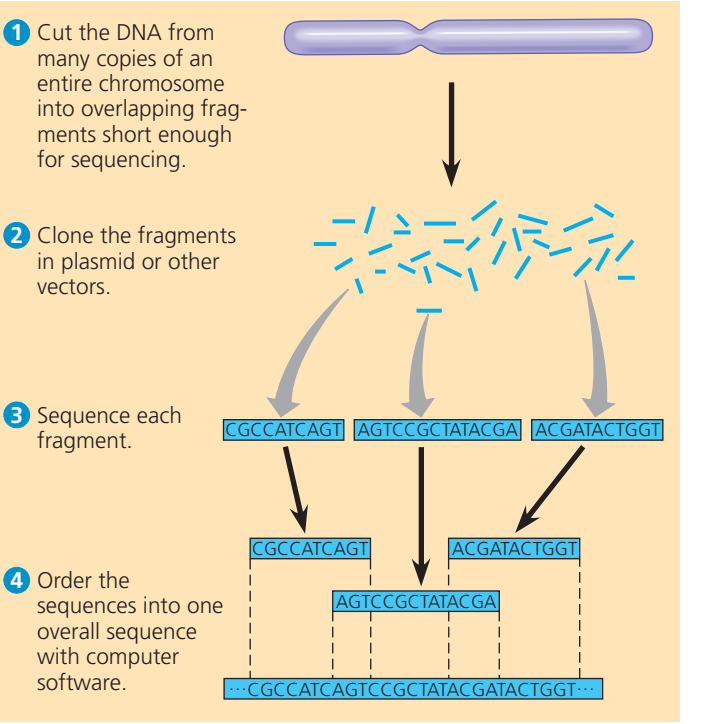
The Human Genome Project fostered development of faster, less expensive sequencing techniques

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 small projects.

After sequencing of the human genome was reported 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. With this refinement, researchers termed the sequencing "virtually complete."

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 sequencing machines. Even with automation, the sequencing of all 3 billion base pairs in a haploid set of human chromosomes presented a formidable challenge. In fact, a major thrust of the Human Genome Project was the development of technology for faster sequencing. 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*, 24 hours a day, seven days a week. Methods that can analyze biological materials very rapidly and produce enormous volumes of data are said to be "high-throughput." Sequencing machines are an example of high-throughput devices.

Two approaches complemented each other in obtaining the complete sequence. The initial approach was a methodical one that built on an earlier storehouse of human genetic information. In 1998, however, molecular biologist J. Craig Venter set up a company (Celera Genomics) and declared his intention to



▲ **Figure 18.2 Whole-genome shotgun approach to sequencing.** In this approach, developed by J. Craig Venter and colleagues at the company he founded, Celera Genomics, random DNA fragments are sequenced and then ordered relative to each other.

? *The fragments in stage 2 of this figure are depicted as scattered, rather than being in an ordered array. How does this depiction accurately reflect the approach?*

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 18.2)**.

Today, the whole-genome shotgun approach is widely used. Also, the development of newer sequencing techniques, generally called *sequencing by synthesis* (see Chapter 13), has resulted in massive increases in speed and decreases in the cost of sequencing entire genomes. In these new techniques, many very small fragments (fewer than 100 base pairs) are sequenced at the same time, and computer software rapidly assembles the complete sequence. Because of the sensitivity of these techniques, the fragments can be sequenced directly; the cloning step (stage 2 in Figure 18.2) is unnecessary. By 2010, the worldwide output was astronomical: close to 100 *billion* bases per day, with the rate estimated to double every 9 months. Whereas sequencing the first human genome took 13 years and cost $100 million, biologist James Watson's genome was sequenced during 4 months in 2007 for about $1 million, and we are rapidly approaching the day when an individual's genome can be sequenced in a matter of hours for less than $1,000!

These technological advances have also facilitated an approach called **metagenomics** (from the Greek *meta*, beyond), in which DNA from a group of species (a *metagenome*) is collected from an environmental sample and sequenced. Again, computer software accomplishes the task of sorting out the partial sequences and assembling them into specific genomes. So far, this approach has been applied to microbial communities found in environments as diverse as the Sargasso Sea and the human intestine. The ability to sequence the DNA of mixed populations eliminates the need to culture each species separately in the lab, a difficulty that has limited the study of many microbial species.

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. Crucial to making sense of this massive amount of data have been new analytical approaches, which we discuss next.

CONCEPT CHECK 18.1

1. Describe the whole-genome shotgun approach to genome sequencing.

For suggested answers, see Appendix A.

CONCEPT 18.2

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 in the 1990s 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 the establishment of banks of data, or databases, and the refining of analytical software. These databases and software programs would then be centralized and made readily accessible on the Internet. Accomplishing this aim has accelerated progress in DNA sequence analysis by making bioinformatics resources available to researchers worldwide and by speeding up the dissemination of information.

Centralized Resources for Analyzing Genome Sequences

Government-funded agencies carried out their mandate to establish databases and provide software with which scientists could analyze the sequence data. For example, in the United States, a joint endeavor between the National Library of Medicine and the National Institutes of Health (NIH) created the National Center for Biotechnology Information (NCBI), which maintains a website (www.ncbi.nlm.nih.gov) with 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 the European Molecular Biology Laboratory, the DNA Data Bank of Japan, and BGI (formerly known as the Beijing Genome Institute) in Shenzhen, China, three genome centers with which NCBI collaborates. 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 2012, it included the sequences of 156 million fragments of genomic DNA, totaling 143 billion base pairs! GenBank is constantly updated, and the amount of data it contains is estimated to double approximately every 18 months. Any sequence in the database can be retrieved and analyzed using software from the NCBI website or elsewhere.

One software program available on the NCBI website, called BLAST, allows the visitor to compare a DNA sequence with every sequence in GenBank, base by base, to look for similar regions. Another program allows comparison of predicted protein sequences. Yet a third can search any protein sequence for 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 18.3)**. There is even a software program that can 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 18.15.)

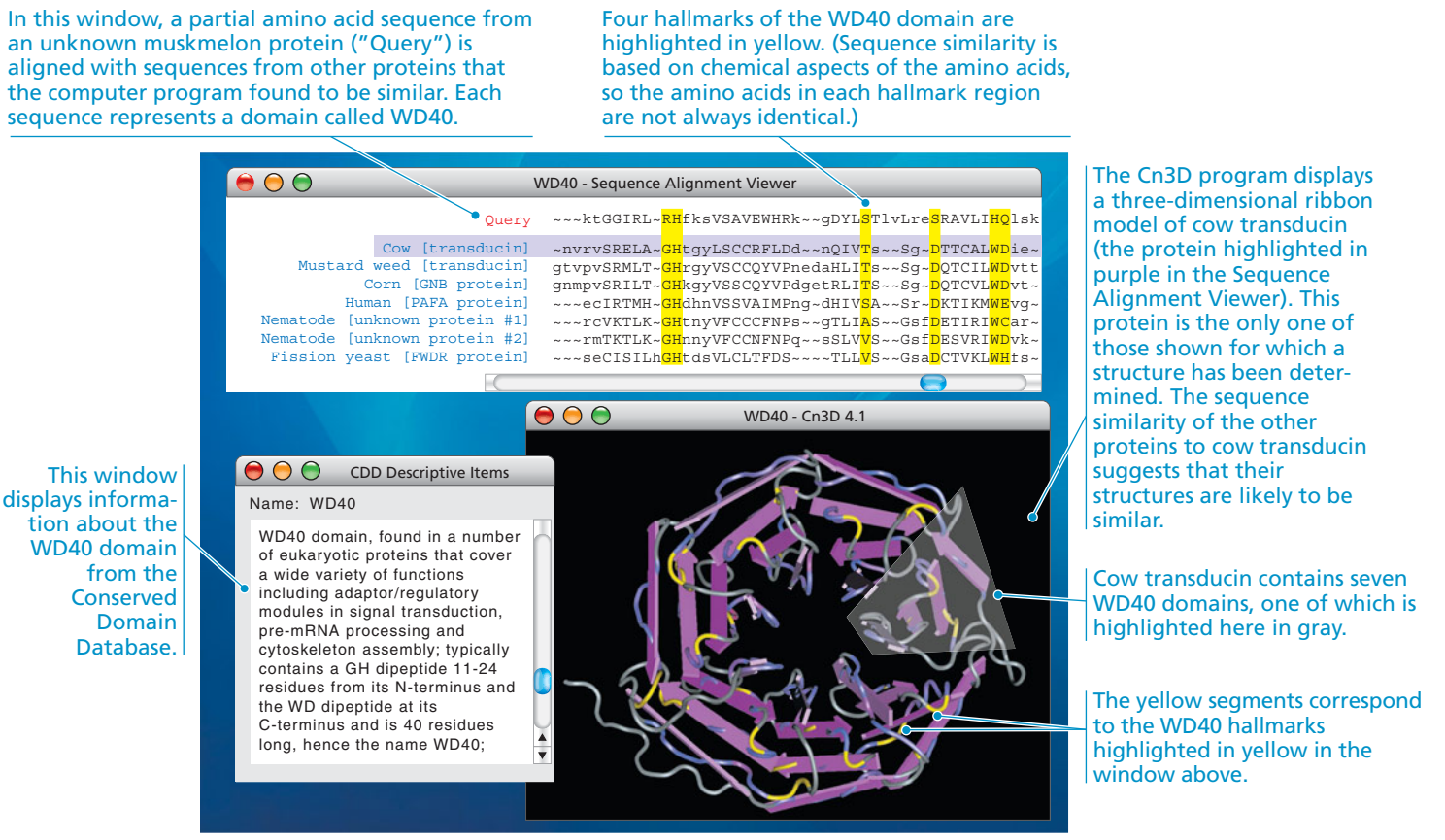
Two research institutions, Rutgers University and the University of California, San Diego, also maintain a worldwide Protein Data Bank, a database of all three-dimensional protein structures that have been determined. (The database is accessible at www.wwpdb.org.) These structures can be rotated by the viewer to show all sides of the protein.

There is a vast array of resources available for researchers anywhere in the world to use. Now let's consider the types of questions scientists can address using these resources.

Understanding the Functions of Protein-Coding Genes

The identities of about half of the human genes were known before the Human Genome Project began. But what about the others, the previously unknown genes revealed by analysis of DNA sequences? Clues about their identities and functions come from comparing sequences that might be genes with known genes from other organisms, using the software described previously. Due to redundancy in the genetic code, the DNA sequence itself may vary more than the protein sequence does. Thus, scientists interested in proteins often compare the predicted amino acid sequence of a protein with that of other proteins.

Sometimes a newly identified sequence will match, at least partially, the sequence of a gene or protein whose function



▲ Figure 18.3 Bioinformatics tools available on the Internet. A website maintained by the National Center for Biotechnology Information allows scientists and the public to access DNA and protein sequences and other stored data. The site includes a link to a protein structure database (Conserved Domain Database, CDD) that can find and describe similar domains in related proteins, as well as software (Cn3D, "See in 3D") that displays three-dimensional models of domains for which the structure has been determined. Some results are shown from a search for regions of proteins similar to an amino acid sequence in a muskmelon protein.

is well known. For example, part of a new gene may match a known gene that encodes an important signaling pathway protein such as a protein kinase (see Chapter 5), suggesting that the new gene does, too. Alternatively, the new gene sequence may 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 *E. coli* when its genome was sequenced. In the last case, 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 blocking or disabling the gene to see how the phenotype is affected.

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, growth and development, and evolution.

One informative approach has been taken by a research project called ENCODE (Encyclopedia of DNA Elements). First, researchers focused intensively on 1% of the human genome and attempted to learn all they could about the functionally important elements in that sequence. They looked for protein-coding genes and genes for noncoding RNAs as well as sequences that regulate DNA replication, gene expression (such as enhancers and promoters), and chromatin modifications. This pilot project, completed in 2007, yielded a wealth of information. One big surprise, discussed in Concept 18.3, was that over 90% of the region was transcribed into RNA, even though less than 2% codes for proteins. The success of this approach led to two follow-up studies, one extending the analysis to the entire human genome and the other analyzing in a similar fashion the genomes of two model organisms, the soil nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. Because genetic and molecular biological experiments can be performed on these species, testing the activities of potentially functional DNA elements in their genomes is expected to reveal much about how the human genome works.

Systems Biology

The success in sequencing genomes and studying entire sets of genes has encouraged scientists to attempt similar systematic study of the full protein sets (*proteomes*) encoded by genomes, an approach called **proteomics**. Proteins, not the genes that encode them, actually carry out most of the activities of the cell. Therefore, we must study when and where proteins are produced in an organism if we are to understand the functioning of cells and organisms.

Genomics and proteomics are enabling molecular biologists to approach the study of life from an increasingly global perspective. Using the tools we have described, biologists have begun to compile catalogs of genes and proteins—listings of 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 to their functional integration in biological systems. This is called the **systems biology** approach, 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, the systems biology approach has really been made possible by advances in computer technology and bioinformatics.

Application of Systems Biology to Medicine

The Cancer Genome Atlas is an example of systems biology in which a large group of interacting genes and gene products are analyzed together. This project, under the joint leadership of the National Cancer Institute and NIH, aims to determine how changes in biological systems lead to cancer. A three-year pilot project that ended in 2010 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 has confirmed the role of several suspected genes and identified a few unknown ones, suggesting possible new targets for therapies. The approach has proved so fruitful for these three types of cancer that it has been extended to ten other types, chosen because they are common and often lethal in humans.

Systems biology has tremendous potential in human medicine that is just starting to be explored. Silicon and glass "chips" have been developed that hold a microarray of most of the known human genes **(Figure 18.4)**. Such chips are being used to analyze gene expression patterns in patients suffering from various cancers and other diseases, with the eventual aim of tailoring their treatment to their unique genetic makeup and the specifics of their cancers. This approach has had modest success in characterizing subsets of several cancers.

Ultimately, people may carry with their medical records a catalog of their DNA sequence, a sort of genetic bar code, with regions highlighted that predispose them to specific diseases. The use of such sequences for personalized medicine—disease prevention and treatment—has great potential.

◀ Figure 18.4 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.

Systems biology is a very efficient way to study emergent properties at the molecular level. Novel properties emerge at each successive level of biological complexity as a result of the arrangement of building blocks at the underlying level (see Chapter 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 will survey what we've learned from genomic studies thus far.

CONCEPT CHECK 18.2

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 more than 90% of the genomic region being studied was transcribed into RNAs, far more than could be accounted for by protein-coding genes. Suggest some roles that these RNAs might play. (Review Concept 15.3.)

For suggested answers, see Appendix A.

CONCEPT 18.3

Genomes vary in size, number of genes, and gene density

By August 2012, the sequencing of about 3,700 genomes had been completed and that of over 7,500 genomes and about 340 metagenomes was in progress. In the completely sequenced group, about 3,300 are genomes of bacteria, and 160 are archaeal genomes. Among the 183 eukaryotic species in the group are vertebrates, invertebrates, protists, fungi, and plants. The accumulated genome sequences contain a wealth of information that we are now beginning to mine. What have we learned so far by comparing the genomes that have been sequenced? In this section, we'll examine the characteristics of genome size, number of genes, and gene density. Because these

characteristics are so broad, we'll focus on general trends, for which there are often exceptions.

Genome Size

Comparing the three domains (Bacteria, Archaea, and Eukarya), we find a general difference in genome size between prokaryotes and eukaryotes **(Table 18.1)**. While there are some exceptions, most bacterial genomes have between 1 and 6 million base pairs (Mb); the genome of *E. coli*, for instance, has 4.6 Mb. Genomes of archaea are, for the most part, 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 genome of *Fritillaria assyriaca*, a flowering plant in the lily family, contains 124 billion base pairs (124,000 Mb), about 40 times the size of the human genome. 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 *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.

Table 18.1 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
Caenorhabditis elegans (nematode)	100	20,100	200
Arabidopsis thaliana (mustard family plant)	120	27,400	228
Drosophila melanogaster (fruit fly)	165	14,000	85
Oryza sativa (rice)	430	40,600	95
Zea mays (corn)	2,300	32,000	14
Mus musculus (house mouse)	2,600	22,000	11
Ailuropoda melanoleuca (giant panda)	2,400	21,000	9
Homo sapiens (human)	3,000	<21,000	7
Fritillaria assyriaca (lily family plant)	124,000	ND	ND

*Some values given here are likely to be revised as genome analysis continues. Mb = million base pairs. ND = not determined.

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 to at least 40,000 for some multicellular eukaryotes (see Table 18.1).

Within the eukaryotes, the number of genes in a species is often lower than expected from simply considering the size of its genome. Looking at Table 18.1, you can see that the genome of the nematode *C. elegans* is 100 Mb in size and contains 20,100 genes. The *Drosophila* genome, in comparison, is much bigger (165 Mb) but has about two-thirds the number of genes—only 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 *Drosophila* 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, and currently, the most reliable count has placed the number at fewer than 21,000. This relatively low number, similar to the number of genes in the nematode *C. elegans*, has surprised biologists, who had clearly expected 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 functional protein from a single gene (see Figure 15.12). 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. It is not yet possible to catalog 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 small RNAs that play regulatory roles have added a new variable to the mix (see Concept 15.3). Some scientists think that this added level of regulation, when present, may contribute to greater organismal complexity for a given number of genes.

Gene Density and Noncoding DNA

In addition to genome size and number of genes, we can compare gene density in different species—in other words, how many genes there 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 18.1). 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 thus far, 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 proceeds from start to finish without interruption 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'll explore the composition and arrangement of these great stretches of DNA in the human genome.

CONCEPT CHECK 18.3

1. According to the best current estimate, the human genome contains fewer than 21,000 genes. However, there is evidence that human cells produce many more than 21,000 different polypeptides. What processes might account for this discrepancy?
2. The number of sequenced genomes is constantly being updated. Go to www.genomesonline.org to find the current number of completed genomes for each domain as well as the number of genomes whose sequencing is in progress. (*Hint*: Click on "Complete Projects" for extra information.)
3. **WHAT IF?** What evolutionary processes might account for prokaryotes having smaller genomes than eukaryotes?

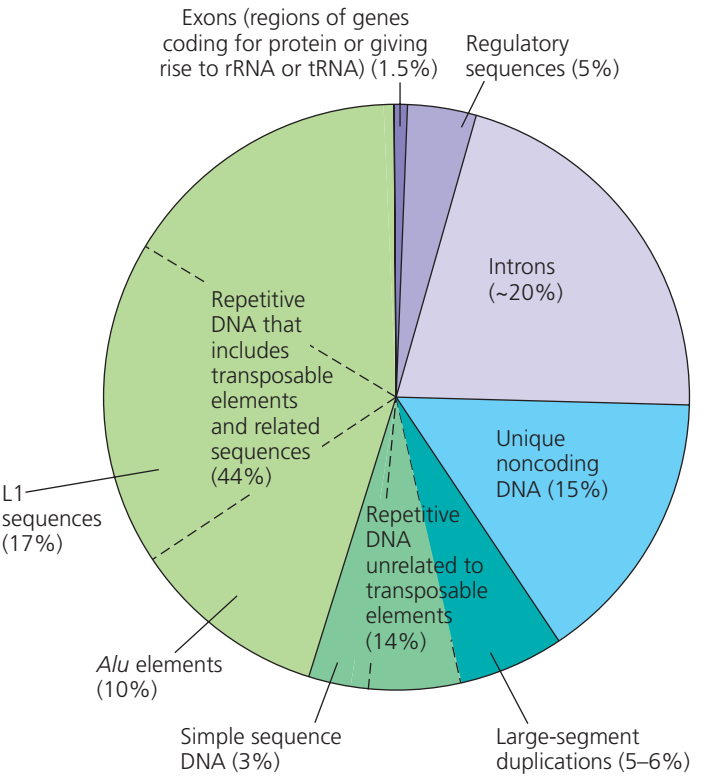
For suggested answers, see Appendix A.

CONCEPT 18.4

Multicellular eukaryotes have much noncoding DNA and many multigene families

We have spent most of this chapter, and indeed this unit, focusing on genes that code for proteins. Yet the coding regions of these genes and the genes for RNA products such as rRNA, tRNA, and miRNA make up only a small portion of the genomes of most multicellular eukaryotes. For example, once the sequencing of the human genome was completed, it became clear that only a tiny part—1.5%—codes for proteins or is transcribed into rRNAs or tRNAs. **Figure 18.5** 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% of the human genome. The rest, located between functional genes, includes some unique noncoding DNA, such as gene fragments and **pseudogenes**, former genes that have accumulated mutations over a long time



▲ Figure 18.5 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 or give rise to known RNAs, and much of it is repetitive DNA (dark and light green and teal). Because repetitive DNA is the most difficult to sequence and analyze, classification of some portions is tentative, and the percentages given here may shift slightly as genome analysis proceeds.

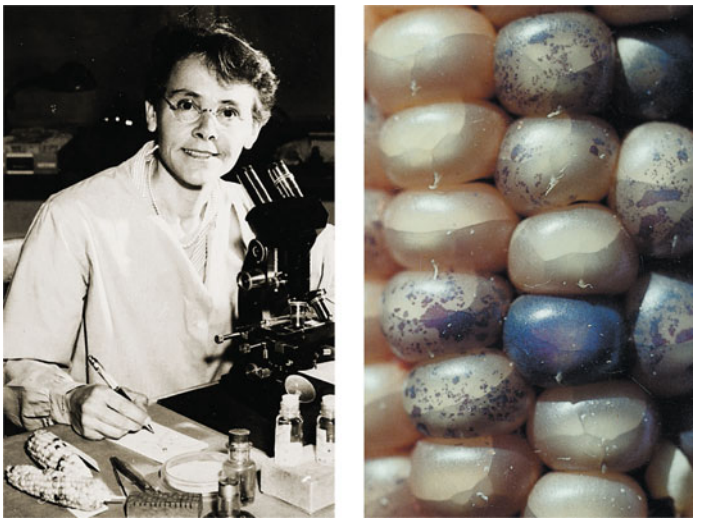
and no longer produce functional proteins. (The genes that produce small noncoding RNAs are a tiny percentage of the genome, distributed between the 20% introns and the 15% unique noncoding DNA.) Most intergenic DNA, however, is **repetitive DNA**, which consists of sequences that are present in multiple copies in the genome. Somewhat surprisingly, about 75% of this repetitive DNA (44% of the entire human genome) is made up of units called transposable elements and sequences related to them.

The bulk of many eukaryotic genomes likewise 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, much evidence is accumulating that this DNA plays important roles in the cell. One measure of its importance is the high degree of sequence conservation between species that diverged many hundreds of generations ago. For example, comparison of the genomes of humans, rats, and mice has revealed the presence of 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. In this section, we'll examine how genes and noncoding DNA sequences are currently organized within genomes of multicellular eukaryotes, using the human genome as our main example. Genome organization tells us much about how genomes have evolved and continue to evolve, the subject of Concept 18.5.

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 it should be kept in mind that they never completely 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.

The first evidence for wandering DNA segments came from American geneticist Barbara McClintock's breeding experiments with Indian corn (maize) in the 1940s and 1950s **(Figure 18.6)**. As she tracked corn plants through multiple generations, McClintock identified changes in the color of corn kernels that made sense only if she postulated the existence of genetic elements capable of moving from other locations in the genome into the genes for kernel color, disrupting the genes so that the kernel color was changed. McClintock's discovery was met with great skepticism and virtually discounted at the time. Her careful work and insightful ideas



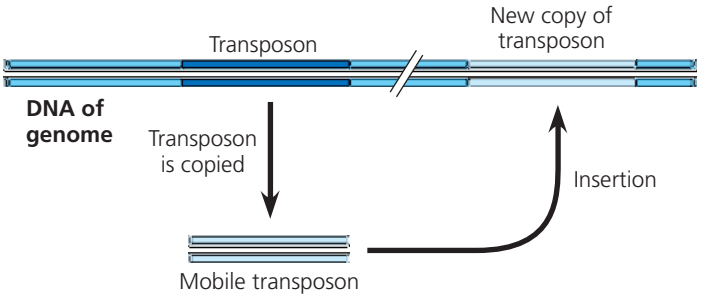
▲ Figure 18.6 The effect of transposable elements on corn kernel color. Barbara McClintock first proposed the idea of mobile genetic elements after observing variegations in corn kernel color (right). She received the Nobel Prize in 1983.

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

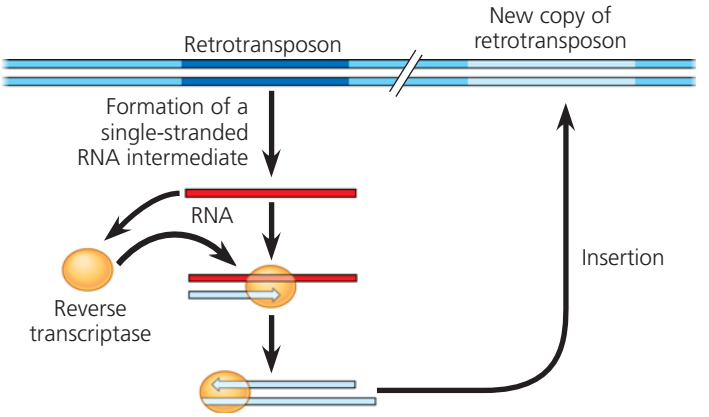
Eukaryotic transposable elements are of two types. The first type, **transposons**, can move within a genome by means of a DNA intermediate. Transposons can move by a "cut-and-paste" mechanism, which removes the element from the original site, or by a "copy-and-paste" mechanism, which leaves a copy behind **(Figure 18.7)**. 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. Retrotransposons always leave a copy at the original site during transposition, since they are initially transcribed



▲ Figure 18.7 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.

? *How would this figure differ if it showed the cut-and-paste mechanism?*



▲ Figure 18.8 Retrotransposon movement. Movement begins with formation of a single-stranded RNA intermediate. The remaining steps are essentially identical to part of the retrovirus replicative cycle (see Figure 17.7).

into an RNA intermediate **(Figure 18.8)**. 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 Chapter 17. In fact, retroviruses may have evolved from retrotransposons.) Another cellular enzyme catalyzes insertion of the reverse-transcribed DNA at a new site.

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 18.5) and even higher percentages in amphibians and many plants. In fact, the very large size of some plant genomes is accounted for not by extra genes, but by extra transposable elements. For example, such sequences 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; its cellular function, if any, is currently unknown.

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 have a low rate of transposition. An accompanying genomic analysis found L1 sequences within the introns of nearly 80% of the human genes that were analyzed, suggesting that L1 may help regulate gene expression. Other researchers have proposed that L1 retrotransposons may have differential effects on gene expression in developing neurons, contributing to the great diversity of neuronal cell types (see Chapter 37).

Although many 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 probably arises due to mistakes during DNA replication or recombination. Such DNA accounts for about 14% of the human genome (see Figure 18.5). 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. The large 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, **simple sequence DNA** contains 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. Repeated units may contain as many as 500 nucleotides, but often contain fewer than 15 nucleotides, as in this example. When the unit contains 2–5 nucleotides, the series of repeats is called a **short tandem repeat (STR)**. 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.

The repeat number also varies from person to person, and since humans are diploid, each person has two alleles per site, which can differ. This diversity produces variation that can be used to identify a unique set of genetic markers for each individual, his or her **genetic profile**. Forensic scientists can use STR analysis on DNA extracted from samples of tissues or body fluids to identify victims of a crime scene or natural disaster. In such an application, STR analysis is performed on STR sites selected because they have relatively few repeats and are easily sequenced. This technique has also been used by The Innocence Project, a nonprofit organization, to free more than 250 wrongly convicted people from prison.

Altogether, simple sequence DNA makes up 3% of the human genome. Much of a genome's simple sequence DNA is

located at chromosomal telomeres and centromeres, suggesting that this DNA plays a structural role for chromosomes. The DNA at centromeres is essential for the separation of chromatids in cell division (see Chapter 9). Centromeric DNA, along with simple sequence DNA located elsewhere, may also help organize the chromatin within the interphase nucleus. The simple sequence DNA located at telomeres, at the tips of chromosomes, binds proteins that protect the ends of a chromosome from degradation and from joining to other chromosomes.

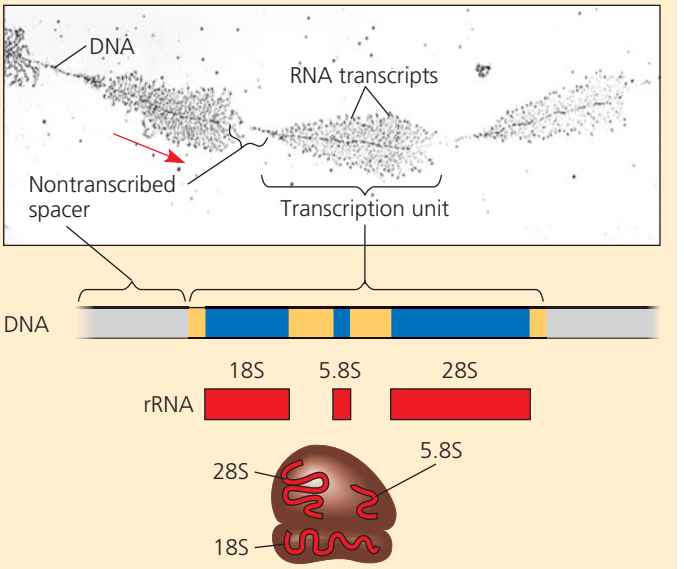
Genes and Multigene Families

We finish our discussion of the various types of DNA sequences in eukaryotic genomes with a closer look at genes. Recall that DNA sequences that code for proteins or give rise to tRNA or rRNA compose a mere 1.5% of the human genome (see Figure 18.5). If we include introns and regulatory sequences associated with genes, 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.

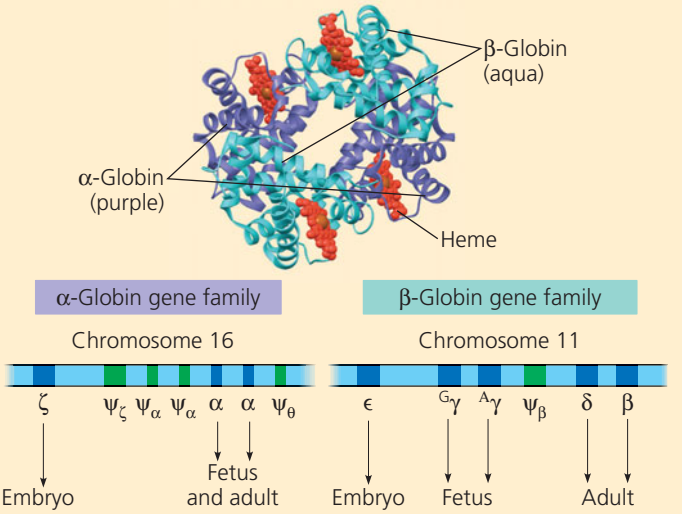
Like the genes of bacteria, many eukaryotic genes are present as unique sequences, with only one copy per haploid set of chromosomes. But in the human genome and the genomes of many other animals and plants, solitary genes make up less than half of the total gene-related DNA. The rest occur in **multigene families**, collections of two or more identical or very similar genes.

In multigene families consisting 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 are the genes for the three largest rRNA molecules **(Figure 18.9a)**. 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 the three rRNA molecules, which combine with proteins and one other kind of rRNA (5S 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 18.9b)**. 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



(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 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. In the diagram of a transcription unit below the TEM, the genes for three types of rRNA (blue) are adjacent to regions that are transcribed but later removed (yellow). 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 (dark blue) encoding α- and β-globins are found in two families, organized as shown here. The noncoding DNA separating the functional genes within each family includes pseudogenes (ψ; green), versions of the original genes that no longer produce functional proteins. Genes and pseudogenes are named with Greek letters. Some genes are expressed only in the embryo or fetus.

▲ Figure 18.9 Gene families.

? *In (a), how could you determine the direction of transcription if it weren't indicated by the red arrow?*

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 (green in Figure 18.9b).

We'll return to the globin gene family to consider the evolutionary history of these gene clusters in the next section. We'll also consider some of the processes that have shaped the genomes of different species over evolutionary time.

CONCEPT CHECK 18.4

1. Discuss the characteristics of mammalian genomes that make them larger than prokaryotic genomes.
2. Which of the three mechanisms described in Figures 18.7 and 18.8 result(s) in a copy remaining at the original site as well as 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.

For suggested answers, see Appendix A.

CONCEPT 18.5

Duplication, rearrangement, and mutation of DNA contribute to genome evolution

EVOLUTION 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 describe 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

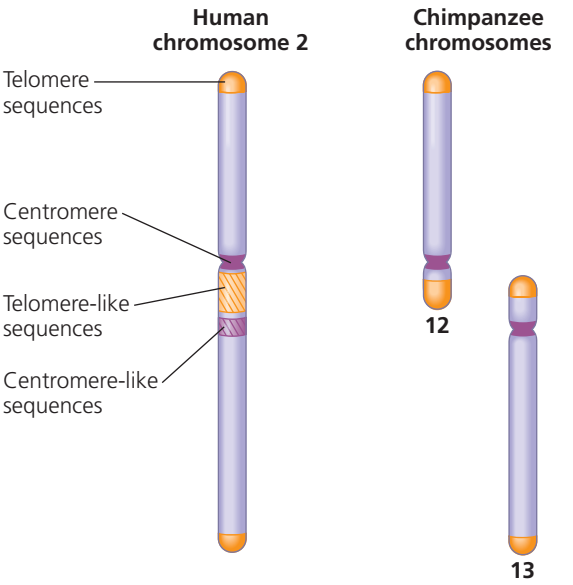
An accident in meiosis can result in one or more extra sets of chromosomes, a condition known as polyploidy. Although such accidents would most often be lethal, in rare cases they could facilitate the evolution of genes. In a polyploid organism, one set of genes can provide essential functions for the 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 be the branching off of a new species, as happens often in flowering plants (see Chapter 22).

Polyploid animals also exist, but they are much rarer; the tetraploid model organism *Xenopus laevis*, the African clawed frog, is an example.

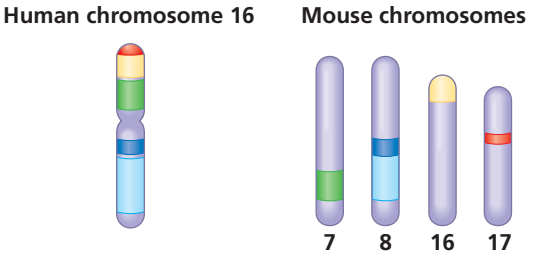
Alterations of Chromosome Structure

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 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 chimp chromosomes 12 and 13 fused end to end, forming chromosome 2 in an ancestor of the human lineage. With the recent explosion in genomic sequence information, we can now compare the chromosomal organizations of many different species on a much finer scale. This information allows us to make inferences about the evolutionary processes that shape chromosomes and may drive speciation. Sequencing and analysis of human chromosome 2 provided very strong supporting evidence for the model we have just described **(Figure 18.10)**.

In another study of broader scope, researchers compared the DNA sequence of each human chromosome with the whole-genome sequence of the mouse. For human chromosome 16, the comparison revealed that large blocks of genes on this chromosome are found on four mouse chromosomes. This



▲ Figure 18.10 Related human and chimpanzee chromosomes. The positions of telomere-like and centromere-like sequences on human chromosome 2 (left) match those of telomeres on chimp 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. (Chimp chromosomes 12 and 13 have since been renamed 2a and 2b, respectively.)



▲ Figure 18.11 A comparison of human and mouse chromosomes. 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 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.

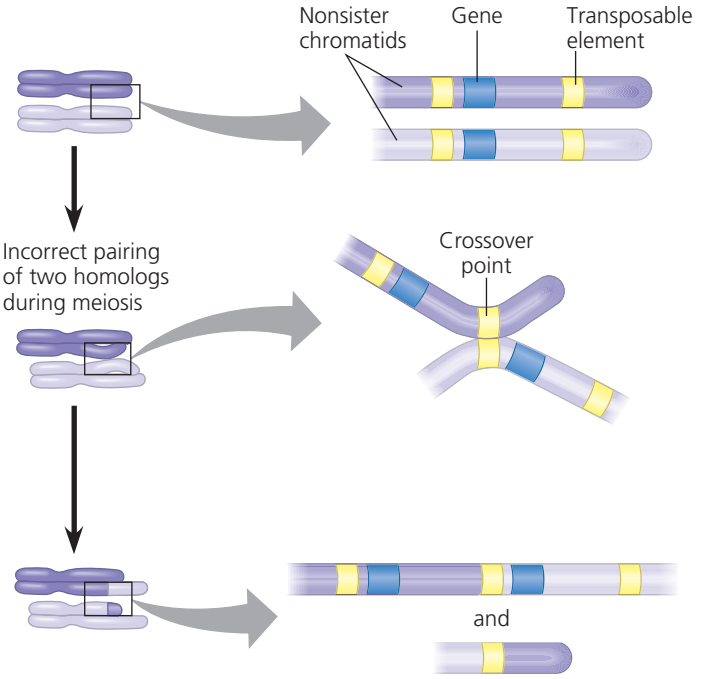
observation suggests that the genes in each block stayed together during the evolution of the mouse and human lineages **(Figure 18.11)**.

Performing the same comparative analysis between 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 mistakes during meiotic recombination in which the DNA broke and was rejoined incorrectly. The rate of these events seems to have accelerated about 100 million years ago, around the time large dinosaurs became extinct and the number of mammalian species increased rapidly. 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 reduce the success of matings between members of the two populations, a step on the way to the populations becoming two separate species. (You'll learn more about this in Chapter 22.)

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. As illustrated in **Figure 18.12**, transposable elements can provide homologous sites where nonsister chromatids can cross over, even when other chromatid regions are not correctly aligned.

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,



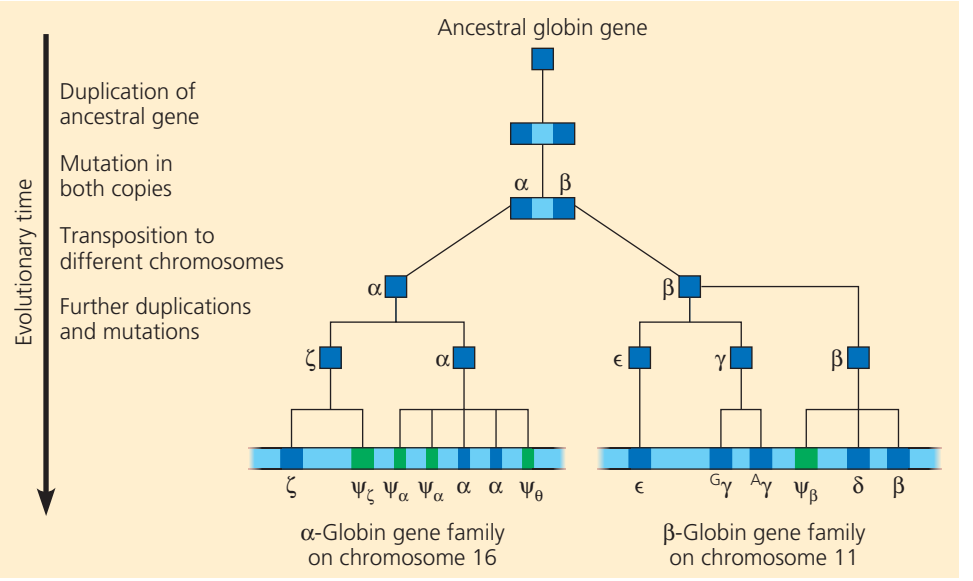
▲ Figure 18.12 Gene duplication due to unequal crossing over. One mechanism by which a gene (or other DNA segment) can be duplicated is recombination during meiosis between copies of a transposable element flanking the gene. Such recombination between misaligned nonsister chromatids of homologous chromosomes produces one chromatid with two copies of the gene and one chromatid with no copy.

MAKE CONNECTIONS *Recall how crossing over occurs (see Figure 10.11). In the middle panel above, draw a line along the portions that result in the upper chromatid in the bottom panel. Use a different color to do the same for the other chromatid.*

a segment of DNA is deleted or duplicated. It is easy to imagine how such errors could occur in regions of repeats. 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.

Evolution of Genes with Related Functions: The Human Globin Genes

Duplication events can lead to the evolution of genes with related functions, such as those of the α-globin and β-globin gene families (see Figure 18.9b). A comparison of gene sequences within a multigene family can suggest the order in which the genes arose. This approach to re-creating the evolutionary history of the globin genes 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 18.13)**. 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



▲ Figure 18.13 A model for the evolution of the human α-globin and β-globin gene families from a single ancestral globin gene.

? *The green elements are pseudogenes. Explain how they could have arisen after gene duplication.*

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, but a few mutations must have altered the function of the protein product in a way that was advantageous to 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 members and see how such comparisons were used to generate the model for globin gene evolution shown in Figure 18.13. The existence of several pseudogenes among the functional globin genes provides additional evidence for this model (see Figure 18.9b): 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 good examples.

Lysozyme is an enzyme that helps protect animals against bacterial infection by hydrolyzing bacterial cell walls; 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. Both genes are found in mammals, whereas 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 different function.

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 discuss next.

Rearrangements of Parts of Genes: Exon Duplication and Exon Shuffling

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.

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 18.12). 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 could 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 with a highly repetitive amino acid sequence, which is reflected in the repetitive pattern of exons in the collagen gene.

Alternatively, 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

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 18.13), researchers compared the amino acid sequences of the polypeptides they encode. In this exercise, you will analyze comparisons of the amino acid sequences of 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 arranged 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.

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

The percent identity value for the α_1- and ζ-globin amino acid sequences was calculated by counting the number of matching amino acids (87, highlighted in yellow), dividing by the total number of amino acid positions (143), and then multiplying by 100. This resulted in a 61% 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.

Interpret the Data

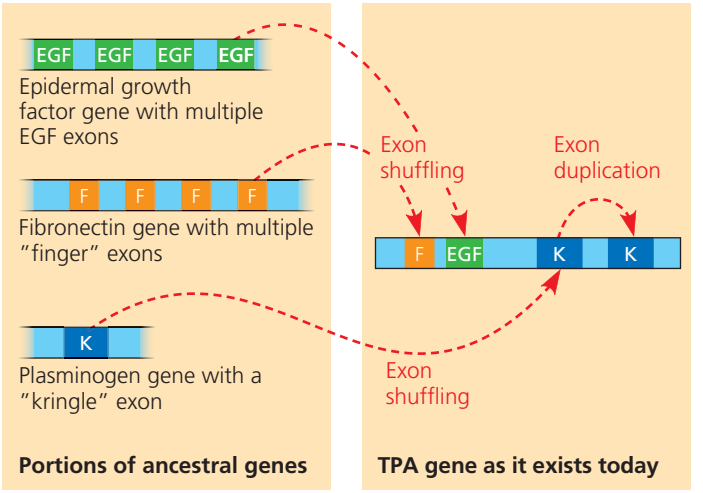
1. Notice that in the amino acid identity table, the data are arranged so each globin pair can be compared. (a) Notice that some cells in the table have dashes. Given the pairs that are being compared for these cells, what percent identity value is implied by the dashes? (b) Notice that the cells in 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?
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 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 18.13 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 a 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 18.13. 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 group to the values for between-group pairs.

Further Reading R. C. Hardison, Globin genes on the move, *Journal of Biology* 7:35.1–35.5 (2008). doi:10.1186/jbiol92

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Amino Acid Identity Table

		α Family			β Family				
		α_1 (alpha 1)	α_2 (alpha 2)	ζ (zeta)	β (beta)	δ (delta)	ε (epsilon)	$^{A}\gamma$ (gamma A)	$^{G}\gamma$ (gamma G)
α Family	α_1	------	100	61	45	44	39	42	42
	α_2		------	61	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
	$^{A}\gamma$							------	99
	$^{G}\gamma$								------



▲ Figure 18.14 Evolution of a new gene by exon shuffling. Exon shuffling 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). Duplication of the "kringle" exon from the plasminogen gene after its movement could account for the two copies of this exon in the TPA gene.

? *How could the presence of transposable elements in introns have facilitated the exon shuffling shown here?*

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; one exon is present in two copies. Because each type of exon is also found in other proteins, the gene for TPA is thought to have arisen by several instances of exon shuffling and duplication **(Figure 18.14)**.

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 chromosomes by providing homologous regions for crossing over. 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, if a transposable element "jumps" into the middle of a protein-coding sequence, it will prevent the production of a normal transcript of the gene. 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 prove beneficial by providing a survival advantage.

During transposition, a transposable element may carry along a gene or group of genes to a new position in the genome. This mechanism 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.

All the processes discussed in this section most often produce either harmful effects, which may be lethal, or no effect at all. In a few cases, however, small beneficial heritable changes 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 and has increased our understanding of how genomes evolve. You'll learn more about these topics in the final section.

CONCEPT CHECK 18.5

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 18.14 (left).
3. What are three ways that transposable elements are thought to contribute to genome evolution?

For suggested answers, see Appendix A.

CONCEPT 18.6

Comparing genome sequences provides clues to evolution and development

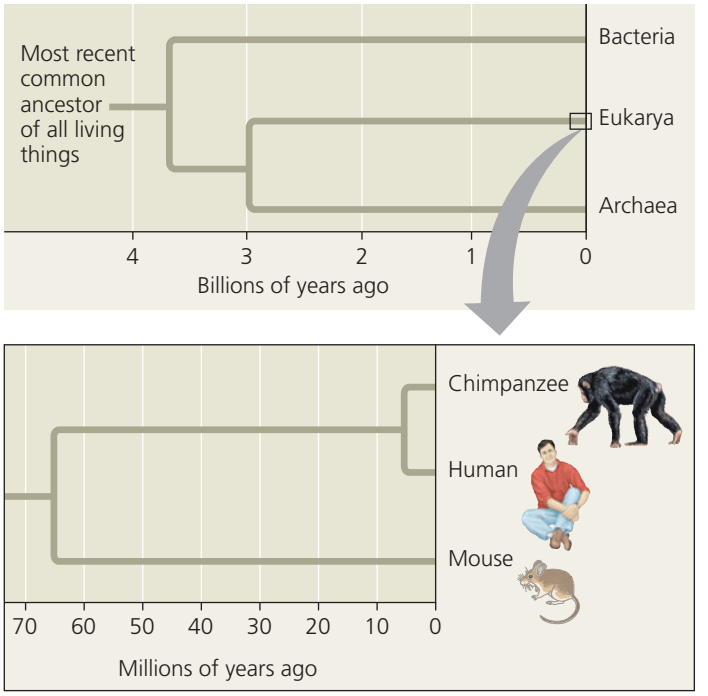
EVOLUTION One researcher has likened the current state of biology to the Age of Exploration in the 15th century after major improvements in navigation and the building of faster ships. In the last 25 years, we have seen rapid advances in genome sequencing and data collection, new techniques for

assessing gene activity across the whole genome, and refined approaches for understanding how genes and their products work together in complex systems. We are truly poised on the brink of a new world.

Comparisons of genome sequences from different species reveal much 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 beginning to clarify the mechanisms that generated the great diversity of life-forms present today. In this final section of the chapter, we'll discuss what has been learned from these two approaches.

Comparing Genomes

The more similar in sequence the genes and genomes of two species are, 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 life-forms 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 Chapter 1). **Figure 18.15** shows the evolutionary relationships of some groups and species we will be discussing. We'll consider comparisons between distantly related species first.



▲ Figure 18.15 Evolutionary relationships of the three domains of life. This tree diagram 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 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 complete genome 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 18.15).

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. Genes that evolved a very long time ago can still be surprisingly similar in disparate species. As a case in point, several genes in yeast are so similar to certain human disease genes that researchers have deduced the functions of the disease genes by studying their yeast counterparts. This striking similarity underscores the common origin of these two distantly related species.

Comparing Closely Related Species

The genomes of two closely related species are likely to be organized similarly because of their relatively recent divergence. This allows the fully sequenced genome of one species to be used as a scaffold for assembling the genomic sequences of a closely related species, accelerating mapping of the second genome. For instance, using the human genome sequence as a guide, researchers were able to quickly sequence the entire chimpanzee genome.

The recent divergence of two closely related species also underlies the small number of gene differences that are found when their genomes are compared. The particular genetic differences can therefore 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 should give clues about what it takes to make a mammal, while finding the genes shared by chimpanzees and humans but not by rodents should tell us something about primates. And, of course, comparing the human genome with that of the chimpanzee should help us answer the tantalizing question we asked at the beginning of the chapter: What genomic information makes a human or a chimpanzee?

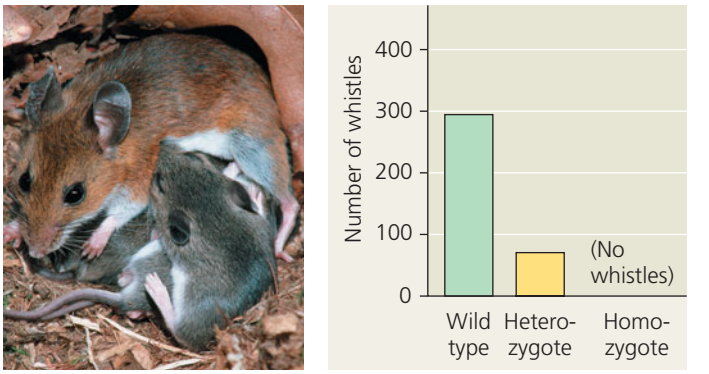
An analysis of the overall composition of the human and chimpanzee genomes, which are thought to have diverged only about 6 million years ago (see Figure 18.15), reveals some general differences. Considering single nucleotide substitutions, the two genomes differ by only 1.2%. When researchers looked

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. There are more *Alu* elements in the human genome than in the chimpanzee genome, and the latter 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. We also don't know how these differences might account for the distinct characteristics of each species.

To discover the basis for the phenotypic differences between the two species, biologists are studying specific genes and types of genes that differ between humans and chimpanzees 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 and 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 whose gene shows evidence of rapid change in the human lineage is called *FOXP2*. Several lines of evidence suggest that the *FOXP2* gene functions in vocalization in vertebrates. For one thing, mutations in this gene can produce severe speech and language impairment in humans. Moreover, the *FOXP2* gene is expressed in the brains of zebra finches and canaries at the time when these songbirds are learning their songs. But perhaps the strongest evidence comes from a "knock-out" experiment in which researchers disrupted the *FOXP2* gene in mice and analyzed the resulting phenotype. Normal mice produce ultrasonic squeaks (whistles) to communicate stress, but mice that were homozygous for a mutated form of *FOXP2* had malformed brains and failed to vocalize normally **(Figure 18.16)**. Heterozygous mice, with one faulty copy of the gene, also showed vocalization defects. These results augmented the evidence from birds and humans, supporting the idea that the *FOXP2* gene product turns on genes involved in vocalization.

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 diverged about 65.5 million years ago (see Figure 18.15) and share about 85% of their genes. This genetic similarity can be exploited in studying



▲ Figure 18.16 The function of *FOXP2*, a gene that is rapidly evolving in the human lineage. (a) Wild-type mice emit ultrasonic squeaks (whistles) to communicate stress. **(b)** Researchers used genetic engineering to produce mice in which one or both copies of *FOXP2* were disrupted, separated each newborn pup from its mother, and recorded the number of ultrasonic whistles produced by the pup. No vocalization was observed in homozygous mutants, and the effect on heterozygotes was also extreme.

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.

Further research efforts are under way to extend genomic studies to many more microbial species, additional primates, and neglected species from diverse branches of the tree of life. These studies will advance our understanding of all aspects of biology, including health and ecology as well as evolution.

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 with that of many other species. Much of our diversity seems to be in the form of **single nucleotide polymorphisms** (**SNPs**, pronounced "snips"), defined as single base-pair sites where variation is found in at least 1% of the population. Usually detected by DNA sequencing, SNPs occur on average about once in 100–300 base pairs in the human genome. Scientists have already identified the location of several million human SNP sites and continue to find more.

In the course of this search, they have also found other variations—including 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 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 casts doubt on the meaning of the phrase "a normal human genome."

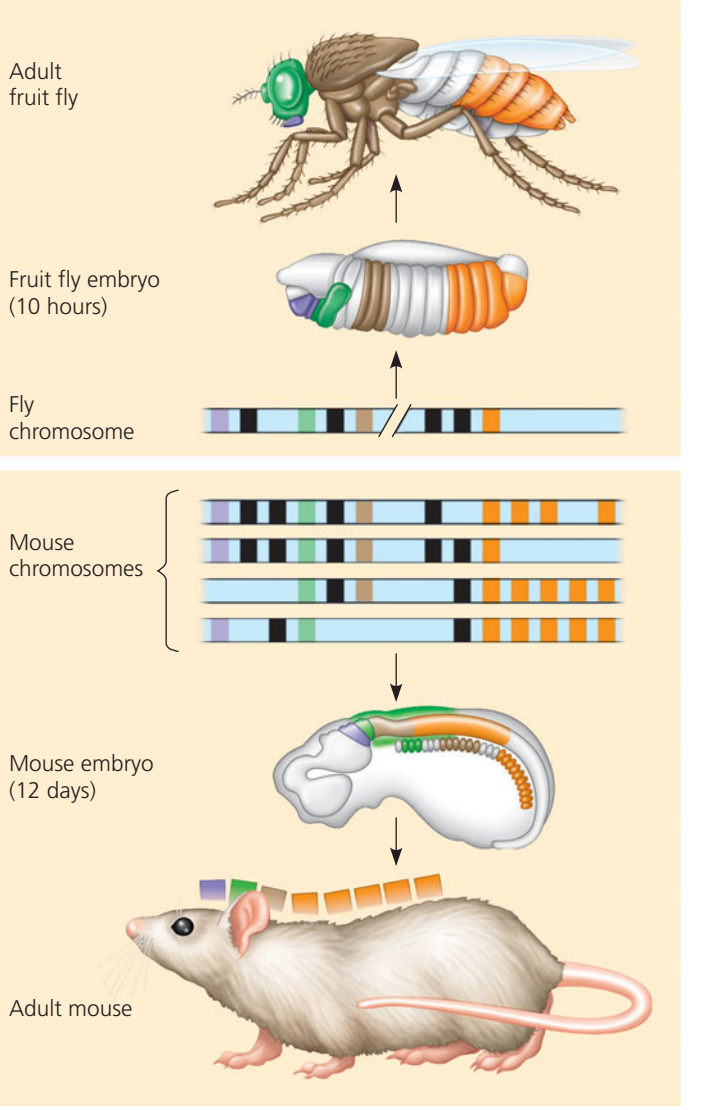
Copy-number variants, SNPs, and variations in repetitive DNA such as short tandem repeats (STRs) will be useful genetic markers for studying human evolution. In 2010, the genomes of two Africans from different communities were sequenced: Archbishop Desmond Tutu, the South African civil rights advocate and a member of the Bantu tribe, the majority population in southern Africa; and !Gubi, a hunter-gatherer from the Khoisan community in Namibia, a minority African population that is probably 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 community members (self-identified Bushmen) living nearby. Remarkably, these four Khoisan genomes differed more from each other than a European would from an Asian. These data highlight the extensive diversity among African genomes. Extending this approach will help us answer important questions about the differences between human populations and the migratory routes of human populations throughout history.

Comparing Developmental Processes

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 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 evolution.

Widespread Conservation of Developmental Genes Among Animals

You may recall that the homeotic genes in *Drosophila* specify the identity of body segments in the fruit fly (see Figure 16.8). Molecular analysis of the homeotic genes in *Drosophila* has shown that they all include a 180-nucleotide sequence called a **homeobox**, which specifies 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 sequences are so similar between humans and fruit flies, in fact, that one researcher has whimsically referred to flies as "little people with wings." 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 18.17)**. 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 in the history of life and was sufficiently beneficial to organisms to have been conserved in animals and plants virtually unchanged for hundreds of millions of years.



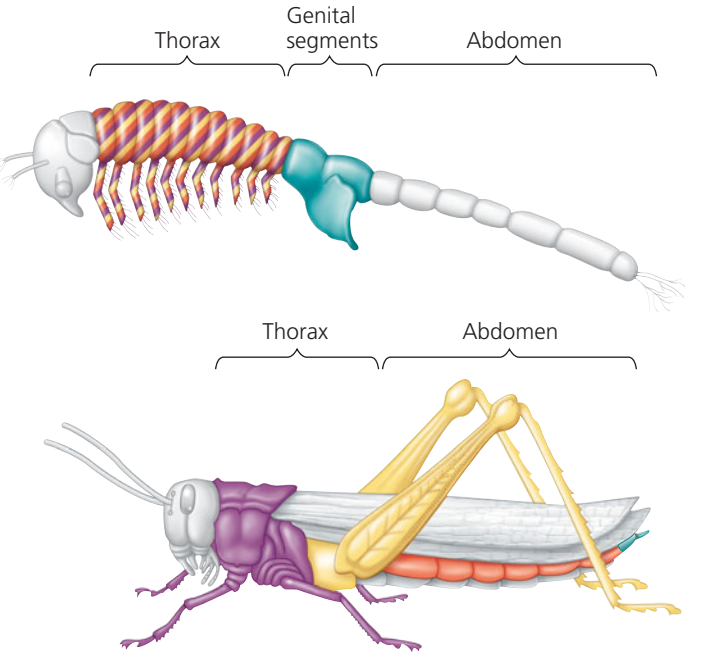
▲ Figure 18.17 Conservation of homeotic genes in a fruit fly and a mouse. Homeotic genes that control the form of anterior and posterior structures of the body occur in the same linear sequence on chromosomes in *Drosophila* and mice. Each colored band on the chromosomes shown here represents a homeotic 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.

Homeotic genes in animals were named *Hox* genes, short for *h*omeob*ox*-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 16.9 and 16.10), in several of the segmentation genes, and in a master regulatory gene for eye development.

Researchers have discovered that the homeobox-encoded homeodomain is the part of a protein that binds to DNA when the protein functions as a transcriptional regulator. However, the shape of the homeodomain allows it to bind to any DNA segment; its own structure is not specific for a particular sequence. Instead, other, more variable domains in a homeodomain-containing protein determine which genes the protein regulates. Interaction of these variable domains with still other transcription factors helps a homeodomain-containing protein recognize specific enhancers in the DNA. 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 particular 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?

Ongoing studies are suggesting answers to this question. 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 insects and crustaceans can explain the variation in the number of leg-bearing segments among these segmented animals **(Figure 18.18)**. Also, recent research suggests that the same *Hox* gene product may have subtly dissimilar effects in different species, turning on new genes or turning on the same genes at higher or lower levels. In other cases, similar genes direct different developmental processes in different 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 (see Figure 6.4). They are among the organisms long used in classical embryological studies (see Chapter 36).



▲ Figure 18.18 Effect of differences in *Hox* gene expression in crustaceans and insects. Changes in the expression patterns of *Hox* genes have occurred over evolutionary time. These changes account in part for the different body plans of the brine shrimp *Artemia*, a crustacean (top), and the grasshopper, an insect. Shown here are regions of the adult body color-coded for expression of four *Hox* genes that determine formation of particular body parts during embryonic development. Each color represents a specific *Hox* gene. Colored stripes on the thorax of *Artemia* indicate co-expression of three *Hox* genes.

In this final chapter of the genetics unit, you have learned how studying genomic composition and comparing the genomes of different species can disclose much about how genomes evolve. Further, 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 remainder of the book, we expand our perspective beyond the level of molecules, cells, and genes to explore this diversity on the organismal level.

CONCEPT CHECK 18.6

1. Would you expect the genome of the macaque (a monkey) to be more similar to the mouse genome or the human genome? Why?
2. The DNA sequences called homeoboxes, which help homeotic genes in animals direct development, are common to flies and mice. Given this similarity, explain why these animals are so different.

For suggested answers, see Appendix A.

18 Chapter Review

SUMMARY OF KEY CONCEPTS

CONCEPT 18.1

The Human Genome Project fostered development of faster, less expensive sequencing techniques (pp. 344–345)

- 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 complete sequence.

? *How did the Human Genome Project result in more rapid, less expensive DNA sequencing?*

CONCEPT 18.2

Scientists use bioinformatics to analyze genomes and their functions (pp. 345–347)

- Computer analysis of genome sequences aids the identification of protein-coding sequences. Methods for determining gene function include comparing the sequences of newly discovered genes with those of known genes in other species, and also observing the phenotypic effects of experimentally inactivating genes whose functions are unknown.
- In **systems biology**, researchers aim to model the dynamic behavior of whole biological systems based on the study of the interactions among the system's parts. For example, scientists use the computer-based tools of **bioinformatics** to compare genomes and to study sets of genes and proteins as whole systems (**genomics** and **proteomics**). These studies include large-scale analyses of functional DNA elements.

? *What was the most significant finding of the ENCODE pilot project? Why has the project been expanded to include other species?*

CONCEPT 18.3

Genomes vary in size, number of genes, and gene density (pp. 347–349)

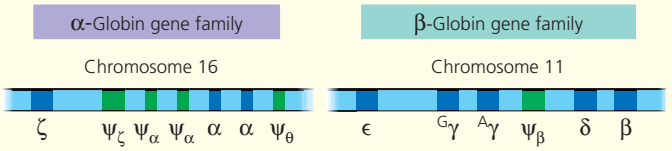
	Bacteria	Archaea	Eukarya
Genome size	Most are 1–6 Mb		Most are 10–4,000 Mb, but a few are much larger
Number of genes	1,500–7,500		5,000–40,000
Gene density	Higher than in eukaryotes		Lower than in prokaryotes (Within eukaryotes, lower density is correlated with larger genomes.)
Introns	None in protein-coding 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 be 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 18.4

Multicellular eukaryotes have much noncoding DNA and many multigene families (pp. 349–353)

- 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 are members of a family of related genes, such as the human globin gene families:



? *Explain how the function of transposable elements might account for their prevalence in human noncoding DNA.*

CONCEPT 18.5

Duplication, rearrangement, and mutation of DNA contribute to genome evolution (pp. 353–357)

- Accidents 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.
- 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 globin proteins evolved from one common ancestral globin gene, which duplicated and diverged into 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.
- Each exon may code for a **domain**, a discrete structural and functional region of a protein. 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, which can alter the functions of genes or their patterns of expression and regulation.

? *How could chromosomal rearrangements lead to the emergence of new species?*

CONCEPT 18.6

Comparing genome sequences provides clues to evolution and development (pp. 357–361)

- Comparative studies of genomes from widely divergent and closely related species provide valuable information about ancient and more recent evolutionary history, respectively. Human and chimpanzee sequences are about 4% different. Along with nucleotide variations in specific genes, these differences may account for the distinct characteristics of the two species. Analysis of **single nucleotide polymorphisms (SNPs)** and copy-number variants (CNVs) within 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. Related sequences are present in the genes of plants and yeasts.

? *What type of information can be obtained by comparing the genomes of closely related species? Of very distantly related species?*

TEST YOUR UNDERSTANDING

Level 1: Knowledge/Comprehension

1. Bioinformatics includes all of the following except
- **a.** using computer programs to align DNA sequences.
- **b.** analyzing protein interactions in a species.
- **c.** using molecular biology to combine DNA from two different sources in a test tube.
- **d.** developing computer-based tools for genome analysis.
- **e.** using mathematical tools to make sense of biological systems.

2. One of the characteristics of retrotransposons is that
- **a.** they code for an enzyme that synthesizes DNA using an RNA template.
- **b.** they are found only in animal cells.
- **c.** they generally move by a cut-and-paste mechanism.
- **d.** they contribute a significant portion of the genetic variability seen within a population of gametes.
- **e.** their amplification is dependent on a retrovirus.

3. 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.
- **e.** are responsible for differentiation in muscle cells.

Level 2: Application/Analysis

4. 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
- **b.** alternative splicing
- **c.** exon shuffling
- **d.** histone modification
- **e.** random point mutations

5. Two eukaryotic proteins are identical except for one domain in each protein, and these two domains are completely different from each other. Which of the following processes is most likely to have contributed to this difference?
- **a.** gene duplication
- **b.** alternative splicing
- **c.** exon shuffling
- **d.** histone modification
- **e.** random point mutations

6. DRAW IT Below are the amino acid sequences (using the single-letter code; see Figure 3.17) of four short segments of the *FOXP2* protein from six species: chimpanzee (C), orangutan (O), gorilla (G), rhesus macaque (R), mouse (M), and human (H). These segments contain all of the amino acid differences between the *FOXP2* proteins of these species.

1. ATETI...PKSSD...TSSTT...NARRD
2. ATETI...PKSSE...TSSTT...NARRD
3. ATETI...PKSSD...TSSTT...NARRD
4. ATETI...PKSSD...TSSNT... SARRD
5. ATETI...PKSSD...TSSTT...NARRD
6. VTETI...PKSSD...TSSTT...NARRD

Use a highlighter to color any amino acid that varies among the species. (Color that amino acid in all sequences.)

- **(a)** The C, G, R sequences are identical. Which lines correspond to those sequences?
- **(b)** The H sequence differs from that of the C, G, R species at two amino acids. Underline the two differences in the H sequence.
- **(c)** The O sequence differs from the C, G, R sequences at one amino acid (having V instead of A) and from the H sequence at three amino acids. Which line is the O sequence?
- **(d)** In the M sequence, circle the amino acid(s) that differ from the C, G, R sequences, and draw a square around those that differ from the H sequence. Describe these differences.
- **(e)** Primates and rodents diverged between 60 and 100 million years ago, and chimpanzees and humans, about 6 million years ago. What can you conclude by comparing the amino acid differences between the mouse and the C, G, R species with those between the human and the C, G, R species?

Level 3: Synthesis/Evaluation

7. SCIENTIFIC INQUIRY
The scientists mapping human SNPs noticed that groups of SNPs tended to be inherited together, in blocks known as haplotypes, ranging from 5,000 to 200,000 base pairs long. There are only four or five commonly occurring combinations of SNPs per haplotype. Propose an explanation, integrating what you've learned throughout this chapter and this unit.

8. FOCUS ON EVOLUTION
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. Why is this?

9. FOCUS ON 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.

For selected answers, see Appendix A.