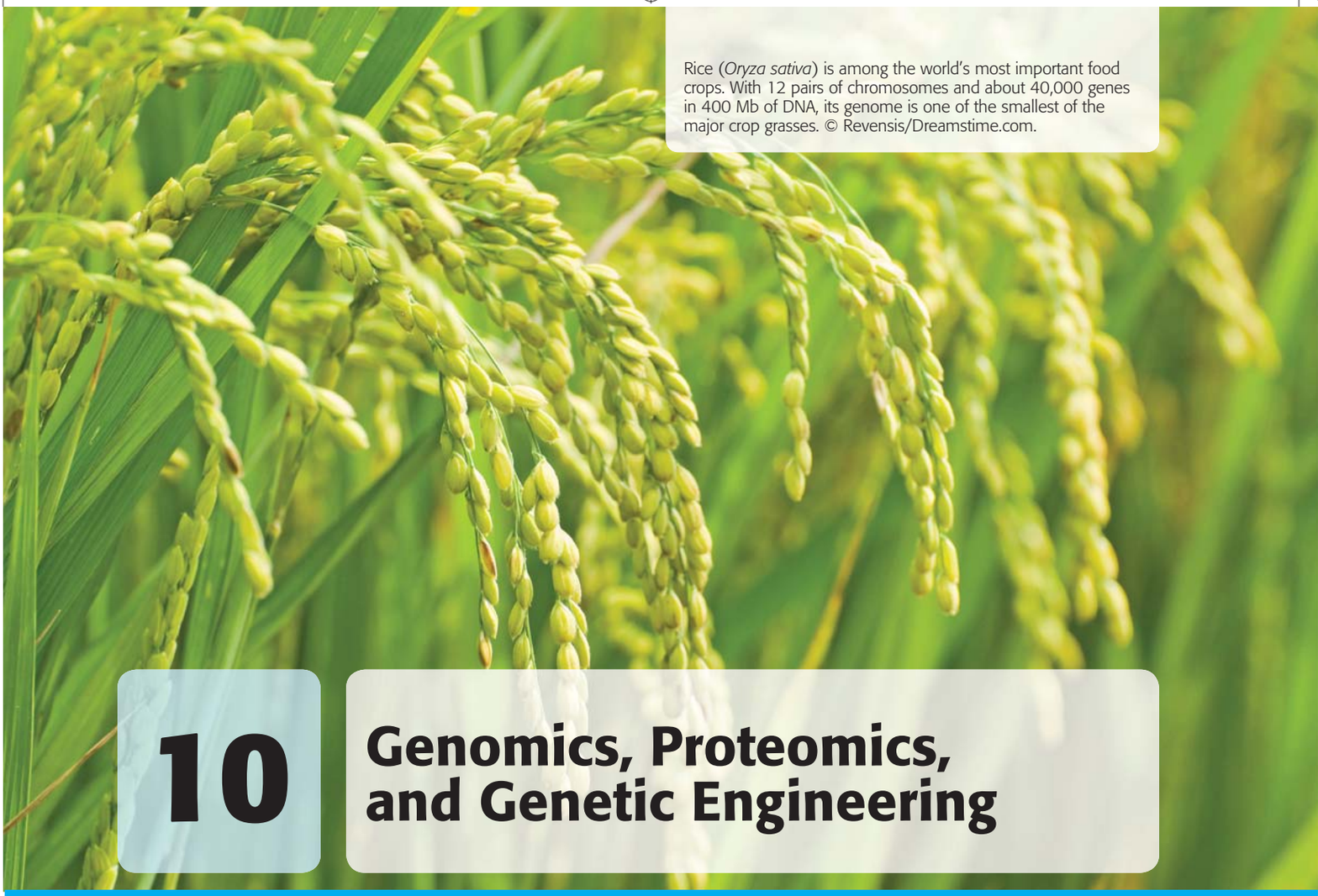


Rice (*Oryza sativa*) is among the world's most important food crops. With 12 pairs of chromosomes and about 40,000 genes in 400 Mb of DNA, its genome is one of the smallest of the major crop grasses. © Revensis/Dreamstime.com.

10 Genomics, Proteomics, and Genetic Engineering

»» CHAPTER ORGANIZATION

Here is an incomplete list of mammalian genomes that have been sequenced: human, chimpanzee, monkey, lemur, mouse, dog, cat, bat, squirrel, rabbit, guinea pig, armadillo, hedgehog, shrew, opposum, horse, elephant, pangolin, sloth, llama, and dolphin. Also sequenced are the genomes of many species of fruit flies, worms, and fungi, hundreds of bacteria, mitochondria, and chloroplasts, and thousands of viruses. Together these genomes represent a colossal amount of sequence data available for analysis and comparison. In addition to the genome sequences, methods are also available for identifying which genes in the genome are transcribed in particular tissue types, at specific times in development, or at different stages of the cell cycle. These are the raw data of **genomics**, which deals with the DNA sequence, organization, function, and evolution of genomes. The counterpart at the level of proteins is **proteomics**, which aims to identify all the proteins in a cell or organism (including any post-translationally modified forms), as well as their cellular localization, functions, and interactions. Proteomics makes use of methods discussed later in this chapter that identify which proteins in the cell undergo physical contact, thereby revealing *networks* of interacting proteins.

Genomics was made possible by the invention of techniques originally devised for the manipulation of genes and the creation of genetically engineered organisms with novel genotypes and phenotypes. We refer to this approach as **recombinant DNA**, but it also goes by the names *gene cloning* or *genetic engineering*. The basic technique is quite simple: DNA is isolated and cut into fragments by one or more restriction enzymes; then the fragments are joined together in a new combination and introduced back into a cell or organism to change its genotype in a directed, predetermined way. Such genetically engineered organisms are called **transgenic organisms**. Transgenics are often created for experimental studies, but an important application is the development of improved varieties of domesticated animals and crop plants, in which case a transgenic organism is often called a *genetically modified organism (GMO)*. Specific examples of genetically modified organisms are considered later in this chapter.

10.1 Cloning a DNA molecule takes place in several steps.

In genetic engineering, the immediate goal of an experiment is usually to insert a *particular* fragment of chromosomal DNA into a plasmid or a viral DNA molecule. This is accomplished by techniques for breaking DNA molecules at specific sites and for isolating particular DNA fragments.

Restriction enzymes cleave DNA into fragments with defined ends.

DNA fragments are usually obtained by the treatment of DNA samples with restriction enzymes. *Restriction enzymes* are nucleases that cleave DNA wherever it contains a particular short sequence of nucleotides that matches the *restriction site* of the enzyme (see Section 6.6). Most restriction sites consist of four or six nucleotides, within which the restriction enzyme makes two single-strand breaks, one in each strand, generating 3′-OH and 5′-P groups at each position. About a thousand restriction enzymes, nearly all with different restriction site specificities, have been isolated from microorganisms.

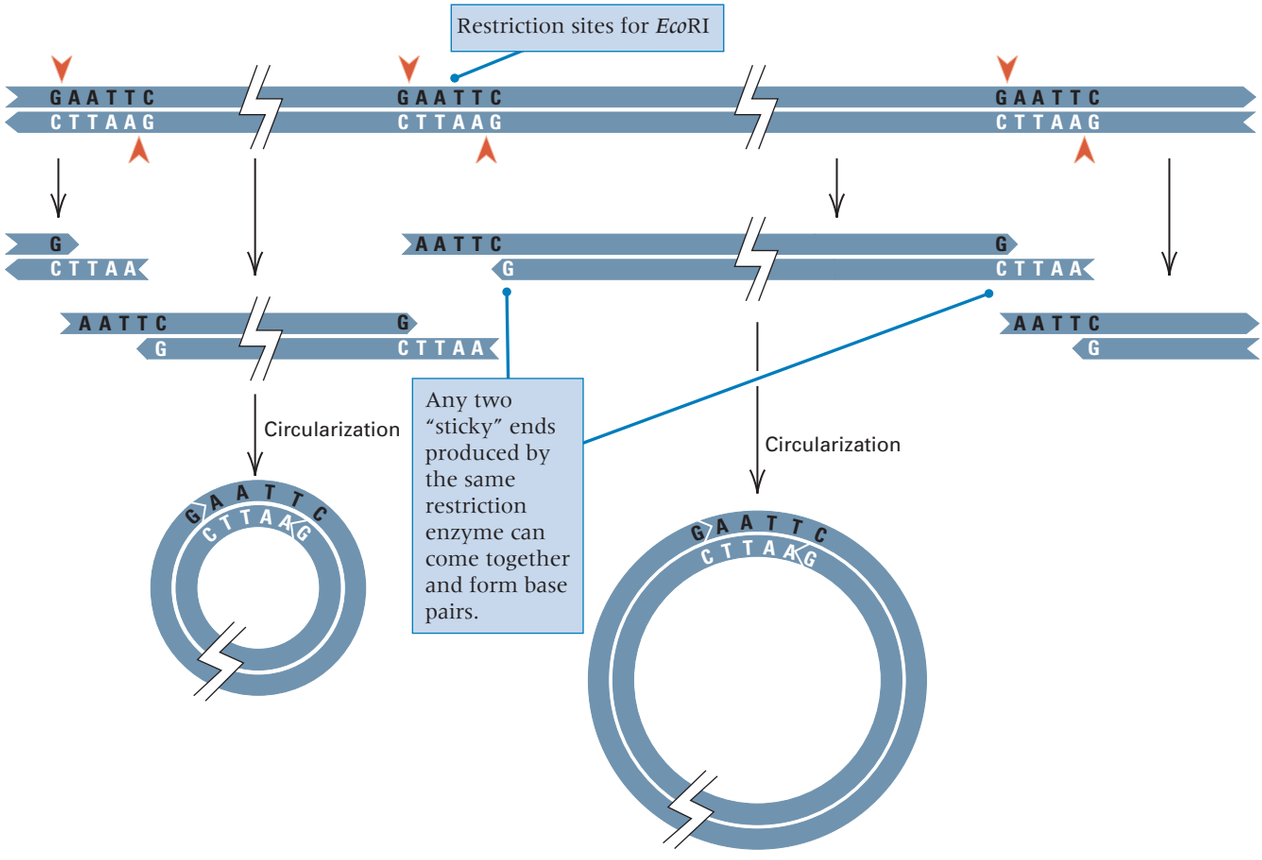


FIGURE 10.1 Circularization of DNA fragments produced by a restriction enzyme. The red arrowheads indicate the *Eco*RI cleavage sites.

Most restriction sites are symmetrical in the sense that the sequence is identical in both strands of the DNA duplex. For example, the restriction enzyme *Eco*RI, isolated from *E. coli*, has the restriction site 5′-GAATTC-3′; the sequence of the other strand is 3′-CTTAAG-5′, which is identical but written with the 3′ end at the left. *Eco*RI cuts each strand between the G and the A. The term *palindrome* is used to denote this type of symmetrical sequence.

Soon after restriction enzymes were discovered, observations with the electron microscope indicated that the fragments produced by many restriction enzymes could spontaneously form circles. The circles could be made linear again by heating. On the other hand, if the circles that formed spontaneously were treated with DNA ligase, which joins 3′-OH and 5′-P groups, then they could no longer be made linear with heat because the ends were covalently linked by the DNA ligase. This observation was the first evidence for three important features of restriction enzymes:

- Restriction enzymes cleave DNA molecules in palindromic sequences.
- The breaks need not be directly opposite one another in the two DNA strands.
- Enzymes that cleave the DNA strands asymmetrically generate DNA fragments with complementary ends.

These properties are illustrated for *Eco*RI in FIGURE 10.1.

Most restriction enzymes are like *Eco*RI in that they make staggered cuts in the DNA strands, producing single-stranded ends called **sticky ends** that can adhere to each other because they contain complementary nucleotide sequences. Some restriction enzymes (such as *Eco*RI) leave a single-stranded overhang at the 5′ end (FIGURE 10.2, part A); others leave a 3′ overhang. A number of restriction enzymes cleave both DNA strands at the center of symmetry, forming **blunt ends**. Part B of Figure 10.2 shows the blunt ends produced by the enzyme *Bal*I. Blunt ends also can be ligated by DNA ligase. However, whereas ligation of sticky ends re-creates the original restriction site, any blunt end can join with any other blunt end and not necessarily create a restriction site.

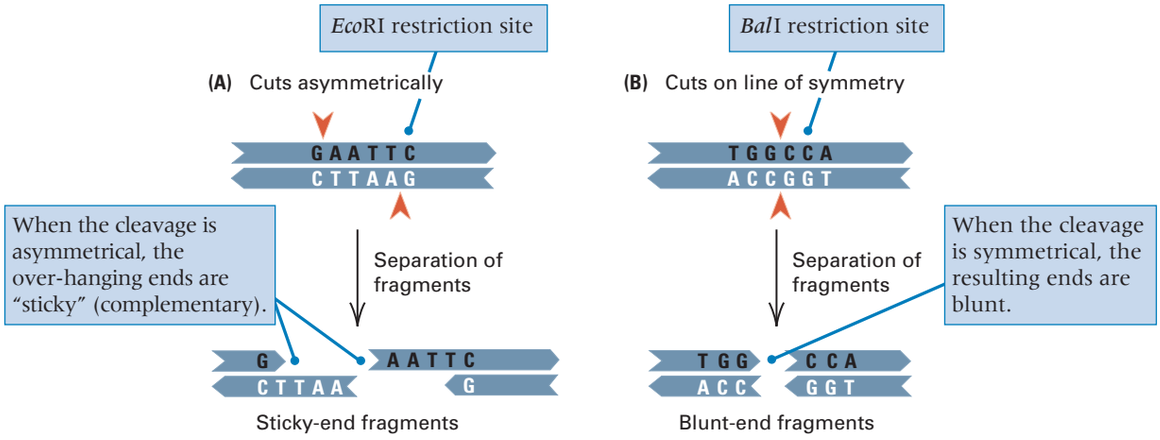


FIGURE 10.2 Two types of cuts made by restriction enzymes. The red arrowheads indicate the cleavage sites. (A) Cuts made in each strand at an equal distance from the center of symmetry of the restriction site. (B) Cuts made in each strand at the center of symmetry of the restriction site.

Most restriction enzymes recognize their restriction sequence without regard to the source of the DNA. Thus,

»» KEY CONCEPT

Restriction fragments of DNA obtained from one organism have the same sticky ends as restriction fragments from another organism if they were produced by the same restriction enzyme.

This principle will be seen to be one of the foundations of recombinant DNA technology.

Because most restriction enzymes recognize a unique sequence, the number of cuts made in the DNA of an organism by a particular enzyme is limited. For example, an *E. coli* DNA molecule contains 4.6×10^6 base pairs, and any enzyme that cleaves a six-base restriction site will cut the molecule into about a thousand fragments. This number of fragments follows from the fact that any particular six-base sequence (including a six-base restriction site) is expected to occur in a random sequence every $4^6 = 4096$ base pairs, on the average, assuming equal frequencies of the four bases. For the same reason, mammalian nuclear DNA would be cut into about a million fragments. These large numbers are still small compared with the number that would be produced if breakage occurred at completely random sequences. Of special interest are the smaller DNA molecules, such as viral or plasmid DNA, which may have from only one to ten sites of cutting (or even none) for particular enzymes. Plasmids that contain a single site for a particular enzyme are especially valuable, as we will see shortly.

Restriction fragments are joined end to end to produce recombinant DNA.

In genetic engineering, a particular DNA fragment of interest is joined to a *vector*, a relatively small DNA molecule that is able to replicate inside a cell and that usually contains one or more sequences able to confer antibiotic resistance (or some other detectable phenotype) on the cell. The simplest types of vectors are plasmids whose DNA is double-stranded and circular (FIGURE 10.3). When the DNA fragment of interest has been joined to the vector, the

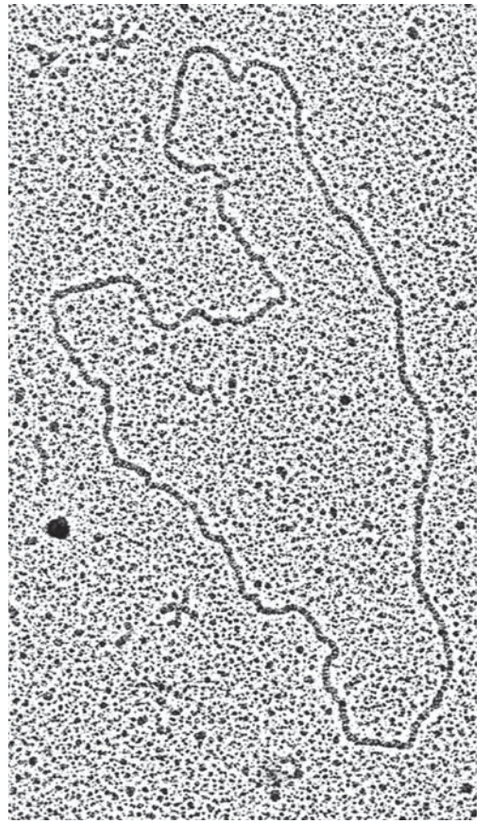


FIGURE 10.3 Electron micrograph of a circular plasmid used as a vector for cloning in *E. coli*.

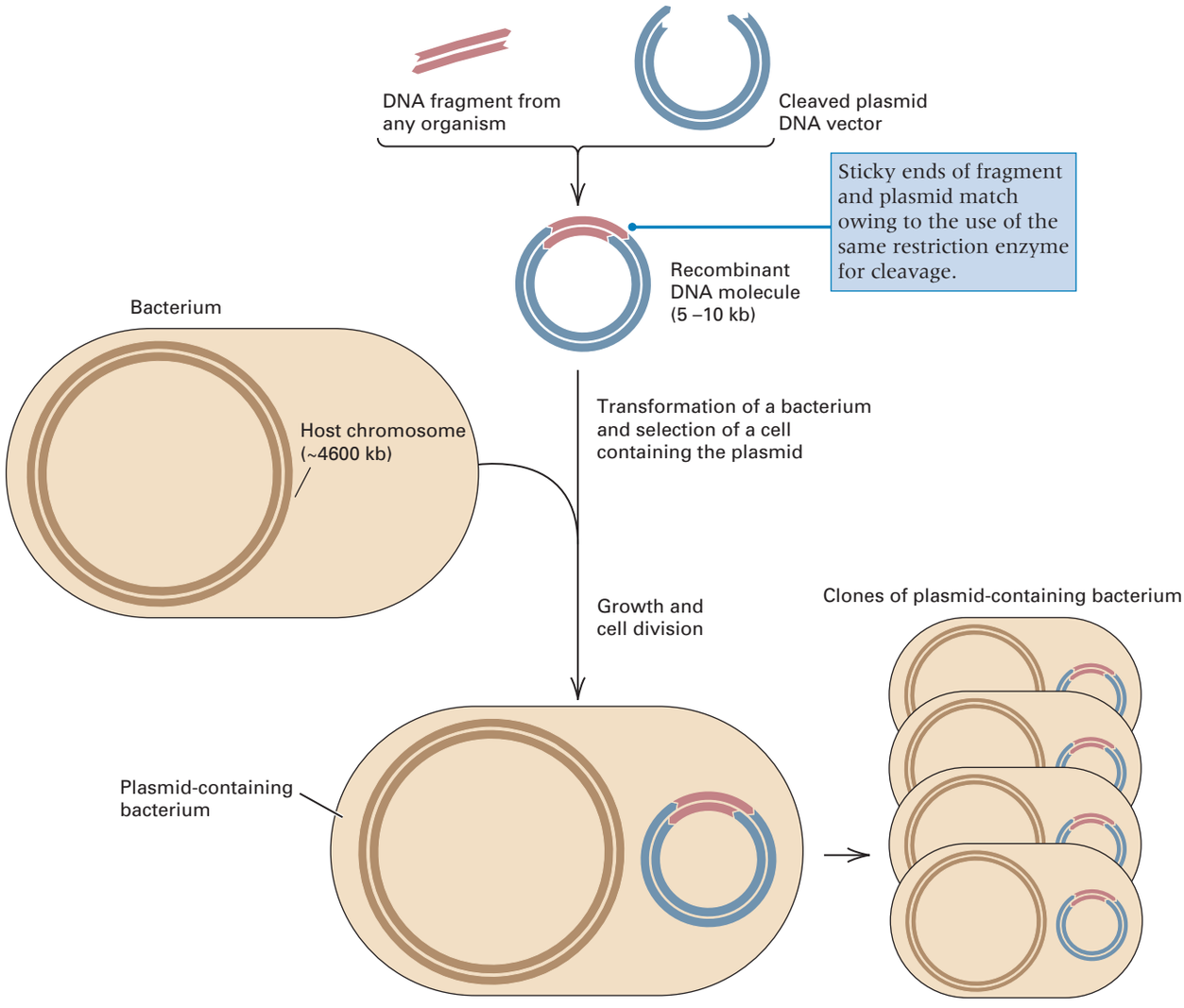


FIGURE 10.4 An example of cloning. A fragment of DNA from any organism is joined to a cleaved plasmid. The recombinant plasmid is then used to transform a bacterial cell, where the recombinant plasmid is replicated and transmitted to the progeny bacteria. The bacterial host chromosome is not drawn to scale. It is typically about 1000 times larger than the plasmid.

recombinant molecule is introduced into a cell by means of DNA transformation (FIGURE 10.4). Inside the cell, the recombinant molecule is replicated as the cell replicates its own DNA, and as the cell divides, the recombinant molecule is transmitted to the progeny cells. When a transformant containing the recombinant molecule has been isolated, the DNA fragment linked to the vector is said to be **cloned**. A **vector** is therefore a DNA molecule into which another DNA fragment can be cloned; it is a carrier for recombinant DNA. In the following sections, several types of vectors are described.

A vector is a carrier for recombinant DNA.

The most generally useful vectors have three properties:

1. The vector DNA can be introduced into a host cell relatively easily.
2. The vector contains a replication origin and so can replicate inside the host cell.
3. Cells containing the vector can usually be selected by a straightforward assay, most conveniently by allowing growth of the host cell on a solid selective medium.

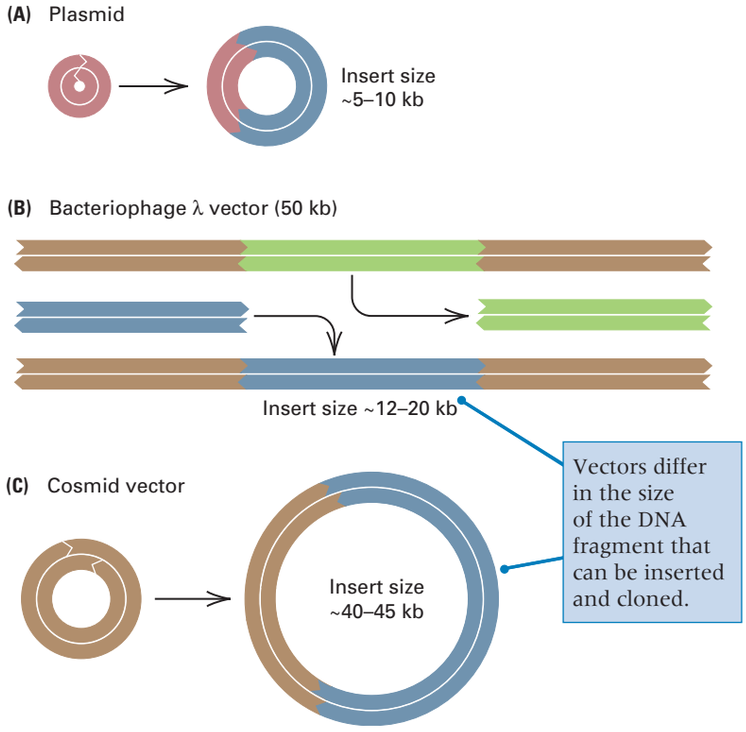


FIGURE 10.5 Common cloning vectors for use with *E. coli*, not drawn to scale. (A) Plasmid vectors are ideal for cloning relatively small fragments of DNA. (B) Bacteriophage λ vectors contain convenient restriction sites for removing the middle section of the phage and replacing it with the DNA of interest. (C) Cosmid vectors are useful for cloning DNA fragments up to about 40 kb; they can replicate as plasmids but contain the cohesive ends of phage λ and so can be packaged in phage particles.

The vectors most commonly used in *E. coli* are plasmids and derivatives of the bacteriophages λ and M13. Many other plasmids and viruses also have been developed for cloning into cells of animals, plants, and other bacteria. Recombinant DNA can be detected in host cells by means of genetic markers or phenotypic characteristics that are evident in the appearance of colonies or plaques. Plasmid and phage DNA can be introduced into cells by transformation, in which cells gain the ability to take up free DNA by exposure to a calcium chloride solution. Recombinant DNA can also be introduced into cells by a kind of electrophoretic procedure called *electroporation*. After introduction of the DNA, the cells that contain the recombinant DNA are plated on a solid medium. If the added DNA is a plasmid, colonies consisting of bacterial cells that contain the recombinant plasmid are formed, and the transformants can usually be detected by the phenotype that the plasmid confers on the host cell. For example, plasmid vectors typically include one or more genes for resistance to antibiotics, and plating the transformed cells on a selective medium with antibiotic prevents all but the plasmid-containing cells from growing (Section 8.2). Alternatively, if the vector is phage DNA, the infected cells are plated in the usual way to yield plaques. Variants of these procedures are used to transform animal or plant cells with suitable vectors, but the technical details may differ considerably.

Three types of vectors commonly used for cloning into *E. coli* are illustrated in FIGURE 10.5. Plasmids (part A) are most convenient for cloning relatively small DNA fragments (5 to 10 kb). Somewhat larger fragments can be cloned with bacteriophage λ (part B). The wildtype phage is approximately 50 kb in length, but the central portion of the genome is not essential for lytic growth and can be removed and replaced with donor DNA. After the donor DNA has been ligated in place, the recombinant DNA is packaged into mature phage *in vitro*, and the phage is used to infect bacterial cells. However, to be packaged into a phage head, the recombinant DNA must be neither too large nor too small, which means that the donor DNA must be roughly the same size as the portion of the λ genome that was removed. Most λ cloning vectors accept inserts ranging in size from 12 to 20 kb. Still larger DNA fragments can be inserted into cosmid vectors (part C). These vectors can exist as plasmids, but they also contain the complementary overhanging single-stranded ends of phage λ, which enables them to be packaged into mature phages. The size limitation on cosmid inserts usually ranges from 40 to 45 kb.

Specialized vectors can carry very large DNA fragments.

Large DNA molecules can be cloned intact in bacterial cells with the use of specialized vectors that can accept large inserts. The vectors that can accept large DNA fragments are called *artificial chromosomes*. Among the most widely used are **bacterial artificial chromosomes (BACs)**. The BAC vector (FIGURE 10.6) is based on the F factor of *E. coli*, which was discussed in Chapter 7 in the context of its role in conjugation. The essential functions included in the 6.8-kb vector are genes for replication (*repE* and *oriS*), for regulating copy number (*parA* and *parB*), and for resistance to the antibiotic chloramphenicol. BAC vectors with inserts greater than 300 kb can be maintained. DNA fragments in the appropriate size

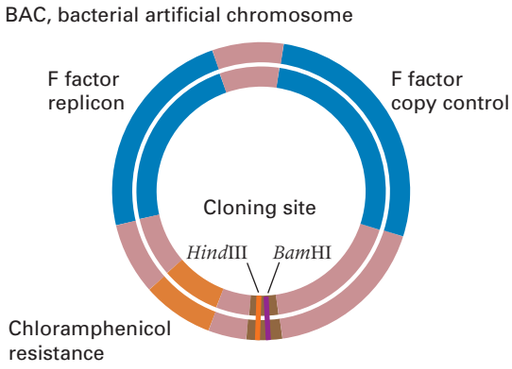


FIGURE 10.6 Bacterial artificial chromosomes are based on a vector that contains the F factor as well as genes for regulating copy number.

range can be produced by breaking larger molecules into fragments of the desired size by physical means, by treatment with restriction enzymes that have infrequent cleavage sites (for example, enzymes such as *Not*I and *Sfi*I), or by treatment with ordinary restriction enzymes under conditions in which only a fraction of the restriction sites are cleaved (*partial digestion*). Cloning the large molecules consists of mixing the large fragments of source DNA with the vector, ligation with DNA ligase, introduction of the recombinant molecules into cells of the host, and selection for the clones of interest. These methods are generally similar to those used for the production of recombinant molecules containing smaller inserts of cloned DNA.

BAC clones play a special role in genomic sequencing. Most genomes are sequenced in the form of many small single-stranded fragments, typically 500 to 1000 bp in length, cloned from random positions. This sequencing strategy is known as *shotgun sequencing*, and it needs to be carried out at high redundancy in order that the short sequence fragments can be recognized by their overlaps and assembled into a finished sequence. Each genomic region covered by overlapping clones is called a *contig*. Typically, a genomic sequence contains many gaps that prevent the contigs from being assembled. BAC clones are important because the sequences at the extreme ends of the cloned fragments give long-range information that allows adjacent contigs to be recognized and assembled in the correct orientation.

Vector and target DNA fragments are joined with DNA ligase.

The circularization of restriction fragments that have terminal single-stranded regions with complementary bases is illustrated in Figure 10.1. Because a particular restriction enzyme produces fragments with *identical* sticky ends, without regard for the source of the DNA, fragments from DNA molecules isolated from two different organisms can be joined, as shown in FIGURE 10.7. In this example, the restriction enzyme *Eco*RI is used to digest DNA from any organism of interest and to cleave a bacterial plasmid that contains only one *Eco*RI restriction site. The donor DNA is digested into many fragments (one of which is shown) and the plasmid into a single linear fragment. When the donor fragment and the linearized plasmid are mixed, recombinant molecules can form by base pairing between the complementary single-stranded ends. At this point, the DNA is treated with DNA ligase to seal the joints, and the donor fragment becomes permanently joined in a combination that may never have existed before. The ability to join a donor DNA fragment of interest to a vector is the basis of recombinant DNA technology.

Joining sticky ends does not always produce a DNA sequence that has functional genes. For example, consider a linear DNA molecule that is cleaved into four fragments—A, B, C, and D—whose sequence in the original molecule was ABCD. Reassembly of the fragments can yield the original molecule, but because B and C have the same pair of sticky ends, molecules with the fragment arrangements ACBD and BADC can also form with the same probability as ABCD. Restriction fragments from the vector can also join together in the wrong order, but this potential problem can be eliminated by using a vector that has only one cleavage site for a particular restriction enzyme. Plasmids of this type are available (most have been created by genetic engineering). Many vectors contain unique sites for several different restriction enzymes, but generally only one enzyme is used at a time.

DNA molecules that lack sticky ends also can be joined. A direct method uses the DNA ligase made by *E. coli* phage T4. This enzyme differs from other DNA ligases in that it not only heals single-stranded breaks in double-stranded DNA but also can join molecules with blunt ends.

A recombinant cDNA contains the coding sequence of a eukaryotic gene.

Many genes in higher eukaryotes are very large. They can extend over hundreds of thousands of base pairs. Much of the length is made up of introns, which are excised from the mRNA in processing (Section 8.4). With such large genes, the length of the spliced mRNA is usually much less than the length of the gene. Even if the large DNA sequence

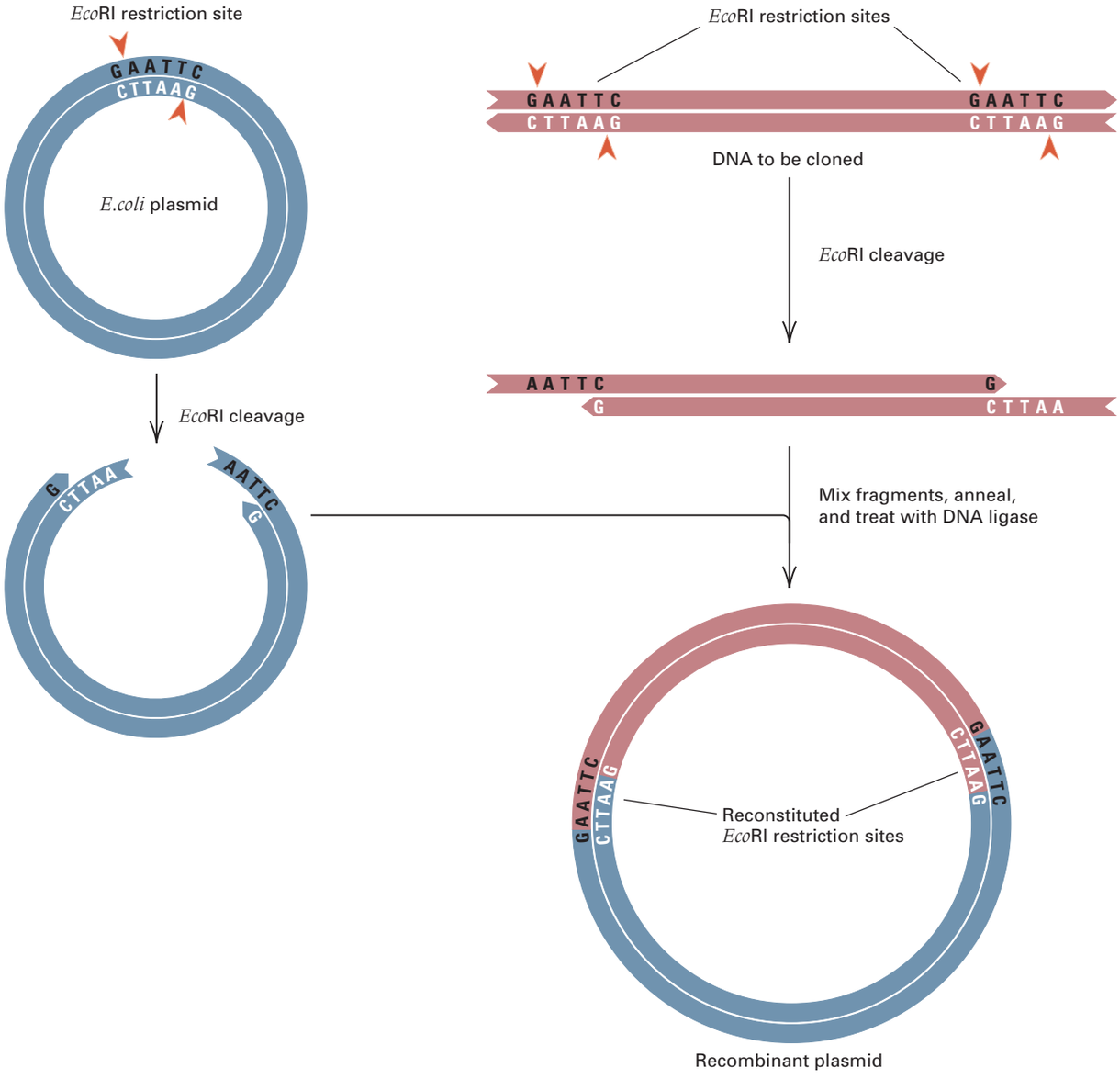


FIGURE 10.7 Construction of recombinant DNA plasmids containing fragments derived from a donor organism, by the use of a restriction enzyme (in this example *Eco*RI) and the joining of complementary (sticky) ends. Red arrowheads indicate cleavage sites.

were cloned, expression of the gene product in bacterial cells would be impossible because bacterial cells are not capable of RNA splicing. Therefore, when a gene is so large that it is difficult to clone and express directly, it would be desirable to clone the coding sequence present in the mRNA to determine the base sequence and study the polypeptide gene product. The method illustrated in FIGURE 10.8 makes possible the direct cloning of any eukaryotic coding sequence from cells in which the mRNA is present.

Cloning from mRNA molecules depends on an unusual polymerase, **reverse transcriptase**, which can use a single-stranded RNA molecule as a template and synthesize a complementary strand of DNA called **complementary DNA**, or **cDNA**. Like other DNA polymerases, reverse transcriptase requires a primer. The stretch of A nucleotides usually found at the 3′ end of eukaryotic mRNA serves as a convenient priming site, because the primer can be an oligonucleotide consisting of poly-T (Figure 10.8). Like any other single-stranded DNA molecule, the single strand of DNA produced from the RNA template can fold back upon itself at the extreme 3′ end to form a "hairpin" structure that includes a very short double-stranded region consisting of a few base pairs. The 3′ end of the hairpin serves as a primer for second-strand synthesis. The second strand can be synthesized either by DNA polymerase

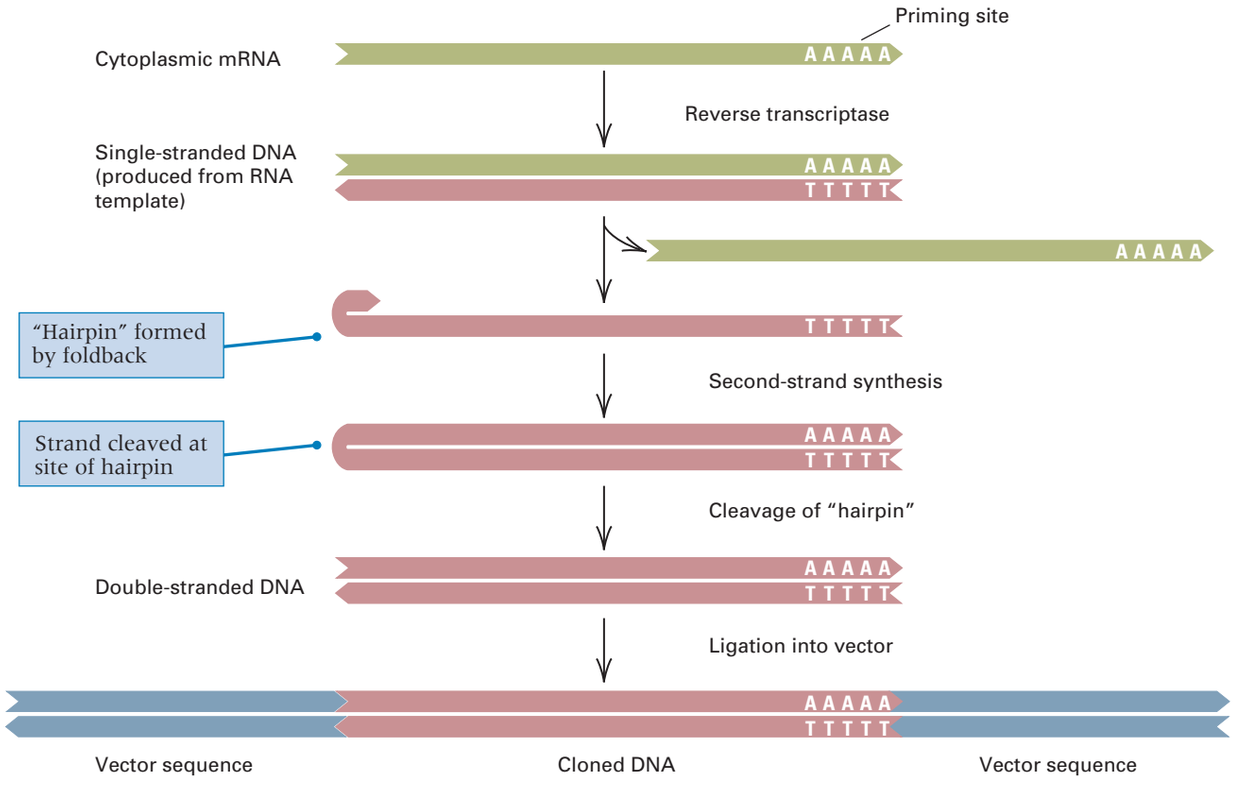


FIGURE 10.8 Reverse transcriptase produces a single-stranded DNA complementary in sequence to a template RNA. In this example, a cytoplasmic mRNA is copied. As indicated here, most eukaryotic mRNA molecules have a tract of consecutive A nucleotides at the 3′ end, which serves as a convenient priming site. After the single-stranded DNA is produced, a foldback at the 3′ end forms a hairpin that serves as a primer for second-strand synthesis. After the hairpin is cleaved, the resulting double-stranded DNA can be ligated into an appropriate vector either immediately or after PCR amplification. The resulting clone contains the entire coding region for the protein product of the gene.

or by reverse transcriptase itself. Reverse transcriptase is the source of the second strand in RNA-based viruses that use reverse transcriptase, such as the human immunodeficiency virus (HIV). Conversion into a conventional double-stranded DNA molecule is achieved by cleavage of the hairpin by a nuclease.

In the reverse transcription of an mRNA molecule, the resulting full-length cDNA contains an uninterrupted coding sequence for the protein of interest. As we saw in Chapter 8, eukaryotic genes often contain DNA sequences, called *introns*, that are initially transcribed into RNA but are removed in the production of the mature mRNA. Because the introns are absent from the mRNA, the cDNA sequence is not identical with that in the genome of the original donor organism. However, if the purpose of forming the recombinant DNA molecule is to identify the coding sequence or to synthesize the gene product in a bacterial cell, then cDNA formed from processed mRNA is the material of choice for cloning. The joining of cDNA to a vector can be accomplished by available procedures for joining blunt-ended molecules (Figure 10.8).

Some specialized animal cells make only one protein, or a very small number of proteins, in large amounts. In these cells, the cytoplasm contains a great abundance of specific mRNA molecules, which constitute a large fraction of the total mRNA synthesized. An example is the mRNA for globin, which is highly abundant in reticulocytes while they are producing hemoglobin. The cDNA produced from purified mRNA from these cells is greatly enriched for the globin cDNA. Genes that are not highly expressed are represented by mRNA molecules whose abundance ranges from low to exceedingly low. The cDNA molecules produced from such rare RNAs will also be rare. The efficiency of cloning rare cDNA molecules can be markedly increased by PCR amplification prior to ligation into the vector. The only limitation on the procedure is the requirement that enough DNA sequence be known at both ends of the cDNA for appropriate oligonucleotide primers to be designed. PCR amplification of the cDNA produced by reverse transcriptase is called **reverse transcriptase PCR (RT-PCR)**. The resulting amplified molecules contain the coding sequence of the gene of interest with very little contaminating DNA.

Loss of β-galactosidase activity is often used to detect recombinant vectors.

When a vector is cleaved by a restriction enzyme and renatured in the presence of many different

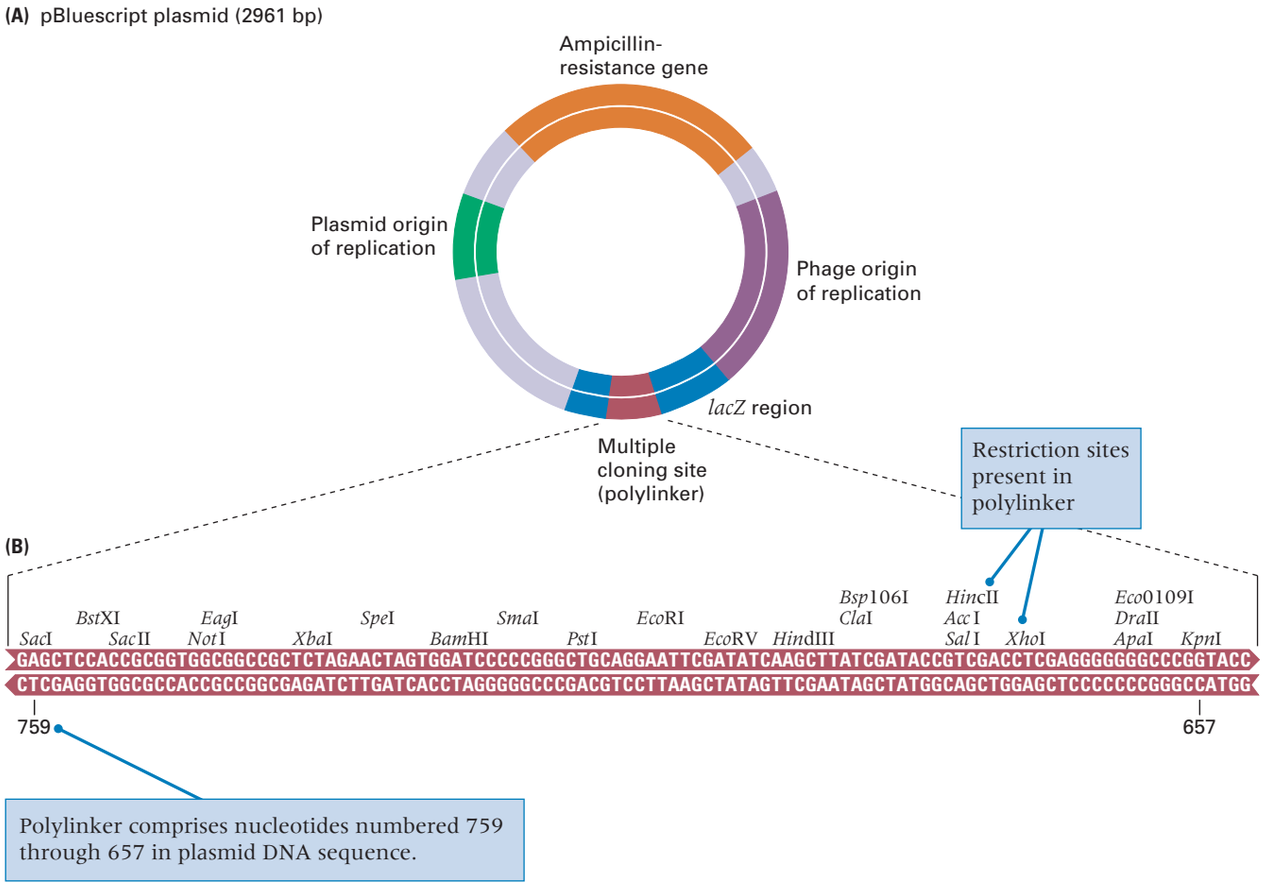


FIGURE 10.9 (A) Diagram of the cloning vector pBluescript II. It contains a plasmid origin of replication, an ampicillin-resistance gene, a multiple cloning site (polylinker) within a fragment of the *lacZ* gene from *E. coli*, and a bacteriophage origin of replication. (B) Sequence of the multiple cloning site showing the unique restriction sites at which the vector can be opened for the insertion of DNA fragments. The numbers 657 and 759 refer to the position of the base pairs in the complete sequence of pBluescript. [Courtesy of Agilent Technologies, Inc., Stratagene Products Division.]

restriction fragments from a particular organism, many types of molecules result, including such examples as a self-joined circular vector that has not acquired any fragments, a vector containing one or more fragments, and a molecule consisting only of many joined fragments. To facilitate the isolation of a vector containing a particular gene, some means is needed to ensure (1) that the vector does indeed possess an inserted DNA fragment, and (2) that the fragment is in fact the DNA segment of interest. This section describes several useful procedures for detecting the correct products.

In the use of transformation to introduce recombinant plasmids into bacterial cells, the initial goal is to isolate bacteria that contain the plasmid from a mixture of plasmid-free and plasmid-containing cells. A common procedure is to use a plasmid that possesses an antibiotic-resistance marker and to grow the transformed bacteria on a medium that contains the antibiotic: Only cells that contain plasmid can form a colony. An example of a cloning vector is the pBluescript plasmid illustrated in FIGURE 10.9, part A. The entire plasmid is 2961 base pairs. Different regions contribute to its utility as a cloning vector.

- The plasmid origin of replication is derived from the *E. coli* plasmid ColE1. The ColE1 is a high-copy-number plasmid, and its origin of replication enables pBluescript and its recombinant derivatives to exist in approximately 300 copies per cell.
- The ampicillin-resistance gene allows for selection of transformed cells in medium containing ampicillin.
- The cloning site is called a *multiple cloning site* (MCS), or *polylinker*, because it contains unique cleavage sites for many different restriction enzymes and enables many types of restriction fragments to be inserted. In pBluescript, the MCS is a 108-bp sequence that contains cloning sites for 23 different restriction enzymes (Figure 10.9, part B).
- The detection of recombinant plasmids is by means of a region containing the *lacZ* gene from *E. coli*, shown in blue in Figure 10.9, part A. The basis of the selection is illustrated in FIGURE 10.10.

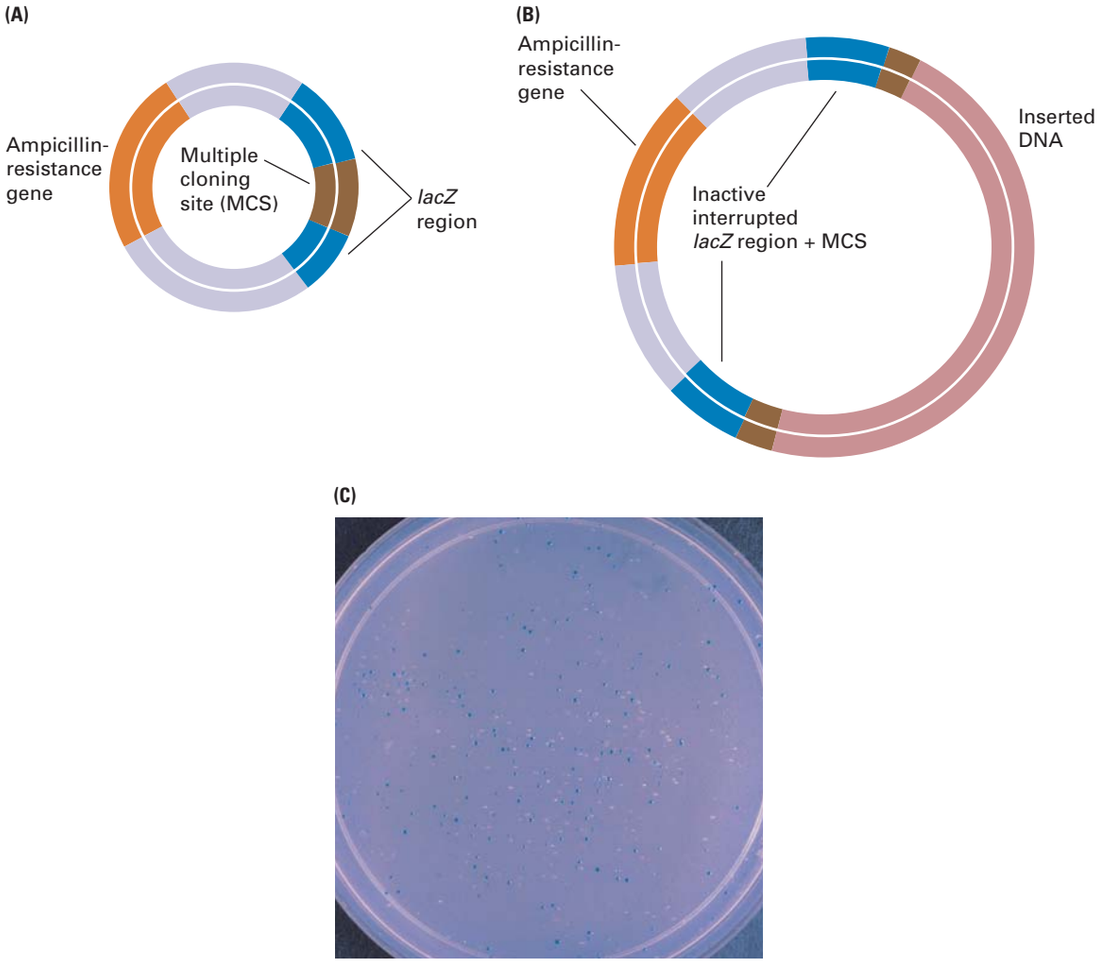
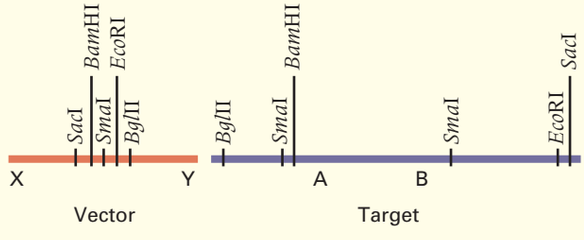


FIGURE 10.10 Detection of recombinant plasmids through insertional inactivation of a fragment of the *lacZ* gene from *E. coli*. (A) Nonrecombinant plasmid containing an uninterrupted *lacZ* region. The multiple cloning site (MCS) within the region (not drawn to scale) is sufficiently small that the plasmid still confers β-galactosidase activity. (B) Recombinant plasmid with donor DNA inserted into the multiple cloning site. This plasmid confers ampicillin resistance but not β-galactosidase activity, because the donor DNA interrupting the *lacZ* region is large enough to render the region nonfunctional. (C) Transformed bacterial colonies. Cells in the white colonies contain plasmids with inserts that disrupt the *lacZ* region; those in the blue colonies do not. [Courtesy of E. R. Lozovsky.]

A Moment to Think

Problem: Presence of a polylinker, or multiple cloning site (MCS), in a vector makes possible *directional cloning*. In this approach, the vector and the target sequence are both cleaved with the same two restriction enzymes, which are chosen so that their complementary sticky ends will ensure that the fragment of interest is inserted in a particular orientation in the vector. Consider, for example, the restriction sites in the vector MCS and the target sequence shown in the accompanying diagram. The vector sequence X to the left of the MCS is a promoter sequence, and the vector sequence to the right of the MCS is a transcriptional terminator. The target sequence is a protein-coding region that, in order to be expressed, must be oriented with A adjacent to, and to the right of, the promoter X. A geneticist therefore wants to create a recombinant molecule with the sequences oriented as sequence X–A–B–Y. The restriction enzymes *Sac*I, *Bam*HI, *Bgl*II, and *Eco*RI all produce sticky ends; *Sma*I produces blunt ends. **(a)** If the vector and target were both digested with *Sma*I and the resulting fragments ligated in such a way that each vector molecule was ligated with one and only one target fragment, in what orientation would the A–B fragment be ligated into the polylinker? **(b)** For directional cloning, what restriction enzyme (or enzymes) would you use to digest the vector and target so that after mixing of the fragments and ligation, the cloned DNA would have the sequence X–A–B–Y? (The answer can be found on page 345.)



When the *lacZ* region is interrupted by a fragment of DNA inserted into the MCS, the recombinant plasmid yields Lac^- cells because the interruption renders the lacZ region nonfunctional. Nonrecombinant plasmids do not contain a DNA fragment in the MCS and yield Lac^+ colonies. The Lac^+ and Lac^- phenotypes can be distinguished by color when the cells are grown on a special β-galactoside compound called X-gal, which releases a deep blue dye when cleaved. On medium containing X-gal, Lac^+ colonies contain nonrecombinant plasmids and are a deep blue, whereas Lac^- colonies contain recombinant plasmids and are white.

- The bacteriophage origin of replication is from the single-stranded DNA phage f1. When cells that contain a recombinant plasmid are infected with an f1 helper phage, the f1 origin enables a single strand of the inserted fragment, starting with *lacZ*, to be packaged in progeny phage. This feature is very convenient because it yields single-stranded DNA for sequencing. The plasmid shown in part A of Figure 10.9 is the SK(+) variety. There is also an SK(−) variety in which the f1 origin is in the opposite orientation and packages the complementary DNA strand.

All good cloning vectors have an efficient origin of replication, at least one unique cloning site for the insertion of DNA fragments, and a second gene whose interruption by inserted DNA yields a phenotype indicative of a recombinant plasmid. Once a **library**, or large set of clones, has been obtained in a particular vector, the next problem is how to identify the particular recombinant clones that contain the gene of interest.

Recombinant clones are often identified by hybridization with labeled probe.

Once a *library* has been obtained in a particular vector, the next problem is how to identify the particular recombinant clones that contain the gene of interest. In situations in which DNA or RNA molecules containing sequences complementary to the gene are available, either can be labeled with a fluorescent tag or radioactivity and used as a *probe* in hybridization experiments to identify the clones containing the gene. The hybridization procedure is known as **colony hybridization** and is outlined in FIGURE 10.11. Colonies to be tested are transferred from a solid medium onto a nitrocellulose or nylon filter by gently pressing the filter onto the surface. A part of each colony remains on the agar medium, which constitutes the reference plate. The filter is treated with sodium hydroxide (NaOH), which simultaneously breaks open the cells

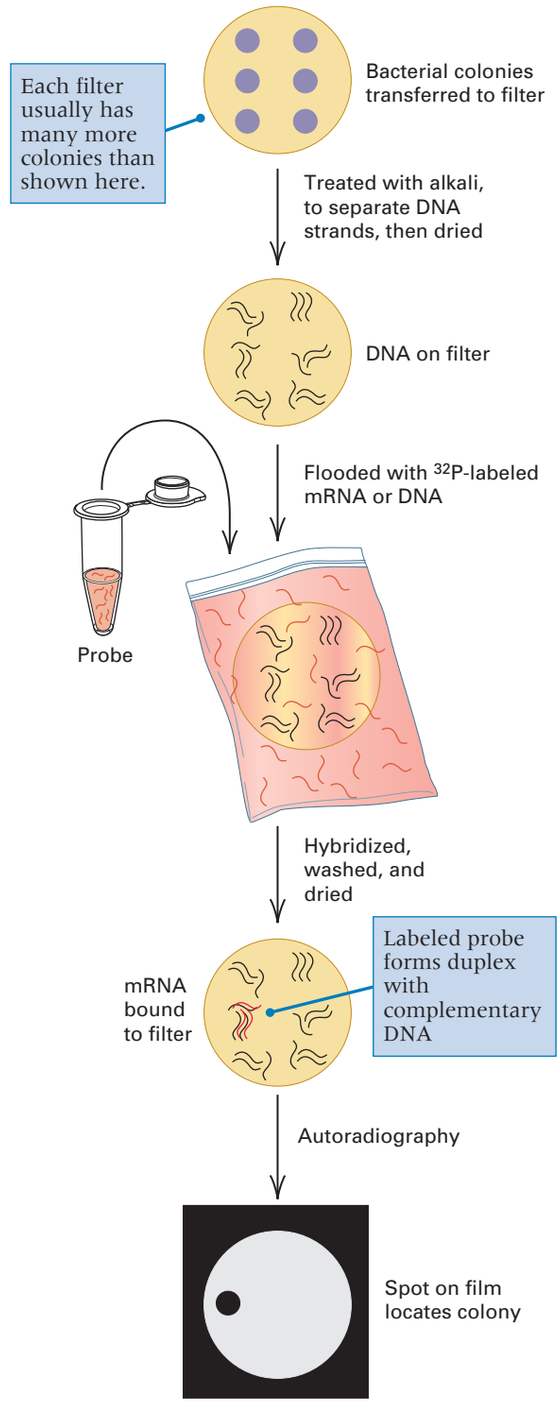


FIGURE 10.11 Colony hybridization.

and separates (*denatures*) the duplex DNA into single strands. The filter is then saturated with labeled probe complementary in base sequence to the gene being sought, and the DNA strands are allowed to form duplex molecules again (*renatured*). After washing to remove unbound probe, the positions of the bound

probe identify the desired colonies. For example, with radioactively labeled probe, the desired colonies are located by means of autoradiography. A similar assay is done with phage vectors, but in this case plaques rather than colonies are lifted onto the filters.

If transformed cells can synthesize the protein product of a cloned gene or cDNA, then immunological techniques may also allow the protein-producing colony to be identified. In one method, the colonies are transferred as in colony hybridization, and the transferred copies are exposed to a labeled antibody directed against the particular protein. Colonies to which the antibody adheres are those that contain the gene of interest.

10.2 A genomic sequence is like a book without an index, and identifying genes and their functions is a major challenge.

In 1985 the idea surfaced that it would be useful to know the complete sequence of all three billion nucleotides in the human genome. This seemed an outlandish idea at the time, because the cost of DNA sequencing was about one dollar per base pair. But proponents of the idea argued that launching such a program would provide incentives for technology development, and sequencing costs would fall. The Human Genome Project was formally inaugurated in 1990, and by the time the human sequence was completed in 2003, sequencing costs had indeed fallen—to about one penny per base pair.

The goals of the Human Genome Project included sequencing not only the human genome but also the genomes of certain key model organisms used in genetic research because of their demonstrated utility in the discovery of gene function. Vast amounts of data would be generated, which would have to be stored and made accessible, and new methods for analyzing such sequences would have to be developed. The information would need to be made available for drug development and other purposes, and ethical issues such as the privacy of one's genetic information had to be dealt with.

The Human Genome Project was a great success, but it will require many years, probably decades, before genome function and regulation are understood in detail. Tools for understanding the human genome include methods for annotating its content, comparative genomics, transcriptional profiling, and studying protein expression, function, and interaction. These are some of the key approaches to genomics and proteomics, and they are discussed in the following sections.

The protein-coding potential of an organism is contained in its genome sequence.

Among the first genomes to be sequenced were those of viruses and bacteria because they are quite small and relatively simple in their layout of regulatory and protein-coding sequences. Small, compact genomes are usually easier to interpret than more complex genomes. The genome of the bacterium *Mycoplasma genitalium* was one of the first to be sequenced. It has the smallest genome of any known free-living organism, a circular DNA molecule 580 kb in length that includes only 471 genes. The organism is a parasite associated with ciliated epithelial cells of the genital and respiratory tracts of primates, including human beings. It belongs to a large group of bacteria (the mycoplasmas) that lack a cell wall and that parasitize a wide range of plant and animal hosts. Analysis of the *M. genitalium* genome has enabled scientists to identify what is probably a minimal set of genes necessary for a free-living cell. The cellular processes in which the gene products this organism participate are summarized in FIGURE 10.12. A substantial fraction of the genome is devoted to the synthesis of macromolecules such as DNA, RNA, and protein; another substantial fraction supports cellular processes and energy metabolism. There are very few genes for biosynthesis of small molecules. However, genes that encode proteins for salvaging and/or for transporting small molecules make up a significant fraction of the total, which underscores the fact that the bacterium is parasitic. The remaining genes are largely devoted to forming the cellular envelope and to helping the organism evade the immune system of the host.

Note in Figure 10.12 that one-third of the genes have no identified function. This finding is typical of genomic sequences. In many genomic sequences, including the human genome, the proportion of genes with no identified function exceeds 50%. Hence, even when the genes in a genome are correctly identified, there are numerous additional issues. Genomic sequencing, therefore, should be thought of as only the initial stage in the quest to understand the higher and more integrated levels of biological organization and function.

A genome sequence without annotation is meaningless.

A genome sequence is not self-explanatory. It is like a book printed in an alphabet of only four letters, without spaces or punctuation, and lacking an index. To be useful, any genomic sequence must be accompanied by **genome annotation**, which refers to explanatory notes that accompany the sequence.

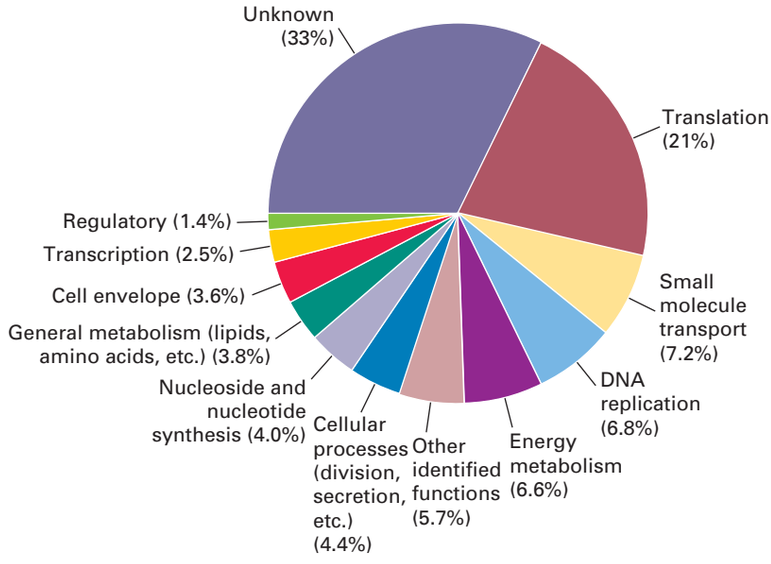


FIGURE 10.12 Genes in the genome of *Mycoplasma genitalium* classified by function. [Data from C. M. Fraser, et al., *Science* 270 (1995): 397–404.]

A genome annotation specifies functional elements, notably sequences in or near coding regions that delineate protein-coding exons and introns, as well as the upstream and downstream binding motifs that are targets of enhancer or silencer elements. Annotations also include sequences encoding functional RNAs such as tRNAs, small nuclear RNAs involved in splicing, and microRNAs. Annotations also identify sequences corresponding to transposable elements, and so forth.

Especially for large, complex genomes in which much of the DNA does not code for proteins, and in which most protein-coding exons are relatively small and interrupted by large introns, it is a daunting challenge to parse a genomic sequence into its protein-coding exons, to identify which protein-coding exons belong to the same gene, and to recognize the upstream and downstream regulatory regions that control gene expression. The annotation of genomic sequences at this level is one aspect of **computational genomics**, defined broadly as the use of computers in the interpretation and management of biological data.

Furthermore, especially in multicellular eukaryotes, even for genes whose functions can be assigned, it is not usually known when during the life cycle each gene is expressed, in which tissues it is expressed, or the presence, patterns, or tissue specificity of alternative splicing. Interactions among genes and gene products are also typically unknown. The greatest challenge is to understand how the genes in the genome function and are coordinately regulated to control development, metabolism, reproduction, behavior, and response to the environment.

Comparison among genomes is an aid to annotation.

In many cases, useful information can be gained by identifying genes with similar sequences in other organisms, but if the organisms diverged from a common ancestor too long ago, there is the problem of recognizing which sequences are sufficiently similar to be regarded as functionally equivalent. One way to get around his problem is to compare the genome sequences of groups of related species that have a graded series of divergence times. This approach is known as **comparative genomics**, which has become one of the most powerful strategies for identifying genetic elements in the human genome and those of model organisms.

The fruits of comparative genomics are exemplified in the genome sequences of 12 *Drosophila* species. FIGURE 10.13 summarizes the evolutionary relationships among the species and their approximate divergence times. The species are very diverse in their geographical origins, global distribution, morphology, behavior, feeding habits, and other phenotypes, yet they share a similar cellular physiology, developmental program, and life cycle. Their genomes show substantial differences in sequence (5 million years in the scale of Figure 10.13 corresponds to about one nucleotide difference per 10 nucleotide sites), and they have also undergone multiple gene rearrangements primarily due to inversions. The 12-genome comparison, therefore, reveals how conserved gene functions are maintained in spite of extensive changes in genome structure and sequence.

Comparative genomics derives its power from the distinctive evolutionary patterns, called *evolutionary signatures*, that different types of functional elements

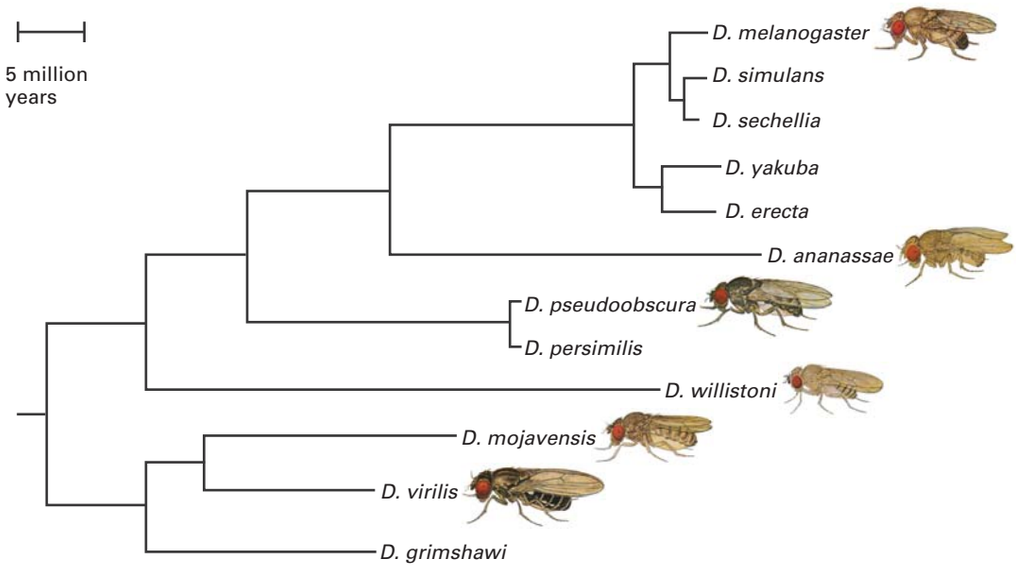


FIGURE 10.13 Evolutionary relationships among twelve *Drosophila* species whose genomes were sequenced for comparative genomics, scaled by their approximate divergence times. [Reproduced from J. T. Patterson, *Studies in the Genetics of Drosophila, Part III* and *Part IV.* University of Texas Publications (1943 and 1944). Used with permission of the School of Biological Sciences, University of Texas at Austin.]

exhibit. Some examples from the 12 *Drosophila* species are illustrated in FIGURE 10.14. Part B shows characteristic evolutionary signatures of protein-coding sequences. Note the pronounced triplet periodicity uninterrupted by stop codons. Many of the nucleotide differences between species are in the third codon position, and the variant codons often encode the same amino acid (green). Deletions, when they occur, remove a number of nucleotides that is a multiple of three (beige), which conserves the proper reading frame. Contrast this pattern with that observed in noncoding regions (part A). Here the nucleotide differences between species are not concentrated in a particular triplet phase (rust), triplets corresponding to chain-terminating (nonsense) codons come and go (yellow), and deletions are not constrained to a multiple of three nuclcotides (gray).

In addition, comparative genomics helps identify regulatory motifs that are the targets of enhancers and silencers. These are often difficult to recognize because they are relatively short, can be present on either DNA strand, and can change position within the gene promoter. The example in Figure 10.14C shows binding sites for the protein Mef-2. The consensus binding site has the sequence YTAWWWWTAR, where Y is any pyrimidine, R is any purine, and W means either A or T. The 12-species comparison shows the differing sequence and location of the Mef-2 biding site in one of its target genes. Some of the species have the binding site toward the 5′ end of the region shown, others have it near the 3′ end, and several species have a Mef-2 binding site at both locations.

RNA transcripts that form foldback secondary structures, such as tRNAs, rRNAs, and some snRNAs, have distinctive evolutionary signature of their own. In Figure 10.14D, for example, the matched parentheses show the conserved base pairs in the paired stem structures, but the identities of the paired bases often differ among species (paired nucleotides are color coded).

A Moment to Think

Answer to Problem: (a) *Sma*I produces blunt ends, and any blunt end can be ligated onto any other blunt end. Hence the *Sma*I fragment A—B can be ligated into the polylinker in either of two orientations. The resulting clones are expected to be X—A—B—Y and X—B—A—Y in equal frequency. **(b)** A restriction enzyme produces fragments whose ends are identical, so either end can be ligated onto a complementary sticky end. To force the orientation X—A—B—Y, one needs to cleave the vector and the target with two restriction enzymes that produce different sticky ends. The site nearest X in the vector must match the site nearest A in the target, and the site nearest Y in the vector must match the site nearest B in the vector. In this case, if the vector and target are both cleaved with *Bam*HI and *Eco*RI, the resulting clone is expected to have the orientation X—A—B—Y. You should be able to convince yourself that no other combination of enzymes will work.

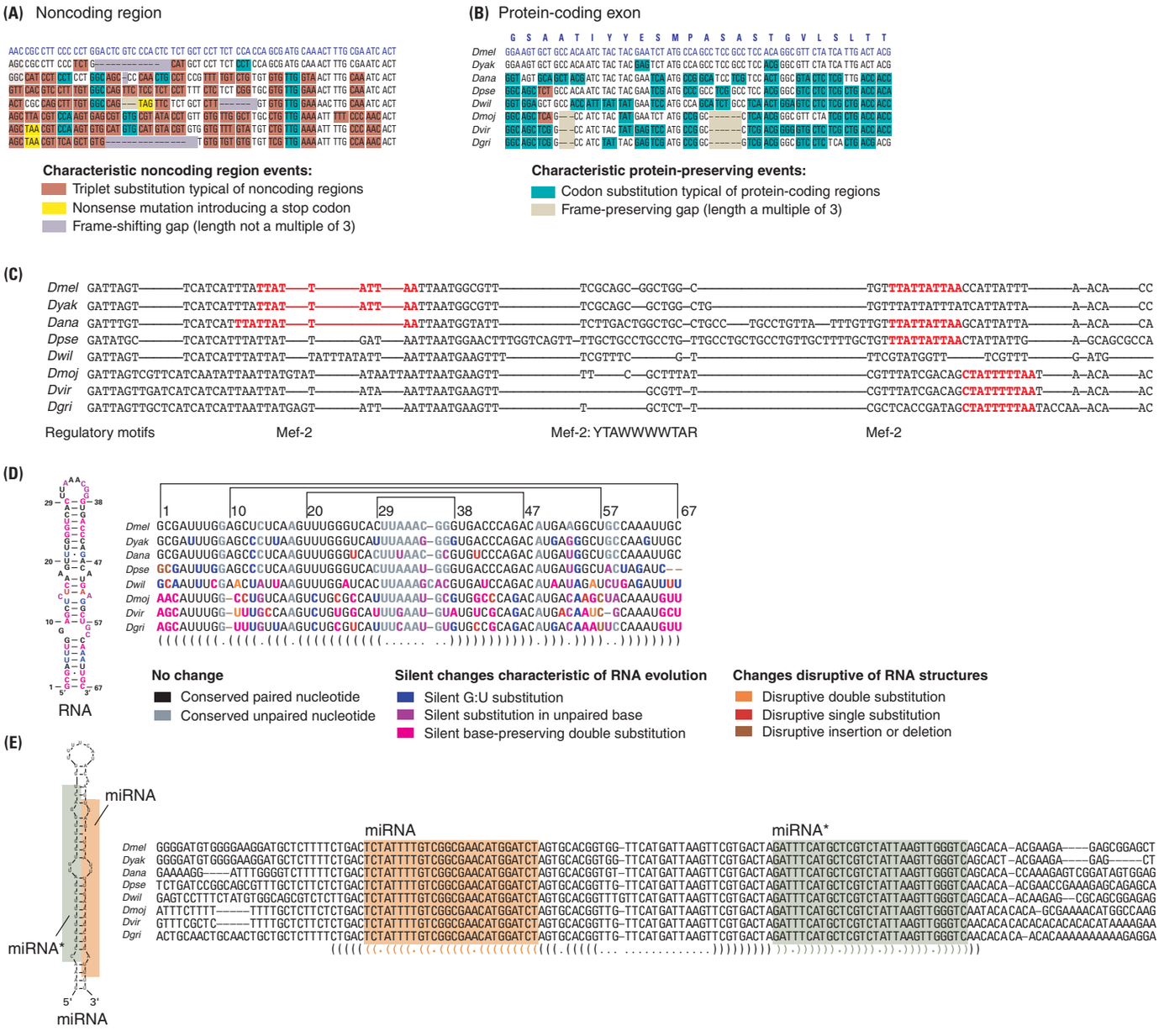


FIGURE 10.14 Evolutionary signatures observed among the twelve *Drosophila* genomes in regions coding for (A) noncoding RNA, (B) protein, (C) transcription-factor binding sites, (D) a stem-loop secondary structure in RNA, and (E) microRNA. [Adapted from A. Stark, et al., *Nature* 450 (2007): 219–232.]

A nice example involves the nucleotides at positions 29 and 38, which in *D. melanogaster* constitute a C-G pair but in *D. yakuba* constitute a U-G pair (U pairs with G as well as with A in double-stranded RNA).

MicroRNAs, important for their regulatory functions in the RNAi pathways, show yet another type of evolutionary signature (part E). In this case changes in the stem regions, even those that are complementary, are not well tolerated, but differences in the loop and other nonpaired regions are found.

The 12-genomes comparisons were instrumental in correctly annotating hundreds of protein-coding

genes in the *D. melanogaster* sequence, predicting the secondary structures of many noncoding RNAs with some likely to be involved in translational regulation, showing that some microRNA genes have multiple functional products that increase their regulatory repertoire, and revealing a network of pretranscriptional and posttranscriptional miRNA regulatory targets. The utility of a graded series of divergence times was also validated by the observation that the optimal divergence time for identifying evolutionary signatures depends on the length of the functional element. Longer functional elements are most easily recognized in closely related species, whereas shorter ones are most efficiently identified in more distantly related species.

10.3 Genomics and proteomics reveal genome-wide patterns of gene expression and networks of protein interactions.

Genomic sequencing has made possible a new approach to genetics called **functional genomics**, which focuses on genome-wide patterns of gene expression and the mechanisms by which gene expression is coordinated. As changes take place in the cellular environment—for example, through development, aging, or changes in the external conditions—the patterns of gene expression change also. But genes are usually deployed in sets, not individually. As the level of expression of one coordinated set is decreased, the level of expression of a different coordinated set may be increased. How can one study tens of thousands of genes all at the same time?

DNA microarrays are used to estimate the relative level of gene expression of each gene in the genome.

The study of genome-wide patterns of gene expression became feasible with the development of the **DNA microarray** (or *chip*), a flat surface about the size of a postage stamp on which 10,000 to 100,000 distinct spots are present, each containing a different immobilized DNA sequence suitable for hybridization with DNA or RNA isolated from cells growing under different conditions, from cells not exposed or cells exposed to a drug or toxic chemical, from different stages of development, or from different types or stages of a disease such as cancer. Two types of DNA chips are presently in use:

- A chip arrayed with oligonucleotides synthesized directly on the chip, one nucleotide at a time, by automated procedures; these chips typically have hundreds of thousands of spots per array.
- A chip arrayed with denatured, double-stranded DNA sequences of 500 to 5000 bp, in which the spots, each about a millionth of a drop in volume, are deposited by capillary action from miniaturized fountain-pen-like devices mounted on the movable head of a flatbed robotic workstation; these chips typically have tens of thousands of spots per array.

FIGURE 10.15 shows one method by which DNA chips are used to assay the genome-wide levels of gene expression in an experimental sample relative to a control. At the upper right are shown six adjacent spots in the microarray, each of which contains a DNA sequence that serves as a probe for a different gene, *A* through *F*. At the left is shown the experimental protocol. Messenger RNA is first extracted from both the experimental and the control samples. This material is then subjected to one or more rounds of reverse transcription, as described in Section 10.1. In the experimental material (sample 2), the primer for reverse transcription includes a red fluorescent label; and in the control material (sample 1), the primer includes a green fluorescent label. When a sufficient quantity of labeled DNA strands have accumulated, the fluorescent samples are mixed and hybridized with the DNA chip.

The result of hybridization is shown in the middle part of Figure 10.15. Because the samples are mixed, the hybridization is competitive, and therefore the density of red or green strands bound to the DNA chip is proportional to the concentration of red or green molecules in the mixture. Genes that are overexpressed in sample 2 relative to sample 1 will have more red strands hybridized to the spot, whereas those that are underexpressed in sample 2 relative to sample 1 will have more green strands hybridized to the spot.

After hybridization, the DNA chip is placed in a confocal fluorescence scanner that scans each *pixel* (the smallest discrete element in a visual image) first to record the intensity of one fluorescent label and then again to record the intensity of the other fluorescent label. These signals are synthesized to

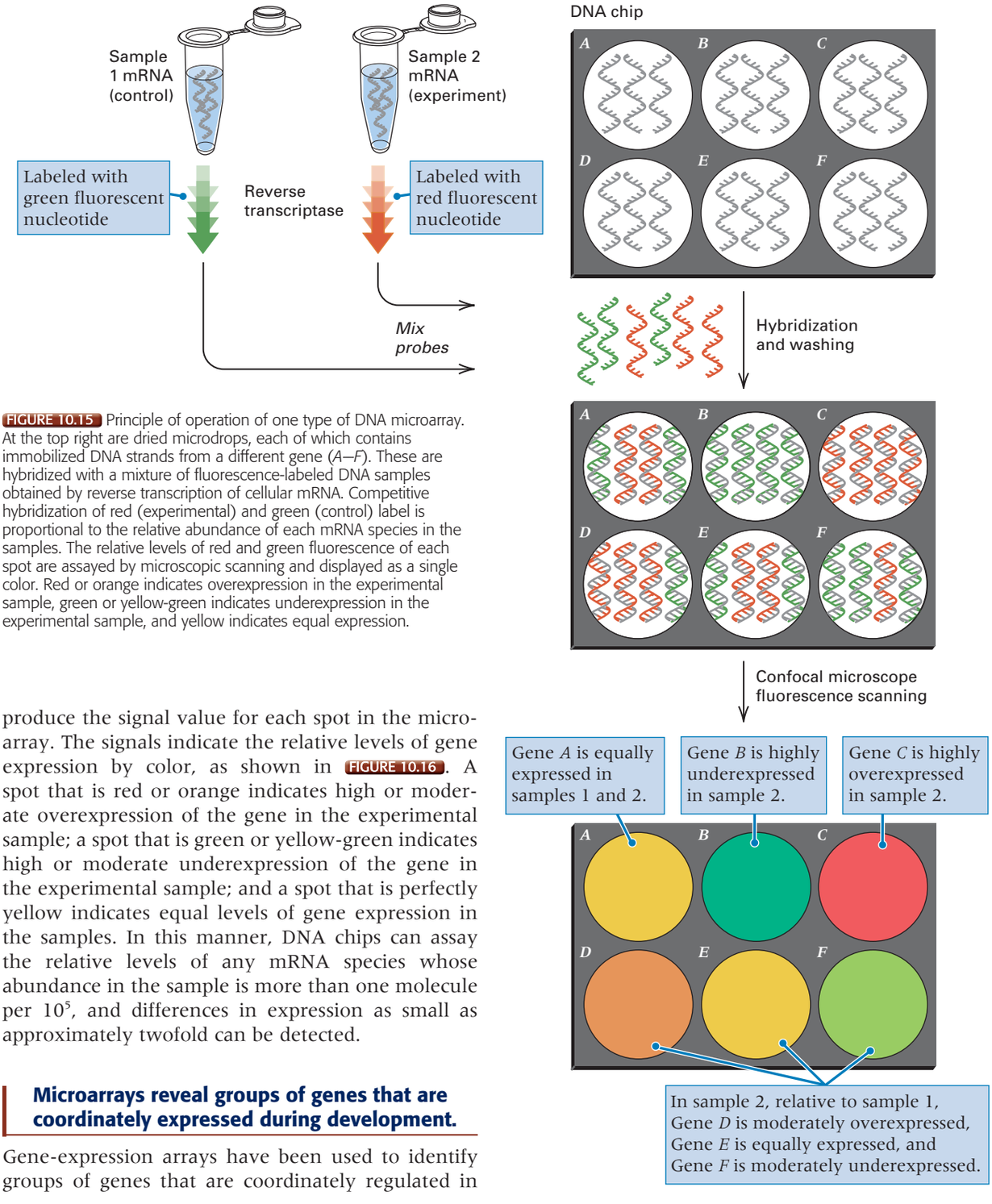


FIGURE 10.15 Principle of operation of one type of DNA microarray. At the top right are dried microdrops, each of which contains immobilized DNA strands from a different gene (*A*–*F*). These are hybridized with a mixture of fluorescence-labeled DNA samples obtained by reverse transcription of cellular mRNA. Competitive hybridization of red (experimental) and green (control) label is proportional to the relative abundance of each mRNA species in the samples. The relative levels of red and green fluorescence of each spot are assayed by microscopic scanning and displayed as a single color. Red or orange indicates overexpression in the experimental sample, green or yellow-green indicates underexpression in the experimental sample, and yellow indicates equal expression.

produce the signal value for each spot in the microarray. The signals indicate the relative levels of gene expression by color, as shown in FIGURE 10.16. A spot that is red or orange indicates high or moderate overexpression of the gene in the experimental sample; a spot that is green or yellow-green indicates high or moderate underexpression of the gene in the experimental sample; and a spot that is perfectly yellow indicates equal levels of gene expression in the samples. In this manner, DNA chips can assay the relative levels of any mRNA species whose abundance in the sample is more than one molecule per 10^5, and differences in expression as small as approximately twofold can be detected.

Microarrays reveal groups of genes that are coordinately expressed during development.

Gene-expression arrays have been used to identify groups of genes that are coordinately regulated in

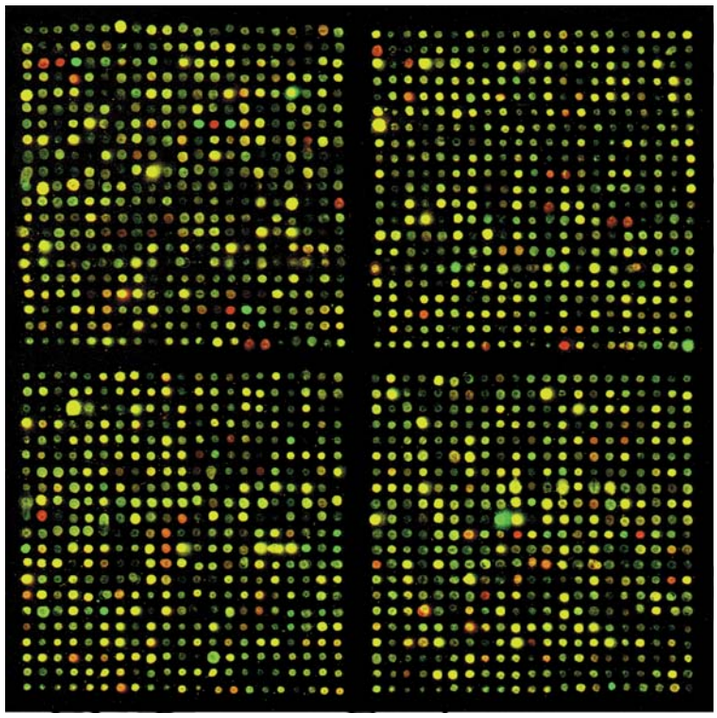


FIGURE 10.16 Small part of a yeast DNA chip showing 1764 spots, each specific for hybridization with a different mRNA sequence. The color of each spot indicates the relative level of gene expression in experimental and control samples. The complete chip for all yeast open reading frames includes over 6200 spots. [Courtesy of Jeffrey P. Townsend, Yale University and Duccio Cavalieri, University of Florence.]

development. The example in **FIGURE 10.17** shows expression profiles for 20 groups of genes in the early stages of development in *C. elegans*. In these experiments, time in development was measured in minutes relative to the four-cell stage. Relative levels of gene expression are plotted on a logarithmic scale and, hence, the changes in relative transcript abundance are often two or three orders of magnitude. Over the time period examined, the embryo undergoes a transition from control through maternal transcripts present in the egg to those transcribed in the embryo itself, and includes the times during which most of the major cell fates are specified. The microarrays used in these experiments allowed detection of transcripts from almost 9000 open reading frames, and the plots include traces for approximately 2500 genes, about 80% of all those that showed significant changes in transcript abundance over the time interval shown.

Up to the four-cell stage of development the patterns of transcription are all quite stable and then begin to change rapidly. Development in the earliest stages is supported largely by maternally derived transcripts. Many of these are cleared rapidly as development proceeds, for example, the transcripts plotted for clusters of 141, 244, and 568 genes in the lower-right panels. Production of transcripts from embryonic cells is clearly induced, as evidenced by the patterns for clusters of 431 and 153 genes in the panels at the upper right. The curves showing the disappearance of the maternal transcripts and appearance of the embryonic transcripts intersect at about the time of gastrulation, indicating a somewhat earlier (mid-blastula) transition from maternal to embryonic control of development. Many of the gene transcription patterns are very complex, with a transient peak of expression suggesting that the transcript (though not necessarily the protein product) is needed for only a brief period

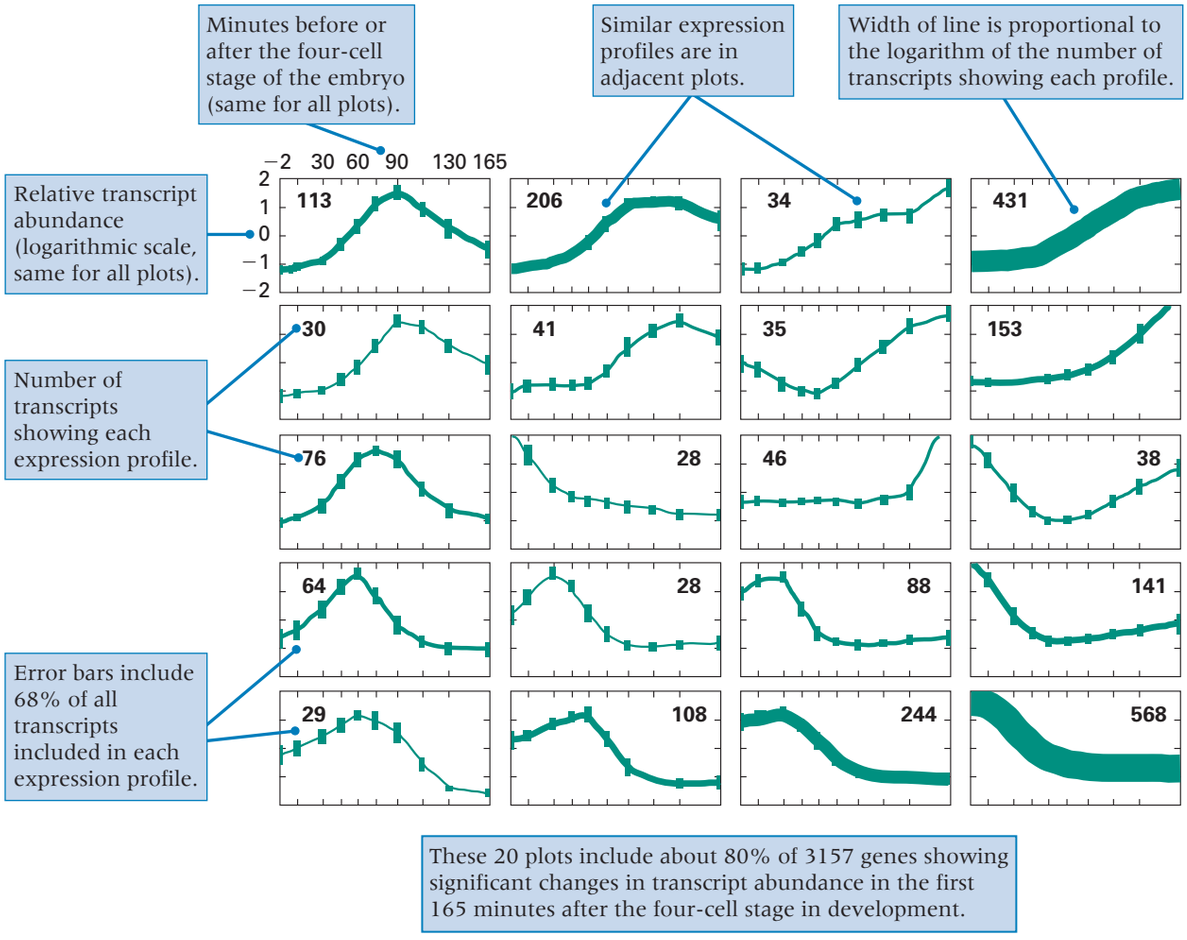


FIGURE 10.17 Patterns of transcriptional regulation of about 2500 genes during the first approximately 2.75 hours of development in *C. elegans*. Complete development requires about 14 hours. [Reproduced from L. R. Baugh, et al., *Development* 130 (2003): 889–900 (http://dev.biologists.org/cgi/content/abstract/130/5/889). Reproduced with permission of the Company of Biologists.]

in development. All five panels along the left-hand side of Figure 10.17 show this kind of pattern.

Although the transcriptional analysis in Figure 10.17 is a rather coarse, bird's-eye view of what takes place during development, the identification of groups of coordinately expressed genes is of considerable value in itself because it suggests that these genes may share common or overlapping *cis*-acting regulatory sequences that are controlled by common or overlapping sets of transcriptional activator proteins.

Yeast two-hybrid analysis reveals networks of protein interactions.

Biological processes can also be explored from the standpoint of proteomics by examining protein–protein interactions. The rationale for studying such interactions is that proteins that participate in related cellular processes often interact with one another; hence, knowing which proteins interact can provide clues to the possible function of otherwise anonymous proteins.

One method for identifying protein–protein interactions makes use of the GAL4 transcriptional activator protein in budding yeast discussed in Section 9.4. The GAL4 protein includes two separate domains or regions, both of which are necessary for transcriptional activation. One domain is the zinc-finger DNA-binding domain that binds with the target site in the promoter of the *GAL* genes that are activated, and the other domain is the transcriptional activation domain that makes contact with the transcriptional complex and actually triggers transcription. In the wildtype GAL4 protein, these domains are tethered together because they are parts of the same polypeptide chain.

The key to identifying protein–protein interactions through the use of GAL4 is that the coding regions for the separate domains can be taken apart and each fused to a coding region for a different protein. The strategy is shown in FIGURE 10.18 part A, where the GAL4 DNA-binding domain and the transcriptional activation domain are depicted as separate entities, each fused to a different polypeptide chain, shown in the vicinity of a *GAL* promoter. The promoter is attached to a **reporter gene** whose transcription can be detected by means of, for example, a color change in the colony, the production

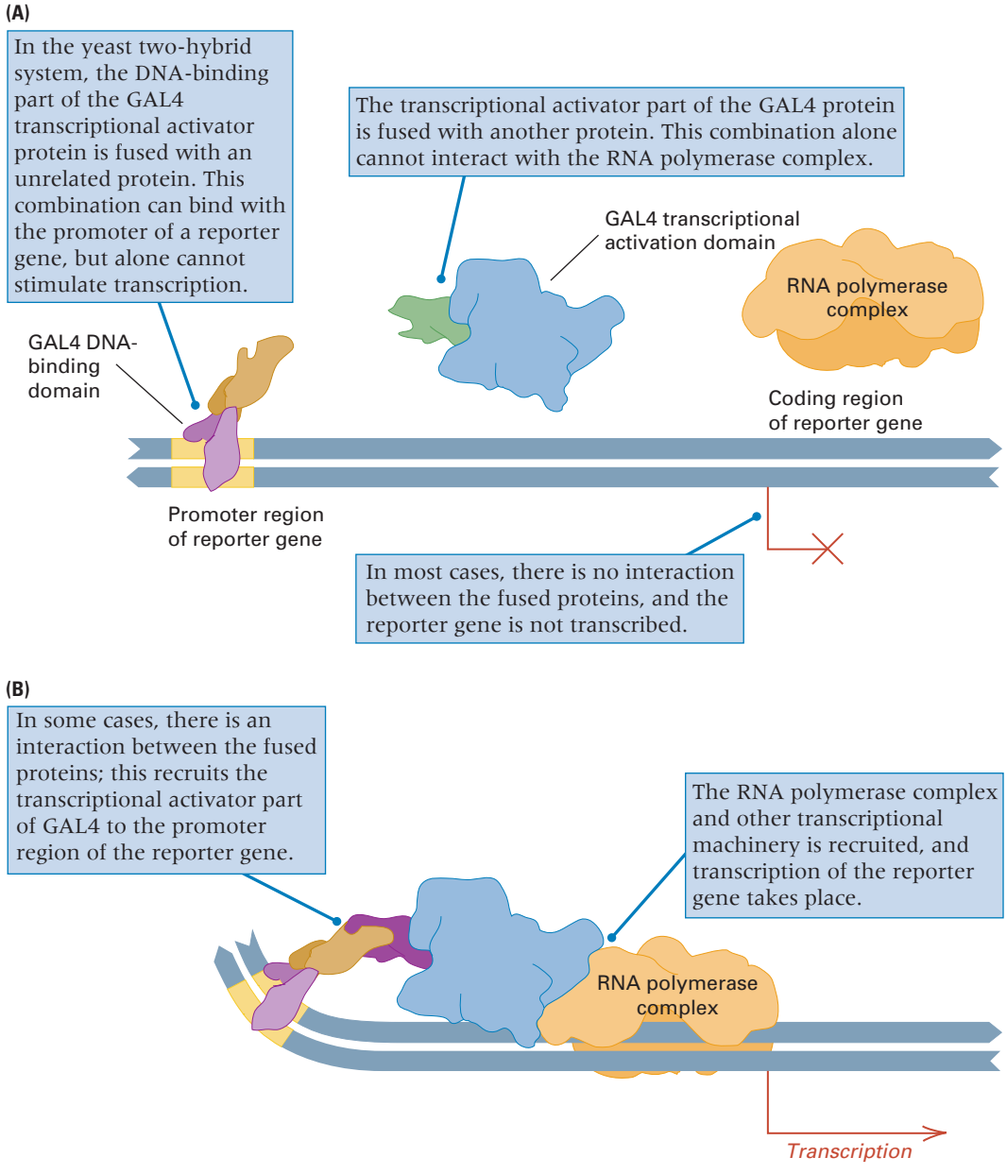


FIGURE 10.18 Two-hybrid analysis by means of the GAL4 protein. (A) When the proteins fused to the GAL4 domains do not interact, transcription of the reporter gene does not take place. (B) When the proteins do interact, the reporter gene is transcribed.

of a fluorescent protein, or the ability of the cells to grow in the presence of an antibiotic. The fused DNA-binding domain and the fused transcriptional activation domain are both hybrid proteins, and for this reason the test system is called a **two-hybrid analysis**. In part A, the proteins fused to the GAL4 domains do not interact within the nucleus. The DNA-binding domain therefore remains separated from the transcriptional activation domain, and transcription of the reporter genes does not occur.

Figure 10.18 part B shows a case in which the protein fused to the GAL4 domains do interact. In this case the DNA-binding domain and the transcriptional activation domain are brought into contact, and transcription of the reporter gene does take place. In this manner, transcription of the reporter gene in the two-hybrid analysis indicates that the proteins fused to the GAL4 domains undergo a physical interaction that brings the two hybrid proteins together.

An example of two-hybrid analysis is shown in FIGURE 10.19, which depicts a network of 318 protein–protein interactions among 329 nuclear proteins in yeast. The purpose of this analysis was to compare the observed network of interactions with random networks containing the same number of interactions but with the interacting partners chosen at random. One interesting property of the network in Figure 10.19 as well as of other protein networks, is that there are fewer than expected interactions between proteins that are already highly connected. In other words, proteins that are highly connected to other proteins through many interactions tend to be connected not to

other highly connected proteins, but to proteins with fewer connections. The systematic suppression of links between highly connected proteins has the effect of minimizing the extent to which random environmental or genetic perturbations in one part of the network spread to other parts of the network.

Two-hybrid analysis affords a powerful approach to discovering protein–protein interactions because it can be performed on a large scale, requires no protein purification, detects interactions that occur in living cells, and requires no information about the function of the proteins being tested. The method, however, does have some limitations. For example, the two-hybrid assay is qualitative, not quantitative, and so weak interactions cannot easily be distinguished from strong ones. The hybrid proteins are usually highly expressed to enhance the reliability of the assay, and so interactions can take place that would not take place at normal concentrations. The two-hybrid assay also requires that the protein–protein interactions take place in the nucleus, whereas some proteins may interact only in the environment of the cytoplasm. Finally, hybrid proteins may fold differently than native proteins, and the misfolded proteins may fail to interact when the native conformations do, or they may interact when the native conformations do not. The conclusion is that results from two-hybrid analyses need to be interpreted with care; nevertheless, the method has already yielded much valuable information.

10.4 Reverse genetics creates an organism with a designed mutation.

Mutation has traditionally provided the raw material needed for genetic analysis. The customary procedure has been to use a mutant phenotype to recognize a mutant gene and then to identify the wildtype allele and its normal function. This approach has proved highly successful, as evidenced by numerous examples throughout this book. But the approach also has its limitations. For example, it may prove difficult or impossible to isolate mutations in genes that duplicate the functions of other genes or that are essential for the viability of the organism.

Recombinant DNA technology has made possible another approach in genetic analysis in which wildtype genes are cloned, intentionally mutated in specific ways, and introduced back into the organism to study the phenotypic effects of the mutations. Because the position and molecular nature of each mutation are precisely defined, a very fine level of resolution is possible in determining the functions of particular regions of nucleotide sequence. This type of analysis has been applied to defining the promoter and enhancer sequences that are necessary for transcription, the sequences necessary for normal RNA splicing, particular amino acids that are essential for protein function, and many other problems. The procedure is often called **reverse genetics** because it reverses the usual flow of study: Instead of starting with a mutant phenotype and trying to identify the wildtype gene, reverse genetics starts by making a mutant gene and

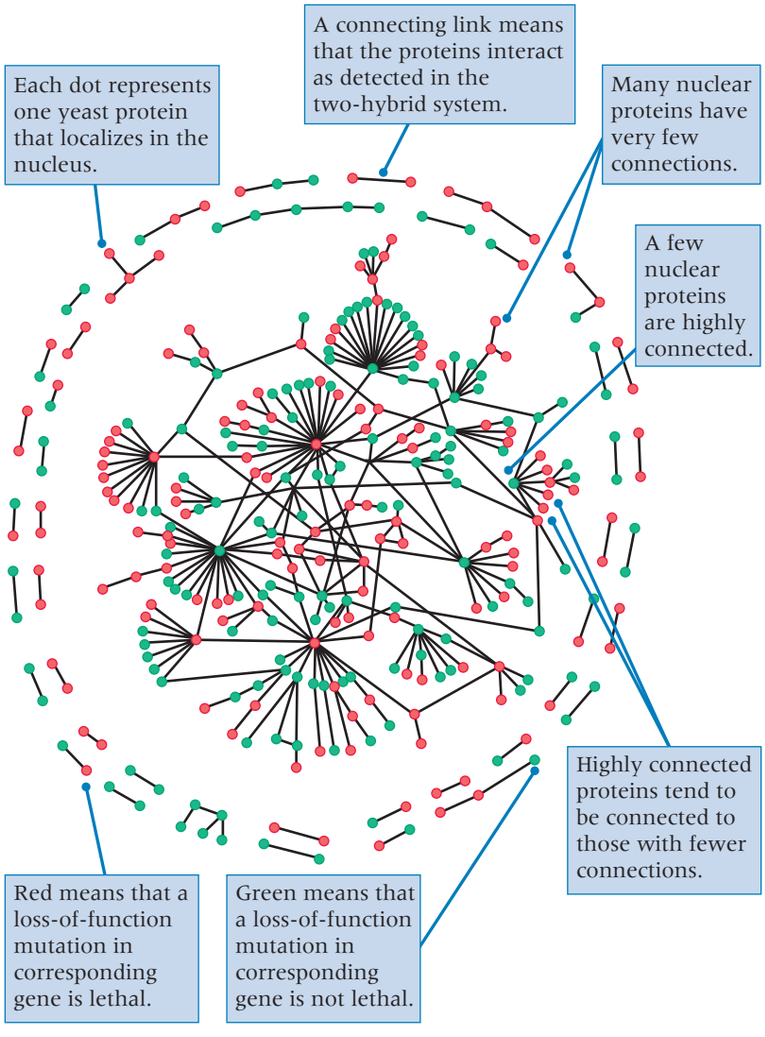


FIGURE 10.19 Physical interactions among nuclear proteins in yeast. Not shown are nuclear proteins that exhibit no interactions. [Reproduced from S. Maslov and K. Sneppen, *Science* 296 (2002): 910–913. Reprinted with permission from AAAS (http://www.sciencemag.org). Illustration courtesy of Sergei Maslov, Brookhaven National Laboratory.]

studies the resulting phenotype. The following sections describe some techniques and applications of reverse genetics.

Recombinant DNA can be introduced into the germ line of animals.

Reverse genetics can be carried out in most organisms that have been extensively studied genetically, including the nematode *Caenorhabditis elegans*, *Drosophila*, the mouse, and many domesticated animals and plants. In nematodes, the basic procedure is to manipulate the DNA of interest in a plasmid that also contains a selectable genetic marker that will alter the phenotype of the transformed animal. The DNA is injected directly into the reproductive organs and sometimes spontaneously becomes incorporated into the chromosomes in the germ line. The result of transformation is observed and can be selected in the progeny of the injected animals because of the phenotype conferred by the selectable marker.

A somewhat more elaborate procedure is necessary for germ-line transformation in *Drosophila*. The usual method makes use of a 2.9-kb transposable element called the ***P* element**, which consists of a central region coding for transposase flanked by 31-base-pair inverted repeats (FIGURE 10.20, part A). A genetically engineered derivative of this *P* element, called *wings clipped*, can make functional transposase but cannot itself transpose because of deletions introduced at the ends of the inverted repeats (Figure 10.20, part B). For germ-line transformation, the vector is a plasmid containing a *P* element that includes, within the inverted repeats, a selectable genetic marker (usually one affecting eye color), as well as a large internal deletion that removes much of the transposase-coding region. By itself, this *P* element cannot transpose because it makes no transposase, but it can be mobilized by the transposase produced by the wings-clipped or other intact *P* elements. In *Drosophila* transformation, any DNA fragment of interest is introduced between the ends of the deleted *P* element. The resulting plasmid and a different plasmid containing the wings-clipped element are injected into the region of the early embryo that contains the germ cells. The DNA is taken up by the germ cells, and the wings-clipped element produces functional transposase (Figure 10.20, part B). The transposase mobilizes the engineered *P* vector and results in its transposition into an essentially random location in the genome. Transformants are detected among the progeny of the injected flies because of the eye color or other genetic marker included in the *P* vector. Integration into the germ line is typically very efficient: From 10 to 20 percent of the injected embryos that survive and are fertile yield one or more transformed progeny. However, the efficiency decreases with the size of the DNA fragment in the *P* element, and the effective upper limit is approximately 20 kb.

Transformation of the germ line in mammals can be carried out in several ways. The most direct is the injection of vector DNA into the nucleus of fertilized eggs, which are then transferred to the uterus of

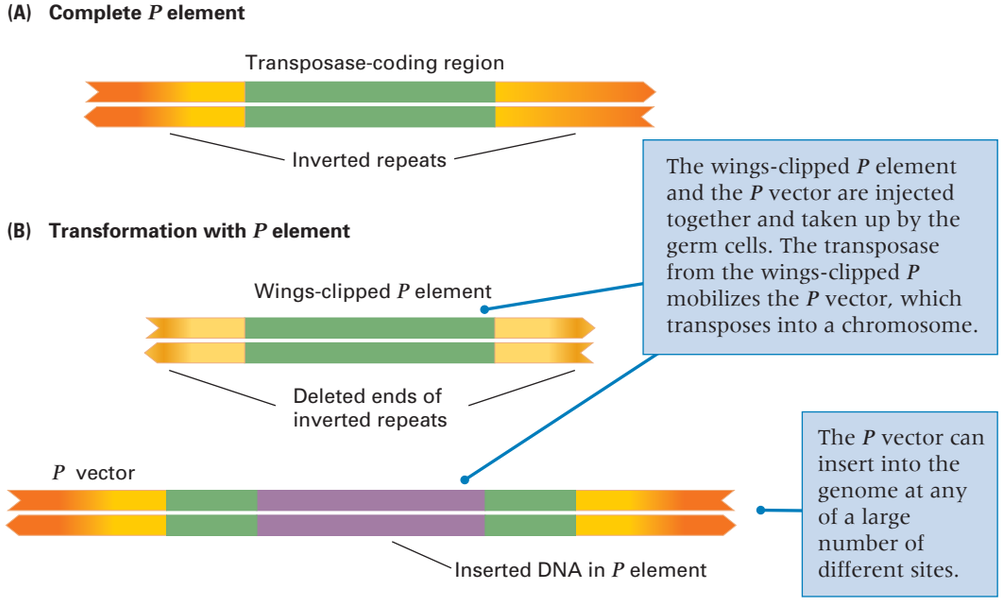


FIGURE 10.20 Transformation in *Drosophila* mediated by the transposable element *P*. (A) Complete *P* element containing inverted repeats at the ends and an internal transposase-coding region. (B) Two-component transformation system. The vector component contains the DNA of interest flanked by the recognition sequences needed for transposition. The wings-clipped component is a modified *P* element that codes for transposase but cannot transpose itself because critical recognition sequences are deleted.

foster mothers for development. The vector is usually a modified retrovirus. **Retroviruses** have RNA as their genetic material and code for a reverse transcriptase that converts the retrovirus genome into double-stranded DNA that becomes inserted into the genome in infected cells. Genetically engineered retroviruses containing inserted genes undergo the same process. Animals that have had new genes inserted into the germ line in this or any other manner are called **transgenic** animals.

Another method of transforming mammals uses **embryonic stem cells** obtained from embryos a few days after fertilization (FIGURE 10.21). Although embryonic stem cells are not very hardy, they can be isolated and then grown and manipulated in culture; mutations in the stem cells can be selected or introduced using recombinant DNA vectors. The mutant stem cells are introduced into another developing embryo and transferred into the uterus of a foster mother (Figure 10.21, part A), where they become incorporated into various tissues of the embryo and often participate in forming the germ line. If the embryonic stem cells carry a genetic marker, such as a gene for black coat color, then mosaic animals can be identified by their spotted coats (part B). Some of these animals, when mated, produce black offspring (part C), which indicates that the embryonic stem cells had become incorporated into the germ line. In this way, mutations introduced into the embryonic stem cells while they were in culture may become incorporated into the germ line of living animals. Embryonic stem cells have been used to create strains of mice with mutations in genes associated with such human genetic diseases as cystic fibrosis. These strains serve as mouse models for studying the disease and for testing new drugs and therapeutic methods.

The procedure for introducing mutations into specific genes is called **gene targeting**. The specificity of gene targeting comes from the DNA sequence

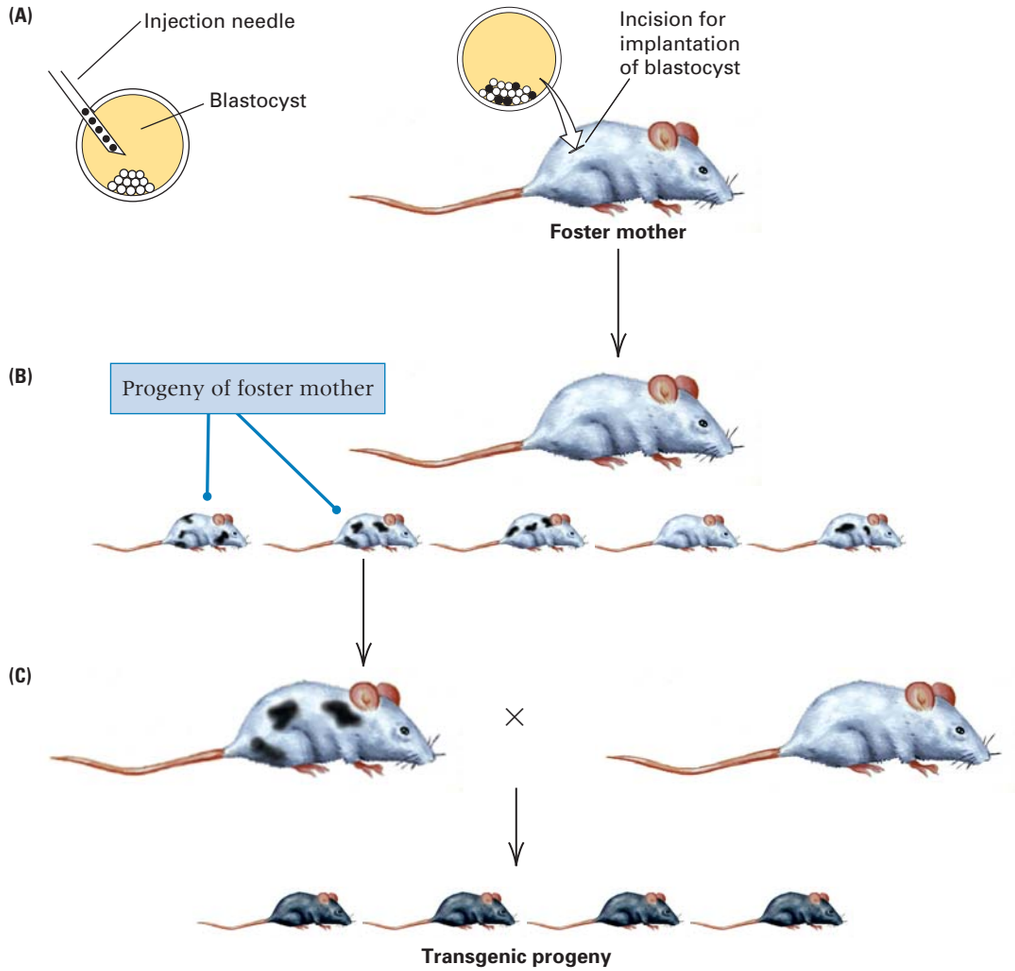


FIGURE 10.21 Transformation of the germ line in the mouse using embryonic stem cells. (A) Stem cells obtained from an embryo of a black strain are isolated and, after genetic manipulation in culture, mixed with the embryonic stem cells from a white strain and introduced into the uterus of a foster mother. (B) The resulting offspring are often mosaics containing cells from both the black and the white strains. (C) If cells from the black strain colonized the germ line, the offspring of the mosaic animal will be black. [Adapted from M. R. Capecchi, *Trends. Genet.* 5 (1989): 70–76.]

the human connection

Pinch of This and a Smidgen of That

Oliver Smithies 2005
University of North Carolina, Chapel Hill, NC
Many Little Things: One Geneticist's View of Complex Diseases

This paper was written on the 25th anniversary of Oliver Smithies's first experimental success in introducing exogeneous DNA into a chosen site in a mammalian genome. The initial applications were to the genetic analysis of phenotypes in which single genes had a major effect. Here Smithies recounts how he came to realize that the approach also could be used to study complex diseases influenced by many genes, each with a relatively small effect. His idea was to create mouse chromosomes either lacking a gene or having an extra copy. By combining these chromosomes through crosses, he could create mice with 0 copies of the gene (if the mouse survived), or 1, 2, 3, or 4 copies. He could then examine the mice and look for associated changes in phenotype. His initial application of this approach was to regulation of blood pressure. This trait was of special interest to Smithies because his father had died from its complications, and because he himself required medication to control his own blood pressure. Smithies's insights and approach were a great success and well deserving of his share of the 2007 Nobel Prize in Physiology and Medicine.

The common multifactorial diseases, such as atherosclerosis, hypertension and diabetes, are caused by complex interactions between multiple genetic and environmental factors. . . . Multifactorial diseases are challenging targets for research because of their complexity. They are attractive because a better understanding of their nature could affect the lives of many people, as more than half of those who live in affluent societies are likely to develop debilitating conditions that have multigenic causes. . . . It seemed likely, as going from two copies to one decreased the level of a gene product, that going from two copies to three copies would increase the level of the product. . . . I realized that this test of the effects of quantitative changes in gene expression could usefully be applied to all genes that potentially alter blood pressure. . . .

> **My mindset had begun to change . . . to the thought that [hypertension] was more likely to be the result of "many little things."**

My mindset had begun to change from the idea that a few major differences might determine this complex phenotype to the thought that it was more likely to be the result of "many little things." . . . I decided . . . [to set up] a computer simulation of the system. . . . Not being a computer buff, I spent a rather unenjoyable six months working off and on with the simulation. But the outcome was most revealing, and I have since, with ever-increasing enjoyment, used relatively simple computer simulations in much of my work. . . . I have found that the greatest value of these simulations is . . . that one is forced to identify crucial elements and define clearly the assumptions required to integrate them into a logical whole. In the case of our hypertension studies, the simulations showed that increases in the number of copies of the *AGT* [angiotensinogen] gene should cause . . . an increase in blood pressure. . . . My experiences during the course of this voyage into the field of multifactorial diseases led me to emphasize the importance of looking for quantitative differences as well as qualitative differences, when searching for genetic factors that determine individual risk of common disorders. Many quantitative differences have accumulated during human evolution—some easily seen as outward differences in our body proportions, some hidden as inward differences in gene expression—that are without sufficient effects to have been fixed or eliminated by selection. These 'many little things' are a joyful source of our individuality. But they are probably also a source of the poorly understood differences in individual susceptibility to [complex diseases].

Source: O. Smithies, *Nat. Rev. Genet.* 6 (2005): 419–425.

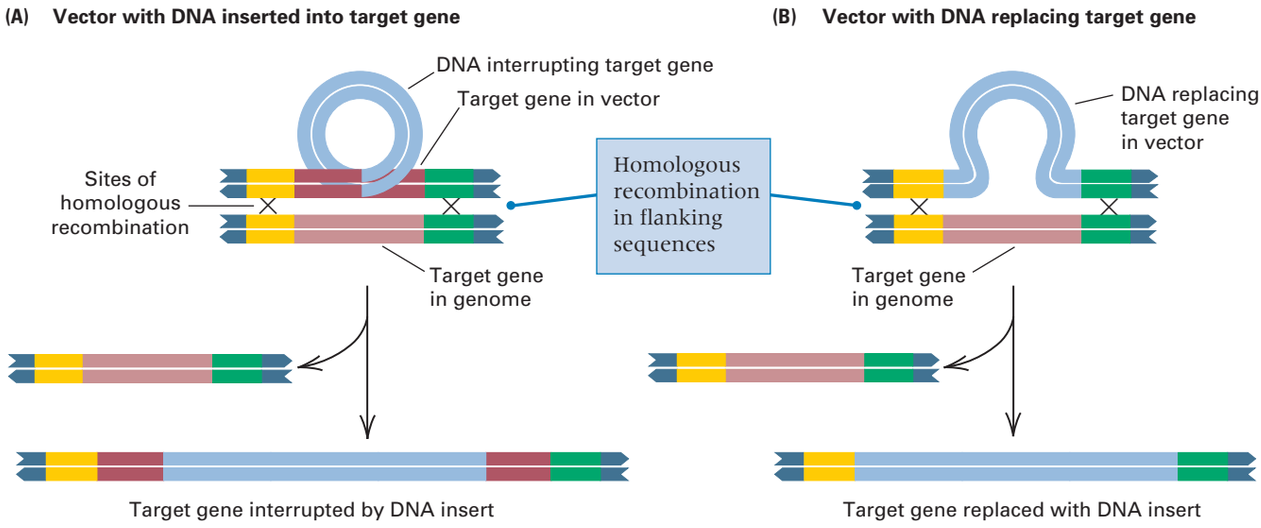


FIGURE 10.22 Gene targeting in embryonic stem cells. (A) The vector (top) contains the targeted sequence (red) interrupted by an insertion. Homologous recombination introduces the insertion into the genome. (B) The vector contains DNA sequences flanking the targeted gene. Homologous recombination results in replacement of the targeted gene with an unrelated DNA sequence. [Adapted from M. R. Capecchi, *Trends. Genet.* 5 (1989): 70–76.]

homology needed for homologous recombination. Two examples are illustrated in FIGURE 10.22, where the DNA sequences present in gene-targeting vectors are shown as looped configurations paired with homologous regions in the chromosome prior to recombination. The targeted gene is shown in pink. In part A of Figure 10.22, the vector contains the targeted gene interrupted by an insertion of a novel DNA sequence, and homologous recombination results in the novel sequence becoming inserted into the targeted gene in the genome. In part B of Figure 10.22, the vector contains only flanking sequences, not the targeted gene, so homologous recombination results in replacement of the targeted gene with an unrelated DNA sequence. In both cases, cells with targeted gene mutations can be selected by including an antibiotic-resistance gene, or other selectable genetic marker, in the sequences that are incorporated into the genome through homologous recombination.

Recombinant DNA can also be introduced into plant genomes.

A procedure for the transformation of plant cells makes use of a plasmid found in the soil bacterium *Agrobacterium tumefaciens* and related species. Infection of susceptible plants with this bacterium results in the growth of what are known as *crown gall tumors* at the entry site, which is usually a wound. Susceptible plants comprise about 160,000 species of flowering plants, known as the dicots, and include the great majority of the most common flowering plants.

The *Agrobacterium* contains a large plasmid of approximately 200 kb called the ***Ti* plasmid**, which includes a smaller (~25 kb) region known as the ***T* DNA** flanked by 25-base-pair direct repeats (FIGURE 10.23, part A). The *Agrobacterium* causes a profound change in the metabolism of infected cells because of transfer of the *T* DNA into the plant genome. The *T* DNA contains genes coding for proteins that stimulate division of infected cells, hence causing the tumor, and also coding for enzymes that convert the amino acid arginine into an unusual derivative, generally *nopaline* or *octopine* (depending on the particular type of *Ti* plasmid), that the bacterium needs in order to grow. The transfer functions are present not in the *T* DNA itself but in another region of the plasmid called the *vir* (stands for *virulence*) region of about 40 kb that includes six genes necessary for transfer.

Transfer of *T* DNA into the host genome is similar in some key respects to bacterial conjugation, which we examined in Chapter 8. In infected cells, transfer begins with the formation of a nick that frees one end of the *T* DNA (Figure 10.23, part A), which peels off the plasmid and is replaced by rolling-circle replication (Figure 10.23, part B). The region of the plasmid that is transferred is delimited by a second nick at the other end of the *T* DNA, but the position of this nick is variable. The resulting single-stranded *T* DNA is bound with molecules of a single-stranded binding protein (SSBP) and is transferred into the plant cell and incorporated into the nucleus. There it is integrated into the chromosomal DNA by a mechanism that is still unclear (Figure 10.23, part C). Although the SSBP has certain similarities in amino acid

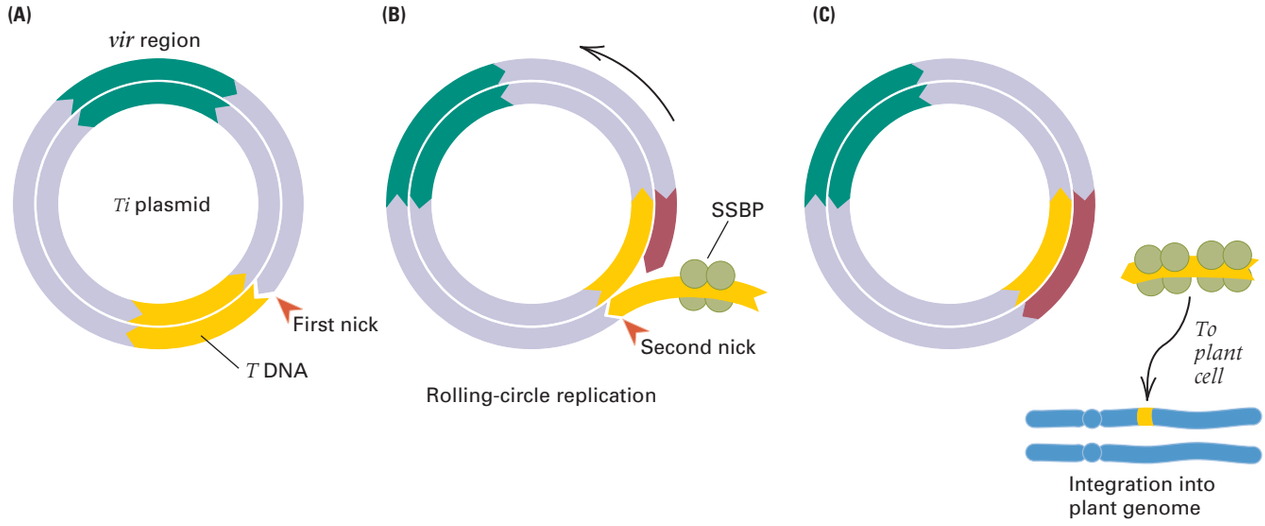


FIGURE 10.23 Transformation of a plant genome by *T* DNA from the *Ti* plasmid. (A) A nick forms at the 5′ end of the *T* DNA. (B) Rolling-circle replication elongates the 3′ end and displaces the 5′ end, which is stabilized by single-stranded binding protein (SSBP). A second nick terminates replication. (C) The SSBP-bound *T* DNA is transferred to a plant cell and inserts into the genome.

sequence to the *recA* protein from *E. coli*, which plays a key role in homologous recombination, it is clear that integration of the *T* DNA does not require homology.

Use of *T* DNA in plant transformation is made possible by engineered plasmids in which the sequences normally present in *T* DNA are removed and replaced with those to be incorporated into the plant genome along with a selectable marker. A second plasmid contains the *vir* genes and permits mobilization of the engineered *T* DNA. In infected tissues, the *vir* functions mobilize the *T* DNA for transfer into the host cells and integration into the chromosome. Transformed cells are selected in culture by the use of the selectable marker and then grown into mature plants in accordance with the methods described in Section 5.4.

Transformation rescue is used to determine experimentally the physical limits of a gene.

One of the important applications of germ-line transformation is to define experimentally the limits of any particular gene along the DNA. Every gene includes upstream and downstream sequences that are necessary for its correct expression, as well as the coding sequences. As we saw in Chapter 9, these regulatory sequences control the time in development, the cell types, and the level at which transcription occurs. In defining the upstream and downstream limits of a gene, even knowing the complete nucleotide sequence of the coding region and the flanking DNA may be insufficient. The main reason is that there is no general method by which to identify regulatory sequences. Regulatory sequences are often composites of short, seemingly nondescript sequences, which are in fact the critical binding sites for regulatory proteins that control transcription.

To see how germ-line transformation is used to define the limits of a functional gene, consider the example of the *Drosophila* gene *white*, which, when mutated, results in flies with white eyes instead of red. The genetic organization of *white* is illustrated in FIGURE 10.24. The primary RNA transcript (part B) is a little more than 6 kb in length and includes five introns that are excised and degraded during RNA processing. The resulting mRNA (part C) is translated into a protein of 720 amino acids (part D), which is a member of a large family of related ATP-dependent transmembrane proteins known as *ABC transporters* that regulate the traffic of their target molecules or ions across the cell membrane. The White transporter is located in the pigment-producing cells in the eye, and its target molecule is one of the key precursors of the eye-color pigments; hence flies with a mutant White transporter have white eyes.

The question addressed by germ-line transformation is this: How much DNA upstream and downstream of *white* is necessary for the fly to produce the wildtype transporter protein in the pigment cells? The experimental approach is first to clone a large fragment of DNA that includes the coding sequence for the wildtype White protein, then to use germ-line transformation to introduce this fragment into the genome of a fly that contains a *white* mutation. If the introduced DNA includes all the 5′ and 3′ regulatory sequences necessary for correct gene expression, then the phenotype of the resulting flies will be wildtype, because the wildtype gene is dominant to the *white* mutation. The ability of an introduced DNA fragment to correct a genetic defect in a

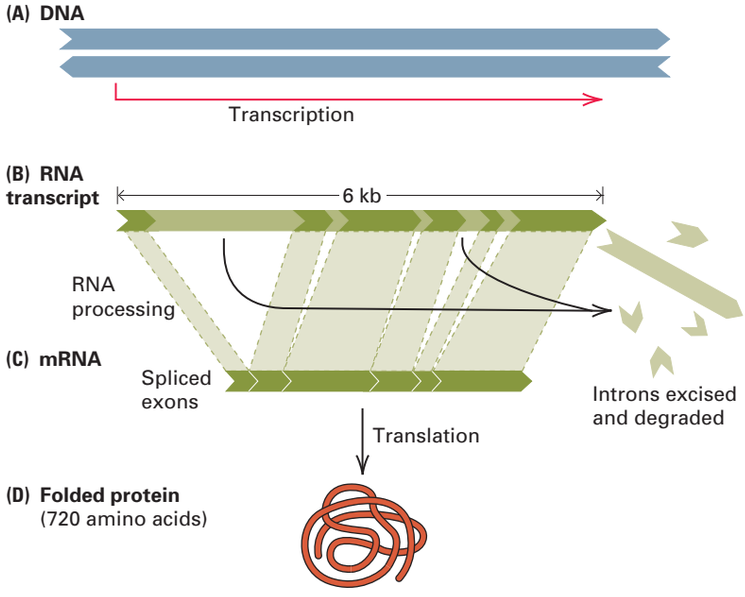
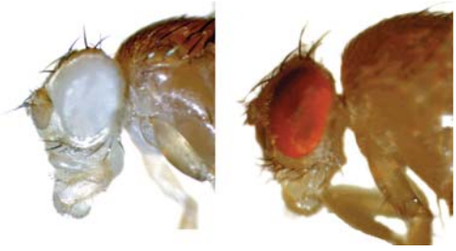


FIGURE 10.24 Genetic organization of the *Drosophila* gene *white*.



(A) White eyes **(B)** Red eyes

FIGURE 10.25 Mutant white-eyed and wildtype red-eyed males of *Drosophila melanogaster*. [Courtesy of E. R. Lozovsky.]

mutant organism is called **transformation rescue** (FIGURE 10.25), and it means that the fragment contains all the essential regulatory sequences. The fragment may also include some sequences that are nonessential, so finding the *minimal* 5′ and 3′ flanking regions requires that the smaller pieces of the original fragment also be assayed for transformation rescue. In the case of *white*, the minimal DNA fragment is about 8.5 kb in length, starting about 2 kb upstream from the transcription start site.

10.5 Genetic engineering is applied in medicine, industry, agriculture, and research.

Recombinant DNA technology has revolutionized modern biology not only by opening up new approaches in basic research but also by making possible the creation of organisms with novel genotypes for practical use in agriculture and industry. In this section we examine a few of many applications of recombinant DNA.

Animal growth rate can be genetically engineered.

In many animals, the rate of growth is controlled by the amount of growth hormone produced. Transgenic animals with a growth-hormone gene under the control of a highly active promoter to drive transcription often grow larger than their normal counterparts. An example of a highly active promoter is found in the gene for *metallothionein*. The metallothioneins are proteins that bind heavy metals. They are ubiquitous in eukaryotic organisms and are encoded by members of a family of related genes. The human genome, for example, includes more than ten metallothionein genes that can be separated into two major groups according to their sequences. The promoter region of a metallothionein gene drives transcription of any gene to which it is attached, in response to heavy metals or steroid hormones. For example, when DNA constructs consisting of a rat growth-hormone gene under metallothionein control are used to produce transgenic mice, the resulting animals grow about twice as large as normal mice.

The effect of another growth-hormone construct is shown in FIGURE 10.26. The fish are coho salmon at 14 months of age. Those on the left are normal, whereas those on the right are transgenic animals that contain a salmon growth-hormone gene driven by a metallothionein regulatory region. Both the growth-hormone gene and the metallothionein gene were cloned from the sockeye salmon. As an indicator of size, the largest transgenic fish on the right has a length of about 42 cm. On average, the transgenic fish are 11 times heavier than their normal counterparts; the largest transgenic fish was 37 times the average weight of the nontransgenic animals. Not only do the transgenic salmon grow faster and become larger than normal salmon; they also mature faster.

Crop plants with improved nutritional qualities can be created.

Beyond the manipulation of single genes, it is also possible to create transgenic organisms that have entire new metabolic pathways introduced. A remarkable example is in the creation of a genetically engineered rice that contains an introduced biochemical pathway for the synthesis of β-carotene, a precursor of vitamin A found primarily in yellow

vegetables and greens. (Deficiency of vitamin A affects some 400 million people throughout the world, predisposing them to skin disorders and night blindness.) The β-carotene pathway includes four enzymes, which in the engineered rice are encoded in genes from different organisms (FIGURE 10.27). Two of the genes come from the common daffodil (*Narcissus pseudonarcissus*), whereas the other two come from the bacterium *Erwinia uredovora*. Each pair of genes was cloned into *T* DNA and transformed into rice using *Agrobacterium tumefaciens* by the mechanism outlined in Figure 10.23. Transgenic plants were then crossed to produce progeny containing all four enzymes. The engineered rice seeds contain enough β-carotene to provide the daily requirement of vitamin A in 300 grams of cooked rice; they even have a yellow tinge (Figure 10.27, part B).

FIGURE 10.26 Normal coho salmon (left) and genetically engineered coho salmon (right) containing a sockeye salmon growth-hormone gene driven by the regulatory region from a metallothionein gene. The transgenic salmon average 11 times the weight of the nontransgenic fish. The smallest fish on the left is about 4 inches long. [Courtesy of R. H. Devlin, Fisheries and Oceans Canada (after Devlin et. al. 1994; *Nature* 371: 209–210).]

People on high-rice diets are also prone to iron deficiency because rice contains a small phosphorus-storage molecule called *phytate*, which binds with iron and interferes with its absorption through the intestine. The transgenic β-carotene rice was also engineered to minimize this problem by introducing the fungal enzyme from *Aspergillus ficuum* that breaks down phytate, along with a gene encoding the iron-storage protein ferritin from the French bean, *Phaseolus vulgaris*, plus yet another gene from basmati rice that encodes a metallothionein-like gene that facilitates iron absorption in the human gut. Altogether, then,

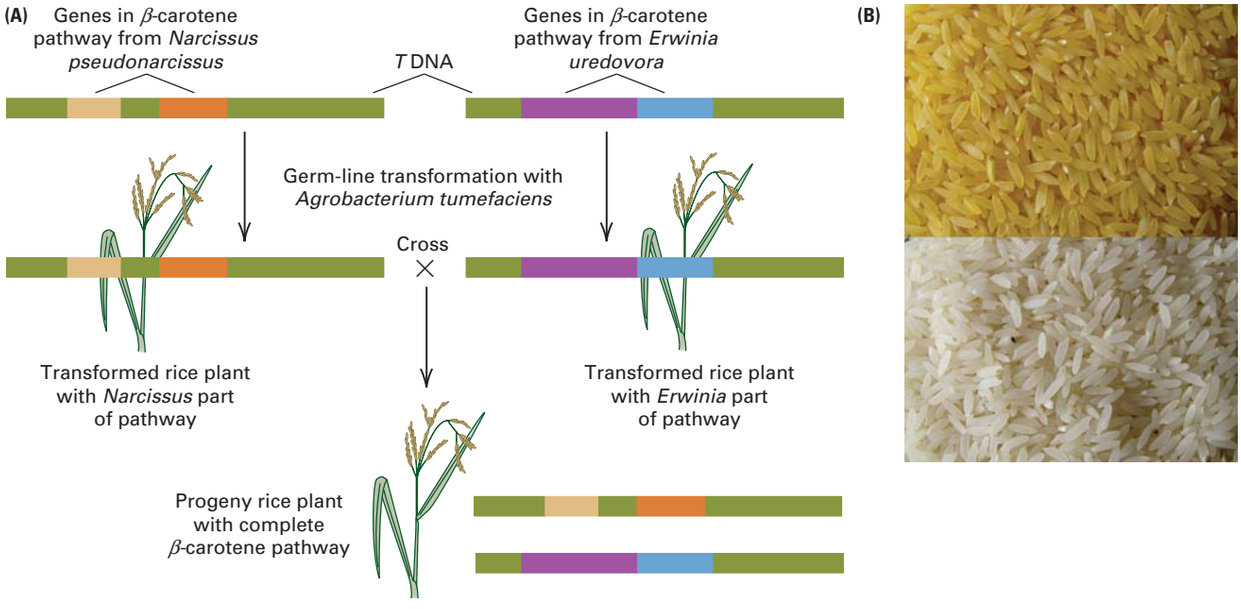


FIGURE 10.27 Genetically engineered rice containing a biosynthetic pathway for β-carotene. (A) Enzymes in the pathway derive from genes in two different species. (B) Rice plants with both parts of the pathway produce grains with a yellowish cast (top) because of the β-carotene they contain, in contrast to the pure white grains (bottom) of normal plants. [B, Courtesy of Ingo Potrykus, Institute für Pflanzenwissenschaften, ETH Zurich.]

the transgenic rice strain rich in β-carotene and available iron contains six new genes taken from four unrelated species plus one gene from a totally different strain of rice!

The production of useful proteins is a primary impetus for recombinant DNA.

Among the most important applications of genetic engineering is the production of large quantities of particular proteins that are otherwise difficult to obtain (for example, proteins that are present in only a few molecules per cell or that are produced in only a small number of cells or only in human cells). The method is simple in principle. A DNA sequence coding for the desired protein is cloned in a vector adjacent to an appropriate regulatory sequence. This step is usually done with cDNA, because cDNA has all the coding sequences spliced together in the right order. Using a vector with a high copy number ensures that many copies of the coding sequence will be present in each bacterial cell, which can result in synthesis of the gene product at concentrations ranging from 1 to 5 percent of the total cellular protein. In practice, the production of large quantities of a protein in bacterial cells is straightforward, but there are often problems that must be overcome, because in the bacterial cell, which is a prokaryotic cell, the eukaryotic protein may be unstable, may not fold properly, or may fail to undergo necessary chemical modification. Many important proteins are currently produced in bacterial cells, including human growth hormone, blood-clotting factors, and insulin. Patent offices in Europe and the United States have already issued tens of thousands of patents for the clinical use of the products of genetically engineered human genes. FIGURE 10.28 gives a breakdown of the numbers of patents issued for various clinical applications.

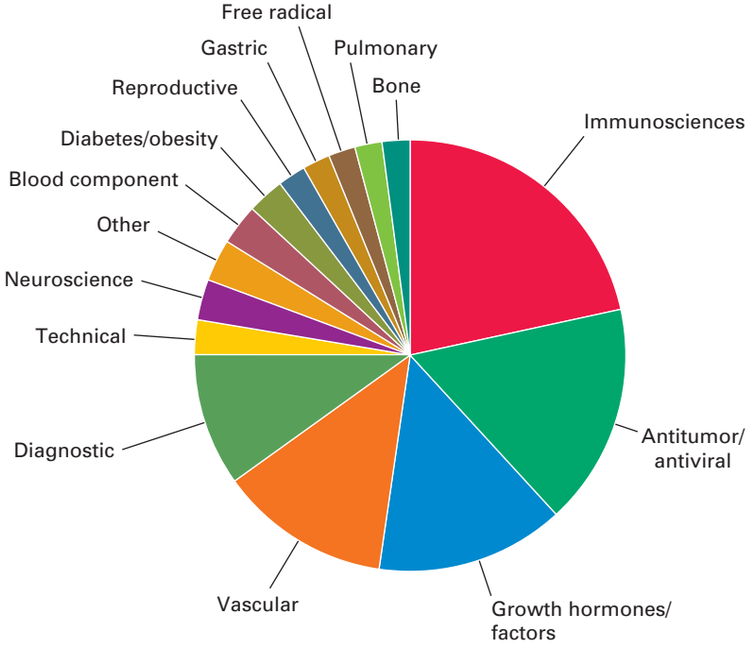


FIGURE 10.28 Relative numbers of patents issued for various clinical applications of the products of genetically engineered human genes. [Data from S. M. Thomas, et al., *Nature* 380 (1996): 387–388.]

Animal viruses may prove useful vectors for gene therapy.

Genetic engineering in animal cells often exploits RNA retroviruses that use reverse transcriptase to make a double-stranded DNA copy of their RNA genome. The DNA copy then becomes inserted into the chromosomes of the cell. Ordinary transcription of DNA to RNA occurs only after the DNA copy is inserted. The infected host cell survives the infection, retaining the DNA copy of the retroviral RNA in its genome. These features of retroviruses make them convenient vectors for the genetic manipulation of animal cells, including those of birds, rodents, monkeys, and human beings.

Genetic engineering with retroviruses allows the possibility of altering the genotypes of animal cells. Because a wide variety of retroviruses are known, including many that infect human cells, genetic defects may be corrected by these procedures in the future. The recombinant DNA procedure employed with retroviruses consists of the *in vitro* synthesis of double-stranded DNA from the viral RNA by means of reverse transcriptase. The DNA is then cleaved with a restriction enzyme and, using techniques already described, any DNA fragment of interest is inserted. Transformation yields cells with the recombinant retroviral DNA permanently inserted into the genome. However, many retroviruses contain genes that result in uncontrolled proliferation of the infected cell, thereby causing a tumor. When retrovirus vectors are used for genetic engineering, the cancer-causing genes are first deleted. The deletion also provides the space needed for incorporation of the desired DNA fragment.

Attempts are currently under way to assess the potential use of retroviral vectors in **gene therapy**, or the correction of genetic defects in somatic cells by genetic engineering. Noteworthy successes so far include the correction of immunological deficiencies in patients with various kinds of inherited disorders. However, a number of major problems stand in the way of gene therapy becoming widely used. At this time, there is no completely reliable way to ensure that a gene will be inserted only

into the appropriate target cell or target tissue. For example, in one of the earliest clinical trials with a group of four patients treated with retroviral vectors for severe combined immunodeficiency disease, one of the patients had a retroviral insertion into a site that caused aberrant expression of a gene, *LMO-2*, associated with acute lymphoblastic leukemia.

A major breakthrough in disease prevention would come through the development of synthetic vaccines produced by recombinant DNA. Production of natural vaccines is often unacceptable because of the extreme hazards of working with large quantities of the active virus—for example, the human immunodeficiency virus (HIV) that causes acquired immune deficiency syndrome (AIDS). The danger can be minimized by cloning and producing viral antigens in a nonpathogenic organism. Vaccinia virus, the agent used in smallpox vaccination, has attracted much attention as a candidate for this application. Viral antigens are often on the surface of virus particles, and some of these antigens can be engineered into the coat of vaccinia. For example, engineered vaccinia virus with certain surface antigens of hepatitis B virus, influenza virus, and vesicular stomatitis virus (which kills cattle, horses, and pigs) has already proved effective in animal tests. One of the great challenges is malaria, which affects approximately 300 million people in Africa, Asia, and Latin America, causing approximately 1 to 1.5 million deaths per year. Control of the disease by drugs is increasingly compromised by the spread of resistance mutations, but a recently developed vaccine shows great promise.

»» CHAPTER SUMMARY

- In recombinant DNA (gene cloning), DNA fragments are isolated, inserted into suitable vector molecules, and introduced into host cells (usually bacteria or yeast), where they are replicated.
- Specialized methods for manipulating and cloning large fragments of DNA have resulted in integrated physical and genetic maps of the DNA in complex genomes.
- High-throughput automated DNA sequencing has resulted in the complete sequence of the genomes of many species of bacteria, archaeons, and eukaryotes, including the genomes of humans and other higher primates.
- Comparisons among genomes of related species helps discover coding sequences and other functional genetic elements.
- Functional genomics using DNA microarrays enables the level of gene expression of all genes in the genome to be assayed simultaneously, which allows global patterns and coordinated regulation of gene expression to be investigated. Proteomics methods, such as two-hybrid analysis of proteins, allows protein–protein interaction networks to be identified.
- Recombinant DNA is widely used in research, medical diagnostics, and the manufacture of drugs and other commercial products.
- Transgenic organisms carry DNA sequences that have been introduced by germ-line transformation or other methods.

»» ISSUES AND IDEAS

- What does the term *recombinant DNA* mean? What are some of the practical uses of recombinant DNA?
- What features are essential in a bacterial cloning vector? How can a vector have more than one cloning site?
- What is the reaction catalyzed by the enzyme reverse transcriptase? How is this enzyme used in recombinant DNA technology?
- What is meant by the term *genome annotation*? Explain why genome sequences need to be annotated.
- What are DNA microarrays and how are they used in functional genomics?
- Describe the two-hybrid system that makes use of the yeast GAL4 protein, and explain how the two-hybrid system detects interaction between proteins.
- What is a transgenic organism? What are some of the practical uses of transgenic organisms?
- Explain the role of recombination in gene targeting in embryonic stem cells. How can gene targeting be used to create a "knockout" mutation?

»» SOLUTIONS: STEP BY STEP

Problem 1

What is the average distance between restriction sites for each of the following restriction enzymes? Assume that the DNA substrate has a random sequence with equal amounts of each base. The symbol N stands for any nucleotide, R for any purine (A or G), and Y for any pyrimidine (T or C).

(a) TCGA (*Taq*I)
(b) GGTACC (*Kpn*I)
(c) GTNAC (*Mae*III)
(d) GGNNCC (*Nla*IV)
(e) GRCGYC (*Acy*I)

Solution. **(a)** The average distance between restriction sites equals the reciprocal of the probability of occurrence of the restriction site. You must, therefore, calculate the probability of occurrence of each restriction site in a random DNA sequence. The probability of the sequence TCGA is

$$1/4 \times 1/4 \times 1/4 \times 1/4 = (1/4)^4 = 1/256$$

and so 256 bases is the average distance between *Taq*I sites. **(b)** By the same reasoning, the probability of a GGTACC site is $(1/4)^6 = 1/4096$, and so 4096 bases is the average distance between *Kpn*I sites. **(c)** The probability of N (any nucleotide at a site) is 1, and, hence, the probability of the sequence GTNAC

$$1/4 \times 1/4 \times 1 \times 1/4 \times 1/4 = (1/4)^4 = 1/256$$

Therefore, 256 is the average distance between *Mae*III sites. **(d)** The same reasoning yields the average distance between GGNNCC (*Nla*IV) sites as $1/4 \times 1/4 \times 1 \times 1 \times 1/4 \times 1/4 = 1/256$ bases. **(e)** The probability of an R (A or G) at a site is 1/2, and the probability of a Y (T or C) at a site is 1/2. Hence, the probability of the sequence GRCGYC is

$$1/4 \times 1/2 \times 1/4 \times 1/4 \times 1/2 \times 1/4 = 1/1024$$

and so the average number of bases between *Acy*I sites is 1024 bases.

Problem 2

How many clones are needed to establish a library of DNA from a species of lemur with a diploid genome size of 6×10^9 base pairs if (1) the clones contain fragments of average size 2×10^4 base pairs, and (2) one wants 99 percent of the genomic sequences to be present in at least one clone in the library? (*Hint*: If a genome is cloned at random into a library with x-fold coverage, the probability that a particular sequence will be missing from the library is e^{-x}.)

Solution. The hint says that if the genome were represented x times in the library, the probability that a particular sequence would be missing is e^{-x}, which we want to equal 0.01. Hence, the required library should have x-fold coverage, where $e^{-x} = 0.01$ or $x = -\ln(0.01) = 4.6$. Because one haploid representation of the genome equals $(6 \times 10^9)/2 = 3 \times 10^9$ base pairs, and the average insert size is 2×10^4 base pairs, the required library should include $[(3 \times 10^9)/(2 \times 10^4)] \times 4.6 = 6.9 \times 10^5$ clones.

»» CONCEPTS IN ACTION: PROBLEMS FOR SOLUTION

10.1 Will the sequences 5′–AATT–3′ and 3′–AATT–5′ in a double-stranded DNA molecule be cut by the same restriction enzyme?

10.2 A circular plasmid has two restriction sites for the enzyme *Zsp2*I, which cleaves the site ATGCA↓T (the arrow indicates the position of the cleavage). After digestion the fragments are ligated together, and a circular product is isolated that includes one copy of each of the fragments. Does this mean that the ligated plasmid is the same as the original? Explain.

10.3 In recombinant DNA, researchers typically prefer ligating restriction fragments that have sticky ends (single-stranded overhangs) rather than those that have blunt ends. Can you propose a reason why?

10.4 A *kan-r tet-r* plasmid is treated with the restriction enzyme *Bgl*I, which cleaves the *kan* (kanamycin) gene. The DNA is annealed with a *Bgl*I digest of *Neurospora* DNA and after ligation used to transform *E. coli*.

(a) What antibiotic would you put into the growth medium to ensure that each colony has the plasmid?
(b) What antibiotic-resistance phenotypes would be found among the resulting colonies?
(c) Which phenotype is expected to contain *Neurospora* DNA inserts?

10.5 You want to introduce the human insulin gene into a bacterial host in hopes of producing a large amount of human insulin. Should you use the genomic DNA or the cDNA? Explain your reasoning.

10.6 You decide to clone your pet dog, which is brown with black spots. You take a few somatic cells from your dog and perform a somatic cell nuclear transfer procedure using an egg from a female dog that is black. In this procedure, the egg nucleus is removed and replaced with that from a somatic cell. What color fur will the puppy clone of your dog have?

10.7 After doing a restriction digest with the enzyme *Sse*I, which has the recognition site 5′-CCTGCA↓GG-3′ (the arrow indicates the position of the cleavage), you wish to separate the fragments in an agarose gel. In order to choose the proper concentration of agarose, you need to know the expected size of the fragments. Assuming equivalent amounts of each of the four nucleotides in the target DNA, what average fragment size would you expect?

10.8 The restriction enzymes *Acc*651 and *Kpn*I have the restriction sites

G↓GTACCT (*Acc*651)
GGTAC↓C (*Kpn*I)

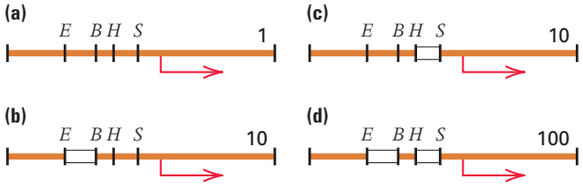
where the 5′ end is written at the left and the arrow indicates the position of the cleavage. Are the sticky ends produced by these restriction enzymes compatible? Explain.

10.9 In cloning into bacterial vectors, why is it useful to insert DNA fragments to be cloned into a restriction site inside an antibiotic-resistance gene? Why is another gene for resistance to a second antibiotic also required?

10.10 A mutant allele is found to express the wildtype gene product but at only about 20 percent of the wildtype level. The mutation is traced to an intron whose size has increased by 3.1 kb because of the presence of a DNA fragment with the restriction map shown here. The symbols *A, B, C, D, E, H, K, P, S,* and *X* represent cleavage sites for the restriction enzymes *Alu*I, *Bam*HI, *Cla*I, *Dde*I, *Eco*RI, *Hind*III, *Kpn*I, *Pst*I, *Sac*I, and *Xho*II, respectively. Does the restriction map of the insertion give any clues to what it is?

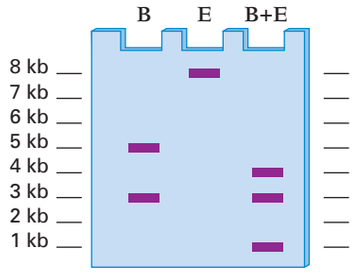
10.11 If the genomic and cDNA sequences of a gene are compared, what information does the cDNA sequence provide that is not obvious from the genomic sequence? What information does the genomic sequence contain that is not in the cDNA?

10.12 In studies of the operator region of an inducible operon in *E. coli*, the four constructs shown below were examined for level of transcription *in vitro*. The number associated with each construct is the relative level of transcription observed in the presence of the repressor protein. The symbols *E, B, H,* and *S* stand for the restriction sites *Eco*RI, *Bam*HI, *Hind*III, and *Sac*I. Construct **(a)** is the wildtype operator region, and in parts **(b–d)** the open boxes indicate restriction fragments that were deleted. What hypothesis about repressor–operator interactions can explain these results? How could this hypothesis be tested?



10.13 *Arabidopsis thaliana* has among the smallest genomes in higher plants, with a haploid genome size of about 100 Mb. If this genome is digested with *Not*I (an eight-base cutter), approximately how many DNA fragments would be produced? Assume equal and random frequencies of the four nucleotides.

10.14 A circular plasmid of 8 kb is digested with *Eco*RI (E) and/or *Bam*HI (B), and the digests are run on an agarose gel and stained. The results are shown below; molecular size standards are shown.



Draw the map of the plasmid.

10.15 How frequently would the restriction enzymes *Taq*I (restriction site TCGA) and *Mae*III (restriction site GTNAC, in which N is any nucleotide) cleave double-stranded DNA molecules containing each of the following random sequences?

(a) 1/6 A, 1/6 T, 1/3 G, and 1/3 C
(b) 1/3 A, 1/3 T, 1/6 G, and 1/6 C

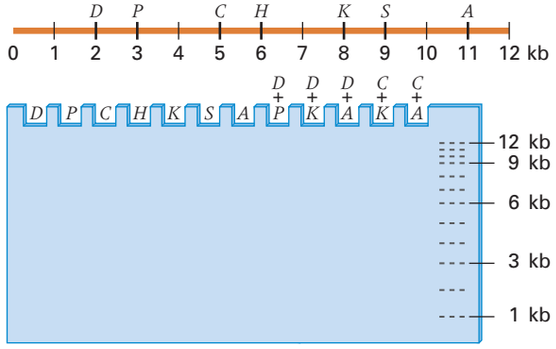
10.16 How many clones are needed to establish a library of DNA from a species of grasshopper with a diploid genome size of 1.8×10^{10} base pairs if (1) fragments of average size 1×10^4 base pairs are used, and (2) one wants 95 percent of the genomic sequences to be in the library? (*Hint*: If the genome is cloned at random with *x*-fold coverage, the probability that a particular sequence will be missing is e^{-x}.)

10.17 Suppose that you digest the genomic DNA of a particular organism with *Sau*3A (↓GATC), where the arrow represents the cleavage site. Then you ligate the resulting fragments into a unique *Bam*HI (G↓GATCC) cloning site of a plasmid vector. Would it be possible to isolate the cloned fragments from the vector using *Bam*HI? From what proportion of clones would it be possible?

10.18 A DNA microarray is hybridized with fluorescently labeled reverse-transcribed DNA as described in the text, where the control mRNA (C) is labeled with a green fluorescent compound and the experimental mRNA (E) with a red fluorescent compound. Indicate what you can conclude about the relative levels of expression of a spot in the microarray that fluoresces:

(a) Red
(b) Green
(c) Yellow
(d) Orange
(e) Lime green

10.19 Shown here is a restriction map of a 12-kb linear plasmid isolated from cells of *Borrelia burgdorferi*, a spirochete bacterium transmitted by the bite of *Ixodes* ticks that causes Lyme disease. The symbols *D, P, C, H, K, S,* and *A* represent cleavage sites for the restriction enzymes *Dde*I, *Pst*I, *Cla*I, *Hind*III, *Kpn*I, *Sac*I, and *Alu*I, respectively. In the accompanying gel diagram, show the positions at which bands would be found after digestion of the plasmid with the indicated restriction enzyme or enzymes.



10.20 A functional genomics experiment is carried out using a DNA microarray to assay levels of gene expression in a species of bacteria. What genes would you expect to find overexpressed in cells grown in minimal medium compared to cells grown in complete medium?

10.21 A functional genomics experiment is carried out in *E. coli* to examine global levels of gene expression in various types of minimal growth medium. RNA extracted from the experimental culture is labeled with a molecule that fluoresces red, and RNA extracted from the control culture is labeled with a molecule that fluoresces green. The experimental and control samples are mixed prior to hybridization. Shown here are spots on the microarray corresponding to five genes: *trpE* (the first gene in the tryptophan biosynthetic operon), *lacI*, *lacZ*, *lacY*, and *crp* (which encodes the cyclic AMP receptor protein). Color the spots red, green, or yellow according to the relative levels of expression of each gene in the experimental and control cultures. (*Hint*: Before answering, think carefully about how the cyclic AMP receptor protein co-regulates the *lac* operon.)

Experimental minimal medium	Control minimal medium	Transcript				
		trpE	*lacI*	*lacZ*	*lacY*	*crp*
Glucose	Glucose	○	○	○	○	○
Glucose	Glycerol	○	○	○	○	○
Glycerol	Glucose	○	○	○	○	○
Lactose	Glucose	○	○	○	○	○
Glucose	Lactose	○	○	○	○	○
Lactose	Glycerol	○	○	○	○	○
Glycerol	Lactose	○	○	○	○	○

GENETICS on the web

GeNETics on the Web will introduce you to some of the most important sites for finding genetics information on the Internet. To explore these sites, visit the Jones and Bartlett companion site to accompany *Essential Genetics: A Genomic Perspective*, *Fifth Edition*, at http://biology.jbpub.com/Hartl/EssentialGenetics.

There you will find a chapter-by-chapter list of highlighted keywords. When you select one of the keywords, you will be linked to a Web site containing information related to that keyword.