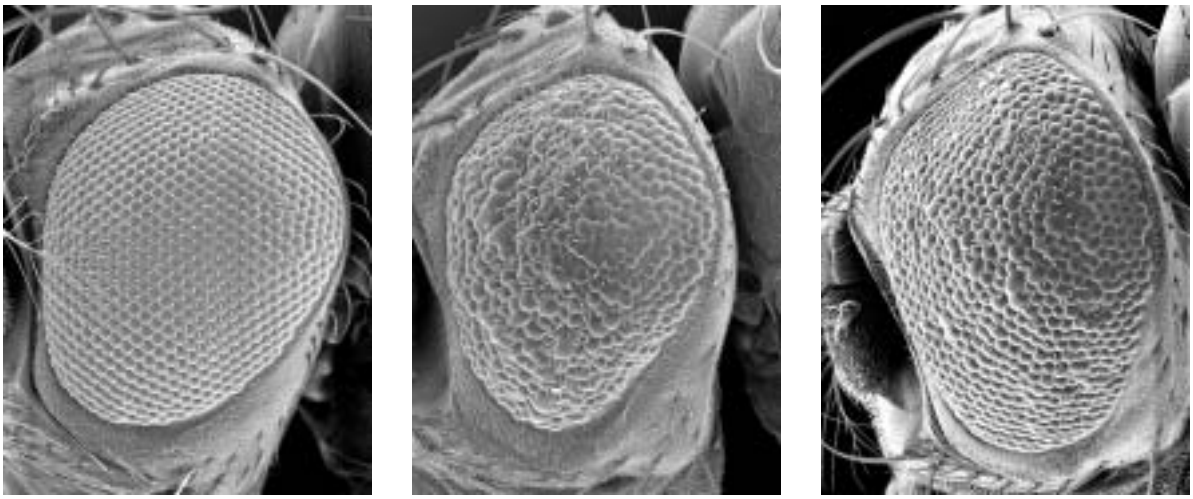
9

MOLECULAR GENETIC TECHNIQUES AND GENOMICS

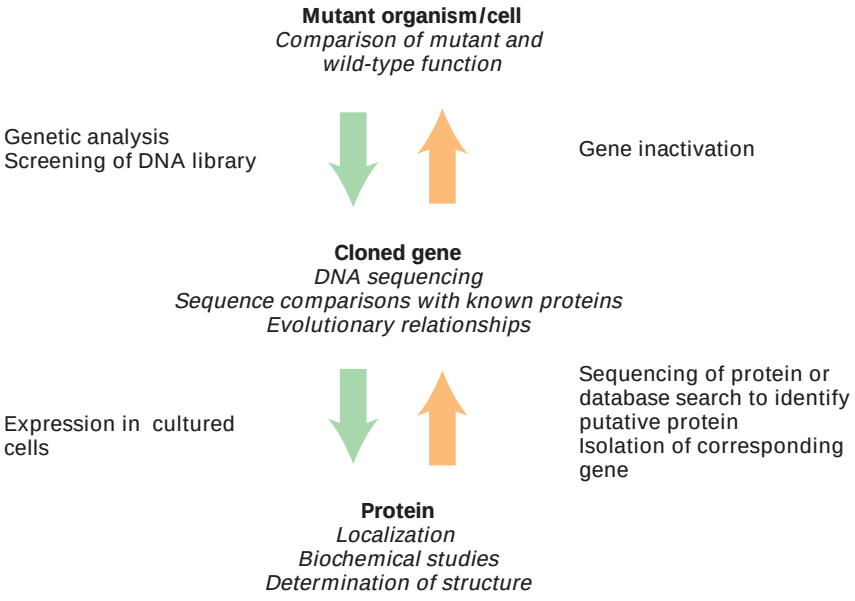


The effect of mutations on *Drosophila* development. Scanning electron micrographs of the eye from (*left*) a wild-type fly, (*middle*) a fly carrying a dominant developmental mutation produced by recombinant DNA methods, and (*right*) a fly carrying a suppresor mutation that partially reverses the effect of the dominant mutation. [Courtesy of Ilaria Rebay, Whitehead Institute, MIT.]

In previous chapters, we were introduced to the variety of tasks that proteins perform in biological systems. How some proteins carry out their specific tasks is described in detail in later chapters. In studying a newly discovered protein, cell biologists usually begin by asking what is its function, where is it located, and what is its structure? To answer these questions, investigators employ three tools: the gene that encodes the protein, a mutant cell line or organism that lacks the function of the protein, and a source of the purified protein for biochemical studies. In this chapter we consider various aspects of two basic experimental strategies for obtaining all three tools (Figure 9-1).

The first strategy, often referred to as *classical genetics*, begins with isolation of a mutant that appears to be defective in some process of interest. Genetic methods then are used to

OUTLINE



▶ FIGURE 9-1 **Overview of two strategies for determining the function, location, and primary structure of proteins.** A mutant organism is the starting point for the classical genetic strategy (green arrows). The reverse strategy (orange arrows) begins with biochemical isolation of a protein or identification of a putative protein based on analysis of stored gene and protein sequences. In both strategies, the actual gene is isolated from a DNA library, a large collection of cloned DNA sequences representing an organism's genome. Once a cloned gene is isolated, it can be used to produce the encoded protein in bacterial or eukaryotic expression systems. Alternatively, a cloned gene can be inactivated by one of various techniques and used to generate mutant cells or organisms.

identify the affected gene, which subsequently is isolated from an appropriate **DNA library,** a large collection of individual DNA sequences representing all or part of an organism's genome. The isolated gene can be manipulated to produce large quantities of the protein for biochemical experiments and to design probes for studies of where and when the encoded protein is expressed in an organism. The second strategy follows essentially the same steps as the classical approach but in reverse order, beginning with isolation of an interesting protein or its identification based on analysis of an organism's genomic sequence. Once the corresponding gene has been isolated from a DNA library, the gene can be altered and then reinserted into an organism. By observing the effects of the altered gene on the organism, researchers often can infer the function of the normal protein.

An important component in both strategies for studying a protein and its biological function is isolation of the corresponding gene. Thus we discuss various techniques by which researchers can isolate, sequence, and manipulate specific regions of an organism's DNA. The extensive collections of DNA sequences that have been amassed in recent years has given birth to a new field of study called **genomics,** the molecular characterization of whole genomes and overall patterns of gene expression. Several examples of the types of information available from such genome-wide analysis also are presented.

9.1 Genetic Analysis of Mutations to Identify and Study Genes

As described in Chapter 4, the information encoded in the DNA sequence of genes specifies the sequence and therefore the structure and function of every protein molecule in a cell. The power of genetics as a tool for studying cells and organisms lies in the ability of researchers to selectively alter every copy of just one type of protein in a cell by making a change in the gene for that protein. Genetic analyses of mutants defective in a particular process can reveal (a) new genes required for the process to occur; (b) the order in which gene products act in the process; and (c) whether the proteins encoded by different genes interact with one another. Before seeing how genetic studies of this type can provide insights into the mechanism of complicated cellular or developmental process, we first explain some basic genetic terms used throughout our discussion.

The different forms, or variants, of a gene are referred to as **alleles.** Geneticists commonly refer to the numerous naturally occurring genetic variants that exist in populations, particularly human populations, as alleles. The term **mutation** usually is reserved for instances in which an allele is known to have been newly formed, such as after treatment of an experimental organism with a **mutagen,** an agent that causes a heritable change in the DNA sequence.

Strictly speaking, the particular set of alleles for all the genes carried by an individual is its **genotype.** However, this term also is used in a more restricted sense to denote just the alleles of the particular gene or genes under examination. For experimental organisms, the term **wild type** often is used to designate a standard genotype for use as a reference in breeding experiments. Thus the normal, nonmutant allele will usually be designated as the wild type. Because of the enormous naturally occurring allelic variation that exists in human populations, the term *wild type* usually denotes an allele that is present at a much higher frequency than any of the other possible alternatives.

Geneticists draw an important distinction between the *genotype* and the **phenotype** of an organism. The phenotype

refers to all the physical attributes or traits of an individual that are the consequence of a given genotype. In practice, however, the term *phenotype* often is used to denote the physical consequences that result from just the alleles that are under experimental study. Readily observable phenotypic characteristics are critical in the genetic analysis of mutations.

Recessive and Dominant Mutant Alleles Generally Have Opposite Effects on Gene Function

A fundamental genetic difference between experimental organisms is whether their cells carry a single set of chromosomes or two copies of each chromosome. The former are referred to as **haploid;** the latter, as **diploid.** Complex multicellular organisms (e.g., fruit flies, mice, humans) are diploid, whereas many simple unicellular organisms are haploid. Some organisms, notably the yeast *Saccharomyces*, can exist in either haploid or diploid states. Many cancer cells and the normal cells of some organisms, both plants and animals, carry more than two copies of each chromosome. However, our discussion of genetic techniques and analysis relates to diploid organisms, including diploid yeasts.

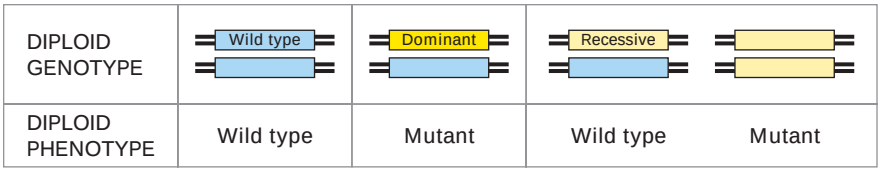
Since diploid organisms carry two copies of each gene, they may carry identical alleles, that is, be **homozygous** for a gene, or carry different alleles, that is, be **heterozygous** for a gene. A **recessive** mutant allele is defined as one in which both alleles must be mutant in order for the mutant phenotype to be observed; that is, the individual must be homozygous for the mutant allele to show the mutant phenotype. In contrast, the phenotypic consequences of a **dominant** mutant allele are observed in a heterozygous individual carrying one mutant and one wild-type allele (Figure 9-2).

Whether a mutant allele is recessive or dominant provides valuable information about the function of the affected gene and the nature of the causative mutation. Recessive alleles usually result from a mutation that inactivates the affected gene, leading to a partial or complete *loss of function*. Such recessive mutations may remove part of or the entire gene from the chromosome, disrupt expression of the gene, or alter the structure of the encoded protein, thereby altering its function. Conversely, dominant alleles are often the consequence of a mutation that causes some kind of *gain of function*. Such dominant mutations may increase the activity of the encoded protein, confer a new activity on it, or lead to its inappropriate spatial or temporal pattern of expression.

Dominant mutations in certain genes, however, are associated with a loss of function. For instance, some genes are *haplo-insufficient*, meaning that both alleles are required for normal function. Removing or inactivating a single allele in such a gene leads to a mutant phenotype. In other rare instances a dominant mutation in one allele may lead to a structural change in the protein that interferes with the function of the wild-type protein encoded by the other allele. This type of mutation, referred to as a *dominant negative*, produces a phenotype similar to that obtained from a loss-of-function mutation.

Some alleles can exhibit both recessive and dominant properties. In such cases, statements about whether an allele is dominant or recessive must specify the phenotype. For example, the allele of the hemoglobin gene in humans designated Hb^s has more than one phenotypic consequence. Individuals who are homozygous for this allele (Hb^s/Hb^s) have the debilitating disease sickle-cell anemia, but heterozygous individuals (Hb^s/Hb^a) do not have the disease. Therefore, Hb^s is *recessive* for the trait of sickle-cell disease. On the other hand, heterozygous (Hb^s/Hb^a) individuals are more resistant to malaria than homozygous (Hb^a/Hb^a) individuals, revealing that Hb^s is *dominant* for the trait of malaria resistance. ▮

A commonly used agent for inducing mutations (mutagenesis) in experimental organisms is ethylmethane sulfonate (EMS). Although this mutagen can alter DNA sequences in several ways, one of its most common effects is to chemically modify guanine bases in DNA, ultimately leading to the conversion of a G·C base pair into an A·T base pair. Such an alteration in the sequence of a gene, which involves only a single base pair, is known as a **point mutation.** A *silent* point mutation causes no change in the amino acid sequence or activity of a gene's encoded protein. However, observable phenotypic consequences due to changes in a protein's activity can arise from point mutations that result in substitution of one amino acid for another (*missense* mutation), introduction of a premature stop codon (*nonsense* mutation), or a change in the reading

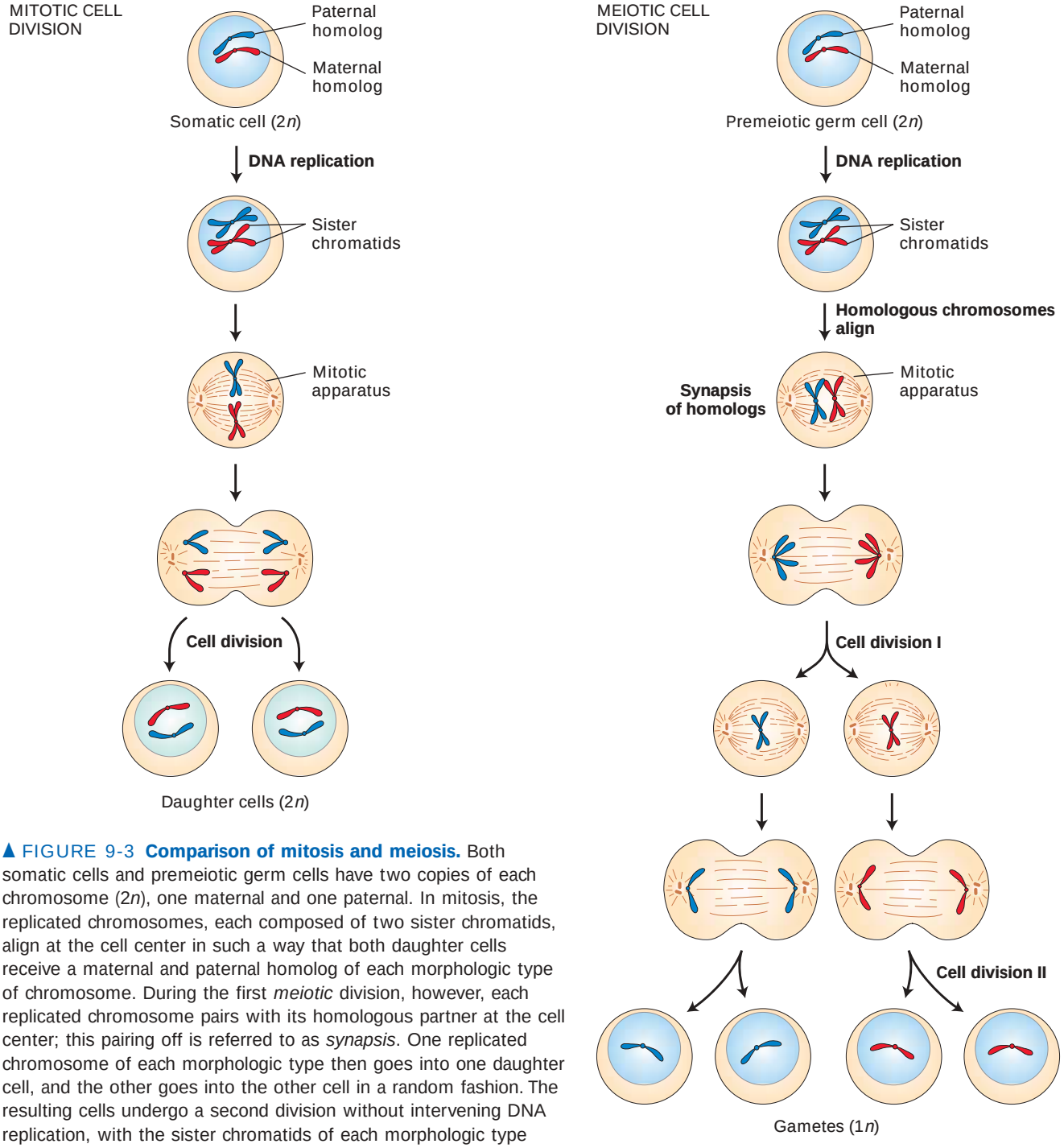


▲ FIGURE 9-2 **Effects of recessive and dominant mutant alleles on phenotype in diploid organisms.** Only one copy of a dominant allele is sufficient to produce a mutant phenotype, whereas both copies of a recessive allele must be present to cause a mutant phenotype. Recessive mutations usually cause a loss of function; dominant mutations usually cause a gain of function or an altered function.

frame of a gene (*frameshift* mutation). Because alterations in the DNA sequence leading to a decrease in protein activity are much more likely than alterations leading to an increase or qualitative change in protein activity, mutagenesis usually produces many more recessive mutations than dominant mutations.

Segregation of Mutations in Breeding Experiments Reveals Their Dominance or Recessivity

Geneticists exploit the normal life cycle of an organism to test for the dominance or recessivity of alleles. To see how



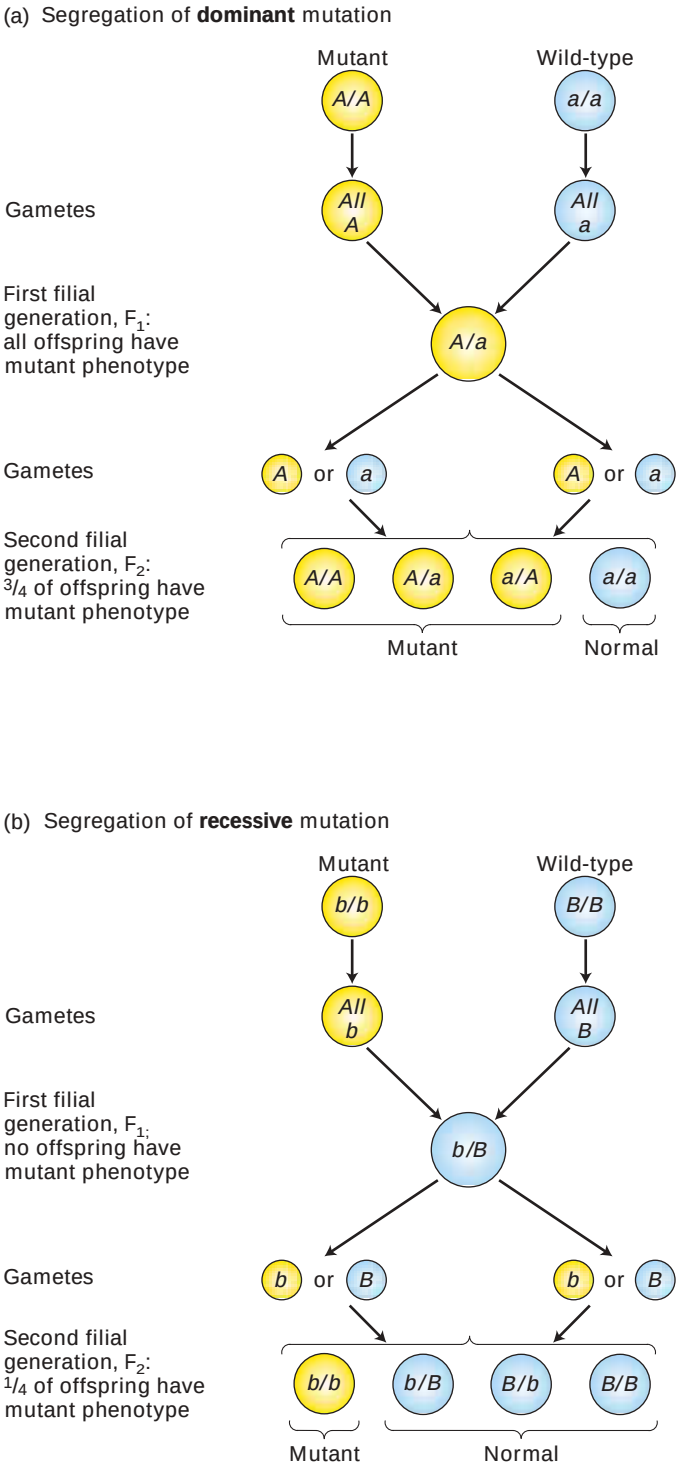
▲ FIGURE 9-3 **Comparison of mitosis and meiosis.** Both somatic cells and premeiotic germ cells have two copies of each chromosome (2*n*), one maternal and one paternal. In mitosis, the replicated chromosomes, each composed of two sister chromatids, align at the cell center in such a way that both daughter cells receive a maternal and paternal homolog of each morphologic type of chromosome. During the first *meiotic* division, however, each replicated chromosome pairs with its homologous partner at the cell center; this pairing off is referred to as *synapsis*. One replicated chromosome of each morphologic type then goes into one daughter cell, and the other goes into the other cell in a random fashion. The resulting cells undergo a second division without intervening DNA replication, with the sister chromatids of each morphologic type being apportioned to the daughter cells. Each diploid cell that undergoes meiosis produces four haploid (1*n*) cells.

MEDIA CONNECTIONS
Overview Animation: Life Cycle of a Cell

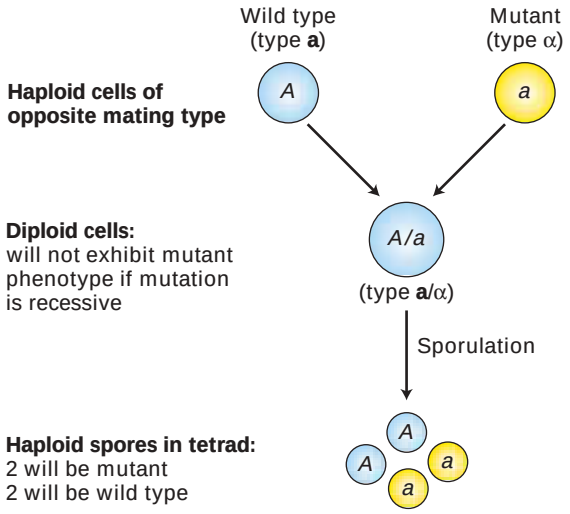
this is done, we need first to review the type of cell division that gives rise to **gametes** (sperm and egg cells in higher plants and animals). Whereas the body (somatic) cells of most multicellular organisms divide by **mitosis,** the **germ cells** that give rise to gametes undergo **meiosis.** Like somatic cells, premeiotic germ cells are diploid, containing two **homologs** of each morphologic type of chromosome. The two homologs constituting each pair of **homologous chromosomes** are descended from different parents, and thus their genes may exist in different allelic forms. Figure 9-3 depicts the major events in mitotic and meiotic cell division. In mitosis DNA replication is always followed by cell division, yielding two diploid daughter cells. In meiosis *one* round of DNA replication is followed by *two* separate cell divisions, yielding four haploid (1*n*) cells that contain only one chromosome of each homologous pair. The apportionment, or **segregation,** of the replicated homologous chromosomes to daughter cells during the first meiotic division is random; that is, maternally and paternally derived homologs segregate independently, yielding daughter cells with different mixes of paternal and maternal chromosomes.

As a way to avoid unwanted complexity, geneticists usually strive to begin breeding experiments with strains that are homozygous for the genes under examination. In such *true-breeding* strains, every individual will receive the same allele from each parent and therefore the composition of alleles will not change from one generation to the next. When a true-breeding mutant strain is mated to a true-breeding wild-type strain, all the first filial (F_1) progeny will be heterozygous (Figure 9-4). If the F_1 progeny exhibit the mutant trait, then the mutant allele is dominant; if the F_1 progeny exhibit the wild-type trait, then the mutant is recessive. Further crossing between F_1 individuals will also reveal different patterns of inheritance according to whether the mutation is dominant or recessive. When F_1 individuals that are heterozygous for a dominant allele are crossed among themselves, three-fourths of the resulting F_2 progeny will exhibit the mutant trait. In contrast, when F_1 individuals that are heterozygous for a recessive allele are crossed among themselves, only one-fourth of the resulting F_2 progeny will exhibit the mutant trait.

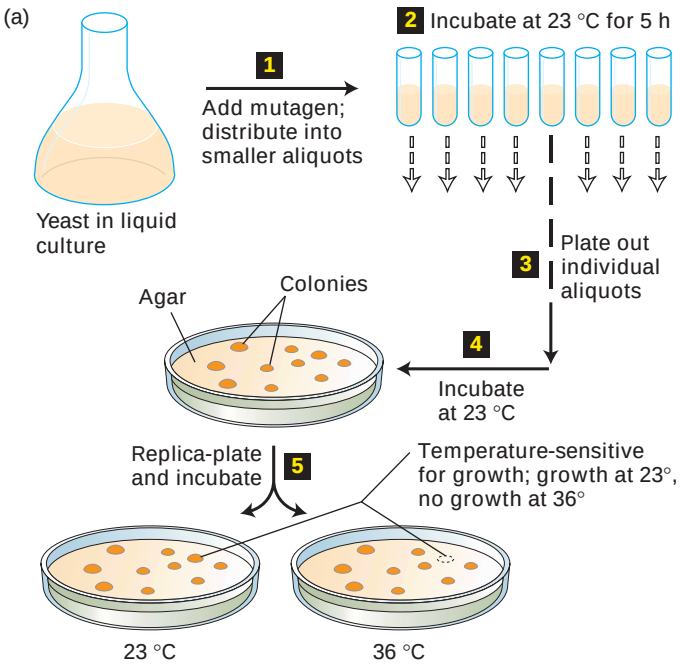
As noted earlier, the yeast *Saccharomyces,* an important experimental organism, can exist in either a haploid or a diploid state. In these unicellular eukaryotes, crosses between haploid cells can determine whether a mutant allele is dominant or recessive. Haploid yeast cells, which carry one copy of each chromosome, can be of two different mating types known as **a** and α. Haploid cells of opposite mating type can mate to produce **a**/α diploids, which carry two copies of each chromosome. If a new mutation with an observable phenotype is isolated in a haploid strain, the mutant strain can be mated to a wild-type strain of the opposite mating type to produce **a**/α diploids that are heterozygous for the mutant allele. If these diploids exhibit the mutant trait, then the mutant allele is dominant, but if the diploids appear as wild-type, then the mutant allele is recessive. When **a**/α diploids are placed under starvation conditions, the cells

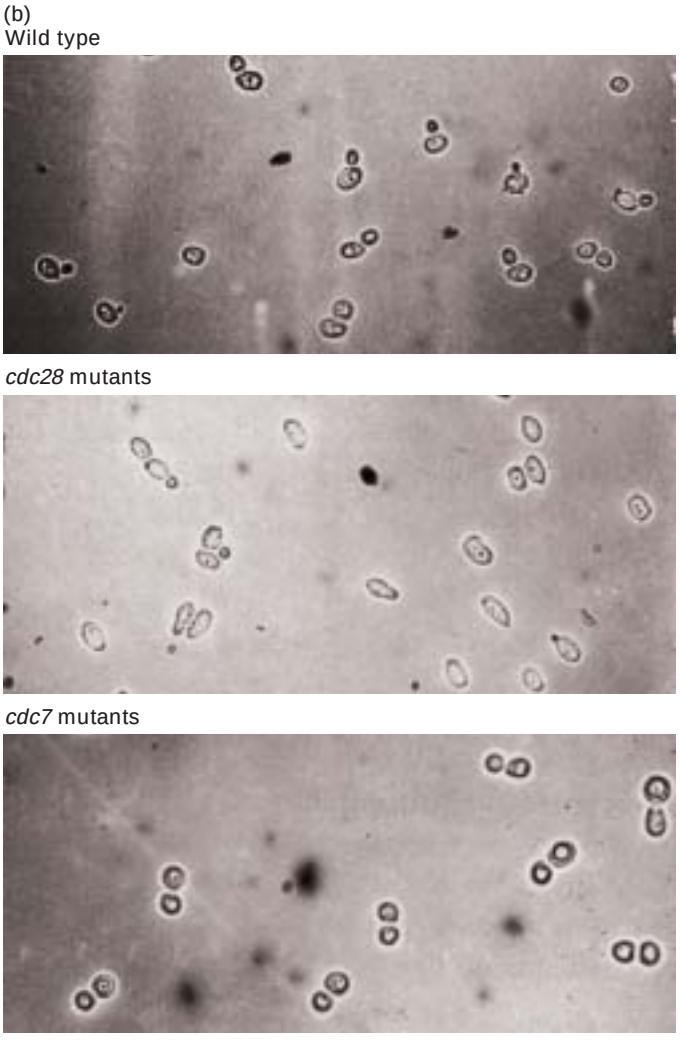


▲ FIGURE 9-4 **Segregation patterns of dominant and recessive mutations in crosses between true-breeding strains of diploid organisms.** All the offspring in the first (F_1) generation are heterozygous. If the mutant allele is dominant, the F_1 offspring will exhibit the mutant phenotype, as in part (a). If the mutant allele is recessive, the F_1 offspring will exhibit the wild-type phenotype, as in part (b). Crossing of the F_1 heterozygotes among themselves also produces different segregation ratios for dominant and recessive mutant alleles in the F_2 generation.



▲ FIGURE 9-5 **Segregation of alleles in yeast.** Haploid *Saccharomyces* cells of opposite mating type (i.e., one of mating type α and one of mating type **a**) can mate to produce an **a**/α diploid. If one haploid carries a dominant mutant allele and the other carries a recessive wild-type allele of the same gene, the resulting heterozygous diploid will express the dominant trait. Under certain conditions, a diploid cell will form a tetrad of four haploid spores. Two of the spores in the tetrad will express the recessive trait and two will express the dominant trait.





undergo meiosis, giving rise to a tetrad of four haploid spores, two of type **a** and two of type α. Sporulation of a heterozygous diploid cell yields two spores carrying the mutant allele and two carrying the wild-type allele (Figure 9-5). Under appropriate conditions, yeast spores will germinate, producing vegetative haploid strains of both mating types.

Conditional Mutations Can Be Used to Study Essential Genes in Yeast

The procedures used to identify and isolate mutants, referred to as *genetic screens*, depend on whether the experimental organism is haploid or diploid and, if the latter, whether the mutation is recessive or dominant. Genes that encode proteins essential for life are among the most interesting and important ones to study. Since phenotypic expression of mutations in essential genes leads to death of the individual, ingenious genetic screens are needed to isolate and maintain organisms with a lethal mutation.

In haploid yeast cells, essential genes can be studied through the use of *conditional mutations*. Among the most common conditional mutations are **temperature-sensitive mutations,** which can be isolated in bacteria and lower eukaryotes but not in warm-blooded eukaryotes. For instance, a mutant protein may be fully functional at one temperature (e.g., 23 °C) but completely inactive at another temperature (e.g., 36 °C), whereas the normal protein would be fully functional at both temperatures. A temperature at which the

◀ EXPERIMENTAL FIGURE 9-6 **Haploid yeasts carrying temperature-sensitive lethal mutations are maintained at permissive temperature and analyzed at nonpermissive temperature.** (a) Genetic screen for temperature-sensitive cell-division cycle (*cdc*) mutants in yeast. Yeasts that grow and form colonies at 23 °C (permissive temperature) but not at 36 °C (nonpermissive temperature) may carry a lethal mutation that blocks cell division. (b) Assay of temperature-sensitive colonies for blocks at specific stages in the cell cycle. Shown here are micrographs of wild-type yeast and two different temperature-sensitive mutants after incubation at the nonpermissive temperature for 6 h. Wild-type cells, which continue to grow, can be seen with all different sizes of buds, reflecting different stages of the cell cycle. In contrast, cells in the lower two micrographs exhibit a block at a specific stage in the cell cycle. The *cdc28* mutants arrest at a point before emergence of a new bud and therefore appear as unbudded cells. The *cdc7* mutants, which arrest just before separation of the mother cell and bud (emerging daughter cell), appear as cells with large buds. [Part (a) see L. H. Hartwell, 1967, *J. Bacteriol.* **93**:1662; part (b) from L. M. Hereford and L. H. Hartwell, 1974, *J. Mol. Biol.* **84**:445.]

mutant phenotype is observed is called *nonpermissive;* a *permissive* temperature is one at which the mutant phenotype is not observed even though the mutant allele is present. Thus mutant strains can be maintained at a permissive temperature and then subcultured at a nonpermissive temperature for analysis of the mutant phenotype.

An example of a particularly important screen for temperature-sensitive mutants in the yeast *Saccharomyces cerevisiae* comes from the studies of L. H. Hartwell and colleagues in the late 1960s and early 1970s. They set out to identify genes important in regulation of the **cell cycle** during which a cell synthesizes proteins, replicates its DNA, and then undergoes mitotic cell division, with each daughter cell receiving a copy of each chromosome. Exponential growth of a single yeast cell for 20–30 cell divisions forms a visible yeast colony on solid agar medium. Since mutants with a complete block in the cell cycle would not be able to form a colony, conditional mutants were required to study mutations that affect this basic cell process. To screen for such mutants, the researchers first identified mutagenized yeast cells that could grow normally at 23 °C but that could not form a colony when placed at 36 °C (Figure 9-6a).

Once temperature-sensitive mutants were isolated, further analysis revealed that they indeed were defective in cell division. In *S. cerevisiae,* cell division occurs through a budding process, and the size of the bud, which is easily visualized by light microscopy, indicates a cell's position in the cell cycle. Each of the mutants that could not grow at 36 °C was examined by microscopy after several hours at the nonpermissive temperature. Examination of many different temperature-sensitive mutants revealed that about 1 percent exhibited a distinct block in the cell cycle. These mutants were therefore designated *cdc* (*c*ell-*d*ivision *c*ycle) mutants. Importantly, these yeast mutants did not simply fail to grow, as they might if they carried a mutation affecting general cellular metabolism. Rather, at the nonpermissive temperature, the mutants of interest grew normally for part of the cell cycle but then arrested at a particular stage of the cell cycle, so that many cells at this stage were seen (Figure 9-6b). Most *cdc* mutations in yeast are recessive; that is, when haploid *cdc* strains are mated to wild-type haploids, the resulting heterozygous diploids are neither temperature-sensitive nor defective in cell division.

Recessive Lethal Mutations in Diploids Can Be Identified by Inbreeding and Maintained in Heterozygotes

In diploid organisms, phenotypes resulting from recessive mutations can be observed only in individuals homozygous for the mutant alleles. Since mutagenesis in a diploid organism typically changes only one allele of a gene, yielding heterozygous mutants, genetic screens must include inbreeding steps to generate progeny that are homozygous for the mutant alleles. The geneticist H. Muller developed a general and efficient procedure for carrying out such inbreeding experiments in the fruit fly *Drosophila.* Recessive lethal mutations in *Drosophila* and other diploid organisms can be maintained in heterozygous individuals and their phenotypic consequences analyzed in homozygotes.

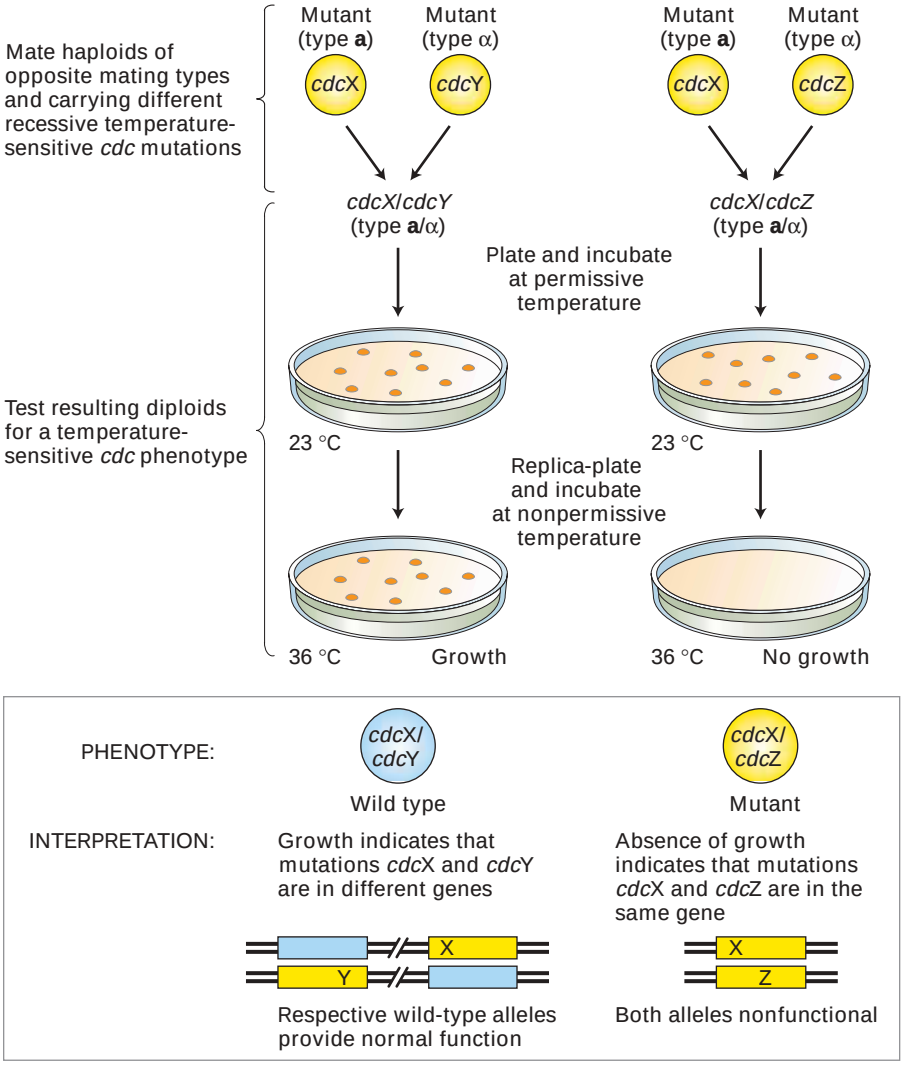
The Muller approach was used to great effect by C. Nüsslein-Volhard and E. Wieschaus, who systematically screened for recessive lethal mutations affecting embryogenesis in *Drosophila*. Dead homozygous embryos carrying recessive lethal mutations identified by this screen were examined under the microscope for specific morphological defects in the embryos. Current understanding of the molecular mechanisms underlying development of multicellular organisms is based, in large part, on the detailed picture of embryonic development revealed by characterization of these *Drosophila* mutants. We will discuss some of the fundamental discoveries based on these genetic studies in Chapter 15.

Complementation Tests Determine Whether Different Recessive Mutations Are in the Same Gene

In the genetic approach to studying a particular cellular process, researchers often isolate multiple recessive mutations that produce the same phenotype. A common test for determining whether these mutations are in the same gene or in different genes exploits the phenomenon of genetic **complementation,** that is, the restoration of the wild-type phenotype by mating of two different mutants. If two recessive mutations, *a* and *b,* are in the *same* gene, then a diploid organism heterozygous for both mutations (i.e., carrying one *a* allele and one *b* allele) will exhibit the mutant phenotype because neither allele provides a functional copy of the gene. In contrast, if mutation *a* and *b* are in *separate* genes, then heterozygotes carrying a single copy of each mutant allele

▶ EXPERIMENTAL FIGURE 9-7

Complementation analysis determines whether recessive mutations are in the same or different genes. Complementation tests in yeast are performed by mating haploid **a** and α cells carrying different recessive mutations to produce diploid cells. In the analysis of *cdc* mutations, pairs of different haploid temperature-sensitive *cdc* strains were systematically mated and the resulting diploids tested for growth at the permissive and nonpermissive temperatures. In this hypothetical example, the *cdcX* and *cdcY* mutants complement each other and thus have mutations in different genes, whereas the *cdcX* and *cdcZ* mutants have mutations in the same gene.



will not exhibit the mutant phenotype because a wild-type allele of each gene will also be present. In this case, the mutations are said to *complement* each other.

Complementation analysis of a set of mutants exhibiting the same phenotype can distinguish the individual genes in a set of functionally related genes, all of which must function to produce a given phenotypic trait. For example, the screen for *cdc* mutations in *Saccharomyces* described above yielded many recessive temperature-sensitive mutants that appeared arrested at the same cell-cycle stage. To determine how many genes were affected by these mutations, Hartwell and his colleagues performed complementation tests on all of the pair-wise combinations of *cdc* mutants following the general protocol outlined in Figure 9-7. These tests identified more than 20 different *CDC* genes. The subsequent molecular characterization of the *CDC* genes and their encoded proteins, as described in detail in Chapter 21, has provided a framework for understanding how cell division is regulated in organisms ranging from yeast to humans.

Double Mutants Are Useful in Assessing the Order in Which Proteins Function

Based on careful analysis of mutant phenotypes associated with a particular cellular process, researchers often can deduce the order in which a set of genes and their protein products function. Two general types of processes are amenable to such analysis: (a) biosynthetic pathways in which a precursor material is converted via one or more intermediates to a final product and (b) signaling pathways that regulate other processes and involve the flow of information rather than chemical intermediates.

Ordering of Biosynthetic Pathways

A simple example of the first type of process is the biosynthesis of a metabolite such as the amino acid tryptophan in bacteria. In this case, each of the enzymes required for synthesis of tryptophan catalyzes the conversion of one of the intermediates in the pathway to the next. In *E. coli*, the genes encoding these enzymes lie adjacent to one another in the genome, constituting the

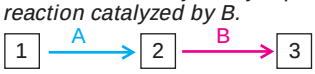
(a) Analysis of a biosynthetic pathway

A mutation in **A** accumulates intermediate 1.

A mutation in **B** accumulates intermediate 2.

PHENOTYPE OF DOUBLE MUTANT: A double mutation in A and B accumulates intermediate 1.

INTERPRETATION: *The reaction catalyzed by A precedes the reaction catalyzed by B.*



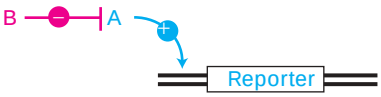
(b) Analysis of a signaling pathway

A mutation in **A** gives repressed reporter expression.

A mutation in B gives constitutive reporter expression.

PHENOTYPE OF DOUBLE MUTANT: A double mutation in A and B gives repressed reporter expression.

INTERPRETATION: *A positively regulates reporter expression and is negatively regulated by B.*



PHENOTYPE OF DOUBLE MUTANT: A double mutation in A and B gives constitutive reporter expression.

INTERPRETATION: *B negatively regulates reporter expression and is negatively regulated by A.*

▶ EXPERIMENTAL FIGURE 9-8 **Analysis of double mutants often can order the steps in biosynthetic or signaling pathways.** When mutations in two different genes affect the same cellular process but have distinctly different phenotypes, the phenotype of the double mutant can often reveal the order in which the two genes must function. (a) In the case of mutations that affect the same biosynthetic pathway, a double mutant will accumulate the intermediate immediately preceding the step catalyzed by the protein that acts earlier in the wild-type organism. (b) Double-mutant analysis of a signaling pathway is possible if two mutations have opposite effects on expression of a reporter gene. In this case, the observed phenotype of the double mutant provides information about the order in which the proteins act and whether they are positive or negative regulators.

trp operon (see Figure 4-12a). The order of action of the different genes for these enzymes, hence the order of the biochemical reactions in the pathway, initially was deduced from the types of intermediate compounds that accumulated in each mutant. In the case of complex synthetic pathways, however, phenotypic analysis of mutants defective in a single step may give ambiguous results that do not permit conclusive ordering of the steps. Double mutants defective in two steps in the pathway are particularly useful in ordering such pathways (Figure 9-8a).

In Chapter 17 we discuss the classic use of the double-mutant strategy to help elucidate the secretory pathway. In this pathway proteins to be secreted from the cell move from their site of synthesis on the rough endoplasmic reticulum (ER) to the Golgi complex, then to secretory vesicles, and finally to the cell surface.

Ordering of Signaling Pathways

As we learn in later chapters, expression of many eukaryotic genes is regulated by signaling pathways that are initiated by extracellular hormones, growth factors, or other signals. Such signaling pathways may include numerous components, and double-mutant analysis often can provide insight into the functions and interactions of these components. The only prerequisite for obtaining useful information from this type of analysis is that the two mutations must have opposite effects on the output of the same regulated pathway. Most commonly, one mutation represses expression of a particular reporter gene even when the signal is present, while another mutation results in reporter gene expression even when the signal is absent (i.e., constitutive expression). As illustrated in Figure 9-8b, two simple regulatory mechanisms are consistent with such single mutants, but the double-mutant phenotype can distinguish between them. This general approach has enabled geneticists to delineate many of the key steps in a variety of different regulatory pathways, setting the stage for more specific biochemical assays.

Genetic Suppression and Synthetic Lethality Can Reveal Interacting or Redundant Proteins

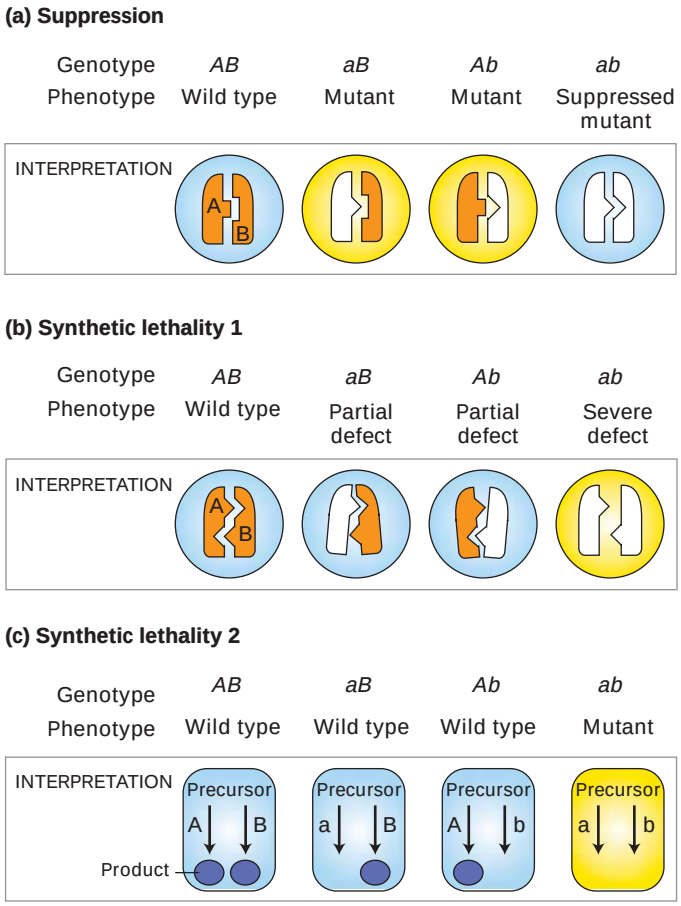
Two other types of genetic analysis can provide additional clues about how proteins that function in the same cellular process may interact with one another in the living cell. Both of these methods, which are applicable in many experimental organisms, involve the use of double mutants in which the phenotypic effects of one mutation are changed by the presence of a second mutation.

Suppressor Mutations

The first type of analysis is based on *genetic suppression*. To understand this phenomenon, suppose that point mutations lead to structural changes in one protein (A) that disrupt its ability to associate with another protein (B) involved in the same cellular process. Similarly, mutations in protein B lead to small structural changes that inhibit its ability to interact with protein A. Assume, furthermore, that the normal functioning of proteins A and B depends on their interacting. In theory, a specific structural change in protein A might be suppressed by compensatory changes in protein B, allowing the mutant proteins to interact. In the rare cases in which such **suppressor mutations** occur, strains carrying both mutant

alleles would be normal, whereas strains carrying only one or the other mutant allele would have a mutant phenotype (Figure 9-9a).

The observation of genetic suppression in yeast strains carrying a mutant actin allele (*act1-1*) and a second mutation (*sac6*) in another gene provided early evidence for a direct interaction in vivo between the proteins encoded by the two genes. Later biochemical studies showed that these two proteins—Act1 and Sac6—do indeed interact in the construction of functional actin structures within the cell.



▲ EXPERIMENTAL FIGURE 9-9 **Mutations that result in genetic suppression or synthetic lethality reveal interacting or redundant proteins.** (a) Observation that double mutants with two defective proteins (A and B) have a wild-type phenotype but that single mutants give a mutant phenotype indicates that the function of each protein depends on interaction with the other. (b) Observation that double mutants have a more severe phenotypic defect than single mutants also is evidence that two proteins (e.g., subunits of a heterodimer) must interact to function normally. (c) Observation that a double mutant is nonviable but that the corresponding single mutants have the wild-type phenotype indicates that two proteins function in redundant pathways to produce an essential product.

Synthetic Lethal Mutations

Another phenomenon, called *synthetic lethality,* produces a phenotypic effect opposite to that of suppression. In this case, the deleterious effect of one mutation is greatly exacerbated (rather than suppressed) by a second mutation in the same or a related gene. One situation in which such **synthetic lethal mutations** can occur is illustrated in Figure 9-9b. In this example, a heterodimeric protein is partially, but not completely, inactivated by mutations in either one of the nonidentical subunits. However, in double mutants carrying specific mutations in the genes encoding both subunits, little interaction between subunits occurs, resulting in severe phenotypic effects.

Synthetic lethal mutations also can reveal nonessential genes whose encoded proteins function in redundant pathways for producing an essential cell component. As depicted in Figure 9-9c, if either pathway alone is inactivated by a mutation, the other pathway will be able to supply the needed product. However, if both pathways are inactivated at the same time, the essential product cannot be synthesized, and the double mutants will be nonviable.

KEY CONCEPTS OF SECTION 9.1

Genetic Analysis of Mutations to Identify and Study Genes

- Diploid organisms carry two copies (alleles) of each gene, whereas haploid organisms carry only one copy.
- Recessive mutations lead to a loss of function, which is masked if a normal allele of the gene is present. For the mutant phenotype to occur, both alleles must carry the mutation.
- Dominant mutations lead to a mutant phenotype in the presence of a normal allele of the gene. The phenotypes associated with dominant mutations often represent a gain of function but in the case of some genes result from a loss of function.
- In meiosis, a diploid cell undergoes one DNA replication and two cell divisions, yielding four haploid cells in which maternal and paternal alleles are randomly assorted (see Figure 9-3).
- Dominant and recessive mutations exhibit characteristic segregation patterns in genetic crosses (see Figure 9-4).
- In haploid yeast, temperature-sensitive mutations are particularly useful for identifying and studying genes essential to survival.
- The number of functionally related genes involved in a process can be defined by complementation analysis (see Figure 9-7).
- The order in which genes function in either a biosynthetic or a signaling pathway can be deduced from the phenotype of double mutants defective in two steps in the affected process.

■ Functionally significant interactions between proteins can be deduced from the phenotypic effects of allele-specific suppressor mutations or synthetic lethal mutations.

9.2 DNA Cloning by Recombinant DNA Methods

Detailed studies of the structure and function of a gene at the molecular level require large quantities of the individual gene in pure form. A variety of techniques, often referred to as *recombinant DNA technology*, are used in **DNA cloning,** which permits researchers to prepare large numbers of identical DNA molecules. **Recombinant DNA** is simply any DNA molecule composed of sequences derived from different sources.

The key to cloning a DNA fragment of interest is to link it to a **vector** DNA molecule, which can replicate within a host cell. After a single recombinant DNA molecule, composed of a vector plus an inserted DNA fragment, is introduced into a host cell, the inserted DNA is replicated along with the vector, generating a large number of identical DNA molecules. The basic scheme can be summarized as follows:

Vector + DNA fragment
↓
Recombinant DNA
↓
Replication of recombinant DNA within host cells
↓
Isolation, sequencing, and manipulation of purified DNA fragment

Although investigators have devised numerous experimental variations, this flow diagram indicates the essential steps in DNA cloning. In this section, we cover the steps in this basic scheme, focusing on the two types of vectors most commonly used in *E. coli* host cells: plasmid vectors, which replicate along with their host cells, and bacteriophage λ vectors, which replicate as lytic viruses, killing the host cell and packaging their DNA into virions. We discuss the characterization and various uses of cloned DNA fragments in subsequent sections.

Restriction Enzymes and DNA Ligases Allow Insertion of DNA Fragments into Cloning Vectors

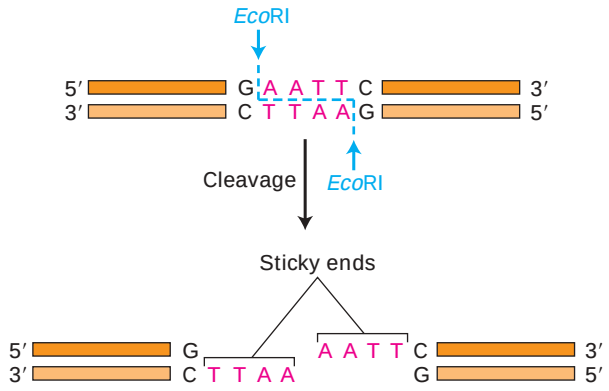
A major objective of DNA cloning is to obtain discrete, small regions of an organism's DNA that constitute specific genes. In addition, only relatively small DNA molecules can be cloned in any of the available vectors. For these reasons, the very long DNA molecules that compose an organism's genome must be cleaved into fragments that can be inserted into the vector DNA. Two types of enzymes—**restriction enzymes** and **DNA ligases**—facilitate production of such recombinant DNA molecules.

Cutting DNA Molecules into Small Fragments Restriction enzymes are endonucleases produced by bacteria that typically recognize specific 4- to 8-bp sequences, called *restriction sites*, and then cleave both DNA strands at this site. Restriction sites commonly are short *palindromic* sequences; that is, the restriction-site sequence is the same on each DNA strand when read in the 5′ → 3′ direction (Figure 9-10).

For each restriction enzyme, bacteria also produce a *modification enzyme*, which protects a bacterium's own DNA from cleavage by modifying it at or near each potential cleavage site. The modification enzyme adds a methyl group to one or two bases, usually within the restriction site. When a methyl group is present there, the restriction endonuclease is prevented from cutting the DNA. Together with the restriction endonuclease, the methylating enzyme forms a restriction-modification system that protects the host DNA while it destroys incoming foreign DNA (e.g., bacteriophage DNA or DNA taken up during transformation) by cleaving it at all the restriction sites in the DNA.

Many restriction enzymes make staggered cuts in the two DNA strands at their recognition site, generating fragments that have a single-stranded "tail" at both ends (see Figure 9-10). The tails on the fragments generated at a given restriction site are complementary to those on all other fragments generated by the same restriction enzyme. At room temperature, these single-stranded regions, often called "sticky ends," can transiently base-pair with those on other DNA fragments generated with the same restriction enzyme. A few restriction enzymes, such as *Alu*I and *Sma*I, cleave both DNA strands at the same point within the restriction site, generating fragments with "blunt" (flush) ends in which all the nucleotides at the fragment ends are base-paired to nucleotides in the complementary strand.

The DNA isolated from an individual organism has a specific sequence, which purely by chance will contain a specific



▲ FIGURE 9-10 **Cleavage of DNA by the restriction enzyme *Eco*RI.** This restriction enzyme from *E. coli* makes staggered cuts at the specific 6-bp inverted repeat (palindromic) sequence shown, yielding fragments with single-stranded, complementary "sticky" ends. Many other restriction enzymes also produce fragments with sticky ends.

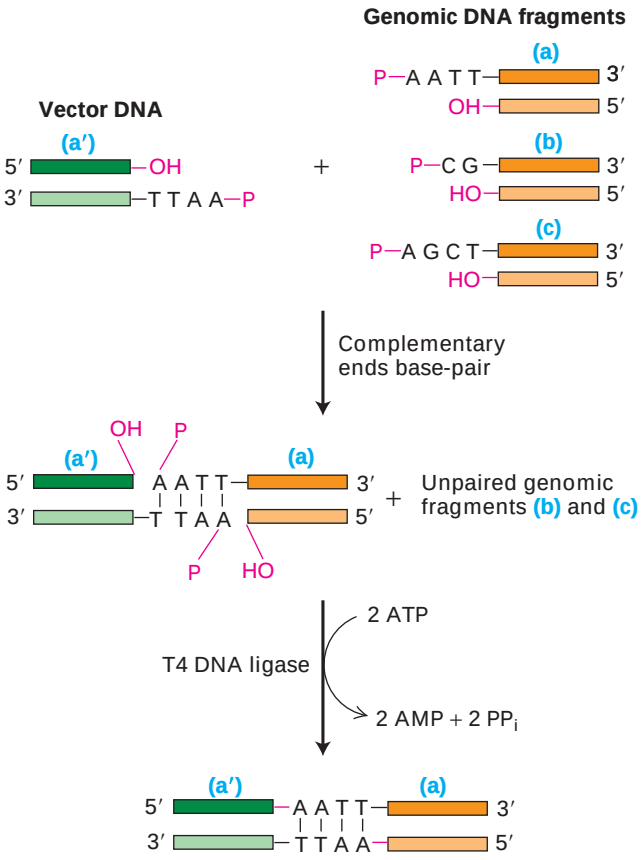
TABLE 9-1 Selected Restriction Enzymes and Their Recognition Sequences

Enzyme	Source Microorganism	Recognition Site*	Ends Produced
BamHI	*Bacillus amyloliquefaciens*	↓ -G-G-A-T-C-C- -C-C-T-A-G-G- ↑	Sticky
*Eco*RI	*Escherichia coli*	↓ -G-A-A-T-T-C- -C-T-T-A-A-G- ↑	Sticky
*Hin*dIII	*Haemophilus influenzae*	↓ -A-A-G-C-T-T- -T-T-C-G-A-A- ↑	Sticky
*Kpn*I	*Klebsiella pneumonia*	↓ -G-G-T-A-C-C- -C-C-A-T-G-G- ↑	Sticky
*Pst*I	*Providencia stuartii*	↓ -C-T-G-C-A-G- -G-A-C-G-T-C- ↑	Sticky
*Sac*I	*Streptomyces achromogenes*	↓ -G-A-G-C-T-C- -C-T-C-G-A-G- ↑	Sticky
*Sal*I	*Streptomyces albue*	↓ -G-T-C-G-A-C- -C-A-G-C-T-G- ↑	Sticky
*Sma*I	*Serratia marcescens*	↓ -C-C-C-G-G-G- -G-G-G-C-C-C- ↑	Blunt
*Sph*I	*Streptomyces phaeochromogenes*	↓ -G-C-A-T-G-C- -C-G-T-A-C-G- ↑	Sticky
*Xba*I	*Xanthomonas badrii*	↓ -T-C-T-A-G-A- -A-G-A-T-C-T- ↑	Sticky

*These recognition sequences are included in a common polylinker sequence (see Figure 9-12).

set of restriction sites. Thus a given restriction enzyme will cut the DNA from a particular source into a reproducible set of fragments called **restriction fragments.** Restriction enzymes have been purified from several hundred different species of bacteria, allowing DNA molecules to be cut at a large number of different sequences corresponding to the recognition sites of these enzymes (Table 9-1).

Inserting DNA Fragments into Vectors DNA fragments with either sticky ends or blunt ends can be inserted into vector DNA with the aid of DNA ligases. During normal DNA replication, DNA ligase catalyzes the end-to-end joining (ligation) of short fragments of DNA, called *Okazaki fragments*. For purposes of DNA cloning, purified DNA ligase is used to covalently join the ends of a restriction fragment and vector DNA that have complementary ends (Figure 9-11). The vector DNA and restriction fragment are covalently ligated together through the standard $3' \rightarrow 5'$ phosphodiester bonds of DNA. In addition to ligating complementary sticky ends, the DNA ligase from bacteriophage T4 can ligate any two



▲ FIGURE 9-11 **Ligation of restriction fragments with complementary sticky ends.** In this example, vector DNA cut with *Eco*RI is mixed with a sample containing restriction fragments produced by cleaving genomic DNA with several different restriction enzymes. The short base sequences composing the sticky ends of each fragment type are shown. The sticky end on the cut vector DNA (a′) base-pairs only with the complementary sticky ends on the *Eco*RI fragment (a) in the genomic sample. The adjacent 3′-hydroxyl and 5′-phosphate groups (red) on the base-paired fragments then are covalently joined (ligated) by T4 DNA ligase.

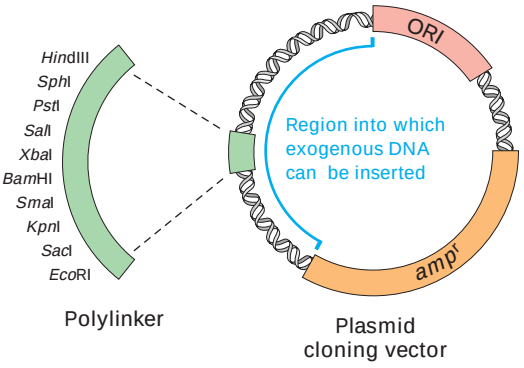
blunt DNA ends. However, blunt-end ligation is inherently inefficient and requires a higher concentration of both DNA and DNA ligase than for ligation of sticky ends.

E. coli Plasmid Vectors Are Suitable for Cloning Isolated DNA Fragments

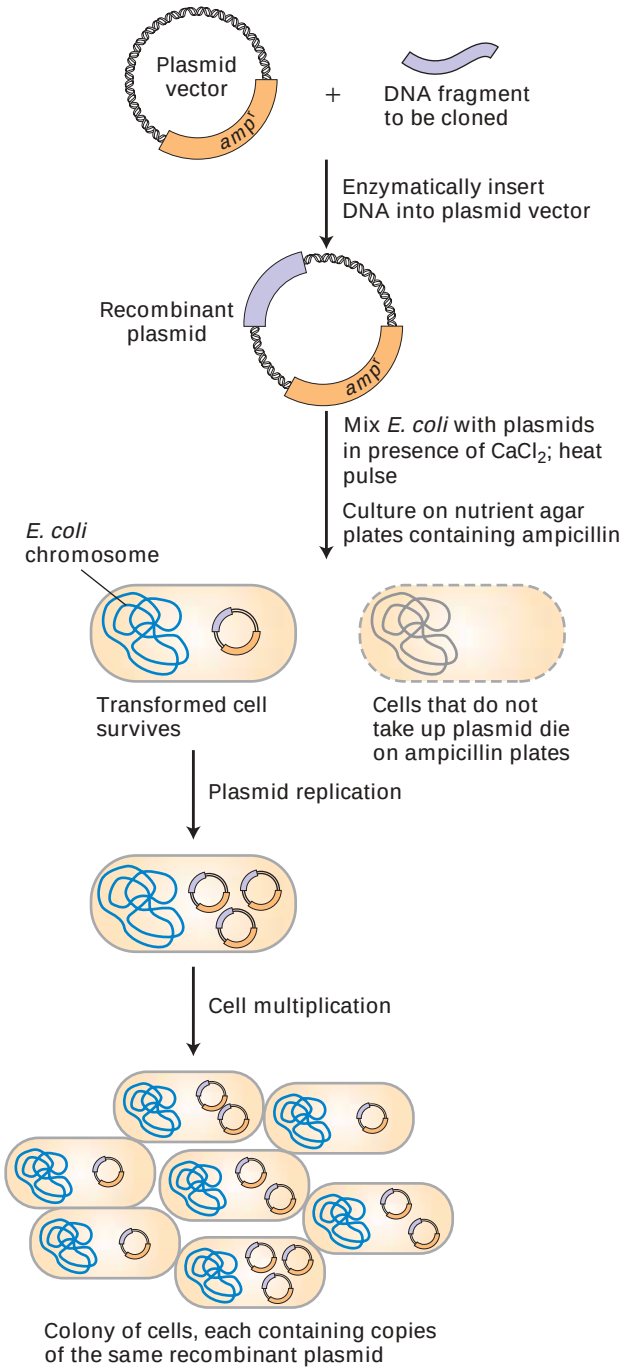
Plasmids are circular, double-stranded DNA (dsDNA) molecules that are separate from a cell's chromosomal DNA. These extrachromosomal DNAs, which occur naturally in bacteria and in lower eukaryotic cells (e.g., yeast), exist in a parasitic or symbiotic relationship with their host cell. Like the host-cell chromosomal DNA, plasmid DNA is duplicated before every cell division. During cell division, copies of the plasmid DNA segregate to each daughter cell, assuring continued propagation of the plasmid through successive generations of the host cell.

The plasmids most commonly used in recombinant DNA technology are those that replicate in *E. coli*. Investigators have engineered these plasmids to optimize their use as vectors in DNA cloning. For instance, removal of unneeded portions from naturally occurring *E. coli* plasmids yields plasmid vectors, ≈1.2–3 kb in circumferential length, that contain three regions essential for DNA cloning: a replication origin; a marker that permits selection, usually a drug-resistance gene; and a region in which exogenous DNA fragments can be inserted (Figure 9-12). Host-cell enzymes replicate a plasmid beginning at the replication origin (ORI), a specific DNA sequence of 50–100 base pairs. Once DNA replication is initiated at the ORI, it continues around the circular plasmid regardless of its nucleotide sequence. Thus any DNA sequence inserted into such a plasmid is replicated along with the rest of the plasmid DNA.

Figure 9-13 outlines the general procedure for cloning a DNA fragment using *E. coli* plasmid vectors. When *E. coli* cells are mixed with recombinant vector DNA under certain conditions, a small fraction of the cells will take up the plasmid DNA, a process known as **transformation.** Typically, 1 cell in about 10,000 incorporates a *single* plasmid DNA molecule and thus becomes transformed. After plasmid vectors are incubated with *E. coli*, those cells that take up the plasmid can be easily selected from the much larger number of cells. For instance, if the plasmid carries a gene that confers resistance to the antibiotic ampicillin, transformed cells



▲ FIGURE 9-12 **Basic components of a plasmid cloning vector that can replicate within an *E. coli* cell.** Plasmid vectors contain a selectable gene such as *amp*^{r}, which encodes the enzyme β-lactamase and confers resistance to ampicillin. Exogenous DNA can be inserted into the bracketed region without disturbing the ability of the plasmid to replicate or express the *amp*^{r} gene. Plasmid vectors also contain a replication origin (ORI) sequence where DNA replication is initiated by host-cell enzymes. Inclusion of a synthetic polylinker containing the recognition sequences for several different restriction enzymes increases the versatility of a plasmid vector. The vector is designed so that each site in the polylinker is unique on the plasmid.



▲ EXPERIMENTAL FIGURE 9-13 **DNA cloning in a plasmid vector permits amplification of a DNA fragment.** A fragment of DNA to be cloned is first inserted into a plasmid vector containing an ampicillin-resistance gene (*amp*^{r}), such as that shown in Figure 9-12. Only the few cells transformed by incorporation of a plasmid molecule will survive on ampicillin-containing medium. In transformed cells, the plasmid DNA replicates and segregates into daughter cells, resulting in formation of an ampicillin-resistant colony.

MEDIA CONNECTIONS
Technique Animation: Plasmid Cloning

can be selected by growing them in an ampicillin-containing medium.

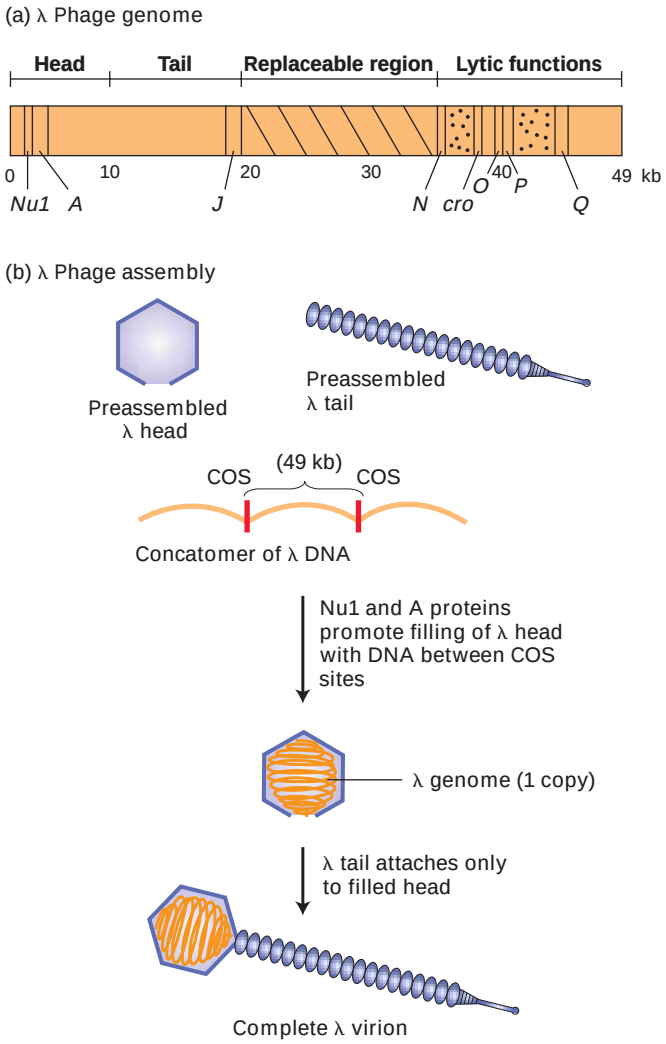
DNA fragments from a few base pairs up to ≈20 kb commonly are inserted into plasmid vectors. If special precautions are taken to avoid manipulations that might mechanically break DNA, even longer DNA fragments can be inserted into a plasmid vector. When a recombinant plasmid with an inserted DNA fragment transforms an *E. coli* cell, all the antibiotic-resistant progeny cells that arise from the initial transformed cell will contain plasmids with the same inserted DNA. The inserted DNA is replicated along with the rest of the plasmid DNA and segregates to daughter cells as the colony grows. In this way, the initial fragment of DNA is replicated in the colony of cells into a large number of identical copies. Since all the cells in a colony arise from a single transformed parental cell, they constitute a **clone** of cells, and the initial fragment of DNA inserted into the parental plasmid is referred to as *cloned DNA* or a *DNA clone*.

The versatility of an *E. coli* plasmid vector is increased by incorporating into it a *polylinker*, a synthetically generated sequence containing one copy of several different restriction sites that are not present elsewhere in the plasmid sequence (see Figure 9-12). When such a vector is treated with a restriction enzyme that recognizes a restriction site in the polylinker, the vector is cut only once within the polylinker. Subsequently any DNA fragment of appropriate length produced with the same restriction enzyme can be inserted into the cut plasmid with DNA ligase. Plasmids containing a polylinker permit a researcher to clone DNA fragments generated with different restriction enzymes using the same plasmid vector, which simplifies experimental procedures.

Bacteriophage λ Vectors Permit Efficient Construction of Large DNA Libraries

Vectors constructed from bacteriophage λ are about a thousand times more efficient than plasmid vectors in cloning large numbers of DNA fragments. For this reason, phage λ vectors have been widely used to generate DNA libraries, comprehensive collections of DNA fragments representing the genome or expressed mRNAs of an organism. Two factors account for the greater efficiency of phage λ as a cloning vector: infection of *E. coli* host cells by λ virions occurs at about a thousandfold greater frequency than transformation by plasmids, and many more λ clones than transformed colonies can be grown and detected on a single culture plate.

When a λ virion infects an *E. coli* cell, it can undergo a cycle of lytic growth during which the phage DNA is replicated and assembled into more than 100 complete progeny phage, which are released when the infected cell lyses (see Figure 4-40). If a sample of λ phage is placed on a lawn of *E. coli* growing on a petri plate, each virion will infect a single cell. The ensuing rounds of phage growth will give rise to a visible cleared region, called a *plaque*, where the cells have been lysed and phage particles released (see Figure 4-39).



▲ FIGURE 9-14 **The bacteriophage λ genome and packaging of bacteriophage λ DNA.** (a) Simplified map of the λ phage genome. There are about 60 genes in the λ genome, only a few of which are shown in this diagram. Genes encoding proteins required for assembly of the head and tail are located at the left end; those encoding additional proteins required for the lytic cycle, at the right end. Some regions of the genome can be replaced by exogenous DNA (diagonal lines) or deleted (dotted) without affecting the ability of λ phage to infect host cells and assemble new virions. Up to ≈25 kb of exogenous DNA can be stably inserted between the *J* and *N* genes. (b) In vivo assembly of λ virions. Heads and tails are formed from multiple copies of several different λ proteins. During the late stage of λ infection, long DNA molecules called *concatomers* are formed; these multimeric molecules consist of multiple copies of the 49-kb λ genome linked end to end and separated by COS sites (red), protein-binding nucleotide sequences that occur once in each copy of the λ genome. Binding of λ head proteins Nu1 and A to COS sites promotes insertion of the DNA segment between two adjacent COS sites into an empty head. After the heads are filled with DNA, assembled λ tails are attached, producing complete λ virions capable of infecting *E. coli* cells.

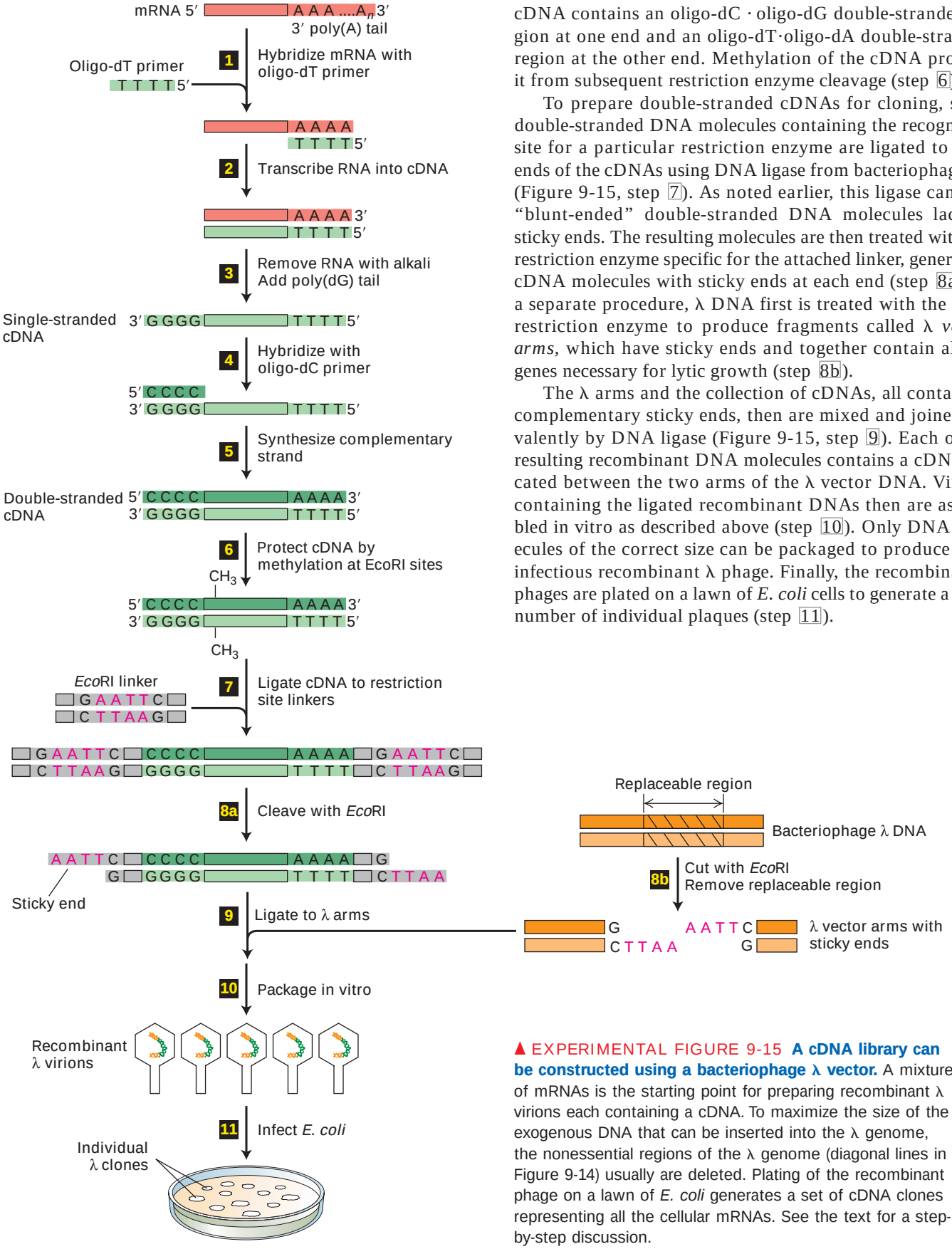
A λ virion consists of a head, which contains the phage DNA genome, and a tail, which functions in infecting *E. coli* host cells. The λ genes encoding the head and tail proteins, as well as various proteins involved in phage DNA replication and cell lysis, are grouped in discrete regions of the ≈50-kb viral genome (Figure 9-14a). The central region of the λ genome, however, contains genes that are not essential for the lytic pathway. Removing this region and replacing it with a foreign DNA fragment up to ≈25 kb long yields a recombinant DNA that can be packaged in vitro to form phage capable of replicating and forming plaques on a lawn of *E. coli* host cells. In vitro packaging of recombinant λ DNA, which mimics the in vivo assembly process, requires preassembled heads and tails as well as two viral proteins (Figure 9-14b).

It is technically feasible to use λ phage cloning vectors to generate a *genomic library*, that is, a collection of λ clones that collectively represent all the DNA sequences in the genome of a particular organism. However, such genomic libraries for higher eukaryotes present certain experimental difficulties. First, the genes from such organisms usually contain extensive intron sequences and therefore are too large to be inserted intact into λ phage vectors. As a result, the sequences of individual genes are broken apart and carried in more than one λ clone (this is also true for plasmid clones). Moreover, the presence of introns and long intergenic regions in genomic DNA often makes it difficult to identify the important parts of a gene that actually encode protein sequences. Thus for many studies, cellular mRNAs, which lack the noncoding regions present in genomic DNA, are a more useful starting material for generating a DNA library. In this approach, DNA copies of mRNAs, called **complementary DNAs (cDNAs),** are synthesized and cloned in phage vectors. A large collection of the resulting cDNA clones, representing all the mRNAs expressed in a cell type, is called a *cDNA library.*

cDNAs Prepared by Reverse Transcription of Cellular mRNAs Can Be Cloned to Generate cDNA Libraries

The first step in preparing a cDNA library is to isolate the total mRNA from the cell type or tissue of interest. Because of their poly(A) tails, mRNAs are easily separated from the much more prevalent rRNAs and tRNAs present in a cell extract by use of a column to which short strings of thymidylate (oligo-dTs) are linked to the matrix.

The general procedure for preparing a λ phage cDNA library from a mixture of cellular mRNAs is outlined in Figure 9-15. The enzyme **reverse transcriptase,** which is found in retroviruses, is used to synthesize a strand of DNA complementary to each mRNA molecule, starting from an oligo-dT primer (steps 1 and 2). The resulting cDNA-mRNA hybrid molecules are converted in several steps to double-stranded cDNA molecules corresponding to all the mRNA molecules in the original preparation (steps 3–5). Each double-stranded



cDNA contains an oligo-dC · oligo-dG double-stranded region at one end and an oligo-dT·oligo-dA double-stranded region at the other end. Methylation of the cDNA protects it from subsequent restriction enzyme cleavage (step 6).

To prepare double-stranded cDNAs for cloning, short double-stranded DNA molecules containing the recognition site for a particular restriction enzyme are ligated to both ends of the cDNAs using DNA ligase from bacteriophage T4 (Figure 9-15, step 7). As noted earlier, this ligase can join "blunt-ended" double-stranded DNA molecules lacking sticky ends. The resulting molecules are then treated with the restriction enzyme specific for the attached linker, generating cDNA molecules with sticky ends at each end (step 8a). In a separate procedure, λ DNA first is treated with the same restriction enzyme to produce fragments called λ *vector arms*, which have sticky ends and together contain all the genes necessary for lytic growth (step 8b).

The λ arms and the collection of cDNAs, all containing complementary sticky ends, then are mixed and joined covalently by DNA ligase (Figure 9-15, step 9). Each of the resulting recombinant DNA molecules contains a cDNA located between the two arms of the λ vector DNA. Virions containing the ligated recombinant DNAs then are assembled in vitro as described above (step 10). Only DNA molecules of the correct size can be packaged to produce fully infectious recombinant λ phage. Finally, the recombinant λ phages are plated on a lawn of *E. coli* cells to generate a large number of individual plaques (step 11).

▲ EXPERIMENTAL FIGURE 9-15 **A cDNA library can be constructed using a bacteriophage λ vector.** A mixture of mRNAs is the starting point for preparing recombinant λ virions each containing a cDNA. To maximize the size of the exogenous DNA that can be inserted into the λ genome, the nonessential regions of the λ genome (diagonal lines in Figure 9-14) usually are deleted. Plating of the recombinant phage on a lawn of *E. coli* generates a set of cDNA clones representing all the cellular mRNAs. See the text for a step-by-step discussion.

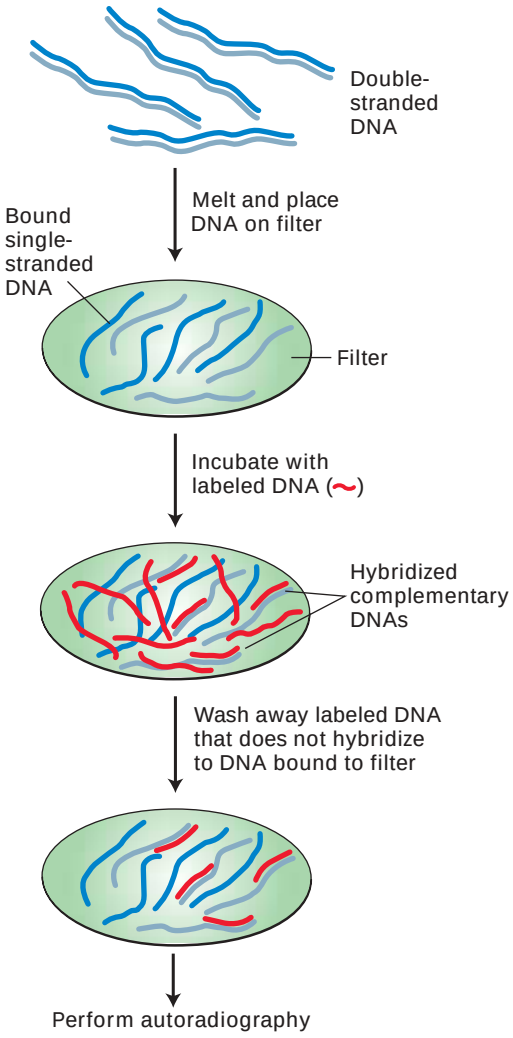
Since each plaque arises from a single recombinant phage, all the progeny λ phages that develop are genetically identical and constitute a clone carrying a cDNA derived from a single mRNA; collectively they constitute a λ cDNA library. One feature of cDNA libraries arises because different genes are transcribed at very different rates. As a result, cDNA clones corresponding to rapidly transcribed genes will be represented many times in a cDNA library, whereas cDNAs corresponding to slowly transcribed genes will be extremely rare or not present at all. This property is advantageous if an investigator is interested in a gene that is transcribed at a high rate in a particular cell type. In this case, a cDNA library prepared from mRNAs expressed in that cell type will be enriched in the cDNA of interest, facilitating screening of the library for λ clones carrying that cDNA. However, to have a reasonable chance of including clones corresponding to slowly transcribed genes, mammalian cDNA libraries must contain 10^6–10^7 individual recombinant λ phage clones.

DNA Libraries Can Be Screened by Hybridization to an Oligonucleotide Probe

Both genomic and cDNA libraries of various organisms contain hundreds of thousands to upwards of a million individual clones in the case of higher eukaryotes. Two general approaches are available for screening libraries to identify clones carrying a gene or other DNA region of interest: (1) detection with oligonucleotide **probes** that bind to the clone of interest and (2) detection based on expression of the encoded protein. Here we describe the first method; an example of the second method is presented in the next section.

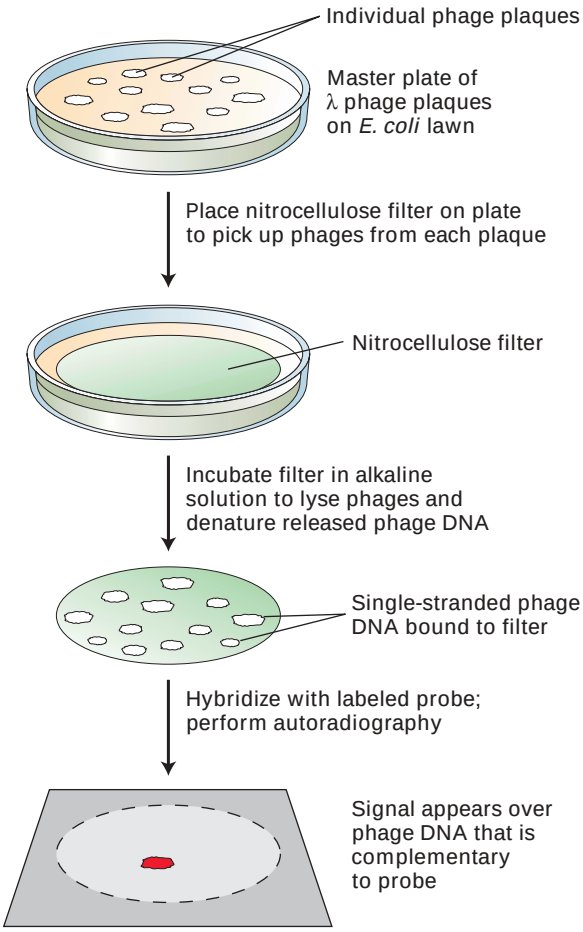
The basis for screening with oligonucleotide probes is **hybridization,** the ability of complementary single-stranded DNA or RNA molecules to associate (hybridize) specifically with each other via base pairing. As discussed in Chapter 4, double-stranded (duplex) DNA can be denatured (melted) into single strands by heating in a dilute salt solution. If the temperature then is lowered and the ion concentration raised, complementary single strands will reassociate (hybridize) into duplexes. In a mixture of nucleic acids, only complementary single strands (or strands containing complementary regions) will reassociate; moreover, the extent of their reassociation is virtually unaffected by the presence of noncomplementary strands.

In the *membrane-hybridization assay* outlined in Figure 9-16, a single-stranded nucleic acid probe is used to detect those DNA fragments in a mixture that are complementary to the probe. The DNA sample first is denatured and the single strands attached to a solid support, commonly a nitrocellulose filter or treated nylon membrane. The membrane is then incubated in a solution containing a radioactively labeled probe. Under hybridization conditions (near neutral pH, 40–65 °C, 0.3–0.6 M NaCl), this labeled probe hybridizes to any complementary nucleic acid strands bound to the membrane. Any excess probe that does not hybridize is washed away, and the labeled hybrids are detected by autoradiography of the filter.



▲ EXPERIMENTAL FIGURE 9-16 **Membrane-hybridization assay detects nucleic acids complementary to an oligonucleotide probe.** This assay can be used to detect both DNA and RNA, and the radiolabeled complementary probe can be either DNA or RNA.

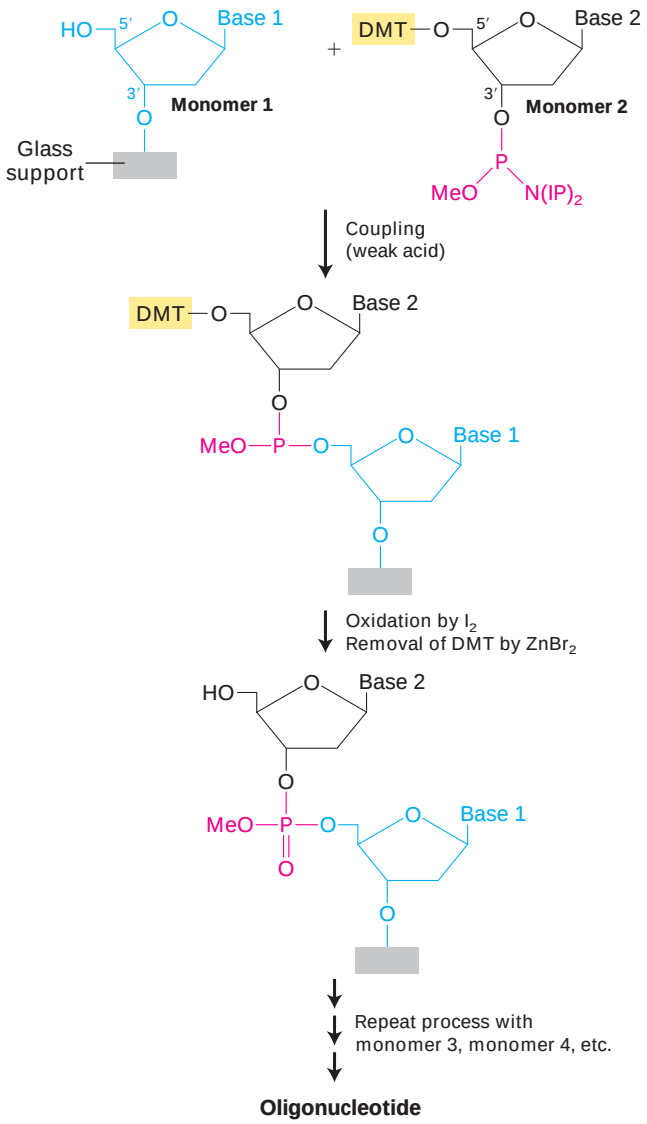
Application of this procedure for screening a λ cDNA library is depicted in Figure 9-17. In this case, a *replica* of the petri dish containing a large number of individual λ clones initially is reproduced on the surface of a nitrocellulose membrane. The membrane is then assayed using a radiolabeled probe specific for the recombinant DNA containing the fragment of interest. Membrane hybridization with radiolabeled oligonucleotides is most commonly used to screen λ cDNA libraries. Once a cDNA clone encoding a particular protein is obtained, the full-length cDNA can be radiolabeled and used to probe a genomic library for clones containing fragments of the corresponding gene.



▲ EXPERIMENTAL FIGURE 9-17 **Phage cDNA libraries can be screened with a radiolabeled probe to identify a clone of interest.** In the initial plating of a library, the λ phage plaques are not allowed to develop to a visible size so that up to 50,000 recombinants can be analyzed on a single plate. The appearance of a spot on the autoradiogram indicates the presence of a recombinant λ clone containing DNA complementary to the probe. The position of the spot on the autoradiogram is the mirror image of the position on the original petri dish of that particular clone. Aligning the autoradiogram with the original petri dish will locate the corresponding clone from which infectious phage particles can be recovered and replated at low density, resulting in well-separated plaques. Pure isolates eventually are obtained by repeating the hybridization assay.

Oligonucleotide Probes Are Designed Based on Partial Protein Sequences

Clearly, identification of specific clones by the membrane-hybridization technique depends on the availability of complementary radiolabeled probes. For an oligonucleotide to be useful as a probe, it must be long enough for its sequence to occur uniquely in the clone of interest and not in any other clones. For most purposes, this condition is satisfied by oligonucleotides containing about 20 nucleotides. This is be-



▲ FIGURE 9-18 **Chemical synthesis of oligonucleotides by sequential addition of reactive nucleotide derivatives.** The first (3′) nucleotide in the sequence (monomer 1) is bound to a glass support by its 3′ hydroxyl; its 5′ hydroxyl is available for addition of the second nucleotide. The second nucleotide in the sequence (monomer 2) is derivatized by addition of 4′,4′-dimethoxytrityl (DMT) to its 5′ hydroxyl, thus blocking this hydroxyl from reacting; in addition, a highly reactive group (red letters) is attached to the 3′ hydroxyl. When the two monomers are mixed in the presence of a weak acid, they form a 5′ → 3′ phosphodiester bond with the phosphorus in the trivalent state. Oxidation of this intermediate increases the phosphorus valency to 5, and subsequent removal of the DMT group with zinc bromide ($ZnBr_2$) frees the 5′ hydroxyl. Monomer 3 then is added, and the reactions are repeated. Repetition of this process eventually yields the entire oligonucleotide. Finally, all the methyl groups on the phosphates are removed at the same time at alkaline pH, and the bond linking monomer 1 to the glass support is cleaved. [See S. L. Beaucage and M. H. Caruthers, 1981, *Tetrahedron Lett.* **22**:1859.]

cause a specific 20-nucleotide sequence occurs once in every 4^{20} ($\approx 10^{12}$) nucleotides. Since all genomes are much smaller ($\approx 3 \times 10^9$ nucleotides for humans), a specific 20-nucleotide sequence in a genome usually occurs only once. Oligonucleotides of this length with a specific sequence can be synthesized chemically and then radiolabeled by using polynucleotide kinase to transfer a ^{32}P-labeled phosphate group from ATP to the 5′ end of each oligonucleotide.

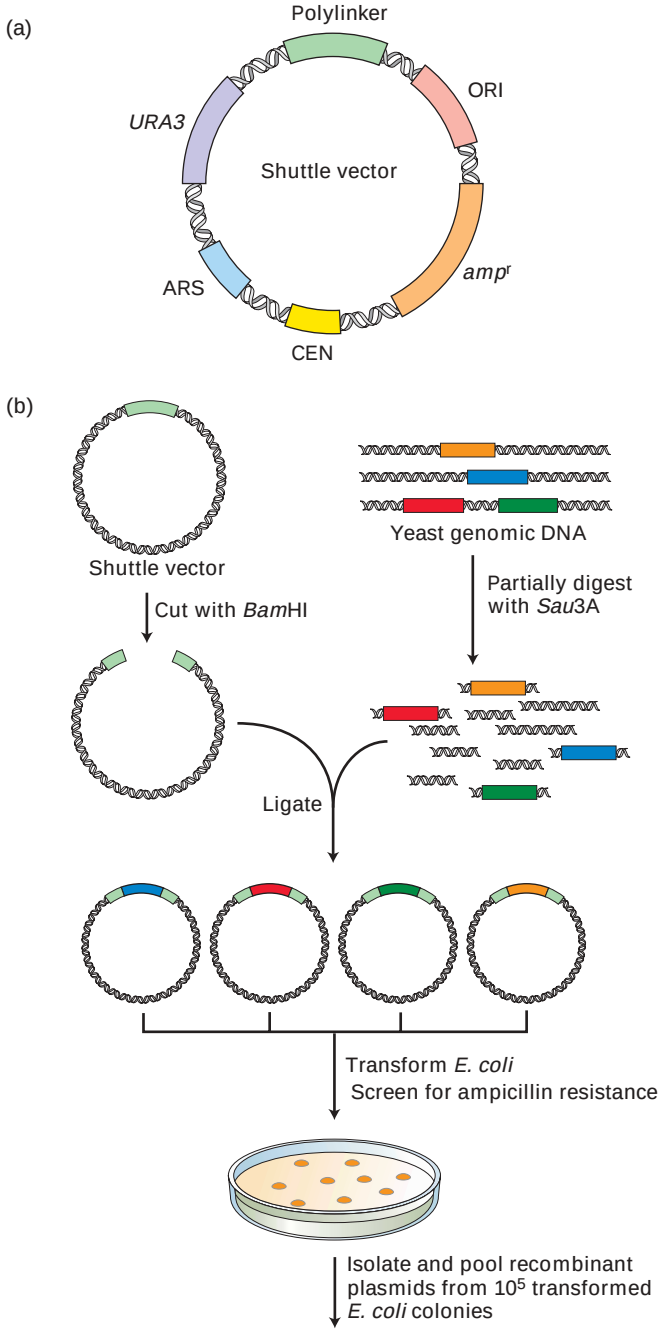
How might an investigator design an oligonucleotide probe to identify a cDNA clone encoding a particular protein? If all or a portion of the amino acid sequence of the protein is known, then a DNA probe corresponding to a small region of the gene can be designed based on the genetic code. However, because the genetic code is degenerate (i.e., many amino acids are encoded by more than one codon), a probe based on an amino acid sequence must include *all* the possible oligonucleotides that could theoretically encode that peptide sequence. Within this mixture of oligonucleotides will be one that hybridizes perfectly to the clone of interest.

In recent years, this approach has been simplified by the availability of the complete genomic sequences for humans and some important model organisms such as the mouse, *Drosophila*, and the roundworm *Caenorhabditis elegans*. Using an appropriate computer program, a researcher can search the genomic sequence database for the coding sequence that corresponds to a specific portion of the amino acid sequence of the protein under study. If a match is found, then a single, unique DNA probe based on this known genomic sequence will hybridize perfectly with the clone encoding the protein under study.

Chemical synthesis of single-stranded DNA probes of defined sequence can be accomplished by the series of reactions shown in Figure 9-18. With automated instruments now available, researchers can program the synthesis of oligonucleotides of specific sequence up to about 100 nucleotides long. Alternatively, these probes can be prepared by the polymerase chain reaction (PCR), a widely used technique for amplifying specific DNA sequences that is described later.

Yeast Genomic Libraries Can Be Constructed with Shuttle Vectors and Screened by Functional Complementation

In some cases a DNA library can be screened for the ability to express a functional protein that complements a recessive mutation. Such a screening strategy would be an efficient way to isolate a cloned gene that corresponds to an interesting recessive mutation identified in an experimental organism. To illustrate this method, referred to as **functional complementation,** we describe how yeast genes cloned in special *E. coli*



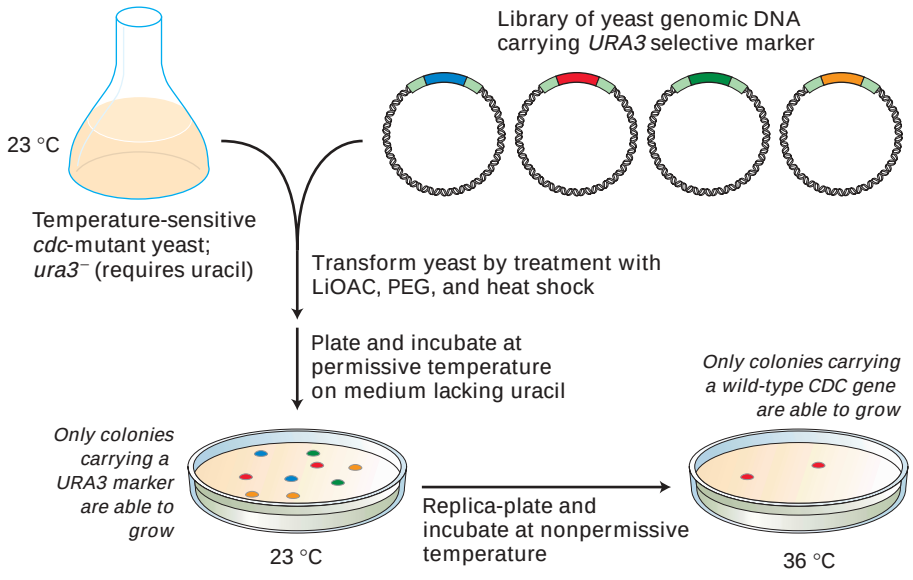
▶ EXPERIMENTAL FIGURE 9-19 **Yeast genomic library can be constructed in a plasmid shuttle vector that can replicate in yeast and *E. coli*.** (a) Components of a typical plasmid shuttle vector for cloning *Saccharomyces* genes. The presence of a yeast origin of DNA replication (ARS) and a yeast centromere (CEN) allows, stable replication and segregation in yeast. Also included is a yeast selectable marker such as *URA3*, which allows a *ura3*$^-$ mutant to grow on medium lacking uracil. Finally, the vector contains sequences for replication and selection in *E. coli* (ORI and *amp*^{r}) and a polylinker for easy insertion of yeast DNA fragments. (b) Typical protocol for constructing a yeast genomic library. Partial digestion of total yeast genomic DNA with *Sau*3A is adjusted to generate fragments with an average size of about 10 kb. The vector is prepared to accept the genomic fragments by digestion with *Bam*HI, which produces the same sticky ends as *Sau*3A. Each transformed clone of *E. coli* that grows after selection for ampicillin resistance contains a single type of yeast DNA fragment.

plasmids can be introduced into mutant yeast cells to identify the wild-type gene that is defective in the mutant strain.

Libraries constructed for the purpose of screening among yeast gene sequences usually are constructed from genomic DNA rather than cDNA. Because *Saccharomyces* genes do not contain multiple introns, they are sufficiently compact so that the entire sequence of a gene can be included in a genomic DNA fragment inserted into a plasmid vector. To construct a plasmid genomic library that is to be screened by functional complementation in yeast cells, the plasmid vector must be capable of replication in both *E. coli* cells and yeast cells. This type of vector, capable of propagation in two different hosts, is called a **shuttle vector.** The structure of a typical yeast shuttle vector is shown in Figure 9-19a (see page 369). This vector contains the basic elements that permit cloning of DNA fragments in *E. coli*. In addition, the shuttle vector contains an autonomously replicating sequence (ARS), which functions as an origin for DNA replication in yeast; a yeast centromere (called CEN), which allows faithful segregation of the plasmid during yeast cell division; and a yeast gene encoding an enzyme for uracil synthesis (*URA3*), which serves as a selectable marker in an appropriate yeast mutant.

To increase the probability that all regions of the yeast genome are successfully cloned and represented in the plasmid library, the genomic DNA usually is only partially digested to yield overlapping restriction fragments of ≈10 kb. These fragments are then ligated into the shuttle vector in which the polylinker has been cleaved with a restriction enzyme that produces sticky ends complementary to those on the yeast DNA fragments (Figure 9-19b). Because the 10-kb restriction fragments of yeast DNA are incorporated into the shuttle vectors randomly, at least 10^5 *E. coli* colonies, each containing a particular recombinant shuttle vector, are necessary to assure that each region of yeast DNA has a high probability of being represented in the library at least once.

Figure 9-20 outlines how such a yeast genomic library can be screened to isolate the wild-type gene corresponding to one of the temperature-sensitive *cdc* mutations mentioned earlier in this chapter. The starting yeast strain is a double mutant that requires uracil for growth due to a *ura3* mutation and is temperature-sensitive due to a *cdc28* mutation identified by its phenotype (see Figure 9-6). Recombinant plasmids isolated from the yeast genomic library are mixed with yeast cells under conditions that promote transformation of the cells with foreign DNA. Since transformed yeast cells carry a plasmid-borne copy of the wild-type *URA3* gene, they can be selected by their ability to grow in the absence of uracil. Typically, about 20 petri dishes, each containing about 500 yeast transformants, are sufficient to represent the entire yeast genome. This collection of yeast transformants can be maintained at 23 °C, a temperature permissive for growth of the *cdc28* mutant. The entire collection on 20 plates is then transferred to replica plates, which are placed at 36 °C, a nonpermissive temperature for



▲ EXPERIMENTAL FIGURE 9-20 **Screening of a yeast genomic library by functional complementation can identify clones carrying the normal form of mutant yeast gene.** In this example, a wild-type *CDC* gene is isolated by complementation of a *cdc* yeast mutant. The *Saccharomyces* strain used for screening the yeast library carries *ura3*⁻ and a temperature-sensitive *cdc* mutation. This mutant strain is grown and maintained at a permissive temperature (23 °C). Pooled recombinant plasmids prepared as shown in Figure 9-19 are incubated with the mutant yeast cells under conditions that promote transformation. The relatively few transformed yeast cells, which contain recombinant plasmid DNA, can grow in the absence of uracil at 23 °C. When transformed yeast colonies are replica-plated and placed at 36 °C (a nonpermissive temperature), only clones carrying a library plasmid that contains the wild-type copy of the *CDC* gene will survive. LiOAC = lithium acetate; PEG = polyethylene glycol.

cdc mutants. Yeast colonies that carry recombinant plasmids expressing a wild-type copy of the *CDC28* gene will be able to grow at 36 °C. Once temperature-resistant yeast colonies have been identified, plasmid DNA can be extracted from the cultured yeast cells and analyzed by subcloning and DNA sequencing, topics we take up in the next section.

KEY CONCEPTS OF SECTION 9.2

DNA Cloning by Recombinant DNA Methods

- In DNA cloning, recombinant DNA molecules are formed in vitro by inserting DNA fragments into vector DNA molecules. The recombinant DNA molecules are then introduced into host cells, where they replicate, producing large numbers of recombinant DNA molecules.
- Restriction enzymes (endonucleases) typically cut DNA at specific 4- to 8-bp palindromic sequences, producing defined fragments that often have self-complementary single-stranded tails (sticky ends).
- Two restriction fragments with complementary ends can be joined with DNA ligase to form a recombinant DNA (see Figure 9-11).
- *E. coli* cloning vectors are small circular DNA molecules (plasmids) that include three functional regions: an origin of replication, a drug-resistance gene, and a site where a DNA fragment can be inserted. Transformed cells carrying a vector grow into colonies on the selection medium (see Figure 9-13).
- Phage cloning vectors are formed by replacing nonessential parts of the λ genome with DNA fragments up to ≈25 kb in length and packaging the resulting recombinant DNAs with preassembled heads and tails in vitro.
- In cDNA cloning, expressed mRNAs are reverse-transcribed into complementary DNAs, or cDNAs. By a series of reactions, single-stranded cDNAs are converted into double-stranded DNAs, which can then be ligated into a λ phage vector (see Figure 9-15).
- A cDNA library is a set of cDNA clones prepared from the mRNAs isolated from a particular type of tissue. A genomic library is a set of clones carrying restriction fragments produced by cleavage of the entire genome.
- The number of clones in a cDNA or genomic library must be large enough so that all or nearly all of the original nucleotide sequences are present in at least one clone.
- A particular cloned DNA fragment within a library can be detected by hybridization to a radiolabeled oligonucleotide whose sequence is complementary to a portion of the fragment (see Figures 9-16 and 9-17).
- Shuttle vectors that replicate in both yeast and *E. coli* can be used to construct a yeast genomic library. Specific genes can be isolated by their ability to complement the corresponding mutant genes in yeast cells (see Figure 9-20).

9.3 Characterizing and Using Cloned DNA Fragments

Now that we have described the basic techniques for using recombinant DNA technology to isolate specific DNA clones, we consider how cloned DNAs are further characterized and various ways in which they can be used. We begin here with several widely used general techniques and examine some more specific applications in the following sections.

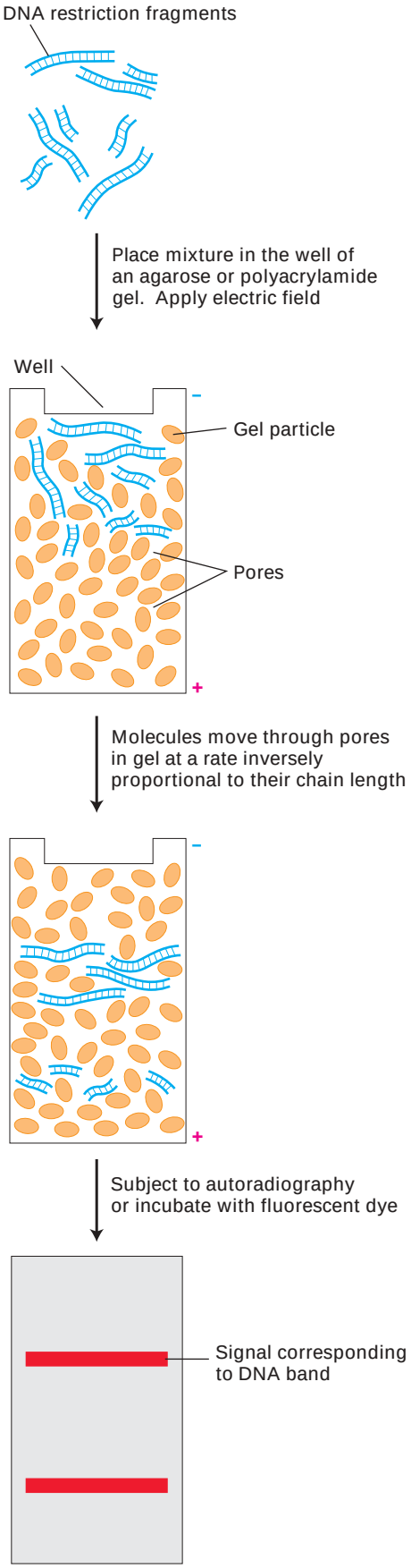
Gel Electrophoresis Allows Separation of Vector DNA from Cloned Fragments

In order to manipulate or sequence a cloned DNA fragment, it first must be separated from the vector DNA. This can be accomplished by cutting the recombinant DNA clone with the same restriction enzyme used to produce the recombinant vectors originally. The cloned DNA and vector DNA then are subjected to gel **electrophoresis,** a powerful method for separating DNA molecules of different size.

Near neutral pH, DNA molecules carry a large negative charge and therefore move toward the positive electrode during gel electrophoresis. Because the gel matrix restricts random diffusion of the molecules, molecules of the same length migrate together as a band whose width equals that of the well into which the original DNA mixture was placed at the start of the electrophoretic run. Smaller molecules move through the gel matrix more readily than larger molecules, so that molecules of different length migrate as distinct bands (Figure 9-21). DNA molecules composed of up to ≈2000 nucleotides usually are separated electrophoretically on *polyacrylamide gels*, and molecules from about 200 nucleotides to more than 20 kb on *agarose gels*.

A common method for visualizing separated DNA bands on a gel is to incubate the gel in a solution containing the fluorescent dye ethidium bromide. This planar molecule binds to DNA by intercalating between the base pairs. Binding concentrates ethidium in the DNA and also increases its intrinsic fluorescence. As a result, when the gel is illuminated with ultraviolet light, the regions of the gel containing DNA fluoresce much more brightly than the regions of the gel without DNA.

Once a cloned DNA fragment, especially a long one, has been separated from vector DNA, it often is treated with various restriction enzymes to yield smaller fragments. After separation by gel electrophoresis, all or some of these smaller fragments can be ligated individually into a plasmid vector and cloned in *E. coli* by the usual procedure. This process, known as *subcloning*, is an important step in rearranging parts of genes into useful new configurations. For instance, an investigator who wants to change the conditions under which a gene is expressed might use subcloning to replace the normal promoter associated with a cloned gene with a DNA segment containing a different promoter. Subcloning also can be used to obtain cloned DNA fragments that are of an appropriate length for determining the nucleotide sequence.

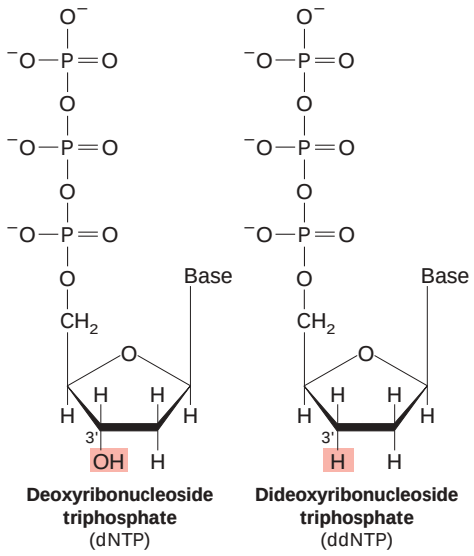


◀ EXPERIMENTAL FIGURE 9-21 **Gel electrophoresis separates DNA molecules of different lengths.** A gel is prepared by pouring a liquid containing either melted agarose or unpolymerized acrylamide between two glass plates a few millimeters apart. As the agarose solidifies or the acrylamide polymerizes into polyacrylamide, a gel matrix (orange ovals) forms consisting of long, tangled chains of polymers. The dimensions of the interconnecting channels, or pores, depend on the concentration of the agarose or acrylamide used to form the gel. The separated bands can be visualized by autoradiography (if the fragments are radiolabeled) or by addition of a fluorescent dye (e.g., ethidium bromide) that binds to DNA.

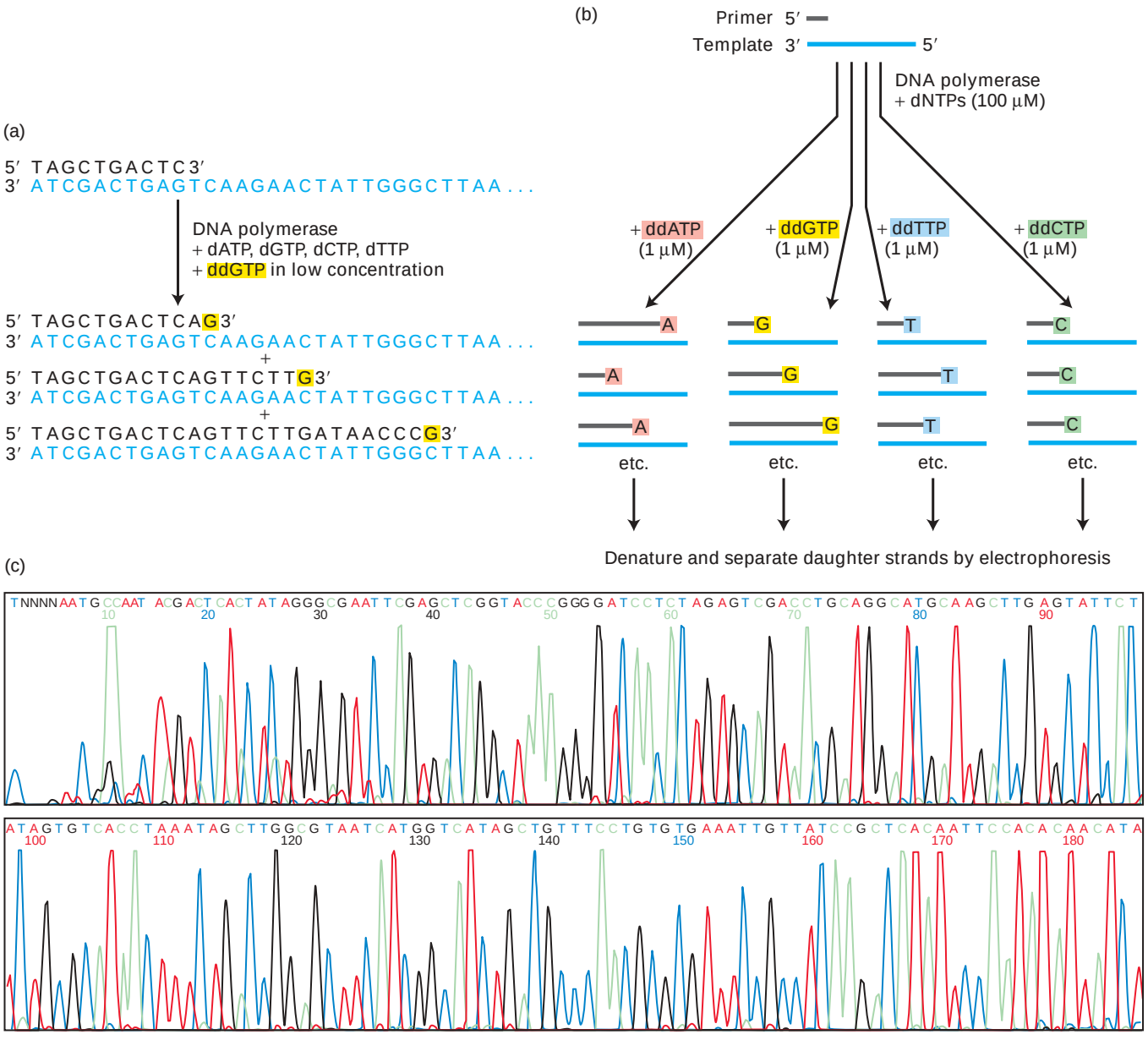
Cloned DNA Molecules Are Sequenced Rapidly by the Dideoxy Chain-Termination Method

The complete characterization of any cloned DNA fragment requires determination of its nucleotide sequence. F. Sanger and his colleagues developed the method now most commonly used to determine the exact nucleotide sequence of DNA fragments up to ≈500 nucleotides long. The basic idea behind this method is to synthesize from the DNA fragment to be sequenced a set of daughter strands that are labeled at one end and differ in length by one nucleotide. Separation of the truncated daughter strands by gel electrophoresis can then establish the nucleotide sequence of the original DNA fragment.

Synthesis of truncated daughter stands is accomplished by use of 2′,3′-dideoxyribonucleoside triphosphates (ddNTPs). These molecules, in contrast to normal deoxyribonucleotides (dNTPs), lack a 3′ hydroxyl group (Figure 9-22). Although ddNTPs can be incorporated into a growing DNA chain by



▲ FIGURE 9-22 **Structures of deoxyribonucleoside triphosphate (dNTP) and dideoxyribonucleoside triphosphate (ddNTP).** Incorporation of a ddNTP residue into a growing DNA strand terminates elongation at that point.

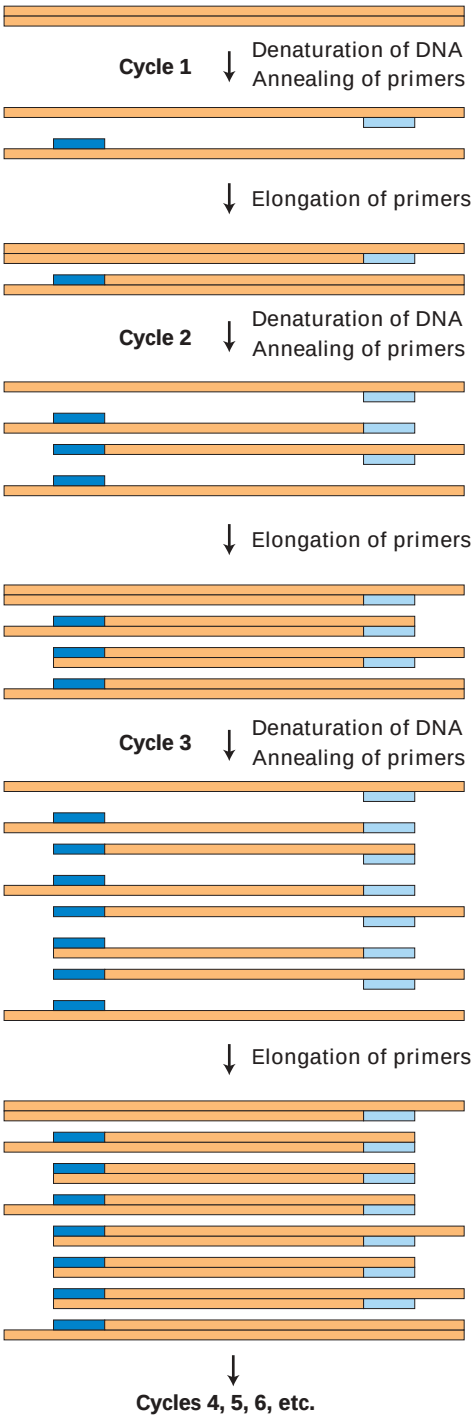


▲ EXPERIMENTAL FIGURE 9-23 **Cloned DNAs can be sequenced by the Sanger method, using fluorescent-tagged dideoxyribonucleoside triphosphates (ddNTPs).** (a) A single (template) strand of the DNA to be sequenced (blue letters) is hybridized to a synthetic deoxyribonucleotide primer (black letters). The primer is elongated in a reaction mixture containing the four normal deoxyribonucleoside triphosphates plus a relatively small amount of one of the four dideoxyribonucleoside triphosphates. In this example, ddGTP (yellow) is present. Because of the relatively low concentration of ddGTP, incorporation of a ddGTP, and thus chain termination, occurs at a given position in the sequence only about 1 percent of the time. Eventually the reaction mixture will contain a mixture of prematurely terminated (truncated) daughter fragments ending at every occurrence of ddGTP. (b) To obtain the complete sequence of a template DNA, four separate reactions are performed, each with a different dideoxyribonucleoside triphosphate (ddNTP). The ddNTP that terminates each truncated fragment can be identified by use of ddNTPs tagged with four different fluorescent dyes (indicated by colored highlights). (c) In an automated sequencing machine, the four reaction mixtures are subjected to gel electrophoresis and the order of appearance of each of the four different fluorescent dyes at the end of the gel is recorded. Shown here is a sample printout from an automated sequencer from which the sequence of the original template DNA can be read directly. N = nucleotide that cannot be assigned. [Part (c) from Griffiths et al., Figure 14-27.]

DNA polymerase, once incorporated they cannot form a phosphodiester bond with the next incoming nucleotide triphosphate. Thus incorporation of a ddNTP terminates chain synthesis, resulting in a truncated daughter strand.

Sequencing using the Sanger *dideoxy chain-termination method* begins by denaturing a double-stranded DNA fragment to generate template strands for in vitro DNA synthesis. A synthetic oligodeoxynucleotide is used as the primer for *four* separate polymerization reactions, each with a low concentration of one of the four ddNTPs in addition to higher concentrations of the normal dNTPs. In each reaction, the ddNTP is randomly incorporated at the positions of the corresponding dNTP, causing termination of polymerization at those positions in the sequence (Figure 9-23a). Inclusion of fluorescent tags of different colors on each of the ddNTPs allows each set of truncated daughter fragments to be distinguished by their corresponding fluorescent label (Figure 9-23b). For example, all truncated fragments that end with a G would fluoresce one color (e.g., yellow), and those ending with an A would fluoresce another color (e.g., red), regardless of their lengths. The mixtures of truncated daughter fragments from each of the four reactions are subjected to electrophoresis on special polyacrylamide gels that can separate single-stranded DNA molecules differing in length by only 1 nucleotide. In automated DNA sequencing machines, a fluorescence detector that can distinguish the four fluorescent tags is located at the end of the gel. The sequence of the original DNA template strand can be determined from the order in which different labeled fragments migrate past the fluorescence detector (Figure 9-23c).

In order to sequence a long continuous region of genomic DNA, researchers often start with a collection of cloned DNA fragments whose sequences overlap. Once the sequence of one of these fragments is determined, oligonucleotides based on that sequence can be chemically synthesized for use as primers in sequencing the adjacent overlapping fragments. In this way, the sequence of a long stretch of DNA is determined incrementally by sequencing of the overlapping cloned DNA fragments that compose it.



◀ EXPERIMENTAL FIGURE 9-24 **The polymerase chain reaction (PCR) is widely used to amplify DNA regions of known sequences.** To amplify a specific region of DNA, an investigator will chemically synthesize two different oligonucleotide primers complementary to sequences of approximately 18 bases flanking the region of interest (designated as light blue and dark blue bars). The complete reaction is composed of a complex mixture of double-stranded DNA (usually genomic DNA containing the target sequence of interest), a stoichiometric excess of both primers, the four deoxynucleoside triphosphates, and a heat-stable DNA polymerase known as Taq *polymerase*. During each PCR cycle, the reaction mixture is first heated to separate the strands and then cooled to allow the primers to bind to complementary sequences flanking the region to be amplified. *Taq* polymerase then extends each primer from its 3′ end, generating newly synthesized strands that extend in the 3′ direction to the 5′ end of the template strand. During the third cycle, two double-stranded DNA molecules are generated equal in length to the sequence of the region to be amplified. In each successive cycle the target segment, which will anneal to the primers, is duplicated, and will eventually vastly outnumber all other DNA segments in the reaction mixture. Successive PCR cycles can be automated by cycling the reaction for timed intervals at high temperature for DNA melting and at a defined lower temperature for the annealing and elongation portions of the cycle. A reaction that cycles 20 times will amplify the specific target sequence 1-million-fold.

MEDIA CONNECTIONS
Technique Animation: Polymerase Chain Reaction

The Polymerase Chain Reaction Amplifies a Specific DNA Sequence from a Complex Mixture

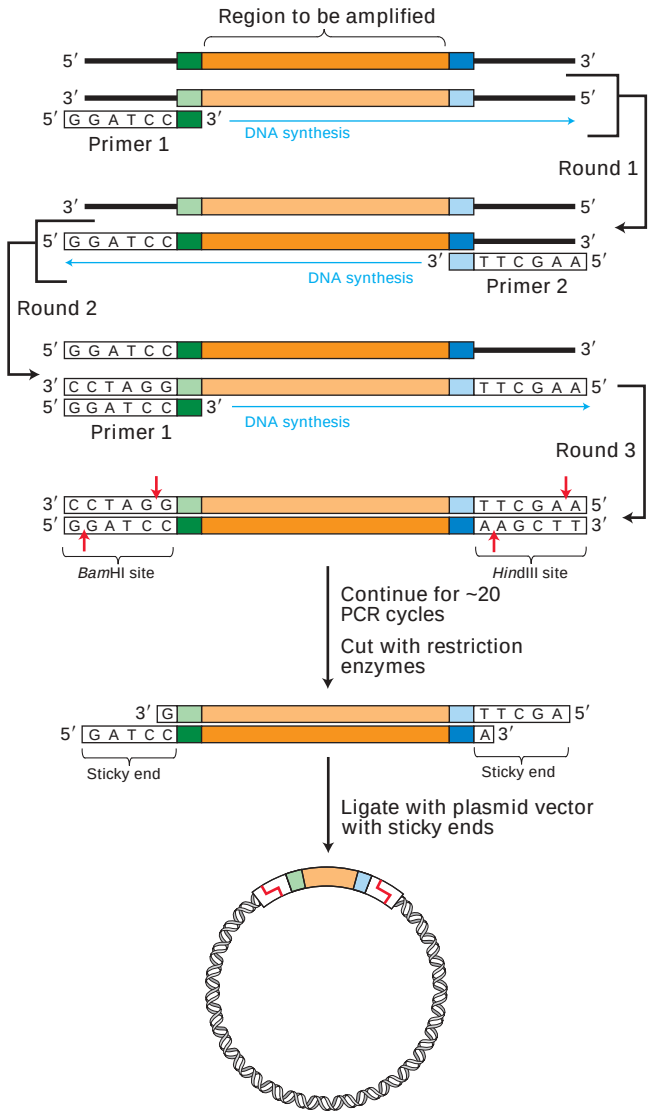
If the nucleotide sequences at the ends of a particular DNA region are known, the intervening fragment can be amplified directly by the **polymerase chain reaction (PCR).** Here we describe the basic PCR technique and three situations in which it is used.

The PCR depends on the ability to alternately denature (melt) double-stranded DNA molecules and renature (anneal) complementary single strands in a controlled fashion. As in the membrane-hybridization assay described earlier, the presence of noncomplementary strands in a mixture has little effect on the base pairing of complementary single DNA strands or complementary regions of strands. The second requirement for PCR is the ability to synthesize oligonucleotides at least 18–20 nucleotides long with a defined sequence. Such synthetic nucleotides can be readily produced with automated instruments based on the standard reaction scheme shown in Figure 9-18.

As outlined in Figure 9-24, a typical PCR procedure begins by heat-denaturation of a DNA sample into single strands. Next, two synthetic oligonucleotides complementary to the 3′ ends of the target DNA segment of interest are added in great excess to the denatured DNA, and the temperature is lowered to 50–60 °C. These specific oligonucleotides, which are at a very high concentration, will hybridize with their complementary sequences in the DNA sample, whereas the long strands of the sample DNA remain apart because of their low concentration. The hybridized oligonucleotides then serve as primers for DNA chain synthesis in the presence of deoxynucleotides (dNTPs) and a temperature-resistant DNA polymerase such as that from *Thermus aquaticus* (a bacterium that lives in hot springs). This enzyme, called *Taq polymerase,* can remain active even after being heated to 95 °C and can extend the primers at temperatures up to 72 °C. When synthesis is complete, the whole mixture is then heated to 95 °C to melt the newly formed DNA duplexes. After the temperature is lowered again, another cycle of synthesis takes place because excess primer is still present. Repeated cycles of melting (heating) and synthesis (cooling) quickly amplify the sequence of interest. At each cycle, the number of copies of the sequence between the primer sites is doubled; therefore, the desired sequence increases exponentially—about a million-fold after 20 cycles—whereas all other sequences in the original DNA sample remain unamplified.

Direct Isolation of a Specific Segment of Genomic DNA

For organisms in which all or most of the genome has been sequenced, PCR amplification starting with the total genomic DNA often is the easiest way to obtain a specific DNA region of interest for cloning. In this application, the two oligonucleotide primers are designed to hybridize to sequences flanking the genomic region of interest and to include sequences that are recognized by specific restriction enzymes (Figure 9-25). After amplification of the desired target sequence for about 20 PCR cycles, cleavage with the appropriate restriction enzymes produces sticky ends that allow efficient ligation of the fragment into a plasmid vector cleaved by the same restriction enzymes in the polylinker. The resulting recombinant plasmids, all carrying the identical genomic DNA segment, can then be cloned in



▲ EXPERIMENTAL FIGURE 9-25 **A specific target region in total genomic DNA can be amplified by PCR for use in cloning.** Each primer for PCR is complementary to one end of the target sequence and includes the recognition sequence for a restriction enzyme that does not have a site within the target region. In this example, primer 1 contains a *Bam*HI sequence, whereas primer 2 contains a *Hin*dIII sequence. (Note that for clarity, in any round, amplification of only one of the two strands is shown, the one in brackets.) After amplification, the target segments are treated with appropriate restriction enzymes, generating fragments with sticky ends. These can be incorporated into complementary plasmid vectors and cloned in *E. coli* by the usual procedure (see Figure 9-13).

E. coli cells. With certain refinements of the PCR, DNA segments >10 kb in length can be amplified and cloned in this way.

Note that this method does not involve cloning of large numbers of restriction fragments derived from genomic DNA and their subsequent screening to identify the specific fragment of interest. In effect, the PCR method inverts this traditional approach and thus avoids its most tedious aspects. The PCR method is useful for isolating gene sequences to be manipulated in a variety of useful ways described later. In addition the PCR method can be used to isolate gene sequences from mutant organisms to determine how they differ from the wild-type.

Preparation of Probes Earlier we discussed how oligonucleotide probes for hybridization assays can be chemically synthesized. Preparation of such probes by PCR amplification requires chemical synthesis of only two relatively short primers corresponding to the two ends of the target sequence. The starting sample for PCR amplification of the target sequence can be a preparation of genomic DNA. Alternatively, if the target sequence corresponds to a mature mRNA sequence, a complete set of cellular cDNAs synthesized from the total cellular mRNA using reverse transcriptase or obtained by pooling cDNA from all the clones in a λ cDNA library can be used as a source of template DNA. To generate a radiolabeled product from PCR, ^{32}P-labeled dNTPs are included during the last several amplification cycles. Because probes prepared by PCR are relatively long and have many radioactive ^{32}P atoms incorporated into them, these probes usually give a stronger and more specific signal than chemically synthesized probes.

Tagging of Genes by Insertion Mutations Another useful application of the PCR is to amplify a "tagged" gene from the genomic DNA of a mutant strain. This approach is a simpler method for identifying genes associated with a particular mutant phenotype than screening of a library by functional complementation (see Figure 9-20).

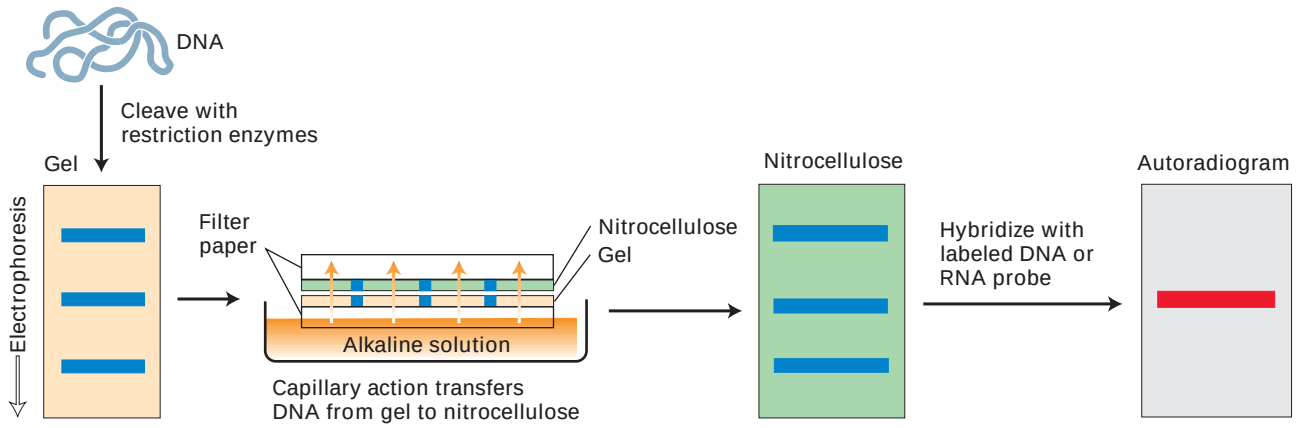
The key to this use of PCR is the ability to produce mutations by insertion of a known DNA sequence into the genome of an experimental organism. Such insertion mutations can be generated by use of **mobile DNA elements,** which can move (or transpose) from one chromosomal site to another. As discussed in more detail in Chapter 10, these DNA sequences occur naturally in the genomes of most organisms and may give rise to loss-of-function mutations if they transpose into a protein-coding region.

For example, researchers have modified a *Drosophila* mobile DNA element, known as the **P element,** to optimize its use in the experimental generation of insertion mutations. Once it has been demonstrated that insertion of a P element causes a mutation with an interesting phenotype, the genomic sequences adjacent to the insertion site can be amplified by a variation of the standard PCR protocol that uses synthetic primers complementary to the known P-element sequence but that allows unknown neighboring sequences to be amplified. Again, this approach avoids the cloning of large numbers of DNA fragments and their screening to detect a cloned DNA corresponding to a mutated gene of interest.

Similar methods have been applied to other organisms for which insertion mutations can be generated using either mobile DNA elements or viruses with sequenced genomes that can insert randomly into the genome.

Blotting Techniques Permit Detection of Specific DNA Fragments and mRNAs with DNA Probes

Two very sensitive methods for detecting a particular DNA or RNA sequence within a complex mixture combine separation by gel electrophoresis and hybridization with a complementary radiolabeled DNA probe. We will encounter

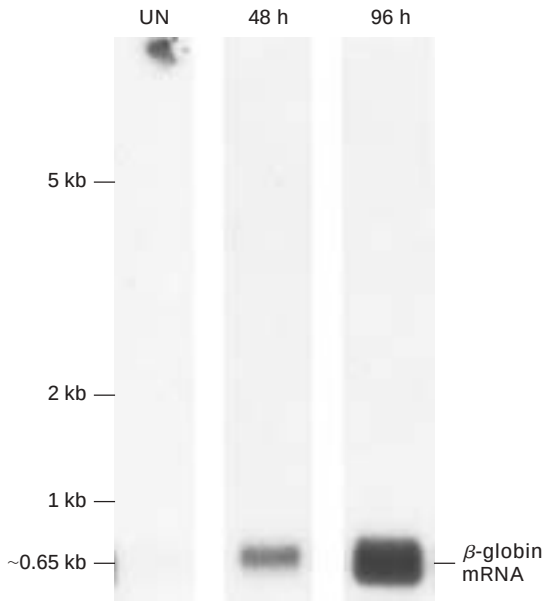


▲ EXPERIMENTAL FIGURE 9-26 **Southern blot technique can detect a specific DNA fragment in a complex mixture of restriction fragments.** The diagram depicts three different restriction fragments in the gel, but the procedure can be applied to a mixture of millions of DNA fragments. Only fragments that hybridize to a labeled probe will give a signal on an autoradiogram. A similar technique called *Northern blotting* detects specific mRNAs within a mixture. [See E. M. Southern, 1975, *J. Mol. Biol.* **98**:508.]

references to both these techniques, which have numerous applications, in other chapters.

Southern Blotting The first blotting technique to be devised is known as **Southern blotting** after its originator E. M. Southern. This technique is capable of detecting a single specific restriction fragment in the highly complex mixture of fragments produced by cleavage of the entire human genome with a restriction enzyme. In such a complex mixture, many fragments will have the same or nearly the same length and thus migrate together during electrophoresis. Even though all the fragments are not separated completely by gel electrophoresis, an individual fragment within one of the bands can be identified by hybridization to a specific DNA probe. To accomplish this, the restriction fragments present in the gel are denatured with alkali and transferred onto a nitrocellulose filter or nylon membrane by blotting (Figure 9-26). This procedure preserves the distribution of the fragments in the gel, creating a replica of the gel on the filter, much like the replica filter produced from clones in a λ library. (The blot is used because probes do not readily diffuse into the original gel.) The filter then is incubated under hybridization conditions with a specific radiolabeled DNA probe, which usually is generated from a cloned restriction fragment. The DNA restriction fragment that is complementary to the probe hybridizes, and its location on the filter can be revealed by autoradiography.

Northern Blotting One of the most basic ways to characterize a cloned gene is to determine when and where in an organism the gene is expressed. Expression of a particular gene can be followed by assaying for the corresponding mRNA by **Northern blotting,** named, in a play on words, after the related method of Southern blotting. An RNA sample, often the total cellular RNA, is denatured by treatment with an agent such as formaldehyde that disrupts the hydrogen bonds between base pairs, ensuring that all the RNA molecules have an unfolded, linear conformation. The individual RNAs are separated according to size by gel electrophoresis and transferred to a nitrocellulose filter to which the extended denatured RNAs adhere. As in Southern blotting, the filter then is exposed to a labeled DNA probe that is complementary to the gene of interest; finally, the labeled filter is subjected to autoradiography. Because the amount of a specific RNA in a sample can be estimated from a Northern blot, the procedure is widely used to compare the amounts of a particular mRNA in cells under different conditions (Figure 9-27).



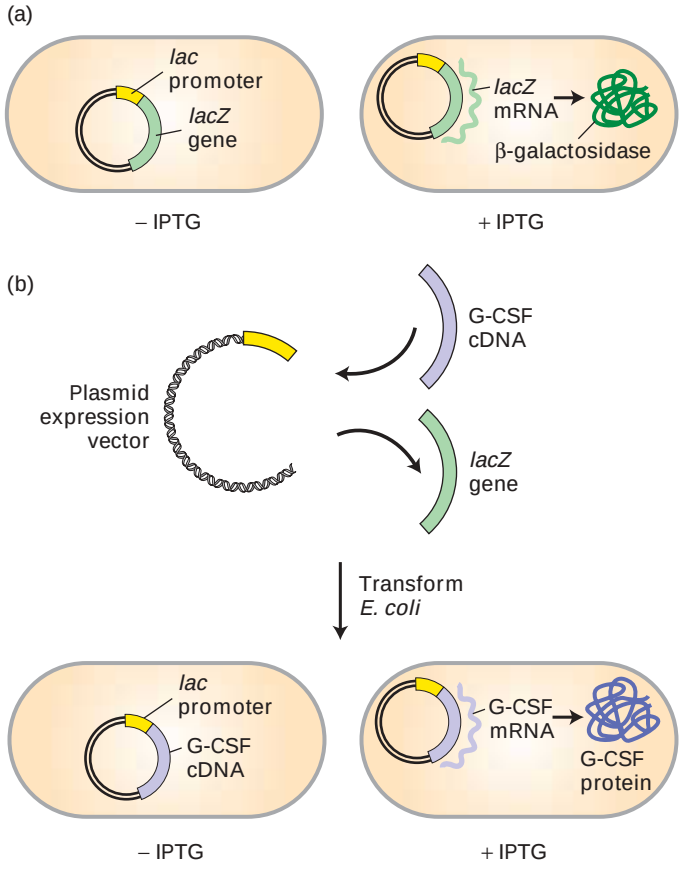
▲ EXPERIMENTAL FIGURE 9-27 **Northern blot analysis reveals increased expression of β-globin mRNA in differentiated erythroleukemia cells.** The total mRNA in extracts of erythroleukemia cells that were growing but uninduced and in cells induced to stop growing and allowed to differentiate for 48 hours or 96 hours was analyzed by Northern blotting for β-globin mRNA. The density of a band is proportional to the amount of mRNA present. The β-globin mRNA is barely detectable in uninduced cells (UN lane) but increases more than 1000-fold by 96 hours after differentiation is induced. [Courtesy of L. Kole.]

E. coli Expression Systems Can Produce Large Quantities of Proteins from Cloned Genes

MEDICINE

Many protein hormones and other signaling or regulatory proteins are normally expressed at very low concentrations, precluding their isolation and purification in large quantities by standard biochemical techniques. Widespread therapeutic use of such proteins, as well as basic research on their structure and functions, depends on efficient procedures for producing them in large amounts at reasonable cost. Recombinant DNA techniques that turn *E. coli* cells into factories for synthesizing low-abundance proteins now are used to commercially produce factor VIII (a blood-clotting factor), granulocyte colony-stimulating factor (G-CSF), insulin, growth hormone, and other human proteins with therapeutic uses. For example, G-CSF stimulates the production of granulocytes, the phagocytic white blood cells critical to defense against bacterial infections. Administration of G-CSF to cancer patients helps offset the reduction in granulocyte production caused by chemotherapeutic agents, thereby protecting patients against serious infection while they are receiving chemotherapy. ▮

The first step in producing large amounts of a low-abundance protein is to obtain a cDNA clone encoding the full-length protein by methods discussed previously. The second step is to engineer plasmid vectors that will express large amounts of the encoded protein when it is inserted into *E. coli* cells. The key to designing such **expression vectors** is



▲ EXPERIMENTAL FIGURE 9-28 **Some eukaryotic proteins can be produced in *E. coli* cells from plasmid vectors containing the *lac* promoter.** (a) The plasmid expression vector contains a fragment of the *E. coli* chromosome containing the *lac* promoter and the neighboring *lacZ* gene. In the presence of the lactose analog IPTG, RNA polymerase normally transcribes the *lacZ* gene, producing *lacZ* mRNA, which is translated into the encoded protein, β-galactosidase. (b) The *lacZ* gene can be cut out of the expression vector with restriction enzymes and replaced by a cloned cDNA, in this case one encoding granulocyte colony-stimulating factor (G-CSF). When the resulting plasmid is transformed into *E. coli* cells, addition of IPTG and subsequent transcription from the *lac* promoter produce G-CSF mRNA, which is translated into G-CSF protein.

inclusion of a **promoter,** a DNA sequence from which transcription of the cDNA can begin. Consider, for example, the relatively simple system for expressing G-CSF shown in Figure 9-28. In this case, G-CSF is expressed in *E. coli* transformed with plasmid vectors that contain the *lac* promoter adjacent to the cloned cDNA encoding G-CSF. Transcription from the *lac* promoter occurs at high rates only when lactose, or a lactose analog such as isopropylthiogalactoside (IPTG), is added to the culture medium. Even larger quantities of a desired protein can be produced in more complicated *E. coli* expression systems.

To aid in purification of a eukaryotic protein produced in an *E. coli* expression system, researchers often modify the cDNA encoding the recombinant protein to facilitate its separation from endogenous *E. coli* proteins. A commonly used modification of this type is to add a short nucleotide sequence to the end of the cDNA, so that the expressed protein will have six histidine residues at the C-terminus. Proteins modified in this way bind tightly to an affinity matrix that contains chelated nickel atoms, whereas most *E. coli* proteins will not bind to such a matrix. The bound proteins can be released from the nickel atoms by decreasing the pH of the surrounding medium. In most cases, this procedure yields a pure recombinant protein that is functional, since addition of short amino acid sequences to either the C-terminus or the N-terminus of a protein usually does not interfere with the protein's biochemical activity.

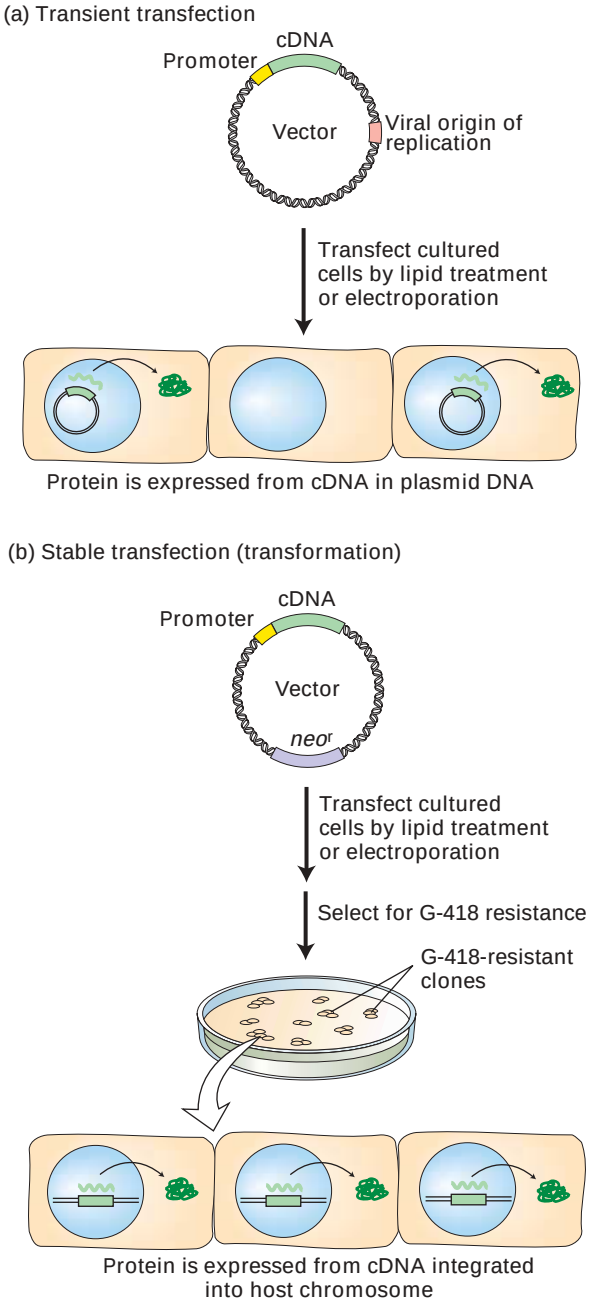
Plasmid Expression Vectors Can Be Designed for Use in Animal Cells

One disadvantage of bacterial expression systems is that many eukaryotic proteins undergo various modifications (e.g., glycosylation, hydroxylation) after their synthesis on ribosomes (Chapter 3). These post-translational modifications generally are required for a protein's normal cellular function, but they cannot be introduced by *E. coli* cells, which lack the necessary enzymes. To get around this limitation, cloned genes are introduced into cultured animal cells, a process called **transfection.** Two common methods for transfecting animal cells differ in whether the recombinant vector DNA is or is not integrated into the host-cell genomic DNA.

In both methods, cultured animal cells must be treated to facilitate their initial uptake of a recombinant plasmid vector. This can be done by exposing cells to a preparation of lipids that penetrate the plasma membrane, increasing its permeability to DNA. Alternatively, subjecting cells to a brief electric shock of several thousand volts, a technique known as *electroporation*, makes them transiently permeable to DNA. Usually the plasmid DNA is added in sufficient concentration to ensure that a large proportion of the cultured cells will receive at least one copy of the plasmid DNA.

Transient Transfection

The simplest of the two expression methods, called *transient transfection*, employs a vector similar to the yeast shuttle vectors described previously. For use in mammalian cells, plasmid vectors are engineered also to carry an origin of replication derived from a virus that infects mammalian cells, a strong promoter recognized by mammalian RNA polymerase, and the cloned cDNA encoding the protein to be expressed adjacent to the promoter (Figure 9-29a). Once such a plasmid vector enters a mammalian cell, the viral origin of replication allows it to replicate efficiently, generating numerous plasmids from which the protein is ex-



◀ EXPERIMENTAL FIGURE 9-29 **Transient and stable transfection with specially designed plasmid vectors permit expression of cloned genes in cultured animal cells.** Both methods employ plasmid vectors that contain the usual elements—ORI, selectable marker (e.g., *amp*^{r}), and polylinker—that permit propagation in *E. coli* and insertion of a cloned cDNA with an adjacent animal promoter. For simplicity, these elements are not depicted. (a) In transient transfection, the plasmid vector contains an origin of replication for a virus that can replicate in the cultured animal cells. Since the vector is not incorporated into the genome of the cultured cells, production of the cDNA-encoded protein continues only for a limited time. (b) In stable transfection, the vector carries a selectable marker such as *neo*^{r}, which confers resistance to G-418. The relatively few transfected animal cells that integrate the exogenous DNA into their genomes are selected on medium containing G-418. These stably transfected, or transformed, cells will continue to produce the cDNA-encoded protein as long as the culture is maintained. See the text for discussion.

pressed. However, during cell division such plasmids are not faithfully segregated into both daughter cells and in time a substantial fraction of the cells in a culture will not contain a plasmid, hence the name *transient transfection*.

Stable Transfection (Transformation) If an introduced vector integrates into the genome of the host cell, the genome is permanently altered and the cell is said to be *transformed*. Integration most likely is accomplished by mammalian enzymes that function normally in DNA repair and recombination. Because integration is a rare event, plasmid expression vectors designed to transform animal cells must carry a selectable marker in order to identify the small fraction of cells that integrate the plasmid DNA. A commonly used selectable marker is the gene for neomycin phosphotransferase (designated *neo*^{r}), which confers resistance to a toxic compound chemically related to neomycin known as G-418. The basic procedure for expressing a cloned cDNA by *stable transfection* is outlined in Figure 9-29b. Only those cells that have integrated the expression vector into the host chromosome will survive and give rise to a clone in the presence of a high concentration of G-418. Because integration occurs at random sites in the genome, individual transformed clones resistant to G-418 will differ in their rates of transcribing the inserted cDNA. Therefore, the stable transfectants usually are screened to identify those that produce the protein of interest at the highest levels.

Epitope Tagging In addition to their use in producing proteins that are modified after translation, eukaryotic expression vectors provide an easy way to study the intracellular localization of eukaryotic proteins. In this method, a cloned cDNA is modified by fusing it to a short DNA sequence encoding an amino acid sequence recognized by a known monoclonal antibody. Such a short peptide that is bound by an antibody is called an **epitope;** hence this method is known as *epitope tagging*. After transfection with a plasmid expression vector containing the fused cDNA, the expressed epitope-tagged form of the protein can be detected by immunofluorescence labeling of the cells with the monoclonal antibody specific for the epitope. Figure 9-30 illustrates the use of this method to localize AP1 adapter proteins, which participate in formation of clathrin-coated vesicles involved in intracellular protein trafficking (Chapter 17). Epitope tagging of a protein so it is detectable with an available monoclonal antibody obviates the time-consuming task of producing a new monoclonal antibody specific for the natural protein.

AP1 Furin Merged

▲ EXPERIMENTAL FIGURE 9-30 **Epitope tagging facilitates cellular localization of proteins expressed from cloned genes.** In this experiment, the cloned cDNA encoding one subunit of the AP1 adapter protein was modified by addition of a sequence encoding an epitope for a known monoclonal antibody. Plasmid expression vectors, similar to those shown in Figure 9-29, were constructed to contain the epitope-tagged AP1 cDNA. After cells were transfected and allowed to express the epitope-tagged version of the AP1 protein, they were fixed and labeled with monoclonal antibody to the epitope and with antibody to furin, a marker protein for the late Golgi and endosomal membranes. Addition of a green fluorescently labeled secondary antibody specific for the anti-epitope antibody visualized the AP1 protein (*left*). Another secondary antibody with a different (red) fluorescent signal was used to visualize furin (*center*). The colocalization of epitope-tagged AP1 and furin to the same intracellular compartment is evident when the two fluorescent signals are merged (*right*). [Courtesy of Ira Mellman, Yale University School of Medicine.]

KEY CONCEPTS OF SECTION 9.3

Characterizing and Using Cloned DNA Fragments

- Long cloned DNA fragments often are cleaved with restriction enzymes, producing smaller fragments that then are separated by gel electrophoresis and subcloned in plasmid vectors prior to sequencing or experimental manipulation.
- DNA fragments up to about 500 nucleotides long are most commonly sequenced in automated instruments based on the Sanger (dideoxy chain termination) method (see Figure 9-23).
- The polymerase chain reaction (PCR) permits exponential amplification of a specific segment of DNA from just a single initial template DNA molecule if the sequence flanking the DNA region to be amplified is known (see Figure 9-24).
- Southern blotting can detect a single, specific DNA fragment within a complex mixture by combining gel electrophoresis, transfer (blotting) of the separated bands to a filter, and hybridization with a complementary radiolabeled DNA probe (see Figure 9-26). The similar technique of Northern blotting detects a specific RNA within a mixture.
- Expression vectors derived from plasmids allow the production of abundant amounts of a protein of interest once a cDNA encoding it has been cloned. The unique feature of these vectors is the presence of a promoter fused to the cDNA that allows high-level transcription in host cells.
- Eukaryotic expression vectors can be used to express cloned genes in yeast or mammalian cells (see Figure 9-29). An important application of these methods is the tagging of proteins with an epitope for antibody detection.

9.4 Genomics: Genome-wide Analysis of Gene Structure and Expression

Using specialized recombinant DNA techniques, researchers have determined vast amounts of DNA sequence including the entire genomic sequence of humans and many key experimental organisms. This enormous volume of data, which is growing at a rapid pace, has been stored and organized in two primary data banks: the GenBank at the National Institutes of Health, Bethesda, Maryland, and the EMBL Sequence Data Base at the European Molecular Biology Laboratory in Heidelberg, Germany. These databases continuously exchange newly reported sequences and make them available to scientists throughout the world on the Internet. In this section, we examine some of the ways researchers use this treasure trove of data to provide insights about gene function and evolutionary relationships, to identify new genes whose encoded proteins have never been isolated, and to determine when and where genes are expressed.

Stored Sequences Suggest Functions of Newly Identified Genes and Proteins

As discussed in Chapter 3, proteins with similar functions often contain similar amino acid sequences that correspond to important functional domains in the three-dimensional structure of the proteins. By comparing the amino acid sequence of the protein encoded by a newly cloned gene with the sequences of proteins of known function, an investigator can look for sequence similarities that provide clues to the function of the encoded protein. Because of the degeneracy in the genetic code, related proteins invariably exhibit more sequence similarity than the genes encoding them. For this reason, protein sequences rather than the corresponding DNA sequences are usually compared.

The computer program used for this purpose is known as BLAST (*b*asic *l*ocal *a*lignment search *t*ool). The BLAST algorithm divides the new protein sequence (known as the *query sequence*) into shorter segments and then searches the database for significant matches to any of the stored sequences. The matching program assigns a high score to

identically matched amino acids and a lower score to matches between amino acids that are related (e.g., hydrophobic, polar, positively charged, negatively charged). When a significant match is found for a segment, the BLAST algorithm will search locally to extend the region of similarity. After searching is completed, the program ranks the matches between the query protein and various known proteins according to their *p-values.* This parameter is a measure of the probability of finding such a degree of similarity between two protein sequences by chance. The lower the *p*-value, the greater the sequence similarity between two sequences. A *p*-value less than about 10^{-3} usually is considered as significant evidence that two proteins share a common ancestor.

To illustrate the power of this approach, we consider *NF1,* a human gene identified and cloned by methods described later in this chapter. Mutations in *NF1* are associated with the inherited disease neurofibromatosis 1, in which multiple tumors develop in the peripheral nervous system, causing large protuberances in the skin (the "elephant-man" syndrome). After a cDNA clone of *NF1* was isolated and sequenced, the deduced sequence of the NF1 protein was checked against all other protein sequences in GenBank. A region of NF1 protein was discovered to have considerable homology to a portion of the yeast protein called Ira (Figure 9-31). Previous studies had shown that Ira is a GTPase-accelerating protein (GAP) that modulates the GTPase activity of the monomeric G protein called Ras (see Figure 3-E). As we examine in detail in Chapters 14 and 15, GAP and Ras proteins normally function to control cell replication and differentiation in response to signals from neighboring cells. Functional studies on the normal NF1 protein, obtained by expression of the cloned wild-type gene, showed that it did, indeed, regulate Ras activity, as suggested by its homology with Ira. These findings suggest that individuals with neurofibromatosis express a mutant NF1 protein in cells of the peripheral nervous system, leading to inappropriate cell division and formation of the tumors characteristic of the disease. ▮

Even when a protein shows no significant similarity to other proteins with the BLAST algorithm, it may nevertheless share a short sequence with other proteins that is functionally important Such short segments recurring in many different proteins, referred to as **motifs,** generally have similar functions. Several such motifs are described in Chapter 3 (see Figure 3-6). To search for these and other motifs in a new protein, researchers compare the query protein sequence with a database of known motif sequences. Table 9-2 summarizes several of the more commonly occurring motifs.

```
NF1   841  TRATFMEVLTKILQQGTEFDTLAETVLADRFERLVELVTMMGDQGELPIA  890
Ira  1500  IRIAFLRVFIDIV...TNYPVNPEKHEMDKMLAIDDFLKYIIKNPILAFF 1546

      891  MALANVVPCSQWDELARVLVTLFDSRHLLYQLLWNMFSKEVELADSMQTL  940
     1547  GSLA..CSPADVDLYAGGFLNAFDTRNASHILVTELLKQEIKRAARSDDI 1594

      941  FRGNSLASKIMTFCFKVYGATYLQKLLDPLLRIVITSSDWQHVSFEVDPT  990
     1595  LRRNSCATRALSLYTRSRGNKYLIKTLRPVLQGIVDNKE....SFEIDE. 1638

      991  RLEPSESLEENQRNLLQMTEKF....FHAIISSSSEFPPQLRSVCHCLYQ 1036
     1639  KMKPG...SENSEKMLDLFEKYMTRLIDAITSSIDDFPIELVDICKTIYN 1685

     1037  VVSQRFPQNSIGAVGSAMFLRFINPAIVSPYEAGILDKKPPPRIERGLKL 1086
     1686  AASVNFPEYAYIAVGSFVFLRFIGPALVSPDSENII.IVTHAHDRKPFIT 1734

     1087  MSKILQSIAN........HVLFTKEEHMRPFND....FVKSNFDAARRFF 1124
     1735  LAKVIQSLANGRENIFKKDILVSKEEFLKTCSDKIFNFLSELCKIPTNNF 1784

     1125  LDIASDCPTSDAVNHSL..............SFISDGNVLALHRLLWNN. 1159
     1785  TVNVREDPTPISFDYSFLHKFFYLNEFTIRKEIINESKLPGEFSFLKNTV 1834

     1160  ..QEKIGQYLSSNRDHKAVGRRPF....DKMATLLAYLGPPEHKPVA  1200
     1835  MLNDKILGVLGQPSMEIKNEIPPFVVENREKYPSLYEFMSRYAFKKVD 1882
```

▲ FIGURE 9-31 **Comparison of the regions of human NF1 protein and *S. cerevisiae* Ira protein that show significant sequence similarity.** The NF1 and the Ira sequences are shown on the top and bottom lines of each row, respectively, in the one-letter amino acid code (see Figure 2-13). Amino acids that are identical in the two proteins are highlighted in yellow. Amino acids with chemically similar but nonidentical side chains are connected by a blue dot. Amino acid numbers in the protein sequences are shown at the left and right ends of each row. Dots indicate "gaps" in the protein sequence inserted in order to maximize the alignment of homologous amino acids. The BLAST *p*-value for these two sequences is 10^{-28}, indicating a high degree of similarity. [From Xu et al., 1990, *Cell* **62**:599.]

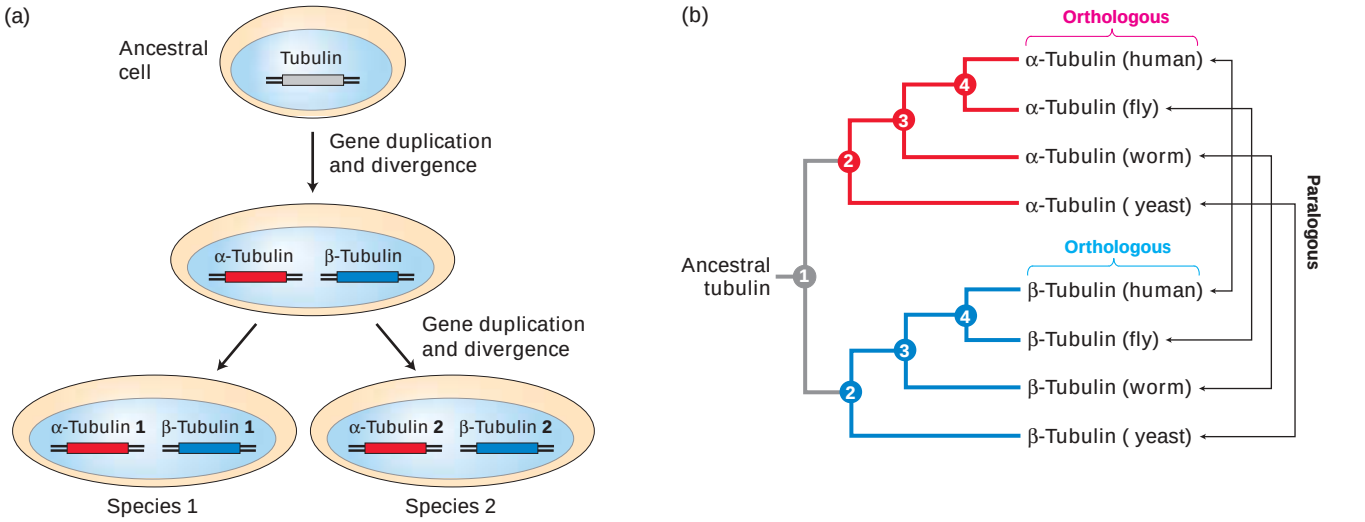
TABLE 9-2 Protein Sequence Motifs

Name	Sequence*	Function
ATP/GTP binding	[A,G]-X_4-G-K-[S,T]	Residues within a nucleotide-binding domain that contact the nucleotide
Prenyl-group binding site	C-Ø-Ø-X (C-terminus)	C-terminal sequence covalently attached to isoprenoid lipids in some lipid-anchored proteins (e.g., Ras)
Zinc finger (C_2H_2 type)	C-X_{2-4}-C-X_3-Ø-X_8-H-X_{3-5}-H	Zn^{2+}-binding sequence within DNA- or RNA-binding domain of some proteins
DEAD box	$Ø_2$-D-E-A-D-[R,K,E,N]-Ø	Sequence present in many ATP-dependent RNA helicases
Heptad repeat	(Ø-X_2-Ø-X_3)$_n$	Repeated sequence in proteins that form coiled-coil structures

*Single-letter amino acid abbreviations used for sequences (see Figure 2-13). X = any residue; Ø = hydrophobic residue. Brackets enclose alternative permissible residues.

Comparison of Related Sequences from Different Species Can Give Clues to Evolutionary Relationships Among Proteins

BLAST searches for related protein sequences may reveal that proteins belong to a **protein family.** (The corresponding genes constitute a **gene family.**) Protein families are thought to arise by two different evolutionary processes, *gene duplication* and *speciation*, discussed in Chapter 10. Consider, for example, the tubulin family of proteins, which constitute the basic subunits of microtubules. According to the simplified scheme in Figure 9-32a, the earliest eukaryotic cells are thought to have contained a single tubulin gene that was duplicated early in evolution; subsequent divergence of the different copies of the



▲ FIGURE 9-32 **The generation of diverse tubulin sequences during the evolution of eukaryotes.** (a) Probable mechanism giving rise to the tubulin genes found in existing species. It is possible to deduce that a gene duplication event occurred before speciation because the α-tubulin sequences from different species (e.g., humans and yeast) are more alike than are the α-tubulin and β-tubulin sequences within a species. (b) A phylogenetic tree representing the relationship between the tubulin sequences. The branch points (nodes), indicated by small numbers, represent common ancestral genes at the time that two sequences diverged. For example, node 1 represents the duplication event that gave rise to the α-tubulin and β-tubulin families, and node 2 represents the divergence of yeast from multicellular species. Braces and arrows indicate, respectively, the orthologous tubulin genes, which differ as a result of speciation, and the paralogous genes, which differ as a result of gene duplication. This diagram is simplified somewhat because each of the species represented actually contains multiple α-tubulin and β-tubulin genes that arose from later gene duplication events.

original tubulin gene formed the ancestral versions of the α- and β-tubulin genes. As different species diverged from these early eukaryotic cells, each of these gene sequences further diverged, giving rise to the slightly different forms of α-tubulin and β-tubulin now found in each species.

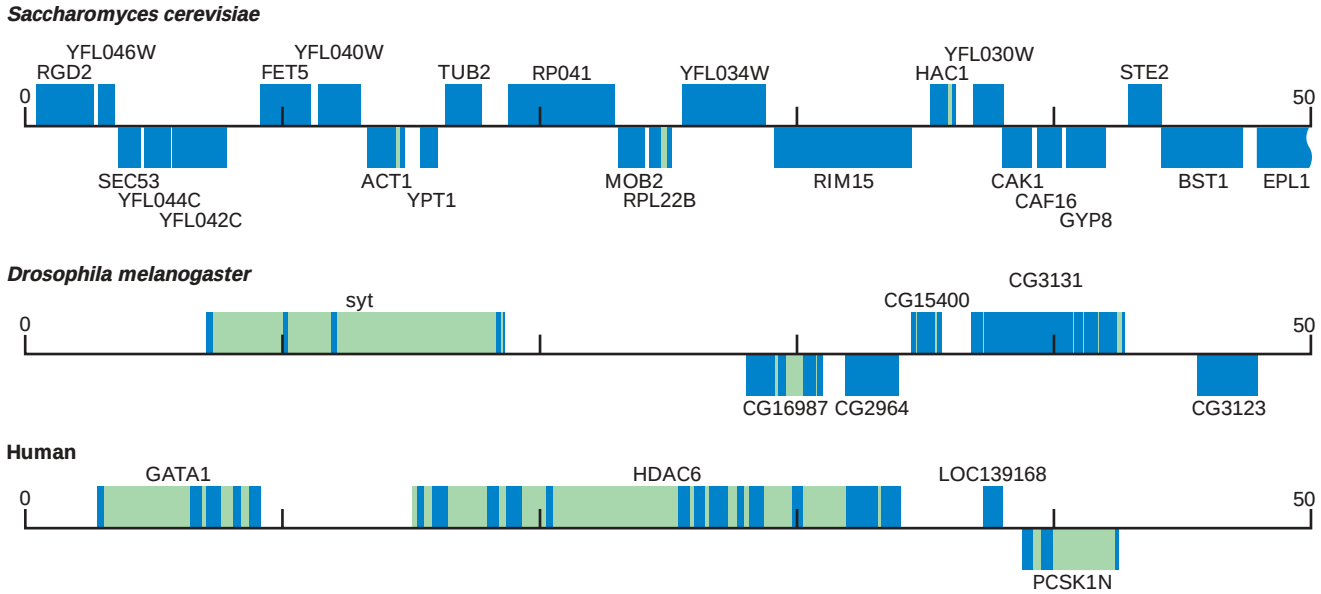
All the different members of the tubulin family are sufficiently similar in sequence to suggest a common ancestral sequence. Thus all these sequences are considered to be *homologous*. More specifically, sequences that presumably diverged as a result of gene duplication (e.g., the α- and β-tubulin sequences) are described as *paralogous*. Sequences that arose because of speciation (e.g., the α-tubulin genes in different species) are described as *orthologous*. From the degree of sequence relatedness of the tubulins present in different organisms today, evolutionary relationships can deduced, as illustrated in Figure 9-32b. Of the three types of sequence relationships, orthologous sequences are the most likely to share the same function.

Genes Can Be Identified Within Genomic DNA Sequences

The complete genomic sequence of an organism contains within it the information needed to deduce the sequence of every protein made by the cells of that organism. For organisms such as bacteria and yeast, whose genomes have few introns and short intergenic regions, most protein-coding sequences can be found simply by scanning the genomic sequence for **open reading frames (ORFs)** of significant length. An ORF usually is defined as a stretch of DNA containing at least 100 codons that begins with a start codon and ends with a stop codon. Because the probability that a random DNA sequence will contain no stop codons for 100 codons in a row is very small, most ORFs encode a protein.

ORF analysis correctly identifies more than 90 percent of the genes in yeast and bacteria. Some of the very shortest genes are missed by this method, and occasionally long open reading frames that are not actually genes arise by chance. Both types of miss assignments can be corrected by more sophisticated analysis of the sequence and by genetic tests for gene function. Of the *Saccharomyces* genes identified in this manner, about half were already known by some functional criterion such as mutant phenotype. The functions of some of the proteins encoded by the remaining putative genes identified by ORF analysis have been assigned based on their sequence similarity to known proteins in other organisms.

Identification of genes in organisms with a more complex genome structure requires more sophisticated algorithms than searching for open reading frames. Figure 9-33 shows a comparison of the genes identified in a representative 50-kb segment from the genomes of yeast, *Drosophila*, and humans. Because most genes in higher eukaryotes, including humans and *Drosophila*, are composed of multiple, relatively short coding regions (**exons**) separated by noncoding



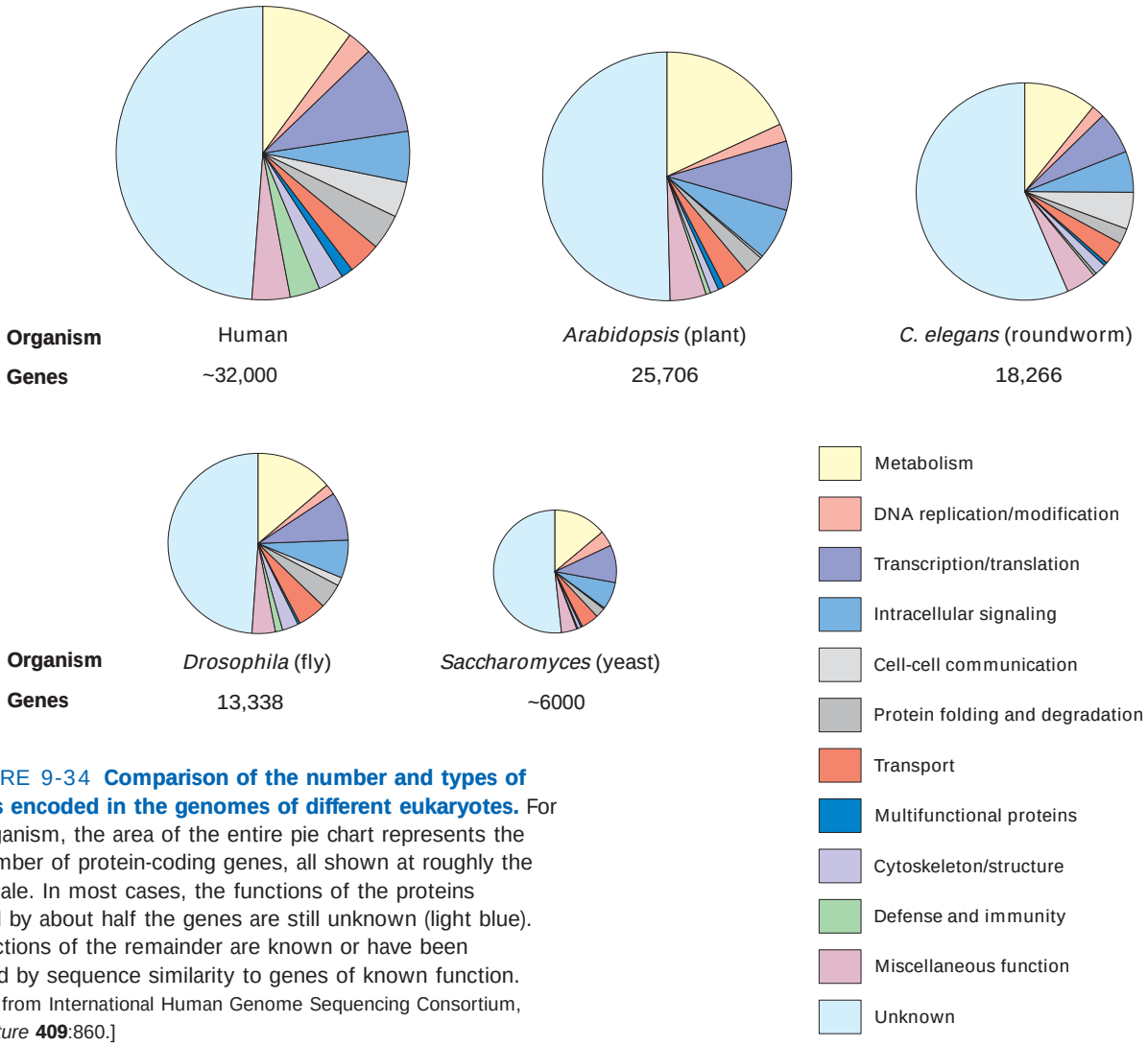
▲ FIGURE 9-33 **Arrangement of gene sequences in representative 50-kb segments of yeast, fruit fly, and human genomes.** Genes above the line are transcribed to the right; genes below the line are transcribed to the left. Blue blocks represent exons (coding sequences); green blocks represent introns (noncoding sequences). Because yeast genes contain few if any introns, scanning genomic sequences for open reading frames (ORFs) correctly identifies most gene sequences. In contrast, the genes of higher eukaryotes typically comprise multiple exons separated by introns. ORF analysis is not effective in identifying genes in these organisms. Likely gene sequences for which no functional data are available are designated by numerical names: in yeast, these begin with Y; in *Drosophila,* with CG; and in humans, with LOC. The other genes shown here encode proteins with known functions.

regions (**introns**), scanning for ORFs is a poor method for finding genes. The best gene-finding algorithms combine all the available data that might suggest the presence of a gene at a particular genomic site. Relevant data include alignment or hybridization to a full-length cDNA; alignment to a partial cDNA sequence, generally 200–400 bp in length, known as an *expressed sequence tag (EST)*; fitting to models for exon, intron, and splice site sequences; and sequence similarity to other organisms. Using these methods computational biologists have identified approximately 35,000 genes in the human genome, although for as many as 10,000 of these putative genes there is not yet conclusive evidence that they actually encode proteins or RNAs.

A particularly powerful method for identifying human genes is to compare the human genomic sequence with that of the mouse. Humans and mice are sufficiently related to have most genes in common; however, largely nonfunctional DNA sequences, such as intergenic regions and introns, will tend to be very different because they are not under strong selective pressure. Thus corresponding segments of the human and mouse genome that exhibit high sequence similarity are likely to be functional coding regions (i.e., exons).

The Size of an Organism's Genome Is Not Directly Related to Its Biological Complexity

The combination of genomic sequencing and gene-finding computer algorithms has yielded the complete inventory of protein-coding genes for a variety of organisms. Figure 9-34 shows the total number of protein-coding genes in several eukaryotic genomes that have been completely sequenced. The functions of about half the proteins encoded in these genomes are known or have been predicted on the basis of sequence comparisons. One of the surprising features of this comparison is that the number of protein-coding genes within different organisms does not seem proportional to our intuitive sense of their biological complexity. For example, the roundworm *C. elegans* apparently has more genes than the fruit fly *Drosophila*, which has a much more complex body plan and more complex behavior. And humans have



▲ FIGURE 9-34 **Comparison of the number and types of proteins encoded in the genomes of different eukaryotes.** For each organism, the area of the entire pie chart represents the total number of protein-coding genes, all shown at roughly the same scale. In most cases, the functions of the proteins encoded by about half the genes are still unknown (light blue). The functions of the remainder are known or have been predicted by sequence similarity to genes of known function. [Adapted from International Human Genome Sequencing Consortium, 2001, *Nature* **409**:860.]

fewer than twice the number of genes as *C. elegans*, which seems completely inexplicable given the enormous differences between these organisms.

Clearly, simple quantitative differences in the genomes of different organisms are inadequate for explaining differences in biological complexity. However, several phenomena can generate more complexity in the expressed proteins of higher eukaryotes than is predicted from their genomes. First, alternative splicing of a pre-mRNA can yield multiple functional mRNAs corresponding to a particular gene (Chapter 12). Second, variations in the post-translational modification of some proteins may produce functional differences. Finally, qualitative differences in the interactions between proteins and their integration into pathways may contribute significantly to the differences in biological complexity among organisms. The specific functions of many genes and proteins identified by analysis of genomic sequences still have not been determined. As researchers unravel the functions of individual proteins in different organisms and further detail their interactions, a more sophisticated understanding of the genetic basis of complex biological systems will emerge.

DNA Microarrays Can Be Used to Evaluate the Expression of Many Genes at One Time

Monitoring the expression of thousands of genes simultaneously is possible with **DNA microarray** analysis. A DNA microarray consists of thousands of individual, closely packed gene-specific sequences attached to the surface of a glass microscopic slide. By coupling microarray analysis with the results from genome sequencing projects, researchers can analyze the global patterns of gene expression of an organism during specific physiological responses or developmental processes.

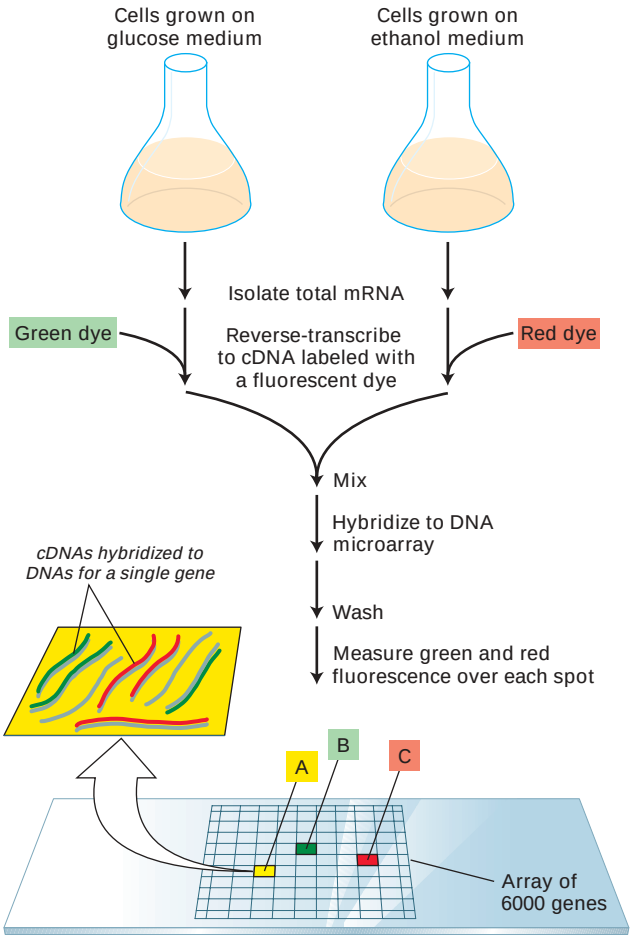
Preparation of DNA Microarrays

In one method for preparing microarrays, a ≈1-kb portion of the coding region of each gene analyzed is individually amplified by PCR. A robotic device is used to apply each amplified DNA sample to the surface of a glass microscope slide, which then is chemically processed to permanently attach the DNA sequences to the glass surface and to denature them. A typical array might contain ≈6000 spots of DNA in a 2 × 2 cm grid.

In an alternative method, multiple DNA oligonucleotides, usually at least 20 nucleotides in length, are synthesized from an initial nucleotide that is covalently bound to the surface of a glass slide. The synthesis of an oligonucleotide of specific sequence can be programmed in a small region on the surface of the slide. Several oligonucleotide sequences from a single gene are thus synthesized in neighboring regions of the slide to analyze expression of that gene. With this method, oligonucleotides representing thousands of genes can be produced on a single glass slide. Because the methods for constructing these arrays of synthetic oligonucleotides were adapted from methods for manufacturing microscopic integrated circuits used in computers, these types of oligonucleotide microarrays are often called *DNA chips*.

Effect of Carbon Source on Gene Expression in Yeast

The initial step in a microarray expression study is to prepare fluorescently labeled cDNAs corresponding to the mRNAs expressed by the cells under study. When the cDNA preparation is applied to a microarray, spots representing genes



▲ EXPERIMENTAL FIGURE 9-35 **DNA microarray analysis can reveal differences in gene expression in yeast cells under different experimental conditions.** In this example, cDNA prepared from mRNA isolated from wild-type *Saccharomyces* cells grown on glucose or ethanol is labeled with different fluorescent dyes. A microarray composed of DNA spots representing each yeast gene is exposed to an equal mixture of the two cDNA preparations under hybridization conditions. The ratio of the intensities of red and green fluorescence over each spot, detected with a scanning confocal laser microscope, indicates the relative expression of each gene in cells grown on each of the carbon sources. Microarray analysis also is useful for detecting differences in gene expression between wild-type and mutant strains.

that are expressed will hybridize under appropriate conditions to their complementary cDNAs and can subsequently be detected in a scanning laser microscope.

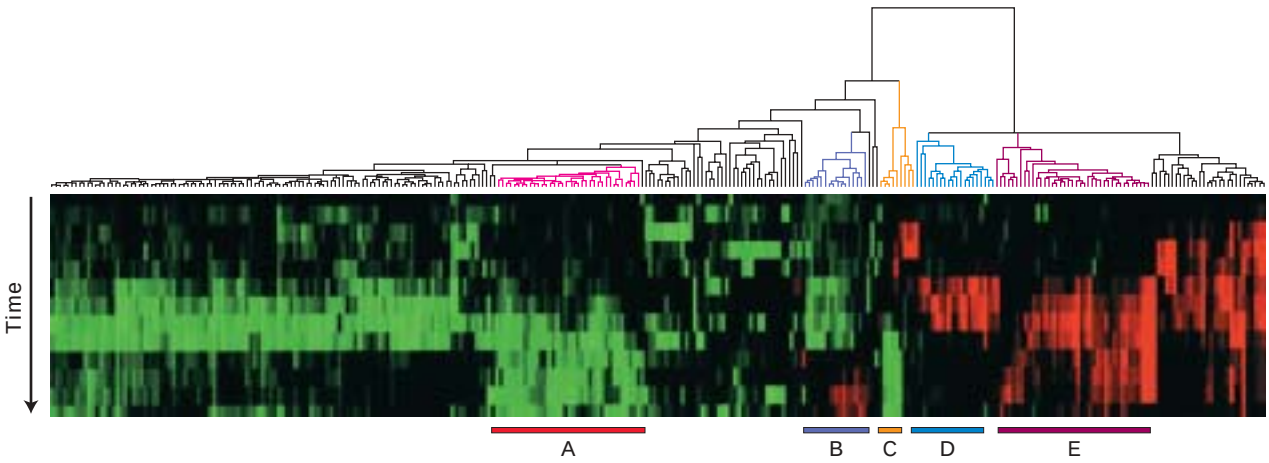
Figure 9-35 depicts how this method can be applied to compare gene expression in yeast cells growing on glucose versus ethanol as the source of carbon and energy. In this type of experiment, the separate cDNA preparations from glucose-grown and ethanol-grown cells are labeled with differently colored fluorescent dyes. A DNA array comprising all 6000 genes then is incubated with a mixture containing equal amounts of the two cDNA preparations under hybridization conditions. After unhybridized cDNA is washed away, the intensity of green and red fluorescence at each DNA spot is measured using a fluorescence microscope and stored in computer files under the name of each gene according to its known position on the slide. The relative intensities of red and green fluorescence signals at each spot are a measure of the relative level of expression of that gene in cells grown in glucose or ethanol. Genes that are not transcribed under these growth conditions give no detectable signal.

Hybridization of fluorescently labeled cDNA preparations to DNA microarrays provides a means for analyzing gene expression patterns on a genomic scale. This type of analysis has shown that as yeast cells shift from growth on glucose to growth on ethanol, expression of 710 genes increases by a factor of two or more, while expression of 1030 genes decreases by a factor of two or more. Although about 400 of the differentially expressed genes have no known function, these results provide the first clue as to their possible function in yeast biology.

Cluster Analysis of Multiple Expression Experiments Identifies Co-regulated Genes

Firm conclusions rarely can be drawn from a single microarray experiment about whether genes that exhibit similar changes in expression are co-regulated and hence likely to be closely related functionally. For example, many of the observed differences in gene expression just described in yeast growing on glucose or ethanol could be indirect consequences of the many different changes in cell physiology that occur when cells are transferred from one medium to another. In other words, genes that appear to be co-regulated in a single microarray expression experiment may undergo changes in expression for very different reasons and may actually have very different biological functions. A solution to this problem is to combine the information from a set of expression array experiments to find genes that are similarly regulated under a variety of conditions or over a period of time.

This more informative use of multiple expression array experiments is illustrated by the changes in gene expression observed after starved human fibroblasts are transferred to a rich, serum-containing, growth medium. In one study, the relative expression of 8600 genes was determined at different



▲ EXPERIMENTAL FIGURE 9-36 **Cluster analysis of data from multiple microarray expression experiments can identify co-regulated genes.** In this experiment, the expression of 8600 mammalian genes was detected by microarray analysis at time intervals over a 24-hour period after starved fibroblasts were provided with serum. The cluster diagram shown here is based on a computer algorithm that groups genes showing similar changes in expression compared with a starved control sample over time. Each column of colored boxes represents a single gene, and each row represents a time point. A red box indicates an increase in expression relative to the control; a green box, a decrease in expression; and a black box, no significant change in expression. The "tree" diagram at the top shows how the expression patterns for individual genes can be organized in a hierarchical fashion to group together the genes with the greatest similarity in their patterns of expression over time. Five clusters of coordinately regulated genes were identified in this experiment, as indicated by the bars at the bottom. Each cluster contains multiple genes whose encoded proteins function in a particular cellular process: cholesterol biosynthesis (A), the cell cycle (B), the immediate-early response (C), signaling and angiogenesis (D), and wound healing and tissue remodeling (E). [Courtesy of Michael B. Eisen, Lawrence Berkeley National Laboratory.]

times after serum addition, generating more than 10^4 individual pieces of data. A computer program, related to the one used to determine the relatedness of different protein sequences, can organize these data and cluster genes that show similar expression over the time course after serum addition. Remarkably, such *cluster analysis* groups sets of genes whose encoded proteins participate in a common cellular process, such as cholesterol biosynthesis or the cell cycle (Figure 9-36).

Since genes with identical or similar patterns of regulation generally encode functionally related proteins, cluster analysis of multiple microarray expression experiments is another tool for deducing the functions of newly identified genes. This approach allows any number of different experiments to be combined. Each new experiment will refine the analysis, with smaller and smaller cohorts of genes being identified as belonging to different clusters.

KEY CONCEPTS OF SECTION 9.4

Genomics: Genome-wide Analysis of Gene Structure and Expression

- The function of a protein that has not been isolated often can be predicted on the basis of similarity of its amino acid sequence to proteins of known function.
- A computer algorithm known as BLAST rapidly searches databases of known protein sequences to find those with significant similarity to a new (query) protein.
- Proteins with common functional motifs may not be identified in a typical BLAST search. These short sequences may be located by searches of motif databases.
- A protein family comprises multiple proteins all derived from the same ancestral protein. The genes encoding these proteins, which constitute the corresponding gene family, arose by an initial gene duplication event and subsequent divergence during speciation (see Figure 9-32).
- Related genes and their encoded proteins that derive from a gene duplication event are paralogous; those that derive from speciation are orthologous. Proteins that are orthologous usually have a similar function.
- Open reading frames (ORFs) are regions of genomic DNA containing at least 100 codons located between a start codon and stop codon.
- Computer search of the entire bacterial and yeast genomic sequences for open reading frames (ORFs) correctly identifies most protein-coding genes. Several types of additional data must be used to identify probable genes in the genomic sequences of humans and other higher eukaryotes because of the more complex gene structure in these organisms.
- Analysis of the complete genome sequences for several different organisms indicates that biological complexity is not directly related to the number of protein-coding genes (see Figure 9-34).
- DNA microarray analysis simultaneously detects the relative level of expression of thousands of genes in different types of cells or in the same cells under different conditions (see Figure 9-35).
- Cluster analysis of the data from multiple microarray expression experiments can identify genes that are similarly regulated under various conditions. Such co-regulated genes commonly encode proteins that have biologically related functions.

9.5 Inactivating the Function of Specific Genes in Eukaryotes

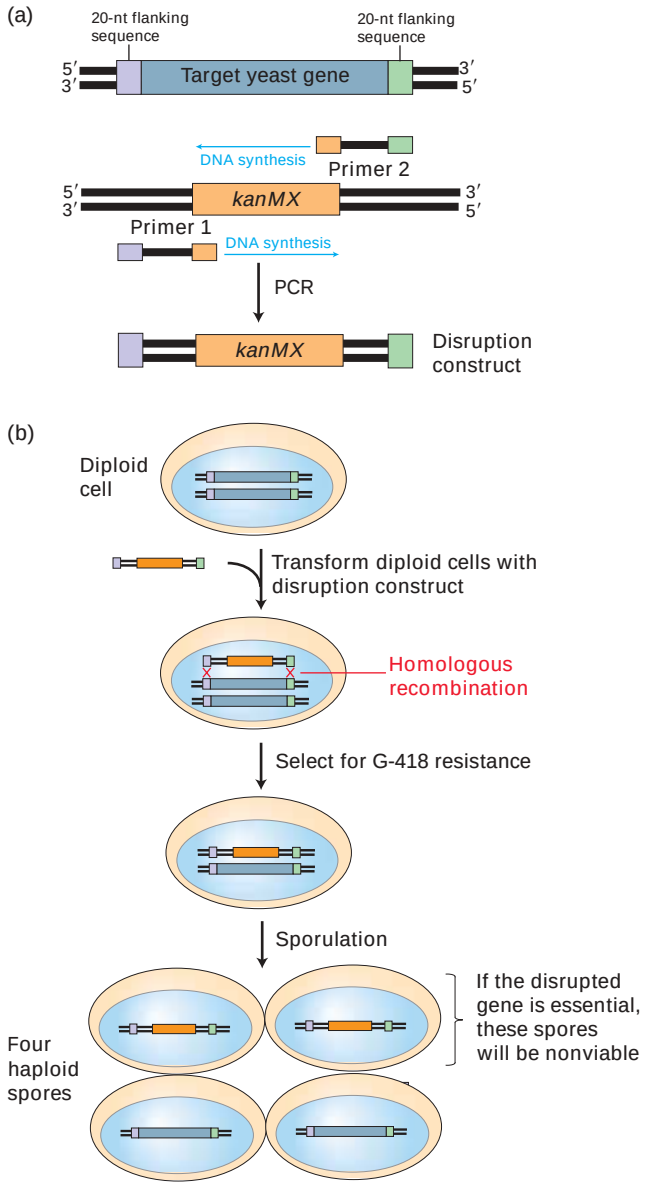
The elucidation of DNA and protein sequences in recent years has led to identification of many genes, using sequence patterns in genomic DNA and the sequence similarity of the encoded proteins with proteins of known function. As discussed in the previous section, the general functions of proteins identified by sequence searches may be predicted by analogy with known proteins. However, the precise in vivo roles of such "new" proteins may be unclear in the absence of mutant forms of the corresponding genes. In this section, we describe several ways for disrupting the normal function of a specific gene in the genome of an organism. Analysis of the resulting mutant phenotype often helps reveal the in vivo function of the normal gene and its encoded protein.

Three basic approaches underlie these gene-inactivation techniques: (1) replacing a normal gene with other sequences; (2) introducing an allele whose encoded protein inhibits functioning of the expressed normal protein; and (3) promoting destruction of the mRNA expressed from a gene. The normal endogenous gene is modified in techniques based on the first approach but is not modified in the other approaches.

Normal Yeast Genes Can Be Replaced with Mutant Alleles by Homologous Recombination

Modifying the genome of the yeast *Saccharomyces* is particularly easy for two reasons: yeast cells readily take up exogenous DNA under certain conditions, and the introduced DNA is efficiently exchanged for the homologous chromosomal site in the recipient cell. This specific, targeted **recombination** of identical stretches of DNA allows any gene in yeast chromosomes to be replaced with a mutant allele. (As we discuss in Section 9.6, recombination between homologous chromosomes also occurs naturally during meiosis.)

In one popular method for disrupting yeast genes in this fashion, PCR is used to generate a *disruption construct* containing a selectable marker that subsequently is transfected into yeast cells. As shown in Figure 9-37a, primers for PCR amplification of the selectable marker are designed to include about 20 nucleotides identical with sequences flanking the yeast gene to be replaced. The resulting amplified construct comprises the selectable marker (e.g., the *kanMX* gene,



▲ EXPERIMENTAL FIGURE 9-37 **Homologous recombination with transfected disruption constructs can inactivate specific target genes in yeast.** (a) A suitable construct for disrupting a target gene can be prepared by the PCR. The two primers designed for this purpose each contain a sequence of about 20 nucleotides (nt) that is homologous to one end of the target yeast gene as well as sequences needed to amplify a segment of DNA carrying a selectable marker gene such as *kanMX*, which confers resistance to G-418. (b) When recipient diploid *Saccharomyces* cells are transformed with the gene disruption construct, homologous recombination between the ends of the construct and the corresponding chromosomal sequences will integrate the *kanMX* gene into the chromosome, replacing the target gene sequence. The recombinant diploid cells will grow on a medium containing G-418, whereas nontransformed cells will not. If the target gene is essential for viability, half the haploid spores that form after sporulation of recombinant diploid cells will be nonviable.

which like *neo*^{r} confers resistance to G-418) flanked by about 20 base pairs that match the ends of the target yeast gene. Transformed diploid yeast cells in which one of the two copies of the target endogenous gene has been replaced by the disruption construct are identified by their resistance to G-418 or other selectable phenotype. These heterozygous diploid yeast cells generally grow normally regardless of the function of the target gene, but half the haploid spores derived from these cells will carry only the disrupted allele (Figure 9-37b). If a gene is essential for viability, then spores carrying a disrupted allele will not survive.

Disruption of yeast genes by this method is proving particularly useful in assessing the role of proteins identified by ORF analysis of the entire genomic DNA sequence. A large consortium of scientists has replaced each of the approximately 6000 genes identified by ORF analysis with the *kanMX* disruption construct and determined which gene disruptions lead to nonviable haploid spores. These analyses have shown that about 4500 of the 6000 yeast genes are not required for viability, an unexpectedly large number of apparently nonessential genes. In some cases, disruption of a particular gene may give rise to subtle defects that do not compromise the viability of yeast cells growing under laboratory conditions. Alternatively, cells carrying a disrupted gene may be viable because of operation of backup or compensatory pathways. To investigate this possibility, yeast geneticists currently are searching for synthetic lethal mutations that might reveal nonessential genes with redundant functions (see Figure 9-9c).

Transcription of Genes Ligated to a Regulated Promoter Can Be Controlled Experimentally

Although disruption of an essential gene required for cell growth will yield nonviable spores, this method provides little information about what the encoded protein actually does in cells. To learn more about how a specific gene contributes to cell growth and viability, investigators must be able to selectively inactivate the gene in a population of growing cells. One method for doing this employs a regulated promoter to selectively shut off transcription of an essential gene.

A useful promoter for this purpose is the yeast *GAL1* promoter, which is active in cells grown on galactose but completely inactive in cells grown on glucose. In this approach, the coding sequence of an essential gene (*X*) ligated to the *GAL1* promoter is inserted into a yeast shuttle vector (see Figure 9-19a). The recombinant vector then is introduced into haploid yeast cells in which gene *X* has been disrupted. Haploid cells that are transformed will grow on galactose medium, since the normal copy of gene *X* on the vector is expressed in the presence of galactose. When the cells are transferred to a glucose-containing medium, gene *X* no longer is transcribed; as the cells divide, the amount of the encoded protein X gradually declines, eventually reaching a state of depletion that mimics a complete loss-of-function mutation. The observed changes in the phenotype of these cells after the shift to glucose medium may suggest

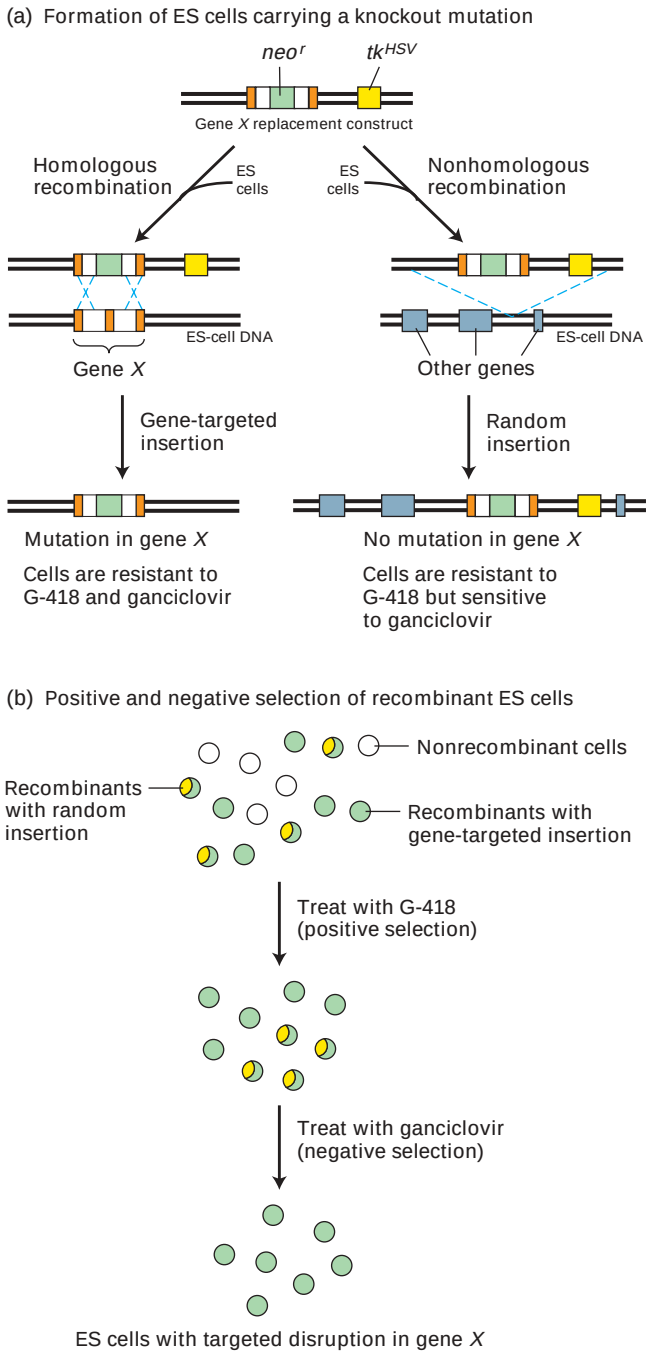
which cell processes depend on the protein encoded by the essential gene *X*.

In an early application of this method, researchers explored the function of cytosolic *Hsc70* genes in yeast. Haploid cells with a disruption in all four redundant *Hsc70* genes were nonviable, unless the cells carried a vector containing a copy of the *Hsc70* gene that could be expressed from the *GAL1* promoter on galactose medium. On transfer to glucose, the vector-carrying cells eventually stopped growing because of insufficient Hsc70 activity. Careful examination of these dying cells revealed that their secretory proteins could no longer enter the endoplasmic reticulum (ER). This study provided the first evidence for the unexpected role of Hsc70 protein in translocation of secretory proteins into the ER, a process examined in detail in Chapter 16.

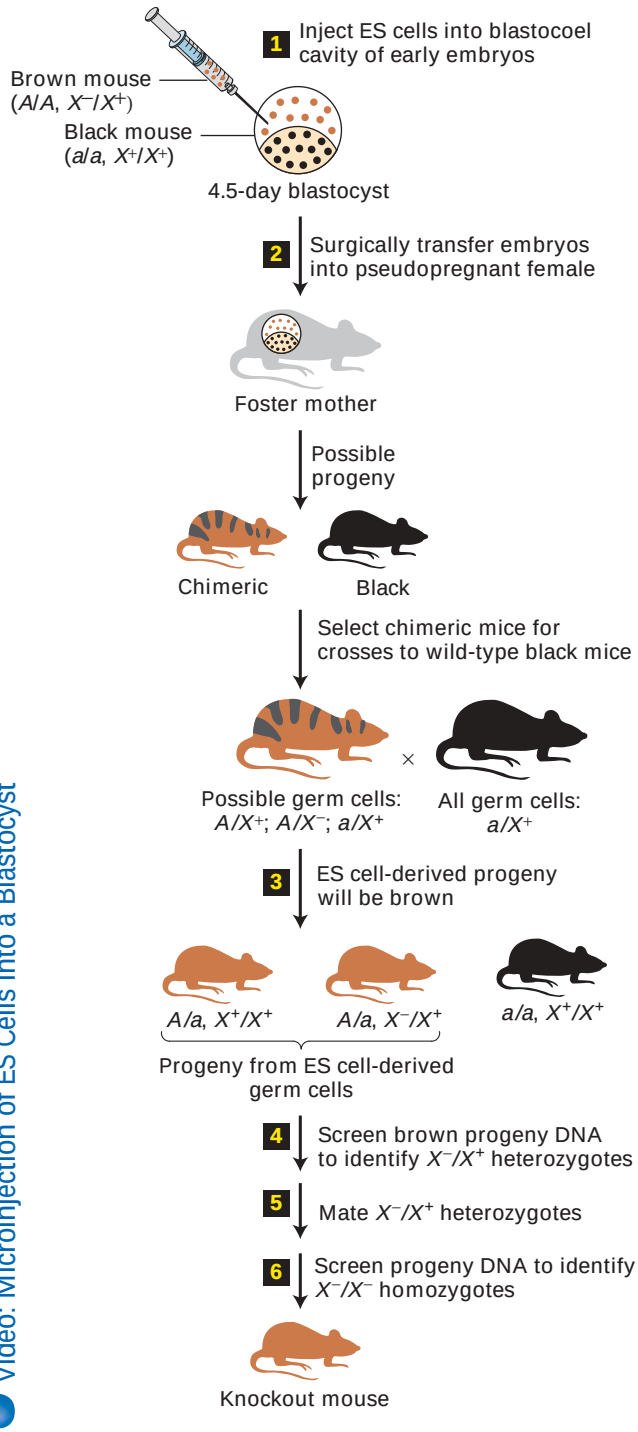
Specific Genes Can Be Permanently Inactivated in the Germ Line of Mice

Many of the methods for disrupting genes in yeast can be applied to genes of higher eukaryotes. These genes can be introduced into the **germ line** via homologous recombination to produce animals with a **gene knockout,** or simply "knockout." Knockout mice in which a specific gene is disrupted are a powerful experimental system for studying mammalian development, behavior, and physiology. They also are useful in studying the molecular basis of certain human genetic diseases.

Gene-targeted knockout mice are generated by a two-stage procedure. In the first stage, a DNA construct containing a disrupted allele of a particular target gene is introduced into *embryonic stem (ES) cells.* These cells, which are derived from the blastocyst, can be grown in culture through many generations (see Figure 22-3). In a small fraction of transfected cells, the introduced DNA undergoes homologous recombination with the target gene, although recombination at nonhomologous chromosomal sites occurs much more frequently. To select for cells in which homologous gene-targeted insertion occurs, the recombinant DNA construct introduced into ES cells needs to include two selectable marker genes (Figure 9-38). One of these genes (*neo*^{r}), which



▶ EXPERIMENTAL FIGURE 9-38 **Isolation of mouse ES cells with a gene-targeted disruption is the first stage in production of knockout mice.** (a) When exogenous DNA is introduced into embryonic stem (ES) cells, random insertion via nonhomologous recombination occurs much more frequently than gene-targeted insertion via homologous recombination. Recombinant cells in which one allele of gene *X* (orange and white) is disrupted can be obtained by using a recombinant vector that carries gene *X* disrupted with *neo*^{r} (green), which confers resistance to G-418, and, outside the region of homology, *tk*^{HSV} (yellow), the thymidine kinase gene from herpes simplex virus. The viral thymidine kinase, unlike the endogenous mouse enzyme, can convert the nucleotide analog ganciclovir into the monophosphate form; this is then modified to the triphosphate form, which inhibits cellular DNA replication in ES cells. Thus ganciclovir is cytotoxic for recombinant ES cells carrying the *tk*^{HSV} gene. Nonhomologous insertion includes the *tk*^{HSV} gene, whereas homologous insertion does not; therefore, only cells with nonhomologous insertion are sensitive to ganciclovir. (b) Recombinant cells are selected by treatment with G-418, since cells that fail to pick up DNA or integrate it into their genome are sensitive to this cytotoxic compound. The surviving recombinant cells are treated with ganciclovir. Only cells with a targeted disruption in gene *X,* and therefore lacking the *tk*^{HSV} gene, will survive. [See S. L. Mansour et al., 1988, *Nature* **336**:348.]



MEDIA CONNECTIONS

Video: Microinjection of ES Cells into a Blastocyst

◀ EXPERIMENTAL FIGURE 9-39 **ES cells heterozygous for a disrupted gene are used to produce gene-targeted knockout mice.** Step 1: Embryonic stem (ES) cells heterozygous for a knockout mutation in a gene of interest (*X*) and homozygous for a dominant allele of a marker gene (here, brown coat color, *A*) are transplanted into the blastocoel cavity of 4.5-day embryos that are homozygous for a recessive allele of the marker (here, black coat color, *a*). Step 2: The early embryos then are implanted into a pseudopregnant female. Those progeny containing ES-derived cells are chimeras, indicated by their mixed black and brown coats. Step 3: Chimeric mice then are backcrossed to black mice; brown progeny from this mating have ES-derived cells in their germ line. Steps 4–6: Analysis of DNA isolated from a small amount of tail tissue can identify brown mice heterozygous for the knockout allele. Intercrossing of these mice produces some individuals homozygous for the disrupted allele, that is, knockout mice. [Adapted from M. R. Capecchi, 1989, *Trends Genet.* **5**:70.]

confers G-418 resistance, is inserted within the target gene (*X*), thereby disrupting it. The other selectable gene, the thymidine kinase gene from herpes simplex virus (tk^{HSV}), confers sensitivity to ganciclovir, a cytotoxic nucleotide analog; it is inserted into the construct outside the target-gene sequence. Only ES cells that undergo homologous recombination can survive in the presence of both G-418 and ganciclovir. In these cells one allele of gene *X* will be disrupted.

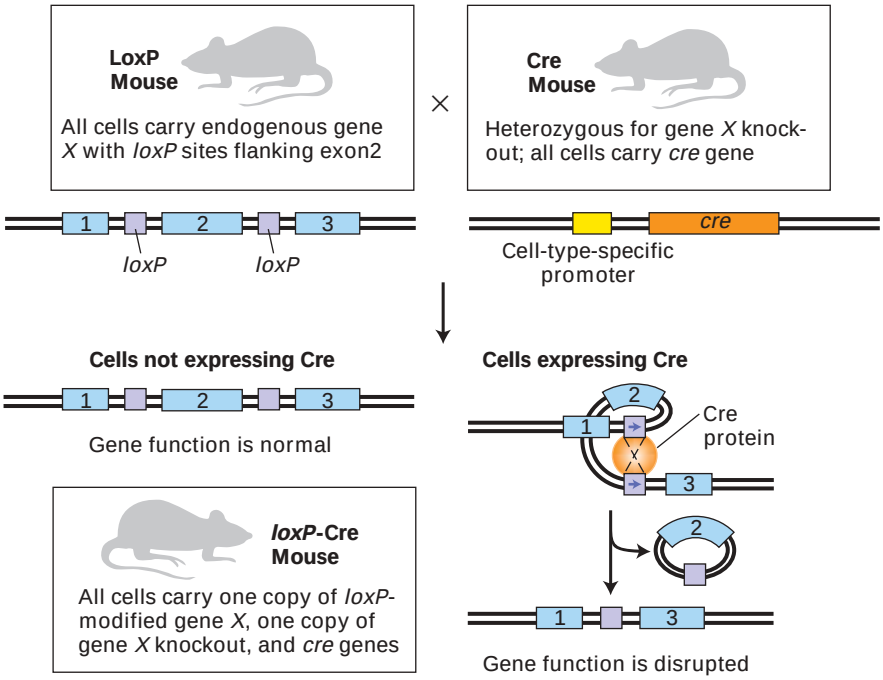
In the second stage in production of knockout mice, ES cells heterozygous for a knockout mutation in gene *X* are injected into a recipient wild-type mouse blastocyst, which subsequently is transferred into a surrogate pseudopregnant female mouse (Figure 9-39). The resulting progeny will be **chimeras,** containing tissues derived from both the transplanted ES cells and the host cells. If the ES cells also are homozygous for a visible marker trait (e.g., coat color), then chimeric progeny in which the ES cells survived and proliferated can be identified easily. Chimeric mice are then mated with mice homozygous for another allele of the marker trait to determine if the knockout mutation is incorporated into the germ line. Finally, mating of mice, each heterozygous for the knockout allele, will produce progeny homozygous for the knockout mutation.

MEDICINE

Development of knockout mice that mimic certain human diseases can be illustrated by cystic fibrosis. By methods discussed in Section 9.6, the recessive mutation that causes this disease eventually was shown to be located in a gene known as *CFTR*, which encodes a chloride channel. Using the cloned wild-type human *CFTR* gene, researchers isolated the homologous mouse gene and subsequently introduced mutations in it. The gene-knockout technique was then used to produce homozygous mutant mice, which showed symptoms (i.e., a phenotype), including disturbances to the functioning of epithelial cells, similar to those of humans with cystic fibrosis. These knockout mice are currently being used as a model system for studying this genetic disease and developing effective therapies. ▮

Somatic Cell Recombination Can Inactivate Genes in Specific Tissues

Investigators often are interested in examining the effects of knockout mutations in a particular tissue of the mouse, at a specific stage in development, or both. However, mice car-



▲ EXPERIMENTAL FIGURE 9-40 **The *loxP-Cre* recombination system can knock out genes in specific cell types.** Two *loxP* sites are inserted on each side of an essential exon (2) of the target gene *X* (blue) by homologous recombination, producing a *loxP* mouse. Since the *loxP* sites are in introns, they do not disrupt the function of *X*. The Cre mouse carries one gene *X* knockout allele and an introduced *cre* gene (orange) from bacteriophage P1 linked to a cell-type-specific promoter (yellow). The *cre* gene is incorporated into the mouse genome by nonhomologous recombination and does not affect the function of other genes. In the *loxP*-Cre mice that result from crossing, Cre protein is produced only in those cells in which the promoter is active. Thus these are the only cells in which recombination between the *loxP* sites catalyzed by Cre occurs, leading to deletion of exon 2. Since the other allele is a constitutive gene *X* knockout, deletion between the *loxP* sites results in complete loss of function of gene *X* in all cells expressing Cre. By using different promoters, researchers can study the effects of knocking out gene *X* in various types of cells.

rying a germ-line knockout may have defects in numerous tissues or die before the developmental stage of interest. To address this problem, mouse geneticists have devised a clever technique to inactivate target genes in specific types of somatic cells or at particular times during development.

This technique employs site-specific DNA recombination sites (called *loxP* sites) and the enzyme Cre that catalyzes recombination between them. The *loxP-Cre recombination system* is derived from bacteriophage P1, but this site-specific recombination system also functions when placed in mouse cells. An essential feature of this technique is that expression of Cre is controlled by a cell-type-specific promoter. In *loxP*-Cre mice generated by the procedure depicted in Figure 9-40, inactivation of the gene of interest (*X*) occurs only in cells in which the promoter controlling the *cre* gene is active.

An early application of this technique provided strong evidence that a particular neurotransmitter receptor is important for learning and memory. Previous pharmacological and physiological studies had indicated that normal learning requires the NMDA class of glutamate receptors in the hippocampus, a region of the brain. But mice in which the gene encoding an NMDA receptor subunit was knocked out died neonatally, precluding analysis of the receptor's role in learning. Following the protocol in Figure 9-40, researchers generated mice in which the receptor subunit gene was inactivated in the hippocampus but expressed in other tissues. These mice survived to adulthood and showed learning and memory defects, confirming a role for these receptors in the ability of mice to encode their experiences into memory.

Dominant-Negative Alleles Can Functionally Inhibit Some Genes

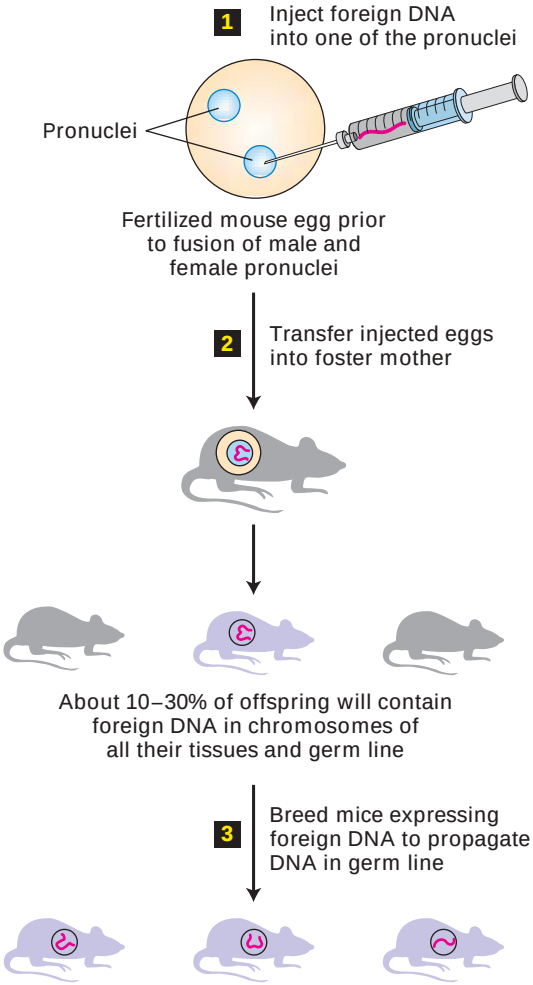
In diploid organisms, as noted in Section 9.1, the phenotypic effect of a recessive allele is expressed only in homozygous individuals, whereas dominant alleles are expressed in heterozygotes. That is, an individual must carry two copies of a recessive allele but only one copy of a dominant allele to exhibit the corresponding phenotypes. We have seen how strains of mice that are homozygous for a given recessive knockout mutation can be produced by crossing individuals that are heterozygous for the same knockout mutation (see Figure 9-39). For experiments with cultured animal cells,

however, it is usually difficult to disrupt both copies of a gene in order to produce a mutant phenotype. Moreover, the difficulty in producing strains with both copies of a gene mutated is often compounded by the presence of related genes of similar function that must also be inactivated in order to reveal an observable phenotype.

For certain genes, the difficulties in producing homozygous knockout mutants can be avoided by use of an allele carrying a **dominant-negative mutation.** These alleles are genetically dominant; that is, they produce a mutant phenotype even in cells carrying a wild-type copy of the gene. But unlike other types of dominant alleles, dominant-negative alleles produce a phenotype equivalent to that of a loss-of-function mutation.

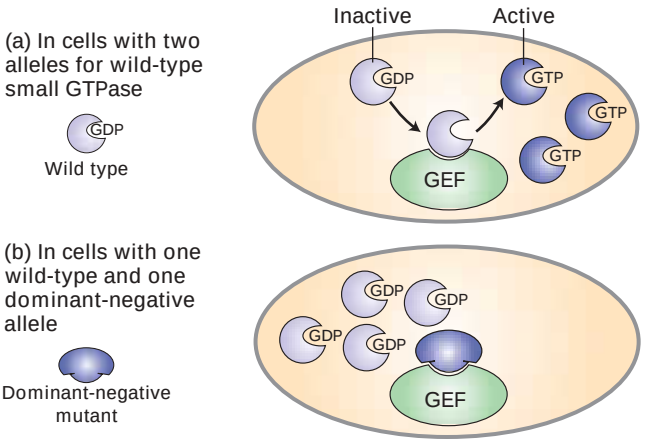
Useful dominant-negative alleles have been identified for a variety of genes and can be introduced into cultured cells by transfection or into the germ line of mice or other organisms. In both cases, the introduced gene is integrated into the genome by nonhomologous recombination. Such randomly inserted genes are called **transgenes;** the cells or organisms carrying them are referred to as **transgenic.** Transgenes carrying a dominant-negative allele usually are engineered so that the allele is controlled by a regulated promoter, allowing expression of the mutant protein in different tissues at different times. As noted above, the random integration of exogenous DNA via nonhomologous recombination occurs at a much higher frequency than insertion via homologous recombination. Because of this phenomenon, the production of transgenic mice is an efficient and straightforward process (Figure 9-41).

Among the genes that can be functionally inactivated by introduction of a dominant-negative allele are those encoding small (monomeric) GTP-binding proteins belonging to the GTPase superfamily. As we will examine in several later chapters, these proteins (e.g., Ras, Rac, and Rab) act as intracellular switches. Conversion of the small GTPases from an inactive GDP-bound state to an active GTP-bound state depends on their interacting with a corresponding guanine nucleotide exchange factor (GEF). A mutant small GTPase



▲ EXPERIMENTAL FIGURE 9-41 **Transgenic mice are produced by random integration of a foreign gene into the mouse germ line.** Foreign DNA injected into one of the two pronuclei (the male and female haploid nuclei contributed by the parents) has a good chance of being randomly integrated into the chromosomes of the diploid zygote. Because a transgene is integrated into the recipient genome by nonhomologous recombination, it does not disrupt endogenous genes. [See R. L. Brinster et al., 1981, *Cell* **27**:223.]

MEDIA CONNECTIONS
Technique Animation: Creating a Transgenic Mouse
Video: DNA Injected into a Pronucleus of a Mouse Zygote



▲ FIGURE 9-42 **Inactivation of the function of a wild-type GTPase by the action of a dominant-negative mutant allele.** (a) Small (monomeric) GTPases (purple) are activated by their interaction with a guanine-nucleotide exchange factor (GEF), which catalyzes the exchange of GDP for GTP. (b) Introduction of a dominant-negative allele of a small GTPase gene into cultured cells or transgenic animals leads to expression of a mutant GTPase that binds to and inactivates the GEF. As a result, endogenous wild-type copies of the same small GTPase are trapped in the inactive GDP-bound state. A single dominant-negative allele thus causes a loss-of-function phenotype in heterozygotes similar to that seen in homozygotes carrying two recessive loss-of-function alleles.

that permanently binds to the GEF protein will block conversion of endogenous wild-type small GTPases to the active GTP-bound state, thereby inhibiting them from performing their switching function (Figure 9-42).

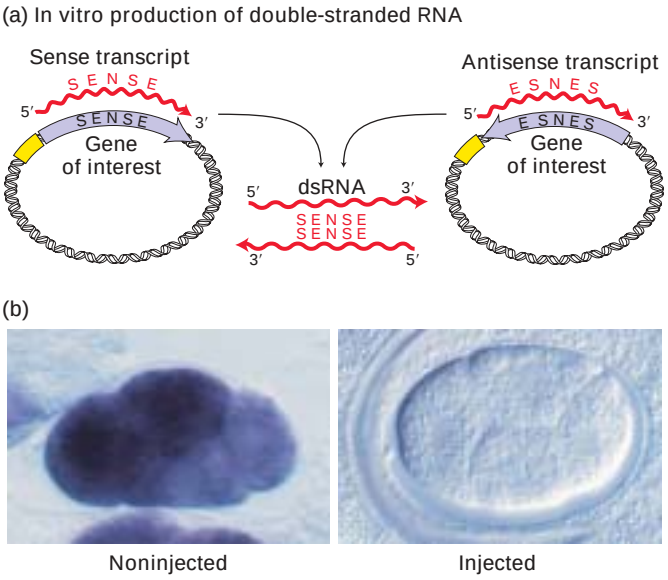
Double-Stranded RNA Molecules Can Interfere with Gene Function by Targeting mRNA for Destruction

Researchers are exploiting a recently discovered phenomenon known as **RNA interference (RNAi)** to inhibit the function of specific genes. This approach is technically simpler than the methods described above for disrupting genes. First observed in the roundworm *C. elegans,* RNAi refers to the ability of a double-stranded (ds) RNA to block expression of its corresponding single-stranded mRNA but not that of mRNAs with a different sequence.

To use RNAi for intentional silencing of a gene of interest, investigators first produce dsRNA based on the sequence of the gene to be inactivated (Figure 9-43a). This dsRNA is injected into the gonad of an adult worm, where it has access to the developing embryos. As the embryos develop, the mRNA molecules corresponding to the injected dsRNA are rapidly destroyed. The resulting worms display a phenotype similar to the one that would result from disruption of the corresponding gene itself. In some cases, entry of just a few molecules of a particular dsRNA into a cell is sufficient to inactivate many copies of the corresponding mRNA. Figure 9-43b illustrates the ability of an injected dsRNA to interfere with production of the corresponding endogenous mRNA in *C. elegans* embryos. In this experiment, the mRNA levels in embryos were determined by incubating the embryos with a fluorescently labeled probe specific for the mRNA of interest. This technique, **in situ hybridization,** is useful in assaying expression of a particular mRNA in cells and tissue sections.

Initially, the phenomenon of RNAi was quite mysterious to geneticists. Recent studies have shown that specialized RNA-processing enzymes cleave dsRNA into short segments, which base-pair with endogenous mRNA. The resulting hybrid molecules are recognized and cleaved by specific nucleases at these hybridization sites. This model accounts for the specificity of RNAi, since it depends on base pairing, and for its potency in silencing gene function, since the complementary mRNA is permanently destroyed by nucleolytic degradation. Although the normal cellular function of RNAi is not understood, it may provide a defense against viruses with dsRNA genomes or help regulate certain endogenous genes. (For a more detailed discussion of the mechanism of RNA interference, see Section 12.4.)

Other organisms in which RNAi-mediated gene inactivation has been successful include *Drosophila,* many kinds of plants, zebrafish, spiders, the frog *Xenopus,* and mice. Although most other organisms do not appear to be as sensitive to the effects of RNAi as *C. elegans,* the method does have general use when the dsRNA is injected directly into embryonic tissues.



▲ EXPERIMENTAL FIGURE 9-43 **RNA interference (RNAi) can functionally inactivate genes in *C. elegans* and some other organisms.** (a) Production of double-stranded RNA (dsRNA) for RNAi of a specific target gene. The coding sequence of the gene, derived from either a cDNA clone or a segment of genomic DNA, is placed in two orientations in a plasmid vector adjacent to a strong promoter. Transcription of both constructs in vitro using RNA polymerase and ribonucleotide triphosphates yields many RNA copies in the sense orientation (identical with the mRNA sequence) or complementary antisense orientation. Under suitable conditions, these complementary RNA molecules will hybridize to form dsRNA. (b) Inhibition of *mex3* RNA expression in worm embryos by RNAi (see the text for the mechanism). (*Left*) Expression of *mex3* RNA in embryos was assayed by in situ hybridization with a fluorescently labeled probe (purple) specific for this mRNA. (*Right*) The embryo derived from a worm injected with double-stranded *mex3* mRNA produces little or no endogenous *mex3* mRNA, as indicated by the absence of color. Each four-cell stage embryo is ≈50 μm in length. [Part (b) from A. Fire et al., 1998, *Nature* **391**:806.]

KEY CONCEPTS OF SECTION 9.5

Inactivating the Function of Specific Genes in Eukaryotes

- Once a gene has been cloned, important clues about its normal function in vivo can be deduced from the observed phenotypic effects of mutating the gene.
- Genes can be disrupted in yeast by inserting a selectable marker gene into one allele of a wild-type gene via homologous recombination, producing a heterozygous mutant. When such a heterozygote is sporulated, disruption of an essential gene will produce two nonviable haploid spores (Figure 9-37).
- A yeast gene can be inactivated in a controlled manner by using the *GAL1* promoter to shut off transcription of a gene when cells are transferred to glucose medium.

- In mice, modified genes can be incorporated into the germ line at their original genomic location by homologous recombination, producing knockouts (see Figures 9-38 and 9-39). Mouse knockouts can provide models for human genetic diseases such as cystic fibrosis.
- The *loxP*-Cre recombination system permits production of mice in which a gene is knocked out in a specific tissue.
- In the production of transgenic cells or organisms, exogenous DNA is integrated into the host genome by nonhomologous recombination (see Figure 9-41). Introduction of a dominant-negative allele in this way can functionally inactivate a gene without altering its sequence.
- In some organisms, including the roundworm *C. elegans*, double-stranded RNA triggers destruction of the all the mRNA molecules with the same sequence (see Figure 9-43). This phenomenon, known as *RNAi* (RNA interference), provides a specific and potent means of functionally inactivating genes without altering their structure.

9.6 Identifying and Locating Human Disease Genes

Inherited human diseases are the phenotypic consequence of defective human genes. Table 9-3 lists several of the most commonly occurring inherited diseases. Although a "disease" gene may result from a new mutation that arose in the preceding generation, most cases of inherited diseases are caused by preexisting mutant alleles that have been passed from one generation to the next for many generations.

Nowadays, the typical first step in deciphering the underlying cause for any inherited human disease is to identify the affected gene and its encoded protein. Comparison of the sequences of a disease gene and its product with those of genes and proteins whose sequence and function are known can provide clues to the molecular and cellular cause of the disease. Historically, researchers have used whatever pheno-

TABLE 9-3 Common Inherited Human Diseases

Disease	Molecular and Cellular Defect	Incidence
AUTOSOMAL RECESSIVE		
Sickle-cell anemia	Abnormal hemoglobin causes deformation of red blood cells, which can become lodged in capillaries; also confers resistance to malaria.	1/625 of sub-Saharan African origin
Cystic fibrosis	Defective chloride channel (CFTR) in epithelial cells leads to excessive mucus in lungs.	1/2500 of European origin
Phenylketonuria (PKU)	Defective enzyme in phenylalanine metabolism (tyrosine hydroxylase) results in excess phenylalanine, leading to mental retardation, unless restricted by diet.	1/10,000 of European origin
Tay-Sachs disease	Defective hexosaminidase enzyme leads to accumulation of excess sphingolipids in the lysosomes of neurons, impairing neural development.	1/1000 Eastern European Jews
AUTOSOMAL DOMINANT		
Huntington's disease	Defective neural protein (huntingtin) may assemble into aggregates causing damage to neural tissue.	1/10,000 of European origin
Hypercholesterolemia	Defective LDL receptor leads to excessive cholesterol in blood and early heart attacks.	1/122 French Canadians
X-LINKED RECESSIVE		
Duchenne muscular dystrophy (DMD)	Defective cytoskeletal protein dystrophin leads to impaired muscle function.	1/3500 males
Hemophilia A	Defective blood clotting factor VIII leads to uncontrolled bleeding.	1–2/10,000 males

typic clues might be relevant to make guesses about the molecular basis of inherited diseases. An early example of successful guesswork was the hypothesis that sickle-cell anemia, known to be a disease of blood cells, might be caused by a defective hemoglobin. This idea led to identification of a specific amino acid substitution in hemoglobin that causes polymerization of the defective hemoglobin molecules, causing the sickle-like deformation of red blood cells in individuals who have inherited two copies of the Hb^s allele for sickle-cell hemoglobin.

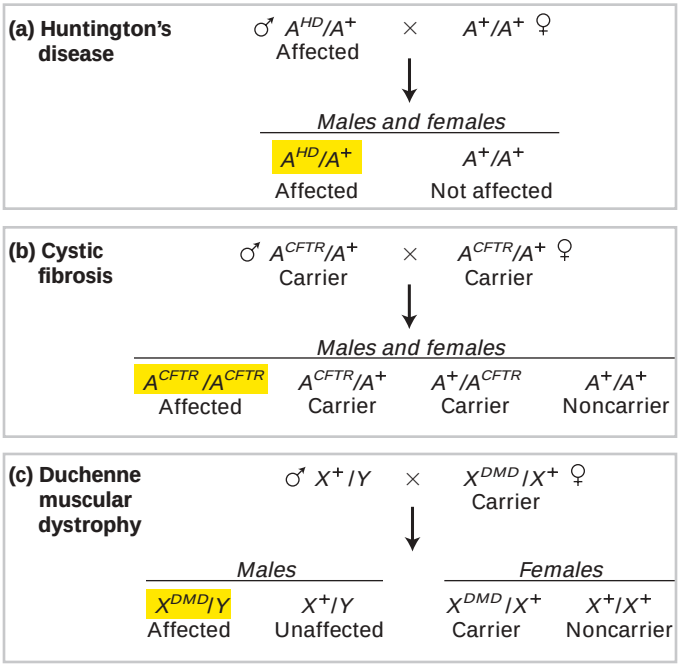
Most often, however, the genes responsible for inherited diseases must be found without any prior knowledge or reasonable hypotheses about the nature of the affected gene or its encoded protein. In this section, we will see how human geneticists can find the gene responsible for an inherited disease by following the segregation of the disease in families. The segregation of the disease can be correlated with the segregation of many other **genetic markers,** eventually leading to identification of the chromosomal position of the affected gene. This information, along with knowledge of the sequence of the human genome, can ultimately allow the affected gene and the disease-causing mutations to be pinpointed.

Many Inherited Diseases Show One of Three Major Patterns of Inheritance

Human genetic diseases that result from mutation in one specific gene exhibit several inheritance patterns depending on the nature and chromosomal location of the alleles that cause them. One characteristic pattern is that exhibited by a dominant allele in an **autosome** (that is, one of the 22 human chromosomes that is not a sex chromosome). Because an *autosomal dominant* allele is expressed in the heterozygote, usually at least one of the parents of an affected individual will also have the disease. It is often the case that the diseases caused by dominant alleles appear later in life after the reproductive age. If this were not the case, natural selection would have eliminated the allele during human evolution. An example of an autosomal dominant disease is Huntington's disease, a neural degenerative disease that generally strikes in mid- to late life. If either parent carries a mutant *HD* allele, each of his or her children (regardless of sex) has a 50 percent chance of inheriting the mutant allele and being affected (Figure 9-44a).

A recessive allele in an autosome exhibits a quite different segregation pattern. For an *autosomal recessive* allele, both parents must be heterozygous **carriers** of the allele in order for their children to be at risk of being affected with the disease. Each child of heterozygous parents has a 25 percent chance of receiving both recessive alleles and thus being affected, a 50 percent chance of receiving one normal and one mutant allele and thus being a carrier, and a 25 percent chance of receiving two normal alleles. A clear example of an autosomal recessive disease is cystic fibrosis, which results from a defective chloride channel gene known as *CFTR* (Figure 9-44b). Related individuals (e.g., first or second cousins) have a relatively high probability of being carriers for the same recessive alleles. Thus children born to related parents are much more likely than those born to unrelated parents to be homozygous for, and therefore affected by, an autosomal recessive disorder.

The third common pattern of inheritance is that of an *X-linked recessive* allele. A recessive allele on the X-chromosome will most often be expressed in males, who receive only one X chromosome from their mother, but not in females who receive an X chromosome from both their mother and father. This leads to a distinctive sex-linked segregation pattern where the disease is exhibited much more frequently in



▲ FIGURE 9-44 **Three common inheritance patterns for human genetic diseases.** Wild-type autosomal (A) and sex chromosomes (X and Y) are indicated by superscript plus signs. (a) In an autosomal dominant disorder such as Huntington's disease, only one mutant allele is needed to confer the disease. If either parent is heterozygous for the mutant *HD* allele, his or her children have a 50 percent chance of inheriting the mutant allele and getting the disease. (b) In an autosomal recessive disorder such as cystic fibrosis, two mutant alleles must be present to confer the disease. Both parents must be heterozygous carriers of the mutant *CFTR* gene for their children to be at risk of being affected or being carriers. (c) An X-linked recessive disease such as Duchenne muscular dystrophy is caused by a recessive mutation on the X chromosome and exhibits the typical sex-linked segregation pattern. Males born to mothers heterozygous for a mutant *DMD* allele have a 50 percent chance of inheriting the mutant allele and being affected. Females born to heterozygous mothers have a 50 percent chance of being carriers.

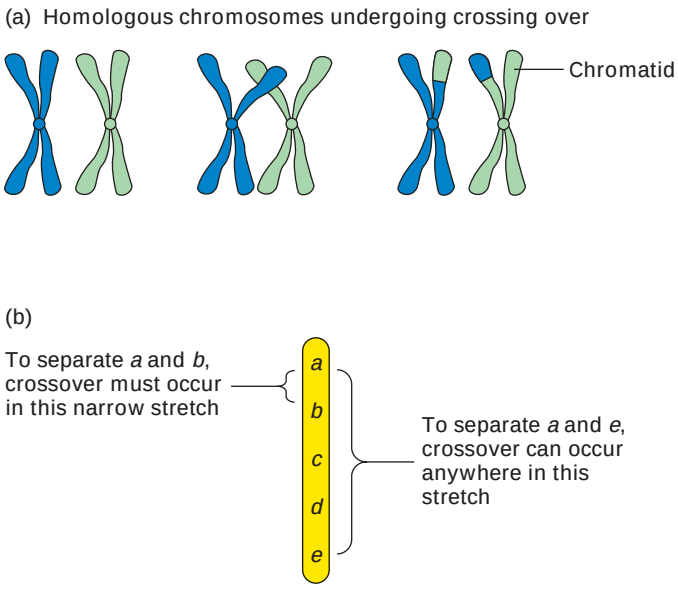
males than in females. For example, Duchenne muscular dystrophy (DMD), a muscle degenerative disease that specifically affects males, is caused by a recessive allele on the X chromosome. DMD exhibits the typical sex-linked segregation pattern in which mothers who are heterozygous and therefore phenotypically normal can act as carriers, transmitting the DMD allele, and therefore the disease, to 50 percent of their male progeny (Figure 9-44c).

Recombinational Analysis Can Position Genes on a Chromosome

The independent segregation of chromosomes during meiosis provides the basis for determining whether genes are on the same or different chromosomes. Genetic traits that segregate together during meiosis more frequently than expected from random segregation are controlled by genes located on the same chromosome. (The tendency of genes on the same chromosome to be inherited together is referred to as genetic **linkage.**) However, the occurrence of recombination during meiosis can separate linked genes; this phenomenon provides a means for locating (mapping) a particular gene relative to other genes on the *same* chromosome.

Recombination takes place before the first meiotic cell division in germ cells when the replicated chromosomes of each homologous pair align with each other, an act called *synapsis* (see Figure 9-3). At this time, homologous DNA sequences on maternally and paternally derived chromatids can exchange with each other, a process known as **crossing over.** The sites of recombination occur more or less at random along the length of chromosomes; thus the closer together two genes are, the less likely that recombination will occur between them during meiosis (Figure 9-45). In other words, *the less frequently recombination occurs between two genes on the same chromosome, the more tightly they are linked and the closer together they are.* The frequency of recombination between two genes can be determined from the proportion of recombinant progeny, whose phenotypes differ from the parental phenotypes, produced in crosses of parents carrying different alleles of the genes.

The presence of many different already mapped genetic traits, or markers, distributed along the length of a chromosome facilitates the mapping of a new mutation by assessing its possible linkage to these marker genes in appropriate crosses. The more markers that are available, the more precisely a mutation can be mapped. As more and more mutations are mapped, the linear order of genes along the length of a chromosome can be constructed. This ordering of genes along a chromosome is called a *genetic map*, or *linkage map*. By convention, one genetic map unit is defined as the distance between two positions along a chromosome that results in one recombinant individual in 100 progeny. The distance corresponding to this 1 percent recombination frequency is called a *centimorgan (cM).* Comparison of the actual physical distances between known genes, determined by molecular analysis, with their recombination frequency indicates that in humans 1 centimorgan on average represents a distance of about 7.5×10^5 base pairs.



▲ FIGURE 9-45 **Recombination during meiosis.** (a) Crossing over can occur between chromatids of homologous chromosomes before the first meiotic division (see Figure 9-3). (b) The longer the distance between two genes on a chromatid, the more likely they are to be separated by recombination.

DNA Polymorphisms Are Used in Linkage-Mapping Human Mutations

Many different genetic markers are needed to construct a high-resolution genetic map. In the experimental organisms commonly used in genetic studies, numerous markers with easily detectable phenotypes are readily available for genetic mapping of mutations. This is not the case for mapping genes whose mutant alleles are associated with inherited diseases in humans. However, recombinant DNA technology has made available a wealth of useful DNA-based *molecular markers.* Because most of the human genome does not code for protein, a large amount of sequence variation exists between individuals. Indeed, it has been estimated that nucleotide differences between unrelated individuals can be detected on an average of every 10^3 nucleotides. If these variations in DNA sequence, referred to as *DNA polymorphisms,* can be followed from one generation to the next, they can serve as genetic markers for linkage studies. Currently, a panel of as many as 10^4 different known polymorphisms whose locations have been mapped in the human genome is used for genetic linkage studies in humans.

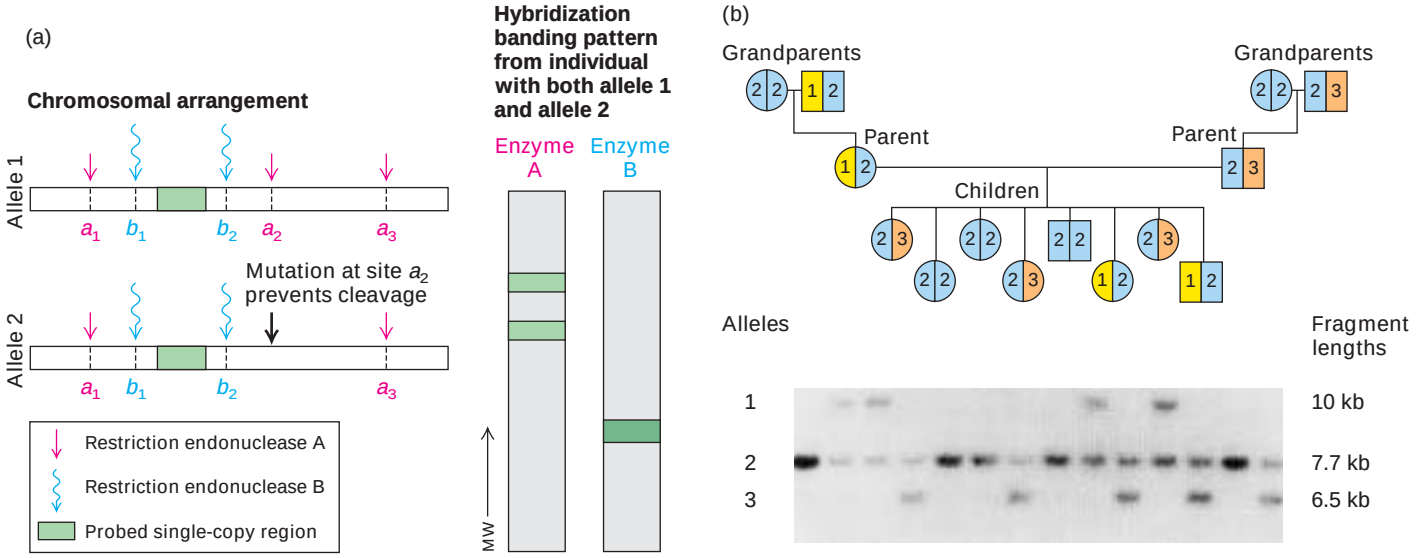
Restriction fragment length polymorphisms (RFLPs) were the first type of molecular markers used in linkage studies. RFLPs arise because mutations can create or destroy the

sites recognized by specific restriction enzymes, leading to variations between individuals in the length of restriction fragments produced from identical regions of the genome. Differences in the sizes of restriction fragments between individuals can be detected by Southern blotting with a probe specific for a region of DNA known to contain an RFLP (Figure 9-46a). The segregation and meiotic recombination of such DNA polymorphisms can be followed like typical genetic markers. Figure 9-46b illustrates how RFLP analysis of a family can detect the segregation of an RFLP that can be used to test for statistically significant linkage to the allele for an inherited disease or some other human trait of interest.

The amassing of vast amounts of genomic sequence information from different humans in recent years has led to identification of other useful DNA polymorphisms. *Single nucleotide polymorphisms* (*SNPs*) constitute the most abundant type and are therefore useful for constructing high-resolution genetic maps. Another useful type of DNA polymorphism consists of a variable number of repetitions of a one- two-, or three-base sequence. Such polymorphisms, known as simple sequence repeats (SSRs), or *microsatellites*, presumably are formed by recombination or a slippage mechanism of either the template or newly synthesized strands during DNA replication. A useful property of SSRs is that different individuals will often have different numbers of repeats. The existence of multiple versions of an SSR makes it more likely to produce an informative segregation pattern in a given pedigree and therefore be of more general use in mapping the positions of disease genes. If an SNP or SSR alters a restriction site, it can be detected by RFLP analysis. More commonly, however, these polymorphisms do not alter restriction fragments and must be detected by PCR amplification and DNA sequencing.

Linkage Studies Can Map Disease Genes with a Resolution of About 1 Centimorgan

Without going into all the technical considerations, let's see how the allele conferring a particular dominant trait (e.g., familial hypercholesterolemia) might be mapped. The first step is to obtain DNA samples from all the members of a family containing individuals that exhibit the disease. The DNA from each affected and unaffected individual then is analyzed to determine the identity of a large number of known DNA polymorphisms (either SSR or SNP markers can be used). The segregation pattern of each DNA polymorphism within the family is then compared with the segregation of the



▲ EXPERIMENTAL FIGURE 9-46 **Restriction fragment length polymorphisms (RFLPs) can be followed like genetic markers.** (a) In the example shown, DNA from an individual is treated with two different restriction enzymes (A and B), which cut DNA at different sequences (*a* and *b*). The resulting fragments are subjected to Southern blot analysis (see Figure 9-26) with a radioactive probe that binds to the indicated DNA region (green) to detect the fragments. Since no differences between the two homologous chromosomes occur in the sequences recognized by the B enzyme, only one fragment is recognized by the probe, as indicated by a single hybridization band. However, treatment with enzyme A produces fragments of two different lengths (two bands are seen), indicating that a mutation has caused the loss of one of the *a* sites in one of the two chromosomes. (b) Pedigree based on RFLP analysis of the DNA from a region known to be present on chromosome 5. The DNA samples were cut with the restriction enzyme *Taq*I and analyzed by Southern blotting. In this family, this region of the genome exists in three allelic forms characterized by *Taq*I sites spaced 10, 7.7, or 6.5 kb apart. Each individual has two alleles; some contain allele 2 (7.7 kb) on both chromosomes, and others are heterozygous at this site. Circles indicate females; squares indicate males. The gel lanes are aligned below the corresponding subjects. [After H. Donis-Keller et al., 1987, *Cell* **51**:319.]

disease under study to find those polymorphisms that tend to segregate along with the disease. Finally, computer analysis of the segregation data is used to calculate the likelihood of linkage between each DNA polymorphism and the disease-causing allele.

In practice, segregation data are collected from different families exhibiting the same disease and pooled. The more families exhibiting a particular disease that can be examined, the greater the statistical significance of evidence for linkage that can be obtained and the greater the precision with which the distance can be measured between a linked DNA polymorphism and a disease allele. Most family studies have a maximum of about 100 individuals in which linkage between a disease gene and a panel of DNA polymorphisms can be tested. This number of individuals sets the practical upper limit on the resolution of such a mapping study to about 1 centimorgan, or a physical distance of about 7.5×10^5 base pairs.

A phenomenon called *linkage disequilibrium* is the basis for an alternative strategy, which in some cases can afford a higher degree of resolution in mapping studies. This approach depends on the particular circumstance in which a genetic disease commonly found in a particular population results from a single mutation that occurred many generations in the past. This ancestral chromosome will carry closely linked DNA polymorphisms that will have been conserved through many generations. Polymorphisms that are farthest away on the chromosome will tend to become separated from the disease gene by recombination, whereas those closest to the disease gene will remain associated with it. By assessing the distribution of specific markers in all the affected individuals in a population, geneticists can identify DNA markers tightly associated with the disease, thus localizing the disease-associated gene to a relatively small region. The resolving power of this method comes from the ability to determine whether a polymorphism and the disease allele were ever separated by a meiotic recombination event at any time since the disease allele first appeared on the ancestral chromosome. Under ideal circumstances linkage disequilibrium studies can improve the resolution of mapping studies to less than 0.1 centimorgan.

Further Analysis Is Needed to Locate a Disease Gene in Cloned DNA

Although linkage mapping can usually locate a human disease gene to a region containing about 7.5×10^5 base pairs, as many as 50 different genes may be located in a region of this size. The ultimate objective of a mapping study is to locate the gene within a cloned segment of DNA and then to determine the nucleotide sequence of this fragment.

One strategy for further localizing a disease gene within the genome is to identify mRNA encoded by DNA in the region of the gene under study. Comparison of gene expression in tissues from normal and affected individuals may suggest tissues in which a particular disease gene normally is expressed. For instance, a mutation that phenotypically affects muscle, but no other tissue, might be in a gene that is expressed only in muscle tissue. The expression of mRNA in both normal and affected individuals generally is determined by Northern blotting or in situ hybridization of labeled DNA or RNA to tissue sections. Northern blots permit comparison of both the level of expression and the size of mRNAs in mutant and wild-type tissues (see Figure 9-27). Although the sensitivity of in situ hybridization is lower than that of Northern blot analysis, it can be very helpful in identifying an mRNA that is expressed at low levels in a given tissue but at very high levels in a subclass of cells within that tissue. An mRNA that is altered or missing in various individuals affected with a disease compared with wild-type individuals would be an excellent candidate for encoding the protein whose disrupted function causes that disease.

In many cases, point mutations that give rise to disease-causing alleles may result in no detectable change in the level of expression or electrophoretic mobility of mRNAs. Thus if comparison of the mRNAs expressed in normal and affected individuals reveals no detectable differences in the candidate mRNAs, a search for point mutations in the DNA regions encoding the mRNAs is undertaken. Now that highly efficient methods for sequencing DNA are available, researchers frequently determine the sequence of candidate regions of DNA isolated from affected individuals to identify point mutations. The overall strategy is to search for a coding sequence that consistently shows possibly deleterious alterations in DNA from individuals that exhibit the disease. A limitation of this approach is that the region near the affected gene may carry naturally occurring polymorphisms unrelated to the gene of interest. Such polymorphisms, not functionally related to the disease, can lead to misidentification of the DNA fragment carrying the gene of interest. For this reason, the more mutant alleles available for analysis, the more likely that a gene will be correctly identified.

Many Inherited Diseases Result from Multiple Genetic Defects

Most of the inherited human diseases that are now understood at the molecular level are *monogenetic traits.* That is, a clearly discernible disease state is produced by the presence of a defect in a single gene. Monogenic diseases caused by mutation in one specific gene exhibit one of the characteristic inheritance patterns shown in Figure 9-44. The genes associated with most of the common monogenic diseases have already been mapped using DNA-based markers as described previously.

However, many other inherited diseases show more complicated patterns of inheritance, making the identification of the underlying genetic cause much more difficult. One type of added complexity that is frequently encountered is *genetic heterogeneity.* In such cases, mutations in

any one of multiple different genes can cause the same disease. For example, retinitis pigmentosa, which is characterized by degeneration of the retina usually leading to blindness, can be caused by mutations in any one of more than 60 different genes. In human linkage studies, data from multiple families usually must be combined to determine whether a statistically significant linkage exists between a disease gene and known molecular markers. Genetic heterogeneity such as that exhibited by retinitis pigmentosa can confound such an approach because any statistical trend in the mapping data from one family tends to be canceled out by the data obtained from another family with an unrelated causative gene.

Human geneticists used two different approaches to identify the many genes associated with retinitis pigmentosa. The first approach relied on mapping studies in exceptionally large single families that contained a sufficient number of affected individuals to provide statistically significant evidence for linkage between known DNA polymorphisms and a single causative gene. The genes identified in such studies showed that several of the mutations that cause retinitis pigmentosa lie within genes that encode abundant proteins of the retina. Following up on this clue, geneticists concentrated their attention on those genes that are highly expressed in the retina when screening other individuals with retinitis pigmentosa. This approach of using additional information to direct screening efforts to a subset of candidate genes led to identification of additional rare causative mutations in many different genes encoding retinal proteins.

A further complication in the genetic dissection of human diseases is posed by diabetes, heart disease, obesity, predisposition to cancer, and a variety of mental disorders that have at least some heritable properties. These and many other diseases can be considered to be *polygenic traits* in the sense that alleles of multiple genes, acting together within an individual, contribute to both the occurrence and the severity of disease. A systematic solution to the problem of mapping complex polygenic traits in humans does not yet exist. Future progress may come from development of refined diagnostic methods that can distinguish the different forms of diseases resulting from multiple causes.

Models of human disease in experimental organisms may also contribute to unraveling the genetics of complex traits such as obesity or diabetes. For instance, large-scale controlled breeding experiments in mice can identify mouse genes associated with diseases analogous to those in humans. The human orthologs of the mouse genes identified in such studies would be likely candidates for involvement in the corresponding human disease. DNA from human populations then could be examined to determine if particular alleles of the candidate genes show a tendency to be present in individuals affected with the disease but absent from unaffected individuals. This "candidate gene" approach is currently being used intensively to search for genes that may contribute to the major polygenic diseases in humans.

KEY CONCEPTS OF SECTION 9.6

Identifying and Locating Human Disease Genes

- Inherited diseases and other traits in humans show three major patterns of inheritance: autosomal dominant, autosomal recessive, and X-linked recessive (see Figure 9-44).
- Genes located on the same chromosome can be separated by crossing over during meiosis, thus producing new recombinant genotypes in the next generation (see Figure 9-45).
- Genes for human diseases and other traits can be mapped by determining their cosegregation with markers whose locations in the genome are known. The closer a gene is to a particular marker, the more likely they are to cosegregate.
- Mapping of human genes with great precision requires thousands of molecular markers distributed along the chromosomes. The most useful markers are differences in the DNA sequence (polymorphisms) among individuals in noncoding regions of the genome.
- DNA polymorphisms useful in mapping human genes include restriction fragment length polymorphisms (RFLPs), single-nucleotide polymorphisms (SNPs), and simple sequence repeats (SSRs).
- Linkage mapping often can locate a human disease gene to a chromosomal region that includes as many as 50 genes. To identify the gene of interest within this candidate region typically requires expression analysis and comparison of DNA sequences between wild-type and disease-affected individuals.
- Some inherited diseases can result from mutations in different genes in different individuals (genetic heterogeneity). The occurrence and severity of other diseases depend on the presence of mutant alleles of multiple genes in the same individuals (polygenic traits). Mapping of the genes associated with such diseases is particularly difficult because the occurrence of the disease cannot readily be correlated to a single chromosomal locus.

PERSPECTIVES FOR THE FUTURE

As the examples in this chapter and throughout the book illustrate, genetic analysis is the foundation of our understanding of many fundamental processes in cell biology. By examining the phenotypic consequences of mutations that inactivate a particular gene, geneticists are able to connect knowledge about the sequence, structure, and biochemical activity of the encoded protein to its function in the context of a living cell or multicellular organism. The classical approach to making these connections in both humans and simpler, experimentally accessible organisms has been to identify new mutations of interest based on their phenotypes and then to isolate the affected gene and its protein product.

Although scientists continue to use this classical genetic approach to dissect fundamental cellular processes and biochemical pathways, the availability of complete genomic sequence information for most of the common experimental organisms has fundamentally changed the way genetic experiments are conducted. Using various computational methods, scientists have identified most of the protein-coding gene sequences in *E. coli*, yeast, *Drosophila, Arabidopsis*, mouse, and humans. The gene sequences, in turn, reveal the primary amino acid sequence of the encoded protein products, providing us with a nearly complete list of the proteins found in each of the major experimental organisms.

The approach taken by most researchers has thus shifted from discovering new genes and proteins to discovering the functions of genes and proteins whose sequences are already known. Once an interesting gene has been identified, genomic sequence information greatly speeds subsequent genetic manipulations of the gene, including its designed inactivation, to learn more about its function. Already all the ≈6000 possible gene knockouts in yeast have been produced; this relatively small but complete collection of mutants has become the preferred starting point for many genetic screens in yeast. Similarly, sets of vectors for RNAi inactivation of a large number of defined genes in the nematode *C. elegans* now allow efficient genetic screens to be performed in this multicellular organism. Following the trajectory of recent advances, it seems quite likely that in the foreseeable future either RNAi or knockout methods will have been used to inactivate every gene in the principal model organisms, including the mouse.

In the past, a scientist might spend many years studying only a single gene, but nowadays scientists commonly study whole sets of genes at once. For example, with DNA microarrays the level of expression of all genes in an organism can be measured almost as easily as the expression of a single gene. One of the great challenges facing geneticists in the twenty-first century will be to exploit the vast amount of available data on the function and regulation of individual genes to gain fundamental insights into the organization of complex biochemical pathways and regulatory networks.

KEY TERMS

alleles *352*
clone *364*
complementary DNAs (cDNAs) *365*
complementation *357*
DNA cloning *361*
DNA library *352*
DNA microarray *385*
dominant *353*
gene knockout *389*
genomics *352*
genotype *352*
heterozygous *353*
homozygous *353*
hybridization *367*
linkage *396*
mutation *352*
Northern blotting *377*
phenotype *352*
plasmids *363*
polymerase chain reaction (PCR) *375*
probes *367*
recessive *353*
recombinant DNA *361*
recombination *387*
restriction enzymes *361*
RNA interference (RNAi) *393*
segregation *355*
Southern blotting *377*
temperature-sensitive mutations *356*
transfection *378*
transformation *363*
transgenes *392*
vectors *361*

REVIEW THE CONCEPTS

1. Genetic mutations can provide insights into the mechanisms of complex cellular or developmental processes. What is the difference between recessive and dominant mutations? What is a temperature-sensitive mutation, and how is this type of mutation useful?

2. A number of experimental approaches can be used to analyze mutations. Describe how complementation analysis can be used to reveal whether two mutations are in the same or in different genes. What are suppressor mutations and synthetic lethal mutations?

3. Restriction enzymes and DNA ligase play essential roles in DNA cloning. How is it that a bacterium that produces a restriction enzyme does not cut its own DNA? Describe some general features of restriction enzyme sites. What are the three types of DNA ends that can be generated after cutting DNA with restriction enzymes? What reaction is catalyzed by DNA ligase?

4. Bacterial plasmids and λ phage serve as cloning vectors. Describe the essential features of a plasmid and a λ phage vector. What are the advantages and applications of plasmids and λ phage as cloning vectors?

5. A DNA library is a collection of clones, each containing a different fragment of DNA, inserted into a cloning vector. What is the difference between a cDNA and a genomic DNA library? How can you use hybridization or expression to screen a library for a specific gene? What oligonucleotide primers could be synthesized as probes to screen a library for the gene encoding the peptide Met-Pro-Glu-Phe-Tyr?

6. In 1993, Kerry Mullis won the Nobel Prize in Chemistry for his invention of the PCR process. Describe the three steps in each cycle of a PCR reaction. Why was the discovery of a thermostable DNA polymerase (e.g., Taq polymerase) so important for the development of PCR?

7. Southern and Northern blotting are powerful tools in molecular biology; describe the technique of each. What are the applications of these two blotting techniques?

8. A number of foreign proteins have been expressed in bacterial and mammalian cells. Describe the essential fea-

tures of a recombinant plasmid that are required for expression of a foreign gene. How can you modify the foreign protein to facilitate its purification? What is the advantage of expressing a protein in mammalian cells versus bacteria?

9. Why is the screening for genes based on the presence of ORFs (open reading frames) more useful for bacterial genomes than for eukaryotic genomes? What are paralogous and orthologous genes? What are some of the explanations for the finding that humans are a much more complex organism than the roundworm *C. elegans*, yet have only less than twice the number of genes (35,000 versus 19,000)?

10. A global analysis of gene expression can be accomplished by using a DNA microarray. What is a DNA microarray? How are DNA microarrays used for studying gene expression? How do experiments with microarrays differ from Northern botting experiments described in question 7?

11. The ability to selectively modify the genome in the mouse has revolutionized mouse genetics. Outline the procedure for generating a knockout mouse at a specific genetic locus. How can the *loxP*-Cre system be used to conditionally knock out a gene? What is an important medical application of knockout mice?

12. Two methods for functionally inactivating a gene without altering the gene sequence are by dominant negative mutations and RNA interference (RNAi). Describe how each method can inhibit expression of a gene.

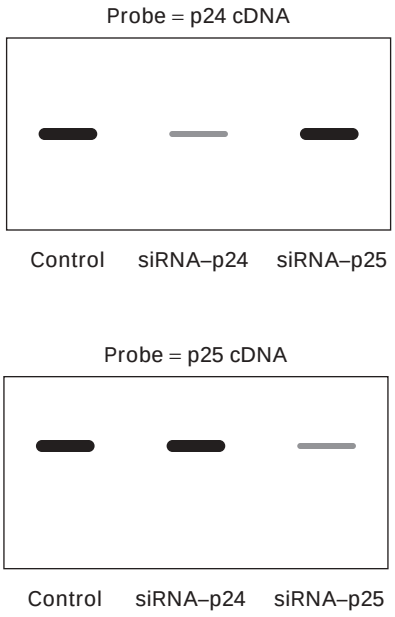
13. DNA polymorphisms can be used as DNA markers. Describe the differences among RFLP, SNP, and SSR polymorphisms. How can these markers be used for DNA mapping studies?

14. Genetic linkage studies can roughly locate the chromosomal position of a "disease" gene. Describe how expression analysis and DNA sequence analysis can be used to identify a "disease" gene.

ANALYZE THE DATA

RNA interference (RNAi) is a process of post-transcriptional gene silencing mediated by short double-stranded RNA molecules called siRNA (small interfering RNAs). In mammalian cells, transfection of 21–22 nucleotide siRNAs leads to degradation of mRNA molecules that contain the same sequence as the siRNA. In the following experiment, siRNA and knockout mice are used to investigate two related cell surface proteins designated p24 and p25 that are suspected to be cellular receptors for the uptake of a newly isolated virus.

a. To test the efficacy of RNAi in cells, siRNAs specific to cell surface proteins p24 (siRNA–p24) and p25 (siRNA–p25) are transfected individually into cultured mouse cells. RNA is extracted from these transfected cells and the mRNA for proteins p24 and p25 are detected on Northern blots using labeled p24 cDNA or p25 cDNA as probes. The control for this experiment is a mock transfection with no siRNA. What do you conclude from this Northern blot about the specificity of the siRNAs for their target mRNAs?



b. Next, the ability of siRNAs to inhibit viral replication is investigated. Cells are transfected with siRNA–p24 or siRNA–p25 or with siRNA to an essential viral protein. Twenty hours later, transfected cells are infected with the virus. After a further incubation period, the cells are collected and lysed. The number of viruses produced by each culture is shown below. The control is a mock transfection with no siRNA. What do you conclude about the role of p24 and p25 in the uptake of the virus? Why might the siRNA to the viral protein be more effective than siRNA to the receptors in reducing the number of viruses?

Cell Treatment	Number of Viruses/ml
Control	1×10^7
siRNA–p24	3×10^6
siRNA–p25	2×10^6
siRNA–p24 and siRNA–p25	1×10^4
siRNA to viral protein	1×10^2

c. To investigate the role of proteins p24 and p25 for viral replication in live mice, transgenic mice that lack genes for p24 or p25 are generated. The *loxP*-Cre conditional knockout system is used to selectively delete the genes in cells of either the liver or the lung. Wild type and knockout mice are infected with virus. After a 24-hour incubation period, mice are killed and lung and liver tissues are removed and examined for the presence (infected) or absence (normal) of virus by immunohistochemistry. What do these data indicate about the cellular requirements for viral infection in different tissues?

	Tissue Examined	
Mouse	**Liver**	**Lung**
Wild type	infected	infected
Knockout of p24 in liver	normal	infected
Knockout of p24 in lung	infected	infected
Knockout of p25 in liver	infected	infected
Knockout of p25 in lung	infected	normal

d. By performing Northern blots on different tissues from wild-type mice, you find that p24 is expressed in the liver but not in the lung, whereas p25 is expressed in the lung but not the liver. Based on all the data you have collected, propose a model to explain which protein(s) are involved in the virus entry into liver and lung cells? Would you predict that the cultered mouse cells used in parts (a) and (b) express p24, p25, or both proteins?

REFERENCES

Genetic Analysis of Mutations to Identify and Study Genes

Adams, A. E. M., D. Botstein, and D. B. Drubin. 1989. A yeast actin-binding protein is encoded by *sac6*, a gene found by suppression of an actin mutation. *Science* **243**:231.

Griffiths, A. G. F., et al. 2000. *An Introduction to Genetic Analysis*, 7th ed. W. H. Freeman and Company.

Guarente, L. 1993. Synthetic enhancement in gene interaction: a genetic tool comes of age. *Trends Genet*. **9**:362–366.

Hartwell, L. H. 1967. Macromolecular synthesis of temperature-sensitive mutants of yeast. *J. Bacteriol.* **93**:1662.

Hartwell, L. H. 1974. Genetic control of the cell division cycle in yeast. *Science* **183**:46.

Nüsslein-Volhard, C., and E. Wieschaus. 1980. Mutations affecting segment number and polarity in *Drosophila. Nature* **287**:795–801.

Simon, M. A., et al. 1991. Ras1 and a putative guanine nucleotide exchange factor perform crucial steps in signaling by the sevenless protein tyrosine kinase. *Cell* **67**:701–716.

Tong, A. H., et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**:2364–2368.

DNA Cloning by Recombinant DNA Methods

Ausubel, F. M., et al. 2002. *Current Protocols in Molecular Biology.* Wiley.

Gubler, U., and B. J. Hoffman. 1983. A simple and very efficient method for generating cDNA libraries. *Gene* **25**:263–289.

Han, J. H., C. Stratowa, and W. J. Rutter. 1987. Isolation of full-length putative rat lysophospholipase cDNA using improved methods for mRNA isolation and cDNA cloning. *Biochem.* **26**:1617–1632.

Itakura, K., J. J. Rossi, and R. B. Wallace. 1984. Synthesis and use of synthetic oligonucleotides. *Ann. Rev. Biochem*. **53**:323–356.

Maniatis, T., et al. 1978. The isolation of structural genes from libraries of eucaryotic DNA. *Cell* **15**:687–701.

Nasmyth, K. A., and S. I. Reed. 1980. Isolation of genes by complementation in yeast: molecular cloning of a cell-cycle gene. *Proc. Nat'l. Acad. Sci. USA* **77**:2119–2123.

Nathans, D., and H. O. Smith. 1975. Restriction endonucleases in the analysis and restructuring of DNA molecules. *Ann. Rev. Biochem*. **44**:273–293.

Roberts, R. J., and D. Macelis. 1997. REBASE—restriction enzymes and methylases. *Nucl. Acids Res.* **25**:248–262. Information on accessing a continuously updated database on restriction and modification enzymes at http://www.neb.com/rebase.

Thomas, M., J. R. Cameron, and R. W. Davis. 1974. Viable molecular hybrids of bacteriophage lambda and eukaryotic DNA. *Proc. Nat'l. Acad. Sci. USA* **71**:4579–4583.

Sambrook, J., and D. Russell. 2001. *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory.

Characterizing and Using Cloned DNA Fragments

Andrews, A. T. 1986. *Electrophoresis,* 2d ed. Oxford University Press.

Erlich, H., ed. 1992. *PCR Technology: Principles and Applications for DNA Amplification.* W. H. Freeman and Company.

Pellicer, A., M. Wigler, R. Axel, and S. Silverstein. 1978. The transfer and stable integration of the HSV thymidine kinase gene into mouse cells. *Cell* **41**:133–141.

Saiki, R. K., et al. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**:487–491.

Sanger, F. 1981. Determination of nucleotide sequences in DNA. *Science* **214**:1205–1210.

Souza, L. M., et al. 1986. Recombinant human granulocyte-colony stimulating factor: effects on normal and leukemic myeloid cells. *Science* **232**:61–65.

Wahl, G. M., J. L. Meinkoth, and A. R. Kimmel. 1987. Northern and Southern blots. *Meth. Enzymol.* **152**:572–581.

Wallace, R. B., et al. 1981. The use of synthetic oligonucleotides as hybridization probes. II: Hybridization of oligonucleotides of mixed sequence to rabbit β-globin DNA. *Nucl. Acids Res.* **9**:879–887.

Genomics: Genome-wide Analysis of Gene Structure and Expression

BLAST Information can be found at: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.htm

Ballester, R., et al. 1990. The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell* **63**:851–859.

Chervitz, S. A., et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282**:2022–2028.

Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Gen*. **25**:25–29.

Lander, E. S., et al. 2001 Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

Rubin, G. M., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287**:2204–2215.

Waterston, R. H., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520–562.

Inactivating the Function of Specific Genes in Eukaryotes

Capecchi, M. R. 1989. Altering the genome by homologous recombination. *Science* **244**:1288–1292.

Deshaies, R. J., et al. 1988. A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides. *Nature* **332**:800–805.

Fire, A., et al. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature***391**:806–811.

Gu, H., et al. 1994. Deletion of a DNA polymerase beta gene segment in T cells using cell type-specific gene targeting. *Science* **265**:103–106.

Zamore, P. D., T. Tuschl, P. A. Sharp, and D. P. Bartel. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**:25–33.

Zimmer, A. 1992. Manipulating the genome by homologous recombination in embryonic stem cells. *Ann. Rev. Neurosci.* **15**:115.

Identifying and Locating Human Disease Genes

Botstein, D., et al. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Genet.* **32**:314–331.

Donis-Keller, H., et al. 1987. A genetic linkage map of the human genome. *Cell* **51**:319–337.

Hartwell, et al. 2000. *Genetics: From Genes to Genomes*. McGraw-Hill.

Hastbacka, T., et al. 1994. The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78**:1073.

Orita, M., et al. 1989. Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* **5**:874.

Tabor, H. K., N. J. Risch, and R. M. Myers. 2002. Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet*. **3**:391–397.