

chapter 9

DNA-BASED INFORMATION TECHNOLOGIES

- 9.1 DNA Cloning: The Basics 306
- 9.2 From Genes to Genomes 317
- 9.3 From Genomes to Proteomes 325
- 9.4 Genome Alterations and New Products of Biotechnology 330

Of all the natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history.

—Emile Zuckerkandl and Linus Pauling, article in *Journal of Theoretical Biology*, 1965

We now turn to a technology that is fundamental to the advance of modern biological sciences, defining present and future biochemical frontiers and illustrating many important principles of biochemistry. Elucidation of the laws governing enzymatic catalysis, macromolecular structure, cellular metabolism, and information pathways allows research to be directed at increasingly complex biochemical processes. Cell division, immunity, embryogenesis, vision, taste, oncogenesis, cognition—all are orchestrated in an elaborate symphony of molecular and macromolecular interactions that we are now beginning to understand with increasing clarity. The real implications of the biochemical journey begun in the nineteenth century are found in the ever-increasing power to analyze and alter living systems.

To understand a complex biological process, a biochemist isolates and studies the individual components *in vitro*, then pieces together the parts to get a coherent picture of the overall process. A major source of molecular insights is the cell's own information archive, its DNA. The sheer size of chromosomes, however, pres-

ents an enormous challenge: how does one find and study a particular gene among the tens of thousands of genes nested in the billions of base pairs of a mammalian genome? Solutions began to emerge in the 1970s.

Decades of advances by thousands of scientists working in genetics, biochemistry, cell biology, and physical chemistry came together in the laboratories of Paul Berg, Herbert Boyer, and Stanley Cohen to yield techniques for locating, isolating, preparing, and studying small segments of DNA derived from much larger chromosomes. Techniques for DNA cloning paved the way to the modern fields of **genomics** and **proteomics**, the study of genes and proteins on the scale of whole cells and organisms. These new methods are transforming basic research, agriculture, medicine, ecology, forensics, and many other fields, while occasionally presenting society with difficult choices and ethical dilemmas.

We begin this chapter with an outline of the fundamental biochemical principles of the now-classic discipline of DNA cloning. Next, after laying the groundwork for a discussion of genomics, we illustrate the range of applications and the potential of these technologies, with a broad emphasis on modern advances in genomics and proteomics.

9.1 DNA Cloning: The Basics

A *clone* is an identical copy. This term originally applied to cells of a single type, isolated and allowed to reproduce to create a population of identical cells. **DNA cloning** involves separating a specific gene or DNA segment from a larger chromosome, attaching it to a small molecule of carrier DNA, and then replicating this modified DNA thousands or millions of times through both an increase in cell number and the creation of multiple

copies of the cloned DNA in each cell. The result is selective amplification of a particular gene or DNA segment. Cloning of DNA from any organism entails five general procedures:

1. *Cutting DNA at precise locations.* Sequence-specific endonucleases (restriction endonucleases) provide the necessary molecular scissors.
2. *Selecting a small molecule of DNA capable of self-replication.* These DNAs are called **cloning vectors** (a vector is a delivery agent). They are typically plasmids or viral DNAs.
3. *Joining two DNA fragments covalently.* The enzyme DNA ligase links the cloning vector and DNA to be cloned. Composite DNA molecules comprising covalently linked segments from two or more sources are called **recombinant DNAs**.
4. *Moving recombinant DNA from the test tube to a host cell that will provide the enzymatic machinery for DNA replication.*
5. *Selecting or identifying host cells that contain recombinant DNA.*

The methods used to accomplish these and related tasks are collectively referred to as **recombinant DNA technology** or, more informally, **genetic engineering**.

Much of our initial discussion will focus on DNA cloning in the bacterium *Escherichia coli*, the first organism used for recombinant DNA work and still the most common host cell. *E. coli* has many advantages: its DNA metabolism (like many other of its biochemical processes) is well understood; many naturally occurring cloning vectors associated with *E. coli*, such as plasmids and bacteriophages (bacterial viruses; also called phages), are well characterized; and techniques are available for moving DNA expeditiously from one bac-



Paul Berg



Herbert Boyer



Stanley N. Cohen

terial cell to another. We also address DNA cloning in other organisms, a topic discussed more fully later in the chapter.

Restriction Endonucleases and DNA Ligase Yield Recombinant DNA

Particularly important to recombinant DNA technology is a set of enzymes (Table 9–1) made available through decades of research on nucleic acid metabolism. Two classes of enzymes lie at the heart of the general approach to generating and propagating a recombinant DNA molecule (Fig. 9–1). First, **restriction endonucleases** (also called restriction enzymes) recognize and cleave DNA at specific DNA sequences (recognition sequences or restriction sites) to generate a set of smaller fragments. Second, the DNA fragment to be cloned can be joined to a suitable cloning vector by using **DNA ligases** to link the DNA molecules together. The recombinant vector is then introduced into a host cell, which amplifies the fragment in the course of many generations of cell division.

Restriction endonucleases are found in a wide range of bacterial species. Werner Arber discovered in the early 1960s that their biological function is to recognize and cleave foreign DNA (the DNA of an infecting virus, for example); such DNA is said to be *restricted*. In the host cell's DNA, the sequence that would be recognized

TABLE 9–1 Some Enzymes Used in Recombinant DNA Technology

Enzyme(s)	Function
Type II restriction endonucleases	Cleave DNAs at specific base sequences
DNA ligase	Joins two DNA molecules or fragments
DNA polymerase I (<i>E. coli</i>)	Fills gaps in duplexes by stepwise addition of nucleotides to 3' ends
Reverse transcriptase	Makes a DNA copy of an RNA molecule
Polynucleotide kinase	Adds a phosphate to the 5'-OH end of a polynucleotide to label it or permit ligation
Terminal transferase	Adds homopolymer tails to the 3'-OH ends of a linear duplex
Exonuclease III	Removes nucleotide residues from the 3' ends of a DNA strand
Bacteriophage λ exonuclease	Removes nucleotides from the 5' ends of a duplex to expose single-stranded 3' ends
Alkaline phosphatase	Removes terminal phosphates from either the 5' or 3' end (or both)

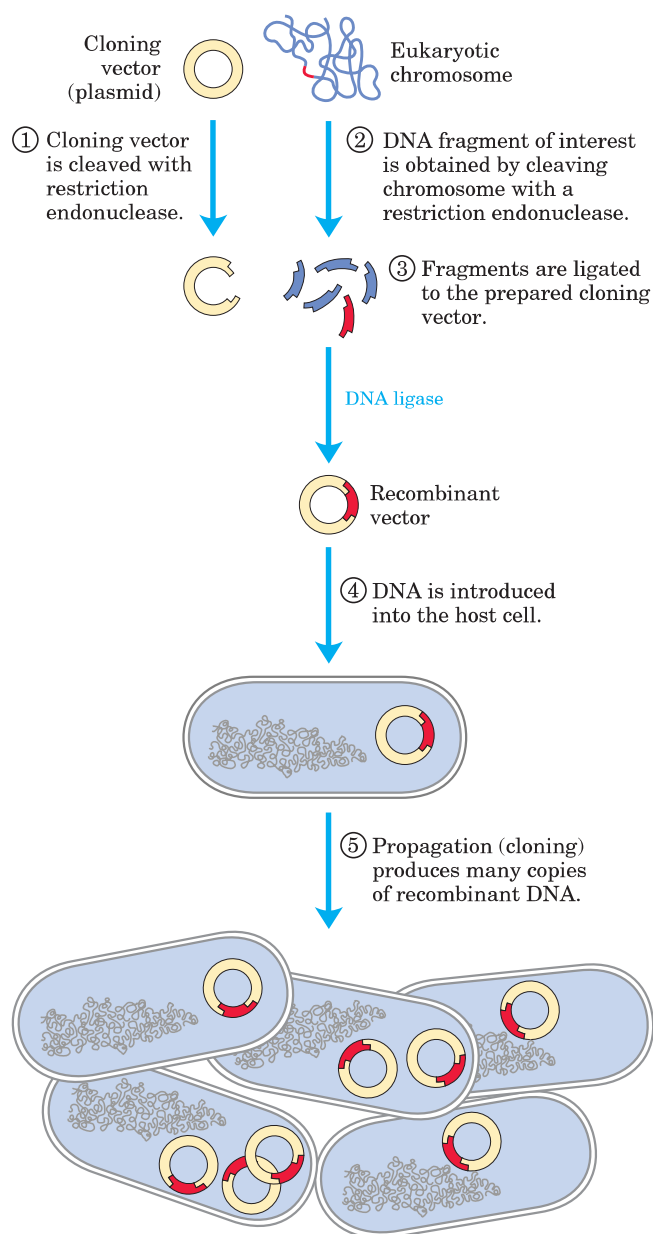


FIGURE 9-1 Schematic illustration of DNA cloning. A cloning vector and eukaryotic chromosomes are separately cleaved with the same restriction endonuclease. The fragments to be cloned are then ligated to the cloning vector. The resulting recombinant DNA (only one recombinant vector is shown here) is introduced into a host cell where it can be propagated (cloned). Note that this drawing is not to scale: the size of the *E. coli* chromosome relative to that of a typical cloning vector (such as a plasmid) is much greater than depicted here.

by its own restriction endonuclease is protected from digestion by methylation of the DNA, catalyzed by a specific DNA methylase. The restriction endonuclease and the corresponding methylase are sometimes referred to as a **restriction-modification system**.

There are three types of restriction endonucleases, designated I, II, and III. Types I and III are generally large, multisubunit complexes containing both the endonucle-

ase and methylase activities. Type I restriction endonucleases cleave DNA at random sites that can be more than 1,000 base pairs (bp) from the recognition sequence. Type III restriction endonucleases cleave the DNA about 25 bp from the recognition sequence. Both types move along the DNA in a reaction that requires the energy of ATP. **Type II restriction endonucleases**, first isolated by Hamilton Smith in 1970, are simpler, require no ATP, and cleave the DNA within the recognition sequence itself. The extraordinary utility of this group of restriction endonucleases was demonstrated by Daniel Nathans, who first used them to develop novel methods for mapping and analyzing genes and genomes.

Thousands of restriction endonucleases have been discovered in different bacterial species, and more than 100 different DNA sequences are recognized by one or more of these enzymes. The recognition sequences are usually 4 to 6 bp long and palindromic (see Fig. 8–20). Table 9–2 lists sequences recognized by a few type II restriction endonucleases. In some cases, the interaction between a restriction endonuclease and its target sequence has been elucidated in exquisite molecular detail; for example, Figure 9–2 shows the complex of the type II restriction endonuclease *EcoRV* and its target sequence.

Some restriction endonucleases make staggered cuts on the two DNA strands, leaving two to four nucleotides of one strand unpaired at each resulting end. These unpaired strands are referred to as **sticky ends** (Fig. 9–3a), because they can base-pair with each other or with complementary sticky ends of other DNA fragments. Other restriction endonucleases cleave both strands of DNA at the opposing phosphodiester bonds, leaving no unpaired bases on the ends, often called **blunt ends** (Fig. 9–3b).

The average size of the DNA fragments produced by cleaving genomic DNA with a restriction endonuclease depends on the frequency with which a particular restriction site occurs in the DNA molecule; this in turn depends largely on the size of the recognition sequence. In a DNA molecule with a random sequence in which all four nucleotides were equally abundant, a 6 bp sequence recognized by a restriction endonuclease such as *Bam*HI would occur on average once every 4^6 (4,096) bp, assuming the DNA had a 50% G≡C content. Enzymes that recognize a 4 bp sequence would produce smaller DNA fragments from a random-sequence DNA molecule; a recognition sequence of this size would be expected to occur about once every 4^4 (256) bp. In natural DNA molecules, particular recognition sequences tend to occur less frequently than this because nucleotide sequences in DNA are not random and the four nucleotides are not equally abundant. In laboratory experiments, the average size of the fragments produced by restriction endonuclease cleavage of a large DNA can be increased by simply terminating the reaction before completion; the result is called a partial digest. Fragment size can also

TABLE 9–2 Recognition Sequences for Some Type II Restriction Endonucleases

<i>Bam</i> HI	$\begin{array}{c} \downarrow \\ (5') \text{ G G A T C C } (3') \\ \text{C C T A G G} \\ \uparrow \end{array}$	<i>Hind</i> III	$\begin{array}{c} \downarrow \\ (5') \text{ A A G C T T } (3') \\ \text{T T C G A A} \\ \uparrow \end{array}$
<i>Cl</i> al	$\begin{array}{c} \downarrow \\ (5') \text{ A T C G A T } (3') \\ \text{T A G C T A} \\ \uparrow \end{array}$	<i>Not</i> I	$\begin{array}{c} \downarrow \\ (5') \text{ G C G G C C G C } (3') \\ \text{C G C C G G C G} \\ \uparrow \end{array}$
<i>Eco</i> RI	$\begin{array}{c} \downarrow \\ (5') \text{ G A A T T C } (3') \\ \text{C T T A A G} \\ \uparrow \end{array}$	<i>Pst</i> I	$\begin{array}{c} \downarrow \\ (5') \text{ C T G C A G } (3') \\ \text{G A C G T C} \\ \uparrow \end{array}$
<i>Eco</i> RV	$\begin{array}{c} \downarrow \\ (5') \text{ G A T A T C } (3') \\ \text{C T A T A G} \\ \uparrow \end{array}$	<i>Pvu</i> II	$\begin{array}{c} \downarrow \\ (5') \text{ C A G C T G } (3') \\ \text{G T C G A C} \\ \uparrow \end{array}$
<i>Hae</i> III	$\begin{array}{c} \downarrow \\ (5') \text{ G G C C } (3') \\ \text{C C G G} \\ \uparrow \end{array}$	<i>Tth</i> 111I	$\begin{array}{c} \downarrow \\ (5') \text{ G A C N N N G T C } (3') \\ \text{C T G N N N C A G} \\ \uparrow \end{array}$

Arrows indicate the phosphodiester bonds cleaved by each restriction endonuclease. Asterisks indicate bases that are methylated by the corresponding methylase (where known). N denotes any base. Note that the name of each enzyme consists of a three-letter abbreviation (in italics) of the bacterial species from which it is derived, sometimes followed by a strain designation and Roman numerals to distinguish different restriction endonucleases isolated from the same bacterial species. Thus *Bam*HI is the first (I) restriction endonuclease characterized from *Bacillus amyloliquefaciens*, strain H.

be increased by using a special class of endonucleases called homing endonucleases (see Fig. 26–34). These recognize and cleave much longer DNA sequences (14 to 20 bp).

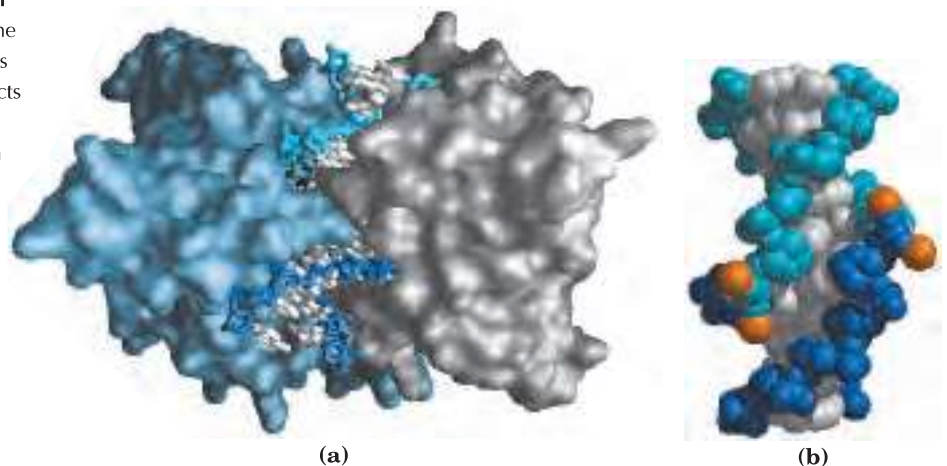
Once a DNA molecule has been cleaved into fragments, a particular fragment of known size can be enriched by agarose or acrylamide gel electrophoresis or by HPLC (pp. 92, 90). For a typical mammalian genome, however, cleavage by a restriction endonuclease usually yields too many different DNA fragments to permit isolation of a particular fragment by electrophoresis or

HPLC. A common intermediate step in the cloning of a specific gene or DNA segment is the construction of a DNA library (as described in Section 9.2).

After the target DNA fragment is isolated, DNA ligase can be used to join it to a similarly digested cloning vector—that is, a vector digested by the *same* restriction endonuclease; a fragment generated by *Eco*RI, for example, generally will not link to a fragment generated by *Bam*HI. As described in more detail in Chapter 25 (see Fig. 25–16), DNA ligase catalyzes the formation of new phosphodiester bonds in a reaction that uses ATP

FIGURE 9–2 Interaction of *Eco*RV restriction endonuclease with its target sequence. (a) The dimeric *Eco*RV endonuclease (its two subunits in light blue and gray) is bound to the products of DNA cleavage at the sequence recognized by the enzyme. The DNA backbone is shown in two shades of blue to distinguish the segments separated by cleavage (PDB ID 1RVC). (b) In this view, showing just the DNA, the DNA segment has been turned 180°. The enzyme creates blunt ends; the cleavage points appear staggered on the two DNA strands because the DNA is kinked. Bound magnesium ions (orange) play a role in catalysis of the cleavage reaction.

 **Restriction Endonucleases**



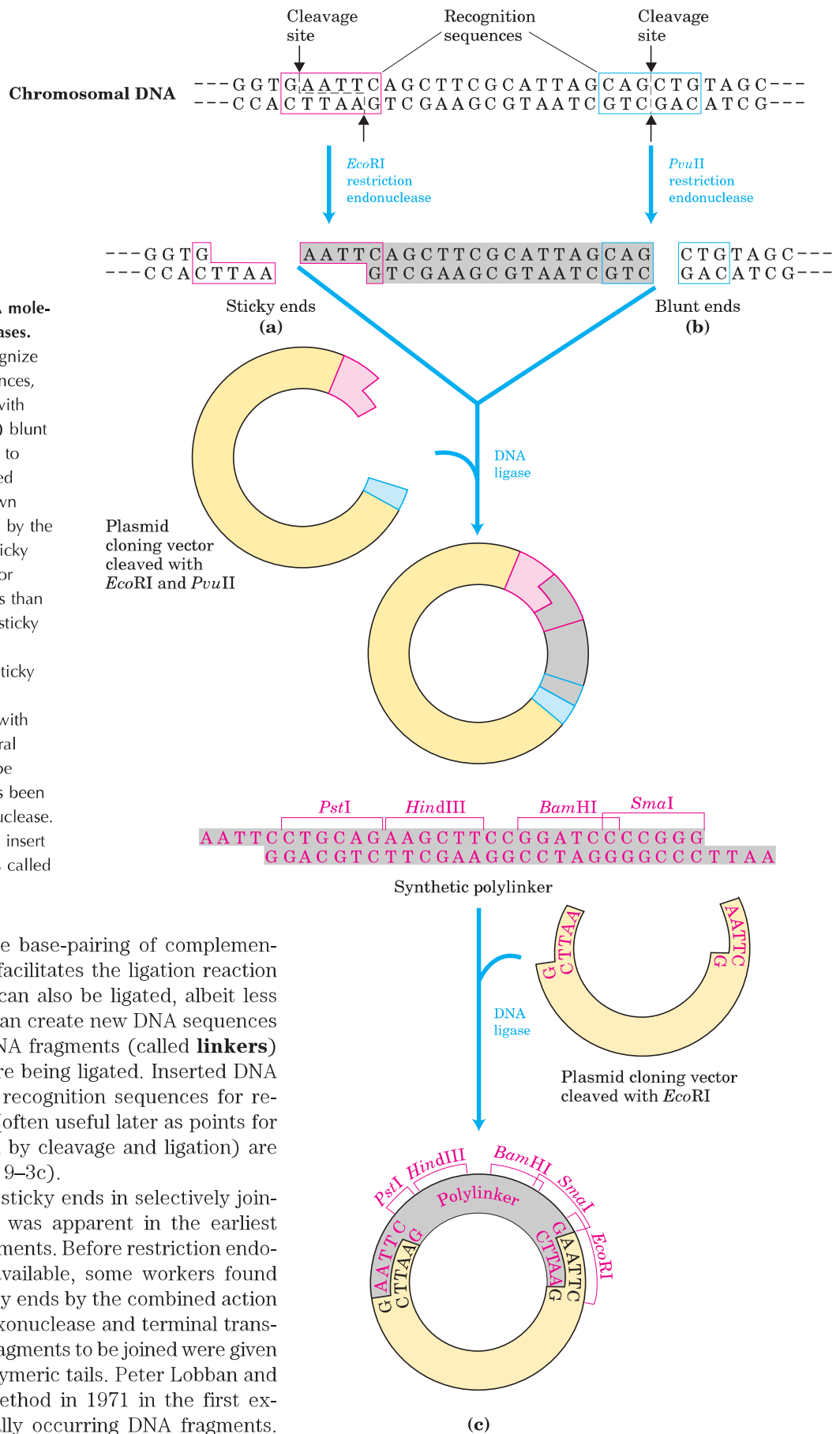


FIGURE 9-3 Cleavage of DNA molecules by restriction endonucleases.

Restriction endonucleases recognize and cleave only specific sequences, leaving either (a) sticky ends (with protruding single strands) or (b) blunt ends. Fragments can be ligated to other DNAs, such as the cleaved cloning vector (a plasmid) shown here. This reaction is facilitated by the annealing of complementary sticky ends. Ligation is less efficient for DNA fragments with blunt ends than for those with complementary sticky ends, and DNA fragments with different (noncomplementary) sticky ends generally are not ligated. (c) A synthetic DNA fragment with recognition sequences for several restriction endonucleases can be inserted into a plasmid that has been cleaved by a restriction endonuclease. The insert is called a linker; an insert with multiple restriction sites is called a polylinker.

or a similar cofactor. The base-pairing of complementary sticky ends greatly facilitates the ligation reaction (Fig. 9-3a). Blunt ends can also be ligated, albeit less efficiently. Researchers can create new DNA sequences by inserting synthetic DNA fragments (called **linkers**) between the ends that are being ligated. Inserted DNA fragments with multiple recognition sequences for restriction endonucleases (often useful later as points for inserting additional DNA by cleavage and ligation) are called **polylinkers** (Fig. 9-3c).

The effectiveness of sticky ends in selectively joining two DNA fragments was apparent in the earliest recombinant DNA experiments. Before restriction endonucleases were widely available, some workers found they could generate sticky ends by the combined action of the bacteriophage λ exonuclease and terminal transferase (Table 9-1). The fragments to be joined were given complementary homopolymeric tails. Peter Lobban and Dale Kaiser used this method in 1971 in the first experiments to join naturally occurring DNA fragments.

Similar methods were used soon after in the laboratory of Paul Berg to join DNA segments from simian virus 40 (SV40) to DNA derived from bacteriophage λ , thereby creating the first recombinant DNA molecule with DNA segments from different species.

Cloning Vectors Allow Amplification of Inserted DNA Segments

The principles that govern the delivery of recombinant DNA in clonable form to a host cell, and its subsequent amplification in the host, are well illustrated by considering three popular cloning vectors commonly used in experiments with *E. coli*—plasmids, bacteriophages, and bacterial artificial chromosomes—and a vector used to clone large DNA segments in yeast.

Plasmids Plasmids are circular DNA molecules that replicate separately from the host chromosome. Naturally occurring bacterial plasmids range in size from 5,000 to 400,000 bp. They can be introduced into bacterial cells by a process called **transformation**. The cells (generally *E. coli*) and plasmid DNA are incubated together at 0 °C in a calcium chloride solution, then subjected to a shock by rapidly shifting the temperature to 37 to 43 °C. For reasons not well understood, some of the cells treated in this way take up the plasmid DNA. Some species of bacteria are naturally competent for DNA uptake and do not require the calcium chloride treatment. In an alternative method, cells incubated with the plasmid DNA are subjected to a high-voltage pulse. This approach, called **electroporation**, transiently renders the bacterial membrane permeable to large molecules.

Regardless of the approach, few cells actually take up the plasmid DNA, so a method is needed to select those that do. The usual strategy is to use a plasmid that includes a gene that the host cell requires for growth under specific conditions, such as a gene that confers resistance to an antibiotic. Only cells transformed by the recombinant plasmid can grow in the presence of that antibiotic, making any cell that contains the plasmid “selectable” under those growth conditions. Such a gene is called a selectable marker.

Investigators have developed many different plasmid vectors suitable for cloning by modifying naturally occurring plasmids. The *E. coli* plasmid pBR322 offers a good example of the features useful in a cloning vector (Fig. 9–4):

1. pBR322 has an origin of replication, *ori*, a sequence where replication is initiated by cellular enzymes (Chapter 25). This sequence is required to propagate the plasmid and maintain it at a level of 10 to 20 copies per cell.
2. The plasmid contains two genes that confer resistance to different antibiotics (*tet^R*, *amp^R*),

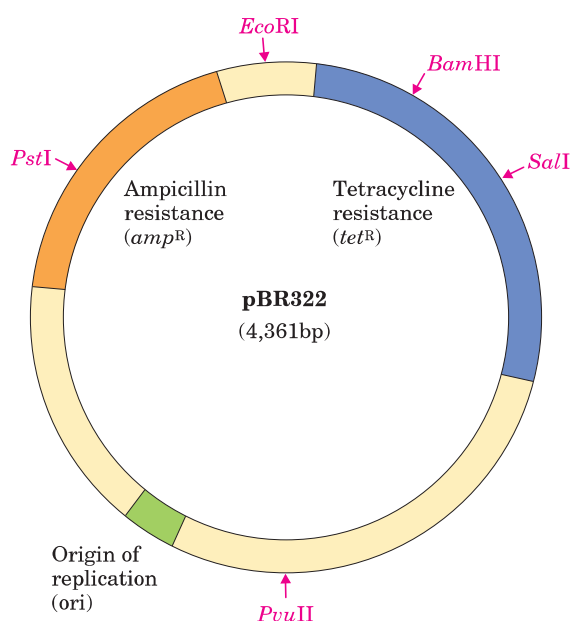


FIGURE 9–4 The constructed *E. coli* plasmid pBR322. Note the location of some important restriction sites—for *Pst*I, *Eco*RI, *Bam*HI, *Sal*I, and *Pvu*II; ampicillin- and tetracycline-resistance genes; and the replication origin (*ori*). Constructed in 1977, this was one of the early plasmids designed expressly for cloning in *E. coli*.

allowing the identification of cells that contain the intact plasmid or a recombinant version of the plasmid (Fig. 9–5).

3. Several unique recognition sequences in pBR322 (*Pst*I, *Eco*RI, *Bam*HI, *Sal*I, *Pvu*II) are targets for different restriction endonucleases, providing sites where the plasmid can later be cut to insert foreign DNA.
4. The small size of the plasmid (4,361 bp) facilitates its entry into cells and the biochemical manipulation of the DNA.

Transformation of typical bacterial cells with purified DNA (never a very efficient process) becomes less successful as plasmid size increases, and it is difficult to clone DNA segments longer than about 15,000 bp when plasmids are used as the vector.

Bacteriophages Bacteriophage λ has a very efficient mechanism for delivering its 48,502 bp of DNA into a bacterium, and it can be used as a vector to clone somewhat larger DNA segments (Fig. 9–6). Two key features contribute to its utility:

1. About one-third of the λ genome is nonessential and can be replaced with foreign DNA.
2. DNA is packaged into infectious phage particles only if it is between 40,000 and 53,000 bp long, a constraint that can be used to ensure packaging of recombinant DNA only.

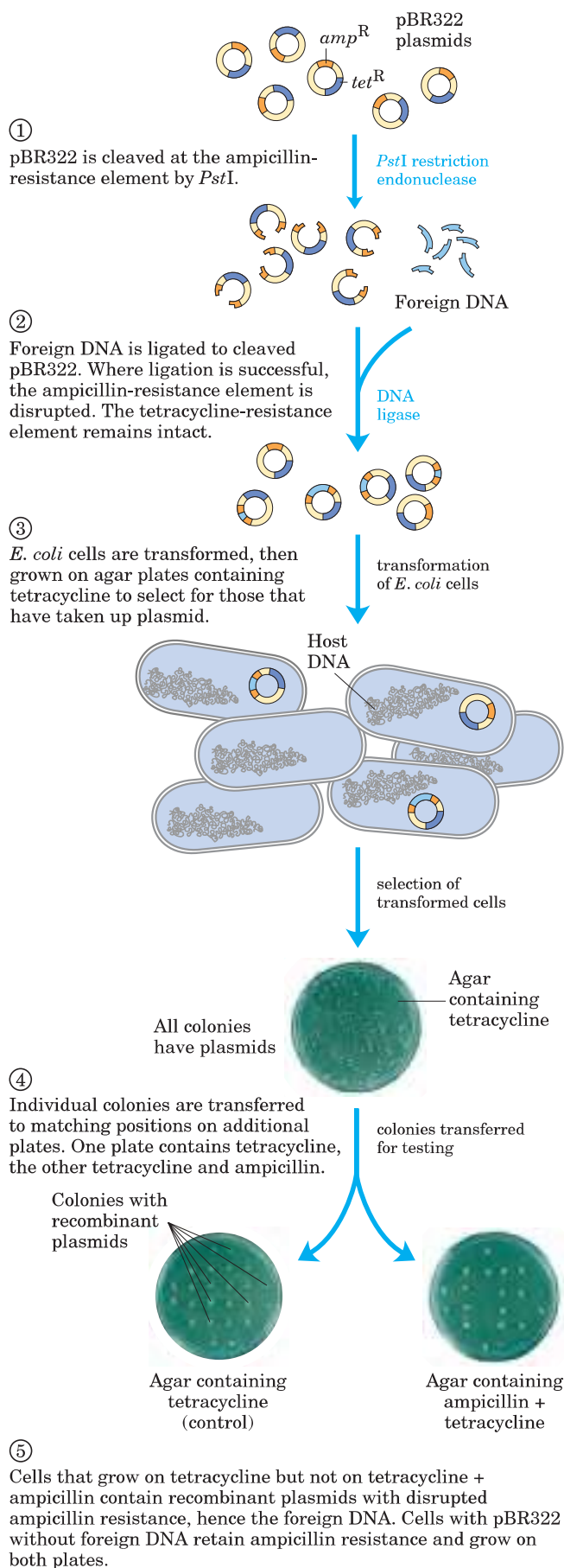


FIGURE 9-5 Use of pBR322 to clone and identify foreign DNA in *E. coli*. **Plasmid Cloning**

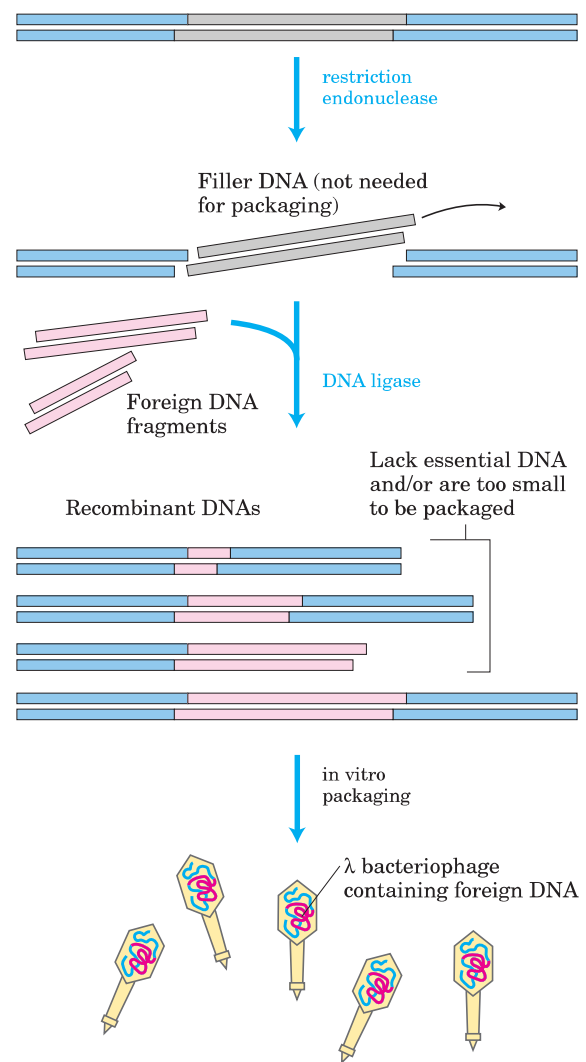


FIGURE 9-6 Bacteriophage λ cloning vectors. Recombinant DNA methods are used to modify the bacteriophage λ genome, removing the genes not needed for phage production and replacing them with “filler” DNA to make the phage DNA large enough for packaging into phage particles. As shown here, the filler is replaced with foreign DNA in cloning experiments. Recombinants are packaged into viable phage particles in vitro only if they include an appropriately sized foreign DNA fragment as well as both of the essential λ DNA end fragments.

Researchers have developed bacteriophage λ vectors that can be readily cleaved into three pieces, two of which contain essential genes but which together are only about 30,000 bp long. The third piece, “filler” DNA, is discarded when the vector is to be used for cloning, and additional DNA is inserted between the two essential segments to generate ligated DNA molecules long enough to produce viable phage particles. In effect, the packaging mechanism *selects for* recombinant viral DNAs.

Bacteriophage λ vectors permit the cloning of DNA fragments of up to 23,000 bp. Once the bacteriophage λ fragments are ligated to foreign DNA fragments of suitable size, the resulting recombinant DNAs can be pack-

aged into phage particles by adding them to crude bacterial cell extracts that contain all the proteins needed to assemble a complete phage. This is called **in vitro packaging** (Fig. 9–6). All viable phage particles will contain a foreign DNA fragment. The subsequent transmission of the recombinant DNA into *E. coli* cells is highly efficient.

Bacterial Artificial Chromosomes (BACs) Bacterial artificial chromosomes are simply plasmids designed for the cloning of very long segments (typically 100,000 to 300,000 bp) of DNA (Fig. 9–7). They generally include selectable markers such as resistance to the antibiotic chloramphenicol (Cm^R), as well as a very stable origin of replication (*ori*) that maintains the plasmid at one or two copies per cell. DNA fragments of several hundred thousand base pairs are cloned into the BAC vector. The large circular DNAs are then introduced into host bacteria by electroporation. These procedures use host bacteria with mutations that compromise the structure of their cell wall, permitting the uptake of the large DNA molecules.

Yeast Artificial Chromosomes (YACs) *E. coli* cells are by no means the only hosts for genetic engineering. Yeasts are particularly convenient eukaryotic organisms for this work. As with *E. coli*, yeast genetics is a well-developed discipline. The genome of the most commonly used yeast, *Saccharomyces cerevisiae*, contains only 14×10^6 bp (a simple genome by eukaryotic standards, less than four times the size of the *E. coli* chromosome), and its entire sequence is known. Yeast is also very easy to maintain and grow on a large scale in the laboratory. Plasmid vectors have been constructed for yeast, employing the same principles that govern the use of *E. coli* vectors described above. Convenient methods are now available for moving DNA into and out of yeast cells, facilitating the study of many aspects of eukaryotic cell biochemistry. Some recombinant plasmids incorporate multiple replication origins and other elements that allow them to be used in more than one species (for example, yeast or *E. coli*). Plasmids that can be propagated in cells of two or more different species are called **shuttle vectors**.

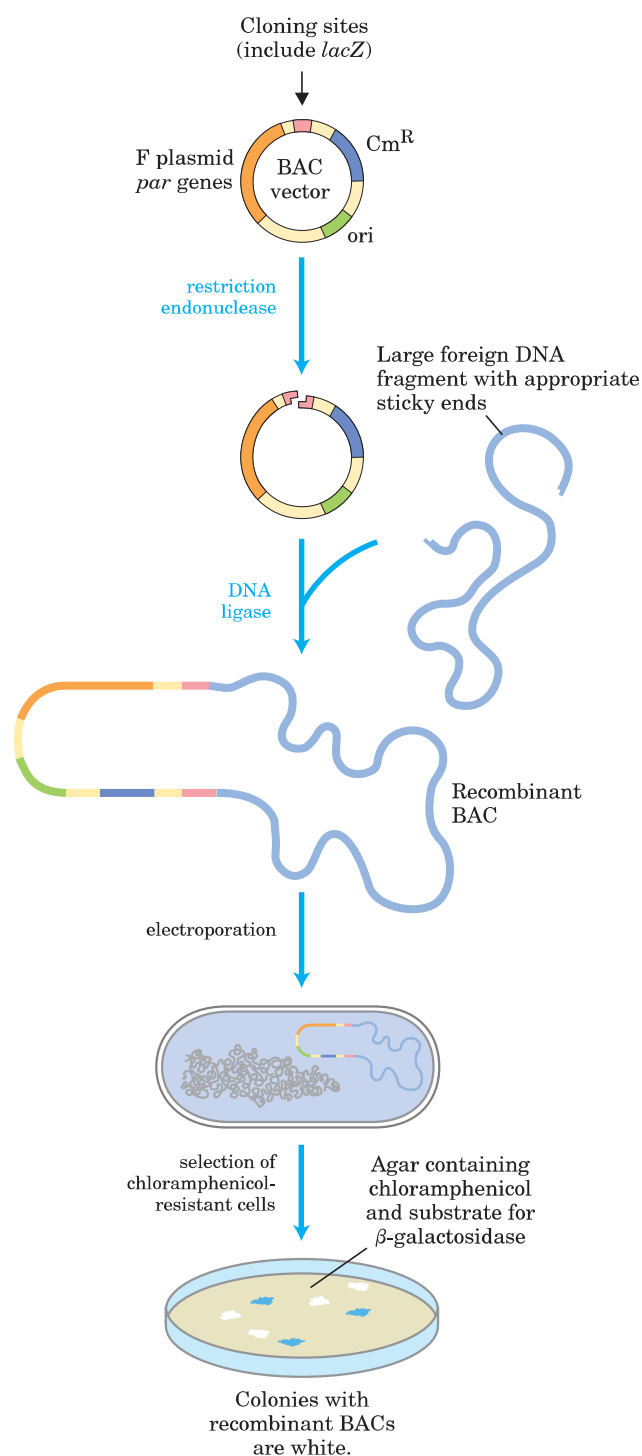


FIGURE 9–7 (above right) Bacterial artificial chromosomes (BACs) as cloning vectors. The vector is a relatively simple plasmid, with a replication origin (*ori*) that directs replication. The *par* genes, derived from a type of plasmid called an F plasmid, assist in the even distribution of plasmids to daughter cells at cell division. This increases the likelihood of each daughter cell carrying one copy of the plasmid, even when few copies are present. The low number of copies is useful in cloning large segments of DNA because it limits the opportunities for unwanted recombination reactions that can unpredictably alter large cloned DNAs over time. The BAC includes selectable markers. A *lacZ*

gene (required for the production of the enzyme β -galactosidase) is situated in the cloning region such that it is inactivated by cloned DNA inserts. Introduction of recombinant BACs into cells by electroporation is promoted by the use of cells with an altered (more porous) cell wall. Recombinant DNAs are screened for resistance to the antibiotic chloramphenicol (Cm^R). Plates also contain a substrate for β -galactosidase that yields a colored product. Colonies with active β -galactosidase and hence no DNA insert in the BAC vector turn blue; colonies without β -galactosidase activity—and thus with the desired DNA inserts—are white.

Research work with large genomes and the associated need for high-capacity cloning vectors led to the development of **yeast artificial chromosomes (YACs; Fig. 9–8)**. YAC vectors contain all the elements needed to maintain a eukaryotic chromosome in the yeast nucleus: a yeast origin of replication, two selectable markers, and specialized sequences (derived from the telomeres and centromere, regions of the chromosome discussed in Chapter 24) needed for stability and

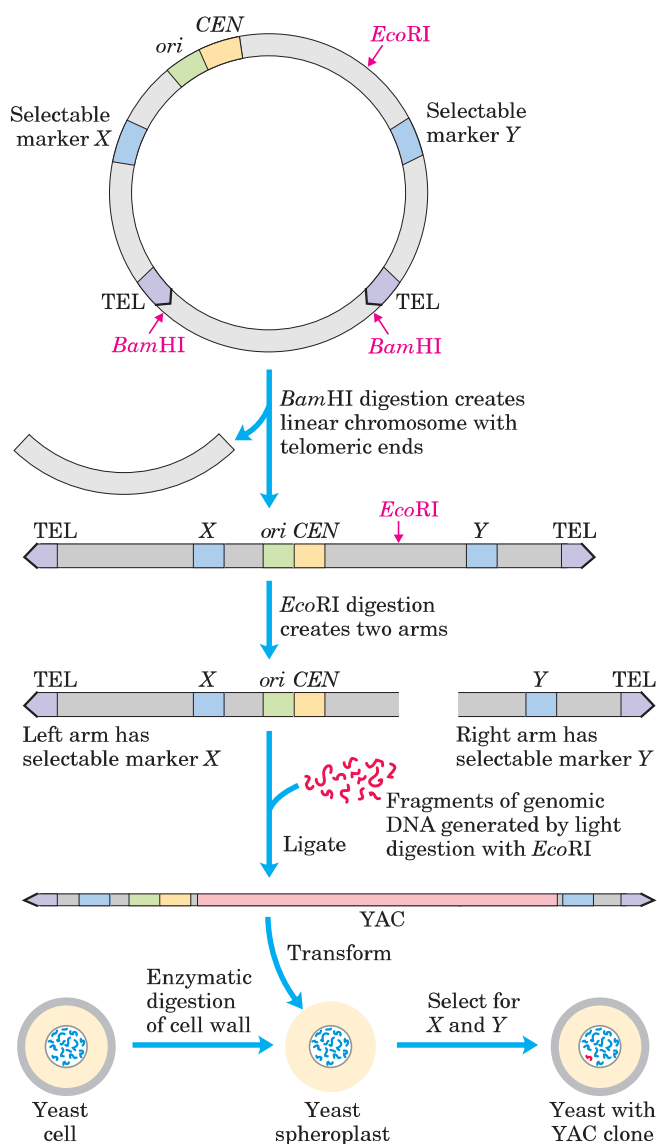


FIGURE 9–8 Construction of a yeast artificial chromosome (YAC). A YAC vector includes an origin of replication (*ori*), a centromere (*CEN*), two telomeres (*TEL*), and selectable markers (*X* and *Y*). Digestion with *Bam*HI and *Eco*RI generates two separate DNA arms, each with a telomeric end and one selectable marker. A large segment of DNA (e.g., up to 2×10^6 bp from the human genome) is ligated to the two arms to create a yeast artificial chromosome. The YAC transforms yeast cells (prepared by removal of the cell wall to form spheroplasts), and the cells are selected for *X* and *Y*; the surviving cells propagate the DNA insert.

proper segregation of the chromosomes at cell division. Before being used in cloning, the vector is propagated as a circular bacterial plasmid. Cleavage with a restriction endonuclease (*Bam*HI in Fig. 9–8) removes a length of DNA between two telomere sequences (*TEL*), leaving the telomeres at the ends of the linearized DNA. Cleavage at another internal site (*Eco*RI in Fig. 9–8) divides the vector into two DNA segments, referred to as vector arms, each with a different selectable marker.

The genomic DNA is prepared by partial digestion with restriction endonucleases (*Eco*RI in Fig. 9–8) to obtain a suitable fragment size. Genomic fragments are then separated by **pulsed field gel electrophoresis**, a variation of gel electrophoresis (see Fig. 3–19) that allows the separation of very large DNA segments. The DNA fragments of appropriate size (up to about 2×10^6 bp) are mixed with the prepared vector arms and ligated. The ligation mixture is then used to transform treated yeast cells with very large DNA molecules. Culture on a medium that requires the presence of both selectable marker genes ensures the growth of only those yeast cells that contain an artificial chromosome with a large insert sandwiched between the two vector arms (Fig. 9–8). The stability of YAC clones increases with size (up to a point). Those with inserts of more than 150,000 bp are nearly as stable as normal cellular chromosomes, whereas those with inserts of less than 100,000 bp are gradually lost during mitosis (so generally there are no yeast cell clones carrying only the two vector ends ligated together or with only short inserts). YACs that lack a telomere at either end are rapidly degraded.

Specific DNA Sequences Are Detectable by Hybridization

DNA hybridization, a process outlined in Chapter 8 (see Fig. 8–32), is the most common sequence-based process for detecting a particular gene or segment of nucleic acid. There are many variations of the basic method, most making use of a labeled (such as radioactive) DNA or RNA fragment, known as a **probe**, complementary to the DNA being sought. In one classic approach to detect a particular DNA sequence within a DNA library (a collection of DNA clones), nitrocellulose paper is pressed onto an agar plate containing many individual bacterial colonies from the library, each colony with a different recombinant DNA. Some cells from each colony adhere to the paper, forming a replica of the plate. The paper is treated with alkali to disrupt the cells and denature the DNA within, which remains bound to the region of the paper around the colony from which it came. Added radioactive DNA probe anneals only to its complementary DNA. After any unannealed probe DNA is washed away, the hybridized DNA can be detected by autoradiography (Fig. 9–9).

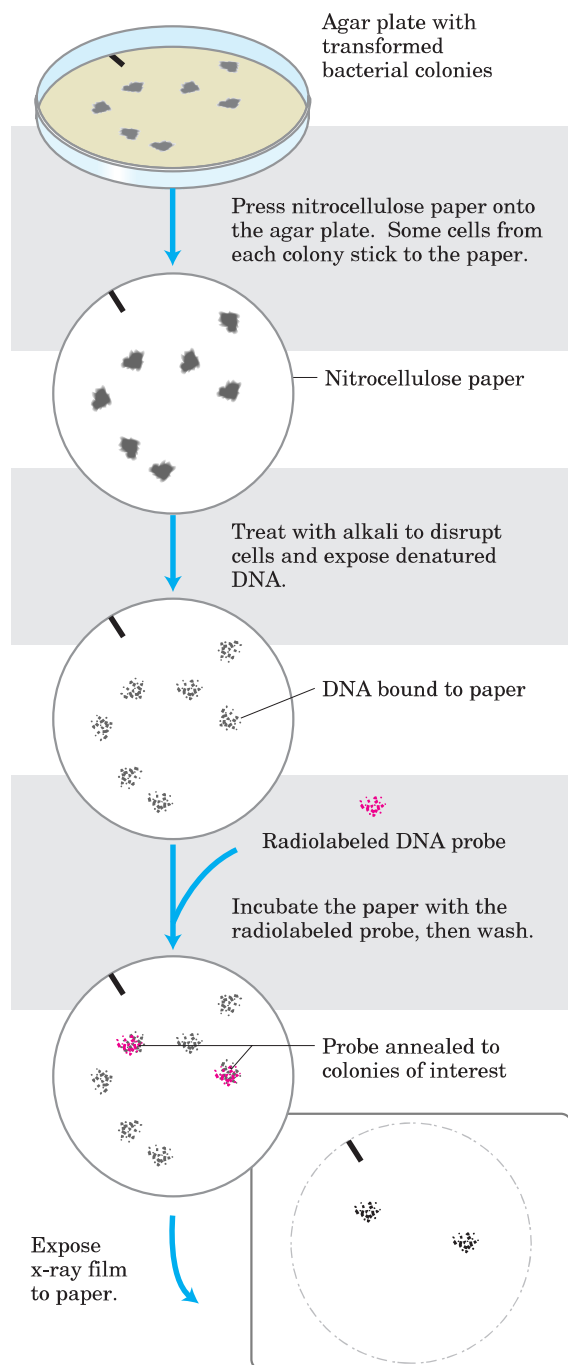


FIGURE 9-9 Use of hybridization to identify a clone with a particular DNA segment. The radioactive DNA probe hybridizes to complementary DNA and is revealed by autoradiography. Once the labeled colonies have been identified, the corresponding colonies on the original agar plate can be used as a source of cloned DNA for further study.

A common limiting step in detecting and cloning a gene is the generation of a complementary strand of nucleic acid to use as a probe. The origin of a probe depends on what is known about the gene under investigation. Sometimes a homologous gene cloned from

another species makes a suitable probe. Or, if the protein product of a gene has been purified, probes can be designed and synthesized by working backward from the amino acid sequence, deducing the DNA sequence that would code for it (Fig. 9–10). Now, researchers typically obtain the necessary DNA sequence information from sequence databases that detail the structure of millions of genes from a wide range of organisms.

Expression of Cloned Genes Produces Large Quantities of Protein

Frequently it is the product of the cloned gene, rather than the gene itself, that is of primary interest—particularly when the protein has commercial, therapeutic, or research value. With an increased understanding of the fundamentals of DNA, RNA, and protein metabolism and their regulation in *E. coli*, investigators can now manipulate cells to express cloned genes in order to study their protein products.

Most eukaryotic genes lack the DNA sequence elements—such as promoters, sequences that instruct RNA polymerase where to bind—required for their expression in *E. coli* cells, so bacterial regulatory sequences for transcription and translation must be inserted at appropriate positions relative to the eukaryotic gene in the vector DNA. (Promoters, regulatory sequences, and other aspects of the regulation of gene expression are discussed in Chapter 28.) In some cases cloned genes are so efficiently expressed that their protein product represents 10% or more of the cellular protein; they are said to be overexpressed. At these concentrations some foreign proteins can kill an *E. coli* cell, so gene expression must be limited to the few hours before the planned harvest of the cells.

Cloning vectors with the transcription and translation signals needed for the regulated expression of a cloned gene are often called **expression vectors**. The rate of expression of the cloned gene is controlled by replacing the gene's own promoter and regulatory sequences with more efficient and convenient versions supplied by the vector. Generally, a well-characterized promoter and its regulatory elements are positioned near several unique restriction sites for cloning, so that genes inserted at the restriction sites will be expressed from the regulated promoter element (Fig. 9–11). Some of these vectors incorporate other features, such as a bacterial ribosome binding site to enhance translation of the mRNA derived from the gene, or a transcription termination sequence.

Genes can similarly be cloned and expressed in eukaryotic cells, with various species of yeast as the usual hosts. A eukaryotic host can sometimes promote post-translational modifications (changes in protein structure made after synthesis on the ribosomes) that might be required for the function of a cloned eukaryotic protein.

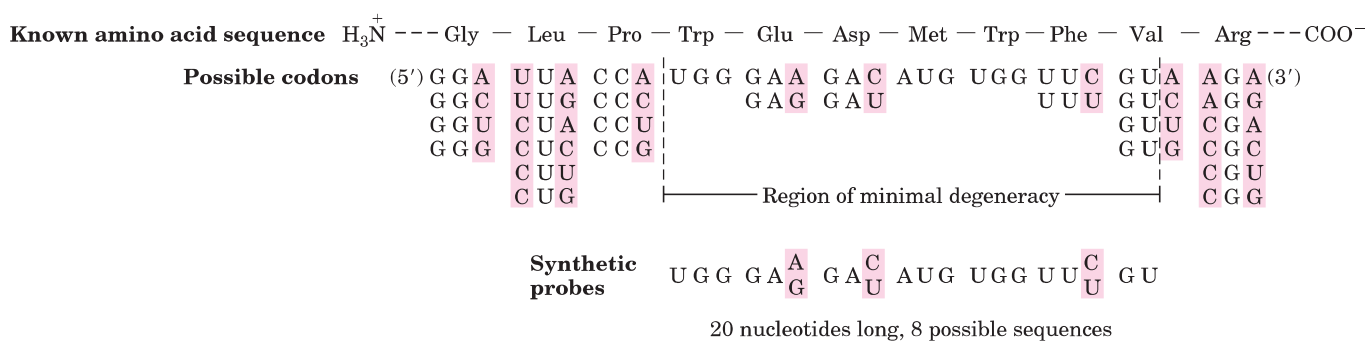


FIGURE 9-10 Probe to detect the gene for a protein of known amino acid sequence. Because more than one DNA sequence can code for any given amino acid sequence, the genetic code is said to be “degenerate.” (As described in Chapter 27, an amino acid is coded for by a set of three nucleotides called a *codon*. Most amino acids have two or more codons; see Fig. 27–7.) Thus the correct DNA sequence for a known amino acid sequence cannot be known in advance. The probe is designed to be complementary to a region of the gene with

minimal degeneracy, that is, a region with the fewest possible codons for the amino acids—two codons at most in the example shown here. Oligonucleotides are synthesized with selectively randomized sequences, so that they contain either of the two possible nucleotides at each position of potential degeneracy (shaded in pink). The oligonucleotide shown here represents a mixture of eight different sequences: one of the eight will complement the gene perfectly, and all eight will match at least 17 of the 20 positions.

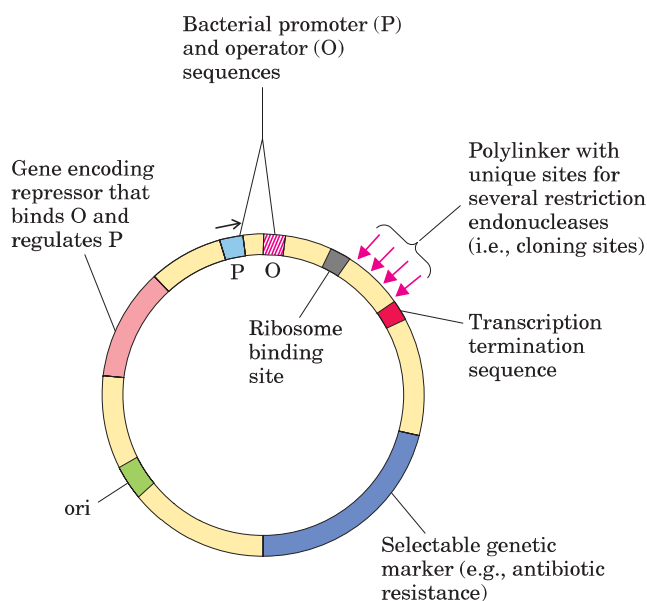


FIGURE 9-11 DNA sequences in a typical *E. coli* expression vector.

The gene to be expressed is inserted into one of the restriction sites in the polylinker, near the promoter (P), with the end encoding the amino terminus proximal to the promoter. The promoter allows efficient transcription of the inserted gene, and the transcription termination sequence sometimes improves the amount and stability of the mRNA produced. The operator (O) permits regulation by means of a repressor that binds to it (Chapter 28). The ribosome binding site provides sequence signals needed for efficient translation of the mRNA derived from the gene. The selectable marker allows the selection of cells containing the recombinant DNA.

Alterations in Cloned Genes Produce Modified Proteins

Cloning techniques can be used not only to overproduce proteins but to produce protein products subtly altered from their native forms. Specific amino acids may be replaced individually by **site-directed mutagenesis**. This powerful approach to studying protein structure and function changes the amino acid sequence of a protein by altering the DNA sequence of the cloned gene. If appropriate restriction sites flank the sequence to be altered, researchers can simply remove a DNA segment and replace it with a synthetic one that is identical to the original except for the desired change (Fig. 9–12a). When suitably located restriction sites are not present, an approach called **oligonucleotide-directed mutagenesis** (Fig. 9–12b) can create a specific DNA sequence change. A short synthetic DNA strand with a specific base change is annealed to a single-stranded copy of the cloned gene within a suitable vector. The mismatch of a single base pair in 15 to 20 bp does not prevent annealing if it is done at an appropriate temperature. The annealed strand serves as a primer for the synthesis of a strand complementary to the plasmid vector. This slightly mismatched duplex recombinant plasmid is then used to transform bacteria, where the mismatch is repaired by cellular DNA repair enzymes (Chapter 25). About half of the repair events will remove and replace the altered base and restore the gene to its original sequence; the other half will remove and

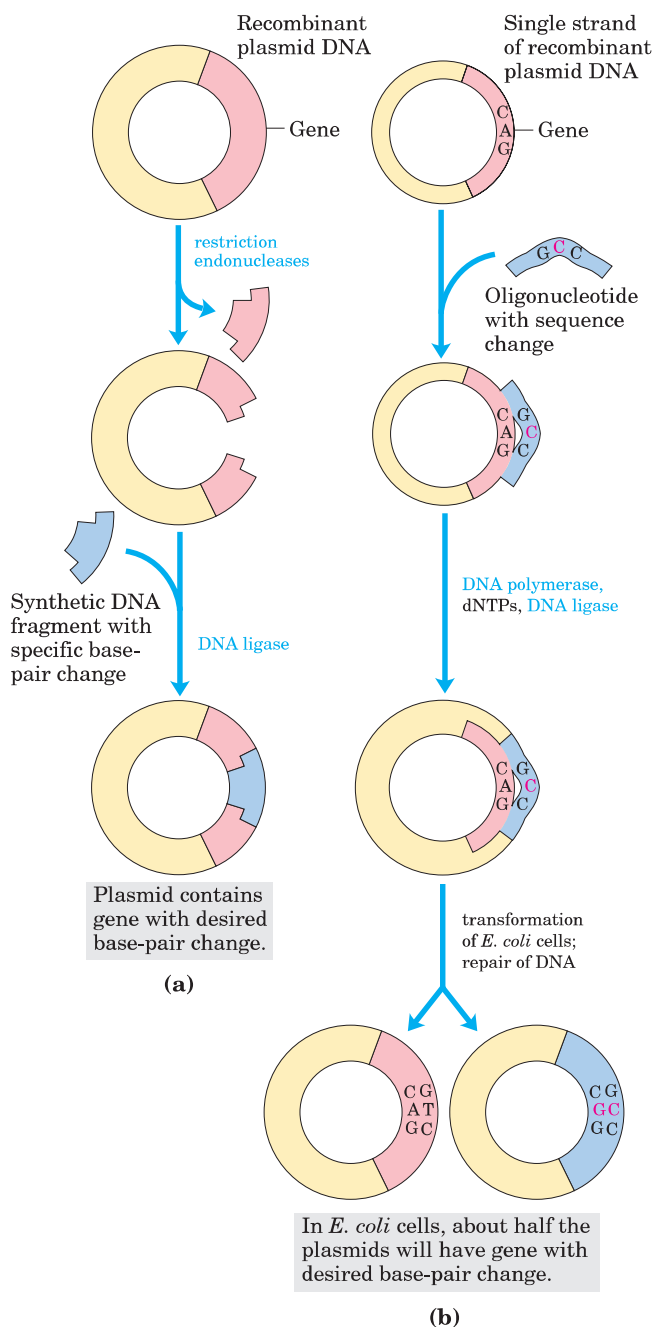


FIGURE 9-12 Two approaches to site-directed mutagenesis. (a) A synthetic DNA segment replaces a DNA fragment that has been removed by cleavage with a restriction endonuclease. (b) A synthetic oligonucleotide with a desired sequence change at one position is hybridized to a single-stranded copy of the gene to be altered. This acts as primer for synthesis of a duplex DNA (with one mismatch), which is then used to transform cells. Cellular DNA repair systems will convert about 50% of the mismatches to reflect the desired sequence change.

replace the *normal* base, retaining the desired mutation. Transformants are screened (often by sequencing their plasmid DNA) until a bacterial colony containing a plasmid with the altered sequence is found.

Changes can also be introduced that involve more than one base pair. Large parts of a gene can be deleted by cutting out a segment with restriction endonucleases and ligating the remaining portions to form a smaller gene. Parts of two different genes can be ligated to create new combinations. The product of such a fused gene is called a **fusion protein**.

Researchers now have ingenious methods to bring about virtually any genetic alteration in vitro. Reintroduction of the altered DNA into the cell permits investigation of the consequences of the alteration. Site-directed mutagenesis has greatly facilitated research on proteins by allowing investigators to make specific changes in the primary structure of a protein and to examine the effects of these changes on the folding, three-dimensional structure, and activity of the protein.

SUMMARY 9.1 DNA Cloning: The Basics

- DNA cloning and genetic engineering involve the cleavage of DNA and assembly of DNA segments in new combinations—recombinant DNA.
- Cloning entails cutting DNA into fragments with enzymes; selecting and possibly modifying a fragment of interest; inserting the DNA fragment into a suitable cloning vector; transferring the vector with the DNA insert into a host cell for replication; and identifying and selecting cells that contain the DNA fragment.
- Key enzymes in gene cloning include restriction endonucleases (especially the type II enzymes) and DNA ligase.
- Cloning vectors include plasmids, bacteriophages, and, for the longest DNA inserts, bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs).
- Cells containing particular DNA sequences can be identified by DNA hybridization methods.
- Genetic engineering techniques manipulate cells to express and/or alter cloned genes.

9.2 From Genes to Genomes

The modern science of **genomics** now permits the study of DNA on a cellular scale, from individual genes to the entire genetic complement of an organism—its genome. Genomic databases are growing rapidly, as one sequencing milestone is superseded by the next. Biology in the twenty-first century will move forward with the aid of informational resources undreamed of only a few years ago. We now turn to a consideration of some of the technologies fueling these advances.

DNA Libraries Provide Specialized Catalogs of Genetic Information

A DNA library is a collection of DNA clones, gathered together as a source of DNA for sequencing, gene discovery, or gene function studies. The library can take a variety of forms, depending on the source of the DNA. Among the largest types of DNA library is a **genomic library**, produced when the complete genome of a particular organism is cleaved into thousands of fragments, and *all* the fragments are cloned by insertion into a cloning vector.

The first step in preparing a genomic library is partial digestion of the DNA by restriction endonucleases, such that any given sequence will appear in fragments of a range of sizes—a range that is compatible with the cloning vector and ensures that virtually all sequences are represented among the clones in the library. Fragments that are too large or too small for cloning are removed by centrifugation or electrophoresis. The cloning vector, such as a BAC or YAC plasmid, is cleaved with the same restriction endonuclease and ligated to the genomic DNA fragments. The ligated DNA mixture is then used to transform bacterial or yeast cells to produce a library of cell types, each type harboring a different recombinant DNA molecule. Ideally, all the DNA in the genome under study will be represented in the library. Each transformed bacterium or yeast cell grows into a colony, or “clone,” of identical cells, each cell bearing the same recombinant plasmid.

Using hybridization methods, researchers can order individual clones in a library by identifying clones with overlapping sequences. A set of overlapping clones represents a catalog for a long contiguous segment of a genome, often referred to as a **contig** (Fig. 9–13). Previously studied sequences or entire genes can be located within the library using hybridization methods to determine which library clones harbor the known sequence. If the sequence has already been mapped on a chromosome, investigators can determine the location (in the genome) of the cloned DNA and any contig of which it is a part. A well-characterized library may contain thousands of long contigs, all assigned to and ordered on particular chromosomes to form a detailed physical map. The known sequences within the library (each called a **sequence-tagged site**, or **STS**) can provide landmarks for genomic sequencing projects.

As more and more genome sequences become available, the utility of genomic libraries is diminishing and investigators are constructing more specialized libraries designed to study gene function. An example is a library that includes only those genes that are *expressed*—that is, are transcribed into RNA—in a given organism or even in certain cells or tissues. Such a library lacks the noncoding DNA that makes up a large portion of many eukaryotic genomes. The researcher first extracts mRNA from an organism or from specific cells of an or-

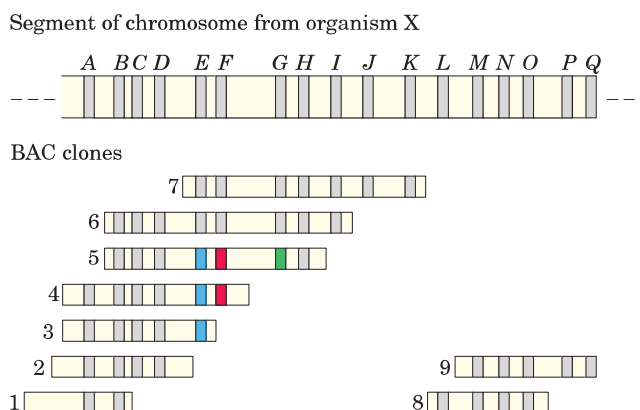


FIGURE 9–13 Ordering of the clones in a DNA library. Shown here is a segment of a chromosome from a hypothetical organism X, with markers A through Q representing sequence-tagged sites (STSs—DNA segments of known sequence, including known genes). Below the chromosome is an array of ordered BAC clones, numbered 1 to 9. Ordering the clones on the genetic map is a many-stage process. The presence or absence of an STS on an individual clone can be determined by hybridization—for example, by probing each clone with PCR-amplified DNA from the STS. Once the STSs on each BAC clone are identified, the clones (and the STSs themselves, if their location is not yet known) can be ordered on the map. For example, compare clones 3, 4, and 5. Marker E (blue) is found on all three clones; F (red) on clones 4 and 5, but not on 3; and G (green) only on clone 5. This indicates that the order of the sites is E, F, G. The clones partially overlap and their order must be 3, 4, 5. The resulting ordered series of clones is called a contig.

ganism and then prepares **complementary DNAs (cDNAs)** from the RNA in a multistep reaction catalyzed by the enzyme reverse transcriptase (Fig. 9–14). The resulting double-stranded DNA fragments are then inserted into a suitable vector and cloned, creating a population of clones called a **cDNA library**. The search for a particular gene is made easier by focusing on a cDNA library generated from the mRNAs of a cell known to express that gene. For example, if we wished to clone globin genes, we could first generate a cDNA library from erythrocyte precursor cells, in which about half the mRNAs code for globins. To aid in the mapping of large genomes, cDNAs in a library can be partially sequenced at random to produce a useful type of STS called an **expressed sequence tag (EST)**. ESTs, ranging in size from a few dozen to several hundred base pairs, can be positioned within the larger genome map, providing markers for expressed genes. Hundreds of thousands of ESTs were included in the detailed physical maps used as a guide to sequencing the human genome.

A cDNA library can be made even more specialized by cloning a cDNA or cDNA fragment into a vector that fuses the cDNA sequence with the sequence for a marker, or reporter gene; the fused genes form a “reporter construct.” Two useful markers are the genes for green fluorescent protein and epitope tags. A target

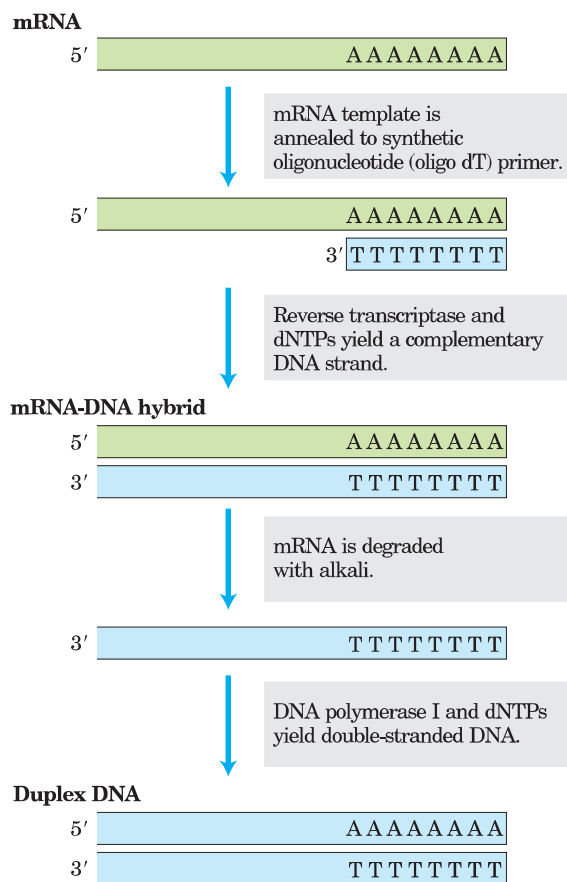


FIGURE 9-14 Construction of a cDNA library from mRNA. A cell's mRNA includes transcripts from thousands of genes, and the cDNAs generated are correspondingly heterogeneous. The duplex DNA produced by this method is inserted into an appropriate cloning vector. Reverse transcriptase can synthesize DNA on an RNA or a DNA template (see Fig. 26–29).

gene fused with a gene for **green fluorescent protein (GFP)** generates a fusion protein that is highly fluorescent—it literally lights up (Fig. 9–15a). Just a few molecules of this protein can be observed microscopically, allowing the study of its location and movements in a cell. An **epitope tag** is a short protein sequence that is bound tightly by a well-characterized monoclonal antibody (Chapter 5). The tagged protein can be specifically precipitated from a crude protein extract by interaction with the antibody (Fig. 9–15b). If any other proteins bind to the tagged protein, those will precipitate as well, providing information about protein-protein interactions in a cell. The diversity and utility of specialized DNA libraries are growing every year.

The Polymerase Chain Reaction Amplifies Specific DNA Sequences

The Human Genome Project, along with the many associated efforts to sequence the genomes of organisms of every type, is providing unprecedented access to gene sequence information. This in turn is simplifying the

process of cloning individual genes for more detailed biochemical analysis. If we know the sequence of at least the flanking parts of a DNA segment to be cloned, we can hugely amplify the number of copies of that DNA segment, using the **polymerase chain reaction (PCR)**, a process conceived by Kary Mullis in 1983. The amplified DNA can be cloned directly or used in a variety of analytical procedures.

The PCR procedure has an elegant simplicity. Two synthetic oligonucleotides are prepared, complementary to sequences on opposite strands of the target DNA at positions just beyond the ends of the segment to be amplified. The oligonucleotides serve as replication primers that can be extended by DNA polymerase. The 3' ends of the hybridized probes are oriented toward each other and positioned to prime DNA synthesis across the desired DNA segment (Fig. 9–16). (DNA polymerases

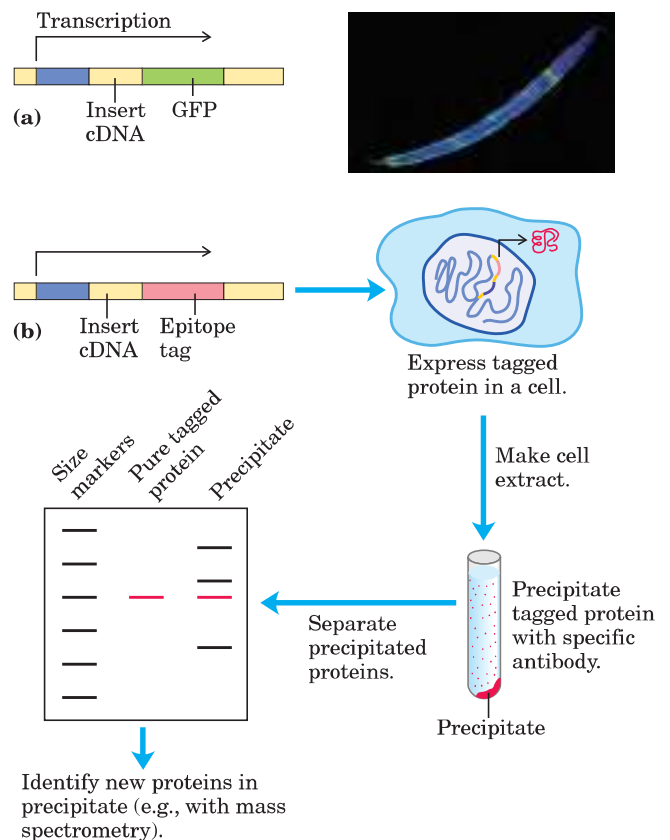


FIGURE 9-15 Specialized DNA libraries. (a) Cloning of cDNA next to a gene for green fluorescent protein (GFP) creates a reporter construct. RNA transcription proceeds through the gene of interest (insert DNA) and the reporter gene, and the mRNA transcript is then expressed as a fusion protein. The GFP part of the protein is visible in the fluorescence microscope. The photograph shows a nematode worm containing a GFP fusion protein expressed only in the four “touch” neurons that run the length of its body. **Reporter Constructs** (b) If the cDNA is cloned next to a gene for an epitope tag, the resulting fusion protein can be precipitated by antibodies to the epitope. Any other proteins that interact with the tagged protein also precipitate, helping to elucidate protein-protein interactions.

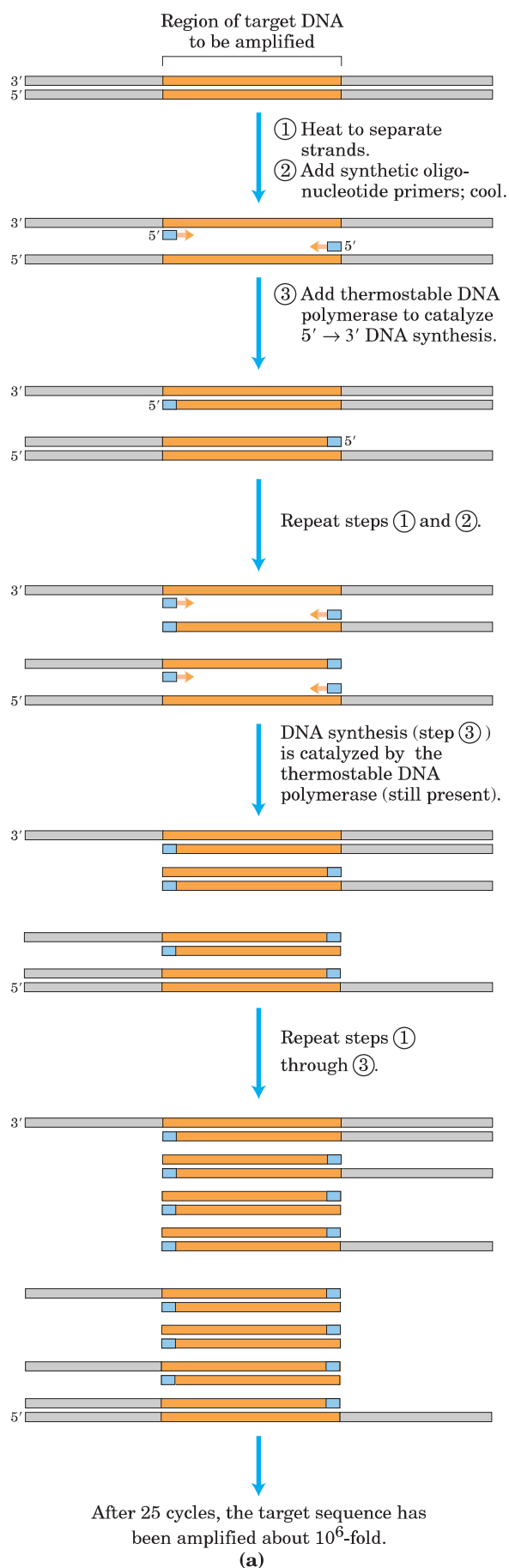
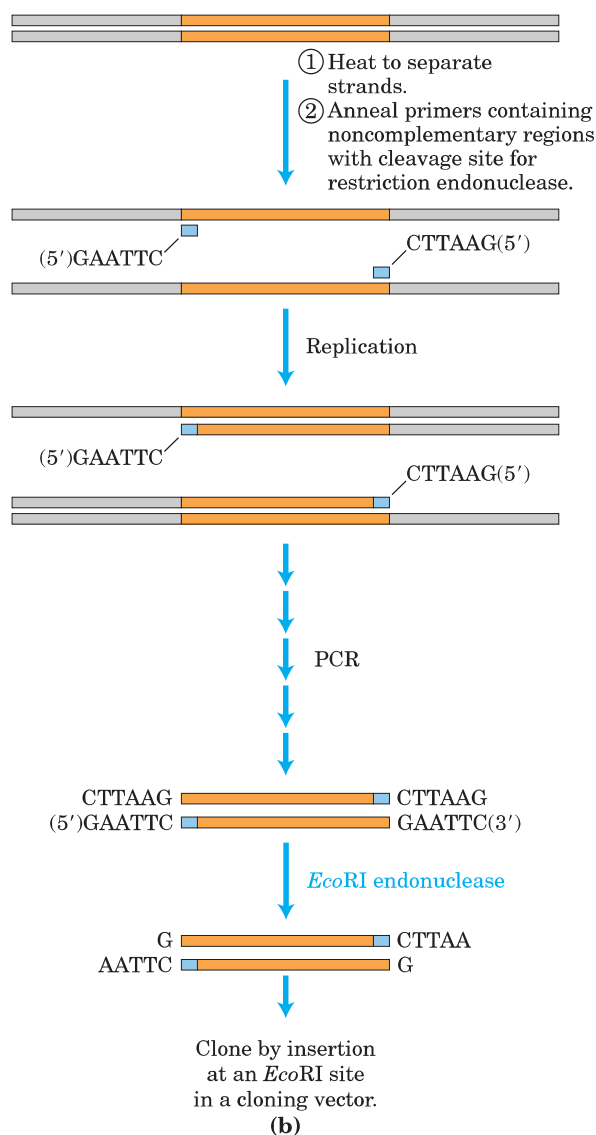


FIGURE 9-16 Amplification of a DNA segment by the polymerase chain reaction. (a) The PCR procedure has three steps. DNA strands are ① separated by heating, then ② annealed to an excess of short synthetic DNA primers (blue) that flank the region to be amplified; ③ new DNA is synthesized by polymerization. The three steps are repeated for 25 or 30 cycles. The thermostable DNA polymerase *TaqI* (from *Thermus aquaticus*, a bacterial species that grows in hot springs) is not denatured by the heating steps. (b) DNA amplified by PCR can be cloned. The primers can include noncomplementary ends that have a site for cleavage by a restriction endonuclease. Although these parts of the primers do not anneal to the target DNA, the PCR process incorporates them into the DNA that is amplified. Cleavage of the amplified fragments at these sites creates sticky ends, used in ligation of the amplified DNA to a cloning vector. Polymerase Chain Reaction



synthesize DNA strands from deoxyribonucleotides, using a DNA template, as described in Chapter 25.) Isolated DNA containing the segment to be amplified is heated briefly to denature it, and then cooled in the presence of a large excess of the synthetic oligonucleotide primers. The four deoxynucleoside triphosphates are then added, and the primed DNA segment is replicated selectively. The cycle of heating, cooling, and replication is repeated 25 or 30 times over a few hours in an automated process, amplifying the DNA segment flanked by the primers until it can be readily analyzed or cloned. PCR uses a heat-stable DNA polymerase, such as the *Taq* polymerase (derived from a bacterium that lives at 90 °C), which remains active after every heating step and does not have to be replenished. Careful design of the primers used for PCR, such as including restriction endonuclease cleavage sites, can facilitate the subsequent cloning of the amplified DNA (Fig. 9-16b).

This technology is highly sensitive: PCR can detect and amplify as little as one DNA molecule in almost any type of sample. Although DNA degrades over time (p. 293), PCR has allowed successful cloning of DNA from samples more than 40,000 years old. Investigators have used the technique to clone DNA fragments from the mummified remains of humans and extinct animals such as the woolly mammoth, creating the new fields of molecular archaeology and molecular paleontology. DNA from burial sites has been amplified by PCR and used to trace ancient human migrations. Epidemiologists can use PCR-enhanced DNA samples from human remains to trace the evolution of human pathogenic viruses. Thus, in addition to its usefulness for cloning DNA, PCR is a potent tool in forensic medicine (Box 9-1). It is also being used for detection of viral infections before they cause symptoms and for prenatal diagnosis of a wide array of genetic diseases.

The PCR method is also important in advancing the goal of whole genome sequencing. For example, the mapping of expressed sequence tags to particular chromosomes often involves amplification of the EST by PCR, followed by hybridization of the amplified DNA to clones in an ordered library. Investigators found many other applications of PCR in the Human Genome Project, to which we now turn.

Genome Sequences Provide the Ultimate Genetic Libraries

The genome is the ultimate source of information about an organism, and there is no genome we are more interested in than our own. Less than 10 years after the development of practical DNA sequencing methods, serious discussions began about the prospects for sequencing the entire 3 billion base pairs of the human genome. The international Human Genome Project got underway with substantial funding in the late 1980s. The effort eventually included significant contributions from

20 sequencing centers distributed among six nations: the United States, Great Britain, Japan, France, China, and Germany. General coordination was provided by the Office of Genome Research at the National Institutes of Health, led first by James Watson and after 1992 by Francis Collins. At the outset, the task of sequencing a 3×10^9 bp genome seemed to be a titanic job, but it gradually yielded to advances in technology. The completed sequence of the human genome was published in April 2003, several years ahead of schedule.

This advance was the product of a carefully planned international effort spanning 14 years. Research teams first generated a detailed physical map of the human genome, with clones derived from each chromosome organized into a series of long contigs (Fig. 9-17). Each contig contained landmarks in the form of STSs at a distance of every 100,000 bp or less. The genome thus mapped could be divided up between the international sequencing centers, each center sequencing the mapped BAC or YAC clones corresponding to its particular segments of the genome. Because many of the

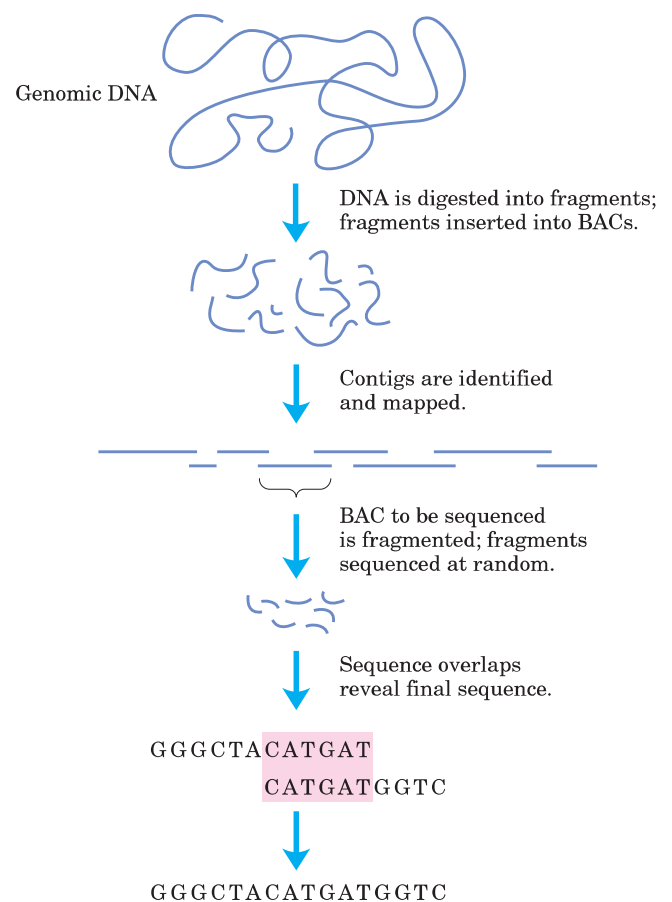


FIGURE 9-17 The Human Genome Project strategy. Clones isolated from a genomic library were ordered into a detailed physical map, then individual clones were sequenced by shotgun sequencing protocols. The strategy used by the commercial sequencing effort eliminated the step of creating the physical map and sequenced the entire genome by shotgun cloning.

BOX 9-1 WORKING IN BIOCHEMISTRY

A Potent Weapon in Forensic Medicine

Traditionally, one of the most accurate methods for placing an individual at the scene of a crime has been a fingerprint. With the advent of recombinant DNA technology, a more powerful tool is now available: **DNA fingerprinting** (also called DNA typing or DNA profiling).

DNA fingerprinting is based on **sequence polymorphisms**, slight sequence differences (usually single base-pair changes) between individuals, 1 bp in every 1,000 bp, on average. Each difference from the prototype human genome sequence (the first one obtained) occurs in some fraction of the human population; every individual has some differences. Some of the sequence changes affect recognition sites for restriction enzymes, resulting in variation in the size of DNA fragments produced by digestion with a particular restriction enzyme. These variations are **restriction fragment length polymorphisms (RFLPs)**.

The detection of RFLPs relies on a specialized hybridization procedure called **Southern blotting** (Fig. 1). DNA fragments from digestion of genomic DNA by restriction endonucleases are separated by size electrophoretically, denatured by soaking the agarose gel in alkali, and then blotted onto a nylon membrane to reproduce the distribution of fragments in the gel. The membrane is immersed in a solution containing a radioactively labeled DNA probe. A probe for a sequence that is repeated several times in the human genome generally identifies a few of the thousands of DNA fragments generated when the human genome is digested with a restriction endonuclease. Autoradiography reveals the fragments to which the probe hybridizes, as in Figure 9-9.

The genomic DNA sequences used in these tests are generally regions containing repetitive DNA

(short sequences repeated thousands of times in tandem), which are common in the genomes of higher eukaryotes (see Fig. 24-8). The number of repeated units in these DNA regions varies among individuals (except between identical twins). With a suitable probe, the pattern of bands produced by DNA fingerprinting is distinctive for each individual. Combining the use of several probes makes the test so selective that it can positively identify a single individual in the entire human population. However, the Southern blot procedure requires relatively fresh DNA samples and larger amounts of DNA than are generally present at a crime scene. RFLP analysis sensitivity is augmented by using PCR (see Fig. 9-16a) to amplify vanishingly small amounts of DNA. This allows investigators to obtain DNA fingerprints from a single hair follicle, a drop of blood, a small semen sample from a rape victim, or samples that might be months or even many years old.

These methods are proving decisive in court cases worldwide. In the example in Figure 1, the DNA from a semen sample obtained from a rape and murder victim was compared with DNA samples from the victim and two suspects. Each sample was cleaved into fragments and separated by gel electrophoresis. Radioactive DNA probes were used to identify a small subset of fragments that contained sequences complementary to the probe. The sizes of the identified fragments varied from one individual to the next, as seen here in the different patterns for the three individuals (victim and two suspects) tested. One suspect's DNA exhibits a banding pattern identical to that of a semen sample taken from the victim. This test used a single probe, but three or four different probes would be used (in separate experiments) to achieve an unambiguous positive identification.

clones were more than 100,000 bp long, and modern sequencing techniques can resolve only 600 to 750 bp of sequence at a time, each clone had to be sequenced in pieces. The sequencing strategy used a shotgun approach, in which researchers used powerful new automated sequencers to sequence random segments of a given clone, then assembled the sequence of the entire clone by computerized identification of overlaps. The number of clone pieces sequenced was determined statistically so that the entire length of the clone was sequenced four to six times on average. The sequenced DNA was then made available in a database covering the entire genome. Construction of the physical map was a

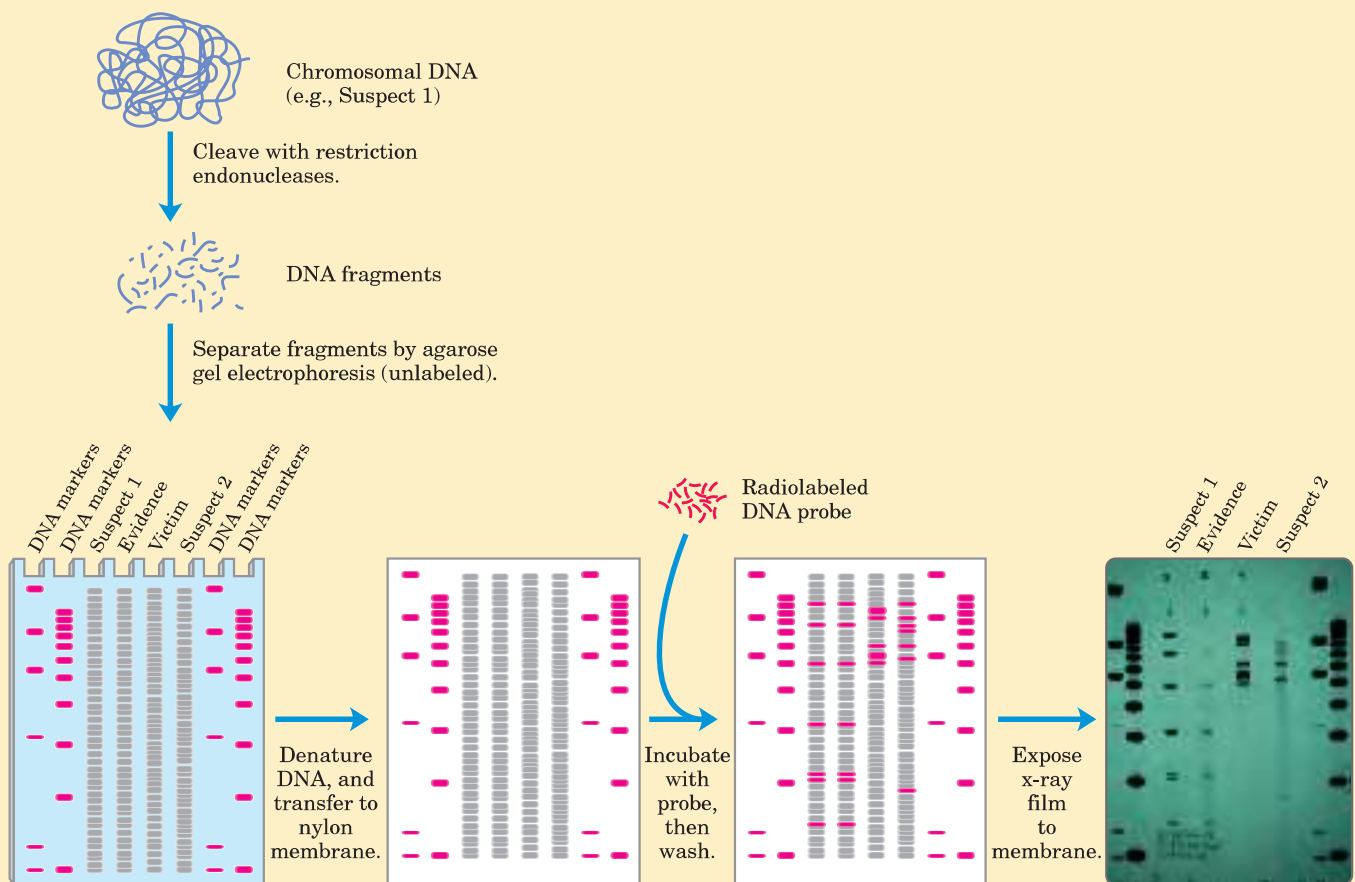
time-consuming task, and its progress was followed in annual reports in major journals throughout the 1990s—by the end of which the map was largely in place. Completion of the entire sequencing project was initially projected for the year 2005, but circumstances and technology intervened to accelerate the process.

A competing commercial effort to sequence the human genome was initiated by the newly established Celera Corporation in 1997. Led by J. Craig Venter, the Celera group made use of a different strategy called “whole genome shotgun sequencing,” which eliminates the step of assembling a physical map of the genome. Instead, teams sequenced DNA segments from through-

Such results have been used to both convict and acquit suspects and, in other cases, to establish paternity with an extraordinary degree of certainty. The impact of these procedures on court cases will continue to grow as societies agree on the standards and as formal methods become widely established in forensic laboratories. Even decades-old murder mysteries

can be solved: in 1996, DNA fingerprinting helped to confirm the identification of the bones of the last Russian czar and his family, who were assassinated in 1918.

FIGURE 1 The Southern blot procedure, as applied to DNA fingerprinting. This procedure was named after Jeremy Southern, who developed the technique.



Francis S. Collins



J. Craig Venter

out the genome at random. The sequenced segments were ordered by the computerized identification of sequence overlaps (with some reference to the public project's detailed physical map). At the outset of the Human Genome Project, shotgun sequencing on this scale had been deemed impractical. However, advances in computer software and sequencing automation had made the approach feasible by 1997. The ensuing race between the private and public sequencing efforts substantially advanced the timeline for completion of the project. Publication of the draft human genome sequence in 2001 was followed by two years of follow-up work to eliminate nearly a thousand discontinuities and

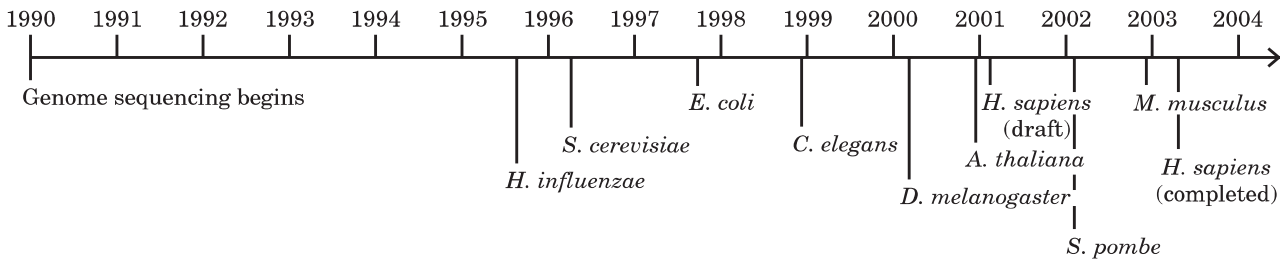


FIGURE 9-18 Genomic sequencing timeline. Discussions in the mid-1980s led to initiation of the project in 1989. Preparatory work, including extensive mapping to provide genome landmarks, occupied much of the 1990s. Separate projects were launched to sequence the genomes of other organisms important to research. The first sequencing efforts to be completed included many bacterial species (such as

Haemophilus influenzae), yeast (*S. cerevisiae*), a nematode worm (*C. elegans*), the fruit fly (*D. melanogaster*), and a plant (*A. thaliana*). Completed sequences for mammalian genomes, including the human genome, began to emerge in 2000. Each genome project has a website that serves as a central repository for the latest data.

to provide high-quality sequence data that are contiguous throughout the genome.

The Human Genome Project marks the culmination of twentieth-century biology and promises a vastly changed scientific landscape for the new century. The human genome is only part of the story, as the genomes of many other species are also being (or have been) sequenced, including the yeasts *Saccharomyces cerevisiae* (completed in 1996) and *Schizosaccharomyces pombe* (2002), the nematode *Caenorhabditis elegans* (1998), the fruit fly *Drosophila melanogaster* (2000), the plant *Arabidopsis thaliana* (2000), the mouse *Mus musculus* (2002), zebrafish, and dozens of bacterial and archaeobacterial species (Fig. 9-18). Most of the early efforts have been focused on species commonly used in laboratories. However, genome sequencing is destined to branch out to many other species as experience grows and technology improves. Broad efforts to map genes, attempts to identify new proteins and disease genes, and many other initiatives are currently under way.

The result is a database with the potential not only to fuel rapid advances in biology but to change the way that humans think about themselves. Early insights provided by the human genome sequence range from the intriguing to the profound. We are not as complicated as we thought. Decades-old estimates that humans possessed about 100,000 genes within the approximately 3.2×10^9 bp in the human genome have been supplanted by the discovery that we have only 30,000 to 35,000 genes. This is perhaps three times more genes than a fruit fly (with 13,000) and twice as many as a nematode worm (18,000). Although humans evolved relatively recently, the human genome is very old. Of 1,278 protein families identified in one early screen, only 94 were unique to vertebrates. However, while we share many protein domain types with plants, worms, and flies, we use these domains in more complex arrangements. Alternative modes of gene expression (Chapter 26) allow the production of more than one protein from a single gene—a process that humans and other vertebrates engage in more than do bacteria, worms, or any

other forms of life. This allows for greater complexity in the proteins generated from our gene complement.

We now know that only 1.1% to 1.4% of our DNA actually encodes proteins (Fig. 9-19). More than 50% of our genome consists of short, repeated sequences, the vast majority of which—about 45% of our genome in all—come from transposons, short movable DNA sequences that are molecular parasites (Chapter 25). Many of the transposons have been there a long time, now altered so that they can no longer move to new genomic locations. Others are still actively moving at low frequencies, helping to make the genome an ever-dynamic and evolving entity. At least a few transposons have been co-opted by their host and appear to serve useful cellular functions.

What does all this information tell us about how much one human differs from another? Within the human population are millions of single-base differences, called **single nucleotide polymorphisms**, or **SNPs** (pronounced “snips”). Each human differs from the next

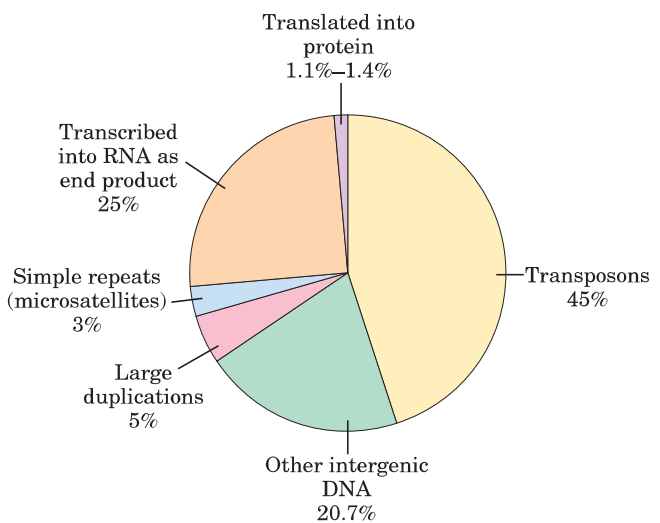


FIGURE 9-19 Snapshot of the human genome. The chart shows the proportions of our genome made up of various types of sequences.

by about 1 bp in every 1,000 bp. From these small genetic differences arises the human variety we are all aware of—differences in hair color, eyesight, allergies to medication, foot size, and even (to some unknown degree) behavior. Some of the SNPs are linked to particular human populations and can provide important information about human migrations that occurred thousands of years ago and about our more distant evolutionary past.

As spectacular as this advance is, the sequencing of the human genome is easy compared with what comes next—the effort to understand all the information in each genome. The genome sequences being added monthly to international databases are roadmaps, parts of which are written in a language we do not yet understand. However, they have great utility in catalyzing the discovery of new proteins and processes affecting every aspect of biochemistry, as will become apparent in chapters to come.

SUMMARY 9.2 From Genes to Genomes

- The science of genomics broadly encompasses the study of genomes and their gene content.
- Genomic DNA segments can be organized in libraries—such as genomic libraries and cDNA libraries—with a wide range of designs and purposes.
- The polymerase chain reaction (PCR) can be used to amplify selected DNA segments from a DNA library or an entire genome.
- In an international cooperative research effort, the genomes of many organisms, including that of humans, have been sequenced in their entirety and are now available in public databases.

9.3 From Genomes to Proteomes

A gene is not simply a DNA sequence; it is information that is converted to a useful product—a protein or functional RNA molecule—when and if needed by the cell. The first and most obvious step in exploring a large sequenced genome is to catalog the products of the genes within that genome. Genes that encode RNA as their final product are somewhat harder to identify than are protein-encoding genes, and even the latter can be very difficult to spot in a vertebrate genome. The explosion of DNA sequence information has also revealed a sobering truth. Despite many years of biochemical advances, there are still thousands of proteins in every eukaryotic cell (and quite a few in bacteria) that we know nothing about. These proteins may have functions in processes not yet discovered, or may contribute in unexpected

ways to processes we think we understand. In addition, the genomic sequences tell us nothing about the three-dimensional structure of proteins or how proteins are modified after they are synthesized. The proteins, with their myriad critical functions in every cell, are now becoming the focus of new strategies for whole cell biochemistry.

The complement of proteins expressed by a genome is called its **proteome**, a term that first appeared in the research literature in 1995. This concept rapidly evolved into a separate field of investigation, called **proteomics**. The problem addressed by proteomics research is straightforward, although the solution is not. Each genome presents us with thousands of genes encoding proteins, and ideally we want to know the structure and function of all those proteins. Given that many proteins offer surprises even after years of study, the investigation of an entire proteome is a daunting enterprise. Simply discovering the function of new proteins requires intensive work. Biochemists can now apply shortcuts in the form of a broad array of new and updated technologies.

Protein function can be described on three levels. **Phenotypic function** describes the effects of a protein on the entire organism. For example, the loss of the protein may lead to slower growth of the organism, an altered development pattern, or even death. **Cellular function** is a description of the network of interactions engaged in by a protein at the cellular level. Interactions with other proteins in the cell can help define the kinds of metabolic processes in which the protein participates. Finally, **molecular function** refers to the precise biochemical activity of a protein, including details such as the reactions an enzyme catalyzes or the ligands a receptor binds.

For several genomes, such as those of the yeast *Saccharomyces cerevisiae* and the plant *Arabidopsis*, a massive effort is underway to inactivate each gene by genetic engineering and to investigate the effect on the organism. If the growth patterns or other properties of the organism change (or if it does not grow at all), this provides information on the phenotypic function of the protein product of the gene.

There are three other main paths to investigating protein function: (1) sequence and structural comparisons with genes and proteins of known function, (2) determination of when and where a gene is expressed, and (3) investigation of the interactions of the protein with other proteins. We discuss each of these approaches in turn.

Sequence or Structural Relationships Provide Information on Protein Function

One of the important reasons to sequence many genomes is to provide a database that can be used to assign gene functions by genome comparisons, an enterprise referred

to as **comparative genomics**. Sometimes a newly discovered gene is related by sequence homologies to a gene previously studied in another or the same species, and its function can be entirely or partly defined by that relationship. Such genes—of different species but possessing a clear sequence and functional relationship to each other—are called **orthologs**. Genes similarly related to each other within a single species are called **paralogs** (see Fig. 1–37). If the function of a gene has been characterized for one species, this information can be used to assign gene function to the ortholog found in the second species. The identity is easiest to make when comparing genomes from relatively closely related species, such as mouse and human, although many clearly orthologous genes have been identified in species as distant as bacteria and humans. Sometimes even the order of genes on a chromosome is conserved over large segments of the genomes of closely related species (Fig. 9–20). Conserved gene order, called **synteny**, provides additional evidence for an orthologous relationship between genes at identical locations within the related segments.

Alternatively, certain sequences associated with particular structural motifs (Chapter 4) may be identified within a protein. The presence of a structural motif may suggest that it, say, catalyzes ATP hydrolysis, binds to DNA, or forms a complex with zinc ions, helping to define molecular function. These relationships are determined with the aid of increasingly sophisticated computer programs, limited only by the current information on gene and protein structure and our capacity to associate sequences with particular structural motifs.

Human 9	Mouse 2
<i>EPB72</i>	<i>Epb7.2</i>
<i>PSMB7</i>	<i>Psmb7</i>
<i>DNM1</i>	<i>Dnm</i>
<i>LMX1B</i>	<i>Lmx1b</i>
<i>CDK9</i>	<i>Cdk9</i>
<i>STXBP1</i>	<i>Stxbp1</i>
<i>AK1</i>	<i>Ak1</i>
<i>LCN2</i>	<i>Lcn2</i>

FIGURE 9–20 Synteny in the mouse and human genomes. Large segments of the mouse and human genomes have closely related genes aligned in the same order on chromosomes, a relationship called synteny. This diagram shows segments of human chromosome 9 and mouse chromosome 2. The genes in these segments exhibit a very high degree of homology as well as the same gene order. The different lettering schemes for the gene names reflect different naming conventions in the two organisms.

To further the assignment of function based on structural relationships, a large-scale structural proteomics project has been initiated. The goal is to crystallize and determine the structure of as many proteins and protein domains as possible, in many cases with little or no existing information about protein function. The project has been assisted by the automation of some of the tedious steps of protein crystallization (see Box 4–4). As these structures are revealed, they will be made available in the structural databases described in Chapter 4. The effort should help define the extent of variation in structural motifs. When a newly discovered protein is found to have structural folds that are clearly related to motifs with known functions in the databases, this information can suggest a molecular function for the protein.

Cellular Expression Patterns Can Reveal the Cellular Function of a Gene

In every newly sequenced genome, researchers find genes that encode proteins with no evident structural relationships to known genes or proteins. In these cases, other approaches must be used to generate information about gene function. Determining which tissues a gene is expressed in, or what circumstances trigger the appearance of the gene product, can provide valuable clues. Many different approaches have been developed to study these patterns.

Two-Dimensional Gel Electrophoresis As shown in Figure 3–22, two-dimensional gel electrophoresis allows the separation and display of up to 1,000 different proteins on a single gel. Mass spectrometry (see Box 3–2) can then be used to partially sequence individual protein spots and assign each to a gene. The appearance and nonappearance (or disappearance) of particular protein spots in samples from different tissues, from similar tissues at different stages of development, or from tissues treated in ways that simulate a variety of biological conditions can help define cellular function.

DNA Microarrays Major refinements of the technology underlying DNA libraries, PCR, and hybridization have come together in the development of **DNA microarrays** (sometimes called **DNA chips**), which allow the rapid and simultaneous screening of many thousands of genes. DNA segments from known genes, a few dozen to hundreds of nucleotides long, are amplified by PCR and placed on a solid surface, using robotic devices that accurately deposit nanoliter quantities of DNA solution. Many thousands of such spots are deposited in a pre-designed array on a surface area of just a few square centimeters. An alternative strategy is to synthesize DNA directly on the solid surface, using photolithography (Fig. 9–21). Once the chip is constructed, it can be probed with mRNAs or cDNAs from a particular cell type

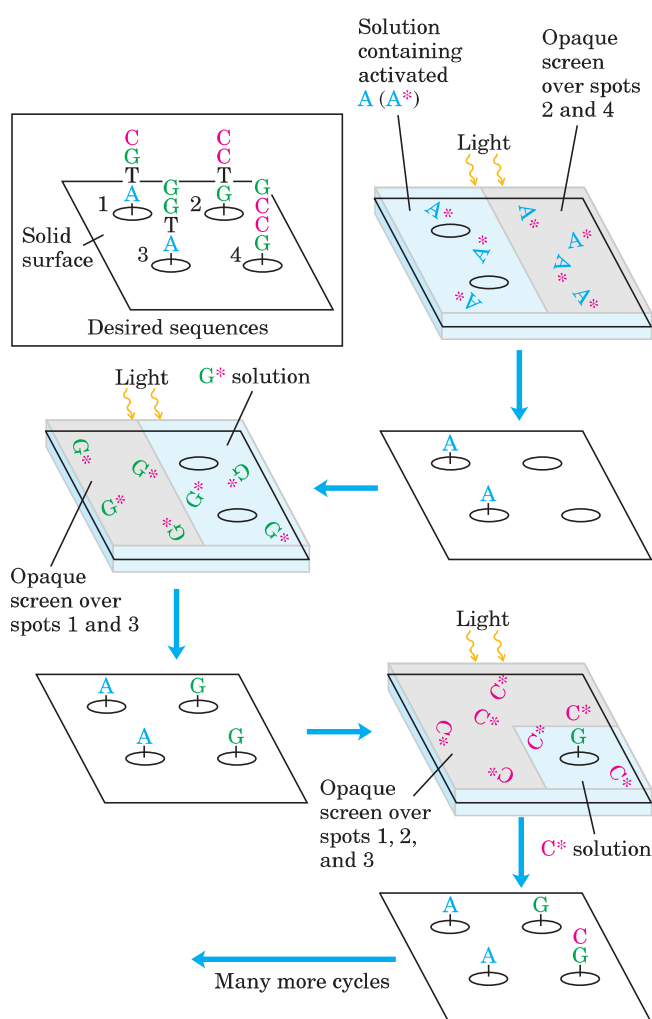


FIGURE 9-21 Photolithography. This technique for preparing a DNA microarray makes use of nucleotide precursors that are activated by light, joining one nucleotide to the next in a photoreaction (as opposed to the chemical process illustrated in Fig. 8-38). A computer is programmed with the oligonucleotide sequences to be synthesized at each point on a solid surface. The surface is washed successively with solutions containing one type of activated nucleotide (A^* , G^* , etc.). As in the chemical synthesis of DNA, the activated nucleotides are blocked so that only one can be added to a chain in each cycle. A screen covering the surface is opened over the areas programmed to receive a particular nucleotide, and a flash of light joins the nucleotide to the polymers in the uncovered areas. This continues until the required sequences are built up on each spot on the surface. Many polymers with the same sequence are generated on each spot, not just the single polymer shown. Also, the surfaces have thousands of spots with different sequences (see Fig. 9-22); this array shows just four spots, to illustrate the strategy.

or cell culture to identify the genes being expressed in those cells.

A microarray can answer such questions as which genes are expressed at a given stage in the development of an organism. The total complement of mRNA is isolated from cells at two different stages of development

and converted to cDNA, using reverse transcriptase and fluorescently labeled deoxynucleotides. The fluorescent cDNAs are then mixed and used as probes, each hybridizing to complementary sequences on the microarray. In Figure 9-22, for example, the labeled nucleotides used to make the cDNA for each sample fluoresce in two different colors. The cDNA from the two samples is mixed and used to probe the microarray. Spots that fluoresce green represent mRNAs more abundant at the single-cell stage; those that fluoresce red represent sequences more abundant later in development. The mRNAs that are equally abundant at both stages of development fluoresce yellow. By using a mixture of two samples to measure relative rather than absolute abundance of sequences, the method corrects for variations in the amounts of DNA originally deposited in each spot on the grid and other possible inconsistencies among spots in the microarray. The spots that fluoresce provide a snapshot of all the genes being expressed in the cells at the moment they were harvested—gene expression examined on a genome-wide scale. For a gene of unknown function, the time and circumstances of its expression can provide important clues about its role in the cell.

An example of this technique is illustrated in Figure 9-23, showing the dramatic results this technique can produce. Segments from each of the more than 6,000 genes in the completely sequenced yeast genome were separately amplified by PCR, and each segment was deposited in a defined pattern to create the illustrated microarray. In a sense, this array provides a snapshot of the entire yeast genome.

Protein Chips Proteins, too, can be immobilized on a solid surface and used to help define the presence or absence of other proteins in a sample. For example, researchers prepare an array of antibodies to particular proteins by immobilizing them as individual spots on a solid surface. A sample of proteins is added, and if the protein that binds any of the antibodies is present in the sample, it can be detected by a solid-state form of the ELISA assay (see Fig. 5-28). Many other types and applications of protein chips are being developed.

Detection of Protein-Protein Interactions Helps to Define Cellular and Molecular Function

A key to defining the function of any protein is to determine what it binds to. In the case of protein-protein interactions, the association of a protein of unknown function with one whose function is known can provide a useful and compelling “guilt by association.” The techniques used in this effort are quite varied.

Comparisons of Genome Composition Although not evidence of direct association, the mere presence of combinations of genes in particular genomes can hint at

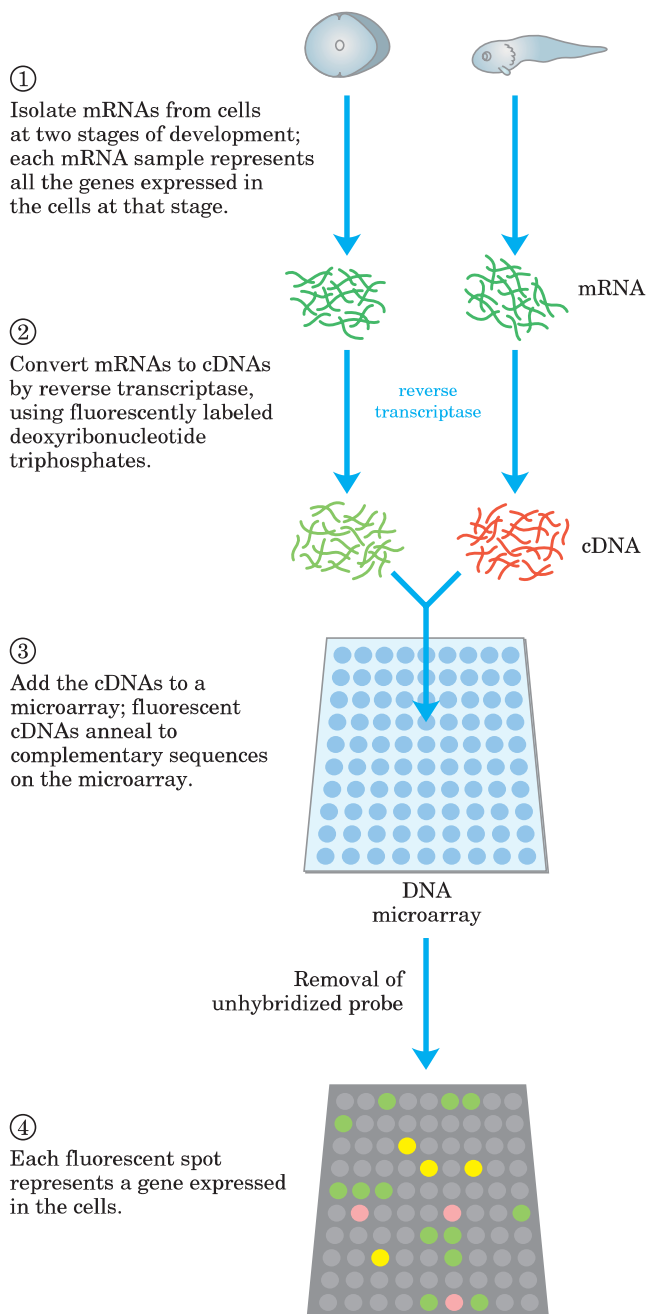



FIGURE 9-22 DNA microarray. A microarray can be prepared from any known DNA sequence, from any source, generated by chemical synthesis or by PCR. The DNA is positioned on a solid surface (usually specially treated glass slides) with the aid of a robotic device capable of depositing very small (nanoliter) drops in precise patterns. UV light cross-links the DNA to the glass slides. Once the DNA is attached to the surface, the microarray can be probed with other fluorescently labeled nucleic acids. Here, mRNA samples are collected from cells at two different stages in the development of a frog. The cDNA probes are made with nucleotides that fluoresce in different colors for each sample; a mixture of the cDNAs is used to probe the microarray. Green spots represent mRNAs more abundant at the single-cell stage; red spots, sequences more abundant later in development. The yellow spots indicate approximately equal abundance at both stages.  **Synthesizing an Oligonucleotide Array**

protein function. We can simply search the genomic databases for particular genes, then determine what other genes are present in the same genomes (Fig. 9–24). When two genes always appear together in a genome, it suggests that the proteins they encode may be functionally related. Such correlations are most useful if the function of at least one of the proteins is known.

Purification of Protein Complexes With the construction of cDNA libraries in which each gene is contiguous with (fused to) an epitope tag, workers can immunoprecipitate the protein product of a gene by using the antibody that binds to the epitope (Fig. 9–15b). If the tagged protein is expressed in cells, other proteins that bind to it may also be precipitated with it. Identification of the associated proteins reveals some of the protein-protein interactions of the tagged protein. There are many variations of this process. For example, a crude extract of cells that express a similarly tagged protein is added to a column containing immobilized antibody. The tagged protein binds to the antibody, and proteins that interact with the tagged protein are sometimes also retained

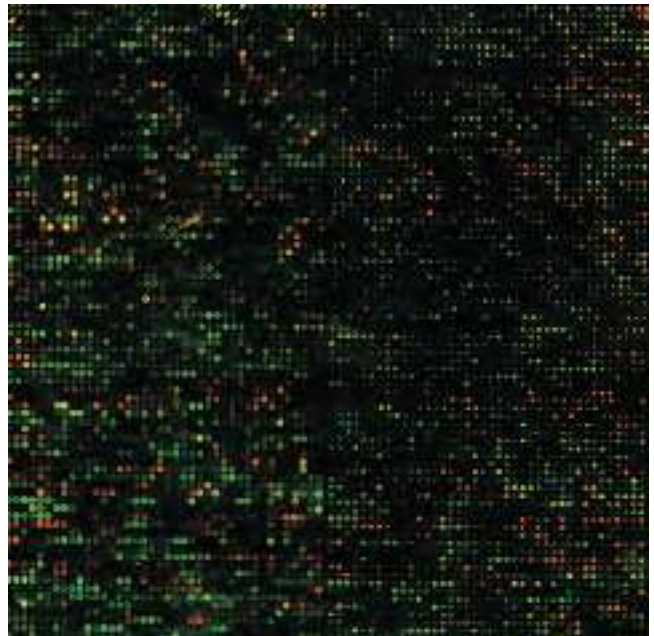



FIGURE 9-23 Enlarged image of a DNA microarray. Each glowing spot in this microarray contains DNA from one of the 6,200 genes of the yeast (*S. cerevisiae*) genome, with every gene represented in the array. The microarray has been probed with fluorescently labeled nucleic acid derived from the mRNAs obtained (1) when the cells were growing normally in culture and (2) five hours after the cells began to form spores. The green spots represent genes expressed at higher levels during normal growth; the red spots, genes expressed at higher levels during sporulation. The yellow spots represent genes that do not change their levels of expression during sporulation. This image is enlarged; the microarray actually measures only 1.8×1.8 cm.  **Screening Oligonucleotide Array for Patterns of Gene Expression**

Protein	Species			
	1	2	3	4
P1	+	—	+	+
P2	—	—	+	—
P3	+	+	—	+
P4	+	—	+	—
P5	+	—	—	—
P6	+	+	—	+
P7	+	+	+	—

FIGURE 9-24 Use of comparative genomics to identify functionally related genes. One use of comparative genomics is to prepare phylogenetic profiles in order to identify genes that always appear together in a genome. This example shows a comparison of genes from four organisms, but in practice, computer searches can look at dozens of species. The designations P1, P2, and so forth refer to proteins encoded by each species. This technique does not require homologous proteins. In this example, because proteins P3 and P6 always appear together in a genome they may be functionally related. Further testing would be needed to confirm this inference.

on the column. The connection between the protein and the tag is cleaved with a specific protease, and the protein complexes are eluted from the column and analyzed. Researchers can use these methods to define complex networks of interactions within a cell.

A variety of useful protein tags are available. A common one is a histidine tag, often just a string of six His residues. A poly-His sequence binds quite tightly to metals such as nickel. If a protein is cloned so that its sequence is contiguous with a His tag, it will have the extra His residues at its carboxyl terminus. The protein can then be purified by chromatography on columns with immobilized nickel. These procedures are convenient but require caution, because the additional amino acid residues in an epitope or His tag can affect protein activity.

Yeast Two-Hybrid Analysis A sophisticated genetic approach to defining protein-protein interactions is based on the properties of the Gal4 protein (Gal4p), which activates transcription of certain genes in yeast (see Fig. 28–28). Gal4p has two domains, one that binds to a specific DNA sequence and another that activates the RNA polymerase that synthesizes mRNA from an adjacent reporter gene. The domains are stable when separated, but activation of the RNA polymerase requires interaction with the activation domain, which in turn requires positioning by the DNA-binding domain. Hence, the domains must be brought together to function correctly (Fig. 9–25a).

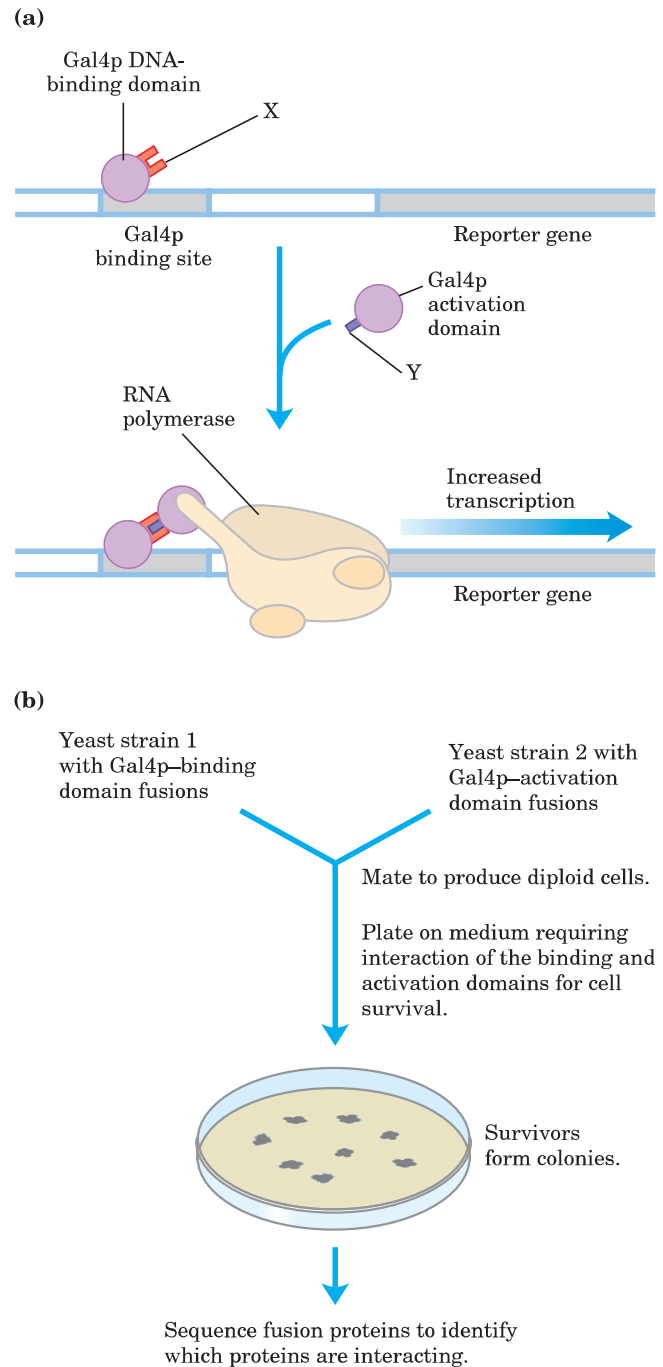



FIGURE 9-25 The yeast two-hybrid system. (a) In this system for detecting protein-protein interactions, the aim is to bring together the DNA-binding domain and the activation domain of the yeast Gal4 protein through the interaction of two proteins, X and Y, to which each domain is fused. This interaction is accompanied by the expression of a reporter gene. (b) The two fusions are created in separate yeast strains, which are then mated. The mated mixture is plated on a medium on which the yeast cannot survive unless the reporter gene is expressed. Thus, all surviving colonies have interacting protein fusion pairs. Sequencing of the fusion proteins in the survivors reveals which proteins are interacting.  **Yeast Two-Hybrid Systems**

In this method, the protein-coding regions of genes to be analyzed are fused to the coding sequences of either the DNA-binding domain or the activation domain of Gal4p, and the resulting genes express a series of fusion proteins. If a protein fused to the DNA-binding domain interacts with a protein fused to the activation domain, transcription is activated. The reporter gene transcribed by this activation is generally one that yields a protein required for growth, or is an enzyme that catalyzes a reaction with a colored product. Thus, when grown on the proper medium, cells that contain such a pair of interacting proteins are easily distinguished from those that do not. Typically, many genes are fused to the Gal4p DNA-binding domain gene in one yeast strain, and many other genes are fused to the Gal4p activation domain gene in another yeast strain, then the yeast strains are mated and individual diploid cells grown into colonies (Fig. 9–25b). This allows for large-scale screening for proteins that interact in the cell.

All these techniques provide important clues to protein function. However, they do not replace classical biochemistry. They simply provide researchers with an expedited entrée into important new biological problems. In the end, a detailed functional understanding of any new protein requires traditional biochemical analyses—such as were used for the many well-studied proteins described in this text. When paired with the simultaneously evolving tools of biochemistry and molecular biology, genomics and proteomics are speeding the discovery not only of new proteins but of new biological processes and mechanisms.

SUMMARY 9.3 From Genomes to Proteomes

- A proteome is the complement of proteins produced by a cell's genome. The new field of proteomics encompasses an effort to catalog and determine the functions of all the proteins in a cell.
- One of the most effective ways to determine the function of a new gene is by comparative genomics, the search of databases for genes with similar sequences. Paralogs and orthologs are proteins (and their genes) with clear functional and sequence relationships in the same or in different species. In some cases, the presence of a gene in combination with certain other genes, observed as a pattern in several genomes, can point toward a possible function.
- Cellular proteomes can be displayed by two-dimensional gel electrophoresis and explored with the aid of mass spectrometry.
- The cellular function of a protein can sometimes be inferred by determining when

and where its gene is expressed. Researchers use DNA microarrays (chips) and protein chips to explore gene expression at the cellular level.

- Several new techniques, including comparative genomics, immunoprecipitation, and yeast two-hybrid analysis, can identify protein-protein interactions. These interactions provide important clues to protein function.

9.4 Genome Alterations and New Products of Biotechnology

We don't need to look far to find practical applications for the new biotechnologies or to find new opportunities for breakthroughs in basic research. Herein lie both the promise and the challenge of genomics. As our knowledge of the genome increases, we will improve our understanding of every aspect of biological function. We will enhance our capacity to engineer organisms and produce new pharmaceutical agents and, as a consequence, will improve human nutrition and health. This promise can be realized only if practical safeguards are in place to ensure responsible application of these techniques.

A Bacterial Plant Parasite Aids Cloning in Plants

We not only can understand genomes, we can change them. This is perhaps the ultimate manifestation of the new technologies. The introduction of recombinant DNA into plants has enormous implications for agriculture, making possible the alteration of the nutritional profile or yield of crops or their resistance to environmental stresses, such as insect pests, diseases, cold, salinity, and drought. Fertile plants of some species may be generated from a single transformed cell, so that an introduced gene passes to progeny through the seeds.

As yet, researchers have not found any naturally occurring plant cell plasmids to facilitate cloning in plants, so the biggest technical challenge is getting DNA into plant cells. An important and adaptable ally in this effort is the soil bacterium *Agrobacterium tumefaciens*. This bacterium can invade plants at the site of a wound, transform nearby cells, and induce them to form a tumor called a crown gall. *Agrobacterium* contains the large (200,000 bp) **Ti plasmid** (Fig. 9–26a). When the bacterium is in contact with a damaged plant cell, a 23,000 bp segment of the Ti plasmid called T DNA is transferred from the plasmid and integrated at a random position in one of the plant cell chromosomes (Fig. 9–26b). The transfer of T DNA from *Agrobacterium* to the plant cell chromosome depends on two 25 bp repeats that flank the T DNA and on the products of the virulence (*vir*) genes on the Ti plasmid (Fig. 9–26a).

The T DNA encodes enzymes that convert plant metabolites to two classes of compounds that benefit

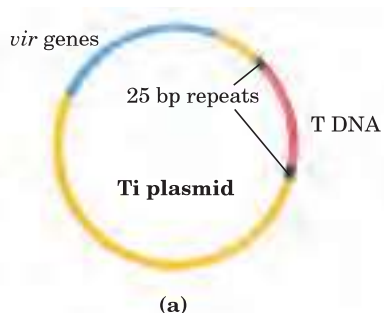
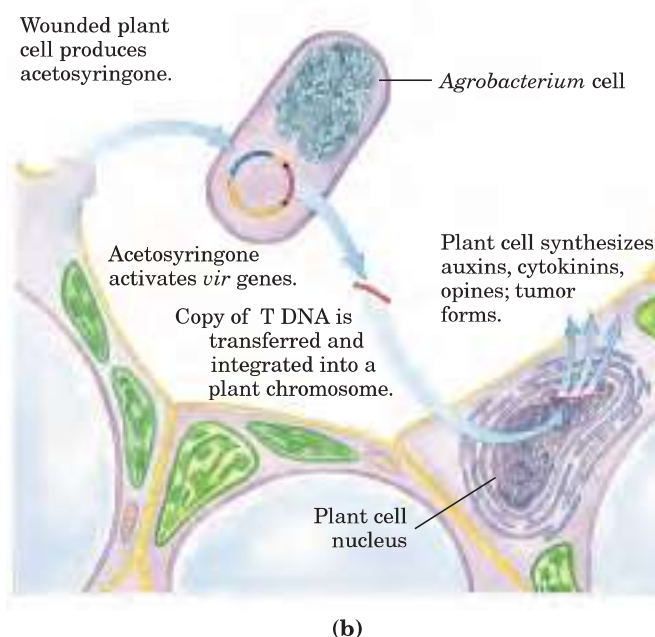


FIGURE 9-26 Transfer of DNA to plant cells by a bacterial parasite.

(a) The Ti (tumor-inducing) plasmid of *Agrobacterium tumefaciens*. (b) Wounded plant cells produce and release the phenolic compound acetosyringone. When *Agrobacterium* detects this compound, the virulence (*vir*) genes on the Ti plasmid are expressed. The *vir* genes encode enzymes needed to introduce the T DNA segment of the Ti plasmid into the genome of nearby plant cells. A single-stranded copy of the T DNA is synthesized and transferred to the plant cell, where it is converted to duplex DNA and integrated into a plant cell chromosome. The T DNA encodes enzymes that synthesize both plant growth hormones and opines (see Fig. 9-27); the latter compounds are metabolized (as a nutrient source) only by *Agrobacterium*. Expression of the T DNA genes by transformed plant cells thus leads to both aberrant plant cell growth (tumor formation) and the diversion of plant cell nutrients to the invading bacteria.



the bacterium (Fig. 9-27). The first group consists of plant growth hormones (auxins and cytokinins) that stimulate growth of the transformed plant cells to form the crown gall tumor. The second constitutes a series of unusual amino acids called opines, which serve as a food source for the bacterium. The opines are produced

in high concentrations in the tumor cells and secreted to the surroundings, where they can be metabolized only by *Agrobacterium*, using enzymes encoded elsewhere on the Ti plasmid. The bacterium thereby diverts plant resources by converting them to a form that benefits only itself.

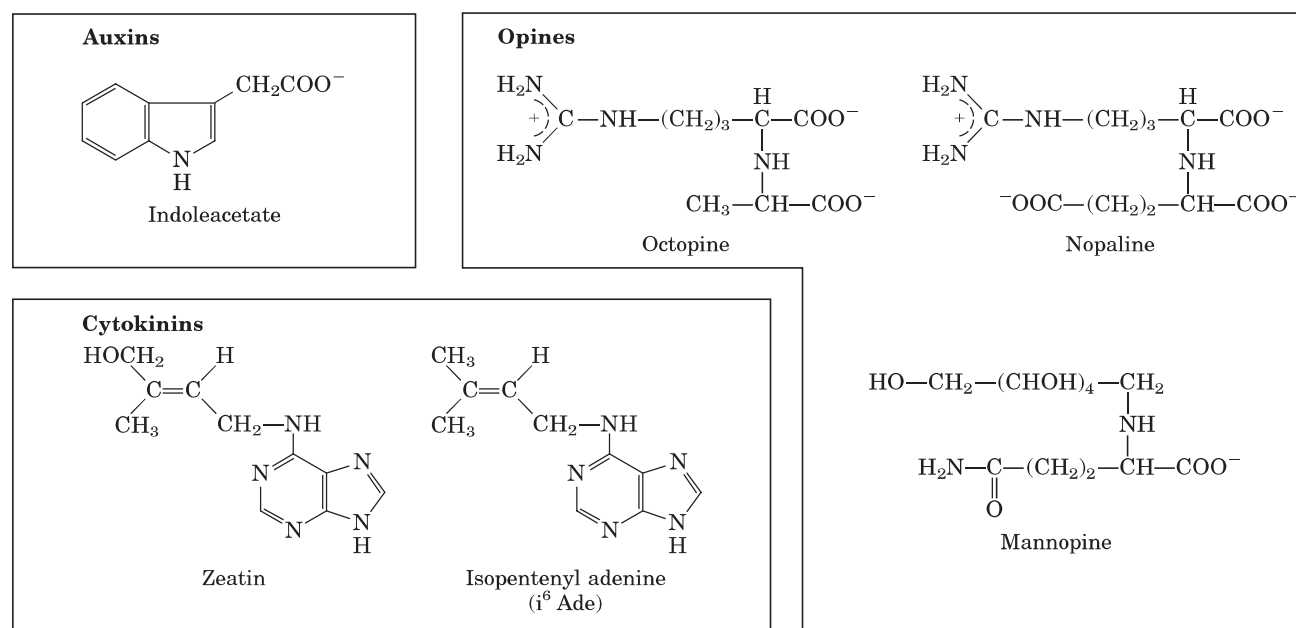


FIGURE 9-27 Metabolites produced in *Agrobacterium*-infected plant cells. Auxins and cytokinins are plant growth hormones. The most common auxin, indoleacetate, is derived from tryptophan. Cytokinins

are adenine derivatives. Opines generally are derived from amino acid precursors; at least 14 different opines are produced by enzymes encoded by the Ti plasmids of different *Agrobacterium* species.

This rare example of DNA transfer from a prokaryote to a eukaryotic cell is a natural genetic engineering process—one that researchers can harness to transfer recombinant DNA (instead of T DNA) to the plant genome. A common cloning strategy employs an *Agrobacterium* with two different recombinant plasmids. The first is a Ti plasmid from which the T DNA segment has been removed in the laboratory (Fig. 9–28a). The second is an *Agrobacterium*–*E. coli* shuttle vector in which the 25 bp repeats of the T DNA flank a foreign gene that the researcher wants to introduce into the plant cell, along with a selectable marker such as resistance to the antibiotic kanamycin (Fig. 9–28b). The engineered *Agrobacterium* is used to infect a leaf, but crown galls are not formed because the T DNA genes for the auxin, cytokinin, and opine biosynthetic enzymes are absent from both plasmids. Instead, the *vir* gene

products from the altered Ti plasmid direct the transformation of the plant cells by the foreign gene—the gene flanked by the T DNA 25 bp repeats in the second plasmid. The transformed plant cells can be selected by growth on agar plates that contain kanamycin, and addition of growth hormones induces the formation of new plants that contain the foreign gene in every cell.

The successful transfer of recombinant DNA into plants was vividly illustrated by an experiment in which the luciferase gene from fireflies was introduced into the cells of a tobacco plant (Fig. 9–29)—a favorite plant for transformation experiments because its cells are particularly easy to transform with *Agrobacterium*. The potential of this technology is not limited to the production of glow-in-the-dark plants, of course. The same approach has been used to produce crop plants that are resistant to herbicides, plant viruses, and insect pests (Fig. 9–30). Potential benefits include increased yields and less need for environmentally harmful agricultural chemicals.

Biotechnology can introduce new traits into a plant much faster than traditional methods of plant breeding. A prominent example is the development of soybeans that are resistant to the general herbicide glyphosate (the active ingredient in the product RoundUp). Glyphosate breaks down rapidly in the environment (glyphosate-sensitive plants can be planted in a treated area after as little as 48 hours), and its use does not generally lead to contamination of groundwater or carryover from one year to the next. A field of glyphosate-resistant soybeans can be treated once with glyphosate during a summer growing season to eliminate essentially all weeds in the field, while leaving the soybeans unaffected (Fig. 9–31). Potential pitfalls of the technology, such as the evolution of glyphosate-resistant weeds or the escape of difficult-to-control recombinant plants, remain a concern of researchers and the public.

FIGURE 9–28 A two-plasmid strategy to create a recombinant plant.

(a) One plasmid is a modified Ti plasmid that contains the *vir* genes but lacks T DNA. (b) The other plasmid contains a segment of DNA that bears both a foreign gene (the gene of interest, e.g., the gene for the insecticidal protein described in Fig. 9–30) and an antibiotic-resistance element (here, kanamycin resistance), flanked by the two 25 bp repeats of T DNA that are required for transfer of the plasmid genes to the plant chromosome. This plasmid also contains the replication origin needed for propagation in *Agrobacterium*.

When bacteria invade at the site of a wound (the edge of the cut leaf), the *vir* genes on the first plasmid mediate transfer into the plant genome of the segment of the second plasmid that is flanked by the 25 bp repeats. Leaf segments are placed on an agar dish that contains both kanamycin and appropriate levels of plant growth hormones, and new plants are generated from segments with the transformed cells. Nontransformed cells are killed by the kanamycin. The foreign gene and the antibiotic-resistance element are normally transferred together, so plant cells that grow in this medium generally contain the foreign gene.

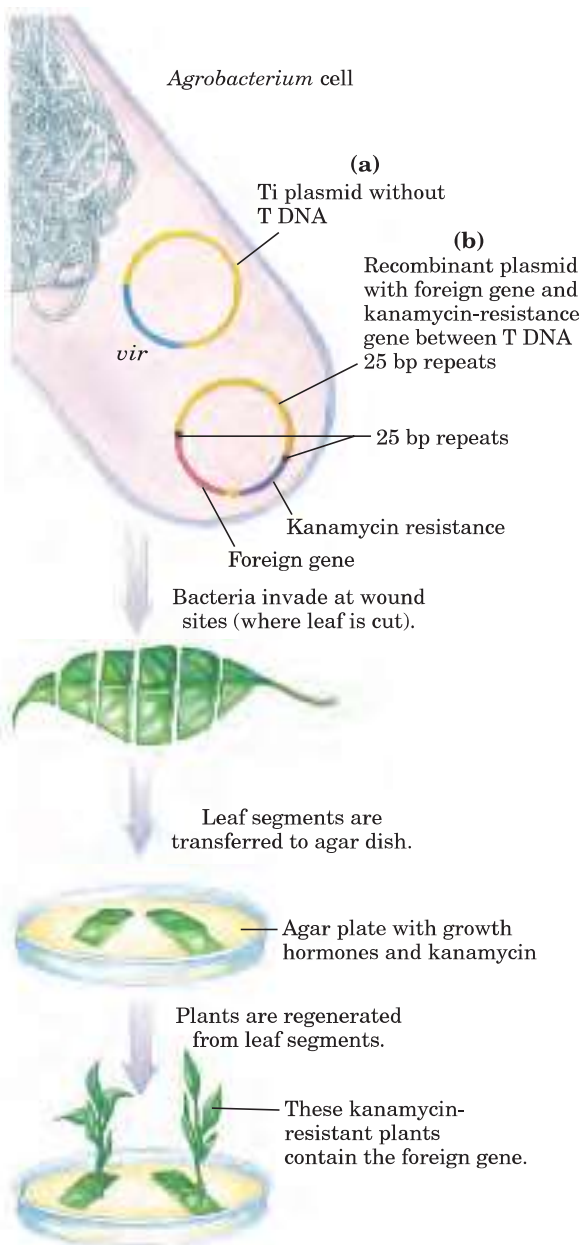




FIGURE 9-29 A tobacco plant expressing the gene for firefly luciferase.

Light was produced after the plant was watered with a solution containing luciferin, the substrate for the light-producing luciferase enzyme (see Box 13-2). Don't expect glow-in-the-dark ornamental plants at your local plant nursery anytime soon. The light is actually quite weak; this photograph required a 24-hour exposure. The real point—that this technology allows the introduction of new traits into plants—is nevertheless elegantly made.



FIGURE 9-30 Tomato plants engineered to be resistant to insect larvae.

Two tomato plants were exposed to equal numbers of moth larvae. The plant on the left has not been genetically altered. The plant on the right expresses a gene for a protein toxin derived from the bacterium *Bacillus thuringiensis*. This protein, introduced by a protocol similar to that depicted in Figure 9-28, is toxic to the larvae of some moth species while being harmless to humans and other organisms. Insect resistance has also been genetically engineered in cotton and other plants.

Manipulation of Animal Cell Genomes Provides Information on Chromosome Structure and Gene Expression

The transformation of animal cells by foreign genetic material offers an important mechanism for expanding our knowledge of the structure and function of animal genomes, as well as for the generation of animals with



(a)



(b)

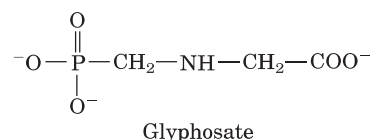


FIGURE 9-31 Glyphosate-resistant soybean plants. The photographs show two areas of a soybean field in Wisconsin. (a) Without glyphosate treatment, this part of the field is overrun with weeds. (b) Glyphosate-resistant soybean plants thrive in the glyphosate-treated section of the field. Glyphosate breaks down rapidly in the environment. Agricultural use of engineered plants such as these proceeds only after considerable deliberation, balancing the extraordinary promise of the technology with the need to select new traits with care. Both science and society as a whole have a stake in ensuring that the use of the resultant plants has no adverse impact on the environment or on human health.

new traits. This potential has stimulated intensive research into more-sophisticated means of cloning animals.

Most work of this kind requires a source of cells into which DNA can be introduced. Although intact tissues are often difficult to maintain and manipulate *in vitro*, many types of animal cells can be isolated and grown in the laboratory if their growth requirements are carefully met. Cells derived from a particular animal tissue and

grown under appropriate **tissue culture** conditions can maintain their differentiated properties (for example, a hepatocyte (liver cell) remains a hepatocyte) for weeks or even months.

No suitable plasmidlike vector is available for introducing DNA into an animal cell, so transformation usually requires the integration of the DNA into a host-cell chromosome. The efficient delivery of DNA to a cell nucleus and integration of this DNA into a chromosome without disrupting any critical genes remain the major technical problems in the genetic engineering of animal cells.

Available methods for carrying DNA into an animal cell vary in efficiency and convenience. Some success has been achieved with spontaneous uptake of DNA or electroporation, techniques roughly comparable to the common methods used to transform bacteria. They are inefficient in animal cells, however, transforming only 1 in 100 to 10,000 cells. **Microinjection**—the injection of DNA directly into a nucleus, using a very fine needle—has a high success rate for skilled practitioners, but the total number of cells that can be treated is small, because each must be injected individually.

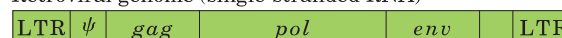
The most efficient and widely used methods for transforming animal cells rely on liposomes or viral vectors. Liposomes are small vesicles consisting of a lipid bilayer that encloses an aqueous compartment (see Fig. 11–4). Liposomes that enclose a recombinant DNA molecule can be fused with the membranes of target cells to deliver DNA into the cell. The DNA sometimes reaches the nucleus, where it can integrate into a chromosome (mostly at random locations). **Viral vectors** are even more efficient at delivering DNA. Animal viruses have effective mechanisms for introducing their nucleic acids into cells, and several types also have mechanisms to integrate their DNA into a host-cell chromosome. Some of these, such as retroviruses (see Fig. 26–30) and adenoviruses, have been modified to serve as viral vectors to introduce foreign DNA into mammalian cells.

The work on retroviral vectors illustrates some of the strategies being used (Fig. 9–32). When an engineered retrovirus enters a cell, its RNA genome is transcribed to DNA by reverse transcriptase and then integrated into the host genome by the enzyme viral integrase. Special regions of DNA are required for this

procedure: long terminal repeat (LTR) sequences to integrate retroviral DNA into the host chromosome and the ψ (psi) sequence to package the viral RNA in viral particles (see Fig. 26–30).

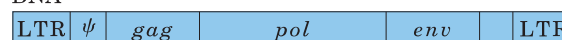
The *gag*, *pol*, and *env* genes of the retroviral genome, required for retroviral replication and assembly of viral particles, can be replaced with foreign DNA. To assemble viruses that contain the recombinant genetic information, researchers must introduce the DNA into cultured cells that are simultaneously infected with a “helper virus” that has the genes to produce viral particles but lacks the ψ sequence required for packaging. Thus the recombinant DNA can be transcribed and its

Retroviral genome (single-stranded RNA)



Reverse transcriptase converts RNA genome to double-stranded DNA.

DNA



Viral genes are replaced with a foreign gene.

Recombinant defective retroviral DNA



Recombinant DNA is introduced into cells in tissue culture.

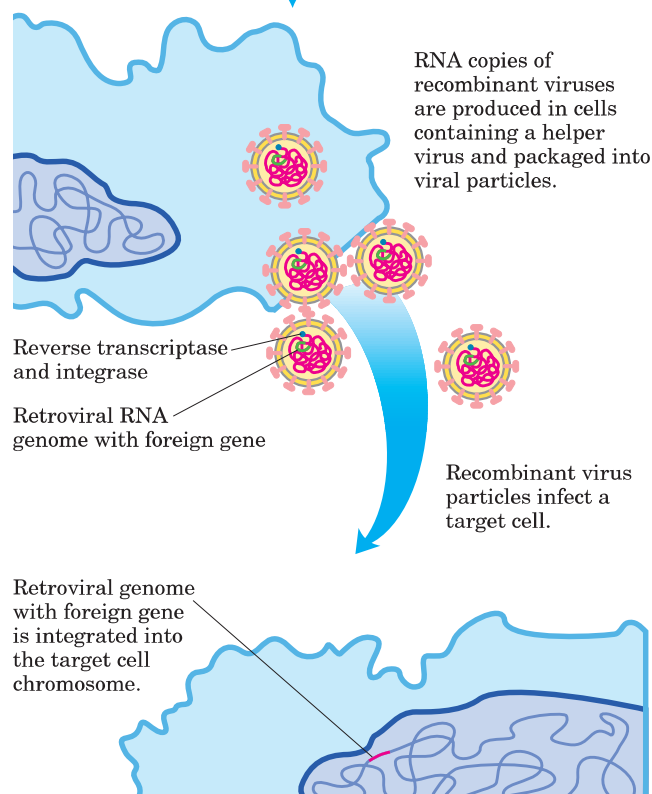


FIGURE 9–32 Use of retroviral vectors in mammalian cell cloning.

A typical retroviral genome (somewhat simplified here), engineered to carry a foreign gene (pink), is added to a host-cell tissue culture. The helper virus (not shown) lacks the packaging sequence, ψ , so its RNA transcripts cannot be packaged into viral particles, but it provides the *gag*, *pol*, and *env* gene products needed to package the engineered retrovirus into functional viral particles. This enables the foreign gene in the recombinant retroviral genome to be introduced efficiently into the target cells.

RNA packaged into viral particles. These particles can act as vectors to introduce the recombinant RNA into target cells. Viral reverse transcriptase and integrase enzymes (produced by the helper virus) are also packaged in the viral particle and introduced into the target cells. Once the engineered viral genome is inside a cell, these enzymes create a DNA copy of the recombinant viral RNA genome and integrate it into a host chromosome. The integrated recombinant DNA then becomes a permanent part of the target cell's chromosome and is replicated with the chromosome at every cell division. The cell itself is not endangered by the integrated viral DNA, because the recombinant virus lacks the genes needed to produce RNA copies of its genome and package them into new viral particles. The use of recombinant retroviruses is often the best method for introducing DNA into large numbers of mammalian cells.


Each type of virus has different attributes, so several classes of animal viruses are being engineered as vectors to transform mammalian cells. Adenoviruses, for example, lack a mechanism for integrating DNA into a chromosome. Recombinant DNA introduced via an adenoviral vector is therefore expressed for only a short time and then destroyed. This can be useful if the objective is transient expression of a gene.

Transformation of animal cells by any of the above techniques has its problems. Introduced DNA is generally integrated into chromosomes at random locations. Even when the foreign DNA contains a sequence similar to a sequence in a host chromosome, allowing targeting to that position, nonhomologous integrants still outnumber the targeted ones by several orders of magnitude. If these integration events disrupt essential genes, they can sometimes alter cellular functions (most cells are diploid or polyploid, however, so an integration usually leaves at least one unaffected copy of any given gene). A particularly poor outcome would involve an integration event that inadvertently activated a gene that stimulated cell division, potentially creating a cancer cell. Although such an event was once thought to be rare, recent trials suggest it is a significant hazard (Box 9-2). Finally, the site of an integration can determine the level of expression of the integrated gene, because integrants are not transcribed equally well everywhere in the genome.


Despite these challenges, the transformation of animal cells has been used extensively to study chromosome structure and the function, regulation, and expression of genes. The successful introduction of recombinant DNA into an animal can be illustrated by an experiment that permanently altered an easily observable inheritable physical trait. Microinjection of DNA into the nuclei of fertilized mouse eggs can produce efficient transformation (chromosomal integration). When the injected eggs are introduced into a female mouse and allowed to develop, the new gene is often expressed in some of the newborn mice. Those in which the germ line has been



FIGURE 9-33 Cloning in mice. The gene for human growth hormone was introduced into the genome of the mouse on the right. Expression of the gene resulted in the unusually large size of this mouse.

altered can be identified by testing *their* offspring. By careful breeding of these mice, researchers can establish a **transgenic** mouse line in which all the mice are homozygous for the new gene or genes. This technology was used to introduce into mice the gene for human growth hormone, under the control of an inducible promoter. When the mice were fed a diet that included the inducer, some of the mice that developed from injected embryos grew to an unusually large size (Fig. 9-33). Transgenic mice have now been produced with a wide range of genetic variations, including many relevant to human diseases and their control, pointing the way to human gene therapy (Box 9-2). A very similar approach is used to generate mice in which a particular gene has been inactivated ("knockout mice"), a way of establishing the function of the inactivated gene.  **Creating a Transgenic Mouse**

New Technologies Promise to Expedite the Discovery of New Pharmaceuticals

 It is difficult to summarize all the ways in which genomics and proteomics might affect the development of pharmaceutical agents, but a few examples illustrate the potential. Hypertension, congestive heart failure, hypercholesterolemia, and obesity are treated by pharmaceutical drugs that alter human physiology. Therapies are arrived at by identifying an enzyme or receptor involved in the process and discovering an inhibitor that interferes with its action. Proteomics will play an increasing role in identifying such potential drug targets. For example, the most potent vasoconstrictor known is the peptide hormone urotensin II. First discovered in fish spinal fluid, urotensin II is a small cyclic peptide, with 11 amino acid residues in humans and 12 or 13 in some other organisms. The vasoconstriction it induces can cause or exacerbate hypertension, congestive heart failure, and coronary artery disease. Some of the methods described in Section 9.3 for elucidating



The Human Genome and Human Gene Therapy

As biotechnology gained momentum in the 1980s, a rational approach to the treatment of genetic diseases became increasingly attractive. In principle, DNA can be introduced into human cells to correct inherited genetic deficiencies. Genetic correction may even be targeted to a specific tissue by inoculating an individual with a genetically engineered, tissue-specific virus carrying a payload of DNA to be incorporated into deficient cells. The goal is entrancing, but the research path is strewn with impediments.

Altering chromosomal DNA entails substantial risk—a risk that cannot be quantified in the early stages of discovery. Consequently, early efforts at human gene therapy were directed at only a small subset of genetic diseases. Panels of scientists and ethicists developed a list of several conditions that should be satisfied to justify the risk involved, including the following. (1) The genetic defect must be a well-characterized, single-gene disorder. (2) Both the mutant and the normal gene must be cloned and sequenced. (3) In the absence of a technique for eliminating the existing mutant gene, the functional gene must function well in the presence of the mutant gene. (4) Finally, and most important, the risks inherent in a new technology must be outweighed by the seriousness of the disease. Protocols for human clinical trials were submitted by scientists in several nations and reviewed for scientific rigor and ethical compliance by carefully selected advisory panels in each country; then human trials commenced.

Early targets of gene therapy included cancer and genetic diseases affecting the immune system. Immunity is mediated by leukocytes (white blood cells) of several different types, all arising from undifferentiated stem cells in the bone marrow. These cells divide quickly and have special metabolic requirements. Differentiation can become blocked in several ways, resulting in a condition called severe combined immune deficiency (SCID). One form of SCID results from genetically inherited defects in the gene encoding adenosine deaminase (ADA), an enzyme involved in nucleotide biosynthesis (discussed in Chapter 22). Another form of SCID arises from a defect in a cell-surface receptor protein that binds chemical signals called cytokines, which trigger differentiation. In both cases, the progenitor stem cells cannot differentiate

into the mature immune system cells, such as T and B lymphocytes (Chapter 5). Children with these rare human diseases are highly susceptible to bacterial and viral infections, and often suffer from a range of related physiological and neurological problems. In the absence of an effective therapy, the children must be confined in a sterile environment. About 20% of these children have a human leukocyte antigen (HLA)-identical sibling who can serve as a bone marrow transplant donor, a procedure that can cure the disease. The remaining children need a different approach.

The first human gene therapy trial was carried out at the National Institutes of Health in Bethesda, Maryland, in 1990. The patient was a four-year-old girl crippled by ADA deficiency. Bone marrow cells from the child were transformed with an engineered retrovirus containing a functional ADA gene; when the alteration of cells is done in this way—in the laboratory rather than in the living patient—the procedure is said to be done *ex vivo*. The treated cells were reintroduced into the patient's marrow. Four years later, the child was leading a normal life, going to school, and even testifying about her experiences before Congress. However, her recovery cannot be uniquely attributed to gene therapy. Before the gene therapy clinical trials began, researchers had developed a new treatment for ADA deficiency, in which synthetic ADA was administered in a complex with polyethylene glycol (PEG). For many ADA-SCID patients, injection of the ADA-PEG complex allowed some immune system development, with weight gain and reduced infection, although not full immune reconstitution. The new gene therapy was risky, and withdrawing the inoculation treatment from patients in the gene therapy trial was judged unethical. So trial participants received both treatments at once, making it unclear which treatment was primarily responsible for the positive clinical outcome. Nevertheless, the clinical trial provided important information: it was feasible to transfer genes *ex vivo* to large numbers of leukocytes, and cells bearing the transferred gene were still detectable years after treatment, suggesting that long-term correction was possible. In addition, the risk associated with use of the retroviral vectors appeared to be low.

Through the 1990s, hundreds of human gene therapy clinical trials were carried out, targeting a variety of genetic diseases, but the results in most cases were

protein-protein interactions have been used to demonstrate that urotensin II is bound by a G-protein-coupled receptor called GPR14. As we shall see in Chapter 12, G proteins play an important role in many signaling pathways. However, GPR14 was an “orphan” receptor,

in that human genome sequencing had identified it as a G-protein-coupled receptor, but with no known function. The association of urotensin II with GPR14 now makes the latter protein a key target for drug therapies aimed at interfering with the action of urotensin II.

discouraging. One major impediment proved to be the inefficiency of introducing new genes into cells. Transformation failed in many cells, and the number of transformed cells often proved insufficient to reverse the disorder. In the ADA trials, achieving a sufficient population of transformed cells was particularly difficult, because of the ongoing ADA-PEG therapy. Normally, stem cells with the correct ADA gene would have a growth advantage over the untreated cells, expanding their population and gradually predominating in the bone marrow. However, the injections of ADA-PEG in the same patients allowed the untransformed (ADA-deficient) cells to live and develop, and the transformed cells did not have the needed growth advantage to expand their population at the expense of the others.

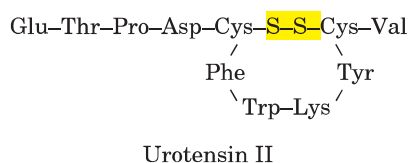
A gene therapy trial initiated in 1999 was successful in correcting a form of SCID caused by defective cytokine receptors (in particular a subunit called γ_c), as reported in 2000 by physician researchers in France, Italy, and Britain. These researchers introduced the corrected gene for the γ_c cytokine-receptor subunit into CD34⁺ cells. (The stem cells that give rise to immune system cells have a protein called CD34 on their surface; these cells can be separated from other bone marrow cells by antibodies to CD34.) The transformed cells were placed back into the patients' bone marrow. In this trial, introduction of the corrected gene clearly conferred a growth advantage over the untreated cells. A functioning immune system was detected in four of the first five patients within 6 to 12 weeks, and levels of mature immune system T lymphocytes reached the levels found in age-matched control subjects (who did not have SCID) within 6 to 8 months. Immune system function was restored, and nearly 4 years later (mid-2003) most of the children are leading normal lives. Similar results have been obtained with four additional patients. This provided dramatic confirmation that human gene therapy could cure a serious genetic disease.

In early 2003 came a setback. One of the original four patients who had received cells with the correct cytokine receptor gene developed a severe form of leukemia. During the gene therapy treatment, one of the introduced retroviruses had by chance inserted itself into a chromosome of one CD34⁺ cell, resulting in abnormally high expression of a gene called LMO-

2. The affected cell differentiated into an immune system T cell, and the elevated expression of LMO-2 led to uncontrolled growth of the cell, giving rise to the leukemia. As of mid-2003 the patient had responded well to chemotherapy, but there may be more chapters to write. The incident shows that early worries about the risk associated with retroviral vectors were well founded. After a review of the gene therapy trial protocols, including consultations with ethicists and parents of children affected by these diseases, further gene therapy trials are still planned for children who are not candidates for bone marrow transplants. The reason is simple enough. The potential benefit to the children with these debilitating conditions has been judged to outweigh the demonstrated risk.

Human gene therapy is not limited to genetic diseases. Cancer cells are being targeted by delivering genes for proteins that might destroy the cell or restore the normal control of cell division. Immune system cells associated with tumors, called tumor-infiltrating lymphocytes, can be genetically modified to produce tumor necrosis factor (TNF; see Fig. 12–50). When these lymphocytes are taken from a cancer patient, modified, and reintroduced, the engineered cells target the tumor, and the TNF they produce causes tumor shrinkage. AIDS may also be treatable with gene therapy; DNA that encodes an RNA molecule complementary to a vital HIV mRNA could be introduced into immune system cells (the targets of HIV). The RNA transcribed from the introduced DNA would pair with the HIV mRNA, preventing its translation and interfering with the virus's life cycle. Alternatively, a gene could be introduced that encodes an inactive form of one subunit of a multisubunit HIV enzyme; with one nonfunctional subunit, the entire enzyme might be inactivated.

Our growing understanding of the human genome and the genetic basis for some diseases brings the promise of early diagnosis and constructive intervention. As the early results demonstrate, however, the road to effective therapies will be a long one, with many detours. We need to learn more about cellular metabolism, more about how genes interact, and more about how to manage the dangers. The prospect of vanquishing life-destroying genetic defects and other debilitating diseases provides the motivation to press on.



Another objective of medical research is to identify new agents that can treat the diseases caused by human pathogens. This now means identifying enzymatic targets in microbial pathogens that can be inactivated with a new drug. The ideal microbial target enzyme

should be (1) essential to the pathogen cell's survival, (2) well-conserved among a wide range of pathogens, and (3) absent or significantly different in humans. The task of identifying metabolic processes that are critical to microorganisms but absent in humans is made much easier by comparative genomics, augmented by the functional information available from genomics and proteomics. ■

Recombinant DNA Technology Yields New Products and Challenges

The products of recombinant DNA technology range from proteins to engineered organisms. The technology can produce large amounts of commercially useful proteins, can design microorganisms for special tasks, and can engineer plants or animals with traits that are useful in agriculture or medicine. Some products of this technology have been approved for consumer or professional use, and many more are in development. Genetic engineering has been transformed over a few years from a promising new technology to a multibillion-dollar industry, with much of the growth occurring in the pharmaceutical industry. Some major classes of new products are listed in Table 9–3.



Erythropoietin is typical of the newer products. This protein hormone (M_r 51,000) stimulates erythrocyte production. People with diseases that com-

promise kidney function often have a deficiency of this protein, resulting in anemia. Erythropoietin produced by recombinant DNA technology can be used to treat these individuals, reducing the need for repeated blood transfusions. ■

Other applications of this technology continue to emerge. Enzymes produced by recombinant DNA technology are already used in the production of detergents, sugars, and cheese. Engineered proteins are being used as food additives to supplement nutrition, flavor, and fragrance. Microorganisms are being engineered with altered or entirely novel metabolic pathways to extract oil and minerals from ground deposits, to digest oil spills, and to detoxify hazardous waste dumps and sewage. Engineered plants with improved resistance to drought, frost, pests, and disease are increasing crop yields and reducing the need for agricultural chemicals. Complete animals can be cloned by moving an entire nucleus and all of its genetic material to a prepared egg from which the nucleus has been removed.

The extraordinary promise of modern biotechnology does not come without controversy. The cloning of mammals challenges societal mores and may be accompanied by serious deficiencies in the health and longevity of the cloned animal. If useful pharmaceutical agents can be produced, so can toxins suitable for biological warfare. The potential for hazards posed by the release of engineered plants and other organisms into

TABLE 9–3 Some Recombinant DNA Products in Medicine	
Product category	Examples/uses
Anticoagulants	Tissue plasminogen activator (TPA); activates plasmin, an enzyme involved in dissolving clots; effective in treating heart attack patients.
Blood factors	Factor VIII; promotes clotting; it is deficient in hemophiliacs; treatment with factor VIII produced by recombinant DNA technology eliminates infection risks associated with blood transfusions.
Colony-stimulating factors	Immune system growth factors that stimulate leukocyte production; treatment of immune deficiencies and infections.
Erythropoietin	Stimulates erythrocyte production; treatment of anemia in patients with kidney disease.
Growth factors	Stimulate differentiation and growth of various cell types; promote wound healing.
Human growth hormone	Treatment of dwarfism.
Human insulin	Treatment of diabetes.
Interferons	Interfere with viral reproduction; used to treat some cancers.
Interleukins	Activate and stimulate different classes of leukocytes; possible uses in treatment of wounds, HIV infection, cancer, and immune deficiencies.
Monoclonal antibodies	Extraordinary binding specificity is used in: diagnostic tests; targeted transport of drugs, toxins, or radioactive compounds to tumors as a cancer therapy; many other applications.
Superoxide dismutase	Prevents tissue damage from reactive oxygen species when tissues briefly deprived of O ₂ during surgery suddenly have blood flow restored.
Vaccines	Proteins derived from viral coats are as effective in “priming” an immune system as is the killed virus more traditionally used for vaccines, and are safer; first developed was the vaccine for hepatitis B.

the biosphere continues to be monitored carefully. The full range of the long-term consequences of this technology for our species and for the global environment is impossible to foresee, but will certainly demand our increasing understanding of both cellular metabolism and ecology.

SUMMARY 9.4 Genome Alterations and New Products of Biotechnology

- Advances in whole genome sequencing and genetic engineering methods are enhancing our ability to modify genomes in all species.

- Cloning in plants, which makes use of the Ti plasmid vector from *Agrobacterium*, allows the introduction of new plant traits.
- In animal cloning, researchers introduce foreign DNA primarily with the use of viral vectors or microinjection. These techniques can produce transgenic animals and provide new methods for human gene therapy.
- The use of genomics and proteomics in basic and pharmaceutical research is greatly advancing the discovery of new drugs. Biotechnology is also generating an ever-expanding range of other products and technologies.

Key Terms

Terms in bold are defined in the glossary.

cloning 306	genomics 317	restriction fragment length
vector 307	genomic library 318	polymorphisms (RFLPs) 322
recombinant DNA 307	contig 318	Southern blot 322
restriction endonucleases 307	sequence-tagged site (STS) 318	single nucleotide polymorphisms (SNPs) 324
DNA ligase 307	complementary DNA (cDNA) 318	proteome 325
plasmid 311	cDNA library 318	proteomics 325
bacterial artificial chromosome (BAC) 313	expressed sequence tag (EST) 318	orthologs 326
yeast artificial chromosome (YAC) 314	epitope tag 319	synteny 326
site-directed mutagenesis 316	polymerase chain reaction (PCR) 319	DNA microarray 326
fusion protein 317	DNA fingerprinting 322	Ti plasmid 330
		transgenic 335

Further Reading

General

Jackson, D.A., Symons, R.H., & Berg, P. (1972) Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **69**, 2904–2909.

The first recombinant DNA experiment linking DNA from two species.

Lobban, P.E. & Kaiser, A.D. (1973) Enzymatic end-to-end joining of DNA molecules. *J. Mol. Biol.* **78**, 453–471.

Report of the first recombinant DNA experiment.

Sambrook, J., Fritsch, E.F., & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Although supplanted by more recent manuals, this three-volume set includes much useful background information on the biological, chemical, and physical principles underlying both classic and still-current techniques.

Gene Cloning

Arnheim, N. & Erlich, H. (1992) Polymerase chain reaction strategy. *Annu. Rev. Biochem.* **61**, 131–156.

Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M., & Paabo, S. (2001) Ancient DNA. *Nat. Rev. Genet.* **2**, 353–359.

Successes and pitfalls in the retrieval of DNA from very old samples.

Ivanov, P.L., Wadhams, M.J., Roby, R.K., Holland, M.M., Weedn, V.W., & Parsons, T.J. (1996) Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. *Nat. Genet.* **12**, 417–420.

Lindahl, T. (1997) Facts and artifacts of ancient DNA. *Cell* **90**, 1–3.

Good description of how nucleic acid chemistry affects the retrieval of DNA in archaeology.

Genomics

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. (1991) Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science* **252**, 1651–1656.

The paper that introduced expressed sequence tags (ESTs).

Bamshad, M. & Wooding, S.P. (2003) Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**, 99A–111A.

Use of the human genome to trace human evolution.

Brenner, S. (2004) Genes to genomics. *Annu. Rev. Genet.* **38**, in press.

Carroll, S.B. (2003) Genetics and the making of *Homo sapiens*. *Nature* **422**, 849–857.

Clark, M.S. (1999) Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* **21**, 121–130.

Useful background on some reasons for the importance of sequencing the genomes of many organisms.

Collins, F.S., Green, E.D., Guttmacher, A.E., & Guyer, M.S. (2003) A vision for the future of genomics research. *Nature* **422**, 835–847.

A wide-ranging overview of the enormous potential of genomics research.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M.,

FitzHugh, W., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Discussion of the draft genome sequence put together by the international Human Genome Project. Many other useful articles are to be found in this issue.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001) The sequence of the human genome. *Science* **291**, 1304–1351.

Description of the draft of the human genome sequence produced by Celera Corporation. Many other articles in the same issue provide insight and additional information.

Proteomics

Brown, P.O. & Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21**, 33–37.

Eisenberg, D., Marcotte, E.M., Xenarios, I., & Yeates, T.O. (2000) Protein function in the post-genomic era. *Nature* **405**, 823–826.

Pandey, A. & Mann, M. (2000) Proteomics to study genes and genomes. *Nature* **405**, 837–846.

An especially good description of the various strategies and methods used to identify proteins and their functions.

Zhu, H., Bilgin, M., & Snyder, M. (2003) Proteomics. *Annu. Rev. Biochem.* **72**, 783–812.

Applying Biotechnology

Foster, E.A., Jobling, M.A., Taylor, P.G., Donnelly, P., de Knijff, P., Mieremet, R., Zerjal, T., & Tyler-Smith, C. (1999) The Thomas Jefferson paternity case. *Nature* **397**, 32.

Last article of a series in an interesting case study of the uses of biotechnology to address historical questions.

Hansen, G. & Wright M.S. (1999) Recent advances in the transformation of plants. *Trends Plant Sci.* **4**, 226–231.

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P., & Lovell-Badge, R. (1991) Male development of chromosomally female mice transgenic for *Sry*. *Nature* **351**, 117–121.

Recombinant DNA technology shows that a single gene directs development of chromosomally female mice into males.

Lapham, E.V., Kozma, C., & Weiss, J. (1996) Genetic discrimination: perspectives of consumers. *Science* **274**, 621–624.

The upside and downside of knowing what is in your genome.

Mahowald, M.B., Verp, M.S., & Anderson, R.R. (1998) Genetic counseling: clinical and ethical challenges. *Annu. Rev. Genet.* **32**, 547–559.

Ohlstein, E.H., Ruffolo, R.R., Jr., & Elliott, J.D. (2000) Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol.* **40**, 177–191.

Palmiter, R.D., Brinster, R.L., Hammer, R.E., Trumbauer, M.E., Rosenfeld, M.G., Birnberg, N.C., & Evans, R.M. (1982) Dramatic growth of mice that develop from eggs microinjected with metallothionein-growth hormone fusion genes. *Nature* **300**, 611–615.

A description of how to make giant mice.

Pfeifer, A. & Verma, I. M. (2001) Gene therapy: promises and problems. *Annu. Rev. Genomics Hum. Genet.* **2**, 177–211.

Thompson, J. & Donkersloot, J.A. (1992) *N*-(Carboxyalkyl) amino acids: occurrence, synthesis, and functions. *Annu. Rev. Biochem.* **61**, 517–557.

A summary of the structure and biological functions of opines.

Wadhwa, P.D., Zielske, S.P., Roth, J.C., Ballas, C.B.,

Bowman, J.E., & Gerson, S.L. (2002) Cancer gene therapy: scientific basis. *Annu. Rev. Med.* **53**, 437–452.

Problems

1. Cloning When joining two or more DNA fragments, a researcher can adjust the sequence at the junction in a variety of subtle ways, as seen in the following exercises.

(a) Draw the structure of each end of a linear DNA fragment produced by an *EcoRI* restriction digest (include those sequences remaining from the *EcoRI* recognition sequence).

(b) Draw the structure resulting from the reaction of this end sequence with DNA polymerase I and the four deoxynucleoside triphosphates (see Fig. 8–36).

(c) Draw the sequence produced at the junction that arises if two ends with the structure derived in (b) are ligated (see Fig. 25–16).

(d) Draw the structure produced if the structure derived in (a) is treated with a nuclease that degrades only single-stranded DNA.

(e) Draw the sequence of the junction produced if an

end with structure (b) is ligated to an end with structure (d).

(f) Draw the structure of the end of a linear DNA fragment that was produced by a *PvuII* restriction digest (include those sequences remaining from the *PvuII* recognition sequence).

(g) Draw the sequence of the junction produced if an end with structure (b) is ligated to an end with structure (f).

(h) Suppose you can synthesize a short duplex DNA fragment with any sequence you desire. With this synthetic fragment and the procedures described in (a) through (g), design a protocol that would remove an *EcoRI* restriction site from a DNA molecule and incorporate a new *BamHI* restriction site at approximately the same location. (See Fig. 9–3.)

(i) Design four different short synthetic double-stranded DNA fragments that would permit ligation of structure (a) with a DNA fragment produced by a *PstI* restriction

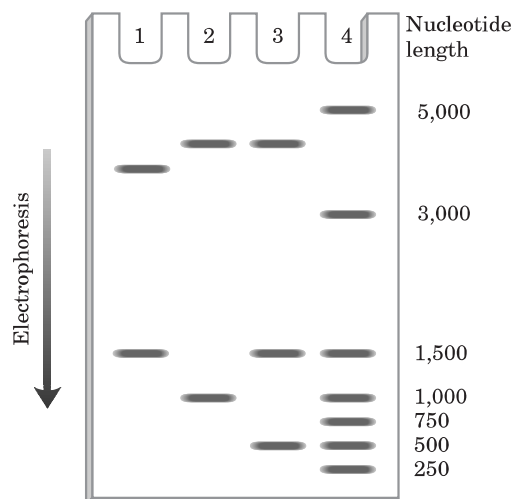
digest. In one of these fragments, design the sequence so that the final junction contains the recognition sequences for both *EcoRI* and *PstI*. In the second and third fragments, design the sequence so that the junction contains only the *EcoRI* and only the *PstI* recognition sequence, respectively. Design the sequence of the fourth fragment so that neither the *EcoRI* nor the *PstI* sequence appears in the junction.

2. Selecting for Recombinant Plasmids When cloning a foreign DNA fragment into a plasmid, it is often useful to insert the fragment at a site that interrupts a selectable marker (such as the tetracycline-resistance gene of pBR322). The loss of function of the interrupted gene can be used to identify clones containing recombinant plasmids with foreign DNA. With a bacteriophage λ vector it is not necessary to do this, yet one can easily distinguish vectors that incorporate large foreign DNA fragments from those that do not. How are these recombinant vectors identified?

3. DNA Cloning The plasmid cloning vector pBR322 (see Fig. 9–4) is cleaved with the restriction endonuclease *PstI*. An isolated DNA fragment from a eukaryotic genome (also produced by *PstI* cleavage) is added to the prepared vector and ligated. The mixture of ligated DNAs is then used to transform bacteria, and plasmid-containing bacteria are selected by growth in the presence of tetracycline.

(a) In addition to the desired recombinant plasmid, what other types of plasmids might be found among the transformed bacteria that are tetracycline resistant? How can the types be distinguished?

(b) The cloned DNA fragment is 1,000 bp long and has an *EcoRI* site 250 bp from one end. Three different recombinant plasmids are cleaved with *EcoRI* and analyzed by gel electrophoresis, giving the patterns shown. What does each pattern say about the cloned DNA? Note that in pBR322, the *PstI* and *EcoRI* restriction sites are about 750 bp apart. The entire plasmid with no cloned insert is 4,361 bp. Size markers in lane 4 have the number of nucleotides noted.



4. Identifying the Gene for a Protein with a Known Amino Acid Sequence Using Figure 27–7 to translate the genetic code, design a DNA probe that would allow you to identify the gene for a protein with the following amino-terminal amino acid sequence. The probe should be 18 to 20 nucleotides long, a size that provides adequate specificity if there is sufficient homology between the probe and the gene.

H_3N^+ –Ala–Pro–Met–Thr–Trp–Tyr–Cys–Met–Asp–Trp–Ile–Ala–Gly–Gly–Pro–Trp–Phe–Arg–Lys–Asn–Thr–Lys–

5. Designing a Diagnostic Test for a Genetic Disease

Huntington's disease (HD) is an inherited neurodegenerative disorder, characterized by the gradual, irreversible impairment of psychological, motor, and cognitive functions. Symptoms typically appear in middle age, but onset can occur at almost any age. The course of the disease can last 15 to 20 years. The molecular basis of the disease is becoming better understood. The genetic mutation underlying HD has been traced to a gene encoding a protein (M_r 350,000) of unknown function. In individuals who will not develop HD, a region of the gene that encodes the amino terminus of the protein has a sequence of CAG codons (for glutamine) that is repeated 6 to 39 times in succession. In individuals with adult-onset HD, this codon is typically repeated 40 to 55 times. In individuals with childhood-onset HD, this codon is repeated more than 70 times. The length of this simple trinucleotide repeat indicates whether an individual will develop HD, and at approximately what age the first symptoms will occur.

A small portion of the amino-terminal coding sequence of the 3,143-codon HD gene is given below. The nucleotide sequence of the DNA is shown in black, the amino acid sequence corresponding to the gene is shown in blue, and the CAG repeat is shaded. Using Figure 27–7 to translate the genetic code, outline a PCR-based test for HD that could be carried out using a blood sample. Assume the PCR primer must be 25 nucleotides long. By convention, unless otherwise specified a DNA sequence encoding a protein is displayed with the coding strand (the sequence identical to the mRNA transcribed from the gene) on top such that it is read 5' to 3', left to right.

```

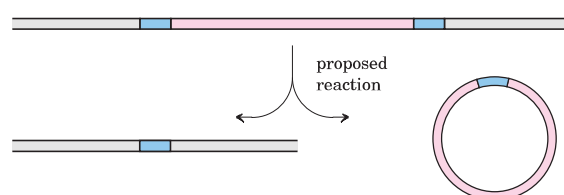
307 ATGGCGACCCCTGGAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTC
1  M A T L E K L M K A F E S L K S F
358 CAGCAGTTCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG
18  Q Q F Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q
409 CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG
35  Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q Q
460 CCGCCTCCTCAGCTTCCTCAGCCGCCGCCGCCGCCGCCGCCGCCGCCG
52  P P P Q L P Q P P P

```

Source: The Huntington's Disease Collaborative Research Group, (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983.

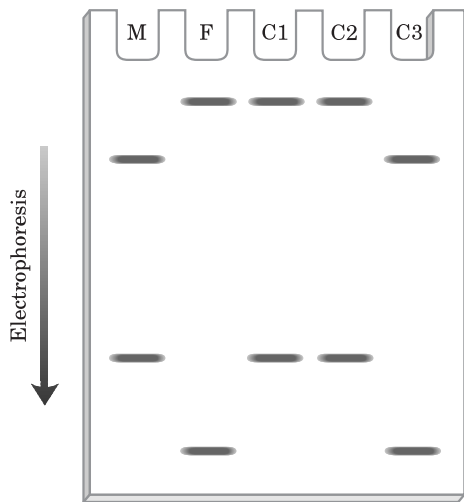
6. Using PCR to Detect Circular DNA Molecules

In a species of ciliated protist, a segment of genomic DNA is sometimes deleted. The deletion is a genetically programmed reaction associated with cellular mating. A researcher proposes that the DNA is deleted in a type of recombination called site-specific recombination, with the DNA on either end of the segment joined together and the deleted DNA ending up as a circular DNA reaction product.

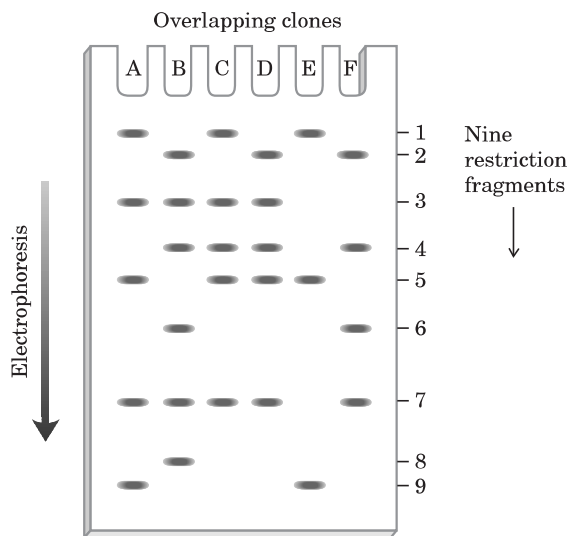


Suggest how the researcher might use the polymerase chain reaction (PCR) to detect the presence of the circular form of the deleted DNA in an extract of the protist.

7. RFLP Analysis for Paternity Testing DNA fingerprinting and RFLP analysis are often used to test for paternity. A child inherits chromosomes from both mother and father, so DNA from a child displays restriction fragments derived from each parent. In the gel shown here, which child, if any, can be excluded as being the biological offspring of the putative father? Explain your reasoning. Lane M is the sample from the mother, F from the putative father, and C1, C2, and C3 from the children.

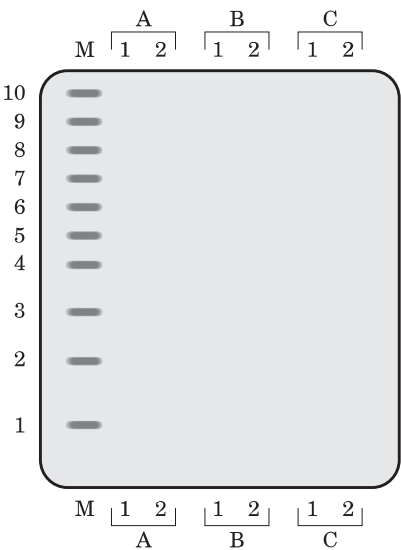
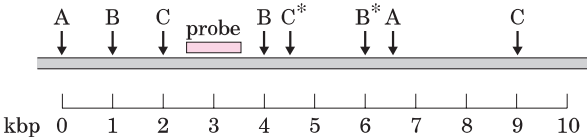


8. Mapping a Chromosome Segment A group of overlapping clones, designated A through F, is isolated from one region of a chromosome. Each of the clones is separately cleaved by a restriction enzyme and the pieces resolved by agarose gel electrophoresis, with the results shown in the figure below. There are nine different restriction fragments in this chromosomal region, with a subset appearing in each clone. Using this information, deduce the order of the restriction fragments in the chromosome.



9. Cloning in Plants The strategy outlined in Figure 9–28 employs *Agrobacterium* cells that contain two separate plasmids. Suggest why the sequences on the two plasmids are not combined on one plasmid.

10. DNA Fingerprinting and RFLP Analysis DNA is extracted from the blood cells of two different humans, individuals 1 and 2. In separate experiments, the DNA from each individual is cleaved by restriction endonucleases A, B, and C, and the fragments separated by electrophoresis. A hypothetical map of a 10,000 bp segment of a human chromosome is shown (1 kbp = 1,000 bp). Individual 2 has point mutations that eliminate restriction recognition sites B* and C*. You probe the gel with a radioactive oligonucleotide complementary to the indicated sequence and expose a piece of x-ray film to the gel. Indicate where you would expect to see bands on the film. The lanes of the gel are marked in the accompanying diagram.



11. Use of Photolithography to Make a DNA Microarray Figure 9–21 shows the first steps in the process of making a DNA microarray, or DNA chip, using photolithography. Describe the remaining steps needed to obtain the desired sequences (a different four-nucleotide sequence on each of the four spots) shown in the first panel of the figure. After each step, give the resulting nucleotide sequence attached at each spot.

12. Cloning in Mammals The retroviral vectors described in Figure 9–32 make possible the efficient integration of foreign DNA into a mammalian genome. Explain how these vectors, which lack genes for replication and viral packaging (*gag*, *pol*, *env*), are assembled into infectious viral particles. Suggest why it is important that these vectors lack the replication and packaging genes.