Chapter 17

# Biotechnology


## Chapter Outline



## Introduction

*Over the past decades, the development of new and powerful techniques for studying and manipulating DNA has revolutionized biology. The knowledge gained in the last 25 years is greater than that accrued during the history of biology. Biotechnology also affects more aspects of everyday life than any other area of biology. From the food on your table to the future of medicine, biotechnology touches your life.*

*Biotechnology is the application of molecular biology principles to numerous aspects of life. The ability to isolate specific DNA sequences arose from the study and use of small DNA molecules found in bacteria, like the plasmid pictured here. In this chapter, we explore these technologies and consider how they apply to specific problems of practical importance.*

## 17.1 DNA Manipulation

### Learning Outcomes

1. *Relate endogenous roles of enzymes to their recombinant DNA applications.*
2. *Explain why DNA fragments can be separated with gel electrophoresis.*

The ability to directly isolate and manipulate genetic material was one of the most profound changes in the field of biology in the late 20th century. The construction of **recombinant DNA** molecules, that is, a single DNA molecule made from two different sources, began in the mid-1970s. The development of this technology, which has led to the entire field of biotechnology, is based on enzymes that can be used to manipulate DNA.

### Restriction enzymes cleave DNA at specific sites

Enzymes called **restriction endonucleases** revolutionalized molecular biology because of their ability to cleave DNA at specific sites. As described in chapter 14, nucleases are enzymes that degrade DNA, and many were known prior to the isolation of the first restriction enzyme (*Hin*dII) in 1970. If a DNA sequence were a rope, then restriction enzymes would be a knife that always cut that rope into specific pieces.

### *Discovery and significance of restriction endonucleases*

This site-specific cleavage activity, long sought by molecular biologists, was discovered from basic research into why bacterial viruses can infect some cells but not others. This phenomenon was termed *host restriction*. The bacteria produce enzymes that can cleave the invading viral DNA at specific sequences. The host cells protect their own DNA from cleavage by modifying it at the cleavage sites; the restriction enzymes do not cleave that modified DNA. Since the initial discovery of these restriction endonucleases, hundreds more have been isolated that recognize and cleave different **restriction sites.**

The ability to cut DNA at specific places is significant in two ways: First, it allows physical maps to be constructed based on the positioning of cleavage sites for restriction enzymes. These restriction maps provide crucial data for identifying and working with DNA molecules.

Second, restriction endonuclease cleavage allows the creation of recombinant molecules. The ability to construct recombinant molecules is critical to research, because many steps in the process of cloning and manipulating DNA require the ability to combine molecules from different sources.

### *How restriction enzymes work*

There are three types of restriction enzymes, but only type II cleaves at precise locations. Types I and III cleave with less precision and are not often used in cloning and manipulating DNA.

Type II enzymes allow creation of recombinant molecules; these enzymes recognize a specific DNA sequence, ranging from 4 bases to 12 bases, and cleave the DNA at a specific base within this sequence (figure 17.1).

The recognition sites for most type II enzymes are palindromes. A linguistic *palindrome* is a word or phrase that reads the same forward and in reverse, such as the sentence: "Madam I'm Adam." The palindromic DNA sequence reads the same from 5′ to 3′ on one strand as it does on the complementary strand (see figure 17.1).

Given this kind of sequence, cutting the DNA at the same base on either strand can lead to staggered cuts that produce "sticky ends." These short, unpaired sequences are the same for any DNA that is cut by this enzyme. Thus, these sticky ends allow DNAs from different sources to be easily joined together (see figure 17.1). While less common, some type II restriction enzymes, including *PvuII*, can cut both strands in the same position, producing blunt, not sticky, ends. Blunt cut ends can be joined with other blunt cut ends.

## DNA ligase allows construction of recombinant molecules

Because the two ends of a DNA molecule cut by a type II restriction enzyme have complementary sequences, the pair can form a duplex. An enzyme is needed, however, to join the two fragments together to create a stable DNA molecule. The enzyme DNA ligase accomplishes this by catalyzing the formation of a phosphodiester bond between adjacent phosphate and hydroxyl groups of DNA nucleotides. The action of ligase is to seal nicks in one or both strands (see figure 17.1). This is the same enzyme that joins Okazaki fragments on the lagging strand during DNA replication (see chapter 14).

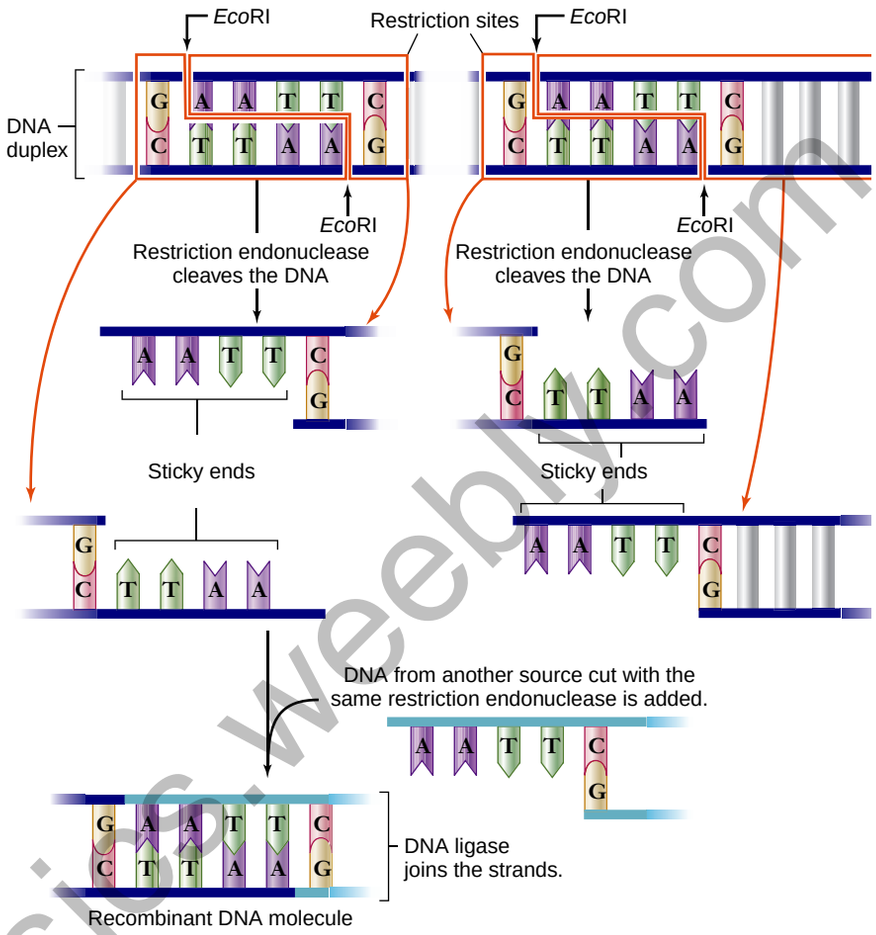



**Figure 17.1 Many restriction endonucleases produce DNA fragments with "sticky ends."** The restriction endonuclease *Eco*RI always cleaves the sequence 5′GAATTC3′ between G and A. Because the same sequence occurs on both strands, both are cut. However, the two sequences run in opposite directions on the two strands. As a result, single-stranded tails called "sticky ends" are produced that are complementary to each other. These complementary ends can then be joined to a fragment from another DNA that is cut with the same enzyme. These two molecules can then be joined by DNA ligase to produce a recombinant molecule.

## Gel electrophoresis separates DNA fragments

The fragments produced by restriction enzymes would not be of much use if we could not also easily separate them for analysis. The most common separation technique used is gel electrophoresis. This technique takes advantage of the negative charge on DNA molecules by using an electrical field to provide the force necessary to separate DNA molecules based on size.

The gel, which is made of either agarose or polyacrylamide and spread thinly on supporting material, provides a three-dimensional matrix that separates molecules based on size (figure 17.2). The gel is submerged in a buffer solution containing ions that can carry current and is subjected to an electrical field.

The strong negative charges from the phosphate groups in the DNA backbone cause it to migrate toward the positive pole (figure 17.2*b*). The gel acts as a sieve to separate DNA molecules based on size: The larger the molecule, the slower it will move through the gel matrix. Over a given period, smaller molecules migrate farther than larger ones. The DNA in gels can be visualized using a fluorescent dye that binds to DNA (figure 17.2*c*, *d*).

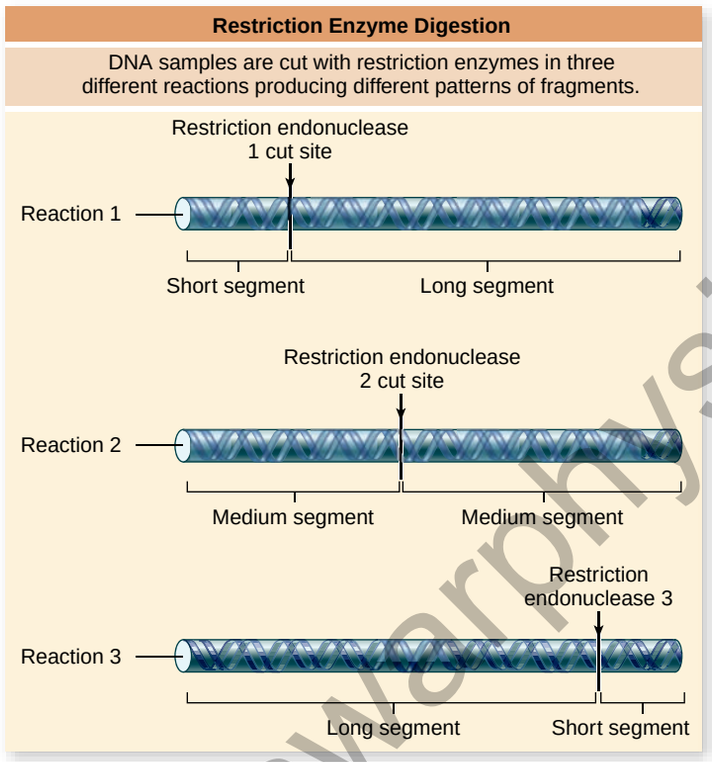
Electrophoresis is one of the most important methods in the toolbox of modern molecular biology, with uses ranging from DNA fingerprinting to DNA sequencing, both of which are described later on.

## Transformation allows introduction of foreign DNA into *E. coli*

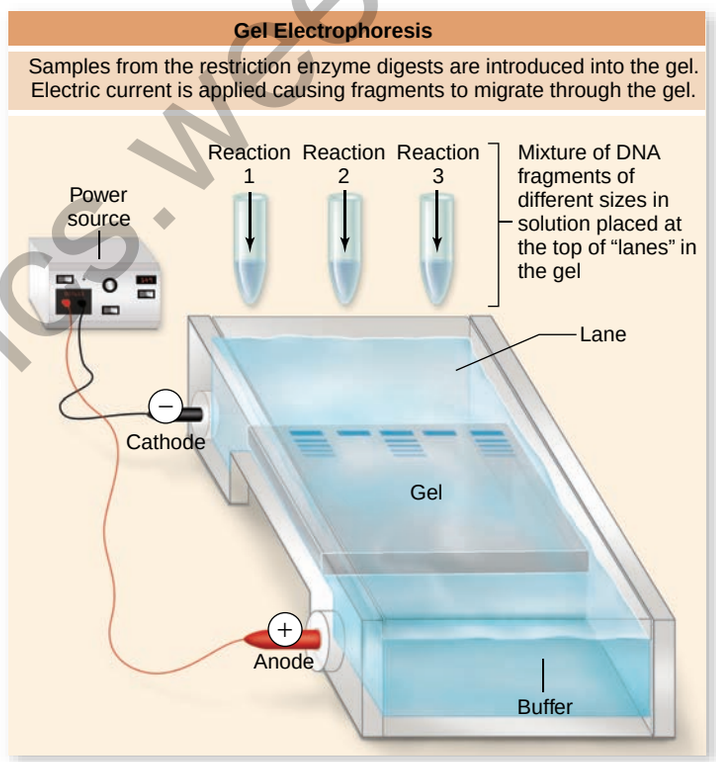
The construction of recombinant molecules is the first step toward genetic engineering. It is also necessary to be able to reintroduce these molecules into cells. In chapter 14 you learned that Frederick Griffith demonstrated that genetic material could be transferred between bacterial cells. This process, called *transformation*, is a natural process in the cells that Griffith was studying.

The bacterium *E. coli*, used routinely in molecular biology laboratories, does not undergo natural transformation; but artificial transformation techniques have been developed to allow introduction of foreign DNA into *E. coli*. Through temperature shifts or an electical charge, the *E. coli* membrane becomes transiently permeable to the foreign DNA. In this way, recombinant molecules can be propagated in a cell that will make many copies of the constructed molecules.

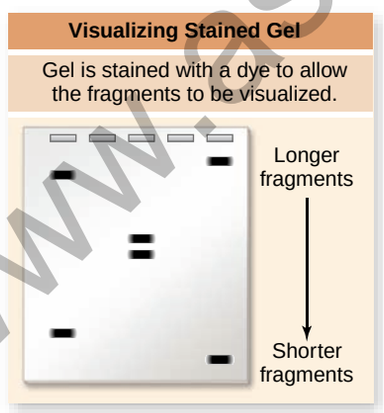
In general, the introduction of DNA from an outside source into a cell is referred to as transformation. This process is important in *E. coli* for molecular cloning and the propagation of cloned DNA. Researchers also want to be able



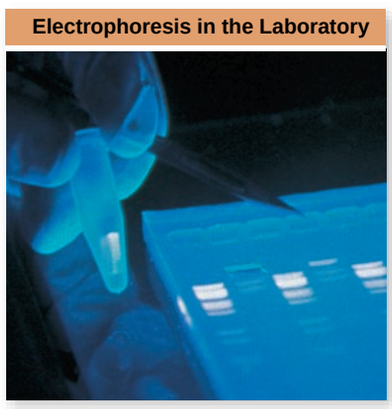

*a.*




*b.*




*c.*




*d.*

**Figure 17.2 Gel electrophoresis.** ***a.*** Three restriction enzymes are used to cut DNA into specific pieces depending on each enzyme's recognition sequence. ***b.*** The fragments are loaded into a gel (agarose or polyacrylamide), and an electrical current is applied. The DNA fragments migrate through the gel based on size, with larger ones moving more slowly. ***c.*** This results in a pattern of fragments separated based on size, with the smaller fragments migrating farther than larger ones. ***d.*** The fragments can be visualized by staining with the dye ethidium bromide. When the gel is exposed to UV light, the DNA with bound dye fluoresces, appearing as pink bands in the gel. In the photograph, one band of DNA has been excised from the gel for further analysis and can be seen glowing in the tube the technician holds.

to reintroduce DNA into the original cells from which it was isolated. A transformed cell that can also be used to form all or part of an organism, is called a **transgenic** organism. Later in this chapter we explore the construction and uses of transgenic plants and animals.

### Learning Outcomes Review 17.1

**Restriction endonucleases are part of bacterial cells' strategies to fight viral infection. Type II endonucleases cleave DNA at specific sites. DNA ligase can be used to link together fragments following action of restriction endonucleases. Gel electrophoresis employs electrical charge to separate DNA fragments according to size. Foreign DNA can be introduced into *E. coli* through artificial transformation, and then propagation can produce cloned DNA.**

- ***Compare and contrast the endogenous roles of* EcoRI *and ligase in* E. coli *with their use in a molecular biology lab.***

## 17.2 Molecular Cloning

### Learning Outcomes

1. ***Explain the role of a vector in molecular cloning.***
2. ***Describe how a DNA library is constructed.***

The term **clone** refers to a genetically identical copy. The technique of propagating plants by growing a new plant from a cutting of a donor plant is an early method of cloning widely used in agriculture and horticulture. The topic of cloning entire organisms is discussed in chapter 19. For now, we explore the idea of molecular cloning.

*Molecular cloning* involves the isolation of a specific sequence of DNA, usually one that encodes a particular protein product. This is sometimes called *gene cloning*, but the term *molecular cloning* is more accurate.

### Host–vector systems allow propagation of foreign DNA in bacteria

Although short sequences of DNA can be synthesized in vitro (in a test tube), the cloning of large unknown sequences requires propagation of recombinant DNA molecules in vivo (in a cell). The enzymes and methods described earlier allow biologists to produce, separate, and then introduce foreign DNA into cells.

The ability to propagate DNA in a host cell requires a **vector** (something to carry the recombinant DNA molecule) that can replicate in the host when it has been introduced. Such host–vector systems are crucial to molecular biology.

The most flexible and common host used for molecular cloning is the bacterium *E. coli*, but many other hosts are now possible. Investigators routinely reintroduce cloned eukaryotic DNA, using mammalian tissue culture cells, yeast cells, and insect cells as host systems. Each kind of host–vector system allows particular uses of the cloned DNA.

The two most commonly used vectors are plasmids and artificial chromosomes. *Plasmids* are small, circular extrachromosomal DNAs that are dispensable to the bacterial cell. Bacterial and eukaryotic artificial chromosomes are used to clone larger pieces of DNA.

#### Plasmid vectors

Plasmid vectors (small, circular chromosomes) are typically used to clone relatively small pieces of DNA, up to a maximum of about 10 kilobases (kb). A plasmid vector must have three components:

1. An *origin of replication* to allow it to be replicated in *E. coli* independently of the host chromosome,
2. A *selectable marker*, usually antibiotic resistance, and
3. *One or more unique restriction sites* where foreign DNA can be added.

The selectable marker allows the presence of the plasmid to be easily identified through genetic selection. For example, cells that contain a plasmid with an antibiotic resistance gene continue to live when plated on antibiotic-containing growth media, whereas cells that lack the plasmid will die (they are killed by the antibiotic).

A fragment of DNA is inserted by the techniques described into a region of the plasmid with restriction sites called the multiple-cloning site (MCS). This region contains a number of unique restriction sites such that when the plasmid is cut with the relevant restriction enzymes, a linear plasmid results. When DNA of interest is cut with the same restriction enzyme, it can then be ligated into this site. The plasmid is then introduced into cells by transformation (see figure 17.3).

This region of the vector often has been engineered to contain another gene that becomes inactivated, so-called *insertional inactivation*, because it is now interrupted by the inserted DNA. One of the first cloning vectors, pBR322, used another antibiotic resistance gene for insertional activation; resistance to one antibiotic and sensitivity to the other indicated the presence of inserted DNA.

More recent vectors use the gene for β-galactosidase, an enzyme that cleaves galactoside sugars such as lactose. When the enzyme cleaves the artificial substrate X-gal, a blue color is produced. In these plasmids, insertion of foreign DNA interrupts the β-galactosidase gene, preventing a functional enzyme from being produced. When transformed cells are plated on medium containing both antibiotic (to select for plasmid-containing cells) and X-gal, they remain white, whereas transformed cells with no inserted DNA are blue (see figure 17.3).

#### Artificial Chromosomes

The size of DNA molecules that can be cloned in plasmid vectors has limited the large-scale analysis of genomes. To deal

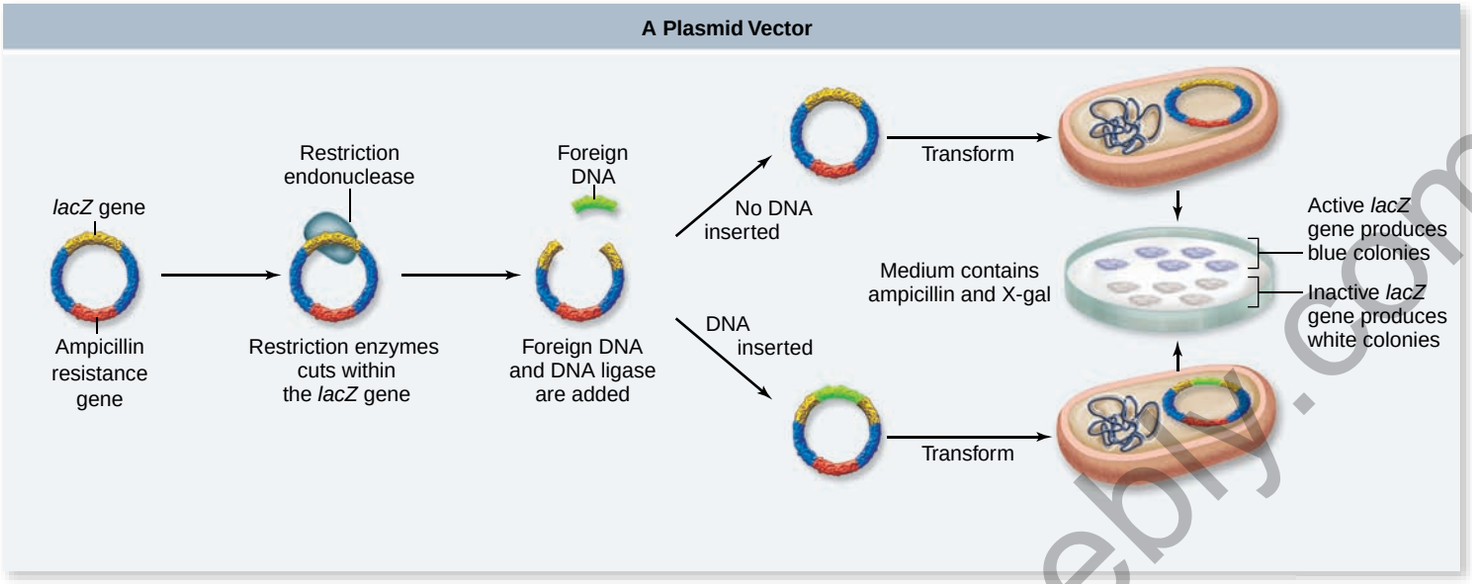



**Figure 17.3 Molecular cloning with vectors.** Plasmids are cut within the β-galactosidase gene *(lacZ)*, and foreign DNA and DNA ligase are added. Foreign DNA inserted into *lacZ* interrupts the coding sequence, thus inactivating the gene. Plating cells on medium containing the antibiotic ampicillin selects for plasmid-containing cells. The medium also contains X-gal, and when *lacZ* is intact *(top)*, the expressed enzyme cleaves the X-gal, producing blue colonies. When *lacZ* is inactivated *(bottom)*, X-gal is not cleaved, and colonies remain white.

with this, geneticists decided to follow the strategy of cells and construct chromosomes, leading to the development of yeast artificial chromosomes (YACs) and bacterial artificial chromosomes (BACs). Progress has also been made on creating mammalian artificial chromosomes. Use of artificial chromosomes is described in the next chapter.

**Inquiry question**

**An investigator wishes to clone a 32-kb recombinant molecule. What do you think is the best vector to use?**

## DNA libraries contain the entire genome of an organism

The idea of molecular cloning depends on the ability to construct a representation of very complex mixtures in DNA, such as an entire genome, in a form that is easier to work with than the enormous chromosomes within a cell. If the huge DNA molecules in chromosomes can be converted into random fragments, and inserted into a vector such as plasmids, then when they are propagated in a host they will together represent the whole genome. This aggregate is termed a **DNA library,** a collection of DNAs in a vector that taken together represent the complex mixture of DNA (figure 17.4).

Conceptually the simplest possible kind of DNA library is a **genomic library**—a representation of the entire genome in a vector. This genome is randomly fragmented by partially digesting it with a restriction enzyme that cuts frequently. By not cutting the DNA to completion, not all sites are cleaved, and which sites are cleaved is random. The random fragments are then inserted into a vector and introduced into host cells. Genomic libraries are usually constructed in bacterial artificial chromosomes (BACs).

A variety of different kinds of libraries can be made depending on the source DNA used. Any particular clone in the library contains only a single DNA, and all of them together make up the library. Keep in mind that unlike a library full of

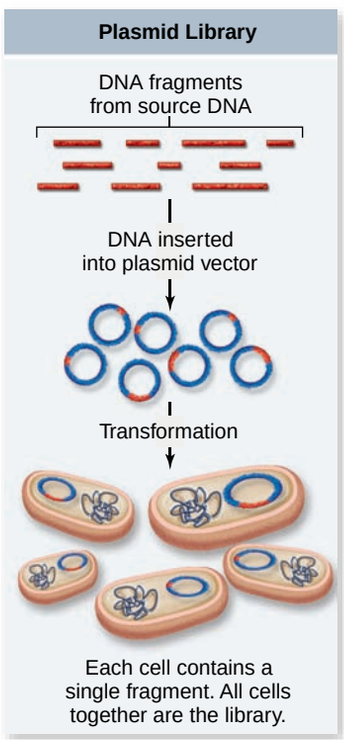



**Figure 17.4 Creating DNA libraries.**

books, which is organized and catalogued, a DNA library is a random collection of overlapping DNA fragments. We explore how to find a sequence of interest in this random collection later in the chapter.

## Reverse transcriptase can make a DNA copy of RNA

In addition to genomic libraries, investigators often wish to isolate only the *expressed* part of genes. The structure of eukaryotic genes is such that the mRNA may be much smaller than the gene itself due to the presence of introns in the gene. After transcription by RNA polymerase II, the primary transcript is spliced to produce the mRNA (chapter 15). Because of this, genomic libraries are crucial to understanding the structure of the gene, but are not of much use if we want to express the gene in a bacterial species, whose genes do not contain introns and which has no mechanism for splicing.

A library of only expressed sequences represents a much smaller amount of DNA than the entire genome. The starting point for a cDNA library is isolated mRNA representing the genes expressed in a specific tissue at a specific developmental stage. Such a library of expressed sequences is made possible by the use of another enzyme: reverse transcriptase.

Reverse transcriptase was isolated from a class of viruses called retroviruses. The life cycle of a retrovirus requires making a DNA copy from its RNA genome. We can take advantage of the activity of the retrovirus enzyme to make DNA copies from isolated mRNA. DNA copies of mRNA are called **complementary DNA (cDNA)** (figure 17.5). A cDNA library is made by first isolating mRNA from genes being expressed and then using the reverse transcriptase enzyme to make cDNA from the mRNA. The cDNA is then used to make a library, as mentioned earlier. These cDNA libraries are extremely useful and are commonly made to represent the genes expressed in many different tissues or cells. While all genomic libraries made from an individual will be identical, cDNA libraries from the same cells at different developmental stages or different tissues will each be distinct.

### Inquiry question

**Suppose you wanted a copy of a section of a eukaryotic genome that included the introns and exons. Would the creation of cDNA be a good way to go about this?**

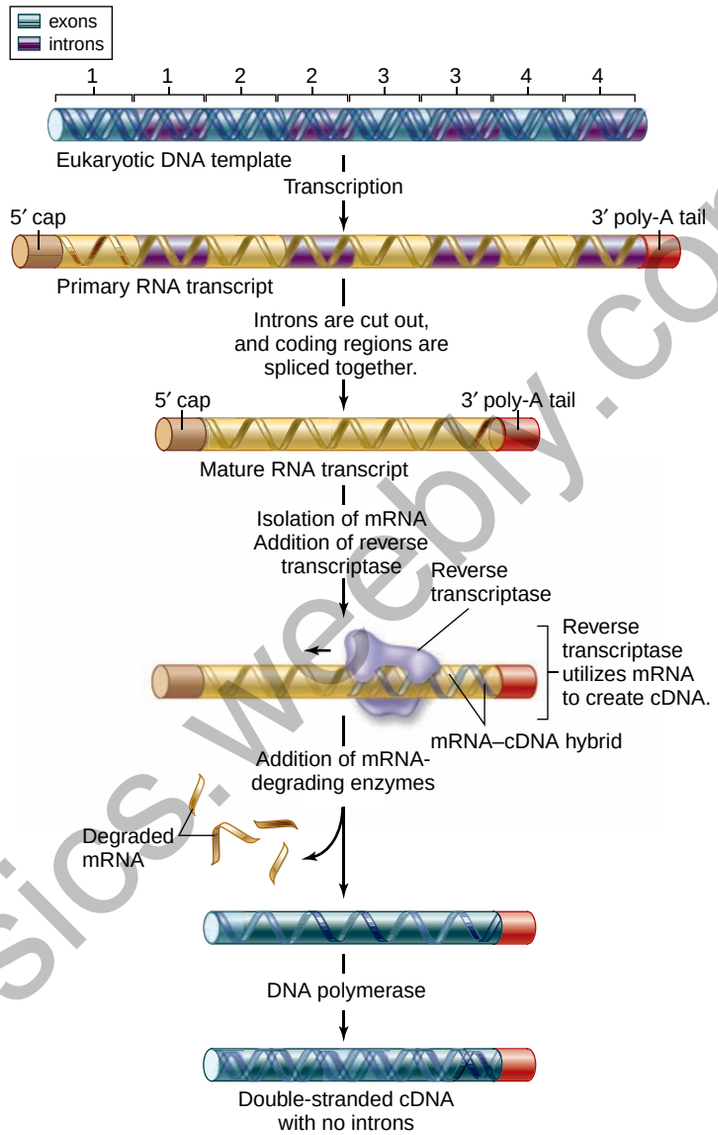



**Figure 17.5 The formation of cDNA.** A mature mRNA transcript is usually much smaller than the gene due to the loss of intron sequences by splicing. mRNA is isolated from the cytoplasm of a cell, which the enzyme reverse transcriptase uses as a template to make a DNA strand complementary to the mRNA. That newly made strand of DNA is the template for the enzyme DNA polymerase, which assembles a complementary DNA strand along it, producing cDNA—a double-stranded DNA version of the intron-free mRNA.

## Hybridization allows identification of specific DNAs in complex mixtures

The technique of **molecular hybridization** is commonly used to identify specific DNAs in complex mixtures such as libraries. Hybridization, also called annealing, takes advantage of the specificity of base-pairing between the two strands of DNA. If a DNA molecule is denatured, that is, the two strands are separated, the strands can only reassociate with partners that have the correct complementary sequence. Molecular biologists can take advantage of this feature experimentally to use a known, specific DNA molecule to find its partner in a complex mixture.

Any single-stranded nucleic acid (DNA or RNA) can be tagged with a radioactive label or with another detectable label, such as a fluorescent dye. This can then be used as a probe to identify its complement in a complex mixture of DNA or RNA. This renaturing is termed *hybridization* because the combination of labeled probe and unlabeled DNA form a hybrid molecule through base-pairing.

Probes have been made historically by a variety of techniques. One technique involved isolating a protein of interest

and then chemically sequencing the protein. With the protein sequence in hand, the DNA sequence could be predicted using the genetic code. This information can then be used to make a synthetic DNA for use as a probe.

## Specific clones can be isolated from a library

The isolation of a specific clone from the random collection that is a DNA library is akin to finding the proverbial needle in a haystack. It requires some information about the gene of interest. For example, many of the first genes isolated were those that are highly expressed in a specific cell type, such as the globin genes that encode the proteins found in the oxygen carrier hemoglobin.

Hybridization is the most common way of identifying a clone within a DNA library. This procedure is outlined for a DNA library in a plasmid vector in figure 17.6.

In the early days of molecular biology, individual investigators made their own DNA libraries, as is shown earlier in figure 17.4. Now, genomic and cDNA libraries are commercially available for a large number of organisms. Screening such a library involves growing the library on agar plates, making a replica of the library, and screening for the cloned sequence of interest.

### *Stage 1: Plating the library*

Physically, the library is either a collection of bacterial viruses that each contain an inserted DNA, or bacterial cells that each harbor a plasmid or artificial chromosome with inserted DNA. To find a specific clone, the library needs to be represented in an organized fashion. Figure 17.6 shows this representation for a plasmid vector. The library of bacteria containing plasmids is grown on agar plates at a high density, but not so high that individual colonies cannot be distinguished.

### *Stage 2: Replicating the library*

Once the library has been grown on plates, a replica can be made by laying a piece of filter paper on the plate; some of the viruses or cells in each colony will stick to the filter, and some will be left on the plate. The result is a copy of the library on a piece of filter paper. The DNA can be affixed to the filter paper by baking or by cross-linking it to the filter using UV light.

### *Stage 3: Screening the library*

Once a replica of the library has been formed on a filter, a specific clone can be identified by hybridization. The probe, which represents the specific sequence of interest, is labeled with a radioactive nucleotide. The probe is then added to the filters with the library replicated on them. Film sensitive to radioactive emissions is then placed in contact with the filters; where radioactivity is present, a dark spot appears on the film. When the film is aligned with the original plate, the clone of interest can be identified (see figure 17.6).

### *Learning Outcomes Review 17.2*

**Molecular cloning is the isolation and amplification of a specific DNA sequence. A vector is a carrier into which a sequence of interest may be introduced. The most common vectors are plasmids and phages. The vector takes the sequence into a cell, which then multiplies, copying its own DNA along with that of the vector. DNA libraries are representations of complex mixtures of DNA, such as an entire genome, stored in a host–vector system. DNA libraries are often screened for specific clones using molecular hybridization.**

- ***How does a gene's sequence in a cDNA library compare with the sequence of the gene itself?***

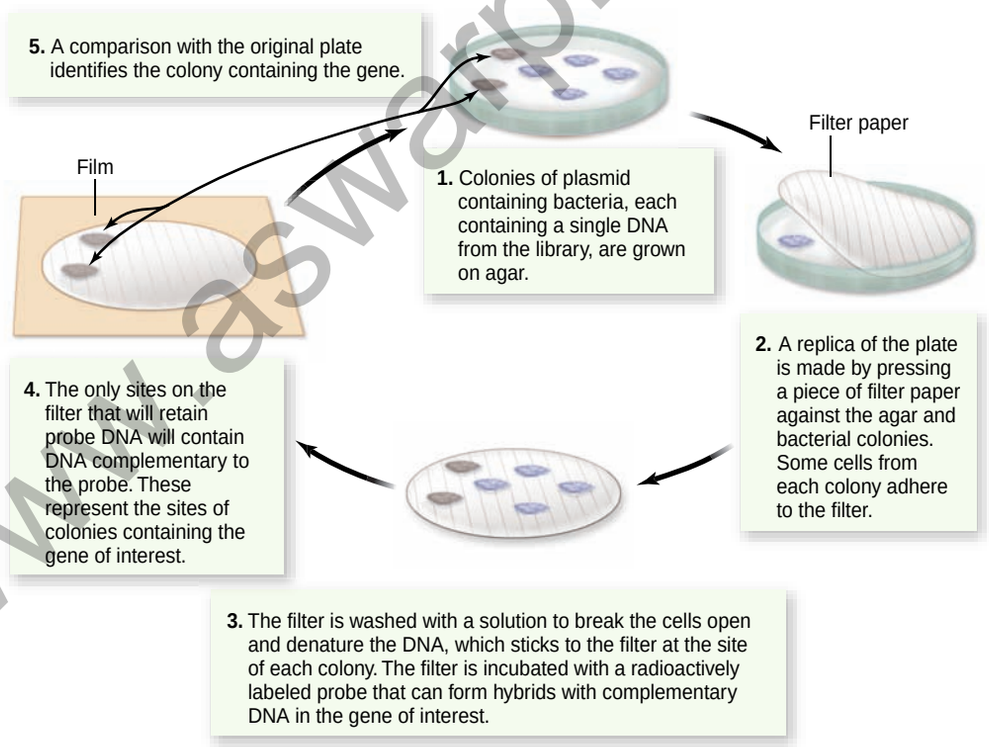
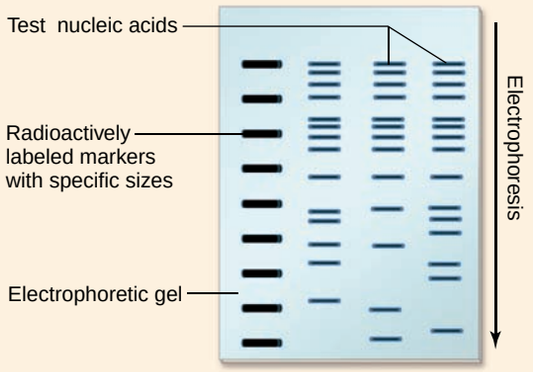
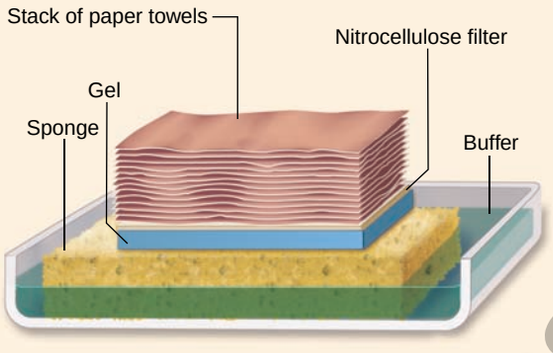



**Figure 17.6 Screening a library using hybridization.** This technique takes advantage of DNA's ability to be denatured and renatured, with complementary strands finding each other. Cells containing the library are plated on agar gel. A replica of the plates is made using special filter paper, nitrocellulose or nylon, which binds to single-stranded DNA. The filter paper with replica colonies is treated to lyse the cells and denature the DNA, producing a pattern of DNA bound to the filter that corresponds to the pattern of colonies. When a radioactive probe is added, it finds complementary DNA and forms hybrids at the site of colonies that contained the gene of interest.

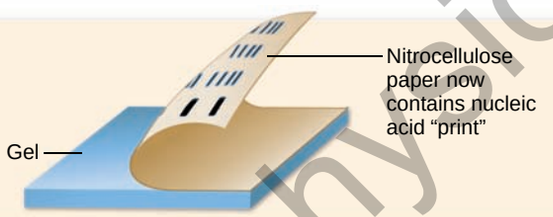
**1.** Electrophoresis is performed, using radioactively labeled markers as a size guide in the first lane.



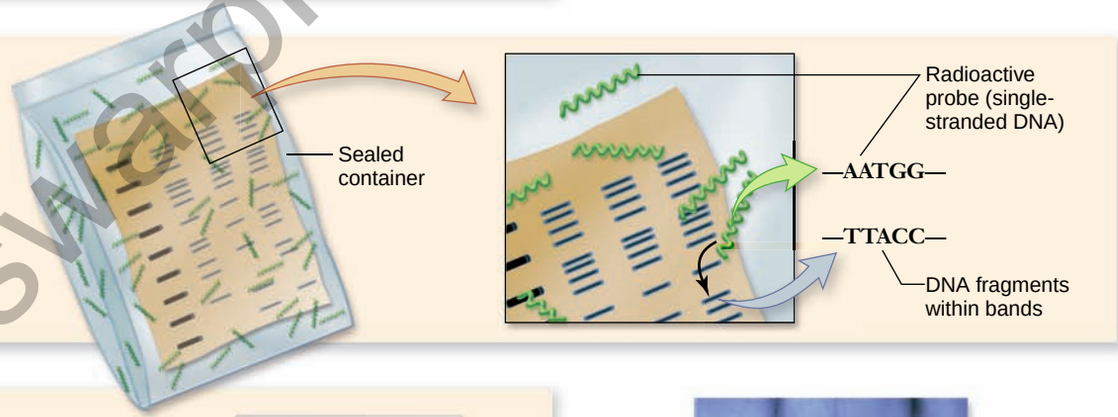

**2.** The gel is covered with a sheet of nitrocellulose and placed in a tray of buffer on top of a sponge. Alkaline chemicals in the buffer denature the DNA into single strands. The buffer wicks its way up through the gel and nitrocellulose into a stack of paper towels placed on top of the nitrocellulose.



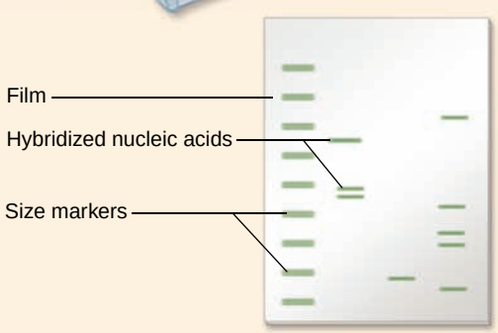

**3.** DNA in the gel is transferred, or "blotted," onto the nitrocellulose.




**4.** Nitrocellulose with bound DNA is incubated with radioactively labeled nucleic acids and is then rinsed.




**5.** Photographic film is laid over the filter and is exposed only in areas that contain radioactivity (autoradiography). Bands on the film represent DNA in the gel that is complementary to the probe sequence.



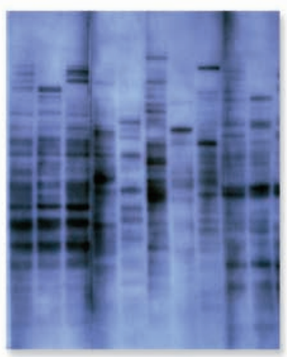



**Figure 17.7 The Southern blot procedure.** Edwin M. Southern developed this procedure in 1975 to enable DNA fragments of interest to be visualized in a complex sample containing many other fragments of similar size. In steps 1–3, the DNA is separated on a gel, and then transferred ("blotted") onto a solid support medium such as nitrocellulose paper or a nylon membrane. Sequences of interest can be detected by using a radioactively labeled probe. This probe (usually several hundred nucleotides in length) of single-stranded DNA (or an mRNA complementary to the gene of interest) is incubated with the filter containing the DNA fragments. All DNA fragments that contain nucleotide sequences complementary to the probe will form hybrids with the probe. Only a short segment of the probe and the complementary sequence are shown in panel 4. The fragments differ in size, with the smallest moving the farthest in the gel. The fragments of interest are then detected using photographic film. A representative image is shown in panel 5. The use of film for detection is being replaced by phosphor imagers, computer-controlled devices that have electronic sensors for light or radioactive emissions.

# 17.3 DNA Analysis

### Learning Outcomes

1. *Explain the Southern blotting method of identifying genes.*
2. *Compare endogenous DNA replication with sequencing and with the polymerase chain reaction.*
3. *Explain how the yeast system is used to study protein–protein interactions.*

Molecular cloning provides specific DNA for further manipulation and analysis. The number of ways that DNA can be manipulated could fill the rest of this book, but for our purposes, we will highlight a few important methods of analysis and uses of molecular clones.

## Restriction maps provide molecular "landmarks"

If you are new to a city, the easiest way to find your way around is to obtain a map and compare that map with your surroundings. In a similar fashion, molecular biologists need maps to analyze and compare cloned DNAs.

The first kind of physical maps were restriction maps that included the location and order of sites cut by the battery of restriction enzymes available. Initially, these maps were created by cutting the DNA with different enzymes, separating the fragments by gel electrophoresis, and analyzing the resulting patterns. Although this method is still in use, many restriction maps are now generated by computer searching of known DNA sequences for the sites cut by restriction enzymes.

## Southern blotting reveals DNA differences

Once a gene has been cloned, it may be used as a probe to identify the same or a similar gene in DNA isolated from a cell or tissue (figure 17.7). In this procedure, called a **Southern blot,** DNA from the sample is cleaved into fragments with a restriction endonuclease, and the fragments are separated by gel electrophoresis. The double-stranded helix of each DNA fragment is then denatured into single strands by making the pH of the gel basic. Then the gel is "blotted" with a sheet of filter paper, transferring some of the DNA strands to the sheet.

Next, the filter is incubated with a labeled probe consisting of purified, single-stranded DNA corresponding to a specific gene (or mRNA transcribed from that gene). Any fragment that has a nucleotide sequence complementary to the probe's sequence hybridizes with the probe (see figure 17.7).

This kind of blotting technique has also been adapted for use with RNA and proteins. When mRNA is separated by electrophoresis, the technique is called a **Northern blot.** The methodology is the same except for the starting material (mRNA instead of DNA) and that no denaturation step is required. Proteins can also be separated by electrophoresis and blotted by a procedure called a **Western blot.** In this case both the electrophoresis and the detection step are different from Southern blotting. The detection, in this case, requires an antibody that can bind to one protein.

The names of these techniques all go back to the original investigator, the British biologist Edwin M. Southern; the Northern and Western blotting names were word play on Southern's name using the cardinal points of the compass.

### *RFLP analysis*

In some cases, an investigator wants to do more than find a specific gene, but instead is looking for variation in the genes of different individuals. One powerful way to do this is by analyzing **restriction fragment length polymorphisms,** or **RFLPs,** using Southern blotting (figure 17.8).

Point mutations that change the sequence of DNA can eliminate sequences recognized by restriction enzymes or create new recognition sequences, changing the pattern of fragments seen in a Southern blot. Sequence repetitions may also occur between the restriction endonuclease sites, and differences in repeat number between individuals can also alter the length of the DNA fragments. These differences can all be detected with Southern blotting.

When a genetic disease has an associated RFLP, the RFLP can be used to diagnose the disease. Huntington disease, cystic fibrosis, and sickle cell anemia all have associated RFLPs that have been used as molecular markers for diagnosis.

### *DNA fingerprinting*

RFLP analysis has been used in **DNA fingerprinting.** When a probe is made for DNA that is repetitive, it often detects a large number of fragments. These fragments are often not identical in different individuals. We say that the population is **polymorphic** for these molecular markers. These markers can be used as

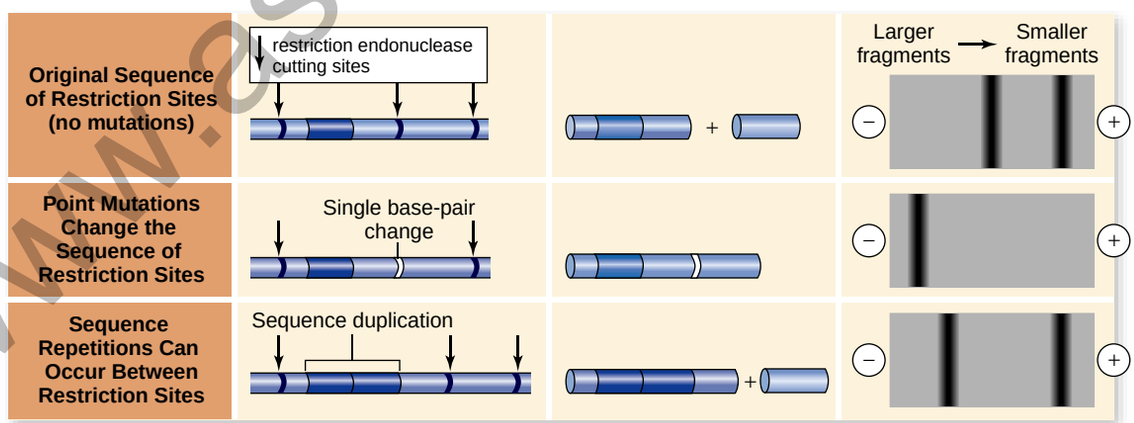



**Figure 17.8 Restriction fragment length polymorphism (RFLP) analysis.** ***a.*** Three samples of DNA differ in their restriction sites due to a single base-pair substitution in one case and a sequence duplication in another case. ***b.*** When the samples are cut with a restriction endonuclease, different numbers and sizes of fragments are produced. ***c.*** Gel electrophoresis separates the fragments, and different banding patterns result.

DNA "fingerprints" in criminal investigations and other identification applications.

Figure 17.9 shows the DNA fingerprints a prosecuting attorney presented in a rape trial in 1987. They consist of autoradiographs, parallel bars on X-ray film. These bars can be thought of as being similar to the product price codes on consumer goods in that they may provide unique identification. Each bar represents the position of a DNA restriction endonuclease fragment produced by techniques similar to those described in figures 17.7 and 17.8. The long dark lane with many bars in figure 17.9 represents a standardized control.

Two different probes were used to identify the restriction fragments. A vaginal swab had been taken from the victim within hours of her attack; from it, semen was collected and its DNA analyzed for restriction endonuclease patterns.

Compare the restriction endonuclease patterns of the semen to that of blood from the suspect. You can see that the suspect's two patterns match that of the rapist (and are not at all like those of the victim). The suspect was Tommie Lee Andrews, and on November 6, 1987, the jury returned a verdict of guilty. Andrews became the first person in the United States to be convicted of a crime based on DNA evidence.

Since the Andrews verdict, DNA fingerprinting evidence is now a determining factor in at least forty percent of the criminal cases in the United States. Although some probes highlight profiles shared by many people, others are quite rare. Using several probes, the probability of identity can be calculated or identity can be ruled out. Laboratory analyses of DNA samples, however, must be carried out properly—sloppy procedures could lead to a wrongful conviction. After widely publicized instances of questionable lab procedures, national standards are being developed.

DNA fingerprinting is also used to identify human remains. After the September 11, 2001 attacks on the World Trade Centers in New York, DNA fingerprinting was the only option for identifying some of the victims of the attack. by 2005, 1585 of the 2792 people who were missing had been identified using DNA fingerprinting. Advances in forensic technology, including improved DNA isolation from very small amounts of tissue, have made it possible to identify additional individuals since 2005.

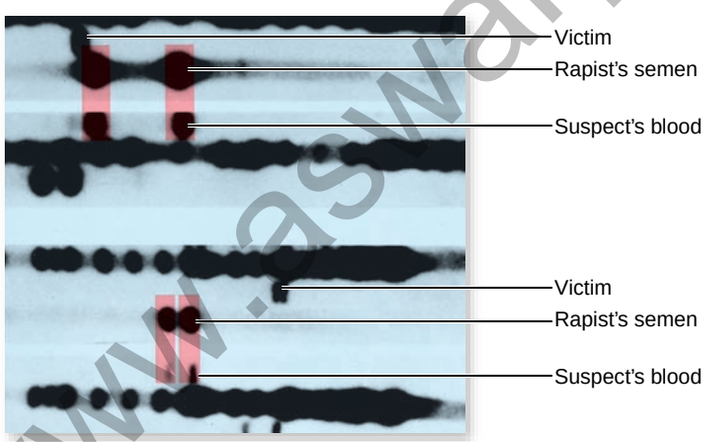



**Figure 17.9 Two of the DNA profiles that led to the conviction of Tommie Lee Andrews for rape in 1987.** The two DNA probes seen here were used to characterize DNA isolated from the victim, the semen left by the rapist, and the suspect. The dark channels are multiband controls. There is a clear match between the suspect's DNA and the DNA of the rapist's semen in these two profiles.

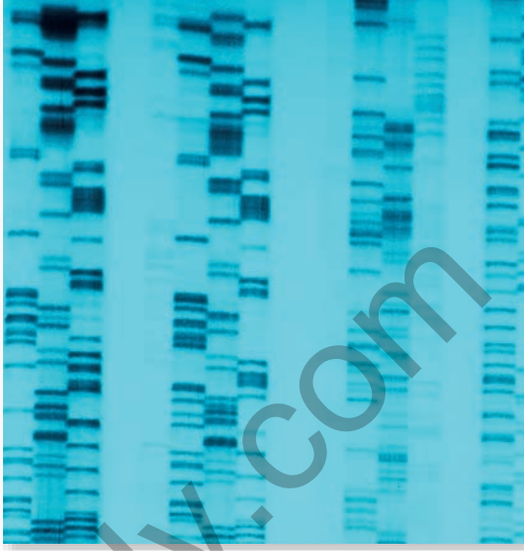


**Figure 17.10 Ladder of fragments used in DNA sequencing.** The photo shows the autoradiograph of the fragments generated by DNA-sequencing reactions. These fragments are generated by either organic reactions that cleave at specific bases or enzymatic reactions that terminate in specific bases. The gel can separate fragments that differ by a single base.

## DNA sequencing provides information about genes and genomes

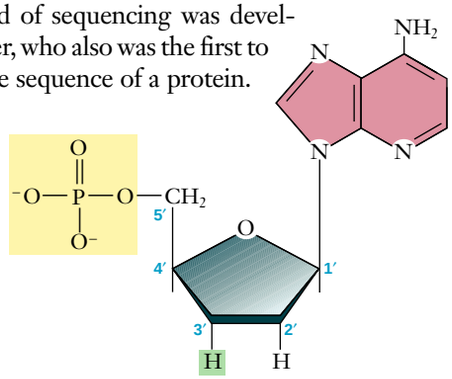
The ultimate level of analysis is determination of the actual sequence of bases in a DNA molecule. The development of sequencing technology has paralleled the advancement of molecular biology. As it became possible to determine the sequence of an entire genome relatively rapidly, the field of genomics emerged.

The basic idea used in DNA sequencing is to generate a set of nested fragments that each begin with the same sequence and end in a specific base. When this set of fragments is separated by high-resolution gel electrophoresis, the result is a "ladder" of fragments (figure 17.10) in which each band consists of fragments that end in a specific base. By starting with the shortest fragment, one can then read the sequence by moving up the ladder.

The problem then became how to generate the sets of fragments that end in specific bases. In the early days of sequencing, both a chemical method and an enzymatic method were utilized. The chemical method involved organic reactions specific for the different bases that made breaks in the DNA chains at specific bases. The enzymatic method used DNA polymerase to synthesize chains, but it also included in the reaction modified nucleotides that could be incorporated but not extended: so-called *chain terminators*. The enzymatic method has proved more versatile, and it is easier to adapt to different uses.

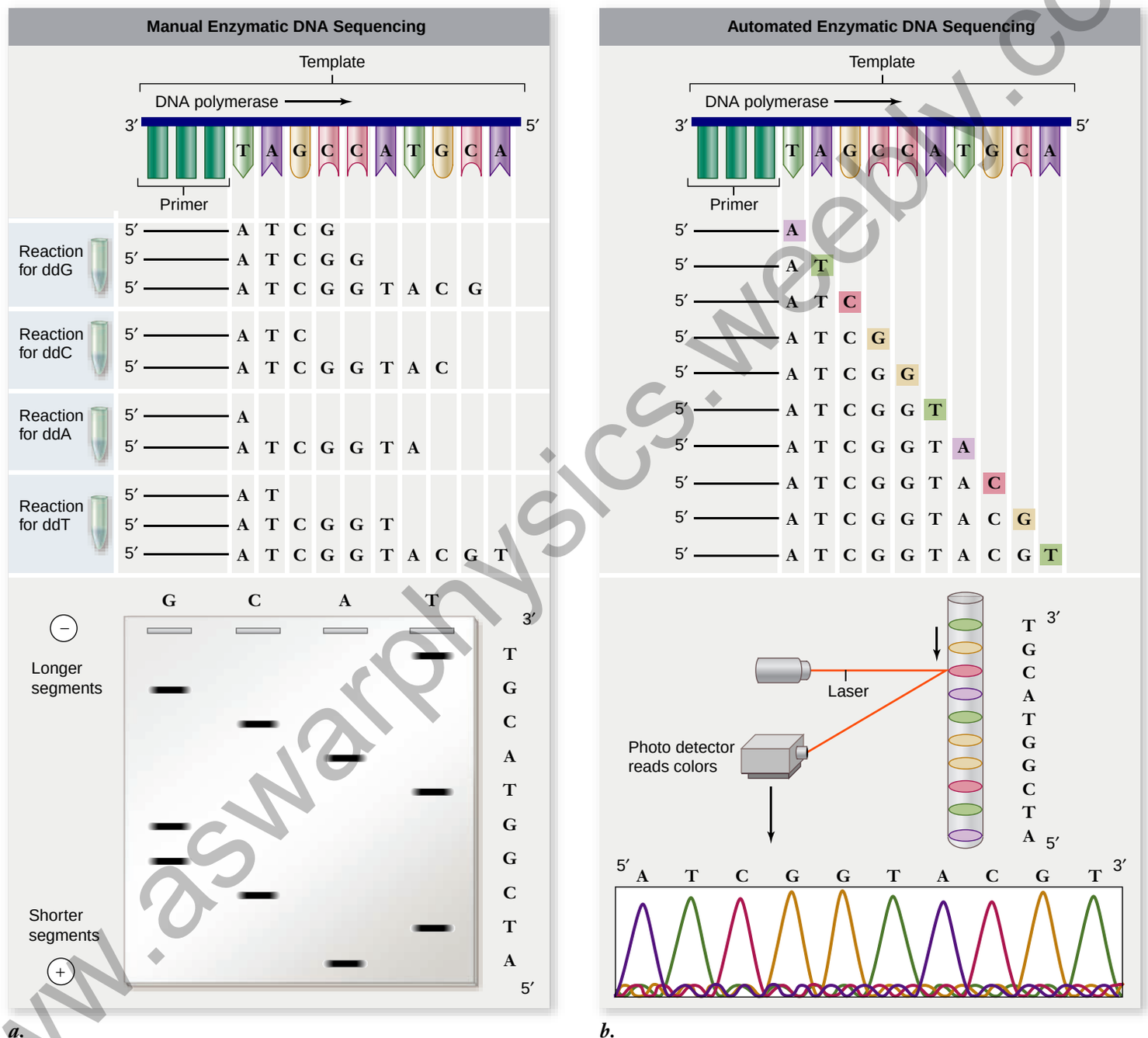
### *Enzymatic sequencing*

The enzymatic method of sequencing was developed by Fredrick Sanger, who also was the first to determine the complete sequence of a protein. This method uses dideoxynucleotides as chain terminators in DNA synthesis reactions. A **dideoxynucleotide** has H in place of OH at both the 2′ position and at the 3′ position.



All DNA nucleotides lack —OH at the 2′ carbon of the sugar, but dideoxynucleotides have no 3′ —OH at which the enzyme can add new nucleotides. Thus the chain is terminated.

The experimenter must perform four separate reactions, each with a single dideoxynucleotide, to generate a set of fragments that terminate in specific bases. Thus all of the fragments produced in the A reaction incorporate dideoxyadenosine and must end in A, and the same for the other three reactions with different terminators. When these fragments are separated by high-resolution gel electrophoresis, each reaction is run in a different track, or lane, to generate a pattern of nested fragments that can be read from the smallest fragment to fragments that are each longer by one base (figure 17.11*a*).




*a.* *b.*

**Figure 17.11 Manual and automated enzymatic DNA sequencing.** The sequence to be determined is shown at the top as a template strand for DNA polymerase with a primer attached. ***a.*** In the manual method, four reactions were done, one for each nucleotide. For example, the A tube would contain dATP, dGTP, dCTP, dTTP, and ddATP. This leads to fragments that end in A due to the dideoxy terminator. The fragments generated in each reaction are shown along with the results of gel electrophoresis. ***b.*** In automated sequencing, each ddNTP is labeled with a different color fluorescent dye, which allows the reaction to be done in a single tube. The fragments generated by the reactions are shown. When these are electrophoresed in a capillary tube, a laser at the bottom of the tube excites the dyes, and each will emit a different color that is detected by a photodetector.

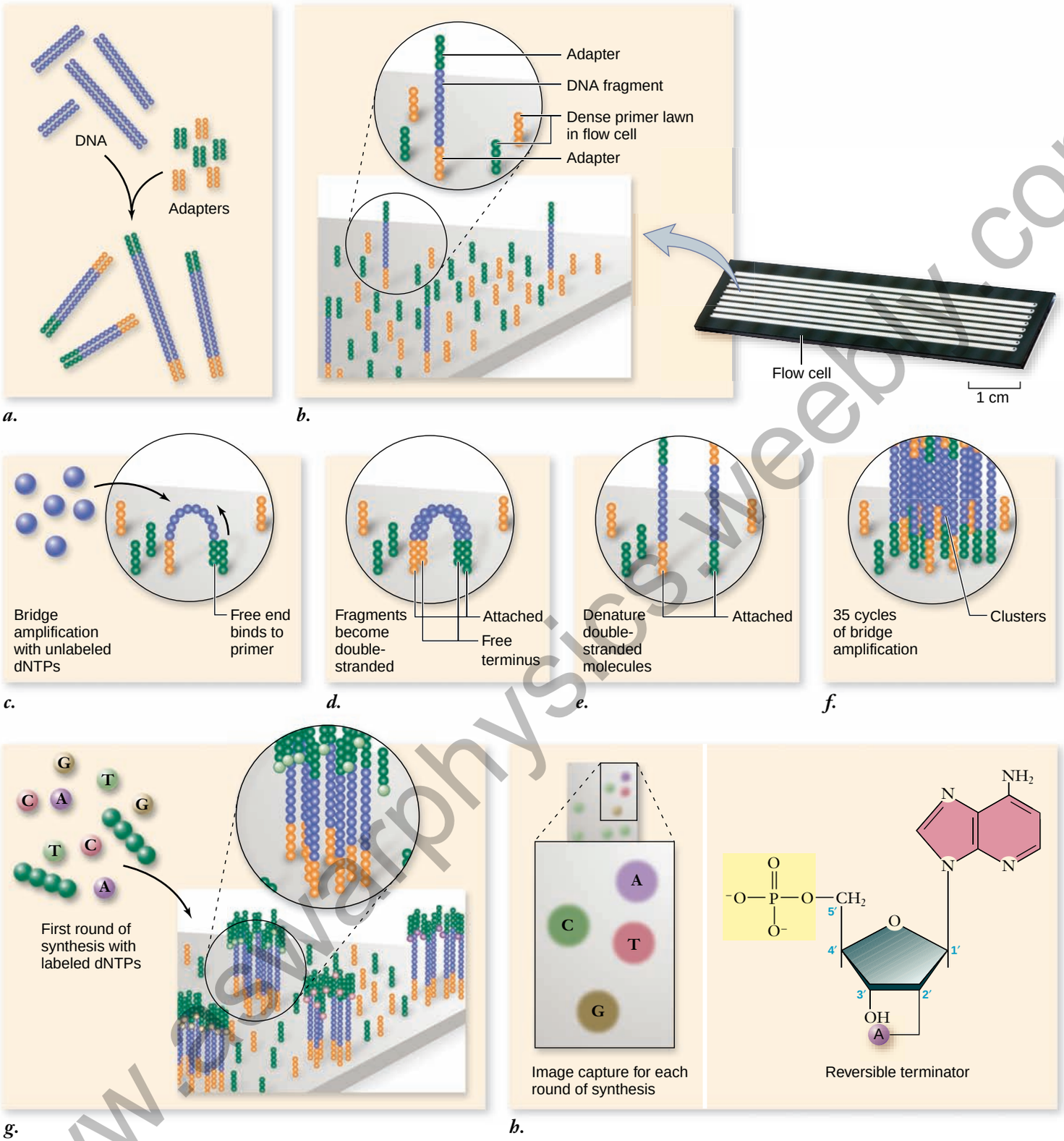



**Figure 17.12 New approach to sequencing.** DNA is cleaved into short fragments that will be sequenced. ***a.*** Adapters are added to the end of the DNA. ***b.*** DNA is denatured and the adapters bind to complementary primers in the flow cell. ***c–f.*** Individual fragments are amplified using dNTPs and polymerase. ***g.*** Fluorescently labeled dNTPs with cleavable dye that blocks the formation of additional phosphodiester bonds are added, and the first fluorescently labeled base is added. ***h.*** A CCD camera records the fluorescence pattern before the fluorescent dye is removed, and the next base is added to each DNA sequence.

Notice that since this is a DNA polymerase reaction, it requires a primer to begin synthesis. The vectors used for DNA sequencing have known regions next to the site where DNA is inserted. Short DNAs that are complementary to these regions are then synthesized and can be used as primers. This serves the dual purposes of providing a primer and ensuring that the first few bases sequenced are known because they are known in the vector itself. This allows the investigator to determine where the sequence of interest begins. As the sequence is generated, new primers can be designed near the end of the known sequence and DNA synthesized to use as a primer to extend the region sequenced in the next set of reactions.

### *Automated sequencing*

The technique of enzymatic sequencing is very powerful, but it is also labor-intensive and takes a significant amount of time. It requires a series of enzymatic manipulations, time for electrophoresis, then time to expose the gel to film. At the end of this, a skilled researcher can read around 300 bases of sequence reliably. The development of automated techniques made sequencing a much more practical and less human-intensive procedure.

Automated sequencing machines use fluorescent dyes instead of a radioactive label and separate the products of the sequencing reactions using gels in thin capillary tubes instead of the large slab gels. The tubes run in front of a laser that excites the dyes, causing them to fluoresce. With a different colored dye for each base, a photodetector can determine the identity of each base by its color.

The data are assembled by a computer that generates a visual image consisting of different colored peaks; these are converted into the raw sequence data (figure 17.11*b*). The sequence data come directly from the electrophoresis, eliminating the time needed for exposing gels to film and for manual reading of the sequences. The use of different colored dyes also reduces handling and allows more sequence to be produced at one time.

With increases in the number of samples per run and the length of sequences able to be read, along with decreases in handling time, the amount of sequence information that can be generated is limited mainly by the number of machines that can be run at once.

### *New sequencing technology*

For over 30 years, the basic chemistry of DNA sequencing did not change. Automation increased the speed of sequencing to the point that sequencing large eukaryotic genomes became possible. In the last few years, however, fundamentally new methods for sequencing have vastly accelerated the rate of sequence generation. Here we explore one new approach, which can generate 20 billion base pairs of sequence in a single run (figure 17.12). DNA is cleaved into smaller pieces, a few hundred base pairs, using a nebulizer—a device that converts the liquid to a very fine spray. Both ends are ligated to adapters that are complementary to specific primers. These DNA fragments are injected into a flow cell, which is like a microscope slide with seven channels, each containing a solid substrate with primers that complement the ligated ends of the DNA fragments. Millions of DNA fragments are placed in these channels, made single-stranded, and then amplified so there are clusters of fragments. Amplification works like DNA replication where a polymerase is added that recognizes the primer and starts copying. The fragments are again denatured to yield single-stranded molecules. They are now ready for sequencing. As with Sanger sequencing, deoxyribonucleotide triphosphates (dNTPs) have a fluorescent tag, but it can be removed. Four colors are used to distinguish each base. The fluorescent tag is reversibly attached to the 2′ position on the deoxyribose sugar and it blocks the 3′ OH so that only a single phosphodiester bond forms, but the blocking group can be removed after each round of DNA extension so the DNA strands continue to elongate. Very powerful charge-coupled device (CCD) cameras, once used exclusively by astronomers, record the pattern of fluorescence in the flow cell after each round of elongation. The technology works because a solid material holds the DNA fragments in place while they are being synthesized so that the repeated CCD images can be compiled and provide information about the sequence of each cluster of fragments. The amount of data generated each time another round of base pairs is added is enormous, so digital storage space and computational power to make sense of the data are the limiting factors.

## The polymerase chain reaction accelerates the process of analysis

The next revolution in molecular biology was the development of the **polymerase chain reaction (PCR).** Kary Mullis developed PCR in 1983 while he was a staff chemist at the Cetus Corporation; in 1993, he was awarded the Nobel Prize in chemistry for his discovery.

The idea of the polymerase chain reaction is simple: Two primers are used that are complementary to the opposite strands of a DNA sequence, oriented toward each other. When DNA polymerase acts on these primers and the sequence of interest, the primers produce complementary strands, each containing the other primer. If this procedure is done cyclically, the result is a large quantity of a sequence corresponding to the DNA that lies between the two primers (figure 17.13).

### *The PCR procedure*

Two developments turned this simple concept into a powerful technique. First, each cycle requires denaturing the DNA after each round of synthesis, which is easily done by raising the temperature; however, this destroys most polymerase enzymes. The solution was to isolate a DNA polymerase from a thermophilic, or heat-loving bacteria, *Thermus aquaticus*. This enzyme, called **Taq polymerase,** allows the reaction mixture to be repeatedly heated without destroying enzyme activity.

The second innovation was the development of machines with heating blocks that can be rapidly cycled over large temperature ranges with very accurate temperature control.

Thus each cycle of PCR involves three steps:

1. Denaturation (high temperature)
2. Annealing of primers (low temperature)
3. Synthesis (intermediate temperature)

Steps 1 to 3 are now repeated, and the two copies become four. It is not necessary to add any more polymerase, because

the heating step does not harm Taq polymerase. Each complete cycle, which takes only 1–2 min, doubles the number of DNA molecules. After 20 cycles, a single fragment produces more than one million ($2^{20}$) copies!

In this way, the process of PCR allows the amplification of a single DNA fragment from a small amount of a complex mixture of DNA. This result is similar to what is isolated using molecular cloning, but in the case of PCR, the DNA cannot be reintroduced directly into a cell. The PCR product can be analyzed using electrophoresis, cloned into a vector for other manipulations, or directly sequenced. There are limitations on the size of the fragment that can be synthesized in this way, but it has been adapted for an amazing number of uses.

### *Applications of PCR*

PCR, now fully automated, has revolutionized many aspects of science and medicine because it allows the investigation of minute samples of DNA. In criminal investigations, DNA fingerprints can now be prepared from the cells in a tiny speck of dried blood or from the tissue at the base of a single human hair. In medicine, physicians can detect genetic defects in very early embryos by collecting a single cell and amplifying its DNA. Due to its sensitivity, speed, and ease of use, technicians now routinely use PCR methods for these applications.

PCR has even been used to analyze mitochondrial DNA from the early human species *Homo neanderthalensis*. This application provides the first glimpse of data from extinct related species. The amplification of ancient DNA has been a controversial field because contamination with modern DNA is difficult to avoid. But it remains an active area of genetic research.

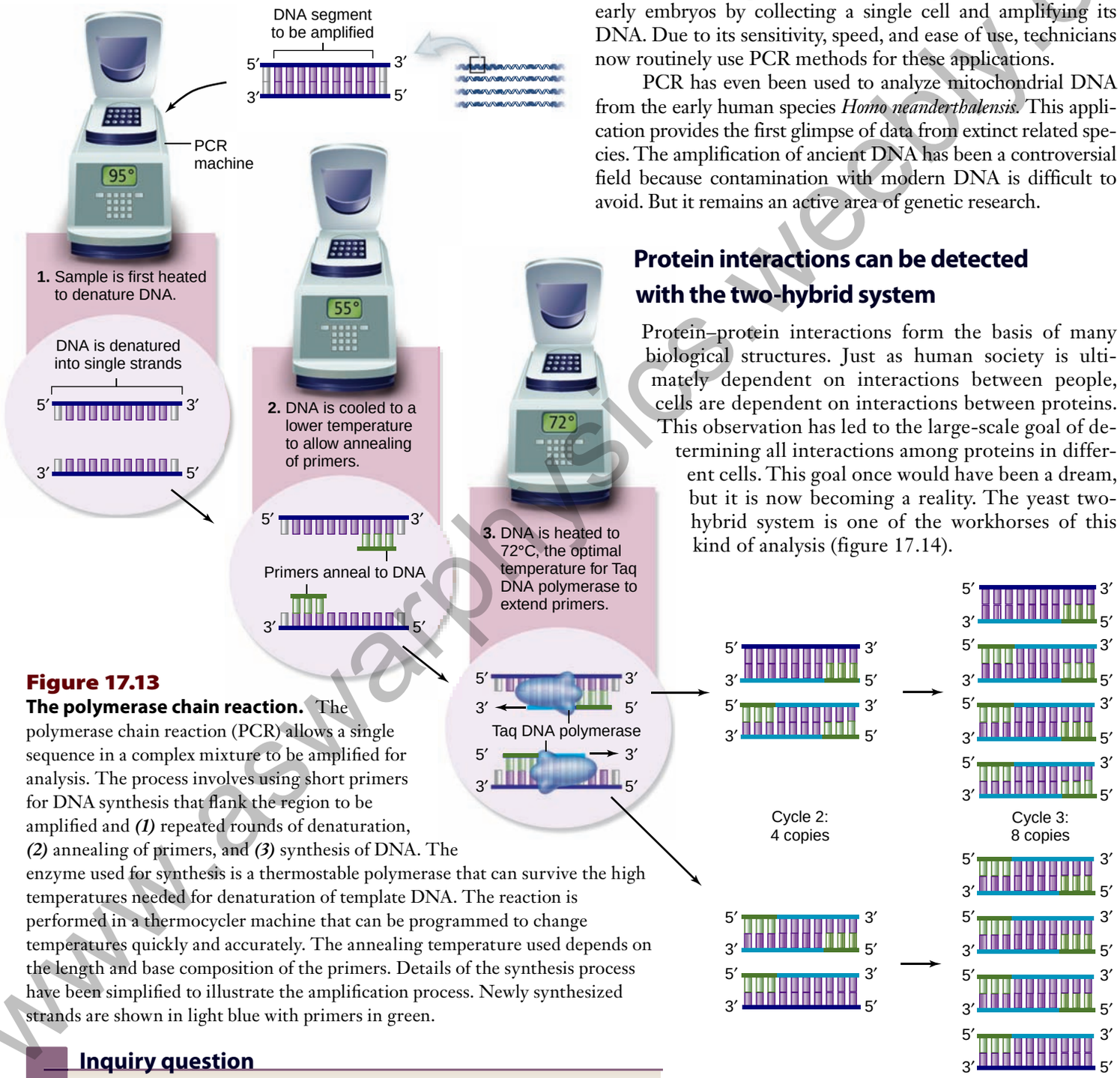



**Figure 17.13**
**The polymerase chain reaction.** The polymerase chain reaction (PCR) allows a single sequence in a complex mixture to be amplified for analysis. The process involves using short primers for DNA synthesis that flank the region to be amplified and *(1)* repeated rounds of denaturation, *(2)* annealing of primers, and *(3)* synthesis of DNA. The enzyme used for synthesis is a thermostable polymerase that can survive the high temperatures needed for denaturation of template DNA. The reaction is performed in a thermocycler machine that can be programmed to change temperatures quickly and accurately. The annealing temperature used depends on the length and base composition of the primers. Details of the synthesis process have been simplified to illustrate the amplification process. Newly synthesized strands are shown in light blue with primers in green.

**Inquiry question**

**?** **Could PCR be used to amplify mRNA?**

## Protein interactions can be detected with the two-hybrid system

Protein–protein interactions form the basis of many biological structures. Just as human society is ultimately dependent on interactions between people, cells are dependent on interactions between proteins. This observation has led to the large-scale goal of determining all interactions among proteins in different cells. This goal once would have been a dream, but it is now becoming a reality. The yeast two-hybrid system is one of the workhorses of this kind of analysis (figure 17.14).

The yeast two-hybrid system integrates much of the technology discussed in this chapter. It takes advantage of one feature of eukaryotic gene regulation, namely that the structure of proteins that turn on eukaryotic gene expression, transcription factors, have a modular structure.

The *Gal4* gene of yeast encodes a transcriptional activator with modular structure consisting of a DNA-binding domain that binds sequences in *Gal4*-responsive promoters, and an activation domain that interacts with the transcription apparatus to turn on transcription. The system uses two vectors: one containing a fragment of the *Gal4* gene that encodes the DNA-binding domain, and another containing a fragment of the *Gal4* gene that encodes the transcription activation domain. Neither of these alone can activate transcription.

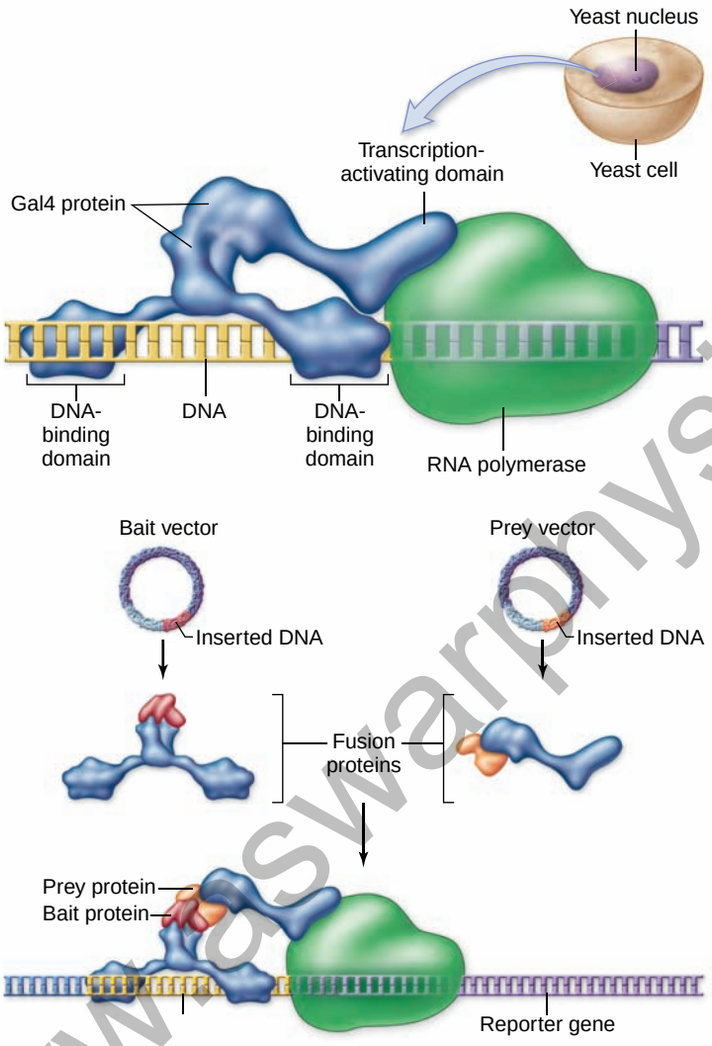



**Figure 17.14 The yeast two-hybrid system detects interacting proteins.** The Gal4 protein is a transcriptional activator *(top)*. The *Gal4* gene has been split and engineered into two different vectors such that one will encode only the DNA-binding domain (bait vector) and the other the transcription-activating domain (prey vector). When other genes are spliced into these vectors, they produce fusion proteins containing part of Gal4 and the proteins to be tested. If the proteins being tested interact, this will restore *Gal4* function and activate expression of a reporter gene.

When cDNAs are inserted into each of these two vectors in the proper reading frame, they are expressed as a single protein consisting of the protein of interest and part of the Gal4 activator protein (see figure 17.13). These hybrid proteins are called *fusion proteins* since they are literally fused in the same polypeptide chain. The DNA-binding hybrid is called the *bait*, and the activating domain hybrid is called the *prey*.

These vectors are inserted into cells of different mating types that can be crossed. One of these vectors also contains a so-called *reporter gene* encoding a protein that can be assayed for enzymatic activity. The reporter gene is under control of a *Gal4*-responsive regulatory region, so that when active *Gal4* is present, the reporter gene is expressed and can be detected by an enzymatic assay.

The DNA-binding hybrid binds to DNA adjacent to the reporter gene. When the two proteins in bait and prey interact, the prey hybrid brings the activating domain into position to turn on gene expression from the reporter gene (see figure 17.13).

The beauty of this system is that it is both simple and flexible. It can be used with two known proteins or with a known protein in the bait vector and entire cDNA libraries in the prey vector. In the latter case, all of the possible interactions in a cell type can be mapped.

It is already clear that even more protein interactions occur in cells than anticipated. In the future these data will form the basis for understanding the networks of protein interactions that make up the normal activities of a cell.

### Learning Outcomes Review 17.3

**The Southern blotting technique allows identification of a target DNA by separating single-stranded DNA fragments and hybridizing fragments of interest with a labeled probe. In living cells, DNA polymerase is a key enzyme in replication. DNA sequencing uses a modified DNA polymerase reaction that contains chain terminators, allowing fragments to be ordered in sequence. The polymerase chain reaction (PCR) produces a large amount of a specific DNA from a small amount of starting material. The yeast system for detecting protein–protein interactions involves a bait protein, a prey protein, and a reporter gene.**

- ***What key component of PCR allows the rapid amplification of a sample?***

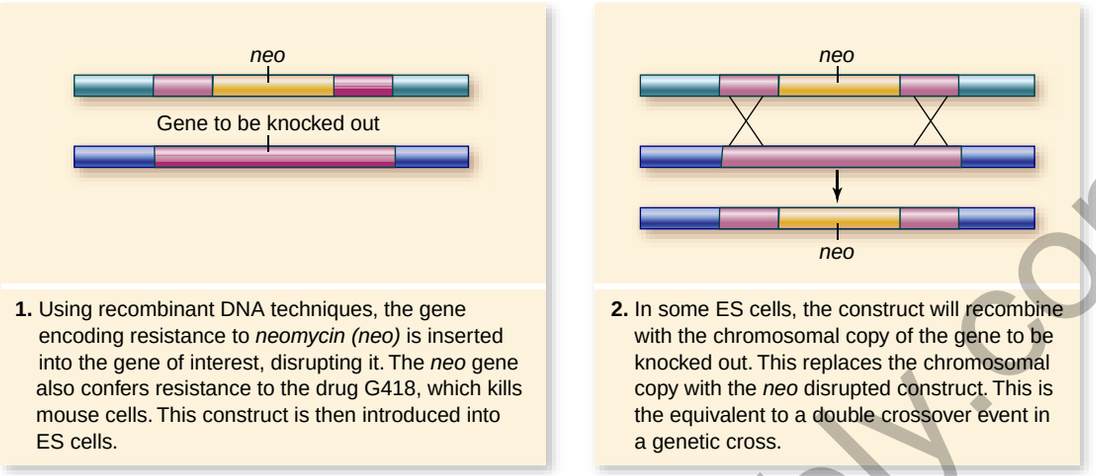
## 17.4 Genetic Engineering

### Learning Outcome

1. ***Describe three applications of cloning technology.***

The ability to clone individual genes for analysis ushered in an era of unprecedented advancement in research. At the time, these advancements were not accompanied by grand announcements of potential medical breakthroughs and other applications. The ability to truly genetically engineer any kind of cell

**Figure 17.15**
**Construction of a knockout mouse.** Steps in the construction of a knockout mouse. Some technical details have been omitted, but the basic concept is shown.




**1.** Using recombinant DNA techniques, the gene encoding resistance to *neomycin (neo)* is inserted into the gene of interest, disrupting it. The *neo* gene also confers resistance to the drug G418, which kills mouse cells. This construct is then introduced into ES cells.

**2.** In some ES cells, the construct will recombine with the chromosomal copy of the gene to be knocked out. This replaces the chromosomal copy with the *neo* disrupted construct. This is the equivalent to a double crossover event in a genetic cross.

or organism was a long way off. But we are now approaching this ability, and it has generated much excitement as well as controversy.

## Expression vectors allow production of specific gene products

A variety of specialized vectors have been constructed since the development of cloning technology. One very important type of vector are the **expression vectors.** These vectors contain the sequences necessary to drive expression of inserted DNA in a specific cell type, namely the correct sequences to permit transcription and translation of the sequences. The production of recombinant proteins in bacteria, for example, uses expression vectors with bacterial promoters and other control regions. The bacteria transformed by such vectors synthesize large amounts of the protein encoded by the inserted DNA. A number of pharmaceuticals have been produced in this way, the first of which was insulin, used to treat diabetes. (This type of application is discussed in more detail in the next section.)

## Genes can be introduced across species barriers

The ability to reintroduce genes into an original host cell, or to introduce genes into another host, is true genetic engineering. An animal containing a gene that has been introduced without the use of conventional breeding is called a **transgenic animal.** We will explore a number of uses of transgenic animals in medicine and agriculture, but it is important to realize that their original use was for basic research.

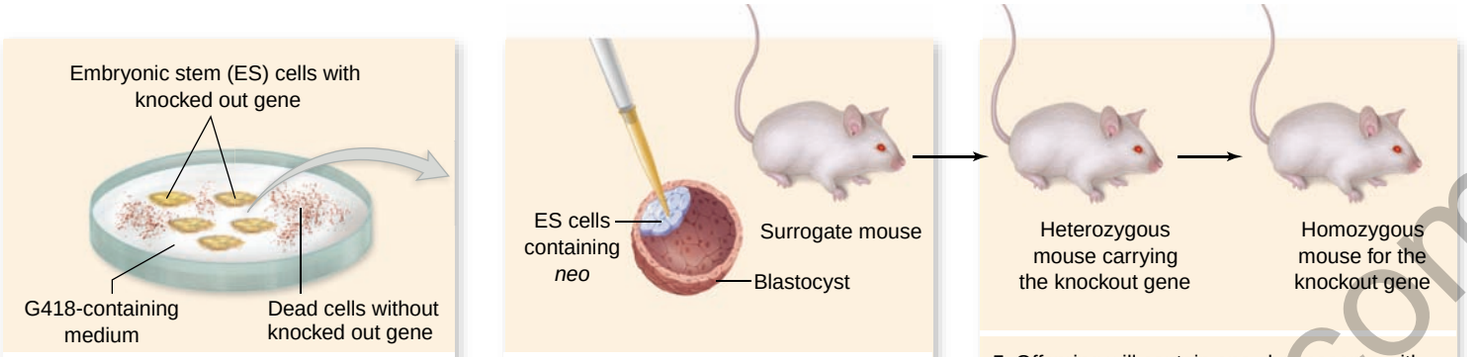
The ability to engineer genes in context or out of context allows an experimenter to ask questions that could never be asked otherwise. A dramatic example was the use of the *eyeless* gene from mice in *Drosophila*. When this mouse gene was introduced into *Drosophila*, it was shown to be able to substitute for a *Drosophila* gene in organizing the formation of eyes. It could even cause the formation of eyes in incorrect locations when expressed in tissue that did not normally form eyes. This amazing result shows that the formation of the compound eye in an insect is not so different from the formation of the complex vertebrate eye. This example is discussed in more detail in chapter 25.

## Cloned genes can be used to construct "knockout" mice

One of the most important technologies for research purposes is **in vitro mutagenesis**—the ability to create mutations at any site in a cloned gene to examine their effect on function. Rather than depending on mutations induced by chemical agents or radiation in intact organisms, which is time- and labor-intensive, the DNA itself is directly manipulated. The ultimate use of this approach is to be able to replace the wild-type gene with a mutant copy to test the function of the mutated gene. Developed first in yeast, this technique has now been extended to the mouse.

In mice, this technique has produced **knockout mice** in which a known gene is inactivated ("knocked out"). The effect of loss of this function is then assessed in the adult mouse, or if it is lethal, the stage of development at which function fails can be determined. The idea is simple, but the technology is quite complex. A streamlined description of the steps in this type of experiment are outlined as follows and illustrated in figure 17.15:

1. The cloned gene is disrupted by replacing it with a marker gene using recombinant DNA techniques. The marker gene codes for resistance to the antibiotic neomycin in bacteria, which allows mouse cells to survive when grown in a medium containing the related drug G418. The construction is done such that the marker gene is flanked by the DNA normally flanking the gene of interest in the chromosome.
2. The interrupted gene is introduced into **embryonic stem cells (ES cells).** These cells are derived from early embryos and can develop into different adult tissues. In these cells, the gene can recombine with the chromosomal copy of the gene based on the flanking DNA. This is the same kind of recombination used to map genes (chapter 13). The knockout gene with the drug resistance




**3.** The ES cells are placed on G418-containing medium. The G418 selects cells that have had a replacement event, and now contain a copy of the knocked out gene.

**4.** The ES cells containing the knocked out gene are injected into a blastocyst stage embryo and then implanted into a female to complete development.

**5.** Offspring will contain one chromosome with the gene of interest knocked out. Genetic crosses can then produce mice homozygous for the knocked out gene to assess the phenotype. This can range from lethality to no visible effect depending on the gene.

gene does not have an origin of replication, and thus it will be lost if no recombination occurs. Cells are grown in medium containing G418 to select for recombination events. (Only those containing the marker gene can grow in the presence of G418.)

3. The ES cells containing the knocked-out gene are injected into a blastocyst stage embryo, which is then implanted into a pseudopregnant female (one that has been mated with a vasectomized male and as a result has a receptive uterus). The pups from this female have one copy of the gene of interest knocked out. Transgenic animals can then be crossed to generate homozygous lines. These homozygous lines can be analyzed for phenotypes.

In conventional genetics, genes are identified based on mutants that show a particular phenotype. Molecular genetic techniques are then used to find the gene and isolate a molecular clone for analysis. The use of knockout mice is an example of **reverse genetics:** A cloned gene of unknown function is used to make a mutant that is deficient in that gene. A geneticist can then assess the effect on the entire organism of eliminating a single gene.

Sometimes this approach leads to surprises, such as happened when the gene for the p53 tumor suppressor was knocked out. Because this protein is found mutated in many human cancers and plays a key role in the regulation of the cell cycle (chapter 10), it was thought to be essential—the knockout was expected to be lethal. Instead, the mice were born normal; that is, development had proceeded normally. These mice do have a phenotype, however; they exhibit an increased incidence of tumors in a variety of tissues as they age.

### *Learning Outcome Review 17.4*

**Expression vectors that contain cloned genes allow the production of known proteins in different cells. This can be done for research purposes or to produce pharmaceuticals.**

- ***Why is recombination an essential factor in creating a "knockout" mouse?***

## 17.5 *Medical Applications*

### *Learning Outcomes*

1. ***Explain how eukaryote proteins can be produced in bacterial cells.***
2. ***Evaluate potential problems of gene therapy.***

The early days of genetic engineering led to a rash of startup companies, many of which are no longer in business. At the same time, all of the major pharmaceutical companies either began research in this area or actively sought to acquire smaller companies with promising technology. The number of applications of this technology are far too numerous to mention here, but we highlight a few; the section following discusses agricultural applications.

### Human proteins can be produced in bacteria

The first and perhaps most obvious commercial application of genetic engineering was the introduction of genes that encode clinically important proteins into bacteria. Because bacterial cells can be grown cheaply in bulk, bacteria that incorporate recombinant genes can synthesize large amounts of the proteins those genes specify, assuming the inserted gene has been designed to be expressed in a bacterial cell. This method has been used to produce several forms of human insulin and the immune system protein interferon, as well as other commercially valuable proteins, such as human growth hormone (figure 17.16) and erythropoietin, which stimulates red blood cell production.

Among the medically important proteins now manufactured by these approaches are **atrial peptides,** small proteins that may provide a new way to treat high blood pressure and kidney failure. Another is **tissue plasminogen activator (TPA),** a human protein synthesized in minute amounts that causes blood clots to dissolve and that if used within the first

**Figure 17.16 Genetically engineered mouse with human growth hormone.** These two mice are from an inbred line and differ only in that the large one has one extra gene: the gene encoding human growth hormone. The gene was added to the mouse's genome and is now a stable part of the mouse's genetic endowment.

3 hr after an ischemic stroke (i.e., one that blocks blood to the brain) can prevent catastrophic disability.

A problem with this approach has been the difficulty of separating the desired protein from the others the bacteria make. The purification of proteins from such complex mixtures is both time-consuming and expensive, but it is still easier than isolating the proteins from bulk processing of the tissues of animals, which is how such proteins used to be obtained. For example, insulin was previously extracted from hog pancreases because hog insulin was similar to human insulin.

## Recombinant DNA may simplify vaccine production

Another area of potential significance involves the use of genetic engineering to produce vaccines against communicable diseases. Two types of vaccines are under investigation: *subunit vaccines* and *DNA vaccines*.

### *Subunit vaccines*

**Subunit vaccines** may be developed against viruses such as those that cause herpes and hepatitis. Genes encoding a part, or subunit, of the protein polysaccharide coat of the herpes simplex virus or hepatitis B virus are spliced into a fragment of the vaccinia (cowpox) virus genome (figure 17.17).

The vaccinia virus, which British physician Edward Jenner used more than 200 years ago in his pioneering vaccinations against smallpox, is now used as a vector to carry the herpes or hepatitis viral coat gene into cultured mammalian cells. These cells produce many copies of the recombinant vaccinia virus, which has the outside coat of a herpes or hepatitis virus. When this recombinant virus is injected into a mouse or rabbit, the immune system of the infected animal produces antibodies directed against the coat of the recombinant virus. The animal then develops an immunity to herpes or hepatitis virus.

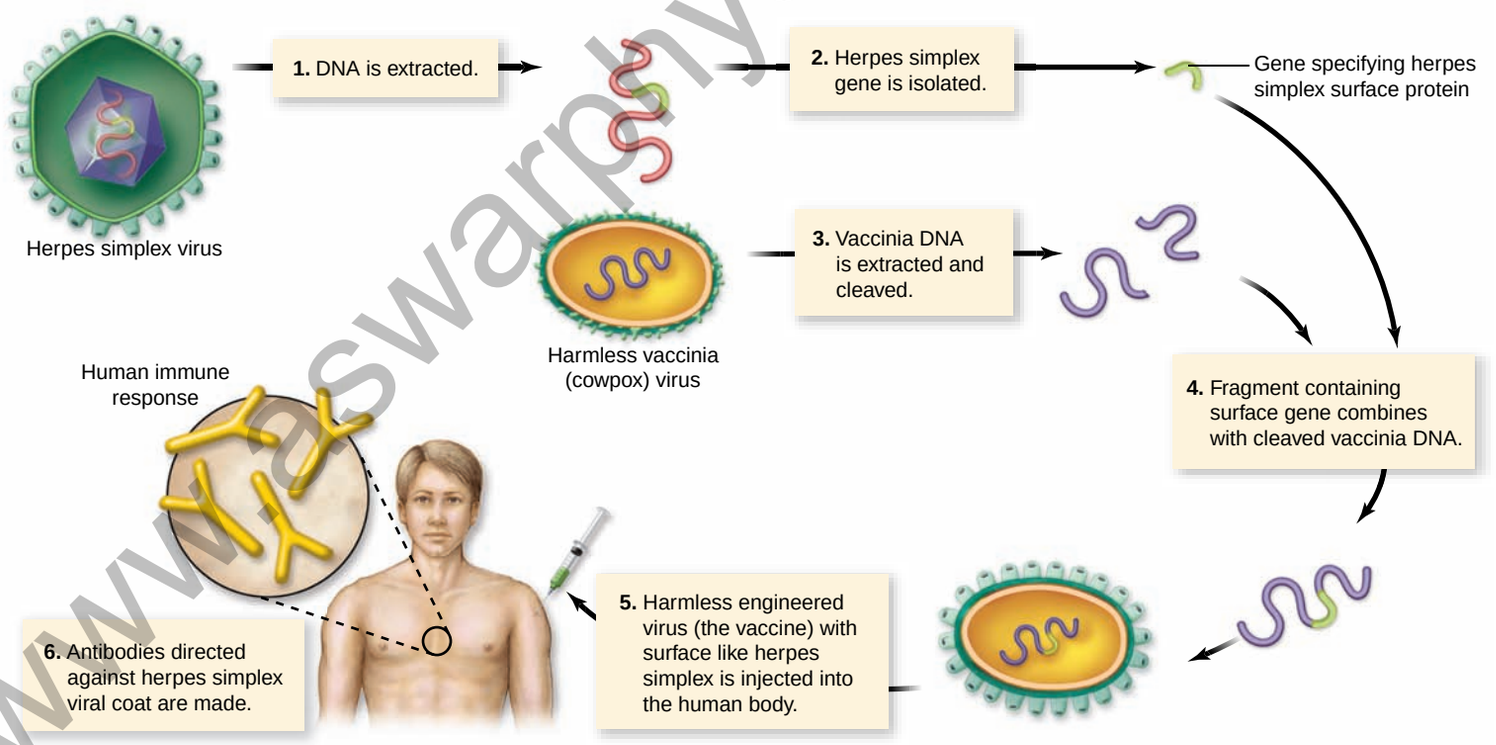



**Figure 17.17 Strategy for constructing a subunit vaccine against herpes simplex.** Recombinant DNA techniques can be used to construct vaccines for a single protein from a virus or bacterium. In this example, the protein is a surface protein from the herpes simplex virus.

Vaccines produced in this way are harmless because the vaccinia virus is benign, and only a small fragment of the DNA from the disease-causing virus is introduced via the recombinant virus.

The great attraction of this approach is that it does not depend on the nature of the viral disease. In the future, similar recombinant viruses may be used in humans to confer resistance to a wide variety of viral diseases.

### *DNA vaccines*

In 1995, the first clinical trials began to test a novel new kind of **DNA vaccine,** one that depends not on antibodies but rather on the second arm of the body's immune defense, the so-called *cellular immune response*, in which blood cells known as killer T cells attack infected cells (chapter 52). The first DNA vaccines spliced an influenza virus gene encoding an internal nucleoprotein into a plasmid, which was then injected into mice. The mice developed a strong cellular immune response to influenza. Although new and controversial, the approach offers great promise.

## Gene therapy can treat genetic diseases directly

In 1990, researchers first attempted to combat genetic defects by the transfer of human genes. When a hereditary disorder is the result of a single defective gene, an obvious way to cure the disorder would be to add a working copy of the gene. This approach is being used in an attempt to combat cystic fibrosis, and it offers the potential of treating muscular dystrophy and a variety of other disorders (table 17.1).

**TABLE 17.1 Diseases Being Treated in Clinical Trials of Gene Therapy**

| Disease |
|---|
| Cancer (melanoma, renal cell, ovarian, neuroblastoma, brain, head and neck, lung, liver, breast, colon, prostate, mesothelioma, leukemia, lymphoma, multiple myeloma) |
| SCID (severe combined immunodeficiency) |
| Cystic fibrosis |
| Gaucher disease |
| Familial hypercholesterolemia |
| Hemophilia |
| Purine nucleoside phosphorylase deficiency |
| $\alpha_1$-Antitrypsin deficiency |
| Fanconi anemia |
| Hunter syndrome |
| Chronic granulomatous disease |
| Rheumatoid arthritis |
| Peripheral vascular disease |
| Acquired immunodeficiency syndrome (AIDS) |
| Duchenne muscular dystrophy |
| Macular degeneration (wet variety) |
| Batten disease (neurological disorder) |

Clinical trials for treating macular degeneration, a genetic eye disease, using an RNAi vector (see chapter 16) are promising. Individuals with a certain type of macular degeneration lose their sight because of the uncontrolled proliferation of blood vessels under the retina. For the patient, it is a lot like looking through a car windshield with broken wipers in the middle of a thunderstorm. RNAi gene therapy involves injection of double-stranded RNA coding for a gene necessary for blood vessel proliferation. The RNAi mechanism has the counterintuitive effect of supressing production of the protein needed for blood vessel development, preventing progression of the disease.

One disease that illustrates both the potential and the problems with gene therapy is **severe combined immunodeficiency disease (SCID).** This disease has multiple forms, including an X-linked form (X-SCID) and a form that lacks the enzyme adenosine deaminase (ADA-SCID).

Recent trials for both of these forms showed great initial promise, with patients exhibiting restoration of immune function. But then problems arose in the case of the X-SCID trial when a patient developed a rare form of leukemia. Since that time, two other patients have developed the same leukemia, and it appears to be due to the gene therapy itself. The vector used to introduce the X-SCID gene integrated into the genome next to a proto-oncogene called *LMO2* in all three cases. Activation of this gene can cause childhood leukemias.

The insertion of a gene during gene therapy has always been a random event, and it has been a concern that the insertion could inactivate an essential gene, or turn on a gene inappropriately. That effect had not been observed prior to the X-SCID trial, despite a large number of genes introduced into blood cells in particular. For leukemia to occur in 15% of the patients treated implies that some influence of the genetic background associated with X-SCID potentiates this development. This possibility is supported by the observation that the ADA-SCID patients treated have not been affected thus far.

On the positive side, 15 children treated successfully are still alive, 14 of them after more than four years, with functioning immune systems. On the negative side, three other children treated have developed leukemia.

When we understand the basis of the preferential integration in the case of X-SCID, it should be possible to overcome this unfortunate result. In the meantime, the investigators have halted the trial and are working on new vectors to reduce the possibility of this preferential integration.

### *Learning Outcomes Review 17.5*

**Recombinant DNA technology has allowed genes from eukaryotes, such as humans, to be isolated, inserted into vectors, and recombined into bacterial genomes, where the genes' products can be mass-produced. Gene therapy is the process of using genetic engineering to replace defective genes; however, in some cases unwanted effects result from random gene insertion.**

- ***What might be some undesirable effects of treating patients with human proteins manufactured in and isolated from other organisms?***

# 17.6 Agricultural Applications

### Learning Outcomes

1. *Compare recombinant technology techniques in plants with those in bacteria.*
2. *Describe the controversial issues in the use of GM plants.*

Perhaps no area of genetic engineering touches all of us so directly as the applications that are being used in agriculture today. Crops are being modified to resist disease, to be tolerant of herbicides, and for changes in nutritional and other content in a variety of ways. Plant systems are also being used to produce pharmaceuticals by "biopharming," and domesticated animals are being genetically modified to produce biologically active compounds.

## The Ti plasmid can transform broadleaf plants

In plants, the primary experimental difficulty has been identifying a suitable vector for introducing recombinant DNA. Plant cells do not possess the many plasmids that bacteria have, so the choice of potential vectors is limited.

### *The Ti plasmid*

The most successful results thus far have been obtained with the **Ti (tumor-inducing) plasmid** of the plant bacterium *Agrobacterium tumefaciens*, which normally infects broadleaf plants such as tomato, tobacco, and soybean. Part of the Ti plasmid integrates into the plant DNA, and researchers have succeeded in attaching other genes to this portion of the plasmid (figure 17.18). The characteristics of a number of plants have been altered using this technique, which should be valuable in improving crops and forests.

Among the features scientists would like to affect are resistance to disease, frost, and other forms of stress; nutritional balance and protein content; and herbicide resistance. All of these traits have either been modified or are being modified. Unfortunately, *Agrobacterium* normally does not infect cereal plants such as corn, rice, and wheat, but alternative methods can be used to introduce new genes into them.

### *Other methods of gene insertion*

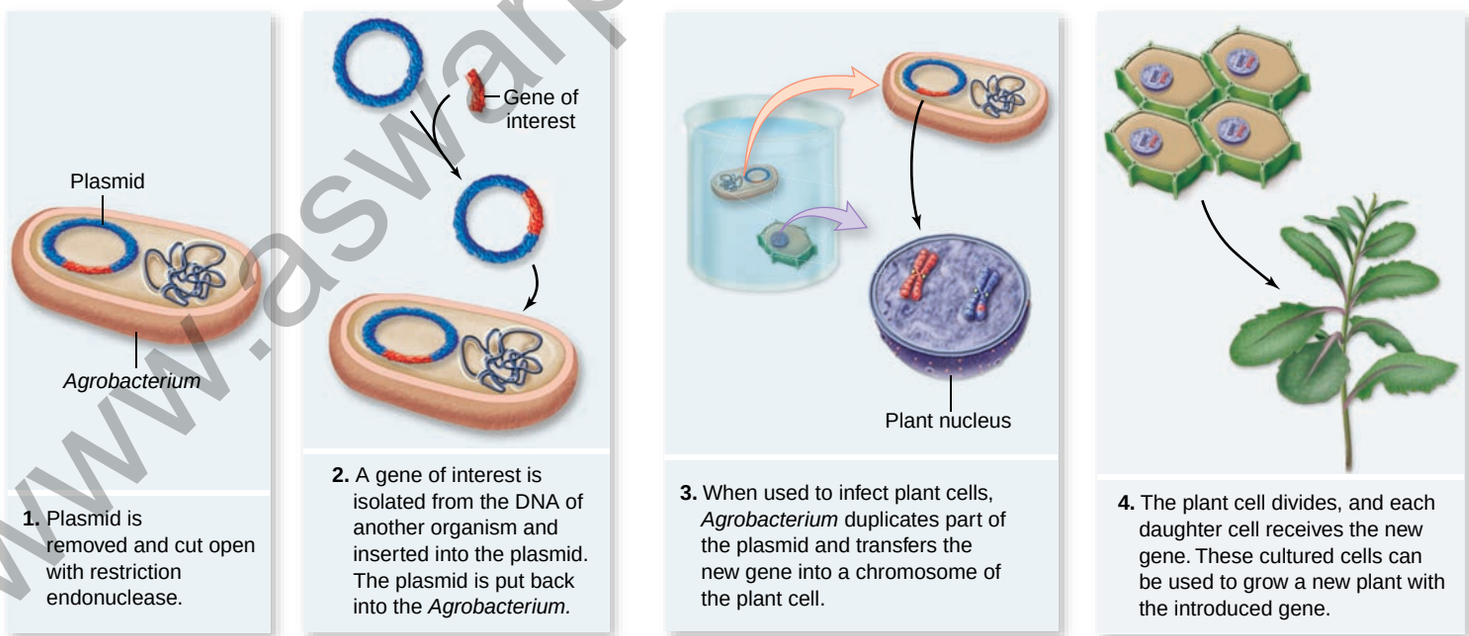
For cereal plants that are not normally infected by *Agrobacterium*, other methods have been used. One popular method, "the gene gun," uses bombardment with tiny gold or tungsten particles coated with DNA. This technique has the advantage of being usable for any species, but it does not allow as precise an engineering because the copy number of introduced genes is much harder to control.

Recently, modifications of the *Agrobacterium* system have allowed it to be used with cereal plants, so the gene gun technology may not be used much in the future. A new bacterium has also been manipulated to function like *Agrobacterium*, offering another potential alternative method of engineering cereal crops.

It is clear that genetic modification of crop plants of all sorts has become a mature technology, which should accelerate the production of a variety of transgenic crops.

## Herbicide-resistant crops allow no-till planting

Recently, broadleaf plants have been genetically engineered to be resistant to **glyphosate,** a powerful, biodegradable herbicide that kills most actively growing plants (figure 17.19). Glyphosate works by inhibiting an enzyme called 5-enolpyruvylshikimate-3-phosphate (EPSP) synthetase, which plants require to produce aromatic amino acids.




1. Plasmid is removed and cut open with restriction endonuclease.
2. A gene of interest is isolated from the DNA of another organism and inserted into the plasmid. The plasmid is put back into the *Agrobacterium.*
3. When used to infect plant cells, *Agrobacterium* duplicates part of the plasmid and transfers the new gene into a chromosome of the plant cell.
4. The plant cell divides, and each daughter cell receives the new gene. These cultured cells can be used to grow a new plant with the introduced gene.

**Figure 17.18 The Ti plasmid.** This *Agrobacterium tumefaciens* plasmid is used in plant genetic engineering.

**SCIENTIFIC THINKING**

**Hypothesis:** *Petunias can acquire tolerance to the herbicide glyphosate by overexpressing EPSP synthase.*

**Prediction:** *Transgenic petunia plants with a chimeric EPSP synthase gene with strong promoter will be glyphosate tolerant.*

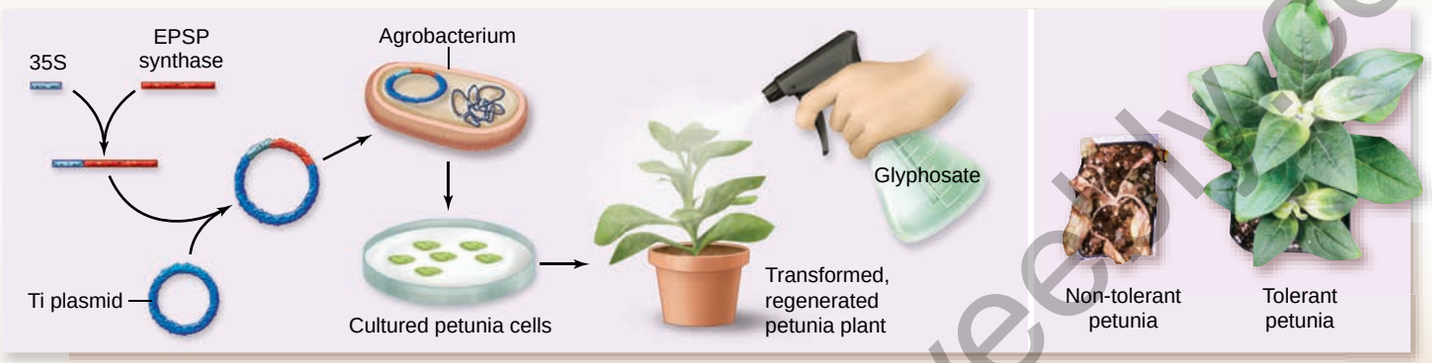
**Test:**

*1. Use restriction enzymes and ligase to "paste" the cauliflower mosaic virus promoter (35S) to the EPSP synthase gene and insert the construct in Ti plasmids.*

*2. Transform* Agrobacterium *with the recombinant plasmid.*

*3. Infect petunia cells and regenerate plants. Regenerate uninfected plants as controls.*

*4. Challenge plants with glyphosate.*




**Result:** *Glyphosate kills control plants, but not transgenic plants.*

**Conclusion:** *Additional EPSP synthase provides glyphosate tolerance.*

**Further Experiments:** *The transgenic plants are tolerant, but not resistant (note bleaching at shoot tip). How could you determine if additional copies of the gene would increase tolerance? Can you think of any downsides to expressing too much EPSP synthase in petunia?*

**Figure 17.19** **Genetically engineered herbicide resistance.**

Humans do not make aromatic amino acids; we get them from our diet, so we are unaffected by glyphosate. To make glyphosate-resistant plants, scientists used a Ti plasmid to insert extra copies of the EPSP synthetase gene into plants. These engineered plants produce 20 times the normal level of EPSP synthetase, enabling them to synthesize proteins and grow despite glyphosate's suppression of the enzyme. In later experiments, a bacterial form of the EPSP synthetase gene that differs from the plant form by a single nucleotide was introduced into plants via Ti plasmids; the bacterial enzyme is not inhibited by glyphosate (see figure 17.19).

These advances are of great interest to farmers because a crop resistant to glyphosate would not have to be weeded—the field could simply be treated with the herbicide. Because glyphosate is a broad-spectrum herbicide, farmers would no longer need to employ a variety of different herbicides, most of which kill only a few kinds of weeds. Furthermore, glyphosate breaks down readily in the environment, unlike many other herbicides commonly used in agriculture. A plasmid is actively being sought for the introduction of the EPSP synthetase gene into cereal plants, making them also glyphosate-resistant.

At this point four important crop plants have been modified to be glyphosate-resistant: maize (corn), cotton, soybeans, and canola. The use of glyphosate-resistant soy has been especially popular, accounting for 60% of the global area of GM (genetically modified) crops grown in nine countries worldwide. In the United States, 90% of soy currently grown is GM soy. Global variation in the use of GM crops has occurred, with the Americas, led by the United States, the largest adopter. The area currently with the largest growth in the use of GM crops is Asia, while Europe has been the slowest to move to their use.

## Bt crops are resistant to some insect pests

Many commercially important plants are attacked by insects, and the usual defense against such attacks has been to apply insecticides. Over 40% of the chemical insecticides used today are targeted against boll weevils, bollworms, and other insects that eat cotton plants. Scientists have produced plants that are resistant to insect pests, removing the need to use many externally applied insecticides.

The approach is to insert into crop plants genes encoding proteins that are harmful to the insects that feed on the plants, but harmless to other organisms. The most commonly used protein is a toxin produced by the soil bacterium *Bacillus thuringiensis (Bt toxin)*. When insects ingest Bt toxin, endogenous enzymes convert it into an insect-specific toxin, causing paralysis and death. Because these enzymes are not found in other animals, the protein is harmless to them.

The same four crops that have been modified for herbicide resistance have also been modified for insect resistance using the Bt toxin. The use of Bt maize is the second most common GM crop globally, representing 14% of global area of GM crops in nine countries. The global distribution of these crops is also similar to the herbicide-resistant relatives.

Given the popularity of both of these types of crop modifications, it is not surprising that they have also been combined,

so-called *stacked GM crops*, in both maize and cotton. Stacked crops now represent 9% of global area of GM crops.

## Golden Rice shows potential of GM crops

One of the successes of GM crops is the development of Golden Rice. This rice has been genetically modified to produce β-carotene (provitamin A). The World Health Organization (WHO) estimates that vitamin A deficiency affects between 140 and 250 million preschool children worldwide. The deficiency is especially severe in developing countries where the major staple food is rice. Provitamin A in the diet can be converted by enzymes in the body to vitamin A, alleviating the deficiency.

Golden Rice is named for its distinctive color imparted by the presence of β-carotene in the endosperm (the outer layer of rice that has been milled). Rice does not normally make β-carotene in endosperm tissue, but does produce a precursor, geranyl geranyl diphosphate, that can be converted by three enzymes, phytoene synthase, phytoene desaturase, and lycopene β-cyclase, to β-carotene. These three genes were engineered to be expressed in endosperm and introduced into rice to complete the biosynthetic pathway producing β-carotene in endosperm (figure 17.20).

This is an interesting case of genetic engineering for two reasons. First, it introduces a new biochemical pathway in tissue of the transgenic plants. Second, it could not have been done by conventional breeding as no rice cultivar known produces these enzymes in endosperm. The original constructs used two genes from daffodil and one from a bacterium (see figure 17.20). There are many reasons to expect failure in the introduction of a biochemical pathway without disrupting normal metabolism. That the original form of Golden Rice makes significant amounts of β-carotene in an otherwise healthy plant is impressive. A second-generation version that makes much higher levels of β-carotene has also been produced by using the gene for phytoene synthase from maize in place of the original daffodil gene.

Golden Rice was originally constructed in a public facility in Switzerland and made available for free with no commercial entanglements. Since its inception, Golden Rice has been improved both by public groups and by industry scientists, and these improved versions are also being made available without commercial strings attached.

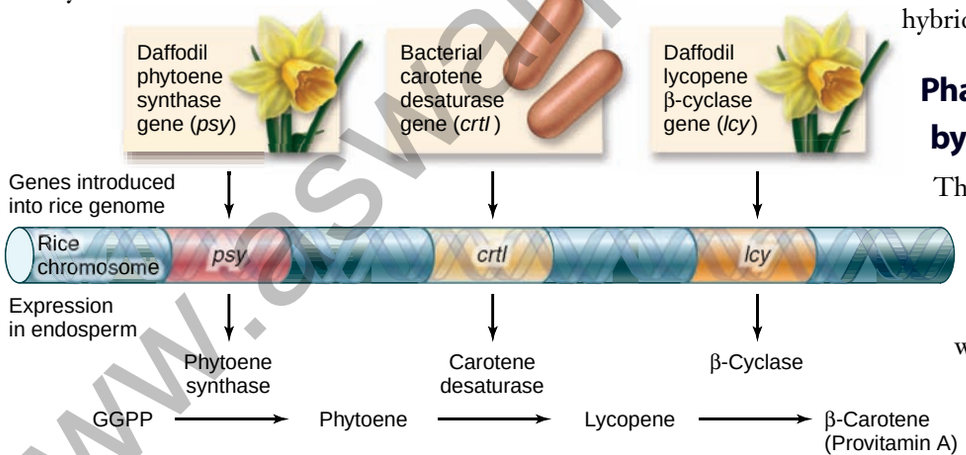



**Figure 17.20 Construction of Golden Rice.** Rice does not normally express the enzymes needed to synthesize β-carotene in endosperm. Three genes were added to the rice genome to allow expression of the pathway for β-carotene in endosperm. The source of the genes and the pathway for synthesis of β-carotene is shown. The result is Golden Rice, which contains enriched levels of β-carotene in endosperm.

## GM crops raise a number of social issues

The adoption of GM crops has been resisted in some places for a variety of reasons. Some people have wondered about the safety of these crops for human consumption, the likelihood of introduced genes moving into wild relatives, and the possible loss of biodiversity associated with these crops.

Powerful forces have aligned on opposing sides in this debate. On the side in favor of the use of GM crops are the multinational companies that are utilizing this technology to produce seeds for the various GM crops. On the other side are a variety of political organizations that are opposed to genetically modified foods. Scientists can be found on both sides of the controversy.

The controversy originally centered on the safety of introduced genes for human consumption. In the United States, this issue has been "settled" for the crops already mentioned, and a large amount of GM soy and maize is consumed in this country. Although some opponents still raise the issue of long-term use and allergic reactions, no negative effects have been documented so far. Existing crops will be monitored for adverse effects, and each new modification will require regulatory approval for human consumption.

Another contention has been the fear that genes might spread outside of the GM crops into wild relatives, a process called introgression. But at this point there is no indication of that happening. One study showed no evidence for the movement of genes from GM crops into native species in Mexico, despite earlier studies indicating significant movement of introduced genes.

This finding does not mean that such movement is impossible, but it does indicate that it seems not to have occurred at present. It is clear that this area requires more study. This issue will likely have to be considered on a case-by-case basis because the number of wild relatives and the ease of hybridization varies greatly among crop plants.

## Pharmaceuticals can be produced by "biopharming"

The medicinal use of plants goes back as far as recorded history. In modern times, the pharmaceutical industry began by isolating biologically active compounds from plants. This approach began to change when in 1897, the Bayer company introduced acetyl salicylic acid, otherwise known as aspirin. This compound was a synthetic version of the compound salicylic acid, which was isolated from the bark of the white willow. The production of pharmaceuticals has since been dominated more by organic synthesis and less by the isolation of plant products.

One exception to this trend is cancer chemotherapeutic agents such as taxol, vinblastine, and vincristine, all of which were isolated from plant sources. In an interesting closing of

the historical loop, the industry is now looking at using transgenic plants for the production of useful compounds.

The first human protein to be produced in plants was human serum albumin, which was produced in 1990 by both genetically engineered tobacco and potato plants. Since that time more than 20 proteins have been produced in transgenic plants. This first crop of transgenic pharmaceuticals are now in the regulatory pipeline.

### *Recombinant subunit vaccines*

One promising aspect of plant genetic engineering is the production of recombinant subunit vaccines, which were discussed earlier. One of these, being produced in genetically modified potatoes, is a vaccine against Norwalk virus. Norwalk virus is not a common source of illness, but it reached the public consciousness when cruise ships were forced to cancel cruises due to outbreaks of the virus among passengers. The vaccine is now in clinical trials. A vaccine against rabies produced in transgenic spinach is also in clinical trials.

One obvious advantage of using plants for vaccine production is scalability. It has been estimated that 250 acres of greenhouse space could produce enough transgenic potato plants to supply Southeast Asia's need for hepatitis B vaccine.

### *Recombinant antibodies*

Molecular cloning and immunology can be combined to produce antibodies in transgenic plants that are normally made by blood cells in vertebrates. The synthesis of monoclonal antibodies in plant systems is a promising use of transgenic plants.

A number of potentially therapeutic antibodies are being produced in plants, and some of these have reached clinical trial stage. One interesting example is an antibody against the bacterium responsible for dental caries, commonly known as tooth decay. It would make a visit to the dentist more pleasant to have a topical antibody applied instead of a drill.

## Domesticated animals can also be genetically modified

Humans have been breeding and selecting domestic animals for thousands of years. With the advent of genetic engineering, this process can be accelerated, and genes can be introduced from other species.

The production of transgenic livestock is in an early stage, and it is hard to predict where it will go. At this point, one of the uses of biotechnology is not to construct transgenic animals, but to use DNA markers to identify animals and to map genes that are involved in such traits as palatability in food animals, texture of hair or fur, and other features of animal products. Molecular techniques combined with the ability to clone domestic animals (chapter 19) could produce improved animals for economically desirable traits.

Transgenic animal technology has not been as successful as initially predicted. Early on, pigs were engineered to overproduce growth hormone in the hope that this would lead to increased and faster growth. These animals proved to have only slightly increased growth, and they had lower fat levels, which reduces flavor, as well as showing other deleterious effects. The main use thus far has been engineering animals to produce pharmaceuticals in milk—another example of the biopharming concept.

One interesting idea for transgenics is the EnviroPig. This animal has been engineered with the gene for phytase under the control of a salivary gland-specific promoter. The enzyme phytase breaks down phosphorus in the feed and can reduce phosphate excretion by up to 70%. Because phosphate is a major problem in pig waste, reducing its excretion could be a large environmental benefit.

As with GM crops, fears exist about the consumption of meat from transgenic animals. At this point, these fears do not seem to be based on sound science; nevertheless, every transgenic animal produced that is intended for consumption will need to be considered on a case-by-case basis.

### *Learning Outcomes Review 17.6*

**Genes can be introduced into plants using the bacterial Ti plasmid and techniques similar to those for bacteria. To date, herbicide resistance, pathogen protection, nutritional enhancement, and vaccine and drug production have been targets of agricultural genetic engineering. Controversy regarding the use of GM plants has centered on the potential of unforeseen effects on human health and on the environment.**

- ***How might a recombinant gene for Bt toxin production "escape" from a crop plant and move into wild plants?***

# Chapter Review

## 17.1 DNA Manipulation

***Restriction enzymes cleave DNA at specific sites.***

DNA molecules fragmented by known type II restriction endonucleases can be ordered into a physical map of DNA.

***DNA ligase allows construction of recombinant molecules.***

Just as in DNA replication, DNA ligase catalyzes formation of a phosphodiester bond between nucleotides, forming a recombinant molecule.

***Gel electrophoresis separates DNA fragments.***

An electric field applied to a gel matrix causes DNA to migrate through the matrix. Smaller fragments migrate farther than large fragments (figure 17.2).

***Transformation allows introduction of foreign DNA into* E. coli.**

Artificial transformation techniques introduce foreign DNA into *E. coli* cells, which are then termed transgenic.

## 17.2 Molecular Cloning

***Host–vector systems allow propagation of foreign DNA in bacteria.***

Plasmids, and artificial chromosomes can be used as vectors. Foreign DNA is inserted using restriction enzymes and DNA ligase; once the vector is inside the host cell, it is replicated during the cell cycle.

***DNA libraries contain the entire genome of an organism.***

A DNA library is a complex mixture of DNAs collected into vectors. These libraries may be probed for a sequence of interest.

***Reverse transcriptase can make a DNA copy of RNA.***

A library can also be created for the expressed parts of the genome by isolating RNA and converting it to cDNA using the enzyme reverse transcriptase (figure 17.5).

***Hybridization allows identification of specific DNAs in complex mixtures.***

DNA can be reversibly denatured and renatured, resulting in single- and then double-stranded DNA. Renaturation of complementary strands from different sources is called hybridization. Known DNA can be labeled to identify complementary strands.

***Specific clones can be isolated from a library.***

Hybridization is the most common way to identify a gene of interest in a library.

## 17.3 DNA Analysis

***Restriction maps provide molecular "landmarks."***

The first physical maps of DNA molecules were based on the sites of restriction enzyme cleavage.

***Southern blotting reveals DNA differences.***

In Southern blotting, a complex mixture is separated by electrophoresis and transferred to filter paper. Specific sequences of DNA can then be identified by hybridization. Similar techniques can identify mRNA (Northern blot) and proteins (Western blot).

Restriction fragment length polymorphisms (RFLPs) identified by Southern blotting reveal individual differences in DNA. DNA fingerprinting uses probes to locate polymorphic DNA fragments.

***DNA sequencing provides information about genes and genomes.***

DNA sequencing uses chain-terminating reagents to identify the order of fragments and from this to infer the sequence of bases (figures 17.11, 17.12).

***The polymerase chain reaction accelerates the process of analysis.***

The polymerase chain reaction (PCR) amplifies a single small DNA fragment using two short primers that flank the region to be amplified. Cyclic replication is accomplished via heating and cooling; a key factor is Taq polymerase, which is not denatured at high temperature.

***Protein interactions can be detected with the two-hybrid system.***

The yeast two-hybrid system relies on fusion proteins and a reporter gene to study protein–protein interactions (figure 17.13).

## 17.4 Genetic Engineering

***Expression vectors allow production of specific gene products.***

Expression vectors contain the promoters and enhancers necessary to drive expression of the inserted DNA.

***Genes can be introduced across species barriers.***

Transgenic organisms can be constructed to express genes in a different species or to create mutations in genes to assess phenotype.

***Cloned genes can be used to construct "knockout" mice.***

In knockout mice, a gene is inactivated by replacing the wild-type version with a mutant copy (figure 17.15). In this way, the function of the gene can be analyzed and clarified.

## 17.5 Medical Applications

***Human proteins can be produced in bacteria.***

Bacterial production of human proteins such as insulin has allowed better results and has increased production to treat disease.

***Recombinant DNA may simplify vaccine production.***

Subunit vaccines produced in cultured cells have been shown to be effective in animals.

DNA vaccines, which alter the cellular immune response, are also promising. Both these approaches require further testing.

***Gene therapy can treat genetic diseases directly.***

Gene therapy involves inserting a normal gene to replace a defective one. Unfortunately, trials of two promising therapies had unintended and fatal consequences in 15% of patients.

## 17.6 Agricultural Applications

***The Ti plasmid can transform broadleaf plants.***

The tumor-inducing (Ti) plasmid from a plant bacterium is used to transfer genes into broad-leaf plants. A number of applications are currently in use.

***Herbicide-resistant crops allow no-till planting.***

***Bt crops are resistant to some insect pests.***

***Golden Rice shows potential of GM crops.***

***GM crops raise a number of social issues.***

Concerns about GM plants include unintended allergic reactions to proteins inserted from a different organism and spread of foreign genes into noncultivated plants in the environment.

***Pharmaceuticals can be produced by "biopharming."***

***Domesticated animals can also be genetically modified.***

To date, results with transgenic animals have been mixed.

# Review Questions

## UNDERSTAND

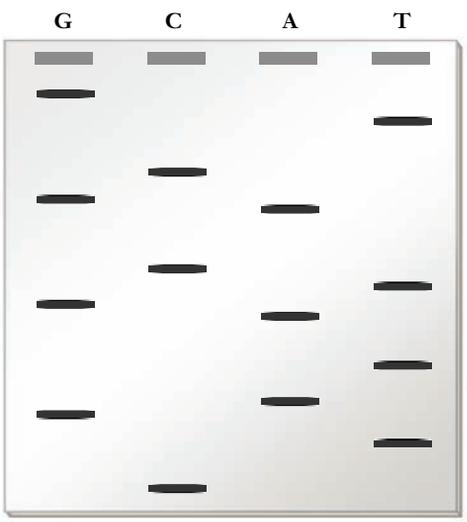
1. A recombinant DNA molecule is one that is
   a. produced through the process of crossing over that occurs in meiosis.
   b. constructed from DNA from different sources.
   c. constructed from novel combinations of DNA from the same source.
   d. produced through mitotic cell division.
2. What is the basis of separation of different DNA fragments by gel electrophoresis?
   a. The negative charge on DNA
   b. The size of the DNA fragments
   c. The sequence of the fragments
   d. The presence of a dye
3. The basic logic of enzymatic DNA sequencing is to produce
   a. a nested set of DNA fragments produced by restriction enzymes.
   b. a nested set of DNA fragments that each begin with different bases.
   c. primers to allow PCR amplification of the region between the primers.
   d. a nested set of DNA fragments that end with known bases.
4. A DNA library is
   a. an orderly array of all the genes within an organism.
   b. a collection of vectors.
   c. the collection of plasmids found within a single *E. coli.*
   d. a collection of DNA fragments representing the entire genome of an organism.
5. Molecular hybridization is used to
   a. generate cDNA from mRNA.
   b. introduce a vector into a bacterial cell.
   c. screen a DNA library.
   d. introduce mutations into genes.
6. How does the yeast two-hybrid system detect protein–protein interactions?
   a. Binding of fusion partners triggers a signal cascade that alters gene expression.
   b. Fusion partners are detected using radioactive probes of Western blots.
   c. Protein–protein binding of fusion partners triggers expression of a reporter gene.
   d. Protein–protein binding of fusion partners triggers expression of the *Gal4* gene.
7. In vitro mutagenesis is used to
   a. produce large quantities of mutant proteins.
   b. create mutations at specific sites within a gene.
   c. create random mutations within multiple genes.
   d. create organisms that carry foreign genes.
8. Insertion of a gene for a surface protein from a medically important virus such as herpes into a harmless virus is an example of
   a. a DNA vaccine.
   b. reverse genetics.
   c. gene therapy.
   d. a subunit vaccine.
9. What is a Ti plasmid?
   a. A vector that can transfer recombinant genes into plant genomes
   b. A vector that can be used to produce recombinant proteins in yeast
   c. A vector that is specific to cereal plants like rice and corn
   d. A vector that is specific to embryonic stem cells

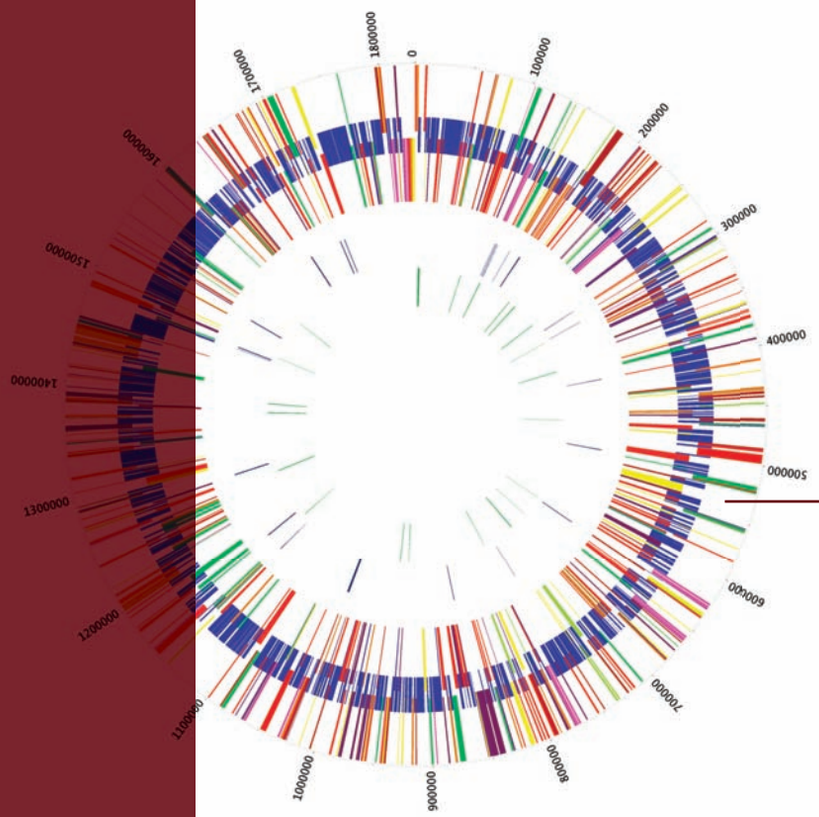
## APPLY

1. How is the gene for β-galactosidase used in the construction of a plasmid?
   a. The gene is a promoter that is sensitive to the presence of the sugar, galactose.
   b. It is an origin of replication.
   c. It is a cloning site.
   d. It is a marker for insertion of DNA.
2. Which of the following statements is accurate for DNA replication in your cells, but not PCR?
   a. DNA primers are required.
   b. DNA polymerase is stable at high temperatures.
   c. Ligase is essential.
   d. dNTPs are necessary.
3. What potential problems must be considered in creating a transgenic bacterium with the human insulin gene to produce insulin?
   a. Introns in the human gene will not be processed after transcription.
   b. The bacterial cell will be unable to post-translationally process the insulin peptide sequence.
   c. There is no way to get the bacterium to transcribe high levels of a human gene.
   d. Both a and b present problems.

## SYNTHESIZE

1. Many human proteins, such as hemoglobin, are only functional as an assembly of multiple subunits. Assembly of these functional units occurs within the endoplasmic reticulum and Golgi apparatus of a eukaryotic cell. Discuss what limitations, if any exist to the large-scale production of genetically engineered hemoglobin.
2. Enzymatic sequencing of a short strand of DNA was completed using dideoxynucleotides. Use the gel shown to determine the sequence of that DNA.






Chapter 18

# Genomics

## Chapter Outline



## Introduction

*The pace of discovery in biology in the last 30 years has been like the exponential growth of a population. Starting with the isolation of the first genes in the mid-1970s, researchers had accomplished the first complete genome sequence by the mid-1990s—that of the bacterial species* Haemophilus influenzae, *shown in the picture (genes with similar functions are shown in the same color). By the turn of the 21st century, the molecular biology community had completed a draft sequence of the human genome. Put another way, scientific accomplishments moved from cloning a single gene, to determining the sequence of a million base pairs in 20 years, then determining the sequence of a billion base pairs in another 5 years, and now sequencing 20 billion base pairs at one time. In the previous chapter you learned about the basic techniques of molecular biology. In this chapter you will see how those techniques have been applied to the analysis of whole genomes. This analysis integrates ideas from classical and molecular genetics with biotechnology, scaled and applied to whole genomes.*

## 18.1 Mapping Genomes

### Learning Outcomes

1. *Distinguish between a genetic map and a physical map.*
2. *Explain how genetic and physical maps can be linked.*

We use maps to find our location, and depending on how accurately we wish to do this, we may use multiple maps with different resolutions. In genomics, we can locate a gene on a chromosome, in a subregion of a chromosome, and finally its precise location in the chromosome's DNA sequence. The DNA sequence level requires knowing the entire sequence of the genome, something that was once out of our reach technologically. Knowing the entire sequence is useless, however,

without other kinds of maps; finding a single gene within the sequence of the human genome is like trying to find your house on a map of the world.

To overcome this difficulty, maps of genomes are constructed at different levels of resolution and using different kinds of information. We can distinguish between *genetic maps* and *physical maps*. **Genetic maps** are abstract maps that place the relative location of genes on chromosomes based on recombination frequency (see chapter 13). **Physical maps** use landmarks within DNA sequences, ranging from restriction sites (described in the preceding chapter) to the ultimate level of detail: the actual DNA sequence.

## Different kinds of physical maps can be generated

To make sense of genome mapping, it is important to have physical landmarks on the genome that are at a lower level of resolution than the entire sequence. In fact, long before the Human Genome Project was even conceived, physical maps of DNA were needed as landmarks on cloned DNA. Two types of physical maps are (1) restriction maps, constructed using restriction enzymes and (2) chromosome-banding patterns, generated by cytological dye methods.

### *Restriction maps*

Distances between "landmarks" on a physical map are measured in base-pairs (1000 base-pairs [bp] equal 1 kilobase, kb). It is not necessary to know the DNA sequence of a segment of DNA in order to create a physical map, or to know whether the DNA encompasses information for a specific gene.

The first physical maps were created by cutting genomic DNA with different restriction enzymes, both singly and with combinations of enzymes (figure 18.1). The analysis of the patterns of fragments generated were used to generate a map.

In terms of larger pieces of DNA, this process is repeated and then used to put the pieces back together, based on size and overlap, into a contiguous segment of the genome, called a **contig.** Coincidently, the very first restriction enzymes to be isolated came from *Haemophilus*, which was also the first free-living genome to be completely sequenced.

### *Chromosome-banding patterns*

Cytologists studying chromosomes with light microscopes found that by using different stains, they could produce reproducible patterns of bands on the chromosomes. In this way, they could identify all of the chromosomes and divide them into subregions based on banding pattern.

The use of different stains allows for the construction of a cytological map of the entire genome. These large-scale physical maps are like a map of an entire country, in that they encompass the whole genome, but at low resolution.

Cytological maps are used to characterize chromosomal abnormalities associated with human diseases, such as chronic myelogenous leukemia. In this disease, a reciprocal translocation occurs between chromosome 9 and chromosome 22 (figure 18.2*a*), resulting in an altered form of tyrosine kinase that is always turned on, causing white blood cell proliferation.

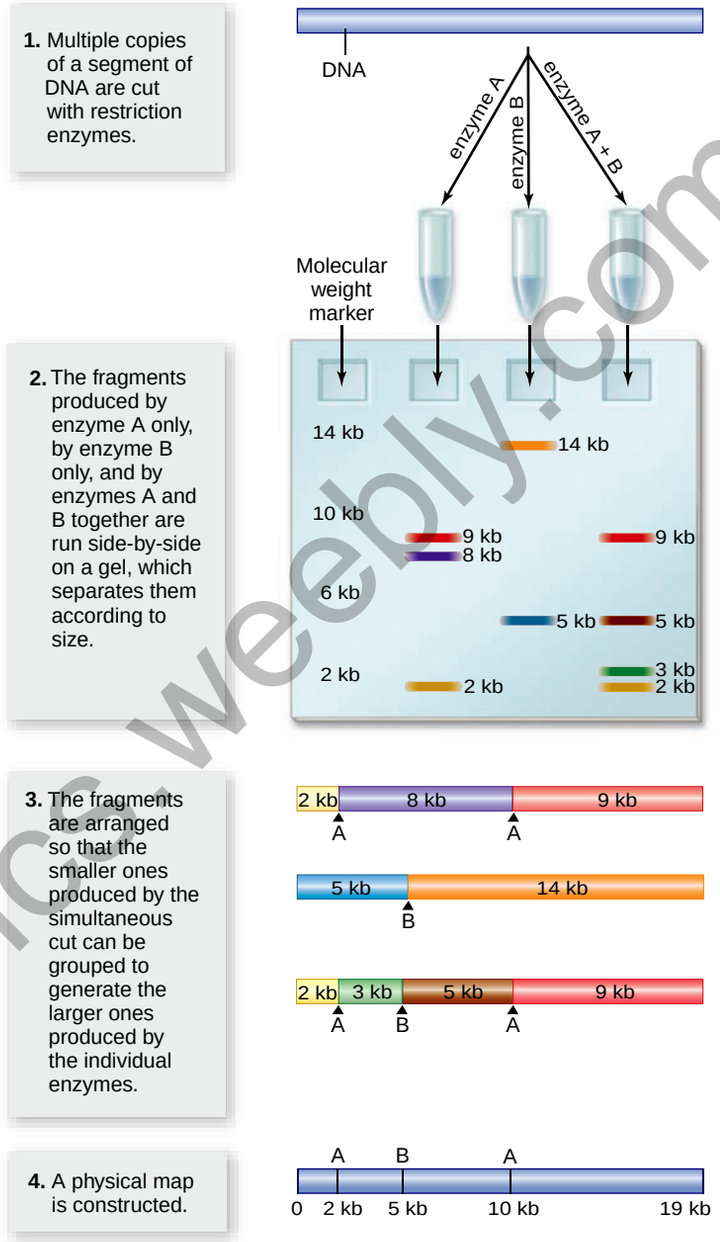
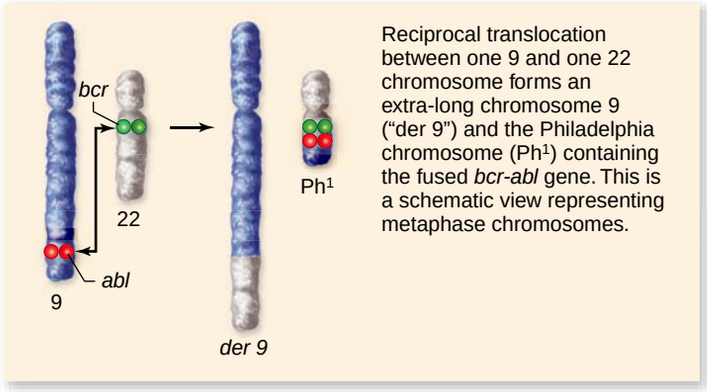
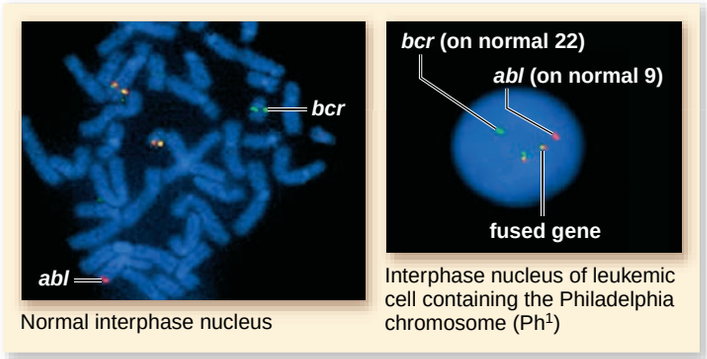



**Figure 18.1 Restriction enzymes can be used to create a physical map.** DNA is digested with two different restriction enzymes singly and in combination, then electrophoresed to separate the fragments. The location of sites can be deduced by comparing the sizes of fragments from the individual reactions with the combined reaction.

The use of hybridization with cloned DNA has added to the utility of chromosome-banding analysis. In this case, because the hybridization involves whole chromosomes, it is called *in situ hybridization*. It is done using fluorescently labeled probes, and so its complete name is **fluorescence in situ hybridization (FISH)** (figure 18.2*b*).




*a.*




*b.*

**Figure 18.2 Use of fluorescence in situ hybridization to correlate cloned DNA with cytological maps.** ***a.*** Karyotype of human chromosomes showing the translocation between chromosomes 9 and 22. ***b.*** FISH using a *bcr (green)* and *abl (red)* probe. The yellow color indicates the fused genes (*red* plus *green* fluorescence combined). The *abl* gene and the fused *bcr-abl* gene both encode a tyrosine kinase, but the fused gene is always expressed.

**Inquiry question**

**Why are there only three colored spots on the karyotype for two different genes?**

## Sequence-tagged sites provide a common language for physical maps

The construction of a physical map for a large genome requires the efforts of many laboratories in different locations. A variety of difficulties arose in comparing data from different labs, as well as integrating different types of landmarks used on physical and genetic maps.

In the early days of the Human Genome Project, this problem was addressed by the creation of a common molecular language that could be used to describe the different types of landmarks.

### *Defining common markers*

Since all genetic information is ultimately based on DNA sequence, it was important for this common language to be sequence-based, but not to require generating a large amount of sequence for any landmark. The solution was the **sequence-tagged site,** or **STS.** This site is a small stretch of DNA that is unique in the genome, that is, it only occurs once.

The boundary of the STS is defined by PCR primers, so the presence of the STS can be identified by PCR using any DNA as a template (see chapter 17). These sites need to be only 200–500 bp long, an amount of sequence that can be determined easily. The STS can contain any other kind of landmark—for example, part of a cloned gene that has been genetically mapped, or a restriction site that is polymorphic. Any marker that has been mapped can be converted to an STS by sequencing only 200–500 bp.

### *The use of STSs*

As maps are generated, new STSs are identified and added to a database, that indicates the sequence of the STS, its location in the genome, and the PCR primers needed to identify it. Any researcher is then able to identify the presence or absence of any STS in the DNA that he or she is analyzing.

Fragments of DNA can be pieced together using STSs by identifying overlapping regions in fragments. Because of the high density of STSs in the human genome and the relative ease of identifying an STS in a DNA clone, investigators were able to develop physical maps on the huge scale of the 3.2-gigabase genome in the mid-1990s (figure 18.3). STSs provide a scaffold for assembling genome sequences.

## Genetic maps provide a link to phenotypes

The first genetic (linkage) map was made in 1911 when Alfred Sturtevant mapped five genes in *Drosophila.* Distances on a genetic map reflect the frequency of recombination between genes and are measured in centimorgans (cM) in honor of the geneticist Thomas Hunt Morgan. One centimorgan corresponds to 1% recombination frequency between two loci. Over 14,000 genes have been mapped on the *Drosophila* genome.

Linkage mapping can be done without knowing the DNA sequence of a gene, as described in chapter 13. Computer programs make it possible to create a linkage map for a thousand genes at a time. But a few limitations to genetic maps still exist. One is that distances between genes determined by recombination frequencies do not directly correspond to physical distance on a chromosome. The conformation of DNA between genes varies, and this conformation can affect the frequency of recombination. Another limitation is that not all genes have obvious phenotypes that can be followed in segregating crosses.

The human genetic map is quite dense, with a marker roughly every 1 cM. This level of detail would have been unheard of 20 years ago, and it was made possible by development of molecular markers that do not cause a phenotype change.

The most common type of markers are short repeated sequences, called short tandem repeats, or STR loci, that differ in repeat length between individuals. These repeats are identified by using PCR to amplify the region containing the

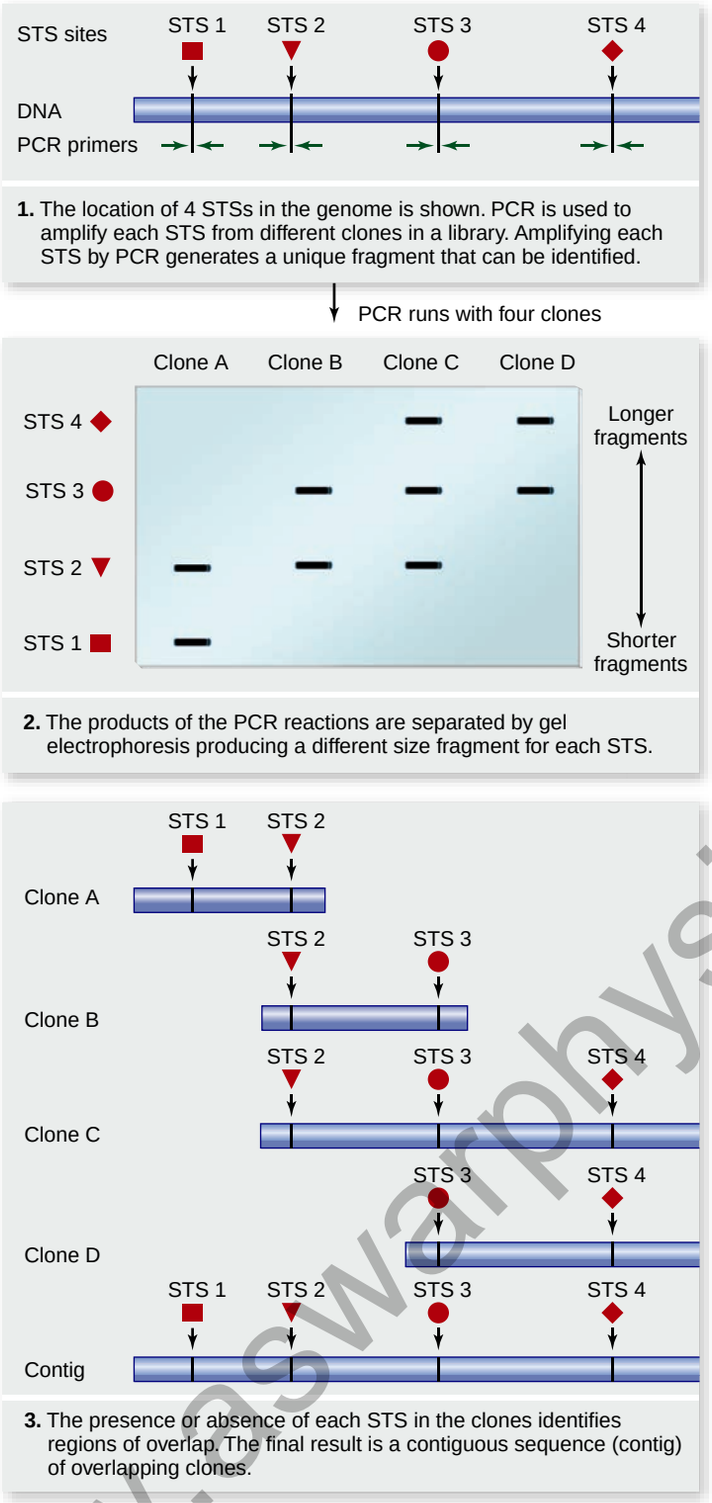
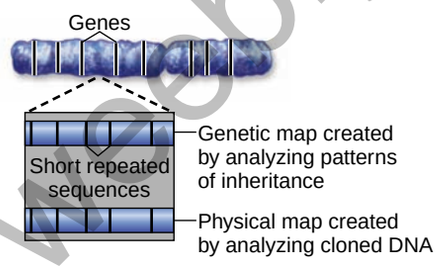



**Figure 18.3 Creating a physical map with sequence-tagged sites.** The presence of landmarks called sequence-tagged sites, or STSs, in the human genome made it possible to begin creating a physical map large enough in scale to provide a foundation for sequencing the entire genome. *(1)* Primers *(green arrows)* that recognize unique STSs are added to cloned DNA, followed by DNA amplification via polymerase chain reaction (PCR). *(2)* PCR products are separated based on size on a DNA gel, and the STSs contained in each clone are identified. *(3)* Cloned DNA segments are aligned based on STSs to create a contig.

repeat, then analyzing the products using electrophoresis. Once a map is constructed using these markers, genes with alleles that cause a disease state can be mapped relative to the molecular landmarks. Thirteen of these STR loci form the basis for modern DNA fingerprinting developed by the FBI. The alleles for these 13 loci are what is cataloged in the Combined DNA Index System (CODIS) database used to identify criminal offenders.

## Physical maps can be correlated with genetic maps

We need to be able to correlate genetic maps with physical maps, particularly genome sequences, to aid in finding physical sequences for genes that have been mapped genetically.




The problem in finding genes is that the resolution of genetic maps at present is not nearly as fine-grained as the genome sequence. Markers that are 1 cM apart may be as much as a million base pairs apart.

Since the markers used to construct genetic maps are now primarily molecular markers, they can be easily located within a genome sequence. Similarly, any gene that has been cloned can be placed within the genome sequence and can also be mapped genetically. This provides an automatic correlation between the two maps. The problem of finding genes that have been mapped genetically but not isolated as molecular clones lies in the nature of genetic maps. Distances measured on genetic maps are not uniform due to variation in recombination frequency along the chromosome. So 1 cM of genetic distance translates to different numbers of base-pairs in different regions.

Different kinds of maps are stored in databases so they can be aligned and viewed. The National Center for Biotechnology Information (NCBI) is a branch of the National Library of Medicine, and it serves as the U.S. repository for these data and more. Similar databases exist in Europe and Japan, and all are kept current. An enormous storehouse of information is available for use by biological researchers worldwide.

### *Learning Outcomes Review 18.1*

**Maps of genomes can by either physical maps or genetic maps. Physical maps include cytogenic maps of chromosome banding or restriction maps. Genetic maps are correlated with physical maps by using DNA markers such as sequence-tagged sites (STSs) unique to each genome.**

- ***What accounts for the difference between the proximity of banding sites on a karyotype and the number of base-pairs separating the two sites?***

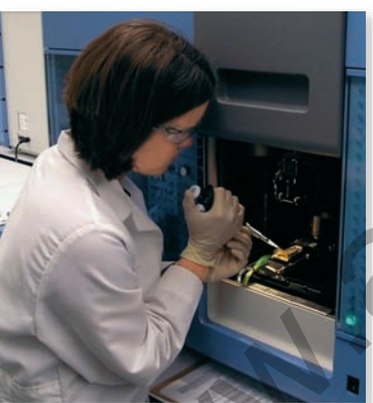
## 18.2 Whole-Genome Sequencing

### Learning Outcomes

1. *Characterize the main hurdle to sequencing an entire genome and how it has been overcome.*
2. *Differentiate between clone-by-clone sequencing and shotgun sequencing.*

The ultimate physical map is the base-pair sequence of an entire genome. In the early days of molecular biology, all sequencing was done manually, and was therefore both time- and labor-intensive. As mentioned in chapter 17, the development of machines to automate this process increased the rate of sequence generation.

*a.*



*b.*


*c.*

**Figure 18.4 Automated sequencing.** *a.* This Sanger sequence facility runs multiple automated sequencers, each processing 96 samples at a time. *b.* The development of new sequencing technologies permit sequencing that is orders of magnitude faster and that can be done in a very small space. *c.* Over 20 billion different DNA segments can be sequenced simultaneously in a flow cell the size of a microscope slide.

Large-scale genome sequencing requires the use of high-throughput automated sequencing and computer analysis (figure 18.4). Genome sequencing is one case in which technology drove the science, rather than the other way around. In a few hours, an automated Sanger sequencer can sequence the same number of base-pairs that a technician could manually sequence in a year—up to 50,000 bp. With the current generation of sequencing technology described in the previous chapter, the rate of sequence generation is now five orders of magnitude greater than when the human genome was sequenced with automated Sanger sequencers. Without the automation of sequencing, it would have been impossible to sequence large, eukaryotic genomes like that of humans.

### Genome sequencing requires larger molecular clones

Although it would be ideal to isolate DNA from an organism, add it to a sequencer, and then come back in a week or two to pick up a computer-generated printout of the genome sequence, the process is not quite that simple. Sequencers provide accurate sequences for DNA segments up to 800 bp long. Even then, errors are possible. So, to reduce errors, each clone is sequenced 5–10 times.

Even with reliable sequence data in hand, each individual sequencing run produces a relatively small amount of sequence. Thus, the genome must be fragmented, and then individual molecular clones isolated for sequencing (see chapter 17).

#### *Artificial chromosomes*

As described in chapter 17, the development of artificial chromosomes has allowed scientists to clone larger pieces of DNA. The first generation of these new vectors were yeast artificial chromosomes (YACs). These are constructed by using a yeast origin of replication and centromere sequence, then adding foreign DNA to it. The origin of replication allows the artificial chromosome to replicate independently of the rest of the genome, and the centromere sequences make the chromosome mitotically stable.

YACs were useful for cloning larger pieces of DNA but they had many drawbacks, including a tendency to rearrange, or to lose portions of DNA by deletion. Despite the difficulties, the YACs were used early on to construct physical maps by restriction enzyme digestion of the YAC DNA.

The artificial chromosomes most commonly used now, particularly for large-scale sequencing, are made in *E. coli*. These bacterial artificial chromosomes (BACs) are a logical extension of the use of bacterial plasmids. BAC vectors accept DNA inserts between 100 and 200 kb long. The downside of BAC vectors is that, like the bacterial chromosome, they are maintained as a single copy whereas plasmid vectors exist at high copy numbers.

#### *Human artificial chromosomes*

Human artificial chromosomes can introduce large segments of human DNA into cultured cells. These artificial chromosomes are usually constructed by fragmentation of chromosomes with centromere sequence. Although circular, some can still segregate correctly during mitosis up to 98% of the time. Construction of linear human artificial chromosomes is not yet possible.

## Whole-genome sequencing is approached in two ways: clone-by-clone and shotgun

Sequencing an entire genome is an enormous task. Two ways of approaching this challenge have been developed: one that approaches the sequencing one step at a time, and another that attempts to take on the whole thing at once and depends on computers to sort out the data. The two techniques grew out of competing projects to sequence the human genome.

### *Clone-by-clone sequencing*

The cloning of large inserts in BACs facilitates the analysis of entire genomes. The strategy most commonly pursued is to construct a physical map first, and then use it to place the site of BAC clones for later sequencing.

Aligning large portions of a chromosome requires identifying regions that overlap between clones. This can be accomplished either by constructing restriction maps of each BAC clone, or by identifying STSs found in clones. If two BAC clones have the same STS, then they must overlap.

The alignment of a number of BAC clones results in a contiguous stretch of DNA called a *contig*. The individual BAC clones can then be sequenced 500 bp at a time to produce the sequence of the entire contig (figure 18.5*a*). This strategy of physical mapping followed by sequencing is called **clone-by-clone sequencing.**

### *Shotgun sequencing*

The idea of **shotgun sequencing** is simply to randomly cut the DNA into small fragments, sequence all cloned fragments, and then use a computer to put together the overlaps (figure 18.5*b*). This terminology actually goes back to the early days of molecular cloning when the construction of a library of randomly cloned fragments was referred to as *shotgun cloning*. This approach is much less labor-intensive than the clone-by-clone method, but it requires much greater computer power to assemble the final sequence and very efficient algorithms to find overlaps.

Unlike the clone-by-clone approach, shotgun sequencing does not tie the sequence to any other information about the genome (figure 18.5*b*). Many investigators have used both clone-by-clone and shotgun-sequencing techniques, and such hybrid approaches are becoming the norm. This combination has the strength of tying the sequence to a physical map while greatly reducing the time involved.

Assembler programs compare multiple copies of sequenced regions in order to assemble a **consensus sequence,** that is, a sequence that is consistent across all copies. Although computer assemblers are incredibly powerful, final human analysis is required after both clone-by-clone and shotgun sequencing to determine when a genome sequence is sufficiently accurate to be useful to researchers.

## The Human Genome Project used both sequencing methods

The vast scale of genomics ushered in a new way of doing biological research involving large teams. Although a single individual can isolate and manually sequence a molecular clone for a single gene, a huge genome like the human genome requires the collaborative efforts of hundreds of researchers.

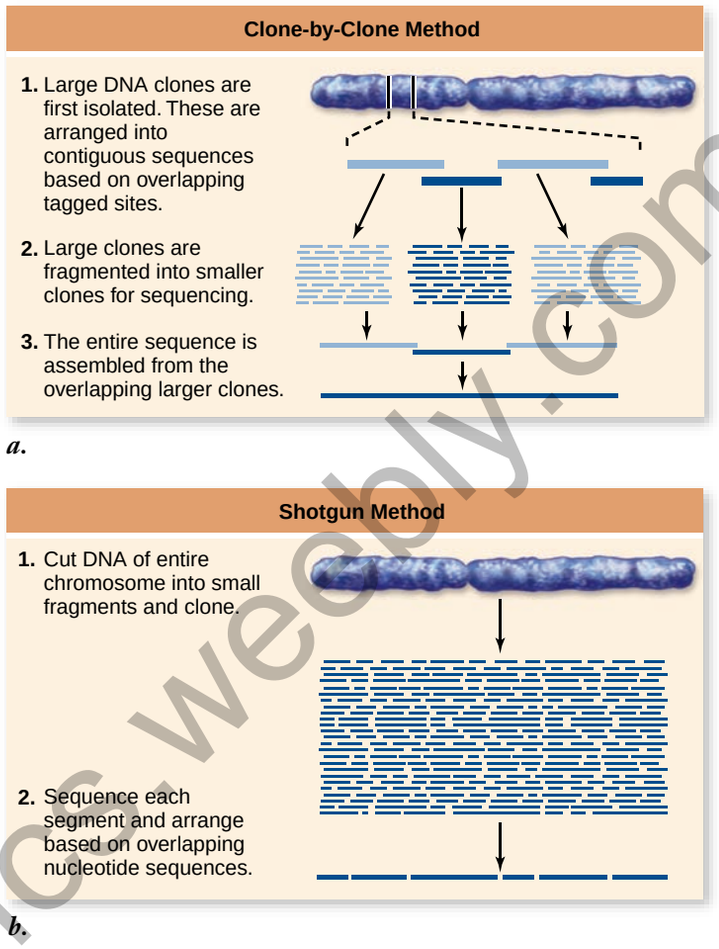



**Figure 18.5 Comparison of sequencing methods.** ***a.*** The clone-by-clone method uses large clones assembled into overlapping regions by STSs. Once assembled, these can be fragmented into smaller clones for sequencing. ***b.*** In the shotgun method the entire genome is fragmented into small clones and sequenced. Computer algorithms assemble the final DNA sequence based on overlapping nucleotide sequences.

The Human Genome Project originated in 1990 when a group of American scientists formed the International Human Genome Sequencing Consortium. The goal of this publicly funded effort was to use a clone-by-clone approach to sequence the human genome. Genetic and physical maps were used as scaffolding to sequence each chromosome.

In May, 1998, Craig Venter, whose research group had sequenced *Haemophilus influenzae*, announced his private company (Celera Genomics) would sequence the human genome. He proposed to shotgun-sequence the 3.2-gigabase genome in only two years. The Consortium rose to the challenge, and the race to sequence the human genome began. The upshot was a tie of sorts. On June 26, 2000, the groups jointly announced success, and each published its findings simultaneously in 2001. The Consortium's draft alone included 248 authors.

The draft sequence of the human genome was just the beginning. Gaps in the sequence are still being filled, and the

map is constantly being refined. The most recent "finished" human sequence is down to only 260 gaps, a 400-fold reduction in gaps, and it now includes 99% of the euchromatic sequence, up from 95%. The reference sequence has an error rate of 1 per 100,000 bases. Newer sequencing technologies are being used to close the remaining gaps. A few individuals, including James Watson who codiscovered the structure of DNA, have now had their personal genomes sequenced. The cost for having one's genome sequenced is predicted to fall to $1000 in the next few years, raising many questions about genome privacy.

Research on the whole genome can move ahead. Now that the ultimate physical map is in place and is being integrated with the genetic map, diseases that result from changes in more than one gene, such as diabetes, can be addressed. Comparisons with other genomes are already changing our understanding of genome evolution (see chapter 24).

### *Learning Outcomes Review 18.2*

**Because of the enormous size of genomes, sequencing requires the use of automated sequencers running many samples in parallel. Two approaches have been developed for whole-genome sequencing: one that uses clones already aligned by physical mapping (clone-by-clone sequencing), and one that involves sequencing random clones and using a computer to assemble the final sequence (shotgun sequencing). In either case, significant computing power is necessary to assemble a final sequence.**

- ***Why would data from a single copy of a genome likely be unreliable?***

## 18.3 *Characterizing Genomes*

### *Learning Outcomes*

1. ***Describe the classes of DNA found in a genome.***
2. ***Explain what an SNP is and why SNPs are helpful in characterizing genomes.***

Automated sequencing technology has produced huge amounts of sequence data, eventually sequencing entire genomes. This has allowed researchers studying complex problems to move beyond approaches restricted to the analysis of individual genes. Sequencing projects in themselves are descriptive analyses that tells us nothing about the organization of genomes, let alone the function of gene products and how they may be interrelated. Additional research and evaluation has given us both answers and new puzzles.

### The Human Genome Project found fewer genes than expected

For many years, geneticists had estimated the number of human genes to be around 100,000. This estimate, although based on some data, was really just a guess. Imagine researchers' surprise when the number turned out to be only around 25,000! This represents only about twice as many genes as *Drosophila* and fewer genes than rice (figure 18.6). Clearly the complexity of an organism is not a simple function of the number of genes in its genome.

### Finding genes in sequence data requires computer searches

Once a genome has been sequenced, the next step is to determine which regions of the genome contain which genes, and what those genes do. A lot of information can be mined from the sequence data. Using markers from physical maps and information from genetic maps, it is possible to find the sequence of the small percentage of genes that are identified by mutations with an observable (phenotypic) effect. Genes can also be found by comparing expressed sequences to genomic sequences. The analysis of expressed sequences is discussed later in this section.

#### *Locating starts and stops*

Information in the nucleotide sequence itself can also be used in the search for genes. A protein-coding gene begins with a start codon, such as ATG, and it contains no stop

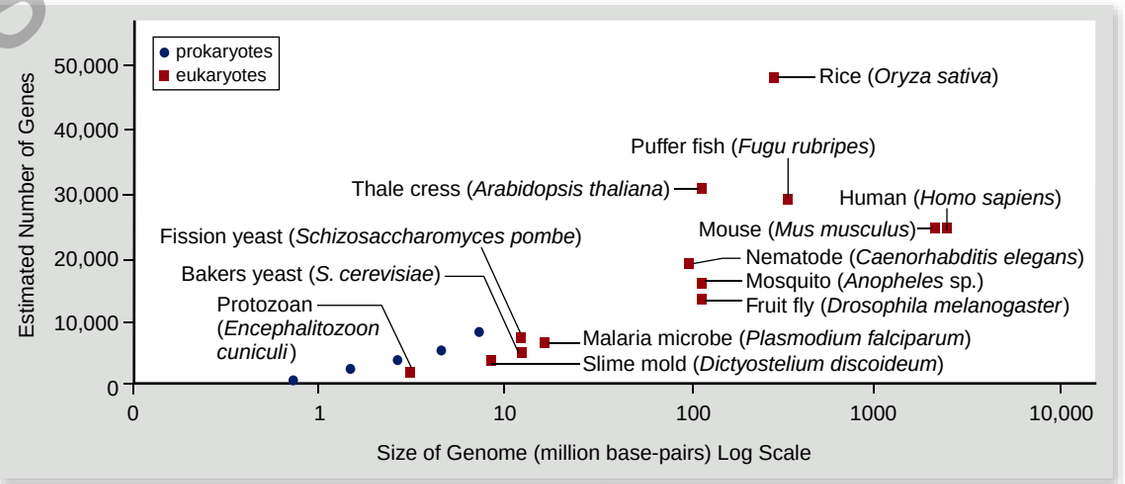



**Figure 18.6 Size and complexity of genomes.** In general, eukaryotic genomes are larger and have more genes than prokaryotic genomes, although the size of the organism is not the determining factor. The mouse genome is nearly as large as the human genome, and the rice genome contains more genes than the human genome.

codons (TAA, TGA, or TAG) for a distance long enough to encode a protein. This coding region is referred to as an **open reading frame (ORF).** Although ORFs are likely to be genes, they may or may not actually be translated into a functional protein. Among putative genes, families of genes can be identified based on common domains. For example, genes in the HOX family have a conserved, 180-bp sequence called the homeobox, which encodes the homeodomain region of certain transcription factors. Sequences for potential genes need to be tested experimentally to determine whether they have a function.

The addition of information to the basic sequence information, like identifying ORFs, is called sequence **annotation.** This process is what converts simple sequence data into something that we can recognize based on landmarks such as regions that are transcribed and regions that are known or thought to encode proteins.

### *Inferring function across species: the BLAST algorithm*

It is also possible to search genome databases for sequences that are homologous to known genes in other species. A researcher who has isolated a molecular clone for a gene of unknown function can search the database for similar sequences to infer function. The tool that makes this possible is a search algorithm called BLAST (which stands for Basic Local Alignment Search Tool). Using a networked computer, one can submit a sequence to the BLAST server and get back a reply with all possible similar sequences contained in the sequence database.

Using these techniques, sequences that are not part of ORFs have been identified that have been conserved over millions of years of evolution. These sequences may be important for the regulation of the genes contained in the genome.

Using computer programs to search for genes, to compare genomes, and to assemble genomes are only a few of the new genomics approaches falling under the heading of **bioinformatics.**

## Genomes contain both coding and noncoding DNA

When genome sequences are analyzed, regions that encode proteins and other regions that do not encode proteins are revealed. For many years investigators had known of the latter, but they did not know the extent and nature of the noncoding DNA. We first consider the types of coding DNA that have been found, then move on to look at types of noncoding DNA.

### *Protein-encoding DNA in eukaryotes*

Four different classes of protein-encoding genes are found in eukaryotic genomes, differing largely in gene copy number.

**Single-copy genes.** Many genes exist as single copies on a particular chromosome. Most mutations in these genes result in recessive Mendelian inheritance.

**Segmental duplications.** Sometimes whole blocks of genes are copied from one chromosome to another, resulting in *segmental duplication.* Blocks of similar genes in the same order are found throughout the human genome. Chromosome 19 seems to have been the biggest borrower, sharing blocks of genes with 16 other chromosomes.

**Multigene families.** As more has been learned about eukaryotic genomes, many genes have been found to exist as parts of *multigene families,* groups of related but distinctly different genes that often occur together in clusters. These genes appear to have arisen from a single ancestral gene that duplicated during an uneven meiotic crossover in which genes were added to one chromosome and subtracted from the other. These multigene families may include silent copies called *pseudogenes,* which are inactivated by mutation.

**Tandem clusters.** Identical copies of genes can also be found in *tandem clusters.* These genes are transcribed simultaneously, increasing the amount of mRNA available for protein production. Tandem clusters also include genes that do not encode proteins, such as clusters of rRNA genes.

### *Noncoding DNA in eukaryotes*

One of the most notable characteristics is the amount of noncoding DNA they possess. The Human Genome Project has revealed a particularly startling picture. Each of your cells has about 6 feet of DNA stuffed into it, but of that, less than 1 inch is devoted to genes! Nearly 99% of the DNA in your cells is non-protein coding DNA.

True genes are scattered about the human genome in clumps among the much larger amount of noncoding DNA, like isolated oases in a desert. Seven major sorts of noncoding human DNA have been described. (Table 18.1 shows the composition of the human genome, including noncoding DNA.)

**Noncoding DNA within genes.** As discussed in chapter 15, a human gene is not simply a stretch of DNA, like the letters of a word. Instead, a human gene is made up of numerous fragments of protein-encoding information (exons) embedded within a much larger matrix of noncoding DNA (introns). Together, introns make up about 24% of the human genome and exons less than 1.5%.

**Structural DNA.** Some regions of the chromosomes remain highly condensed, tightly coiled, and untranscribed throughout the cell cycle. Called *constitutive heterochromatin,* these portions tend to be localized around the centromere or located near the ends of the chromosome, at the telomeres.

**Simple sequence repeats.** Scattered about chromosomes are **simple sequence repeats (SSRs).** An SSR is a 1- to 6-nt sequence such as CA or CGG, repeated like a broken record thousands and thousands of times. SSRs can arise from DNA replication errors. SSRs make up about 3% of the human genome.

**Segmental duplications.** Blocks of genomic sequences composed of from 10,000 to 300,000 bp have duplicated and moved either within a chromosome or to a nonhomologous chromosome.

**TABLE 18.1** Classes of DNA Sequences Found in the Human Genome

| Class | Estimated Frequency (%) | Description |
|---|---|---|
| Protein-encoding genes | 1.5 | Translated portions of the 25,000 genes scattered about the chromosomes |
| Introns | 24 | Noncoding DNA that constitutes the great majority of each human gene |
| Segmental duplications | 5 | Regions of the genome that have been duplicated |
| Pseudogenes (inactive genes) | 2 | Sequence that has characteristics of a gene but is not a functional gene |
| Structural DNA | 20 | Constitutive heterochromatin, localized near centromeres and telomeres |
| Simple sequence repeats | 3 | Stuttering repeats of a few nucleotides such as CGG, repeated thousands of times |
| Transposable elements | 45 | 21%: Long interspersed elements (LINEs), which are active transposons<br>13%: Short interspersed elements (SINEs), which are active transposons<br>8%: Retrotransposons, which contain long terminal repeats (LTRs) at each end<br>3%: DNA transposon fossils |
| microRNA | 0.03 | Code for RNAs complementary to one or more mature mRNAs |

**Pseudogenes.** These are inactive genes that may have lost function because of mutation.

**Transposable elements.** Fully 45% of the human genome consists of mobile bits of DNA called *transposable elements*. Some of these elements code for proteins, but many do not. Because of the significance of these elements, we describe them more fully in the following section.

**microRNA genes.** Hidden within the nonprotein-coding DNA lies an extraordinary mechanism for controlling gene expression and development. Compact regulatory RNAs have a much larger role in directing development in complex organisms than we imagined even a few years ago. Specifically, DNA that was once considered "junk" has been shown to encode microRNAs, or miRNAs, which are processed after transcription to lengths of 21 to 23 nt, but never translated. About 10,000 unique miRNAs have been identified that are complementary to one or more mature mRNAs.

### *Transposable elements: mobile DNA*

Discovered by Barbara McClintock in 1950, **transposable elements,** also termed *transposons* and *mobile genetic elements*, are bits of DNA that are able to move from one location on a chromosome to another. Barbara McClintock received the 1983 Nobel Prize in physiology or medicine for discovery of these elements and their unexpected ability to change location.

Transposable elements move around in different ways. In some cases, the transposon is duplicated, and the duplicated DNA moves to a new place in the genome, so the number of copies of the transposon increases. Other types of transposons are excised without duplication and insert themselves elsewhere in the genome. The role of transposons in genome evolution is discussed in chapter 24.

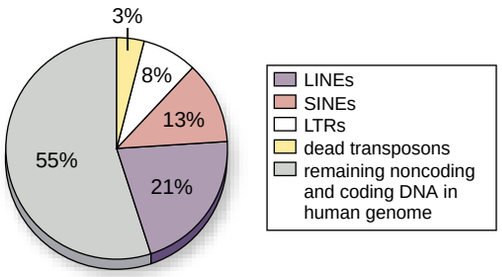
Human chromosomes contain four sorts of transposable elements. Fully 21% of the genome consists of **long interspersed elements (LINEs).** These ancient and very successful elements are about 6000 bp long, and they contain all the equipment needed for transposition. LINEs encode a reverse transcriptase enzyme that can make a cDNA copy of the transcribed LINE RNA. The result is a double-stranded segment that can reinsert into the genome rather than undergo translation into a protein. Since these elements use an RNA intermediate, they are termed *retrotransposons*.

**Short interspersed elements (SINEs)** are similar to LINEs, but they cannot transpose without using the transposition machinery of LINEs. Nested within the genome's LINEs are over half a million copies of a SINE element called Alu (named for a restriction enzyme that cuts within the sequence). The Alu SINE is 300 bp and represents 10% of the human genome. Like a flea on a dog, Alu moves with the LINE it resides within. Just as a flea sometimes jumps to a different dog, so Alu sometimes uses the enzymes of its LINE to move to a new chromosome location. Alu can also jump right into genes, causing harmful mutations.

Two other sorts of transposable elements are also found in the human genome: 8% of the human genome is devoted to retrotransposons called **long terminal repeats (LTRs).** Although the transposition mechanism is a bit different from

that of LINEs, LTRs also use reverse transcriptase to ensure that copies are double-stranded and can reintegrate into the genome.

Some 3% of the genome is devoted to dead transposons, elements that have lost the signals for replication and can no longer move.




**Inquiry question**

**How do you think these repetitive elements would affect the determination of gene order?**

## Expressed sequence tags identify genes that are transcribed

Given the complexity of coding and noncoding DNA, it is important to be able to recognize regions of the genome that are actually expressed—that is, transcribed and then translated.

Because DNA is easier to work with than protein, one approach is to isolate mRNA, use this to make cDNA, then sequence one or both ends of as many cDNAs as possible. With automated sequencing, this task is not difficult, and these short sections of cDNA have been named **expressed sequence tags (ESTs).** An EST is another form of STS, and thus it can be included in physical maps. This technique does not tell us anything about the function of any particular EST, but it does provide one view, at the whole-genome level, of what genes are expressed, at least as mRNAs.

ESTs have been used to identify 87,000 cDNAs in different human tissues. About 80% of these cDNAs were previously unknown. You may wonder at this point how the estimated 25,000 genes of the human genome can result in 87,000 different cDNAs. The answer lies in the modularity of eukaryotic genes, which consist of exons interspersed with introns, as described in chapter 15.

Following transcription in eukaryotes, the introns are removed, and exons are spliced together. In some cells, some of the splice sites are skipped, and one or more exons is removed along with the introns. This process, called *alternative splicing* (figure 18.7), yields different proteins that can have different functions. Thus, the added complexity of proteins in the human genome comes not from additional genes, but from new ways to put existing parts of genes together.

## SNPs are single-base differences between individuals

One fact becoming clear from analysis of the human genome is that a huge amount of genetic variation exists in our species. This information has practical use.

Single-nucleotide polymorphisms (SNPs) are sites where individuals differ by only a single nucleotide. To be classified as a polymorphism, an SNP must be present in at least 1% of the population. SNPs occur about every 100 to 300 bp in the 3 billion bp human genome. As of January 2009, 1.5 million nonredundant human SNPs had been identified, about 10% of the variation available. These SNPs are being used to look for associations between genes. We expect that the genetic recombination occurring during meiosis randomizes all but the most tightly linked genes. We call the tendency for genes *not* to be randomized **linkage disequilibrium.** This kind of association can be used to map genes.

The preliminary analysis of SNPs shows that many are in linkage disequilibrium. This unexpected result has led to the idea of genomic **haplotypes,** or regions of chromosomes that are not being exchanged by recombination. The existence of haplotypes allows the genetic characterization of genomic regions by describing a small number of SNPs (figure 18.8). If these haplotypes stand up to further analysis, they could greatly aid in mapping the genetic basis of disease. The Human Genome Project is now working on a haplotype map of the genome.

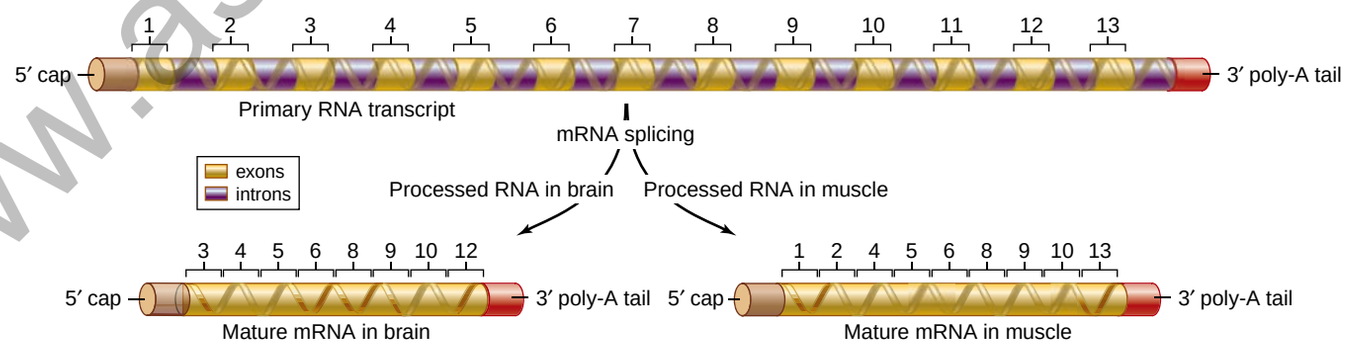



**Figure 18.7 Alternative splicing can result in the production of different mRNAs from the same coding sequence.** In some cells, exons can be excised along with neighboring introns, resulting in different proteins. Alternative splicing explains why 25,000 human genes can code for three to four times as many proteins.

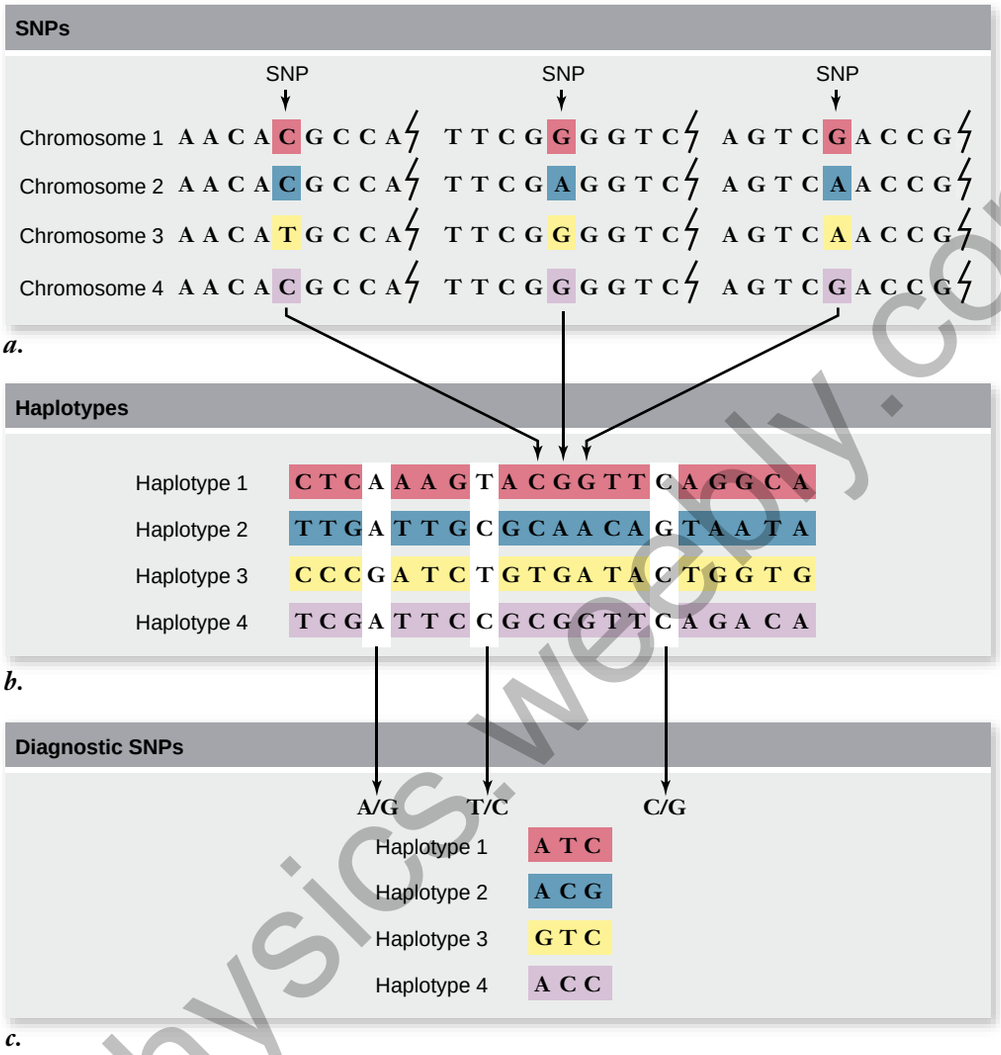



**Figure 18.8 Construction of a haplotype map.** Single-nucleotide polymorphisms (SNPs) are single-base differences between individuals. Sections of DNA sequences from four individuals are shown in ***(a)***, with three SNPs indicated by arrows. ***b.*** These three SNPs are shown aligned along with 17 other SNPs from this chromosomal region. This represents a haplotype map for this region of the chromosome. Haplotypes are regions of the genome that are not exchanged by recombination during meiosis. ***c.*** Haplotypes can be identified using a small number of diagnostic SNPs that differ between the different haplotypes. In this case, 3 SNPs out of the 20 in this region are all that are needed to uniquely identify each haplotype. This greatly facilitates locating disease-causing genes, as haplotypes represent large regions of the genome that behave as a single site during meiosis.

### *Learning Outcomes Review 18.3*

**Coding sequences in a genome can be found as a single copy, as repeated clusters, as part of segmental duplications, or as part of a gene family. A significant amount of noncoding DNA is found in all eukaryotic organisms. Transposable elements are capable of movement in the genome and are found in all eukaryotic genomes. Single-nucleotide polymorphisms (SNPs) provide a way of identifying individual variation, and they have also revealed cases of nonrandom recombination (genomic haplotypes).**

- ***What explanation could you suggest, based on principles of natural selection, for the many repeated transposable elements in the human genome?***

## 18.4 *Genomics and Proteomics*

### *Learning Outcomes*

1. ***Describe the advances that have come from comparative genomics.***
2. ***Distinguish between genomics and proteomics.***

To fully understand how genes work, we need to characterize the proteins they produce. This information is essential to understanding cell biology, physiology, development, and evolution. In many ways, we continue to ask the same questions that Mendel asked, but at a much different level of organization.

### Comparative genomics reveals conserved regions in genomes

With the large number of sequenced genomes, it is now possible to make comparisons at both the gene and genome level. The flood of information from different genomes has given rise to a new field: *comparative genomics*. One of the striking lessons learned from the sequence of the human genome is how very similar humans are to other organisms. More than half of the genes of *Drosophila* have human counterparts. Among mammals, the similarities are even greater. Humans have only 300 genes that have no counterpart in the mouse genome.

The use of comparative genomics to ask evolutionary questions is also a field of great promise. The comparison of the many prokaryotic genomes already indicates a greater degree of lateral gene transfer than was previously suspected. The latest round of animal genomes sequenced has included the chimpanzee, our closest living relative. The draft sequence of the

chimp *(Pan troglodytes)* genome has just been completed, and comparisons between the chimp and human genome may allow us to unravel what makes us uniquely human.

The early returns from this sequencing effort confirm that our genomes differ by only 1.23% in terms of nucleotide substitutions. At first glance, the largest difference between our genomes actually appears to be in transposable elements. In humans, the SINEs have been threefold more active than in the chimp, but the chimp has acquired two elements that are not found in the human genome. The differences due to insertion and deletion of bases are fewer than substitutions but account for about 1.5% of the euchromatic sequence being unique in each genome.

## Synteny allows comparison of unsequenced genomes

Similarities and differences between highly conserved genes can be investigated on a gene-by-gene basis between species. Genome science allows for a much larger scale approach to comparing genomes by taking advantage of synteny.

*Synteny* refers to the conserved arrangements of segments of DNA in related genomes. Physical mapping techniques can be used to look for synteny in genomes that have not been sequenced. Comparisons with the sequenced, syntenous segment in another species can be very helpful.

To illustrate this, consider rice, already sequenced, and its grain relatives maize, barley, and wheat, none of which have been fully sequenced. Even though these plants diverged more than 50 million years ago, the chromosomes of rice, corn, wheat, and other grass crops show extensive synteny (figure 18.9). In a genomic sense, "rice is wheat."

By understanding the rice genome at the level of its DNA sequence, identification and isolation of genes from grains with larger genomes should be much easier. DNA sequence analysis of cereal grains could be valuable for identifying genes associated with disease resistance, crop yield, nutritional quality, and growth capacity.

As mentioned earlier, the rice genome has more genes than the human genome. However, rice still has a much smaller genome than its grain relatives, which also represent a major food source for humans.

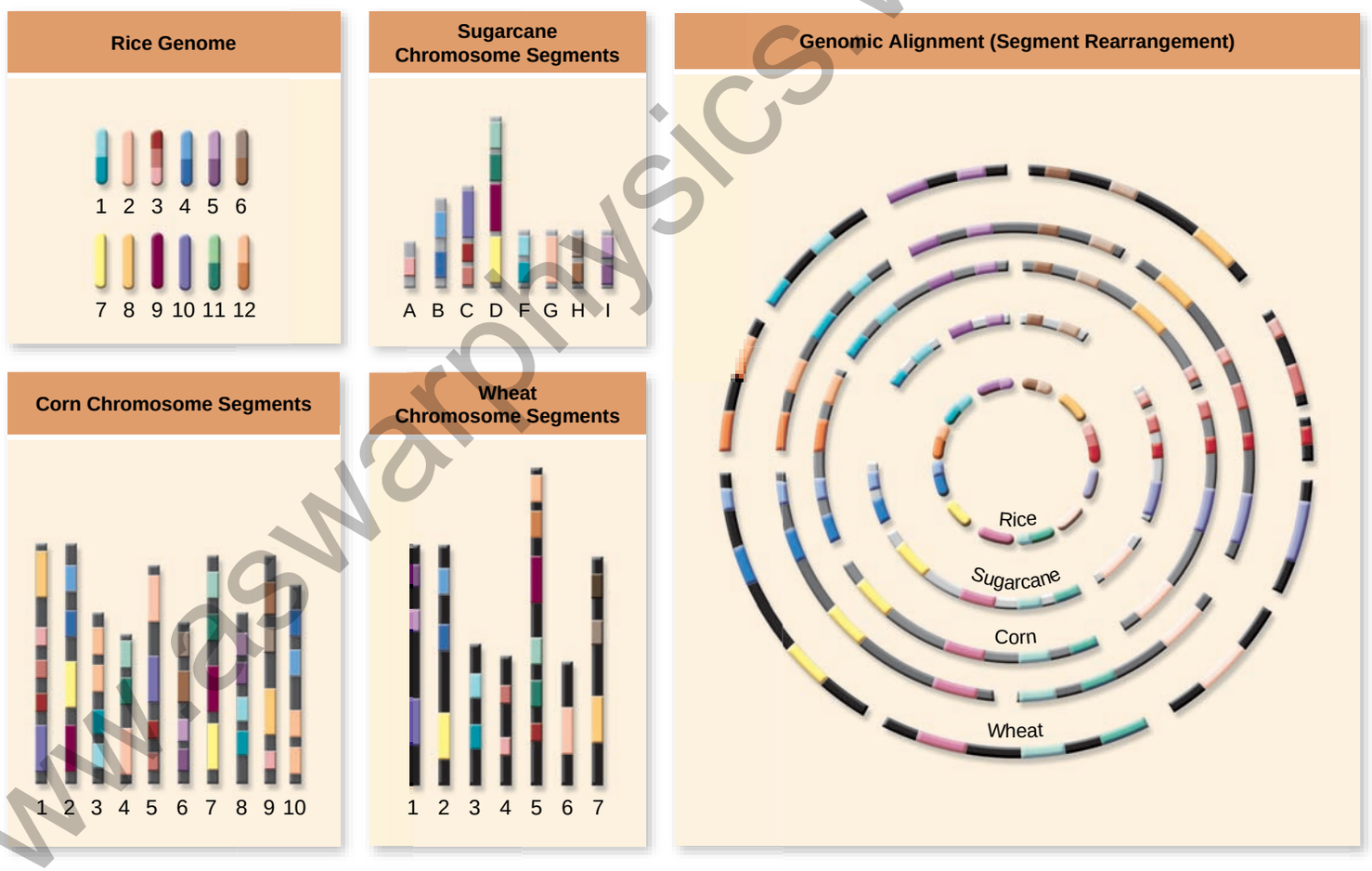



**Figure 18.9 Grain genomes are rearrangements of similar chromosome segments.** Shades of the same color represent pieces of DNA that are conserved among the different species but have been rearranged. By splitting the individual chromosomes of major grass species into segments and rearranging the segments, researchers have found that the genome components of rice, sugarcane, corn, and wheat are highly conserved. This implies that the order of the segments in the ancestral grass genome has been rearranged by recombination as the grasses have evolved.

## Organelle genomes have exchanged genes with the nuclear genome

Mitochondria and chloroplasts are considered to be descendants of ancient bacterial cells living in eukaryotes as a result of endosymbiosis (chapter 4). Their genomes have been sequenced in some species, and they are most like prokaryotic genomes. The chloroplast genome, having about 100 genes, is minute compared with the rice genome, with 32,000 to 55,000 genes.

### *The chloroplast genome*

The chloroplast, a plant organelle that functions in photosynthesis, can independently replicate in the plant cell because it has its own genome. The DNA in the chloroplasts of all land plants have about the same number of genes, and they are present in about the same order. In contrast to the evolution of the DNA in the plant cell nucleus, chloroplast DNA has evolved at a more conservative pace and therefore shows a more easily interpretable evolutionary pattern when scientists study DNA sequence similarities. Chloroplast DNA is also not subject to modification caused by transposable elements or to mutations due to recombination.

Over time, some genetic exchange appears to have occurred between the nuclear and chloroplast genomes. For example, Rubisco, the key enzyme in the Calvin cycle of photosynthesis (chapter 8), consists of large and small subunits. The small subunit is encoded in the nuclear genome. The protein it encodes has a targeting sequence that allows it to enter the chloroplast and combine with large subunits, which are coded for and produced by the chloroplast.

### *The mitochondrial genome*

Mitochondria are also constructed of components encoded by both the nuclear genome and the mitochondrial genome. For example, the electron transport chain (chapter 7) is made up of proteins that are encoded by both nuclear and mitochondrial genomes—and the pattern varies with different species. This observation implies a movement of genes from the mitochondria to the nuclear genome with some lineage-specific variation.

The evolutionary history of the localization of these genes is a puzzle. Comparative genomics and their evolutionary implications are explored in detail in chapter 24, after we have established the fundamentals of evolutionary theory.

## Functional genomics reveals gene function at the genome level

Bioinformatics takes advantage of high-end computer technology to analyze the growing gene databases, look for relationships among genomes, and then hypothesize functions of genes based on sequence. Genomics is now shifting gears and moving back to hypothesis-driven science, to **functional genomics,** the study of the function of genes and their products.

Like sequencing whole genomes, finding how these genomes work requires the efforts of a large team. For example, an international community of researchers has come together with a plan to assign function to all of the 20,000–25,000 *Arabidopsis* genes by 2010 (Project 2010). One of the first steps is to determine when and where these genes are expressed. Each step beyond that will require additional improvements in technology.

### *DNA microarrays*

The earlier description of ESTs indicated that we could locate sequences that are transcribed on our DNA maps—but this tells us nothing about when and where these genes are turned on. To be able to analyze gene expression at the whole-genome level requires a representation of the genome that can be manipulated experimentally. This has led to the creation of **DNA microarrays,** or "gene chips" (figure 18.10).

**Preparation of a microarray** To prepare a particular microarray, fragments of DNA are deposited on a microscope slide by a robot at indexed locations (i.e., an array). Silicon chips instead of slides can also be arrayed. These chips can then be used in hybridization experiments with labeled mRNA from different sources. This gives a high-level view of genes that are active and inactive in specific tissues.

Researchers are currently using a chip with 24,000 *Arabidopsis* genes on it to identify genes that are expressed developmentally in certain tissues or in response to environmental factors. RNA from these tissues can be isolated and used as a probe for these microarrays. Only those sequences that are expressed in the tissues will be present and will hybridize to the microarray.

**Microarray analysis and cancer.** One of the most exciting uses of microarrays has been the profiling of gene expression patterns in human cancers. Microarray analysis has revealed that different cancers have different gene expression patterns. These findings are already being used to diagnose and design specific treatments for particular cancers.

From a large body of data, several patterns emerge:

1. Specific cancer types can be reliably distinguished from other cancer types and from normal tissue based on microarray data.
2. Subtypes of particular cancers often have different gene expression patterns in microarray data.
3. Gene expression patterns from microarray data can be used to predict disease recurrence, tendency to metastasize, and treatment response.

This represents an important step forward in both the diagnosis and treatment of human cancers.

### *Microarray analysis and genome-wide association mapping*

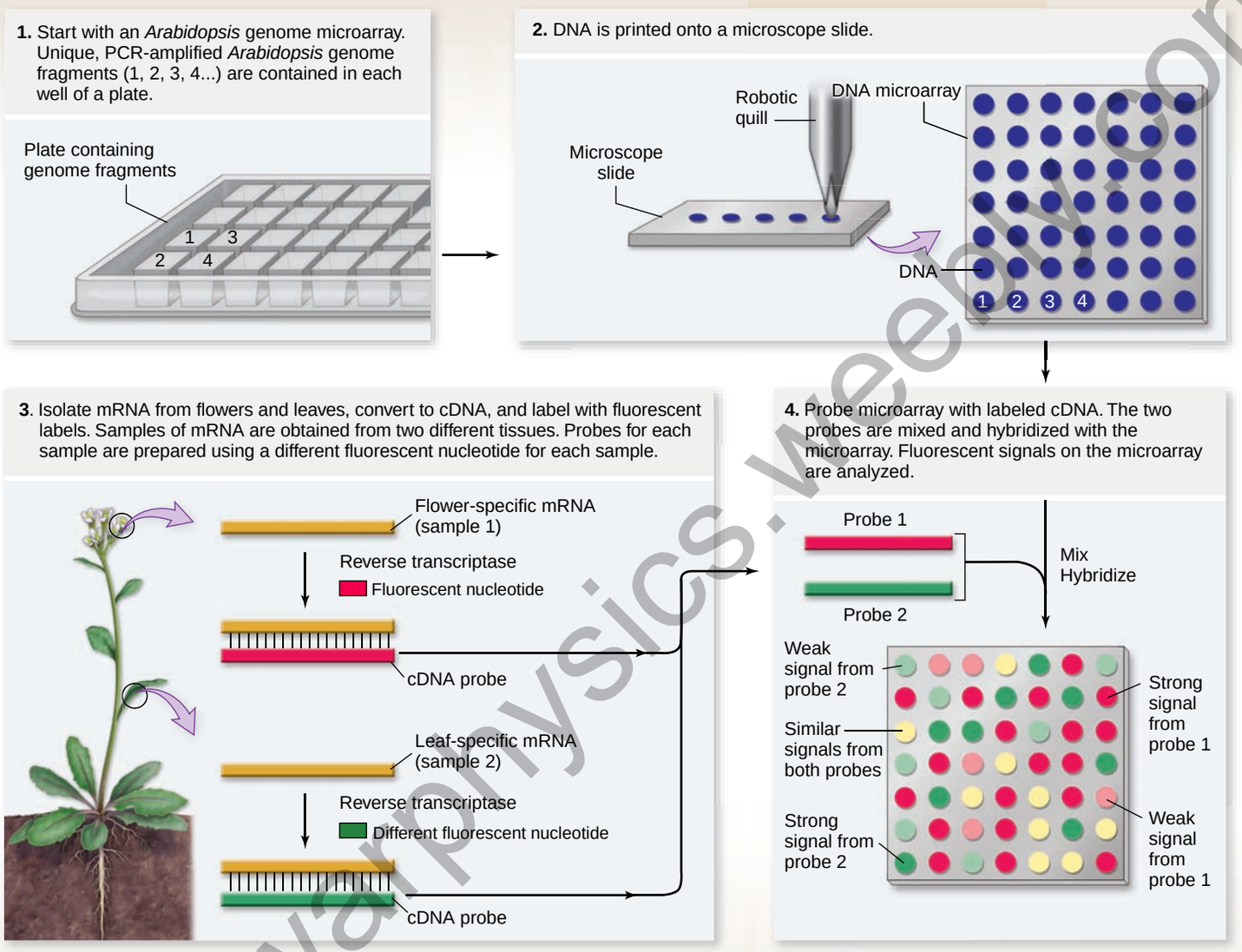
Genome-wide association (GWA) is an approach that compares SNPs throughout the genome between members in a population with and without a specific trait. The goal is to find a SNP that correlates with a specific trait as a way to map the trait. The dog genome exemplifies the value of GWA mapping. Using microarrays that distinguish between 15,000 SNP variants,

## SCIENTIFIC THINKING

**Hypothesis:** *Flowers and leaves will express some of the same genes.*

**Prediction:** *When mRNAs isolated from* Arabidopsis *flowers and from leaves are used as probes on an* Arabidopsis *genome microarray, the two different probe sets will hybridize to both common and unique sequences.*

**Test:**




**Result:** *Yellow spots represent sequences that hybridized to cDNA from both flowers and leaves. Red spots represent genes expressed only in flowers. Green spots represent genes expressed only in leaves.*

**Conclusion:** *Some* Arabidopsis *genes are expressed in both flowers and leaves, but there are genes expressed in flowers but not leaves and leaves but not flowers.*

**Further Experiments:** *How could you use microarrays to determine whether the genes expressed in both flowers and leaves are housekeeping genes or are unique to flowers and leaves?*

**Figure 18.10** **Microarrays.**

disease alleles for a recessive trait can be mapped by comparing 20 purebred dogs exhibiting the disease with 20 healthy dogs.

### *Transgenics*

How can we determine whether two genes from different species having similar sequences have the same function? And, how can we be sure that a gene identified by an annotation program actually functions as a gene in the organism? One way to address these questions is through transgenics—the creation of organisms containing genes from other species (transgenic organisms).

The technology for creating transgenic organisms was discussed in chapter 17; it is illustrated for plants in

figure 18.11. Different markers can be incorporated into the gene so that its protein product can be visualized or isolated in the transgenic plant, demonstrating that the inserted gene is being transcribed. In some cases, the transgene (inserted foreign gene) may affect a visible phenotype. Of course, transgenics are but one of many ways to address questions about gene function.

## Proteomics moves from genes to proteins

Proteins are much more difficult to study than DNA because of posttranslational modification and formation of protein complexes. And, as already mentioned, a single gene can code for multiple proteins using alternative splicing. Although all the DNA in a genome can be isolated from a single cell, only a portion of the proteome is expressed in a single cell or tissue.

**Proteomics** is the study of the **proteome**—all of the proteins encoded by the genome. Understanding the proteome for even a single cell will be a much more difficult task than determining the sequence of a genome. Because a single gene can produce more than one protein by alternative splicing, the first step is to characterize the **transcriptome**—all of the RNA that is present in a cell or tissue. Because of alternative splicing, both the transcriptome and the proteome are larger and more complex than the simple number of genes in the genome.

To make matters worse, a single protein can be modified posttranslationally to produce functionally different forms. The function of a protein can also depend on its association with other proteins. Nonetheless, since proteins perform most of the major functions of cells, understanding their diversity is essential.

**Inquiry question**

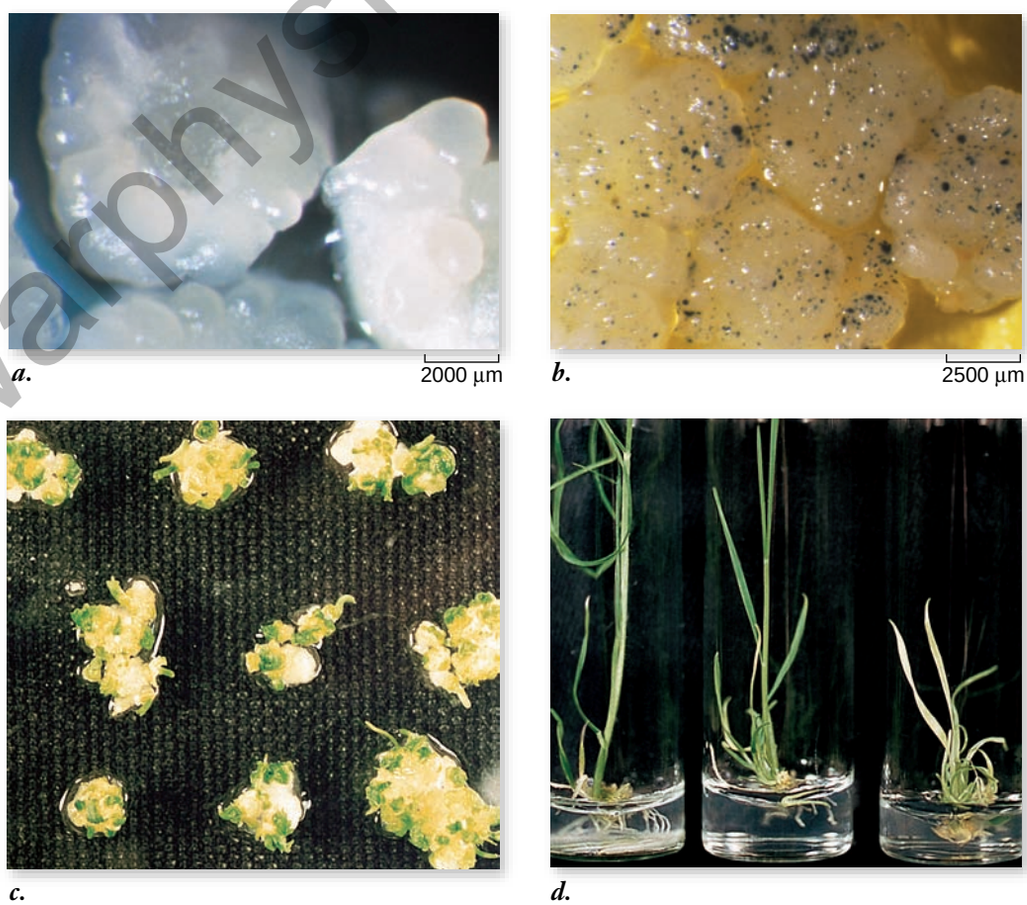
**?** **Why is the "proteome" likely to be different from simply the predicted protein products found in the complete genome sequence?**

### *Predicting protein function*

The use of new methods to quickly identify and characterize large numbers of proteins is the distinguishing feature between traditional protein biochemistry and proteomics. As with genomics, the challenge is one of scale.

Ideally, a researcher would like to be able to examine a nucleotide sequence and know what sort of functional protein the sequence specifies. Databases of protein structures in different organisms can be searched to predict the structure and function of genes known only by sequence, as identified in genome projects. Analysis of these data provides a clearer picture of how gene sequence relates to protein structure and function. Having a greater number of DNA sequences available allows for more extensive comparisons as well as identification of common structural patterns as groups of proteins continue to emerge.

Although there may be as many as a million different proteins, most are just variations on a handful of themes.




***a.*** ***b.*** ***c.*** ***d.***

**Figure 18.11 Growth of a transgenic plant.** DNA containing a gene for herbicide resistance was transferred into wheat *(Triticum aestivum)*. The DNA also contains the *GUS* gene, which is used as a tag or label. The *GUS* gene produces an enzyme that catalyzes the conversion of a staining solution from clear to blue. ***a.*** Embryonic tissue just prior to insertion of foreign DNA. ***b.*** Following DNA transfer, callus cells containing the foreign DNA are indicated by color from the *GUS* gene *(blue spots)*. ***c.*** Shoot formation in the transgenic plants growing on a selective medium. Here, the gene for herbicide resistance in the transgenic plants allows growth on the selective medium containing the herbicide. ***d.*** Comparison of growth on the selection medium for transgenic plants bearing the herbicide resistance gene *(left)* and a nontransgenic plant *(right)*.

The same shared structural motifs—barrels, helices, molecular zippers—are found in the proteins of plants, insects, and humans (figure 18.12; also see chapter 3 for more information on protein motifs). The maximum number of distinct motifs has been estimated to be fewer than 5000. About 1000 of these motifs have already been cataloged. Efforts are now under way to detail the shapes of all the common motifs.

### *Protein microarrays*

**Protein microarrays,** comparable to DNA microarrays, are being used to analyze large numbers of proteins simultaneously. Making a protein microarray starts with isolating the transcriptome of a cell or tissue. Then cDNAs are constructed and reproduced by cloning them into bacteria or viruses. Transcription and translation occur in the prokaryotic host, and micromolar quantities of protein are isolated and purified. These are then spotted onto glass slides.

Protein microarrays can be probed in at least three different ways. First, they can be screened with antibodies to specific proteins. Antibodies are labeled so that they can be detected, and the patterns on the protein array can be determined by computer analysis.

An array of proteins can also be screened with another protein to detect binding or other protein interactions. Thousands of interactions can be tested simultaneously. For example, calmodulin (which mediates $Ca^{2+}$ function; see chapter 9) was labeled and used to probe a yeast proteome array with 5800 proteins. The screen revealed 39 proteins that bound calmodulin. Of those 39, 33 were previously unknown!

A third type of screen uses small molecules to assess whether they will bind to any of the proteins on the array. This approach shows promise for discovering new drugs that will inhibit proteins involved in disease.

## Large-scale screens reveal protein–protein interactions

We often study proteins in isolation, compared with their normal cellular context. This approach is obviously artificial. One immediate goal of proteomics, therefore, is to map all the physical interactions between proteins in a cell. This is a daunting task that requires tools that can be automated, similarly to the way that genome sequencing was automated.

One approach is to use the yeast two-hybrid system discussed in the preceding chapter. This system can be automated once libraries of known cDNAs are available in each of the two vectors used. The use of two-hybrid screens has been applied to budding yeast to generate a map of all possible interacting proteins. This method is difficult to apply to more complex multicellular organisms, but in a technical tour-de-force, it has been applied to *Drosophila melanogaster* as well.

For vertebrates, the two-hybrid system is being applied more selectively, by concentrating on a biologically significant process, such as signal transduction. The technique can then be used to map all of the interacting proteins in a specific signaling pathway.

**Figure 18.12 Computer-generated model of an enzyme.** Searchable databases contain known protein structures, including human aldose reductase shown here. Secondary structural motifs are shown in different colors.

### Inquiry question

**What is the relationship among genome, transcriptome, and proteome?**

### *Learning Outcomes Review 18.4*

**Comparisons of different genomes allows geneticists to infer structural, functional, and evolutionary relationships between genes and proteins as well as relationships between species. Microarrays enable evaluation of gene expression for many genes at once. Proteomics involves similar analysis of all the proteins coded by a genome, that is, an organism's proteome. Because of alternative splicing, this task is much more complex.**

- ***Why is establishment of a species' transcriptome an important step in studying its proteome?***

# 18.5 *Applications of Genomics*

### *Learning Outcomes*

1. ***List ways in which genomics could be applied to infectious disease research.***
2. ***Explain how genomics could enhance crop production and nutritional yield.***
3. ***Evaluate the issues of genome ownership and privacy.***

Space allows us to highlight only a few of the myriad applications of genomics to show the possibilities. The tools being developed truly represent a revolution in biology that will likely have a lasting influence on the way that we think about living systems.

## Genomics can help to identify infectious diseases

The genomics revolution has yielded millions of new genes to be investigated. The potential of genomics to improve human health is enormous. Mutations in a single gene can explain some, but not most, hereditary diseases. With entire genomes to search, the probability of unraveling human, animal, and plant diseases is greatly improved.

Although proteomics will likely lead to new pharmaceuticals, the immediate effect of genomics is being seen in diagnostics. Both improved technology and gene discovery are enhancing the diagnosis of genetic abnormalities.

Diagnostics are also being used to identify individuals. For example, short tandem repeats (STRs), discovered through genomic research, were among the forensic diagnostic tools used to identify remains of victims of the September 11, 2001, terrorist attack on the World Trade Center in New York City.

The September 11 attacks were followed by an increased awareness and concern about biological weapons. Five people died and 17 more were infected with anthrax after envelopes containing anthrax spores were sent through the U.S. mail. A massive FBI investigation initially focused on the wrong individual, Steven J. Hatfill, a government scientist. Genome sequencing allowed exploration of possible sources of the deadly bacteria. A difference of only 10 bp between strains allowed the FBI to trace the source to a single vial of the bacteria used in a vaccine research program at U.S. Army Medical Research Institute for Infectious Diseases. By 2008, Hatfield was exonerated. Another researcher, Bruce E. Ivins, committed suicide just before being formally charged by the FBI with criminal activity in the 2001 anthrax attacks. Ivins had been working on vaccine development. In addition, substantial effort has been turned toward the use of genomic tools to distinguish between naturally occurring infections and intentional outbreaks of disease. The Centers for Disease Control and Prevention (CDC) have ranked bacteria and viruses that are likely targets for bioterrorism (table 18.2).

**TABLE 18.2** High-Priority Pathogens for Genomic Research

| Pathogen | Disease | Genome* |
|---|---|---|
| *Variola major* | Smallpox | Complete |
| *Bacillus anthracis* | Anthrax | Complete |
| *Yersinia pestis* | Plague | Complete |
| *Clostridium botulinum* | Botulism | In progress |
| *Francisella tularensis* | Tularemia | Complete |
| Filoviruses | Ebola and Marburg hemorrhagic fever | Both are complete |
| Arenaviruses | Lassa fever and Argentine hemorrhagic fever | Both are complete |

*There are multiple strains of these viruses and bacteria. "Complete" indicates that at least one has been sequenced. For example, the Florida strain of anthrax was the first to be sequenced.

## Genomics can help improve agricultural crops

Globally speaking, poor nutrition is the greatest impediment to human health. Much of the excitement about the rice genome project is based on its potential for improving the yield and nutritional quality of rice and other cereals worldwide. The development of Golden Rice (chapter 17) is an example of improved nutrition through genetic approaches. About one third of the world population obtains half its calories from rice (figure 18.13). In some regions, individuals consume up to 1.5 kg of rice daily. More than 500 million tons of rice is produced each year, but this may not be adequate to provide enough rice for the world in the future.

Due in large part to scientific advances in crop breeding and farming techniques, in the last 50 years world grain production has more than doubled, with an increase in cropland of only 1%. The world now farms a total area the size of South America, but without the scientific advances of the past 50 years, an area equal to the entire western hemisphere would need to be farmed to produce enough food for the world.

Unfortunately, water usage for crops has tripled in that time period, and quality farmland is being lost to soil erosion. Scientists are also concerned about the effects of global climate

**Figure 18.13 Rice field.** Most of the rice grown globally is directly consumed by humans and is the dietary mainstay of 2 billion people.

change on agriculture worldwide. Increasing the yield and quality of crops, especially on more marginal farmland, will depend on many factors—but genetic engineering, built on the findings of genomics projects, can contribute significantly to the solution.

Most crops grown in the United States produce less than half of their genetic potential because of environmental stresses (salt, water, and temperature), herbivores, and pathogens (figure 18.14). Identifying genes that can provide resistance to stress and pests is the focus of many current genomics research projects. Having access to entire genomic sequences will enhance the probability of identifying critical genes.

## Genomics raises ethical issues over ownership of genomic information

Genome science is also a source of ethical challenges and dilemmas. One example is the issue of gene patents. Actually, it is the use of a gene, not the gene itself, that is patentable. For a patent to be granted for a gene's use, the product and its function must be known.

The public genome consortia, supported by federal funding, have been driven by the belief that the sequence of genomes should be freely available to all and should not be patented. Private companies patent gene functions, but they often make sequence data available with certain restrictions. The physical sciences have negotiated the landscape of public and for-profit research for decades, but this is relatively new territory for biologists.

Another ethical issue involves privacy. How sequence data are used is the focus of thoughtful and ongoing discussions. The Universal Declaration on the Human Genome and Human Rights states, "The human genome underlies the fundamental unity of all members of the human family, as well as the recognition of their inherent dignity and diversity. In a symbolic sense, it is the heritage of humanity."

Although we talk about "the" human genome, each of us has subtly different genomes that can be used to identify us. Genetic disorders such as cystic fibrosis and Huntington disease can already be identified by screening, but genomics will greatly increase the number of identifiable traits. The Genetic Information Nondiscrimination Act (GINA) was signed into law in 2008 to prevent discrimination based on genotype. Employers and health insurance companies may not request genetic tests or discriminate based on someone's genetic code. Life, disability, and long term care insurance are not covered by GINA. Members of the military are excluded from GINA's privacy protection. The U.S. Armed Forces require DNA samples from members of the military for possible casualty identification. The genome privacy debate continues.

Behavioral genomics is an area that is also rich with possibilities and dilemmas. Very few behavioral traits can be accounted for by single genes. Two genes have been associated with fragile-X mental retardation, and three with early-onset Alzheimer disease. Comparisons of multiple genomes will likely lead to the identification of multiple genes controlling a range of behaviors. Will this change the way we view acceptable behavior?

**Figure 18.14 Corn crop productivity well below its genetic potential due to drought stress.** Corn production can be limited by water deficiencies due to the drought that occurs during the growing season in dry climates. Global climate change may increase drought stress in areas where corn is the major crop.

### Inquiry question

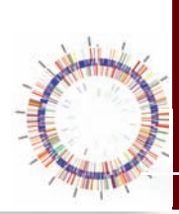
**As of February 2008 a draft version of the corn genome has been sequenced. How could you use information from the corn and rice genome sequences to try to improve drought tolerance in corn?**

In Iceland, the parliament has voted to have a private company create a database from pooled medical, genetic, and genealogical information about all Icelanders, a particularly fascinating population from a genetic perspective. Because minimal migration or immigration has occurred there over the last 800 years, the information that can be mined from the Icelandic database is phenomenal. Ultimately, the value of that information has to be weighed, however, against any possible discrimination or stigmatization of individuals or groups.

### *Learning Outcomes Review 18.5*

**Genomics is one approach to better diagnosis, based on knowledge of infectious agents' genetic makeup; it also allows identification of individual disease strains. Genomics has enhanced DNA identification of remains. Agricultural crop yields and nutritional content could be improved if genes that confer disease resistance or increased synthesis can be identified.**

- ***Suppose you produced an engineered form of potato that had twice the amount of protein. Would you seek a patent on this plant?***



# Chapter Review

## 18.1 Mapping Genomes

***Different kinds of physical maps can be generated.***

Physical genetic maps include fully sequenced genomes, restriction maps, and maps of chromosome banding patterns.

***Sequence-tagged sites provide a common language for physical maps.***

Any physical site can be used as a sequence-tagged site (STS), based on a small stretch of a unique DNA sequence that allows unambiguous identification of a fragment.

***Genetic maps provide a link to phenotypes.***

Short tandem repeats (STRs) are the most common type of markers for distinguishing regions of the genome and assessing its phenotypic effects.

***Physical maps can be correlated with genetic maps.***

Physical and genetic maps can be correlated. Any gene that can be cloned can be placed within the genome sequence and mapped. However, absolute correspondence of distances cannot be accomplished.

## 18.2 Whole-Genome Sequencing

***Genome sequencing requires larger molecular clones.***

Yeast artificial chromosomes (YACs) have allowed cloning of larger pieces of DNA, although their use has some drawbacks. Bacterial artificial chromosomes (BACs) are most commonly used now.

***Whole-genome sequencing is approached in two ways: clone-by-clone and shotgun.***

Clone-by-clone sequencing starts with known clones, often in BACs that can be aligned with each other.

Shotgun sequencing involves sequencing random clones, then using a computer to assemble the finished sequence.

***The Human Genome Project used both sequencing methods.***

By 2004, the "finished" sequence was announced, and it includes 99% of the euchromatic human DNA sequence.

## 18.3 Characterizing Genomes

***The Human Genome Project found fewer genes than expected.***

Although eukaryotic genomes are larger and have more genes than those of prokaryotes, the size of the organism is not always correlated with the size of the genome. The human genome contains only around 25,000 genes, fewer than found in rice.

***Finding genes in sequence data requires computer searches.***

In a sequenced genome, protein-coding genes are identified by looking for open-reading frames (ORFs). An ORF begins with a start codon and contains no stop codon for a distance long enough to encode a protein. Genes are then grouped based on conserved regions.

***Genomes contain both coding and noncoding DNA.***

Protein-encoding DNA includes single-copy genes, segmental duplications, multigene families, and tandem clusters. Noncoding DNA in eukaryotes makes up about 99% of DNA. Approximately 45% of the human genome is composed of mobile transposable elements, including LINEs, SINEs, and LTRs.

***Expressed sequence tags identify genes that are transcribed.***

The number and location of expressed genes can be estimated by sequencing the ends of randomly selected cDNAs to produce expressed sequence tags (ESTs).

***SNPs are single-base differences between individuals.***

Single-nucleotide difference between individuals are called single-nucleotide polymorphisms (SNPs). To be classified as a polymorphism, an SNP must be present in at least 1% of the population. At least 50,000 SNPs are currently known in coding regions.

Genomic haplotypes are regions of chromosomes that are not exchanged by recombination. These regions can be used to map genes by association (see figure 18.8).

## 18.4 Genomics and Proteomics

***Comparative genomics reveals conserved regions in genomes.***

More than half of the genes of *Drosophila* have human counterparts. The biggest difference between our genome and the chimpanzee genome is in transposable elements.

***Synteny allows comparison of unsequenced genomes.***

Synteny refers to the conserved arrangements of segments of DNA in related genomes (see figure 18.9). Many separate species have been found to have large regions of synteny.

***Organelle genomes have exchanged genes with the nuclear genome.***

Both chloroplasts and mitochondria contain components that indicate exchange of genetic material with the nuclear genome.

***Functional genomics reveals gene function at the genome level.***

Functional genomics uses high-end computer technology to analyze gene function and gene products. DNA microarrays allow the expression of all of the genes in a cell to be monitored at once (see figure 18.10).

***Proteomics moves from genes to proteins.***

Proteomics characterizes all of the proteins produced by a cell. The transcriptome is all the mRNAs present in a cell at a specific time. Protein microarrays can identify and characterize large numbers of proteins.

***Large-scale screens reveal protein–protein interactions.***

The yeast two-hybrid system is used to generate large-scale maps of interacting proteins; however, the scope of this task is daunting in humans, mice, and other vertebrates. Selective applications in specific areas, such as signal transduction, have been undertaken.

## 18.5 Applications of Genomics

***Genomics can help to identify infectious diseases.***

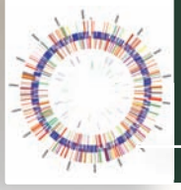
Genomics can help identify naturally occurring and intentional outbreaks of infectious diseases and tracing of disease strains.

***Genomics can help improve agricultural crops.***

Genomics can potentially increase the nutritional value of crops and alter their responses to environmental stresses, potentially helping to feed a growing population.

***Genomics raises ethical issues over ownership of genomic information.***

Questions regarding profit and ownership of genomic data provide ongoing challenges for the ethical use of scientific knowledge.



# Review Questions

## UNDERSTAND

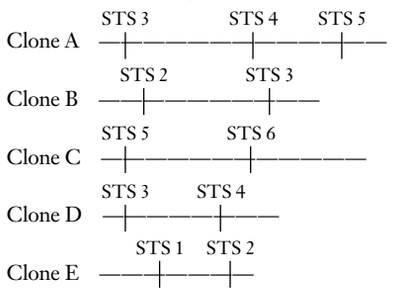
1. A genetic map is based on the
   a. sequence of the DNA.
   b. relative position of genes on chromosomes.
   c. location of sites of restriction enzyme cleavage.
   d. banding pattern on a chromosome.
2. What is an STS?
   a. A unique sequence within the DNA that can be used for mapping
   b. A repeated sequence within the DNA that can be used for mapping
   c. An upstream element that allows for mapping of the 3′ region of a gene
   d. Both b and c
3. Which number represents the total number of genes in the human genome?
   a. 2500 b. 10,000
   c. 25,000 d. 100,000
4. An open reading frame (ORF) is distinguished by the presence of
   a. a stop codon.
   b. a start codon.
   c. a sequence of DNA long enough to encode a protein.
   d. All of the above
5. What is a BLAST search?
   a. A mechanism for aligning consensus regions during whole-genome sequencing
   b. A search for similar gene sequences from other species
   c. A method of screening a DNA library
   d. A method for identifying ORFs
6. Which of the following is *not* an example of a protein-encoding gene?
   a. Single-copy gene b. Tandem clusters
   c. Pseudogene d. Multigene family
7. What is a proteome?
   a. The collection of all genes encoding proteins
   b. The collection of all proteins encoded by the genome
   c. The collection of all proteins present in a cell
   d. The amino acid sequence of a protein
8. Which of the following is *not* an example of noncoding DNA?
   a. Promoter b. Intron
   c. Pseudogene d. Exon

## APPLY

1. An artificial chromosome is useful because it
   a. produces more consistent results than a natural chromosome.
   b. allows for the isolation of larger DNA sequences.
   c. provides a high copy number of a DNA sequence.
   d. is linear.
2. Comparisons between genomes is made easier because of
   a. synteny. b. haplotypes.
   c. transposons. d. expressed sequence tags.
3. Which of the following techniques relies on prior knowledge of overlapping sequences?
   a. Yeast two-hybrid system
   b. Shotgun method of genome sequencing
   c. FISH
   d. Clone-by-clone method of genome sequencing
4. The duplication of a gene due to uneven meiotic crossing over is thought to lead to the production of a
   a. segmental duplication. c. simple sequence repeat.
   b. tandem duplication. d. multigene family.
5. What information can be obtained from a DNA microarray?
   a. The sequence of a particular gene.
   b. The presence of genes within a specific tissue.
   c. The pattern of gene expression.
   d. Differences between genomes.
6. Which of the following is true regarding microarray technology and cancer?
   a. A DNA microarray can determine the type of cancer.
   b. A DNA microarray can measure the response of a cancer to therapy.
   c. A DNA microarray can be used to predict whether the cancer will metastasize.
   d. All of the above
7. Which of the following techniques could be used to examine protein–protein interactions in a cell?
   a. Two-hybrid screens c. Protein microarrays
   b. Protein structure databases d. Both a and c

## SYNTHESIZE

1. You are in the early stages of a genome-sequencing project. You have isolated a number of clones from a BAC library and mapped the inserts in these clones using STSs. Use the STSs to align the clones into a contiguous sequence of the genome (a contig).



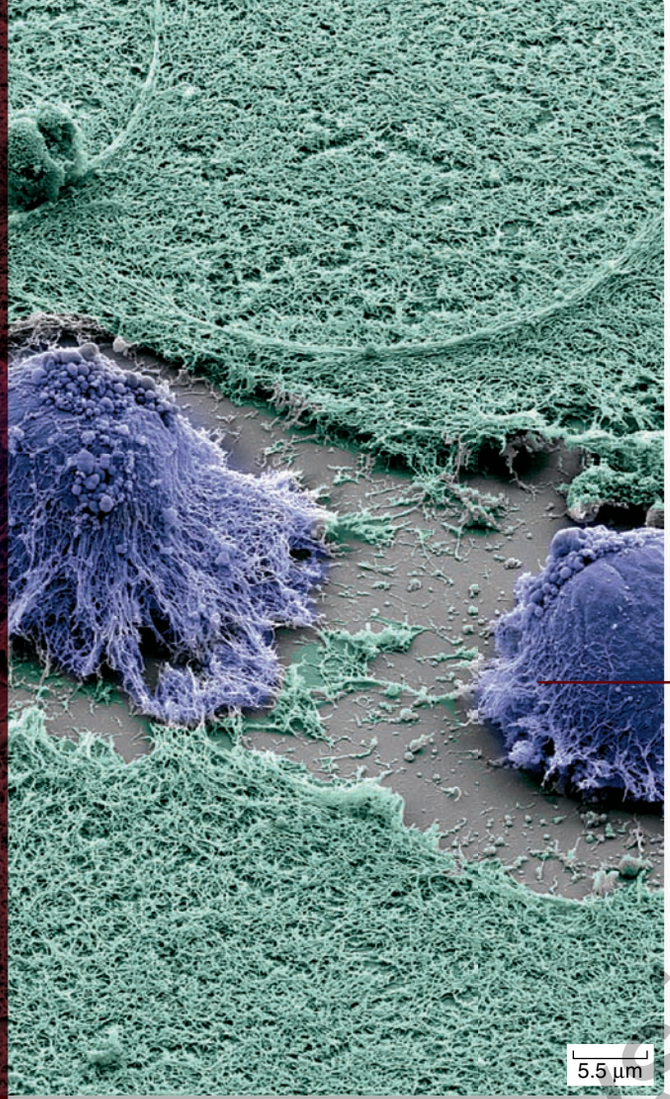

2. Genomic research can be used to determine if an outbreak of an infectious disease is natural or "intentional." Explain what a genomic researcher would be looking for in a suspected intentional outbreak of a disease like anthrax.

## ONLINE RESOURCE

www.ravenbiology.com



Understand, Apply, and Synthesize—enhance your study with animations that bring concepts to life and practice tests to assess your understanding. Your instructor may also recommend the interactive eBook, individualized learning tools, and more.




Chapter 19

# Cellular Mechanisms of Development

## Chapter Outline



## Introduction

*Recent work with different kinds of stem cells, like those pictured, have captured the hopes and imagination of the public. For thousands of years, humans have wondered how organisms arise, grow, change, and mature. We are now in an era when long-standing questions may be answered, and new possibilities for regenerative medicine seem on the horizon.*

*We have explored gene expression from the perspective of individual cells, examining the diverse mechanisms cells employ to control the transcription of particular genes. Now we broaden our perspective and look at the unique challenge posed by the development of a single cell, the fertilized egg, into a multicellular organism. In the course of this developmental journey, a pattern of decisions about gene expression takes place that causes particular lines of cells to proceed along different paths, spinning an incredibly complex web of cause and effect. Yet, for all its complexity, this developmental program works with impressive precision. In this chapter, we explore the mechanisms of development at the cellular and molecular level.*

## 19.1 The Process of Development

Development can be defined as the process of systematic, gene-directed changes through which an organism forms the successive stages of its life cycle. Development is a continuum, and explorations of development can be focused on any point along this continuum. The study of development plays a central role in unifying the understanding of both the similarities and diversity of life on Earth.

We can divide the overall process of development into four subprocesses:

- **Cell Division.** A developing plant or animal begins as a fertilized egg, or zygote, that must undergo cell division to produce the new individual. In all cases early development involves extensive cell division, but in many cases it does not include much growth as the egg cell itself is quite large.
- **Differentiation.** As cells divide, orchestrated changes in gene expression result in differences between cells that ultimately result in cell specialization. In differentiated

cells, certain genes are expressed at particular times, but other genes may not be expressed at all.

- **Pattern Formation.** Cells in a developing embryo must become oriented to the body plan of the organism the embryo will become. Pattern formation involves cells' abilities to detect positional information that guides their ultimate fate.
- **Morphogenesis.** As development proceeds, the form of the body—its organs and anatomical features—is generated. Morphogenesis may involve cell death as well as cell division and differentiation.

Despite the overt differences between groups of plants and animals, most multicellular organisms develop according to molecular mechanisms that are fundamentally very similar. This observation suggests that these mechanisms evolved very early in the history of multicellular life.

## 19.2 Cell Division

### Learning Outcomes

1. *Explain the importance of cell division to early development.*
2. *Describe the use of* C. elegans *to track cell lineages.*
3. *Distinguish differences in cell division between animals and plants.*

When a frog tadpole hatches out of its protective coats, it is roughly the same overall mass as the fertilized egg from which it came. Instead of being made up of just one cell, however, the tadpole consists of about a million cells, which are organized into tissues and organs with different functions. Thus, the very first process that must occur during embryogenesis is cell division.

Immediately following fertilization, the diploid zygote undergoes a period of rapid mitotic divisions that ultimately result in an early embryo comprised of dozens to thousands of diploid cells. In animal embryos, the timing and number of these divisions are species-specific and are controlled by a set of molecules that we examined in chapter 10: the *cyclins* and *cyclin-dependent kinases (Cdks)*. These molecules exert control over checkpoints in the cycle of mitosis.

### Development begins with cell division

In animal embryos, the period of rapid cell division following fertilization is called **cleavage.** During cleavage, the enormous mass of the zygote is subdivided into a larger and larger number of smaller and smaller cells, called **blastomeres** (figure 19.1). Hence, cleavage is not accompanied by any increase in the overall size of the embryo. The $G_1$ and $G_2$ phases of the cell cycle, during which a cell increases its mass and size, are extremely shortened or eliminated during cleavage (figure 19.2).

Because of the absence of the two gap/growth phases, the rapid rate of mitotic divisions during cleavage is never again approached in the lifetime of any animal. For example, zebrafish blastomeres divide once every several minutes during cleavage, to create an embryo with a thousand cells in just under 3 hr! In contrast, cycling adult human intestinal epithelial cells divide on average only once every 19 hr. A comparison of the different patterns of cleavage can be found in chapter 54.

When external sources of nutrients become available—for example, during larval feeding stages or after implantation of mammalian embryos—daughter cells can increase in size following cytokinesis, and an overall increase in the size of the organism occurs as more cells are produced.

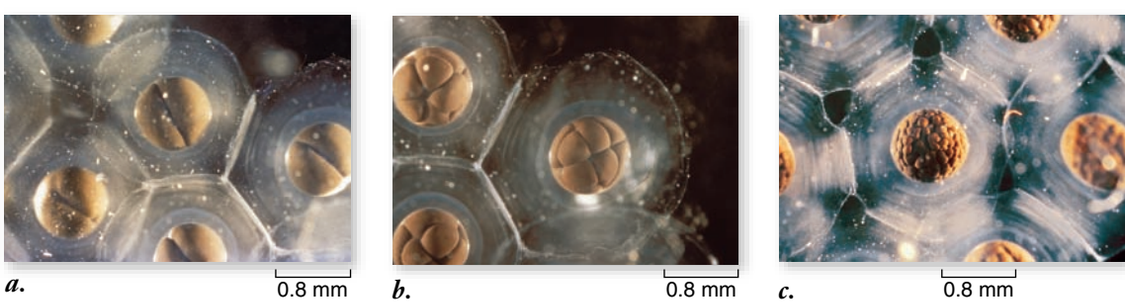
### Every cell division is known in the development of C. *elegans*

One of the most completely described models of development is the tiny nematode *Caenorhabditis elegans*. Only about 1 mm long, the adult worm consists of 959 somatic cells.

Because *C. elegans* is transparent, individual cells can be followed as they divide. By observing them, researchers have learned how each of the cells that make up the adult worm is derived from the fertilized egg. As shown on the lineage map in figure 19.3*a*, the egg divides into two cells, and these daughter cells continue to divide. Each horizontal line on the map represents one round of cell division. The length of each vertical line represents the time between cell divisions, and the end of each vertical line represents one fully differentiated cell. In figure 19.3*b*, the major organs of the worm are color-coded to match the colors of the corresponding groups of cells on the lineage map.

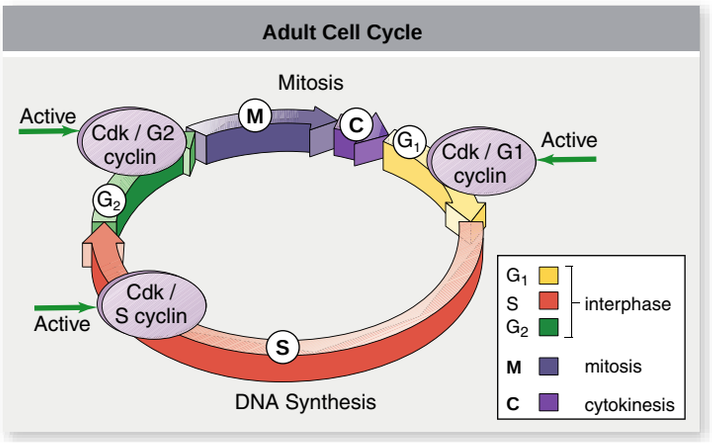
Some of these differentiated cells, such as some cells that generate the worm's external cuticle, are "born" after only 8 rounds of cell division; other cuticle cells require as many as 14 rounds. The cells that make up the worm's pharynx, or feeding organ, are born after 9 to 11 rounds of division, whereas cells in the gonads require up to 17 divisions.

Exactly 302 nerve cells are destined for the worm's nervous system. Exactly 131 cells are programmed to die, mostly within minutes of their "birth." The fate of each cell is the same in every *C. elegans* individual, except for the cells that will become eggs and sperm.

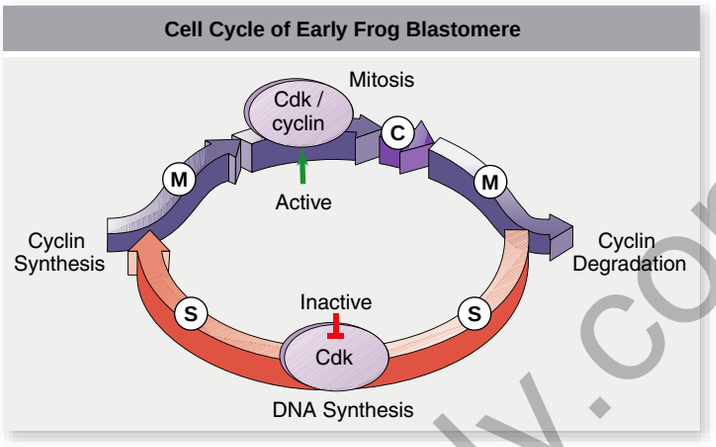



***a.*** ***b.*** ***c.***

**Figure 19.1 Cleavage divisions in a frog embryo.** ***a.*** The first cleavage division divides the egg into two large blastomeres. ***b.*** After two more divisions, four small blastomeres sit on top of four large blastomeres, each of which continues to divide to produce ***(c)*** a compact mass of cells.




*a.*




*b.*

**Figure 19.2 Cell cycle of adult cell and embryonic cell.** In contrast to the cell cycle of adult somatic cells ***(a)***, the dividing cells of early frog embryos lack $G_1$ and $G_2$ stages ***(b)***, enabling the cleavage stage nuclei to rapidly cycle between DNA synthesis and mitosis. Large stores of cyclin mRNA are present in the unfertilized egg. Periodic degradation of cyclin proteins correlates with exiting from mitosis. Cyclin degradation and Cdk inactivation allow the cell to complete mitosis and initiate the next round of DNA synthesis.

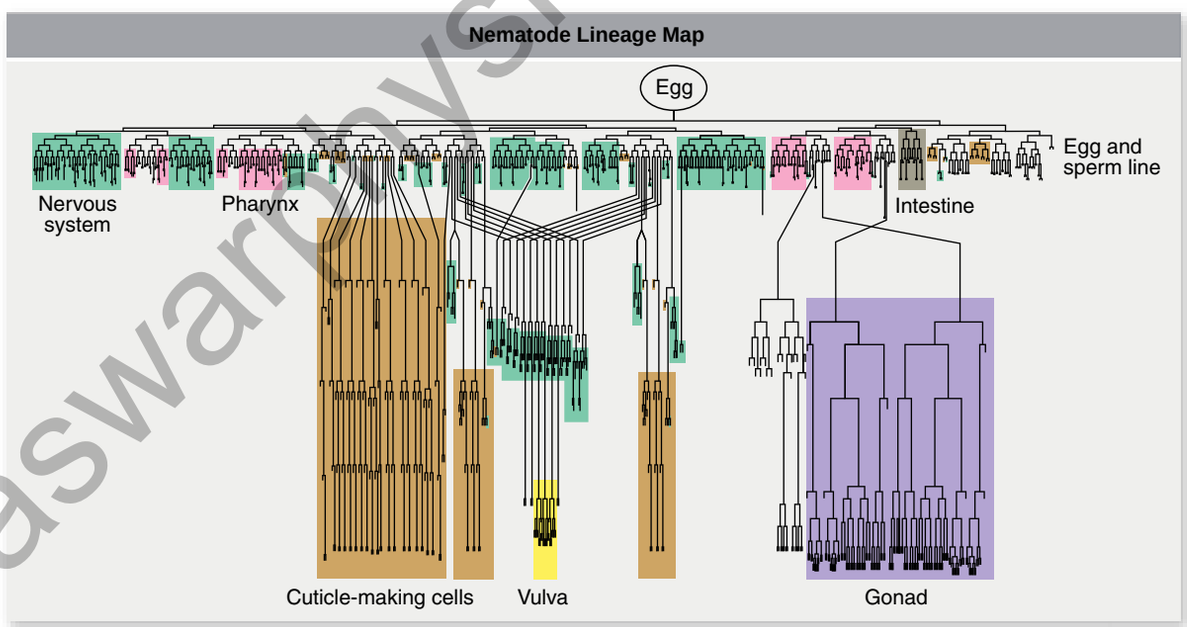
## Plant growth occurs in specific areas called meristems

A major difference between animals and plants is that most animals are mobile, at least in some phase of their life cycles, and therefore they can move away from unfavorable circumstances. Plants, in contrast, are anchored in position and must simply endure whatever environment they experience. Plants compensate for this restriction by allowing development to accommodate local circumstances.

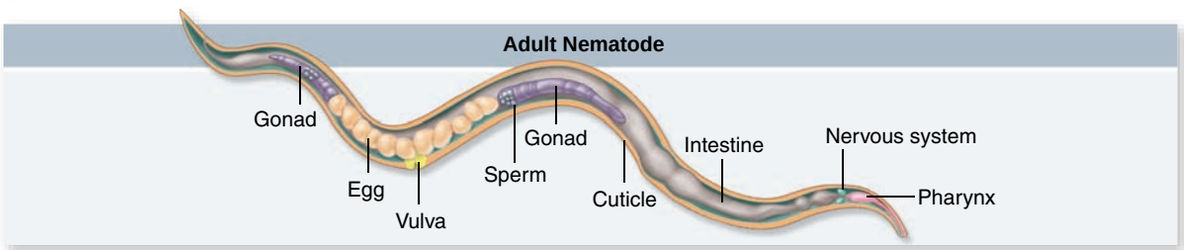
Instead of creating a body in which every part is specified to have a fixed size and location, a plant assembles its body throughout its life span from a few types of modules, such as leaves, roots, branch nodes, and flowers. Each module has a rigidly controlled structure and organization, but how the modules are utilized is quite flexible—they can be adjusted to environmental conditions.

**Figure 19.3 Studying embryonic cell division and development in the nematode.** Development in *C. elegans* has been mapped out such that the fate of each cell from the single egg cell has been determined. ***a.*** The lineage map shows the number of cell divisions from the egg, and the color coding links their placement in ***(b)*** the adult organism.

M. E. Challinor illustration. From Howard Hughes Medical Institute © as published in *From Egg to Adult*, 1992. Reprinted by permission.




*a.*




*b.*

Plants develop by building their bodies outward, creating new parts from groups of stem cells that are contained in structures called **meristems.** As meristematic stem cells continually divide, they produce cells that can differentiate into the tissues of the plant.

This simple scheme indicates a need to control the process of cell division. We know that cell-cycle control genes are present in both yeast (fungi) and animal cells, implying that these are a eukaryotic innovation—and in fact, the plant cell cycle is regulated by the same mechanisms, namely through cyclins and cyclin-dependent kinases. In one experiment, overexpression of a Cdk inhibitor in transgenic *Arabidopsis thaliana* plants resulted in strong inhibition of cell division in leaf meristems, leading to significant changes in leaf size and shape.

### *Learning Outcomes Review 19.2*

**In animal embryos, a series of rapid cell divisions that skip the $G_1$ and $G_2$ phases convert the fertilized egg into many cells with no change in size. In the nematode *C. elegans,* every cell division leading to the adult form is known, and this pattern is invariant, allowing biologists to trace development in a cell-by-cell fashion. In plants, growth is restricted to specific areas called meristems, where undifferentiated stem cells are retained.**

- ***How are early cell divisions in an embryo different from in an adult organism?***

## 19.3 Cell Differentiation

### *Learning Outcomes*

1. *Describe the progressive nature of determination.*
2. *List the ways in which cells become committed to developmental pathways on the molecular level.*
3. *Differentiate between the different types of stem cells.*

In chapter 16, we examined the mechanisms that control eukaryotic gene expression. These processes are critical for the development of multicellular organisms, in which life functions are carried out by different tissues and organs. In the course of development, cells become different from one another because of the differential expression of subsets of genes—not only at different times, but in different locations of the growing embryo. We now explore some of the mechanisms that lead to differential gene expression during development.

### Cells become determined prior to differentiation

A human body contains more than 210 major types of differentiated cells. These differentiated cells are distinguishable from one another by the particular proteins that they synthesize, their morphologies, and their specific functions. A molecular decision to become a particular type of differentiated cell occurs prior to any overt changes in the cell. This molecular decision-making process is called **cell determination,** and it commits a cell to a particular developmental pathway.

#### *Tracking determination*

Determination is often not visible in the cell and can only be "seen" by experiment. The standard experiment to test whether a cell or group of cells is determined is to move the donor cell(s) to a different location in a host (recipient) embryo. If the cells of the transplant develop into the same type of cell as they would have if left undisturbed, then they are judged to be already determined (figure 19.4).

Determination has a time course; it depends on a series of intrinsic or extrinsic events, or both. For example, a cell in the prospective brain region of an amphibian embryo at the early gastrula stage (see chapter 54) has not yet been determined; if transplanted elsewhere in the embryo, it will develop according to the site of transplant. By the late gastrula stage, however,

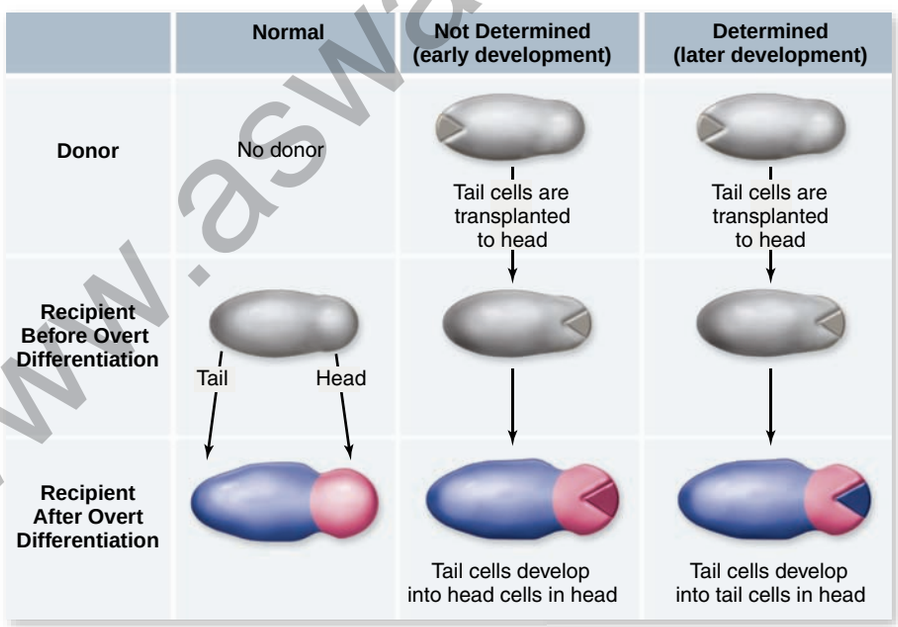



**Figure 19.4 The standard test for determination.** The gray ovals represent embryos at early stages of development. The cells to the right normally develop into head structures, whereas the cells to the left usually form tail structures. If prospective tail cells from an early embryo are transplanted to the opposite end of a host embryo, they develop according to their new position into head structures. These cells are not determined. At later stages of development, the tail cells are determined since they now develop into tail structures after transplantation into the opposite end of a host embryo!

additional cell interactions have occurred, determination has taken place, and the cell will develop as neural tissue no matter where it is transplanted.

Determination often takes place in stages, with a cell first becoming partially committed, acquiring positional labels that reflect its location in the embryo. These labels can have a great influence on how the pattern of the body subsequently develops. In a chicken embryo, tissue at the base of the leg bud normally gives rise to the thigh. If this tissue is transplanted to the tip of the identical-looking wing bud, which would normally give rise to the wing tip, the transplanted tissue will develop into a toe rather than a thigh. The tissue has already been determined as leg, but it is not yet committed to being a particular part of the leg. Therefore, it can be influenced by the positional signaling at the tip of the wing bud to form a tip (but in this case, a tip of leg).

### *The molecular basis of determination*

Cells initiate developmental changes by using transcription factors to change patterns of gene expression. When genes encoding these transcription factors are activated, one of their effects is to reinforce their own activation. This reinforcement makes the developmental switch deterministic, initiating a chain of events that leads down a particular developmental pathway.

Cells in which a set of regulatory genes have been activated may not actually undergo differentiation until some time later, when other factors interact with the regulatory protein and cause it to activate still other genes. Nevertheless, once the initial "switch" is thrown, the cell is fully committed to its future developmental path.

Cells become committed to follow a particular developmental pathway in one of two ways:

1. via the differential inheritance of cytoplasmic determinants, which are maternally produced and deposited into the egg during oogenesis; or
2. via cell–cell interactions.

The first situation can be likened to a person's social status being determined by who his or her parents are and what he or she has inherited. In the second situation, the person's social standing is determined by interactions with his or her neighbors. Clearly both can be powerful factors in the development and maturation of that individual.

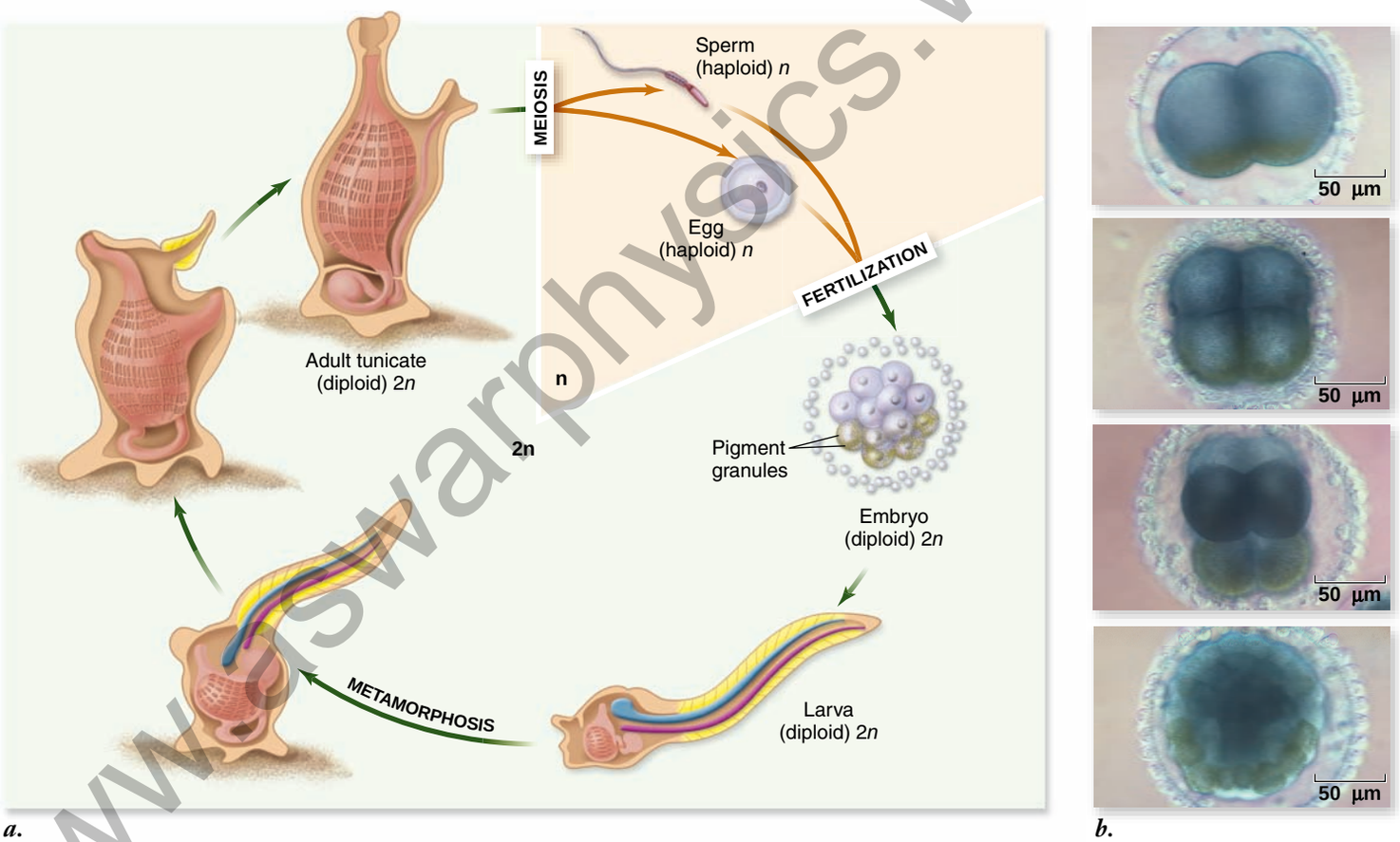



**Figure 19.5 Muscle determinants in tunicates.** *a.* The life cycle of a solitary tunicate. Muscle cells that move the tail of the swimming tadpole are arranged on either side of the notochord and nerve cord. The tail is lost during metamorphosis into the sedentary adult. *b.* The egg of the tunicate *Styela* contains bright yellow pigment granules. These become asymmetrically localized in the egg following fertilization, and cells that inherit the yellow granules during cleavage will become the larval muscle cells. Embryos at the 2-cell, 4-cell, 8-cell, and 64-cell stages are shown. The tadpole tail will grow out from the lower region of the embryo in the bottom panel.

## Determination can be due to cytoplasmic determinants

Many invertebrate embryos provide good visual examples of cell determination through the differential inheritance of cytoplasmic determinants. Tunicates are marine invertebrates (see chapter 35), and most adults have simple, saclike bodies that are attached to the underlying substratum. Tunicates are placed in the phylum Chordata, however, due to the characteristics of their swimming, tadpolelike larval stage, which has a dorsal nerve cord and notochord (figure 19.5*a*). The muscles that move the tail develop on either side of the notochord.

In many tunicate species, colored pigment granules become asymmetrically localized in the egg following fertilization and subsequently segregate to the tail muscle cell progenitors during cleavage (figure 19.5*b*). When these pigment granules are shifted experimentally into other cells that normally do not develop into muscle, their fate is changed and they become muscle cells. Thus, the molecules that flip the switch for muscle development appear to be associated with the pigment granules.

The next step is to determine the identity of the molecules involved. Experiments indicate that the female parent provides the egg with mRNA encoded by the *macho-1* gene. The elimination of *macho-1* function leads to a loss of tail muscle in the tadpole, and the misexpression of *macho-1* mRNA leads to the formation of additional (ectopic) muscle cells from nonmuscle lineage cells. The *macho-1* gene product has been shown to be a transcription factor that can activate the expression of several muscle-specific genes.

## Induction can lead to cell differentiation

In chapter 9, we examined a variety of ways by which cells communicate with one another. We can demonstrate the importance of cell–cell interactions in development by separating the cells of an early frog embryo and allowing them to develop independently.

Under these conditions, blastomeres from one pole of the embryo (the "animal pole") develop features of ectoderm, and blastomeres from the opposite pole of the embryo (the "vegetal pole") develop features of endoderm. None of the two separated groups of cells ever develop features characteristic of mesoderm, the third main cell type. If animal-pole cells and vegetal-pole cells are placed next to each other, however, some of the animal-pole cells develop as mesoderm. The interaction between the two cell types triggers a switch in the developmental path of these cells. This change in cell fate due to interaction with an adjacent cell is called **induction.** Signaling molecules act to alter gene expression in the target cells, in this case, some of the animal-pole cells.

Another example of inductive cell interactions is the formation of the notochord and mesenchyme, a specific tissue, in tunicate embryos. Muscle, notochord, and mesenchyme all arise from mesodermal cells that form at the vegetal margin of the 32-cell stage embryo. These prospective mesodermal cells receive signals from the underlying endodermal precursor cells that lead to the formation of notochord and mesenchyme (figure 19.6).

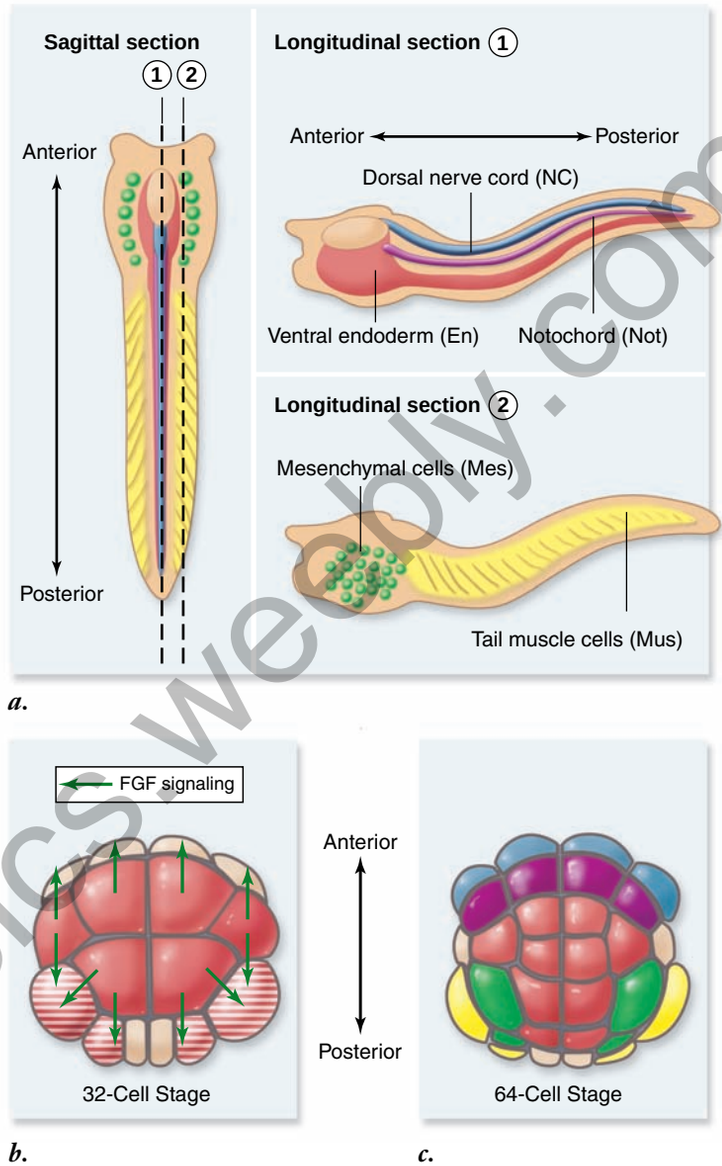



**Figure 19.6 Inductive interactions contribute to cell fate specification in tunicate embryos.** ***a.*** Internal structures of a tunicate larva. To the left is a sagittal section through the larva with dotted lines indicating two longitudinal sections. Section 1, through the midline of a tadpole, shows the dorsal nerve cord (NC), the underlying notochord (Not) and the ventral endoderm cells (En). Section 2, a more lateral section, shows the mesenchymal cells (Mes) and the tail muscle cells (Mus). ***b.*** View of the 32-cell stage looking up at the endoderm precursor cells. FGF secreted by these cells is indicated with light-green arrows. Only the surfaces of the marginal cells that directly border the endoderm precursor cells bind FGF signal molecules. Note that the posterior vegetal blastomeres also contain the *macho-1* determinants (red and white stripes). ***c.*** Cell fates have been fixed by the 64-cell stage. Colors are as in ***(a).*** Cells on the anterior margin of the endoderm precursor cells become notochord and nerve cord, respectively, whereas cells that border the posterior margin of the endoderm cells become mesenchyme and muscle cells, respectively.

The chemical signal is a member of the *fibroblast growth factor (FGF)* family of signaling molecules. It induces the overlying marginal zone cells to differentiate into either notochord (anterior) or mesenchyme (posterior). The FGF receptor on the marginal zone cells is a receptor tyrosine kinase that signals through a MAP kinase cascade to activate a transcription factor that turns on gene expression resulting in differentiation (figure 19.7).

This example is also a case of two cells responding differently to the same signal. The presence or absence of the *macho-1* muscle determinant discussed earlier controls this difference in cell fate. In the presence of *macho-1*, cells differentiate into mesenchyme; in its absence, cells differentiate into notochord. Thus, the combination of *macho-1* and FGF signaling leads to four different cell types (see figure 19.7)

## Stem cells can divide and produce cells that differentiate

It is important, both during development, and even in the adult animal, to have cells set aside that can divide but are not determined for only a single cell fate. We call cells that are capable of continued division but that can also give rise to differentiated cells, **stem cells.** These cells can be characterized based on the degree to which they have become determined.

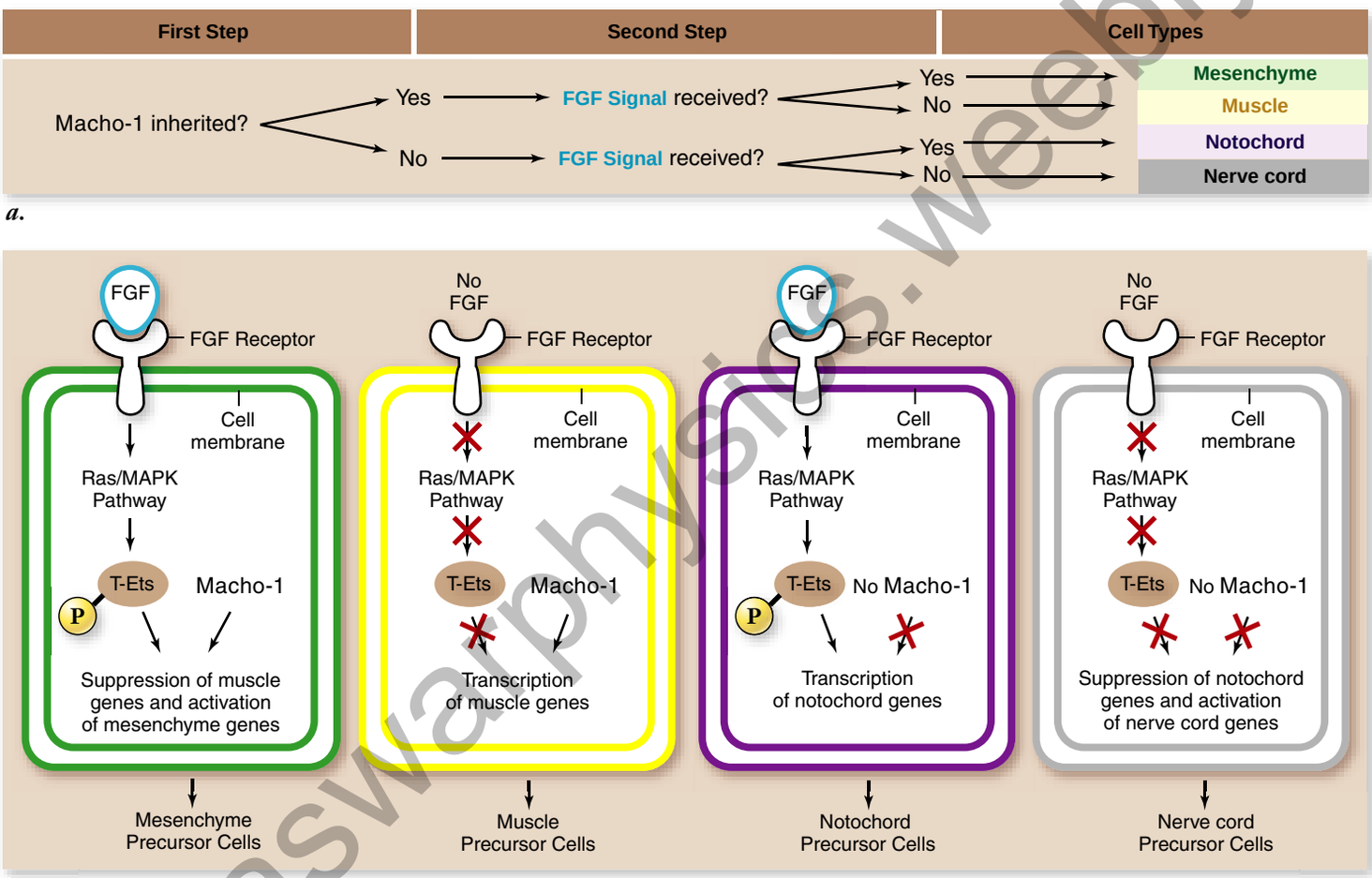



**Figure 19.7 Model for cell fate specification by Macho-1 muscle determinant and FGF signaling.** ***a.*** Two-step model of cell fate specification in vegetal marginal cells of the tunicate embryo. The first step is inheritance (or not) of muscle *macho-1* mRNA. The second step is FGF signaling from the underlying endoderm precursor cells. ***b.*** Posterior vegetal margin cells inherit *macho-1* mRNA. Signaling by FGF activates a Ras/MAP kinase pathway that produces the transcription factor T-Ets. Macho-1 protein and T-Ets suppress muscle-specific genes and turn on mesenchyme specific genes *(green cells).* In cells with Macho-1 that do not receive the FGF signal, Macho-1 alone turns on muscle-specific cells *(yellow cells).* Anterior vegetal margin cells do not inherit *macho-1* mRNA. If these cells receive the FGF signal, T-Ets turns on notochord-specific genes *(purple cells).* In cells that lack Macho-1 and FGF, notochord-specific genes are suppressed and nerve cord-specific genes are activated *(gray cells).*

**Inquiry question**

**What dictates whether Macho-1 acts as a transcriptional repressor or a transcriptional activator?**

At one extreme, we call a cell that can give rise to any tissue in an organism **totipotent.** In mammals, the only cells that can give rise to both the embryo and the extraembryonic membranes are the zygote and early blastomeres from the first few cell divisions. Cells that can give rise to all of the cells in the organism's body are called **pluripotent.** A stem cell that can give rise to a limited number of cell types, such as the cells that give rise to the different blood cell types, are called **multipotent.** Then at the other extreme, **unipotent** stem cells give rise to only a single cell type, such as the cells that give rise to sperm cells in males.

## Embryonic stem cells are pluripotent cells derived from embryos

A form of pluripotent stem cells that has been derived in the laboratory are called embryonic stem cells (ES cells). These cells are made from mammalian embryos that have undergone the cleavage stage of development to produce a ball of cells called a blastocyst. The blastocyst consists of an outer ball of cells, the trophectoderm, which will become the placenta, and the inner cell mass that will go on to form the embryo (see chapter 54 for details). Embryonic stem cells can be isolated from the inner cell mass and grown in culture (figure 19.8). In mice, these cells have been studied extensively and have been shown to be able to develop into any type of cell in the tissues of the adult. However, these cells cannot give rise to the extraembryonic tissues that arise during development, so they are pluripotent, but not totipotent.

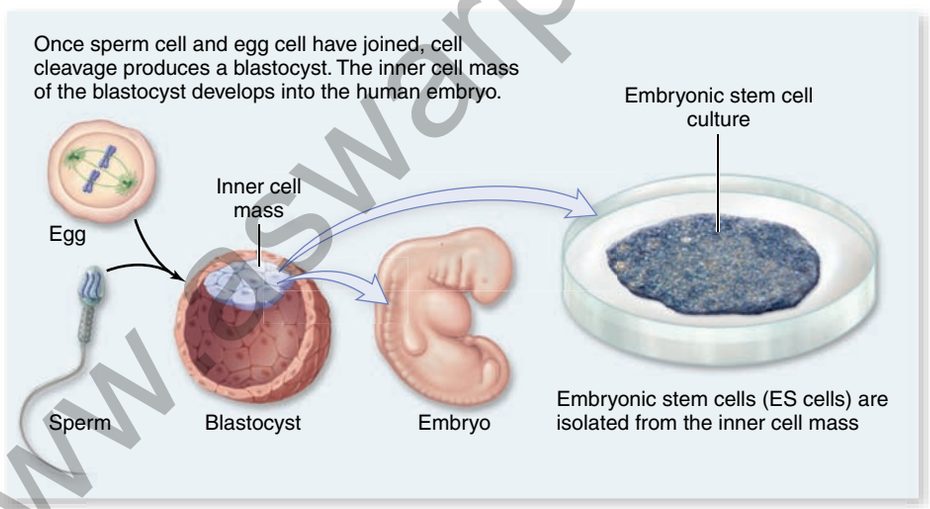
Once these cells were found in mice, it was only a matter of time before human ES cells were derived as well. In 1998, the first human ES cells (hES cells) were isolated and grown in culture. While there are differences between human and mouse ES cells, there are also substantial similarities. These embryonic stem cells hold great promise for regenerative medicine based on their potential to produce any cell type as described below. These cells have also been the source of much controversy and ethical discussion due to their embryonic origin.

### *Differentiation in culture*

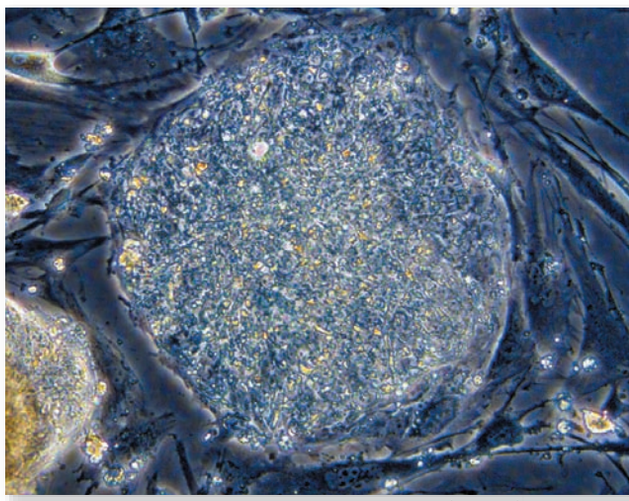
In addition to their possible therapeutic uses, ES cells offer a way to study the differentiation process in culture. The manipulation of these cells by additions to the culture media will allow us to tease out the factors involved in differentiation at the level of the actual cell undergoing the process. Early attempts at assessing differentiation in culture was plagued by the culture conditions. The medium in early experiments contained fetal calf serum (common in tissue culture), which is ill-defined, and varies lot-to-lot. More recently, more defined culture conditions have been found that allow greater reproducibility in controlling differentiation in culture.

Using more defined media, ES cells have been used to recapitulate in culture the early events in mouse development. Thus mouse ES cells can be used to first give rise to ectoderm, endoderm, and mesoderm, then these three cell types will give rise to the different cells each germ layer is determined to become. This work is in early stages but is tremendously exciting as it offers the promise of understanding the molecular cues that are involved in the stepwise determination of different cell types.

In humans, ES cells have been used to give rise to a variety of cell types in culture. For example, human ES cells have been shown to give rise to different kinds of blood cells in culture. Work is underway to produce hematopoeitic stem cells in culture, which could be used to replace such cells in patients with diseases that affect blood cells. Human ES cells have also been used to produce cardiomyocytes in culture. These cells could be used to replace damaged heart tissue after heart attacks.




*a.*




*b.*

**Figure 19.8 Isolation of embryonic stem cells.** ***a.*** Early cell divisions lead to the blastocyst stage that consists of an outer layer and an inner cell mass, which will go on to form the embryo. Embryonic stem cells (ES cells) can be isolated from this stage by disrupting the embryo and plating the cells. Stem cells removed from a six-day blastocyst can be established in culture and maintained indefinitely in an undifferentiated state. ***b.*** Human embryonic stem cells. This mass in the photograph is a colony of undifferentiated human embryonic stem cells being studied in the developmental biologist James Thomson's research lab at the University of Wisconsin–Madison.

### Learning Outcomes Review 19.3

**Cell differentiation is preceded by determination, where the cell becomes committed to a developmental pathway, but has not yet differentiated. Differential inheritance of cytoplasmic factors can cause determination and differentiation, as can interactions between neighboring cells (induction). Inductive changes are mediated by signaling molecules that trigger transduction pathways. Stem cells are able to divide indefinitely, and they can give rise to differentiated cells. Embryonic stem cells are pluripotent cells that can give rise to all adult structures.**

- ***How could you distinguish whether a cell becomes determined by induction or because of cytoplasmic factors?***

## 19.4 Nuclear Reprogramming

### Learning Outcomes

1. ***Define nuclear reprogramming and describe ways in which it has been accomplished.***
2. ***Differentiate between reproductive and therapeutic cloning.***

The study of the process of determination and differentiation leads quite naturally to questions about whether this process can be reversed. This is of interest both in terms of the experimental possibilities to understand the basic process, and the prospect of creating patient-specific populations of specific cell types to replace cells lost to disease or trauma. This has led to a fascinating path with many twists and turns that has accelerated in the recent past. We will briefly consider the history of this topic, then look at the most recent results available.

### Reversal of determination has allowed cloning

Experiments carried out in the 1950s showed that single cells from fully differentiated tissue of an adult plant could develop into entire, mature plants. The cells of an early cleavage stage mammalian embryo are also totipotent. When mammalian embryos naturally split in two, identical twins result. If individual blastomeres are separated from one another, any one of them can produce a completely normal individual. In fact, this type of procedure has been used to produce sets of four or eight identical offspring in the commercial breeding of particularly valuable lines of cattle.

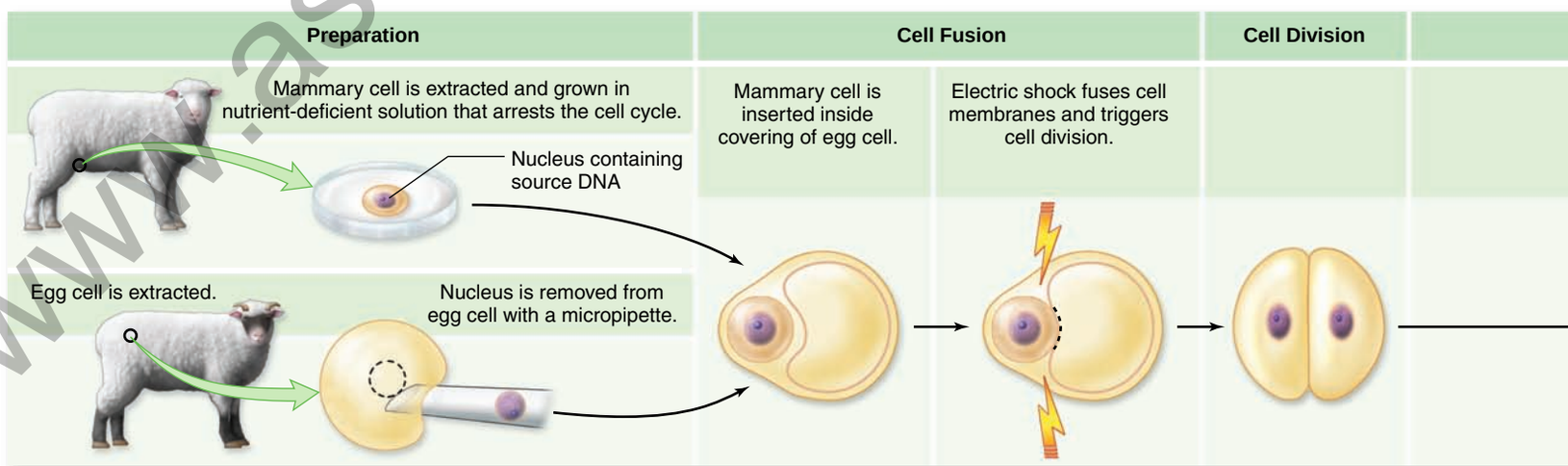
#### Early research in amphibians

An early question in developmental biology was whether the production of differentiated cells during development involved irreversible changes to cells. Experiments carried out in the 1950s by Briggs and King, and by John Gurdon in the 1960s and 1970s showed nuclei could be transplanted between cells. Using very fine pipettes (hollow glass tubes), these researchers sucked the nucleus out of a frog or toad egg and replaced the egg nucleus with a nucleus sucked out of a body cell taken from another individual.

The conclusions from these experiments are somewhat contradictory. On the one hand, cells do not appear to undergo any truly irreversible changes, such as loss of genes. On the other hand, the more differentiated the cell type, the less successful the nucleus in directing development when transplanted. This led to the concept of *nuclear reprogramming*, that is, a nucleus from a differentiated cell undergoes **epigenetic** changes that must be reversed to allow the nucleus to direct development. Epigenetic changes do not change a cell's DNA but are stable through cell divisions. The early work on amphibians showed that tadpoles' intestinal cell nuclei could be reprogrammed to produce viable adult frogs. These animals not only can be considered clones, but they show that tadpole nuclei can be completely reprogrammed. However, nuclei from adult differentiated cells could only be reprogrammed to produce tadpoles, but not viable, fertile adults. Thus this work showed that adult nuclei have remarkable developmental potential, but cannot be reprogrammed to be totipotent.

#### Early research in mammals

Given the work done in amphibians, much effort was put into nuclear transfer in mammals, primarily mice and cattle. Not only did this not result in reproducible production of cloned



animals, but this work led to the discovery of imprinting through the production of embryos with only maternal or paternal input (see chapter 13 for more information on imprinting). These embryos never developed, and showed different kinds of defects depending on whether the maternal or paternal genome was the sole contributor.

### *Successful nuclear transplant in mammals*

These results stood until a sheep was cloned using the nucleus from a cell of an early embryo in 1984. The key to this success was in picking a donor cell very early in development. This exciting result was soon replicated by others in a host of other organisms, including pigs and monkeys. Only early embryo cells seemed to work, however.

Geneticists at the Roslin Institute in Scotland reasoned that the egg and donated nucleus would need to be at the same stage of the cell cycle for successful development. To test this idea, they performed the following procedure (figure 19.9):

1. They removed differentiated mammary cells from the udder of a six-year-old sheep. The cells were grown in tissue culture, and then the concentration of serum nutrients was substantially reduced for five days, causing them to pause at the beginning of the cell cycle.
2. In parallel preparation, eggs obtained from a ewe were enucleated.
3. Mammary cells and egg cells were surgically combined in a process called **somatic cell nuclear transfer (SCNT)** in January of 1996. Mammary cells and eggs were fused to introduce the mammary nucleus into egg.
4. Twenty-nine of 277 fused couplets developed into embryos, which were then placed into the reproductive tracts of surrogate mothers.
5. A little over five months later, on July 5, 1996, one sheep gave birth to a lamb named Dolly, the first clone generated from a fully differentiated animal cell.

Dolly matured into an adult ewe, and she was able to reproduce the old-fashioned way, producing six lambs. Thus, Dolly established beyond all dispute that determination in animals is reversible—that with the right techniques, the nucleus of a fully differentiated cell *can* be reprogrammed to be totipotent.

## Reproductive cloning has inherent problems

The term **reproductive cloning** refers to the process just described, in which scientists use SCNT to create an animal that is genetically identical to another animal. Since Dolly's birth in 1997, scientists have successfully cloned one or more cats, rabbits, rats, mice, cattle, goats, pigs, and mules. All of these procedures used some form of adult cell.

### *Low success rate and age-associated diseases*

The efficiency in all reproductive cloning is quite low—only 3–5% of adult nuclei transferred to donor eggs result in live births. In addition, many clones that are born usually die soon thereafter of liver failure or infections. Many become oversized, a condition known as *large offspring syndrome (LOS)*. In 2003, three of four cloned piglets developed to adulthood, but all three suddenly died of heart failure at less than 6 months of age.

Dolly herself was euthanized at the relatively young age of six. Although she was put down because of virally induced lung cancer, she had been diagnosed with advanced-stage arthritis a year earlier. Thus, one difficulty in using genetic engineering and cloning to improve livestock is production of enough healthy animals.

### *Lack of imprinting*

The reason for these problems lies in a phenomenon discussed in chapter 13: *genomic imprinting*. Imprinted genes are expressed differently depending on parental origin—that is, they are turned off in either egg or sperm, and this "setting" continues through development into the adult. Normal mammalian development depends on precise genomic imprinting.

The chemical reprogramming of the DNA, which occurs in adult reproductive tissue, takes months for sperm and years for eggs. During cloning, by contrast, the reprogramming of the donor DNA must occur within a few hours. The organization of the chromatin in a somatic cell is also quite different from that in a newly fertilized egg. Significant chromatin remodeling of the transferred donor nucleus must also occur if the cloned embryo is to survive. Cloning fails because there is likely not enough time in these few hours to get the remodeling and reprogramming jobs done properly.

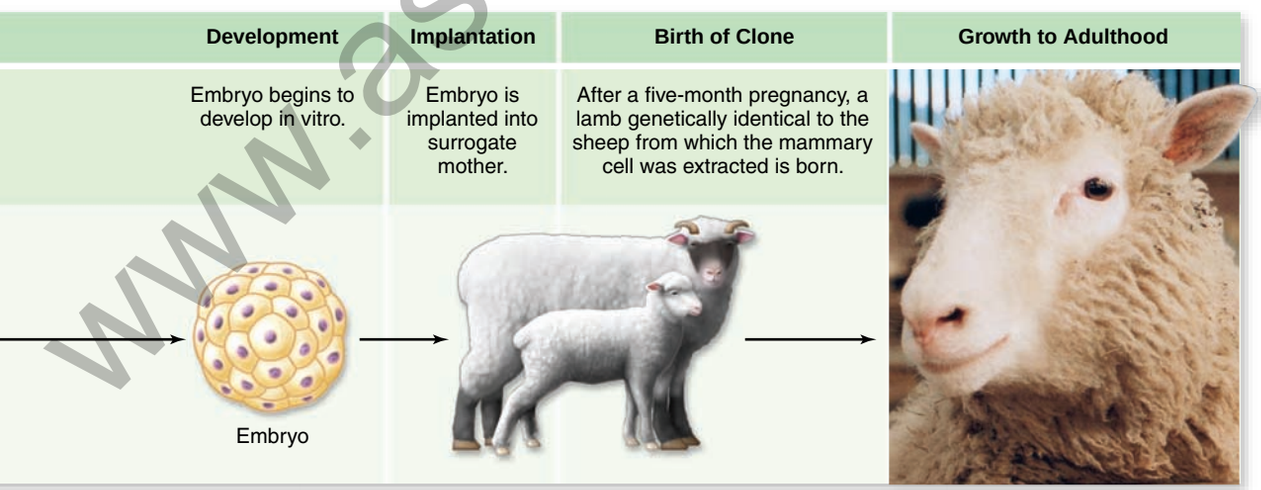



**Figure 19.9 Proof that determination in animals is reversible.** Scientists combined a nucleus from an adult mammary cell with an enucleated egg cell to successfully clone a sheep, named Dolly, who grew to be a normal adult and bore healthy offspring. This experiment, the first successful cloning of an adult animal, shows that a differentiated adult cell can be used to drive all of development.

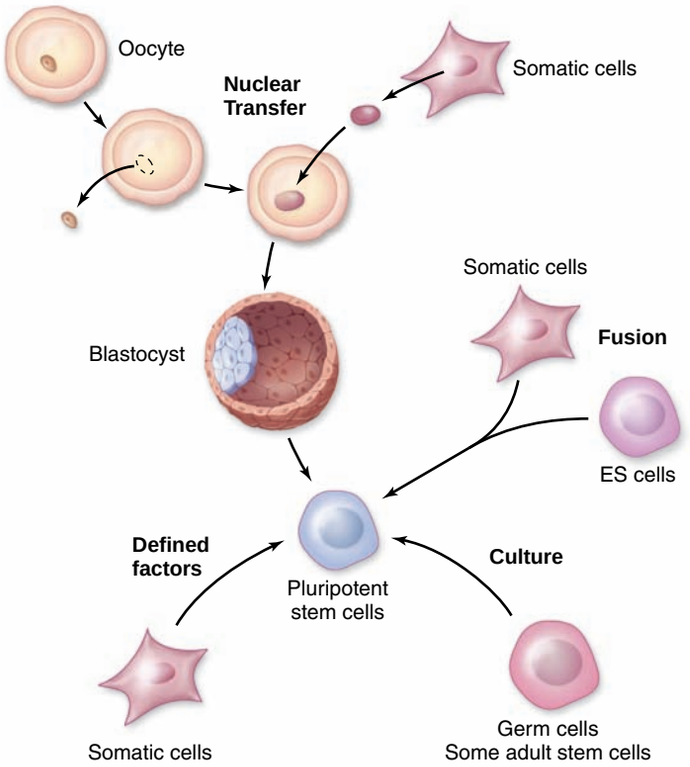



**Figure 19.10 Methods to reprogram adult cell nuclei.**
Cells taken from adult organisms can be reprogrammed to pluripotent cells in a number of different ways. Nuclei from somatic cells can be transplanted into oocytes as during cloning. Somatic cells can be fused to ES cells created by some other means. Germ cells, and some adult stem cells, after prolonged culture appear to be reprogrammed. Recent work has shown that somatic cells in culture can be reprogrammed by introduction of specific factors.

## Nuclear reprogramming has been accomplished by use of defined factors

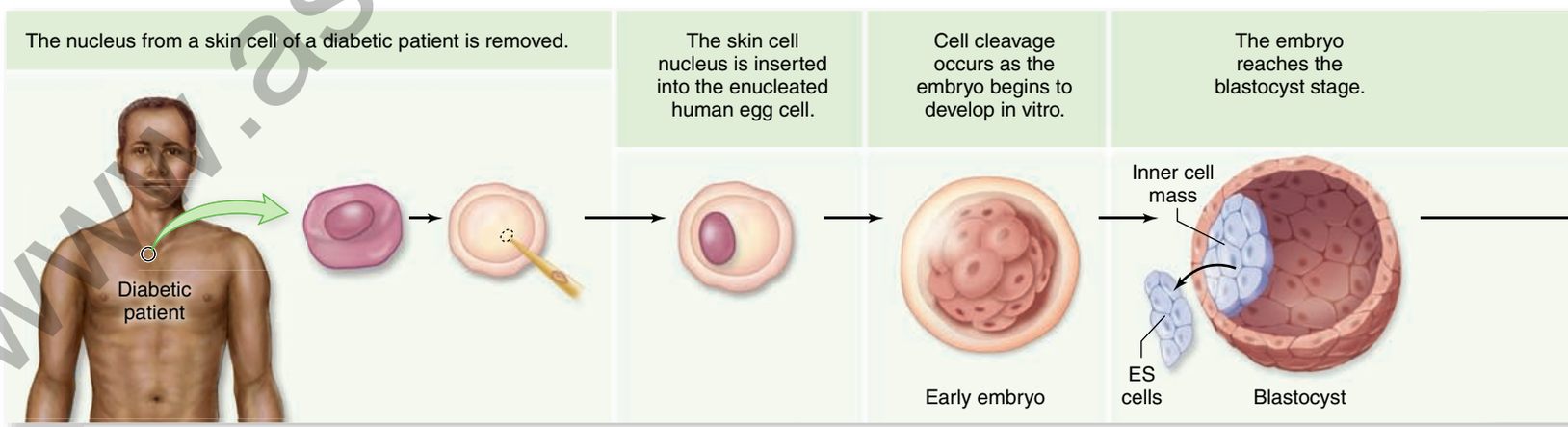
Stimulated by the discovery of ES cells and success in the reproductive cloning of mammals, much work has been put into trying to find ways to reprogram adult cells to become pluripotent cells without the use of embryos (figure 19.10). One approach was to fuse an ES cell to a differentiated cell. These fusion experiments showed that the nucleus of the differentiated cell could be reprogrammed by exposure to ES cell cytoplasm. Of course, the resulting cells are tetraploid (4 copies of the genome), which limits their experimental and practical utility. Another line of research showed that primordial germ cells explanted into culture can give rise to cells that act similar to ES cells after extended time in culture. There are also reports that some adult stem cells become pluripotent cells with prolonged culture, but this is still controversial.

All of these different lines of inquiry showed that reprogramming of somatic nuclei was possible. The next obvious step was to reprogram nuclei using defined factors. While it was assumed that this was possible, it was not accomplished until 2006 when the genes for four different transcription factors, Oct4, Sox2, c-Myc, and Klf4 were introduced into fibroblast cells in culture. These cells were then selected for expression of a gene that is a target for Oct4 and Sox2, and these cells appear to be pluripotent. These were named induced pluripotent stem cells, or iPS cells. This protocol has been improved by selection for a different target gene, Nanog. These Nanog expressing iPS cells appear to be similar to ES cells in terms of developmental potential, as well as gene expression pattern.

This technology has now been used to construct ES cells from patients with the inherited neurological disorder spinal muscular atrophy. These ES cells will differentiate in culture into motor neurons that show the phenotype expected for the disease. The ability to derive disease specific stem cells will be an incredible advance for researchers studying such diseases. This will allow the creation of in vitro systems to study directly the cells affected by genetic diseases, and to screen for possible therapeutics.

Pluripotent cell types have potential for therapeutic applications. One way to solve the problem of graft rejection, such as in skin grafts in severe burn cases, is to produce patient-specific lines of embryonic stem cells. Early in 2001, a research team at Rockefeller University devised a way to accomplish this feat.

First, skin cells are isolated; then, using the same SCNT procedure that created Dolly, an embryo is assembled. After removing the nucleus from the skin cell, they insert it into an egg whose nucleus has already been removed. The egg with



its skin cell nucleus is allowed to form a blastocyst stage embryo. This artificial embryo is then destroyed, and its cells are used as embryonic stem cells for transfer to injured tissue (figure 19.11).

Using this procedure, termed **therapeutic cloning,** the researchers succeeded in converting cells from the tail of a mouse into the dopamine-producing cells of the brain that are lost in Parkinson disease. Therapeutic cloning successfully addresses the key problem that must be solved before stem cells can be used to repair human tissues damaged by heart attack, nerve injury, diabetes, or Parkinson disease—the problem of immune acceptance. Since stem cells are cloned from a person's own tissues in therapeutic cloning, they pass the immune system's "self" identity check, and the body readily accepts them.

These early attempts at therapeutic cloning may become obsolete before they are even refined with the work described above on iPS cells. The potential to produce pluripotent cells from adult skin cells removes the ethical problems of embryo destruction, and the practical problem of the requirement for oocytes for therapeutic cloning. However, these cell types are not without problems of their own. Two of the genes introduced to reprogram the nuclei of cells are oncogenes, and although these experiments have been reproduced without c-Myc, the efficiency is greatly reduced. These also require the introduction of new DNA, which can induce mutations by integration into the genome.

### *Learning Outcomes Review 19.4*

**Cloning has long been practiced in plants. In animals, cells from early-stage embryos are also totipotent, but attempts to use adult nuclei for cloning led to mixed results. The nucleus of a differentiated cell requires reprogramming to be totipotent. This appears to be necessary at least in part because of genomic imprinting. Nuclei may be reprogrammed by fusion with an embryonic stem cell, which produces a tetraploid cell, or through the introduction of four important transcription factors. That reprogramming is possible was shown by reproductive cloning via somatic cell nuclear transfer (SCNT). In therapeutic cloning, the goal is to produce replacement tissue using a patient's own cells.**

- ***What changes must occur to produce a totipotent cell from a differentiated nucleus?***

## 19.5 *Pattern Formation*

### *Learning Outcomes*

1. ***Describe A/P axis formation in* Drosophila.**
2. ***Describe D/V axis formation in* Drosophila.**
3. ***Explain the importance of homeobox-containing genes in development.***

For cells in multicellular organisms to differentiate into appropriate cell types, they must gain information about their relative locations in the body. All multicellular organisms seem to use positional information to determine the basic pattern of body compartments and, thus, the overall architecture of the adult body. This positional information then leads to intrinsic changes in gene activity, so that cells ultimately adopt a fate appropriate for their location.

Pattern formation is an unfolding process. In the later stages, it may involve morphogenesis of organs (to be discussed later), but during the earliest events of development, the basic body plan is laid down, along with the establishment of the anterior–posterior (A/P, head-to-tail) axis and the dorsal–ventral (D/V, back-to-front) axis. Thus, pattern formation can be considered the process of taking a radially symmetrical cell and imposing two perpendicular axes to define the basic body plan, which in this way becomes bilaterally symmetrical. Developmental biologists use the term **polarity** to refer to the acquisition of axial differences in developing structures.

The fruit fly *Drosophila melanogaster* is the best understood animal in terms of the genetic control of early patterning. We will concentrate on the Drosophila system here, and later in chapter 54 we will examine axis formation in vertebrates in the context of their overall development.

A hierarchy of gene expression that begins with maternally expressed genes controls the development of *Drosophila*. To understand the details of these gene interactions, we first need to briefly review the stages of *Drosophila* development.

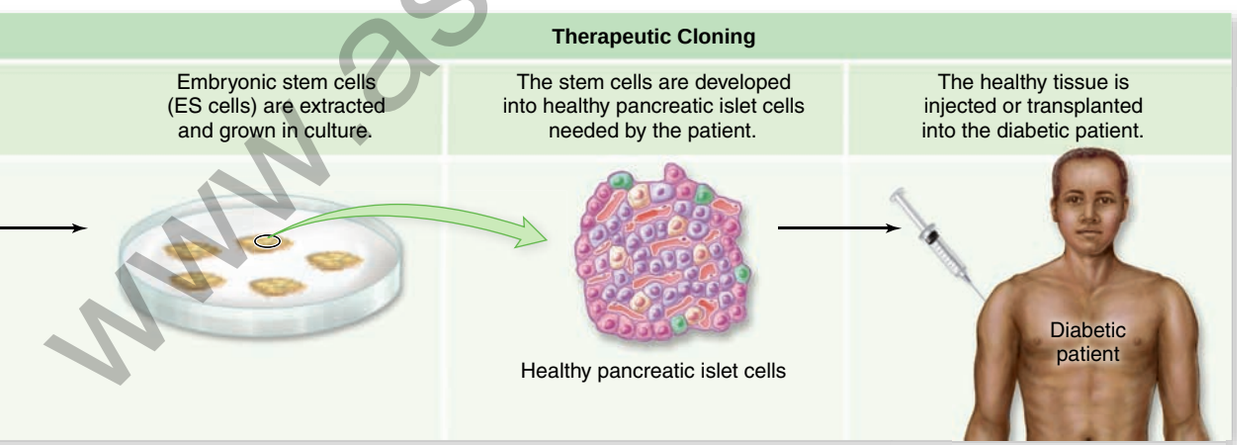



**Figure 19.11 How human embryos might be used for therapeutic cloning.** In therapeutic cloning, after initial stages to reproductive cloning, the embryo is broken apart and its embryonic stem cells are extracted. These are grown in culture and used to replace the diseased tissue of the individual who provided the DNA. This is useful only if the disease in question is not genetic as the stem cells are genetically identical to the patient.

## *Drosophila* embryogenesis produces a segmented larva

*Drosophila* and many other insects produce two different kinds of bodies during their development: the first, a tubular eating machine called a **larva,** and the second, an adult flying sex machine with legs and wings. The passage from one body form to the other, called **metamorphosis,** involves a radical shift in development (figure 19.12). In this chapter, we concentrate on the process of going from a fertilized egg to a larva, which is termed *embryogenesis.*

### *Prefertilization maternal contribution*

The development of an insect like *Drosophila* begins before fertilization, with the construction of the egg. Specialized *nurse cells* that help the egg grow move some of their own maternally encoded mRNAs into the maturing oocyte (figure 19.12*a*).

Following fertilization, the maternal mRNAs are transcribed into proteins, which initiate a cascade of sequential gene activations. Embryonic nuclei do not begin to function (that is, to direct new transcription of genes) until approximately 10 nuclear divisions have occurred. Therefore, the action of maternal, rather than zygotic, genes determines the initial course of *Drosophila* development.

### *Postfertilization events*

After fertilization, 12 rounds of nuclear division without cytokinesis produce about 4000 nuclei, all within a single cytoplasm. All of the nuclei within this **syncytial blastoderm** (figure 19.12*b*) can freely communicate with one another, but nuclei located in different sectors of the egg encounter different maternal products.

Once the nuclei have spaced themselves evenly along the surface of the blastoderm, membranes grow between them to form the **cellular blastoderm.** Embryonic folding and primary tissue development soon follow, in a process fundamentally similar to that seen in vertebrate development. Within a day of fertilization, embryogenesis creates a segmented, tubular body—which is destined to hatch out of the protective coats of the egg as a larva.

## Morphogen gradients form the basic body axes in *Drosophila*

Pattern formation in the early *Drosophila* embryo requires positional information encoded in labels that can be read by cells. The unraveling of this puzzle, work that earned the 1995 Nobel Prize for researchers Christiane Nüsslein-Volhard and Eric Wieschaus, is summarized in figure 19.13. We now know that two different genetic pathways control the establishment of A/P and D/V polarity in *Drosophila*.

### *Anterior–posterior axis*

Formation of the A/P axis begins during maturation of the oocyte and is based on opposing gradients of two different proteins: **Bicoid** and **Nanos.** These protein gradients are established by an interesting mechanism.

Nurse cells in the ovary secrete maternally produced *bicoid* and *nanos* mRNAs into the maturing oocyte where they are differentially transported along microtubules to opposite poles of the oocyte (figure 19.14*a*). This differential transport comes

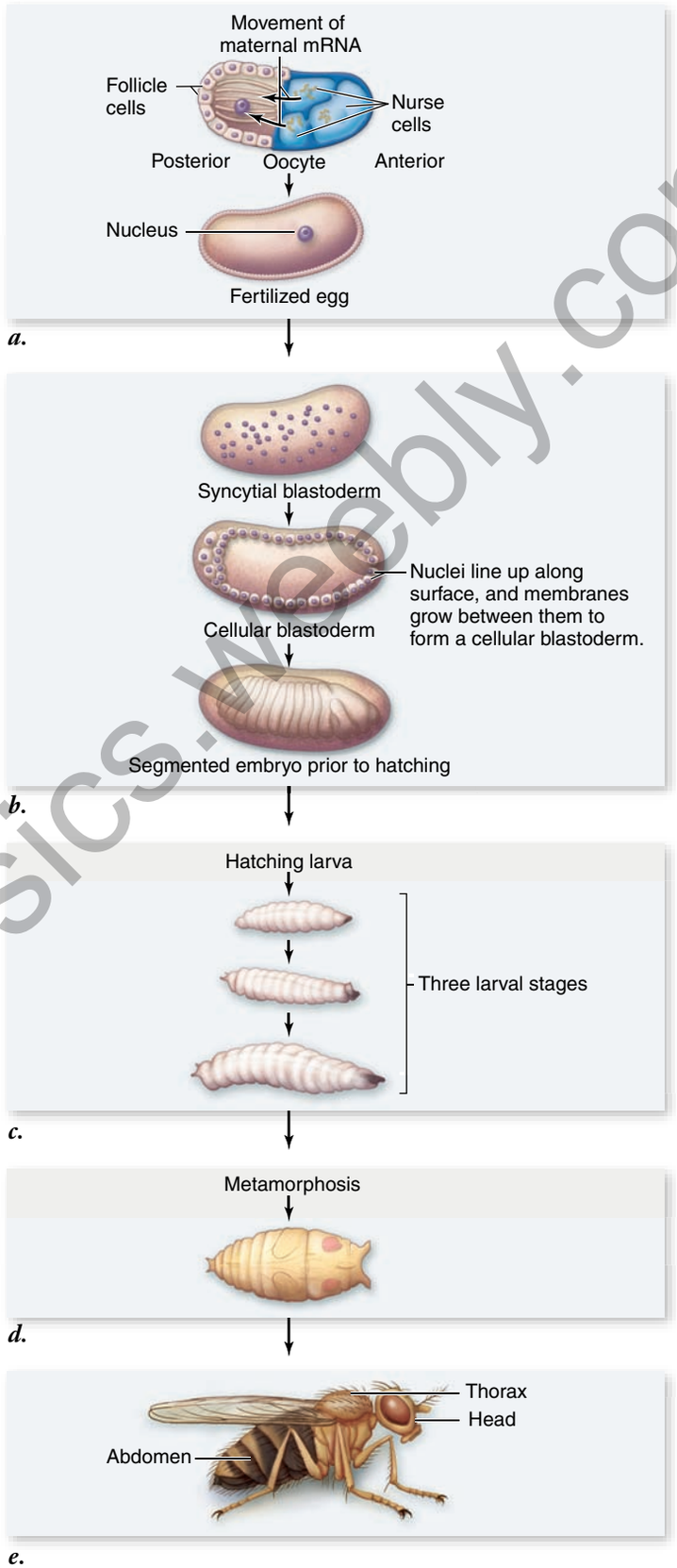
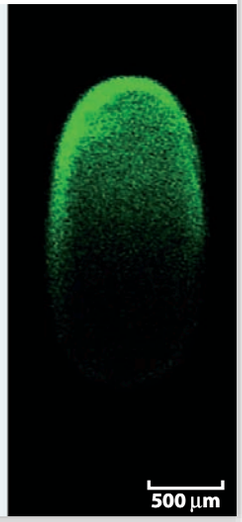



**Figure 19.12 The path of fruit fly development.** Major stages in the development of *Drosophila melanogaster* include formation of the ***(a)*** egg, ***(b)*** syncytial and cellular blastoderm, ***(c)*** larval instars, ***(d)*** pupa and metamorphosis into a ***(e)*** sexually mature adult.

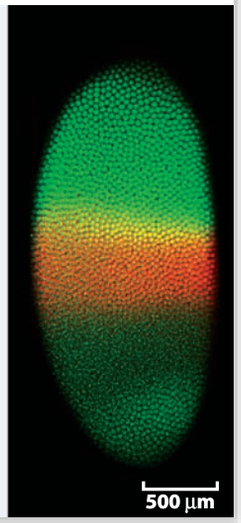
### Establishing the Polarity of the Embryo

Fertilization of the egg triggers the production of Bicoid protein from maternal RNA in the egg. The Bicoid protein diffuses through the egg, forming a gradient. This gradient determines the polarity of the embryo, with the head and thorax developing in the zone of high concentration (*green* fluorescent dye in antibodies that bind bicoid protein allows visualization of the gradient).




### Setting the Stage for Segmentation

About 2½ hours after fertilization, Bicoid protein turns on a series of brief signals from so-called gap genes. The gap proteins act to divide the embryo into large blocks. In this photo, fluorescent dyes in antibodies that bind to the gap proteins Krüppel (*orange*) and Hunchback (*green*) make the blocks visible; the region of overlap is yellow.




### Laying Down the Fundamental Regions

About 0.5 hr later, the gap genes switch on the "pair-rule" genes, which are each expressed in seven stripes. This is shown for the pair-rule gene *hairy*. Some pair-rule genes are only required for even-numbered segments while others are only required for odd numbered segments.


### Forming the Segments

The final stage of segmentation occurs when a "segment-polarity" gene called *engrailed* divides each of the seven regions into halves, producing 14 narrow compartments. Each compartment corresponds to one segment of the future body. There are three head segments (H, *bottom right*), three thoracic segments (T, *upper right*), and eight abdominal segments (A, *from top right to bottom left*).

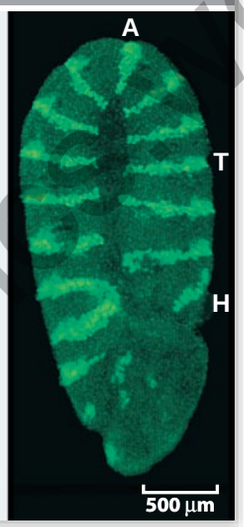
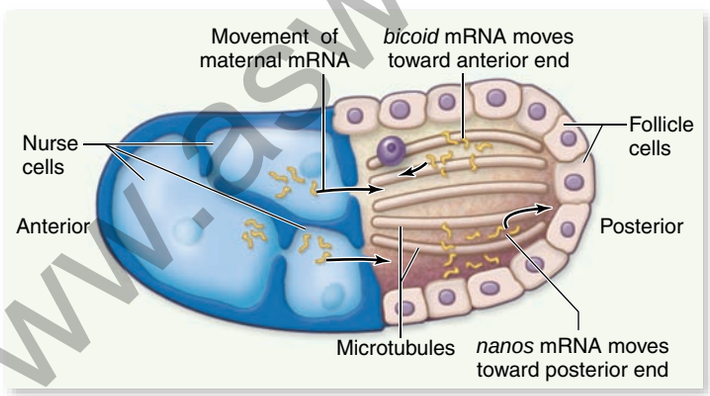
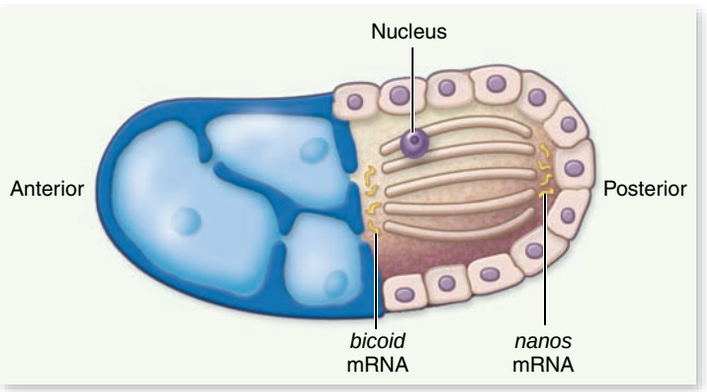



**Figure 19.13 Body organization in an early *Drosophila* embryo.** In these fluorescent microscope images by 1995 Nobel laureate Christiane Nüsslein-Volhard and Sean Carroll, we watch a *Drosophila* egg pass through the early stages of development, in which the basic segmentation pattern of the embryo is established. The proteins in the photographs were made visible by binding fluorescent antibodies to each specific protein.




*a.*




*b.*

**Figure 19.14 Specifying the A/P axis in *Drosophila* embryos I.** ***a.*** In the ovary, nurse cells secrete maternal mRNAs into the cytoplasm of the oocyte. Clusters of microtubules direct oocyte growth and maturation. Motor proteins travel along the microtubules transporting molecules in two directions. *Bicoid* mRNAs are transported toward the anterior pole of the oocyte, *nanos* mRNA is transported toward the posterior pole of the oocyte. ***b.*** A mature oocyte, showing localization of *bicoid* mRNAs to the anterior pole and *nanos* mRNAs to the posterior pole.

about due to the use of different motor proteins to move the two mRNAs. The *bicoid* mRNA then becomes anchored in the cytoplasm at the end of the oocyte closest to the nurse cells, and this end will develop into the anterior end of the embryo. *Nanos* mRNA becomes anchored to the opposite end of the oocyte, which will become the posterior end of the embryo. Thus, by the end of oogenesis, the *bicoid* and *nanos* mRNAs are already set to function as cytoplasmic determinants in the fertilized egg (figure 19.14*b*).

Following fertilization, translation of the anchored mRNA and diffusion of the proteins away from their respective sites of synthesis create opposing gradients of each protein: Highest levels of Bicoid protein are at the anterior pole of the embryo (figure 19.15*a*), and highest levels of the Nanos protein are at the posterior pole. Concentration gradients of soluble molecules can specify different cell fates along an axis, and proteins that act in this way, like Bicoid and Nanos, are called **morphogens.**

The Bicoid and Nanos proteins control the translation of two other maternal messages, *hunchback* and *caudal*, that encode transcription factors. **Hunchback** activates genes required for the formation of anterior structures, and **Caudal** activates genes required for the development of posterior (abdominal) structures. The *hunchback* and *caudal* mRNAs are evenly distributed across the egg (figure 19.15*b*), so how is it that proteins translated from these mRNAs become localized?

The answer is that Bicoid protein binds to and inhibits translation of *caudal* mRNA. Therefore, *caudal* is only translated in the posterior regions of the egg where Bicoid is absent. Similarly, Nanos protein binds to and prevents translation of the *hunchback* mRNA. As a result, *hunchback* is only translated in the anterior regions of the egg (figure 19.15*c*). Thus, shortly after fertilization, four protein gradients exist in the embryo: anterior–posterior gradients of Bicoid and Hunchback proteins, and posterior–anterior gradients of Nanos and Caudal proteins (figure 19.15*c*).

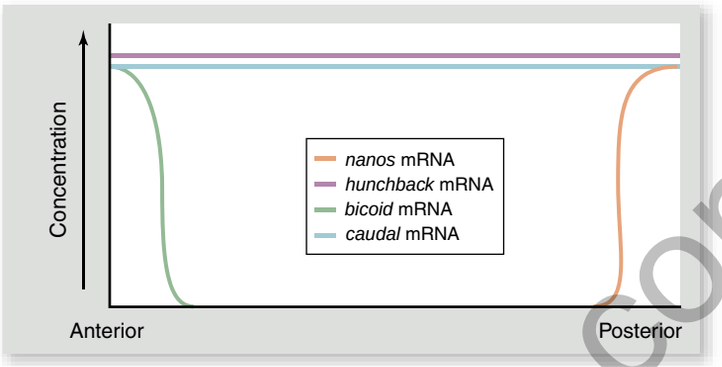
### *Dorsal–ventral axis*

The dorsal–ventral axis in *Drosophila* is established by actions of the *dorsal* gene product. Once again the process begins in the ovary, when maternal transcripts of the *dorsal* gene are put into the oocyte. However, unlike *bicoid* or *nanos,* the *dorsal* mRNA does not become asymmetrically localized. Instead, a series of steps are required for Dorsal to carry out its function.

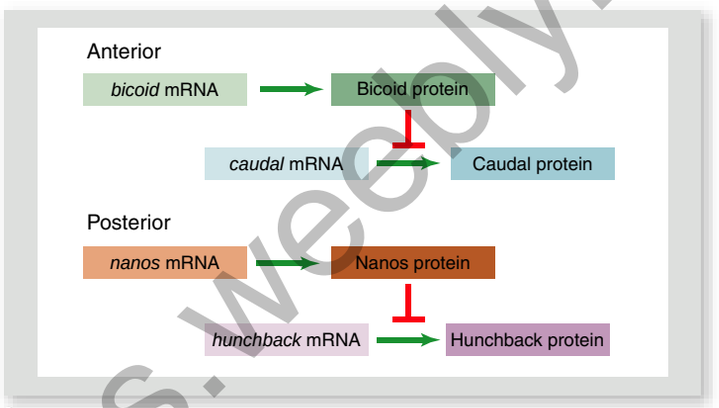
First, the oocyte nucleus, which is located to one side of the oocyte, synthesizes *gurken* mRNA. The *gurken* mRNA then accumulates in a crescent between the nucleus and the membrane on that side of the oocyte (figure 19.16*a*). This will be the future dorsal side of the embryo.

The Gurken protein is a soluble cell-signaling molecule, and when it is translated and released from the oocyte, it binds to receptors in the membranes of the overlying follicle cells (figure 19.16*b*). These cells then differentiate into a dorsal morphology. Meanwhile, no Gurken signal is released from the other side of the oocyte, and the follicle cells on that side of the oocyte adopt a ventral fate.

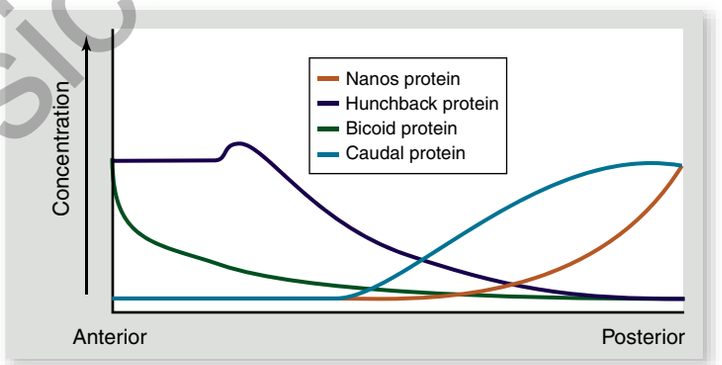
Following fertilization, a signaling molecule is differentially activated on the ventral surface of the embryo in a complex sequence of steps. This signaling molecule then binds to a membrane receptor in the ventral cells of the embryo and activates a signal transduction pathway in those cells. Activation of this pathway results in the selected transport of the Dorsal protein (which is everywhere) into ventral nuclei, forming a gradient along the D/V axis. The Dorsal protein levels are highest in the nuclei of ventral cells (figure 19.16*c*).




***a.*** Oocyte mRNAs




***b.*** After fertilization




***c.*** Early cleavage embryo proteins

**Figure 19.15 Specifying the A/P axis in *Drosophila* embryos II.** ***a.*** Unlike *bicoid* and *nanos*, *hunchback* and *caudal* mRNAs are evenly distributed throughout the cytoplasm of the oocyte. ***b.*** Following fertilization, *bicoid* and *nanos* mRNAs are translated into protein, making opposing gradients of each protein. Bicoid binds to and represses translation of *caudal* mRNAs (in anterior regions of the egg). Nanos binds to and represses translation of *hunchback* mRNAs (in posterior regions of the egg). ***c.*** Translation of *hunchback* mRNAs in anterior regions of the egg will create a Hunchback gradient that mirrors the Bicoid gradient. Translation of *caudal* mRNAs in posterior regions of the embryo will create a Caudal gradient that mirrors the Nanos gradient.

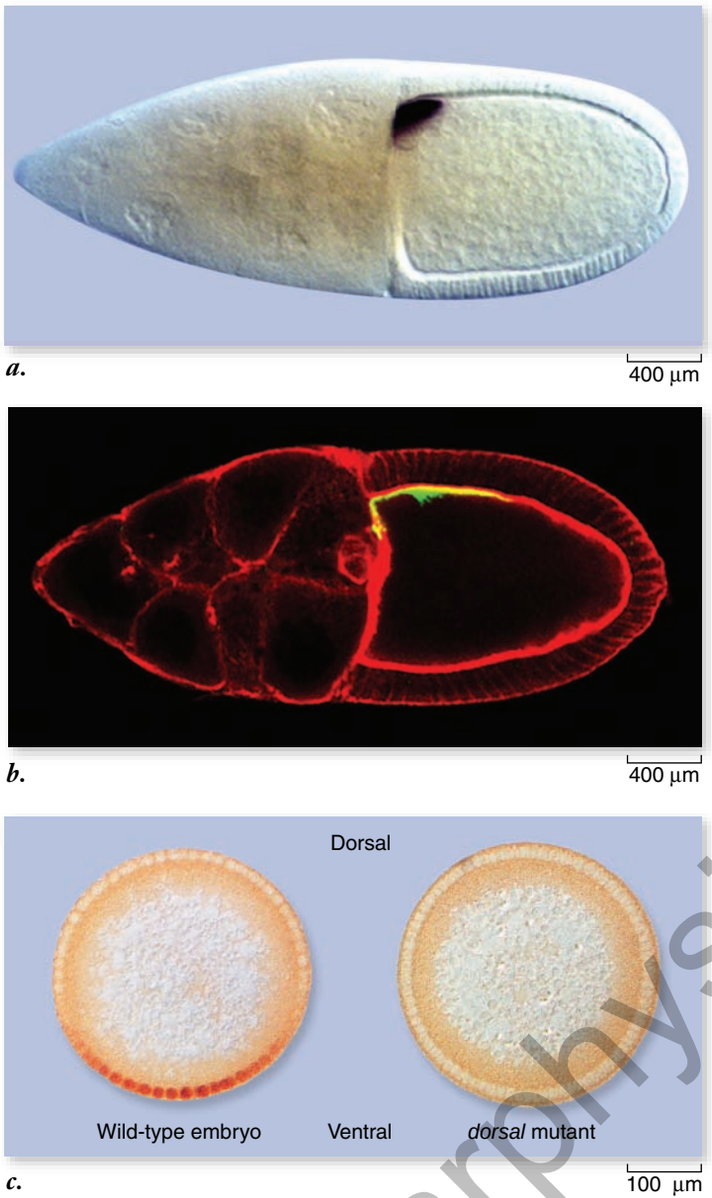



**Figure 19.16 Specifying the D/V axis in *Drosophila* embryos.** ***a.*** The *gurken* mRNA *(dark stain)* is concentrated between the oocyte nucleus (not visible) and the dorsal, anterior surface of the oocyte. ***b.*** In a more mature oocyte, Gurken protein *(yellow stain)* is secreted from the dorsal anterior surface of the oocyte, forming a gradient along the dorsal surface of the egg. Gurken then binds to membrane receptors in the overlying follicle cells. Double staining for actin *(red)* shows the cell boundaries of the oocyte, nurse cells, and follicle cells. ***c.*** For these images, cellular blastoderm stage embryos were cut in cross section to visualize the nuclei of cells around the perimeter of the embryos. Dorsal protein *(dark stain)* is localized in nuclei on the ventral surface of the blastoderm in a wild-type embryo *(left)*. The *dorsal* mutant on the right will not form ventral structures, and Dorsal is not present in ventral nuclei of this embryo.

The Dorsal protein is a transcription factor, and once it is transported into nuclei, it activates genes required for the proper development of ventral structures, simultaneously repressing genes that specify dorsal structures. Hence, the product of the *dorsal* gene ultimately directs the development of ventral structures.

(Note that many *Drosophila* genes are named for the mutant phenotype that results from a loss of function in that gene. A lack of *dorsal* function produces dorsalized embryos with no ventral structures.)

Although profoundly different mechanisms are involved, the unifying factor controlling the establishment of both A/P and D/V polarity in *Drosophila* is that *bicoid*, *nanos*, *gurken*, and *dorsal* are all maternally expressed genes. The polarity of the future embryo in both instances is therefore laid down in the oocyte using information coming from the maternal genome.

The preceding discussion simplifies events, but the outline is clear: Polarity is established by the creation of morphogen gradients in the embryo based on maternal information in the egg. These gradients then drive the expression of the zygotic genes that will actually pattern the embryo. This reliance on a hierarchy of regulatory genes is a unifying theme for all of development.

## The body plan is produced by sequential activation of genes

Let us now return to the process of pattern formation in *Drosophila* along the A/P axis. Determination of structures is accomplished by the sequential activation of three classes of **segmentation genes.** These genes create the hallmark segmented body plan of a fly, which consists of three fused head segments, three thoracic segments, and eight abdominal segments (see figure 19.12*e*).

To begin, Bicoid protein exerts its profound effect on the organization of the embryo by activating the translation and transcription of *hunchback* mRNA (which is the first mRNA to be transcribed after fertilization). *Hunchback* is a member of a group of nine genes called the **gap genes.** These genes map out the initial subdivision of the embryo along the A/P axis (see figure 19.13).

All of the gap genes encode transcription factors, which, in turn, activate the expression of eight or more **pair-rule genes.** Each of the pair-rule genes, such as *hairy*, produces seven distinct bands of protein, which appear as stripes when visualized with fluorescent reagents (see figure 19.13). These bands subdivide the broad gap regions and establish boundaries that divide the embryo into seven zones. When mutated, each of the pair-rule genes alters every other body segment.

All of the pair-rule genes also encode transcription factors, and they, in turn, regulate the expression of each other and of a group of nine or more **segment polarity genes.** The segment polarity genes are each expressed in 14 distinct bands of cells, which subdivide each of the seven zones specified by the pair-rule genes (see figure 19.13). The *engrailed* gene, for example, divides each of the seven zones established by *hairy* into anterior and posterior compartments. The segment polarity genes encode proteins that function in cell–cell signaling pathways. Thus, they function in inductive events—which occur *after* the syncytial blastoderm is divided into cells—to fix the anterior and posterior fates of cells within each segment.

In summary, within 3 hr after fertilization, a highly orchestrated cascade of segmentation gene activity transforms the broad gradients of the early embryo into a periodic, segmented

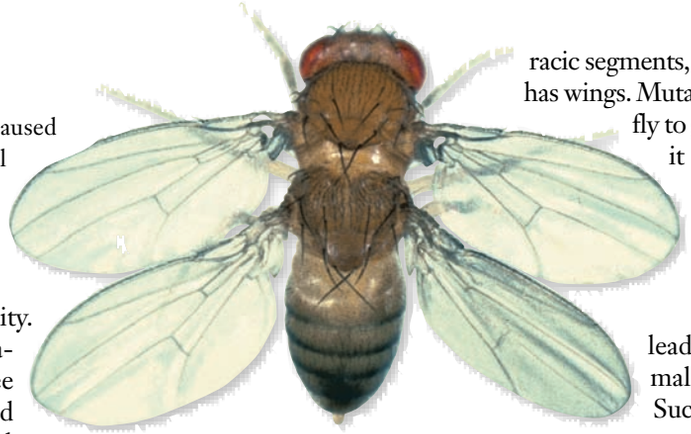


**Figure 19.17 Mutations in homeotic genes.** Three separate mutations in the bithorax complex caused this fruit fly to develop an additional second thoracic segment, with accompanying wings.

structure with A/P and D/V polarity. The activation of the segmentation genes depends on the free diffusion of maternally encoded morphogens, which is only possible within the syncytial blastoderm of the early *Drosophila* embryo.

## Segment identity arises from the action of homeotic genes

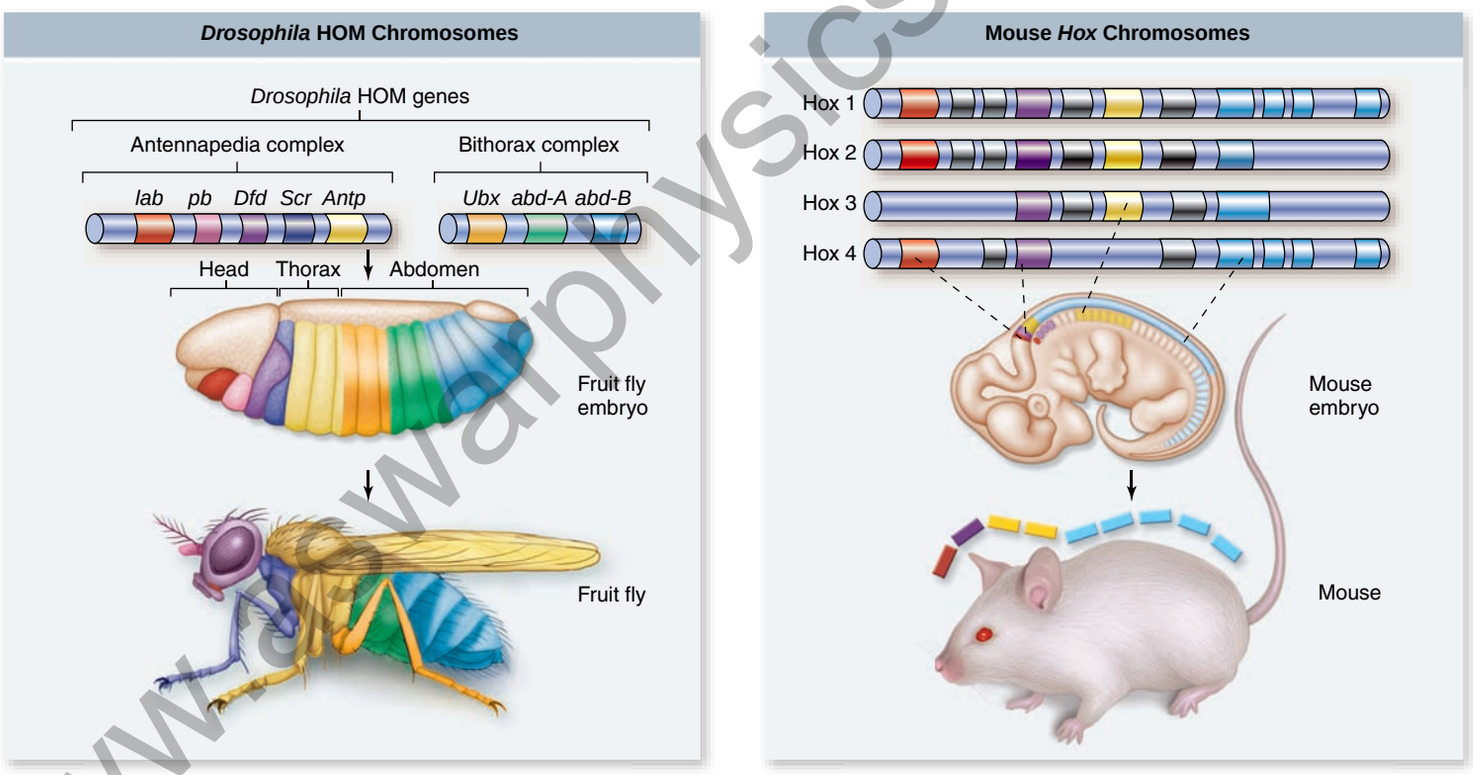
With the basic body plan laid down, the next step is to give identity to the segments of the embryo. A highly interesting class of *Drosophila* mutants has provided the starting point for understanding the creation of segment identity.

In these mutants, a particular segment seems to have changed its identity—that is, it has characteristics of a different segment. In wild-type flies, a pair of legs emerges from each of the three thoracic segments, but only the second thoracic segment has wings. Mutations in the *Ultrabithorax* gene cause a fly to grow an extra pair of wings, as though it has two second thoracic segments (figure 19.17). Even more bizarre are mutations in *Antennapedia*, which cause legs to grow out of the head in place of antennae!

Thus, mutations in these genes lead to the appearance of perfectly normal body parts in inappropriate places. Such mutants are termed *homeotic mutants* because the transformed body part looks similar (homeotic) to another. The genes in which such mutants occur are therefore called **homeotic genes.**

### *Homeotic gene complexes*

In the early 1950s, geneticist and Nobel laureate Edward Lewis discovered that several homeotic genes, including *Ultrabithorax*, map together on the third chromosome of *Drosophila* in a tight cluster called the **bithorax complex.** Mutations in these genes all affect body parts of the thoracic and abdominal segments, and Lewis concluded that the genes of the bithorax complex control the development of body parts in the rear half of the thorax and all of the abdomen.




*a.* *b.*

**Figure 19.18 A comparison of homeotic gene clusters in the fruit fly *Drosophila melanogaster* and the mouse *Mus musculus*.** ***a.*** *Drosophila* homeotic genes. Called the homeotic gene complex, or HOM complex, the genes are grouped into two clusters: the Antennapedia complex (anterior) and the bithorax complex (posterior). ***b.*** The *Drosophila* HOM genes and the mouse *Hox* genes are related genes that control the regional differentiation of body parts in both animals. These genes are located on a single chromosome in the fly and on four separate chromosomes in mammals. In this illustration, the genes are color-coded to match the parts of the body along the A/P axis in which they are expressed. Note that the order of the genes along the chromosome(s) is mirrored by their pattern of expression in the embryo and in structures in the adult fly.

Interestingly, the order of the genes in the bithorax complex mirrors the order of the body parts they control, as though the genes are activated serially. Genes at the beginning of the cluster switch on development of the thorax; those in the middle control the anterior part of the abdomen; and those at the end affect the posterior tip of the abdomen.

A second cluster of homeotic genes, the **Antennapedia complex,** was discovered in 1980 by Thomas Kaufmann. The Antennapedia complex governs the anterior end of the fly, and the order of genes in this complex also corresponds to the order of segments they control (figure 19.18*a*).

### *The homeobox*

An interesting relationship was discovered after the genes of the bithorax and Antennapedia complexes were cloned and sequenced. These genes all contain a conserved sequence of 180 nucleotides that codes for a 60-amino-acid, DNA-binding domain. Because this domain was found in all of the homeotic genes, it was named the *homeodomain*, and the DNA that encodes it is called the homeobox. Thus, the term ***Hox* gene** now refers to a homeobox-containing gene that specifies the identity of a body part. These genes function as transcription factors that bind DNA using their homeobox domain.

Clearly, the homeobox distinguishes portions of the genome that are devoted to pattern formation. How the *Hox* genes do this is the subject of much current research. Scientists believe that the ultimate targets of *Hox* gene function must be genes that control cell behaviors associated with organ morphogenesis.

### *Evolution of homeobox-containing genes*

A large amount of research has been devoted to analyzing the clustered complexes of *Hox* genes in other organisms. These investigations have led to a fairly coherent view of homeotic gene evolution.

It is now clear that the *Drosophila* bithorax and Antennapedia complexes represent two parts of a single cluster of genes. In vertebrates, there are four copies of *Hox* gene clusters. As in *Drosophila*, the spatial domains of *Hox* gene expression correlate with the order of the genes on the chromosome (figure 19.18*b*). The existence of four *Hox* clusters in vertebrates is viewed by many as evidence that two duplication events of the entire genome have occurred in the vertebrate lineage.

This idea raises the issue of when the original cluster arose. To answer this question, researchers have turned to more primitive organisms, such as *Amphioxus* (now called *Branchiostoma*), a lancelet chordate (see chapter 35). The finding of only one cluster of *Hox* genes in *Amphioxus* implies that indeed there have been two duplications in the vertebrate lineage, at least of the *Hox* cluster. Given the single cluster in arthropods, this finding implies that the common ancestor to all animals with bilateral symmetry had a single *Hox* cluster as well.

The next logical step is to look at even more-primitive animals: the radially symmetrical cnidarians such as *Hydra* (see chapter 33). Thus far, *Hox* genes have been found in a number of cnidarian species, and recent sequence analyses suggest that cnidarian *Hox* genes are also arranged into clusters. Thus, the appearance of the ancestral *Hox* cluster likely preceded the divergence between radial and bilateral symmetries in animal evolution.

## Pattern formation in plants is also under genetic control

The evolutionary split between plant and animal cell lineages occurred about 1.6 BYA, before the appearance of multicellular organisms with defined body plans. The implication is that multicellularity evolved independently in plants and animals. Because of the activity of meristems, additional modules can be added to plant bodies throughout their lifetimes. In addition, plant flowers and roots have a radial organization, in contrast to the bilateral symmetry of most animals. We may therefore expect that the genetic control of pattern formation in plants is fundamentally different from that of animals.

Although plants have homeobox-containing genes, they do not possess complexes of *Hox* genes similar to the ones that determine regional identity of developing structures in animals. Instead, the predominant homeotic gene family in plants appears to be the ***MADS*-box** genes.

*MADS*-box genes are a family of transcriptional regulators found in most eukaryotic organisms, including plants, animals, and fungi. The MADS-box is a conserved DNA-binding and dimerization domain, named after the first five genes to be discovered with this domain. Only a small number of *MADS*-box genes are found in animals, where their functions include the control of cell proliferation and tissue-specific gene expression in postmitotic muscle cells. They do not appear to play a role in the patterning of animal embryos.

In contrast, the number and functional diversity of *MADS*-box genes increased considerably during the evolution of land plants, and there are more than 100 *MADS*-box genes in the *Arabidopsis* genome. In flowering plants, the *MADS*-box genes dominate the control of development, regulating such processes as the transition from vegetative to reproductive growth, root development, and floral organ identity.

Although distinct from genes in the *Hox* clusters of animals, homeodomain-containing transcription factors in plants do have important developmental functions. One such example is the family of *knottedlike homeobox (knox)* genes, which are important regulators of shoot apical meristem development in both seed-bearing and nonseed-bearing plants. Mutations that affect expression of *knox* genes produce changes in leaf and petal shape, suggesting that these genes play an important role in generating leaf form.

### *Learning Outcomes Review 19.5*

**Pattern formation in animals involves the coordinated expression of a hierarchy of genes. Gradients of morphogens in *Drosophila* specify A/P and D/V axes, then lead to sequential activation of segmentation genes. Bicoid and Nanos protein gradients determine the A/P axis. The protein Dorsal determines the D/V axis, but activation requires a series of steps beginning with the oocyte's Gurken protein. The action of homeotic genes provide segment identity. These genes, which include a DNA-binding homeodomain sequence, are called *Hox* genes (for *homeobox* genes), and they are organized into clusters. Plants use a different set of developmental control genes called *MADS*-box genes.**

- ***Why would you expect homeotic genes to be conserved across species evolution?***

# 19.6 Morphogenesis

### Learning Outcomes

1. *Discuss the importance of cell shape changes and cell migration in development.*
2. *Explain how cell death can contribute to morphogenesis.*
3. *Describe the role of the extracellular matrix in cell migration.*

At the end of cleavage, the *Drosophila* embryo still has a relatively simple structure: It comprises several thousand identical-looking cells, which are present in a single layer surrounding a central yolky region. The next step in embryonic development is **morphogenesis**—the generation of ordered form and structure.

Morphogenesis is the product of changes in cell structure and cell behavior. Animals regulate the following processes to achieve morphogenesis:

- The number, timing, and orientation of cell divisions;
- Cell growth and expansion;
- Changes in cell shape;
- Cell migration; and
- Cell death.

Plant and animal cells are fundamentally different in that animal cells have flexible surfaces and can move, but plant cells are immotile and encased within stiff cellulose walls. Each cell in a plant is fixed into position when it is created. Thus, animal cells use cell migration extensively during development while plants use the other four mechanisms but lack cell migration. We consider the morphogenetic changes in animals first, and then those that occur in plants.

## Cell division during development may result in unequal cytokinesis

The orientation of the mitotic spindle determines the plane of cell division in eukaryotic cells. The coordinated function of microtubules and their motor proteins determines the respective position of the mitotic spindle within a cell (see chapter 10). If the spindle is centrally located in the dividing cell, two equal-sized daughter cells will result. If the spindle is off to one side, one large daughter cell and one small daughter cell will result.

The great diversity of cleavage patterns in animal embryos is determined by differences in spindle placement. In many cases, the fate of a cell is determined by its relative placement in the embryo during cleavage. For example, in preimplantation mammalian embryos, cells on the outside of the embryo usually differentiate into trophectoderm cells, which form only extra-embryonic structures later in development (for example, a part of the placenta). In contrast, the embryo proper is derived from the inner cell mass, cells which, as the name implies, are in the interior of the embryo.

## Cells change shape and size as morphogenesis proceeds

In animals, cell differentiation is often accompanied by profound changes in cell size and shape. For example, the large nerve cells that connect your spinal cord to the muscles in your big toe develop long processes called *axons* that span this entire distance. The cytoplasm of an axon contains microtubules, which are used for motor-driven transport of materials along the length of the axon.

As another example, muscle cells begin as *myoblasts*, undifferentiated muscle precursor cells. They eventually undergo conversion into the large, multinucleated *muscle fibers* that make up mammalian skeletal muscles. These changes begin with the expression of the *MyoD1* gene, which encodes a transcription factor that binds to the promoters of muscle-determining genes to initiate these changes.

## Programmed cell death is a necessary part of development

Not every cell produced during development is destined to survive. For example, human embryos have webbed fingers and toes at an early stage of development. The cells that make up the webbing die in the normal course of morphogenesis. As another example, vertebrate embryos produce a very large number of neurons, ensuring that enough neurons are available to make the necessary synaptic connections, but over half of these neurons never make connections and die in an orderly way as the nervous system develops.

Unlike accidental cell deaths due to injury, these cell deaths are planned—and indeed required—for proper development and morphogenesis. Cells that die due to injury typically swell and burst, releasing their contents into the extracellular fluid. This form of cell death is called necrosis. In contrast, cells programmed to die shrivel and shrink in a process called apoptosis, which means "falling away," and their remains are taken up by surrounding cells.

### *Genetic control of apoptosis*

Apoptosis occurs when a "death program" is activated. All animal cells appear to possess such programs. In *C. elegans*, the same 131 cells always die during development in a predictable and reproducible pattern.

Work on *C. elegans* showed that three genes are central to this process. Two (*ced-3* and *ced-4*) activate the death program itself; if either is mutant, those 131 cells do not die, and go on instead to form nervous tissue and other tissue. The third gene *(ced-9)* represses the death program encoded by the other two: All 1090 cells of the *C. elegans* embryo die in *ced-9* mutants. In *ced-9/ced-3* double mutants, all 1090 cells live, which suggests that *ced-9* inhibits cell death by functioning prior to *ced-3* in the apoptotic pathway (figure 19.19*a*).

The mechanism of apoptosis appears to have been highly conserved during the course of animal evolution. In human nerve cells, the *Apaf1* gene is similar to *ced-4* of *C. elegans* and activates the cell death program, and the human *bcl-2* gene acts

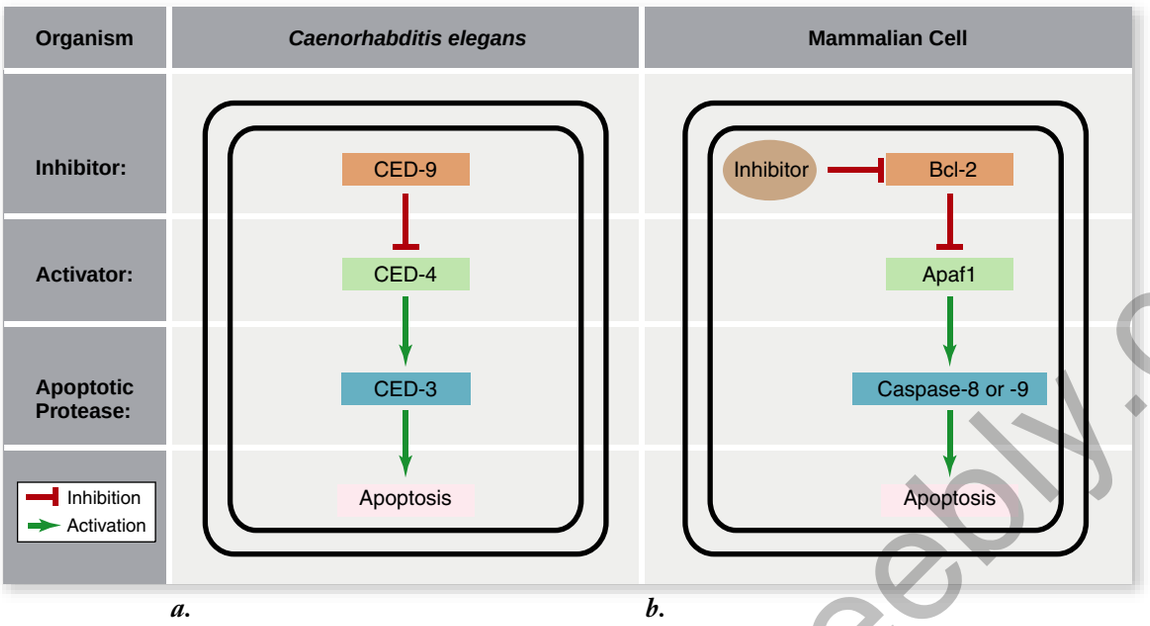



**Figure 19.19 Programmed cell death pathway.** Apoptosis, or programmed cell death, is necessary for the normal development of all animals. ***a.*** In the developing nematode, for example, two genes, *ced-3* and *ced-4*, code for proteins that cause the programmed cell death of 131 specific cells. In the other (surviving) cells of the developing nematode, the product of a third gene, *ced-9*, represses the death program encoded by *ced-3* and *ced-4*. ***b.*** The mammalian homologues of the apoptotic genes in *C. elegans* are *bcl-2* (*ced-9* homologue), *Apaf1* (*ced-4* homologue), and *caspase-8* or *-9* (*ced-3* homologues). In the absence of any cell survival factor, Bcl-2 is inhibited and apoptosis occurs. In the presence of nerve growth factor (NGF) and NGF receptor binding, Bcl-2 is activated, thereby inhibiting apoptosis.

similarly to *ced-9* to repress apoptosis. If a copy of the human *bcl-2* gene is transferred into a nematode with a defective *ced-9* gene, *bcl-2* suppresses the cell death program of *ced-3* and *ced-4*.

### *The mechanism of apoptosis*

The product of the *C. elegans ced-4* gene is a protease that activates the product of the *ced-3* gene, which is also a protease. The human *Apaf1* gene is actually named for its role: *A*poptotic *p*rotease *a*ctivating *f*actor. It activates two proteases called caspases that have a role similar to the Ced-3 protease in *C. elegans* (figure 19.19*b*). When the final proteases are activated, they chew up proteins in important cellular structures such as the cytoskeleton and the nuclear lamina, leading to cell fragmentation.

The role of Ced-9/Bcl-2 is to inhibit this program. Specifically, it inhibits the activating protease, preventing the activation of the destructive proteases. The entire process is thus controlled by an inhibitor of the death program.

Both internal and external signals control the state of the Ced-9/Bcl-2 inhibitor. For example, in the human nervous system, neurons have a cytoplasmic inhibitor of Bcl-2 that allows the death program to proceed (see figure 19.19*b*). In the presence of nerve growth factor, a signal transduction pathway leads to the cytoplasmic inhibitor being inactivated, allowing Bcl-2 to inhibit apoptosis and the nerve cell to survive.

## Cell migration gets the right cells to the right places

The migration of cells is important during many stages of animal development. The movement of cells involves both adhesion and the loss of adhesion. Adhesion is necessary for cells to get "traction," but cells that are initially attached to others must lose this adhesion to be able to leave a site.

Cell movement also involves cell-to-substrate interactions, and the extracellular matrix may control the extent or route of cell migration. The central paradigm of morphogenetic cell movements in animals is a change in cell adhesiveness, which is mediated by changes in the composition of macromolecules in the plasma membranes of cells or in the extracellular matrix. Cell-to-cell interactions are often mediated through cadherins, but cell-to-substrate interactions often involve integrin-to-extracellular-matrix (ECM) interactions.

### *Cadherins*

Cadherins are a large gene family, with over 80 members identified in humans. In the genomes of *Drosophila*, *C. elegans*, and humans, the cadherins can be sorted into several subfamilies that exist in all three genomes.

The cadherin proteins are all transmembrane proteins that share a common motif, the *cadherin domain*, a 110-amino-acid domain in the extracellular portion of the protein that mediates $Ca^{2+}$-dependent binding between like cadherins (homophilic binding).

Experiments in which cells are allowed to sort in vitro illustrate the function of cadherins. Cells with the same cadherins adhere specifically to one another, while not adhering to other cells with different cadherins. If cell populations with different cadherins are dispersed and then allowed to reaggregate, they sort into two populations of cells based on the nature of the cadherins on their surface.

An example of the action of cadherins can be seen in the development of the vertebrate nervous system. All surface ectoderm cells of the embryo express E-cadherin. The formation of the nervous system begins when a central strip of cells on the dorsal surface of the embryo turns off E-cadherin expression and turns on N-cadherin expression. In the process of **neurulation,** the formation of the neural tube (see chapter 54), the central strip of N-cadherin-expressing cells folds up to form the tube. The neural tube pinches off from the overlying cells, which continue to express E-cadherin. The surface cells outside the tube differentiate into the epidermis of the skin, whereas the neural tube develops into the brain and spinal cord of the embryo.

### *Integrins*

In some tissues, such as connective tissue, much of the volume of the tissue is taken up by the spaces *between* cells. These spaces are filled with a network of molecules secreted by surrounding cells, termed a *matrix*. In connective tissue such as cartilage, long polysaccharide chains are covalently linked to proteins (proteoglycans), within which are embedded strands of fibrous protein (collagen, elastin, and fibronectin). Migrating cells traverse this matrix by binding to it with cell surface proteins called integrins.

Integrins are attached to actin filaments of the cytoskeleton and protrude out from the cell surface in pairs, like two hands. The "hands" grasp a specific component of the matrix, such as collagen or fibronectin, thus linking the cytoskeleton to the fibers of the matrix. In addition to providing an anchor, this binding can initiate changes within the cell, alter the growth of the cytoskeleton, and activate gene expression and the production of new proteins.

The process of **gastrulation** (described in detail in chapter 54), during which the hollow ball of animal embryonic cells folds in on itself to form a multilayered structure, depends on fibronectin–integrin interactions. For example, injection of antibodies against either fibronectin or integrins into salamander embryos blocks binding of cells to fibronectin in the ECM and inhibits gastrulation. The result is like a huge traffic jam following a major accident on a freeway: Cells (cars) keep coming, but they get backed up since they cannot get beyond the area of inhibition (accident site) (figure 19.20). Similarly, a targeted knockout of the fibronectin gene in mice resulted in gross defects in the migration, proliferation, and differentiation of embryonic mesoderm cells.

Thus, cell migration is largely a matter of changing patterns of cell adhesion. As a migrating cell travels, it continually extends projections that probe the nature of its environment. Tugged this way and that by different tentative attachments, the cell literally feels its way toward its ultimate target site.

## In seed plants, the plane of cell division determines morphogenesis

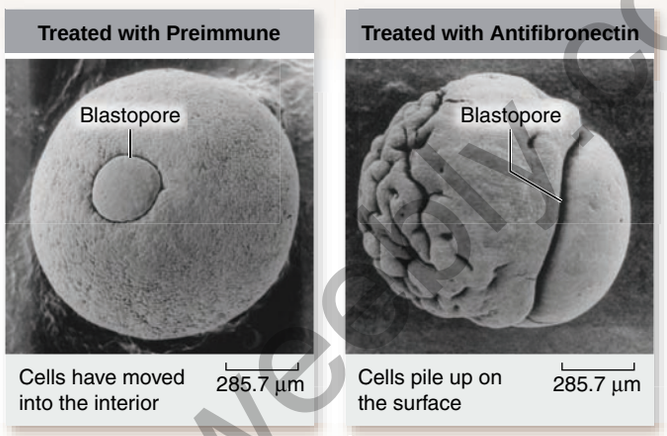
The form of a plant body is largely determined by the plane in which cells divide. The first division of the fertilized egg in a flowering plant is off-center, so that one of the daughter cells is small, with dense cytoplasm (figure 19.21*a*). That cell, the future embryo, begins to divide repeatedly, forming a ball of cells. The other daughter cell also divides repeatedly, forming an elongated structure called a *suspensor*, which links the embryo to the nutrient tissue of the seed. The suspensor also provides a route for nutrients to reach the developing embryo.

Just as many animal embryos acquire their initial axis as a cell mass formed during cleavage divisions, so the plant embryo forms its root–shoot axis at this time. Cells near the suspensor are destined to form a root, whereas those at the other end of the axis ultimately become a shoot, the aboveground portion of the plant.

The relative position of cells within the plant embryo is also a primary determinant of cell differentiation. The outermost cells in a plant embryo become epidermal cells. The bulk of the embryonic interior consists of ground tissue cells that eventually function in food and water storage. Finally, cells at the core of the embryo are destined to form the future vascular tissue (figure 19.21*b*). (Plant tissues and development are described in detail in chapters 36 and 37.)

**SCIENTIFIC THINKING**

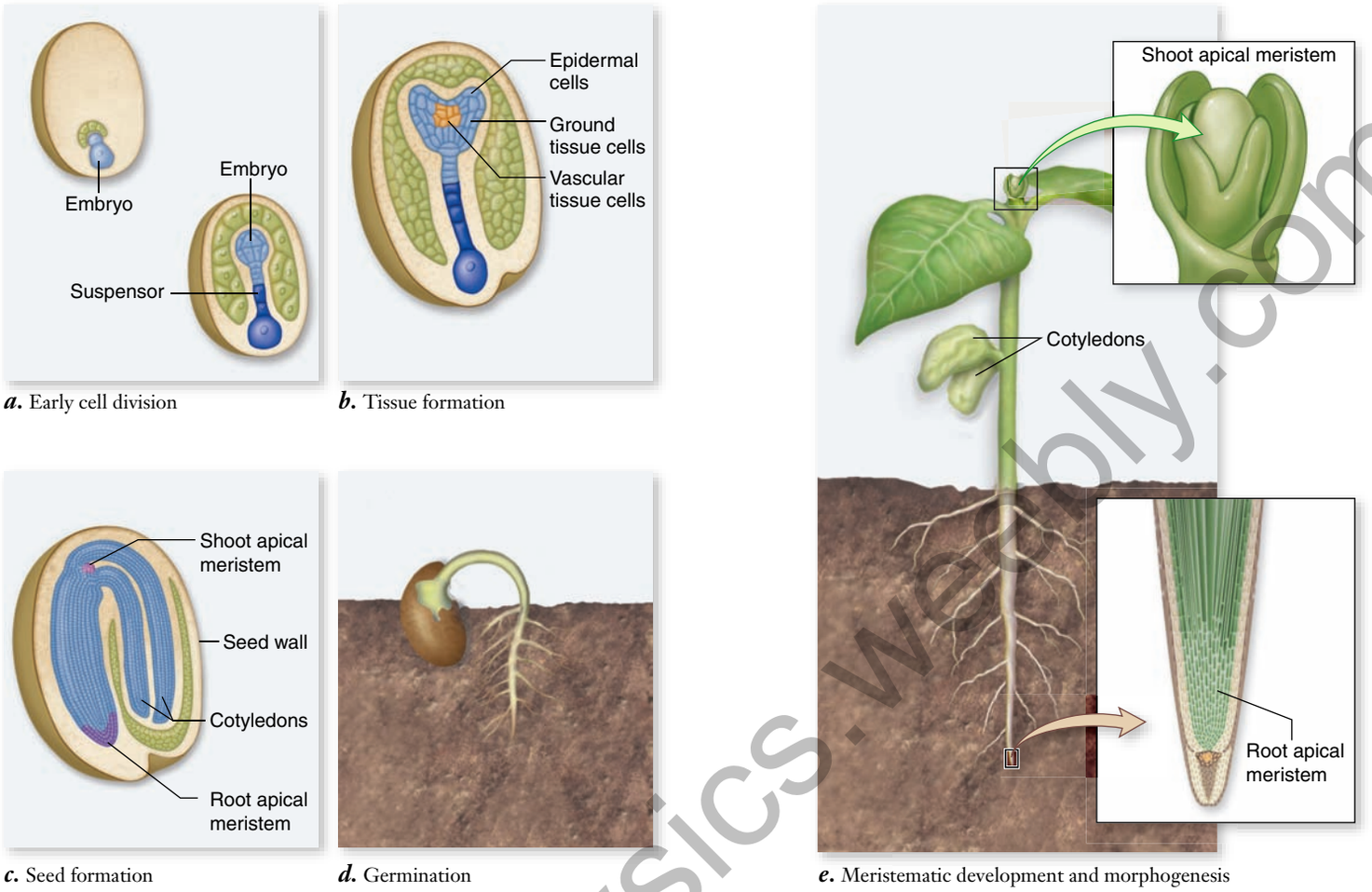
**Hypothesis:** *Fibronectin is required for cell migration during gastrulation.*
**Prediction:** *Blocking fibronectin with antifibronectin antibodies before gastrulation should prevent cell movement.*
**Test:** *Staged salamander embryos were injected either with antifibronectin antibody, or with preimmune serum as a control, prior to gastrulation. Cell movements were then monitored photographically.*




*a.* *b.*

**Result:** *The experimental embryos injected with antifibronectin antibody show extremely aberrant gastrulation where cells pile up and do not enter the interior of the embryo. Control embryos gastrulate normally.*
**Conclusion:** *Fibronectin is required for cells to migrate into the interior of the embryo during gastrulation.*
**Further Experiments:** *How can this same system be used to analyze the role of fibronectin in other early morphogenetic events?*

**Figure 19.20 Fibronectin is necessary for cell migration during gastrulation.**




***a.*** Early cell division

***b.*** Tissue formation

***c.*** Seed formation

***d.*** Germination

***e.*** Meristematic development and morphogenesis

**Figure 19.21 The path of plant development.** The developmental stages of *Arabidopsis thaliana* are ***(a)*** early embryonic cell division, ***(b)*** embryonic tissue formation, ***(c)*** seed formation, ***(d)*** germination, and ***(e)*** meristematic development and morphogenesis.

Soon after the three basic tissues form, a flowering plant embryo develops one or two seed leaves called *cotyledons*. At this point, development is arrested, and the embryo is either surrounded by nutritive tissue or has amassed stored food in its cotyledons (figure 19.21*c*). The resulting package, known as a *seed*, is resistant to drought and other unfavorable conditions.

A seed germinates in response to favorable changes in its environment. The embryo within the seed resumes development and grows rapidly, its roots extending downward and its leaf-bearing shoots extending upward (figure 19.21*d*). Plant development exhibits its great flexibility during the assembly of the modules that make up a plant body. Apical meristems at the root and shoot tips generate the large numbers of cells needed to form leaves, flowers, and all other components of the mature plant (figure 19.21*e*).

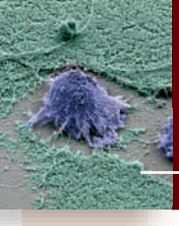
Growth within the developing flower is controlled by a cascade of transcription factors. A key member of this cascade is the *AINTEGUMENTA (ANT)* gene. Loss of ANT function reduces the number and size of floral organs, and inappropriate expression leads to larger floral organs.

Plant body form is also established by controlled changes in cell shape as cells expand osmotically after they form. Plant growth-regulating hormones and other factors influence the orientation of bundles of microtubules on the interior of the plasma membrane. These microtubules seem to guide cellulose deposition as the cell wall forms around the outside of a new cell. The orientation of the cellulose fibers, in turn, determines how the cell will elongate as it increases in volume due to osmosis, and so determines the cell's final shape.

### *Learning Outcomes Review 19.6*

**Morphogenesis is the generation of ordered form and structure. This process proceeds along with cell differentiation. The primary mechanisms of morphogenesis are cell shape change and cell migration. Apoptosis is programmed cell death that is a necessary part of morphogenesis. Cell migration in animals involves alternating changes in adhesion brought about by cadherins and integrins. In plants, which cannot move, cell division and cell expansion are the primary morphogenetic processes.**

- ***Why is cell death important to morphogenesis?***



# Chapter Review

## 19.1 The Process of Development

Development is the sequence of systematic, gene-directed changes throughout a life cycle. The four subprocesses of development are growth, cell differentiation, pattern formation, and morphogenesis.

## 19.2 Cell Division

***Development begins with cell division.***

In animals, cleavage stage divisions divide the fertilized egg into numerous smaller cells called blastomeres. During cleavage the $G_1$ and $G_2$ phases of the cell cycle are shortened or eliminated (figure 19.2).

***Every cell division is known in the development of* C. elegans.**

The lineage of 959 adult somatic *Caenorhabditis elegans* cells is invariant. Knowledge of the differentiation sequence and outcome allows study of developmental mechanisms.

***Plant growth occurs in specific areas called meristems.***

Plant growth continues throughout the life span from meristematic stem cells that can divide and differentiate into any plant tissue.

## 19.3 Cell Differentiation

***Cells become determined prior to differentiation.***

The process of determination commits a cell to a particular developmental pathway prior to its differentiation. This is not visible but can be tracked experimentally. Determination is due to differential inheritance of cytoplasmic factors or cell-to-cell interactions.

***Determination can be due to cytoplasmic determinants.***

In tunicates, determination of tail muscle cells depends on the presence of mRNA for the Macho-1 transcription factor, which is deposited in the egg cytoplasm during gamete formation.

***Induction can lead to cell differentiation.***

Induction occurs when one cell type produces signal molecules that induce gene expression in neighboring target cells.

In frogs, cells from animal and vegetal poles do not develop into mesoderm when isolated. In tunicates, signaling by the growth factor FGF induces mesoderm development.

***Stem cells can divide and produce cells that differentiate.***

Stem cells replace themselves by division and produce cells that differentiate. Totipotent stem cells can give rise to any cell type including extraembryonic tissues; pluripotent cells can give rise to all cells of an organism; and multipotent stem cells can give rise to many kinds of cells.

***Embryonic stem cells are pluripotent cells derived from embryos.***

Embryonic stem cells are derived from the inner cell mass of the blastocyst (figure 19.8). They can differentiate into any adult tissue in a mouse.

## 19.4 Nuclear Reprogramming

***Reversal of determination has allowed cloning.***

Cells undergo no irreversible changes during development. However, transplanted nuclei from older donors are less able to direct complete development. The cloning of the sheep Dolly showed that the nucleus of an adult cell can be reprogrammed to be totipotent (figure 19.9).

***Reproductive cloning has inherent problems.***

Reproductive cloning has a low success rate, and clones often develop age-associated diseases.

***Nuclear reprogramming has been accomplished by use of defined factors.***

Adult cells can be converted into pluripotent cells by introduction of four genes for transcription factors. These induced pluripotent cells appear to be similar to ES cells.

The use of cells cloned from a patient's cells to replace damaged tissue could avoid the problem of transplant rejection.

## 19.5 Pattern Formation

**Drosophila *embryogenesis produces a segmented larva.***

The maternal contribution of mRNA along with the postfertilization events of cellular blastoderm formation produce a segmented embryo.

***Morphogen gradients form the basic body axes in* Drosophila.**

Pattern formation produces two perpendicular axes in a bilaterally symmetrical organism. Positional information leads to changes in gene activity so cells adopt a fate appropriate for their location.

Formation of the anterior/posterior (A/P) axis is based on opposing gradients of morphogens, Bicoid and Nanos, synthesized from maternal mRNA (figures 19.14, 19.15).

The dorsal/ventral (D/V) axis is established by a gradient of the Dorsal transcription factor. Successive action of transcription factors divides the embryo into segments.

***The body plan is produced by sequential activation of genes.***

***Segment identity arises from the action of homeotic genes.***

Homeotic genes, called *Hox* genes because they contain a DNA sequence called the homeobox, give identity to embryo segments.

*Hox* genes are found in four clusters in vertebrates.

***Pattern formation in plants is also under genetic control.***

Plants have *MADS*-box genes that control the transition from vegetative to reproductive growth, root development, and floral organ identity.

## 19.6 Morphogenesis

***Cell division during development may result in unequal cytokinesis.***

***Cells change shape and size as morphogenesis proceeds.***

Depending on the orientation of the mitotic spindle, cells of equal or different sizes can arise. Morphogenesis involves changes in cell shape and size and cell migration.

***Programmed cell death is a necessary part of development.***

Apoptosis, the programmed death of cells, removes structures once they are no longer needed (figure 19.19).

***Cell migration gets the right cells to the right places.***

The migration of cells requires both adhesion and loss of adhesion between cells and their substrate.

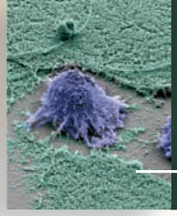
Cell-to-cell interactions are often mediated by cadherin proteins, whereas cell-to-substrate interactions may involve integrin-to-extracellular-matrix interactions.

Integrins bind to fibers found in the extracellular matrix. This action can alter the cytoskeleton and activate gene expression.

***In seed plants, the plane of cell division determines morphogenesis.***

In plants, the primary morphogenetic processes are cell division, relative position of cells within the embryo, and changes in cell shape. Plant development stages begin with cell division and end with meristematic development and morphogenesis (figure 19.21).

Relative position of cells in the plant embryo is the main determinant of cell differentiation.



# Review Questions

## UNDERSTAND

1. During development, cells become
   a. differentiated before they become determined.
   b. determined before they become differentiated.
   c. determined by the loss of genetic material.
   d. differentiated by the loss of genetic material.
2. Determination can occur by
   a. the action of cytoplasmic determinants.
   b. induction by other cells.
   c. the loss of chromosomes during cell division.
   d. both a and b.
3. The rapid divisions that occur early in development are made possible by shortening
   a. M phase.
   b. S phase.
   c. $G_1$ and $G_2$ phases.
   d. all of the above.
4. A pluripotent cell is one that can
   a. become any cell type in an organism.
   b. produce an indefinite supply of a single cell type.
   c. produce a limited amount of a specific cell type.
   d. produce multiple cell types.
5. Plant meristems
   a. are only present during development.
   b. contain stem cells.
   c. undergo meiosis.
   d. all of the above
6. Pattern formation involves cells determining their position in the embryo. One mechanism that can accomplish this is
   a. the loss of genetic material.
   b. alterations of chromosome structure.
   c. gradients of morphogens.
   d. changes in the cell cycle.
7. The process of nuclear reprogramming
   a. is a normal part of pattern formation.
   b. reverses the changes that occur during differentiation.
   c. requires the introduction of new DNA.
   d. is not possible with mammalian cells.

## APPLY

1. What is the common theme in cell determination by induction or cytoplasmic determinants?
   a. The activation of transcription factors
   b. The activation cell division
   c. A change in gene expression
   d. Both a and c
2. The process of reproductive cloning
   a. shows that nuclear reprogramming is possible.
   b. is very efficient in mammals.
   c. always produces adult animals that are identical to the donor.
   d. is both a and b.
3. Production of anterior–posterior and dorsal–ventral axes in the fruit fly *Drosophila*
   a. both use gradients of mRNA.
   b. are conceptually similar but mechanistically different.
   c. use the exact same mechanisms.
   d. both use gradients of protein.
4. For pattern formation to occur, the cells in the developing embryo must
   a. "know" their position in the embryo.
   b. be determined during the earliest divisions.
   c. differentiate as they are "born."
   d. must all be reprogrammed after each cell division.
5. The genes that encode the morphogen gradients in *Drosophila* were all identified in mutant screens. A mutation that removes the gradient necessary for the A/P morphogen gradient would be expected to
   a. affect the larvae but not the adult.
   b. affect the adult but not the larvae.
   c. be lethal and lead to an abnormal embryo.
   d. produce replacement of one adult structure with another.
6. What would be the likely result of a mutation of the *bcl-2* gene on the level of apoptosis?
   a. No change
   b. A decrease in apoptosis
   c. An increase in apoptosis
   d. An initial decrease, followed by a increase in apoptosis
7. *MADS*-box, and *Hox* genes are
   a. found only in plants and animals, respectively.
   b. found only in animals and plants, respectively.
   c. have similar roles in development in plants and animals, respectively.
   d. have similar roles in development in animals and plants, respectively.

## SYNTHESIZE

1. The fate map for *C. elegans* (refer to figure 19.3) diagrams development of a multicellular organism from a single cell. Use this fate map to determine the number of cell divisions required to establish the population of cells that will become (a) the nervous system and (b) the gonads.
2. Carefully examine the *C. elegans* fate map in figure 19.3. Notice that some of the branchpoints (daughter cells) do *not* go on to produce more cells. What is the cellular mechanism underlying this pattern?
3. You have generated a set of mutant embryonic mouse cells. Predict the developmental consequences for each of the following mutations.
   a. Knockout mutation for N-cadherin
   b. Knockout mutation for integrin
   c. Deletion of the cytoplasmic domain of integrin
4. Assume you have the factors in hand necessary to reprogram an adult cell, and the factors necessary to induce differentiation to any cell type. How could these be used to replace a specific damaged tissue in a human patient?

## ONLINE RESOURCE

www.ravenbiology.com


Understand, Apply, and Synthesize—enhance your study with animations that bring concepts to life and practice tests to assess your understanding. Your instructor may also recommend the interactive eBook, individualized learning tools, and more.

Chapter 20

# Genes Within Populations

## Chapter Outline





## Introduction

*No other human being is exactly like you (unless you have an identical twin). Often the particular characteristics of an individual have an important bearing on its survival, on its chances to reproduce, and on the success of its offspring. Evolution is driven by such factors, as different alleles rise and fall in populations. These deceptively simple matters lie at the core of evolutionary biology, which is the topic of this chapter and chapters 21 through 25.*

## 20.1 Genetic Variation and Evolution

### Learning Outcomes

1. ***Define* evolution *and* population genetics.**
2. ***Explain the difference between evolution by natural selection and the inheritance of acquired characteristics.***

Genetic variation, that is, differences in alleles of genes found within individuals of a population, provides the raw material for natural selection, which will be described shortly. Natural populations contain a wealth of such variation. In plants, insects, and vertebrates, many genes exhibit some level of variation. In this chapter, we explore genetic variation in natural populations and consider the evolutionary forces that cause allele frequencies in natural populations to change.

The word *evolution* is widely used in the natural and social sciences. It refers to how an entity—be it a social system, a gas,

or a planet—changes through time. Although development of the modern concept of evolution in biology can be traced to Darwin's landmark work, *On the Origin of Species*, the first five editions of his book never actually used the term. Rather, Darwin used the phrase "descent with modification."

Although many more complicated definitions have been proposed, Darwin's words probably best capture the essence of biological evolution: Through time, species accumulate differences; as a result, descendants differ from their ancestors. In this way, new species arise from existing ones.

## Many processes can lead to evolutionary change

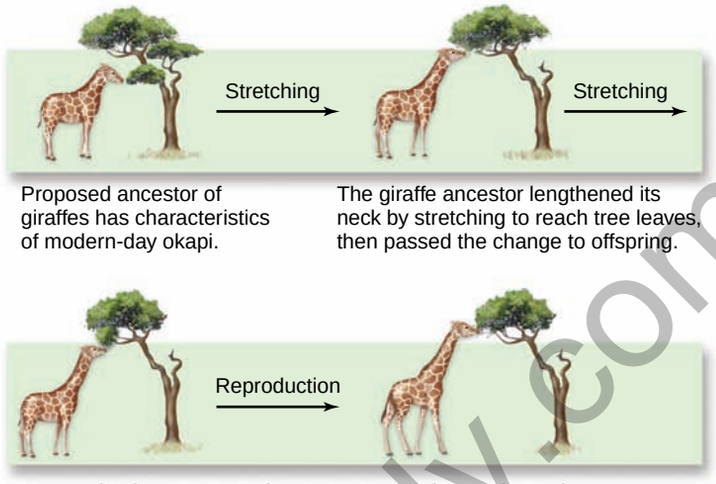
You have already learned about the development of Darwin's ideas in chapter 1. Darwin was not the first to propose a theory of evolution. Rather, he followed a long line of earlier philosophers and naturalists who deduced that the many kinds of organisms around us were produced by a process of evolution.

Unlike his predecessors, however, Darwin proposed natural selection as the mechanism of evolution. Natural selection produces evolutionary change when some individuals in a population possess certain inherited characteristics and then produce more surviving offspring than individuals lacking these characteristics. As a result, the population gradually comes to include more and more individuals with the advantageous characteristics. In this way, the population evolves and becomes better adapted to its local circumstances.

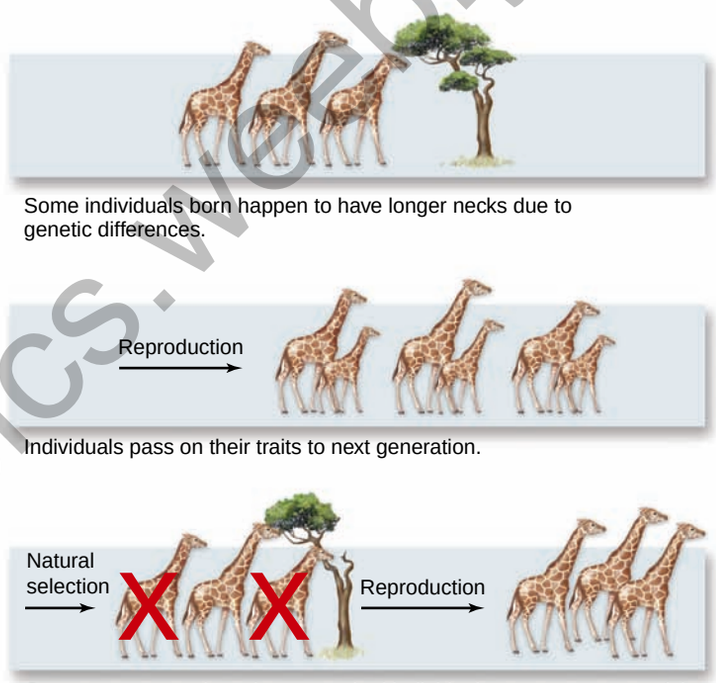
A rival theory, championed by the prominent biologist Jean-Baptiste Lamarck, was that evolution occurred by the **inheritance of acquired characteristics.** According to Lamarck, changes that individuals acquired during their lives were passed on to their offspring. For example, Lamarck proposed that ancestral giraffes with short necks tended to stretch their necks to feed on tree leaves, and this extension of the neck was passed on to subsequent generations, leading to the long-necked giraffe (figure 20.1*a*). In Darwin's theory, by contrast, the variation is not created by experience, but is the result of preexisting genetic differences among individuals (figure 20.1*b*).

One way to monitor how populations change through time is to look at changes in the frequencies of alleles of a gene from one generation to the next. Natural selection, by favoring individuals with certain alleles, can lead to change in such *allele frequencies*, but it is not the only process that can do so. Allele frequencies can also change when mutations occur repeatedly, changing one allele to another, and when migrants bring alleles into a population. In addition, when populations are small, the frequencies of alleles can change randomly as the result of chance events. Often, natural selection overwhelms the effects of these other processes, but as you will see later in this chapter, this is not always the case.

Evolution can result from any process that causes a change in the genetic composition of a population. We cannot talk about evolution, therefore, without also considering **population genetics,** the study of the properties of genes in populations.




***a.*** Lamarck's theory: acquired variation is passed on to descendants.




***b.*** Darwin's theory: natural selection or genetically-based variation leads to evolutionary change.

**Figure 20.1 Two ideas of how giraffes might have evolved long necks.**

## Populations contain ample genetic variation

It is best to start by looking at the genetic variation present among individuals within a species. This is the raw material available for the selective process.

As you saw in chapter 12, a natural population can contain a great deal of genetic variation. How much variation usually occurs? Humans are representative of most—but not all—species in that human populations contain substantial amounts of genetic variation. For example:

1. **Genes that influence blood groups.** Chemical analysis has revealed the existence of more than 30 blood group genes in humans, in addition to the ABO locus. At least

one third of these genes are routinely found in several alternative allelic forms in human populations. In addition to these, more than 45 variable genes encode other proteins in human blood cells and plasma that are not considered blood groups. In short, many genetically variable genes are present in this one system alone.

2. **Genes that influence enzymes.** Alternative alleles of genes specifying particular enzymes are easy to distinguish by measuring how fast the alternative proteins migrate in an electrical field (a process called *electrophoresis*—see chapter 17). A great deal of variation exists at enzyme-specifying loci. About 5% of the enzyme loci of a typical human are heterozygous: If you picked an individual at random, and in turn selected one of the enzyme-encoding genes of that individual at random, the chances are 1 in 20 (5%) that the gene you selected would be heterozygous in that individual.

Considering the entire genome, it is fair to say that all humans are different from one another except for identical twins. This is also true of other organisms, except for those that reproduce asexually. In nature, genetic variation is the rule.

### Enzyme polymorphism

Many loci in a particular population have more than one allele at frequencies significantly greater than would occur due to mutation alone. Researchers refer to such a locus as **polymorphic** (figure 20.2). The extent of such variation within natural populations was not even suspected a few decades ago, when modern techniques such as protein electrophoresis made it possible to examine enzymes and other proteins directly.

We now know that most populations of insects and plants are polymorphic at more than half of their enzyme-encoding loci, that is, the loci have more than one allele occurring at a frequency greater than 5%. Vertebrates are somewhat less polymorphic. Heterozygosity, the probability that a randomly selected gene will be heterozygous in a randomly selected individual, is about 15% in *Drosophila* and other invertebrates, between 5% and 8% in vertebrates, and around 8% in outcrossing plants (values of heterozygosity tend to be lower than the proportion of loci that are polymorphic because for loci that are polymorphic, many individuals within the population will be homozygous). These high levels of genetic variability provide ample supplies of raw material for evolution.

### DNA sequence polymorphism

The advent of gene technology has made it possible to assess genetic variation even more directly by sequencing the DNA itself. For example, when the *ADH* genes (which encode for alcohol dehydrogenase) of 11 *Drosophila melanogaster* individuals were sequenced, scientists found 43 variable sites, only 1 of which had been detected by protein electrophoresis.

Numerous other studies of variation at the DNA level have confirmed these findings: Abundant variation exists in both the coding regions of genes and in their nontranslated introns—considerably more variation than we can detect by examining enzymes with electrophoresis.

**Figure 20.2 Polymorphic variation.** This natural population of loosestrife, *Lythrum salicaria*, exhibits considerable variation in flower color. Individual differences are inherited and passed on to offspring.

#### Learning Outcomes Review 20.1

**Evolution can be described as descent with modification. Natural selection occurs when individuals carrying certain alleles leave more offspring than those without the alleles. Natural populations contain more genetic variation than can be accounted for by mutation alone. Population genetics studies this variability through statistical analyses.**

- ***Why is genetic variation in a population necessary for evolution to occur?***

## 20.2 Changes in Allele Frequency

#### Learning Outcomes

1. *Explain the Hardy–Weinberg principle.*
2. *Describe the characteristics of a population that is in Hardy–Weinberg equilibrium.*
3. *Demonstrate how the operation of evolutionary processes can be detected.*

Genetic variation within natural populations was a puzzle to Darwin and his contemporaries in the mid-1800s. The way in which meiosis produces genetic segregation among the progeny of a hybrid had not yet been discovered. And, although Mendel performed his experiments during this same time period, his work was largely unknown. Selection, scientists then thought, should always favor an optimal form, and so tend to eliminate variation. Moreover, the theory of *blending inheritance*—in which offspring were expected to be phenotypically intermediate relative to their parents—was widely accepted. If blending inheritance were correct, then the effect of any new genetic variant would quickly be diluted to the point of disappearance in subsequent generations.

## The Hardy–Weinberg principle allows prediction of genotype frequencies

Following the rediscovery of Mendel's research, two people in 1908 solved the puzzle of why genetic variation persists—Godfrey H. Hardy, an English mathematician, and Wilhelm Weinberg, a German physician. These workers were initially confused about why, after many generations, a population didn't come to be composed solely of individuals with the dominant phenotype. The conclusion they independently came to was that the original proportions of the genotypes in a population will remain constant from generation to generation, as long as the following assumptions are met:

1. No mutation takes place.
2. No genes are transferred to or from other sources (no immigration or emigration takes place).
3. Random mating is occurring.
4. The population size is very large.
5. No selection occurs.

Because the genotypes' proportions do not change, they are said to be in **Hardy–Weinberg equilibrium.**

### *The Hardy–Weinberg equation with two alleles: A binomial expansion*

In algebraic terms, the Hardy–Weinberg principle is written as an equation. Consider a population of 100 cats in which 84 are black and 16 are white. The frequencies of the two phenotypes would be 0.84 (or 84%) black and 0.16 (or 16%) white. Based on these phenotypic frequencies, can we deduce the underlying frequency of genotypes?

If we assume that the white cats are homozygous recessive for an allele we designate as *b*, and the black cats are either homozygous dominant *BB* or heterozygous *Bb*, we can calculate the **allele frequencies** of the two alleles in the population from the proportion of black and white individuals, assuming that the population is in Hardy–Weinberg equilibrium.

Let the letter $p$ designate the frequency of the *B* allele and the letter $q$ the frequency of the alternative allele. Because there are only two alleles, $p$ plus $q$ must always equal 1 (that is, the total population). In addition, we know that the sum of the three genotype frequencies must also equal 1. If the frequency of the *B* allele is $p$, then the probability that an individual will have two *B* alleles is simply the probability that each of its alleles is a *B*. The probability of two events happening independently is the product of the probability of each event; in this case, the probability that the individual received a *B* allele from its father is $p$, and the probability the individual received a *B* allele from its mother is also $p$, so the probability that both happened is $p * p = p^2$ (figure 20.3). By the same reasoning, the probability that an individual will have two *b* alleles is $q^2$.

What about the probability that an individual will be a heterozygote? There are two ways this could happen: The individual could receive a *B* from its father and a *b* from its mother, or vice versa. The probability of the first case is $p * q$ and the probability of the second case is $q * p$. Because the result in either case is that the individual is a heterozygote, the probability of that outcome is the sum of the two probabilities, or $2pq$.

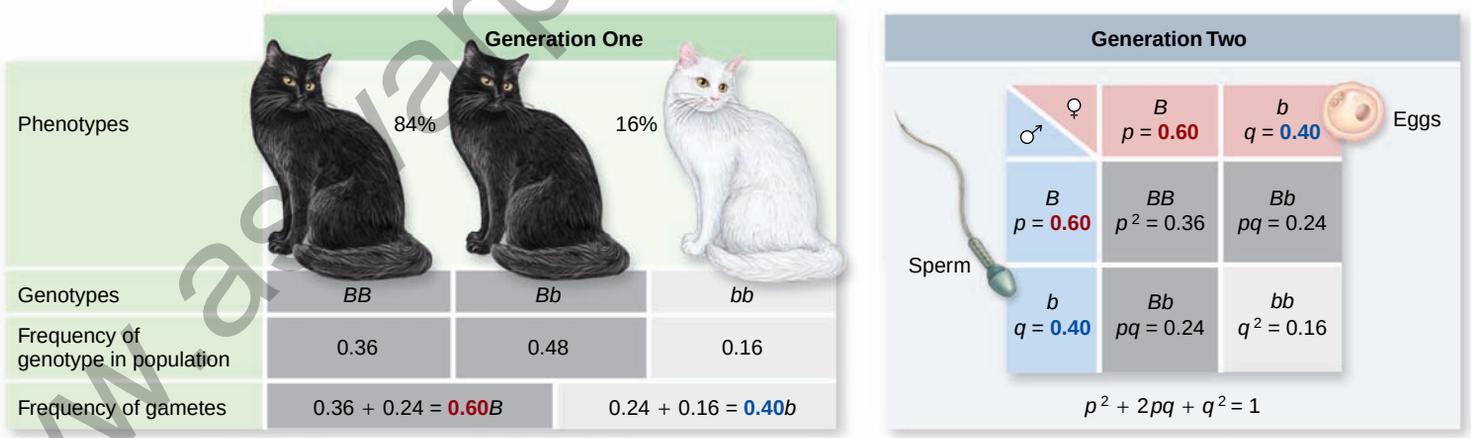



**Figure 20.3 The Hardy–Weinberg equilibrium.** In the absence of factors that alter them, the frequencies of gametes, genotypes, and phenotypes remain constant generation after generation.

**Inquiry question**

**If all white cats died, what proportion of the kittens in the next generation would be white?**

So, to summarize, if a population is in Hardy–Weinberg equilibrium with allele frequencies of $p$ and $q$, then the probability that an individual will have each of the three possible genotypes is $p^2 + 2pq + q^2$. You may recognize this as the *binomial expansion:*

$$(p + q)^2 = p^2 + 2pq + q^2$$

Finally, we may use these probabilities to predict the distribution of genotypes in the population, again assuming that the population is in Hardy–Weinberg equilibrium. If the probability that any individual is a heterozygote is $2pq$, then we would expect the proportion of heterozygous individuals in the population to be $2pq$; similarly, the frequency of *BB* and *bb* homozygotes would be expected to be $p^2$ and $q^2$.

Let us return to our example. Remember that 16% of the cats are white. If white is a recessive trait, then this means that such individuals must have the genotype *bb*. If the frequency of this genotype is $q^2 = 0.16$ (the frequency of white cats), then $q$ (the frequency of the *b* allele) = 0.4. Because $p + q = 1$, therefore, $p$, the frequency of allele *B*, would be 1.0 – 0.4 = 0.6 (remember, the frequencies must add up to 1). We can now easily calculate the expected **genotype frequencies:** homozygous dominant *BB* cats would make up the $p^2$ group, and the value of $p^2 = (0.6)^2 = 0.36$, or 36 homozygous dominant *BB* individuals in a population of 100 cats. The heterozygous cats have the *Bb* genotype and would have the frequency corresponding to $2pq$, or (2 * 0.6 * 0.4) = 0.48, or 48 heterozygous *Bb* individuals.

### *Using the Hardy–Weinberg equation to predict frequencies in subsequent generations*

The Hardy–Weinberg equation is another way of expressing the Punnett square described in chapter 12, with two alleles assigned frequencies, $p$ and $q$. Figure 20.3 allows you to trace genetic reassortment during sexual reproduction and see how it affects the frequencies of the *B* and *b* alleles during the next generation.

In constructing this diagram, we have assumed that the union of sperm and egg in these cats is random, so that all combinations of *b* and *B* alleles occur. The alleles are therefore mixed randomly and are represented in the next generation in proportion to their original occurrence. Each individual egg or sperm in each generation has a 0.6 chance of receiving a *B* allele ($p = 0.6$) and a 0.4 chance of receiving a *b* allele ($q = 0.4$).

In the next generation, therefore, the chance of combining two *B* alleles is $p^2$, or 0.36 (that is, 0.6 * 0.6), and approximately 36% of the individuals in the population will continue to have the *BB* genotype. The frequency of *bb* individuals is $q^2$ (0.4 * 0.4) and so will continue to be about 16%, and the frequency of *Bb* individuals will be $2pq$ (2 * 0.6 * 0.4), or on average, 48%.

Phenotypically, if the population size remains at 100 cats, we would still see approximately 84 black individuals (with either *BB* or *Bb* genotypes) and 16 white individuals (with the *bb* genotype). Allele, genotype, and phenotype frequencies have remained unchanged from one generation to the next, despite the reshuffling of genes that occurs during meiosis and sexual reproduction. Dominance and recessiveness of alleles can therefore be seen only to affect how an allele is expressed in an individual and not how allele frequencies will change through time.

## Hardy–Weinberg predictions can be applied to data to find evidence of evolutionary processes

The lesson from the example of black and white cats is that if all five of the assumptions listed earlier hold true, the allele and genotype frequencies will not change from one generation to the next. But in reality, most populations in nature will not fit all five assumptions. The primary utility of this method is to determine whether some evolutionary process or processes are operating in a population and, if so, to suggest hypotheses about what they may be.

Suppose, for example, that the observed frequencies of the *BB*, *bb*, and *Bb* genotypes in a different population of cats were 0.6, 0.2, and 0.2, respectively. We can calculate the allele frequencies for *B* as follows: 60% (0.6) of the cats have two *B* alleles, 20% have one, and 20% have none. This means that the average number of *B* alleles per cat is 1.4 [(0.6 × 2) + (0.2 × 1) + (0.2 × 0) = 1.4]. Because each cat has two alleles for this gene, the frequency is 1.4/2.0 = 0.7. Similarly, you should be able to calculate that the frequency of the *b* allele = 0.3.

If the population were in Hardy–Weinberg equilibrium, then, according to the equation earlier in this section, the frequency of the *BB* genotype would be $0.7^2 = 0.49$, lower than it really is. Similarly, you can calculate that there are fewer heterozygotes and more *bb* homozygotes than expected; then clearly, the population is not in Hardy–Weinberg equilibrium.

What could cause such an excess of homozygotes and deficit of heterozygotes? A number of possibilities exist, including (1) natural selection favoring homozygotes over heterozygotes, (2) individuals choosing to mate with genetically similar individuals (because *BB* * *BB* and *bb* * *bb* matings always produce homozygous offspring, but only half of *Bb* * *Bb* produce heterozygous offspring, such mating patterns would lead to an excess of homozygotes), or (3) an influx of homozygous individuals from outside populations (or conversely, emigration of heterozygotes to other populations). By detecting a lack of Hardy–Weinberg equilibrium, we can generate potential hypotheses that we can then investigate directly.

The operation of evolutionary processes can be detected in a second way. As discussed previously, if all of the Hardy–Weinberg assumptions are met, then allele frequencies will stay the same from one generation to the next. Changes in allele frequencies between generations would indicate that one of the assumptions is not met.

Suppose, for example, that the frequency of *b* was 0.53 in one generation and 0.61 in the next. Again, there are a number of possible explanations: For example, (1) selection favoring individuals with *b* over *B*, (2) immigration of *b* into the population or emigration of *B* out of the population, or (3) high rates of mutation that more commonly occur from *B* to *b* than vice versa. Another possibility is that the population is a small one, and that the change represents the random fluctuations that result because, simply by chance, some individuals pass on more of their genes than others. We will discuss how each of these processes is studied in the rest of the chapter.

### Learning Outcomes Review 20.2

**The Hardy–Weinberg principle states that in a large population with no selection and random mating, the proportion of alleles does not change through the generations. Finding that a population is not in Hardy–Weinberg equilibrium indicates that one or more evolutionary agents are operating.**

- ***If you know the genotype frequencies in a population, how can you determine whether the population is in Hardy–Weinberg equilibrium?***

## 20.3 Five Agents of Evolutionary Change

### Learning Outcomes

1. ***Define the five processes that can cause evolutionary change.***
2. ***Explain how these processes can cause populations to deviate from Hardy-Weinberg Equilibrium***

The five assumptions of the Hardy–Weinberg principle also indicate the five agents that can lead to evolutionary change in populations. They are mutation, gene flow, nonrandom mating, genetic drift in small populations, and the pressures of natural selection. Any one of these may bring about changes in allele or genotype proportions.

### Mutation changes alleles

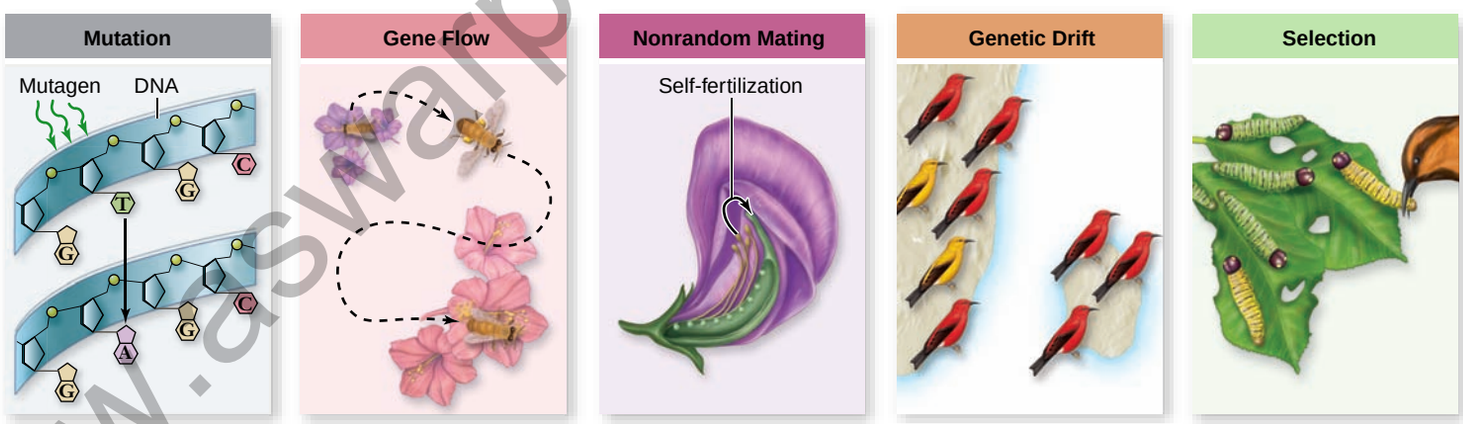
Mutation from one allele to another can obviously change the proportions of particular alleles in a population. Mutation rates are generally so low that they have little effect on the Hardy–Weinberg proportions of common alleles. A typical gene mutates about once per 100,000 cell divisions. Because this rate is so low, other evolutionary processes are usually more important in determining how allele frequencies change.

Nonetheless, mutation is the ultimate source of genetic variation and thus makes evolution possible (figure 20.4*a*). It is important to remember, however, that the likelihood of a particular mutation occurring is not affected by natural selection; that is, mutations do not occur more frequently in situations in which they would be favored by natural selection.

### Gene flow occurs when alleles move between populations

**Gene flow** is the movement of alleles from one population to another. It can be a powerful agent of change. Sometimes gene flow is obvious, as when an animal physically moves from one place to another. If the characteristics of the newly arrived individual differ from those of the animals already there, and if the newcomer is adapted well enough to the new area to survive and mate successfully, the genetic composition of the receiving population may be altered.

Other important kinds of gene flow are not as obvious. These subtler movements include the drifting of gametes or the immature stages of plants or marine animals from one place to another (figure 20.4*b*). Pollen, the male gamete of flowering plants, is often carried great distances by insects and other animals that visit flowers. Seeds may also blow in the wind or be carried by animals to new populations far from their place of




***a.*** The ultimate source of variation. Individual mutations occur so rarely that mutation alone usually does not change allele frequency much.

***b.*** A very potent agent of change. Individuals or gametes move from one population to another.

***c.*** Inbreeding is the most common form. It does not alter allele frequency but reduces the proportion of heterozygotes.

***d.*** Statistical accidents. The random fluctuation in allele frequencies increases as population size decreases.

***e.*** The only agent that produces *adaptive* evolutionary changes.

**Figure 20.4 Five agents of evolutionary change.** ***a.*** Mutation, ***(b)*** gene flow, ***(c)*** nonrandom mating, ***(d)*** genetic drift, and ***(e)*** selection.

origin. In addition, gene flow may also result from the mating of individuals belonging to adjacent populations.

Consider two populations initially different in allele frequencies: In population 1, $p = 0.2$ and $q = 0.8$; in population 2, $p = 0.8$ and $q = 0.2$. Gene flow will tend to bring the rarer allele into each population. Thus, allele frequencies will change from generation to generation, and the populations will not be in Hardy–Weinberg equilibrium. Only when allele frequencies reach 0.5 for both alleles in both populations will equilibrium be attained. This example also indicates that gene flow tends to homogenize allele frequencies among populations.

## Nonrandom mating shifts genotype frequencies

Individuals with certain genotypes sometimes mate with one another more commonly than would be expected on a random basis, a phenomenon known as *nonrandom mating* (figure 20.4*c*). **Assortative mating,** in which phenotypically similar individuals mate, is a type of nonrandom mating that causes the frequencies of particular genotypes to differ greatly from those predicted by the Hardy–Weinberg principle.

Assortative mating does not change the frequency of the individual alleles, but rather increases the proportion of homozygous individuals because phenotypically similar individuals are likely to be genetically similar and thus are also more likely to produce offspring with two copies of the same allele. This is why populations of self-fertilizing plants consist primarily of homozygous individuals.

By contrast, **disassortative mating,** in which phenotypically different individuals mate, produces an excess of heterozygotes.

## Genetic drift may alter allele frequencies in small populations

In small populations, frequencies of particular alleles may change drastically by chance alone. Such changes in allele frequencies occur randomly, as if the frequencies were drifting from their values. These changes are thus known as **genetic drift** (figure 20.4*d*). For this reason, a population must be large to be in Hardy–Weinberg equilibrium.

If the gametes of only a few individuals form the next generation, the alleles they carry may by chance not be representative of the parent population from which they were drawn, as illustrated in figure 20.5. In this example, a small number of individuals are removed from a bottle containing many. By chance, most of the individuals removed are green, so the new population has a much higher population of green individuals than the parent generation had.

A set of small populations that are isolated from one another may come to differ strongly as a result of genetic drift, even if the forces of natural selection are the same for both. Because of genetic drift, sometimes harmful alleles may increase in frequency in small populations, despite selective disadvantage, and favorable alleles may be lost even though they are selectively advantageous. It is interesting to realize that humans have lived in small groups for much of the course of their evolution; consequently, genetic drift may have been a particularly important factor in the evolution of our species.

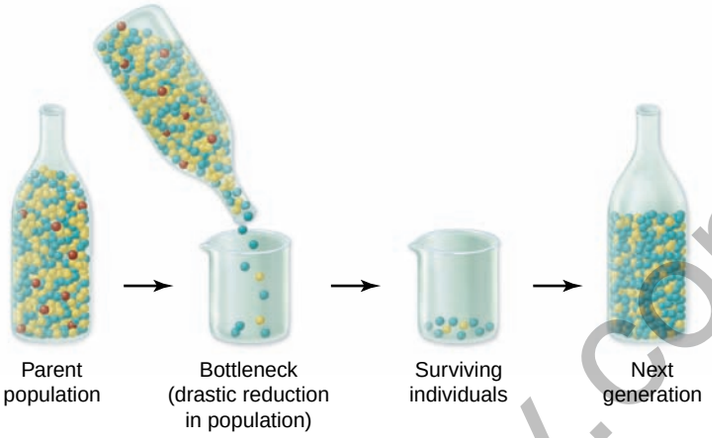



**Figure 20.5 Genetic drift: a bottleneck effect.** The parent population contains roughly equal numbers of green and yellow individuals and a small number of red individuals. By chance, the few remaining individuals that contribute to the next generation are mostly green. The bottleneck occurs because so few individuals form the next generation, as might happen after an epidemic or a catastrophic storm.

Larger populations also experience the effect of genetic drift, but to a lesser extent than smaller populations—the magnitude of genetic drift is inversely related to population size. However, large populations may have been much smaller in the past, and genetic drift may have greatly altered allele frequencies at that time. Imagine a population containing only two alleles of a gene, *B* and *b*, in equal frequency (that is, $p = q = 0.50$). In a large Hardy–Weinberg population, the genotype frequencies are expected to be 0.25 *BB*, 0.50 *Bb*, and 0.25 *bb*. If only a small sample of individuals produces the next generation, large deviations in these genotype frequencies can occur simply by chance.

Suppose, for example, that four individuals form the next generation, and that by chance they are two *Bb* heterozygotes and two *BB* homozygotes—that is, the allele frequencies in the next generation would be $p = 0.75$ and $q = 0.25$. In fact, if you were to replicate this experiment 1000 times, each time randomly drawing four individuals from the parental population, then in about 8 of the 1000 experiments, one of the two alleles would be missing entirely.

This result leads to an important conclusion: Genetic drift can lead to the loss of alleles in isolated populations. Alleles that initially are uncommon are particularly vulnerable (see figure 20.5).

Although genetic drift occurs in any population, it is particularly likely in populations that were founded by a few individuals or in which the population was reduced to a very small number at some time in the past.

### *The founder effect*

Sometimes one or a few individuals disperse and become the founders of a new, isolated population at some distance from their place of origin. These pioneers are not likely to carry all the alleles present in the source population. Thus, some alleles may be lost from the new population, and others may change