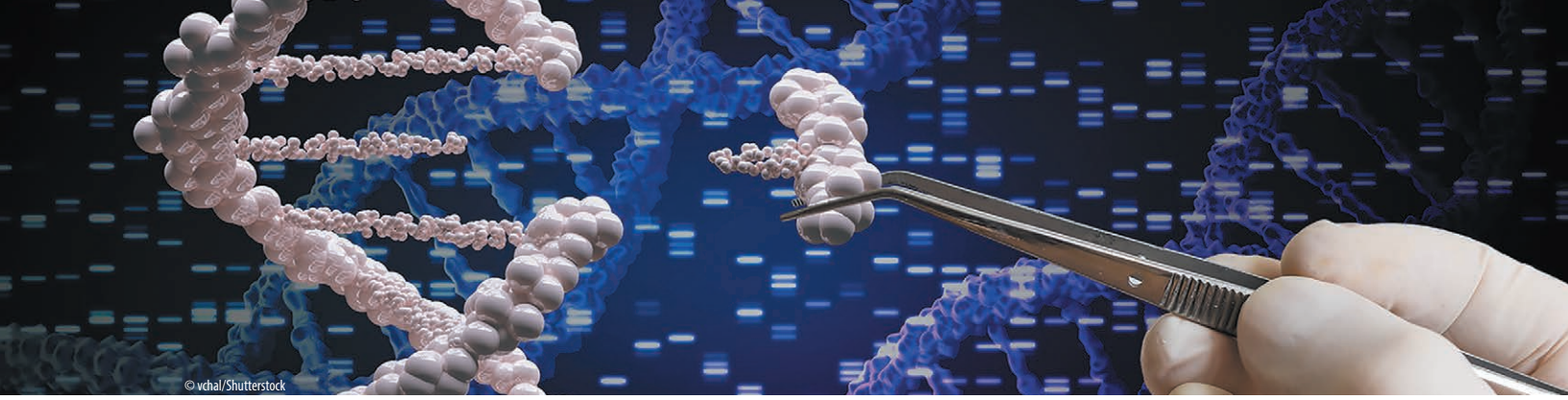


CHAPTER 2

Methods in Molecular Biology and Genetic Engineering

CHAPTER OUTLINE

2.1 Introduction

Today, the field of molecular biology focuses on the mechanisms by which cellular processes are carried out by the various biological macromolecules in the cell, with a particular emphasis on the structure and function of genes and genomes. Molecular biology as a field, however, was originally born from the development of tools and methods that allow the direct manipulation of deoxyribonucleic acid (DNA) both *in vitro* and *in vivo* in numerous organisms.

Two of the first and most basic items in the molecular biologist's toolkit are **restriction endonucleases** (which allow DNA to be cut into precise

restriction endonuclease
An enzyme that recognizes specific short sequences of DNA and cleaves the duplex (sometimes at the target site and sometimes elsewhere, depending on type).

cloning vectors DNA (often derived from a plasmid or a bacteriophage genome) that can be used to propagate an incorporated DNA sequence in a host cell; vectors contain selectable markers and replication origins to allow identification and maintenance in the host.

pieces) and **cloning vectors** (such as plasmids or phages) used to "carry" inserted foreign DNA fragments for the purposes of producing more DNA, ribonucleic acid (RNA), or protein products, or for gene therapy and genome editing techniques. The term *genetic engineering* was originally used to describe the range of manipulations of DNA that become possible with the ability to clone a gene by placing its DNA into another context in which it could be propagated. From this beginning, when recombinant DNA was used as a tool to analyze gene structure and expression, we moved to the ability to change the DNA content of bacteria and eukaryotic cells by directly introducing cloned DNA that could become part of the genome. Then, by changing the genetic content in conjunction with the ability to develop an animal from an embryonic cell, it became possible to generate multicellular eukaryotes with deletions or additions of specific genes that are inherited via the germline. We now use genetic engineering to describe a range of activities, including the manipulation of DNA, the introduction of changes into specific somatic cells within an animal or plant, and even changes in the germline itself.

As research has advanced, more and more sensitive methods for detecting and amplifying DNA have been developed. Now that we have entered the era of routine whole-genome sequencing, methods to assess the content, function, and expression of entire genomes have become commonplace. This chapter will discuss some of the most common methods used in molecular biology, ranging from the very first tools developed by molecular biologists to some of the most recently developed methods now in use. We can only imagine the new technologies that lie over the horizon that will allow the students of today to ask questions not even dreamed of yet.

2.2 Nucleases

nuclease An enzyme that can break a phosphodiester bond.

phosphatase An enzyme that breaks a phosphomonoester bond, cleaving a terminal phosphate.

ligate To covalently link two ends of nucleic acid chains; they may be two ends of one chain or two ends of different chains, either DNA or RNA.

endonuclease An enzyme that cleaves bonds within a nucleic acid chain; it may be specific for RNA or for single-stranded or double-stranded DNA.

exonuclease An enzyme that cleaves nucleotides one at a time from the end of a polynucleotide chain; it may be specific for either the 5′ or 3′ end of DNA or RNA.

Nucleases are one of the most fundamental tools in a molecular biology laboratory. One class of enzymes, the restriction endonucleases that we discuss below, was critical for the cloning revolution. **Nucleases** are enzymes that degrade nucleic acids, the opposite function of polymerases. They hydrolyze, or break, an ester bond in a phosphodiester linkage between adjacent nucleotides in a polynucleotide chain, as shown in **FIGURE 2.1**.

There is another, related class of enzymes that can hydrolyze an ester bond in a nucleotide chain (a monoesterase, usually called a **phosphatase**). The critical difference between a phosphatase and a nuclease is shown in Figure 2.1. A phosphatase can hydrolyze only a terminal ester bond linking a phosphate (or di- or triphosphate) to a terminal nucleotide at the 3′ or 5′ end, whereas a nuclease can hydrolyze an internal ester bond in a diester link between adjacent bases.

Phosphatases are important enzymes in the laboratory because they allow the removal of a terminal phosphate from a polynucleotide chain. This is often required to subsequently **ligate** chains together. This also allows replacement of the phosphate with a labelled phosphate (e.g., ^{32}P).

Nucleases are divided into different groups based on a number of different features. First, **endonucleases** can hydrolyze internal bonds within a polynucleotide chain, while **exonucleases** must start at the end of a chain and hydrolyze from that end position, as shown in Figure 2.1.

The specificity of nucleases ranges from none to extreme. Nucleases may be specific for DNA, as DNases, or RNA, as RNases, or even be specific for a DNA/RNA hybrid, as RNase H (which cleaves the RNA strand of a hybrid

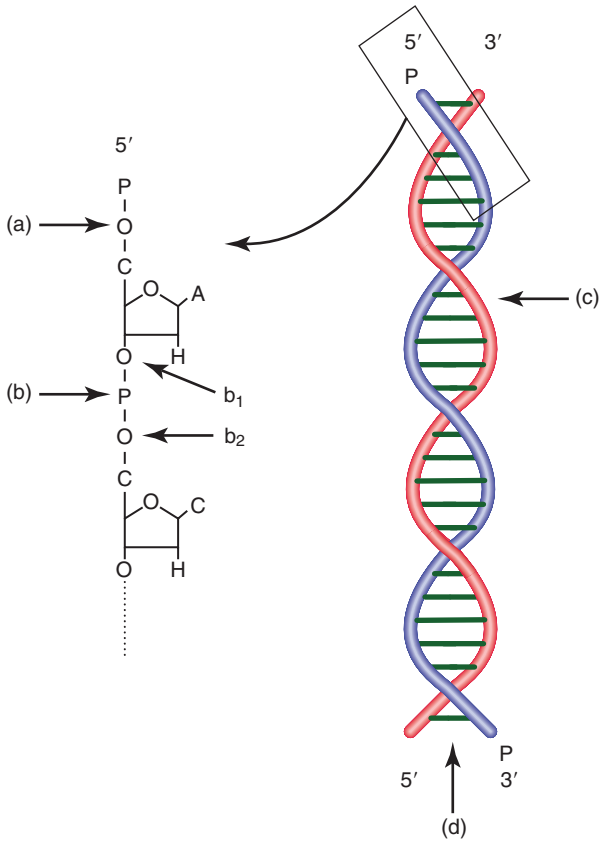


FIGURE 2.1 The target of a phosphatase is shown in (a), a terminal phosphomonoester bond. The target of a nuclease is shown in (b), the phosphodiester bond between two adjacent nucleotides. Note that the nuclease can cleave either the first ester bond from the 3′ end of the terminal nucleotide (b_1) or the second ester bond from the 5′ end of the next nucleotide (b_2). Nucleases can cleave internal bonds (c) as an endonuclease or start from an end and progress into the fragment (d) as an exonuclease.

RNA–DNA duplex). Nucleases may be specific for single-strand nucleotide chains, duplex chains, or both.

When a nuclease, either endo- or exo-, hydrolyzes an ester bond in a phosphodiester linkage, it will have specificity for either of the two ester bonds, generating either 5′ nucleotides or 3′ nucleotides, as seen in Figure 2.1. An exonuclease may either attack a polynucleotide chain from the 5′ end and hydrolyze 5′ to 3′ or attack from the 3′ end and hydrolyze 3′ to 5′.

Nucleases may have a sequence preference, such as pancreatic RNase A, which preferentially cuts after a pyrimidine, or T1 RNase, which cuts single-stranded RNA chains after a G. At the extreme end of sequence specificity lie the *restriction endonucleases*, usually called *restriction enzymes*. These are endonucleases from eubacteria and archaea that recognize a specific DNA sequence. Their name typically derives from the species in which they were discovered. For example, EcoR1 is the first restriction enzyme from an *Escherichia coli R* strain.

Broadly speaking, there are three different classes of restriction enzymes and several subclasses. In 1978, the Nobel Prize in Medicine was awarded to Daniel Nathans, Werner Arber, and Hamilton Smith for the discovery of restriction endonucleases and their application to problems in molecular genetics. It was this discovery that enabled scientists to develop the methods to clone DNA, as we will see in the next section. Thousands of restriction enzymes are known, many of which are now commercially available.

Restriction enzymes have to do two things: (1) recognize a specific sequence and (2) cut, or restrict, that sequence.

The type II restriction enzymes (with several subgroups) are the most common. Type II enzymes are distinguished because the recognition site and cleavage site are the same. These sites range in length from 4 to 8 bp. The sites are typically *inversely palindromic*, that is, reading the same forward and backward on complementary strands, as shown in **FIGURE 2.2**. Restriction enzymes can cut the DNA in two different ways. The first, and more common, is a staggered cut, which leaves single-stranded overhangs, or "sticky ends." The overhang may be a 3′ or a 5′ overhang. The second way is a blunt double-stranded cut, which does not leave an overhang. An additional level of specificity determines whether or not the enzyme will cut DNA containing a methylated base. The degree of specificity in the site also varies. Most enzymes are very specific, whereas some will allow multiple bases at one or two positions within the site.

Restriction enzymes from different bacteria may have the same recognition site but cut the DNA differently. One may make a blunt cut and the other may make a staggered cut, or one may leave a 3′ overhang but the second may leave a 5′ overhang. These different enzymes are called *isoschizomers*.

Types I and III enzymes differ from type II enzymes in that the recognition site and cleavage site are different and are usually not palindromes. With a type I enzyme, the cleavage site can be up to 1,000 bp away from the recognition site. Type III enzymes have closer cleavage sites, usually 20–30 bp away.

A *restriction map* represents a linear sequence of the sites at which particular restriction enzymes find their targets. When a DNA molecule is cut

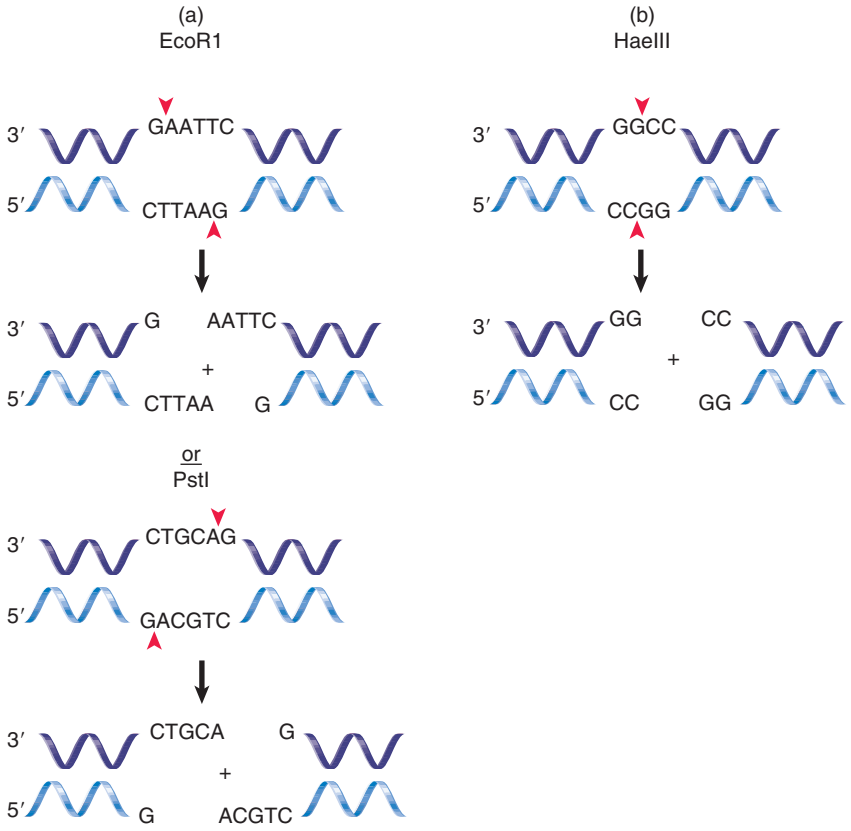


FIGURE 2.2 (a) A restriction endonuclease may cleave its recognition site and make a staggered cut, leaving a 5′ overhang or a 3′ overhang. (b) A restriction endonuclease may cleave its recognition site and make a blunt end cut.

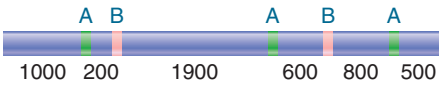


FIGURE 2.3 A restriction map is a linear sequence of sites separated by defined distances on DNA. The map identifies the three sites cleaved by enzyme A and the two sites cleaved by enzyme B. Thus, A produces four fragments, which overlap those of B, and B produces three fragments, which overlap those of A; digestion with both A and B will result in six fragments.

with a suitable restriction enzyme, it is cleaved into distinct negatively charged fragments. These fragments can be separated on the basis of their size by gel electrophoresis (described later in this chapter; see Figure 2.14). By analyzing the restriction fragments of DNA, we can generate a map of the original molecule in the form shown in **FIGURE 2.3**. The map shows the positions at which particular restriction enzymes cut DNA. *So the DNA is divided into a series of regions of defined lengths that lie between sites recognized by the restriction enzymes.* A restriction map can be obtained for any sequence of DNA, irrespective of whether we have any knowledge of its function. If the sequence of the DNA is known, a restriction map can be generated *in silico* by simply searching for the recognition sites of known enzymes. Knowing the restriction map of a DNA sequence of interest is extremely valuable in DNA cloning, which is described in the next section.

KEY CONCEPTS

- Nucleases hydrolyze an ester bond within a phosphodiester bond.
- Phosphatases hydrolyze the ester bond in a phosphomonoester bond.
- Nucleases have a multiplicity of specificities.
- Restriction endonucleases can be used to cleave DNA into defined fragments.
- A map can be generated by different restriction enzymes.

CONCEPT AND REASONING CHECK

What would be the advantage of a naturally produced restriction endonuclease in a bacterial cell?

2.3 Cloning

Cloning has a very simple definition: to *clone* is to make identical copies, whether it is done by a copy machine for a piece of paper, cloning Dolly the sheep, or cloning DNA, which is what we will discuss here. Cloning can also be considered an amplification process, in which we currently have one copy and we want many identical copies. Cloning DNA typically involves **recombinant DNA**. This also has a very simple definition: a DNA molecule from two (or more) different sources.

In order to clone a fragment of DNA, a recombinant DNA molecule must be created and copied many times. Two different DNAs are needed: a ***cloning vector*** and an **insert**, or the molecule to be cloned. The two most popular classes of vectors are derived from plasmids and viruses, respectively.

Over the years, vectors have been specifically engineered for safety, selection ability, and high growth rate. "Safety" means that the vector will not integrate into a genome (unless engineered specifically for that purpose) and the

cloning Propagation of a DNA sequence by incorporating it into a hybrid construct that can be replicated in a host cell.

recombinant DNA An artificial DNA molecule created by joining two (or more) DNA molecules from different sources.

vector A plasmid or phage chromosome that is used to perpetuate a cloned DNA segment.

insert A fragment of DNA that is to be cloned in a vector.

recombinant vector will not autotransfer to another cell. (We discuss selection later.) In the simplest approach, both the vector and insert should be restricted with the same restriction endonuclease to create compatible DNA ends.

Let us now examine the details and the variables that will affect the process. The insert (the DNA fragment to be amplified) could come from one of many different sources, such as restricted genomic DNA; a larger fragment from another clone to be **subcloned** (meaning taking a smaller part of the larger fragment); a polymerase chain reaction (PCR) product; or a DNA fragment synthesized *in vitro*. The insert fragments may be blunt-ended or may have overhanging single strands, depending how the fragments were created (e.g., what restriction enzyme(s) were used to cut the DNA or what PCR primers were used to amplify the DNA).

subclone The process of breaking a cloned fragment into smaller fragments for further cloning.

The vector is selected based on the answers to these questions. For this exercise, we will use a common type of plasmid cloning vector called a *blue/white selection vector*, as shown in **FIGURE 2.4**. This vector has been constructed with a number of important elements. It has an *ori*, or origin of replication (see the *Extrachromosomal Replication* chapter), to allow plasmid replication, which will provide the actual amplification step in a bacterial cell. It contains a gene (*amp*ρ) that encodes resistance to the antibiotic ampicillin, which will allow selection of bacteria that contain the vector. It also contains the *E. coli lacZ* gene (see the chapter titled *The Operon*), which will allow selection of vectors with inserted DNA.

multiple cloning site An artificial DNA sequence in a cloning vector containing multiple restriction endonuclease sites for cloning.

The *lacZ* gene has been engineered to contain a **multiple cloning site**, or MCS. This is an oligonucleotide sequence with a series of different restriction endonuclease recognition sites arranged in tandem in the same reading frame as the *lacZ* gene itself. This is the heart of blue/white selection. The *lacZ* gene encodes the β-galactosidase (β-gal) enzyme, which cleaves the galactoside bond in lactose. It will also cleave the galactoside bond in an artificial substrate called X-gal (5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside), which can be added to bacterial growth media and has a blue color when cleaved by the intact enzyme. *If a fragment of DNA is cloned (inserted) into the MCS, the lacZ gene will be disrupted, inactivating it, and the resulting β-gal will no longer be able to cleave X-gal, resulting in white bacterial colonies rather than blue colonies.* This is the blue/white selection mechanism.

Let us now begin the cloning experiment. Following along in Figure 2.4, both the vector and the insert are cut with the same restriction enzyme in order to generate compatible single-stranded sticky ends. The variables here are the abilities to select different enzymes that recognize different restriction sites as long as they generate the same overhang sequence. An enzyme that makes a blunt cut can also be used, although that will make the next step, ligation, less efficient. Two completely different ends with different overhangs can also be used if an exonuclease is used to trim the ends and produce blunt ends. (Continuing with the same reasoning, randomly sheared DNA can also be used if the ends are then blunted for ligation.) If we are forced to use a type I or type III restriction enzyme, the ends must also be blunted. An important alternative is to use two different restriction enzymes that leave different overhangs on each end. The advantages to this are that neither the vector nor the insert will self-circularize, and the orientation of how the insert goes into the vector can be controlled; this is called *directional cloning*. We will select the vector that has the appropriate restriction endonuclease sites.

The next step is to combine the two pools of DNA fragments, vector and insert, in order to connect or ligate them, usually with a 5- or 10-to-1 molar

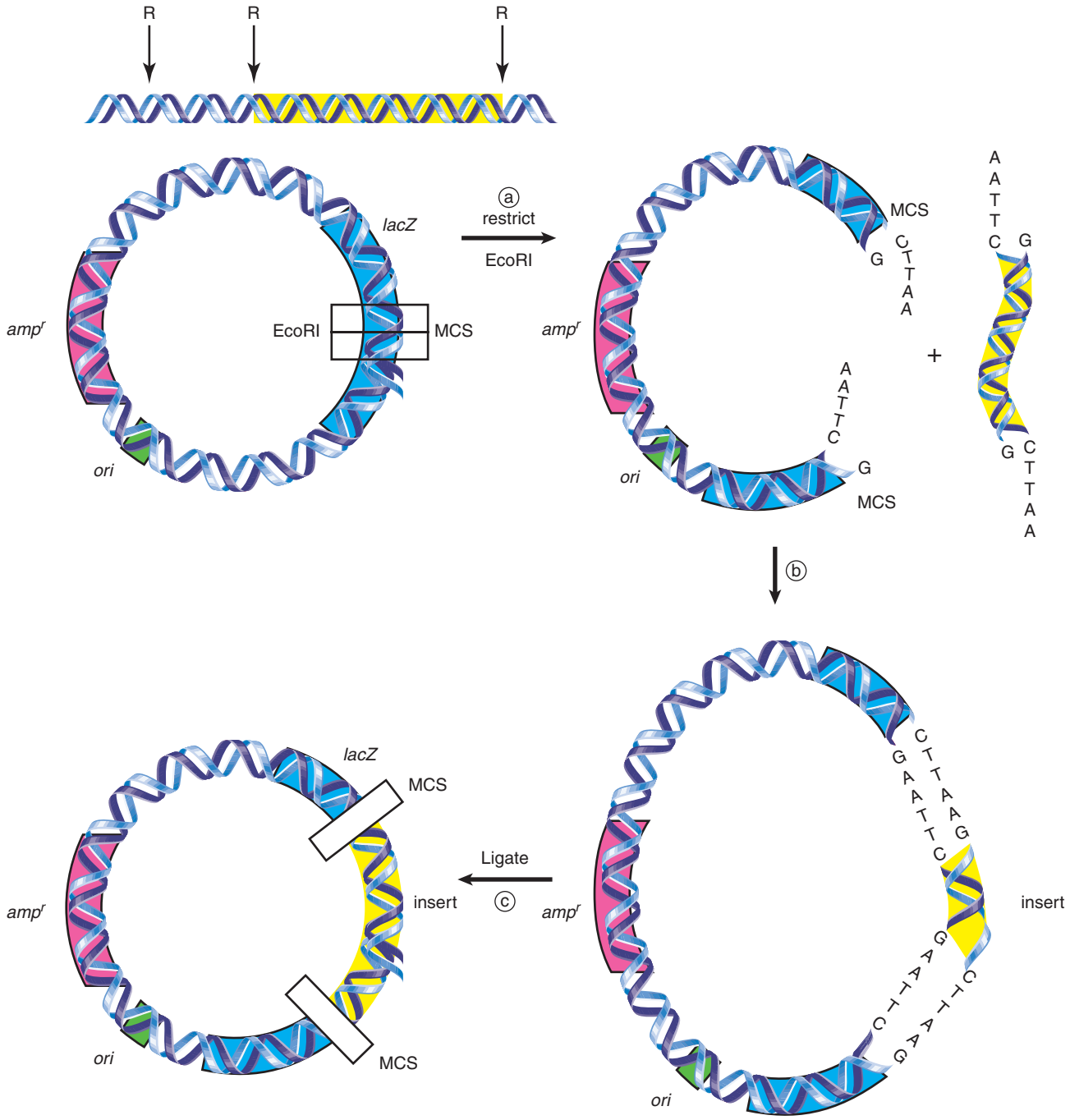


FIGURE 2.4 (a) A plasmid that contains three key sites (an origin of replication, *ori*; a gene for ampicillin resistance, *amp*^{r}; and *lacZ* with an MCS), together with the insert DNA to be cloned, is restricted with EcoR1. (b) Restricted insert fragments and the vector will be combined and (c) ligated together. The final pool of this DNA will be transformed into *E. coli*.

ratio of insert to vector. Large inserts (over 10 kb) cannot be efficiently cloned in a plasmid vector (these are generally cloned into an alternative virus-based vector or artificial chromosomes).

The pool of randomly generated ligated DNA molecules is now used to "transform" *E. coli*. **Transformation** is the process by which DNA is introduced into a host cell. *E. coli* does not normally undergo physiological transformation, so DNA must be forced into the cell using methods such as high $CaCl_2$ treatment or application of an electric current ("electroporation"). Both methods create small holes in the cell wall.

transformation The acquisition of new exogenous genetic material by a cell.

Transformation results in a pool that contains bacteria that have not taken up any DNA at all, those that have taken up vector with no insert, or the minority that have taken up the desired vector with the correct insert. To select the handful of bacteria that contain recombinant plasmid from the millions that do not, the cells are spread on an agar plate containing both the antibiotic ampicillin and an artificial β-gal inducer called IPTG (isopropyl thiogalactoside). The ampicillin will kill the bacterial cells that have not been transformed with the *amp*ρ plasmid. The remaining bacteria can now grow and form visible colonies. As shown in **FIGURE 2.5**, two different types of colonies are seen—blue ones that contain a vector without an insert (because active β-gal cleaved X-gal into a blue compound) and the desired white ones in which the β-gal inactivated by an insert did not cleave X-gal and so remained colorless. Plasmids from the white colonies can then be screened to eliminate false positives and to ensure the correct insert has been obtained.

KEY CONCEPTS

- Cloning a fragment of DNA requires a specially engineered vector.
- Blue/white selection allows the identification of bacteria that contain the vector plasmid and vector plasmids that contain an insert.

CONCEPT AND REASONING CHECK

If you digest a vector and donor DNA with restriction enzymes that make blunt ends, do you have to use the same enzyme for both vector and donor DNA? Why or why not?

cosmid Cloning vector derived from a bacterial plasmid by incorporating the *cos* sites of phage lambda, which make the plasmid DNA a substrate for the lambda packaging system.

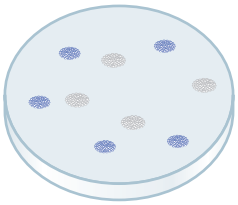


FIGURE 2.5 After transformation into *E. coli* of restricted and ligated vector plus insert DNA, the bacterial cells are plated onto agar plates containing ampicillin, IPTG, and the color indicator, X-gal. Overnight incubation at 37°C will yield both blue and white colonies. The white colonies will be used to prepare DNA for further analysis.

2.4 Cloning Vectors Can Be Specialized

In the example in the previous section, we described the use of a vector that is designed simply for amplifying insert DNA, with inserts up to about 10 kb. It is often desirable to clone larger inserts, though, and sometimes the goal is not just to amplify the DNA but also to express cloned genes in cells, investigate properties of a promoter, create various fusion proteins, or a plethora of other downstream applications. **FIGURE 2.6** summarizes the properties of common classes of cloning vectors. These include vectors based on bacteriophage genomes, which can be used in bacteria but have the disadvantage that only a limited amount of DNA can be packaged into the viral coat (although more than can be carried in a plasmid). The advantages of plasmids and phages are combined in the **cosmid**, which propagates like a plasmid but uses the packaging mechanism of phage lambda to deliver the DNA to the bacterial cells. Cosmids can carry inserts of up to 47 kb (the maximum length of DNA that can be packaged into the phage head).

Two vectors used for cloning the largest possible DNA inserts are the **yeast artificial chromosome (YAC)** and the newer human artificial chromosome (HAC). A YAC has a yeast origin to support replication, a centromere to ensure proper segregation, and telomeres to afford stability. In effect, it is propagated just like a yeast chromosome and can carry inserts measured in the megabase (Mb) length range. The HAC is the roomiest

Vector	Features	Isolation of DNA	DNA limit
Plasmid	High copy number	Physical	10 kb
Phage	Infects bacteria	Via phage packaging	20 kb
Cosmid	High copy number	Via phage packaging	48 kb
BAC	Based on F plasmid	Physical	300 kb
YAC	Origin + centromere + telomere	Physical	3000 kb

FIGURE 2.6 Cloning vectors may be based on plasmids or phages, or may mimic eukaryotic chromosomes.

yeast artificial chromosome (YAC) A cloning vector used to clone very large DNA fragments, up to 3,000 kb in size, containing yeast telomeres, a centromere, and a replication origin so that it can propagate in yeast cells.

addition to the line of vectors, and it offers the advantage of functioning in human cells in addition to its expanded capacity for 10 Mb or more inserts.

The extremely useful class of vectors known as **shuttle vectors** can be used in more than one species of host cell. The example shown in **FIGURE 2.7** contains origins of replication and selectable markers for both *E. coli* and the yeast *Saccharomyces cerevisiae*. It can replicate as a circular multicopy plasmid in *E. coli*. It has a yeast centromere and also has yeast telomeres adjacent to BamH1 restriction sites, so that cleavage with BamH1 generates a YAC that can be propagated in yeast.

shuttle vector A cloning vector that can replicate in two different species.

Other vectors, such as **expression vectors**, may contain promoters to drive expression of genes. Any open reading frame (ORF) can be inserted into the vector and expressed without further modification. These promoters can be continuously active or may be inducible, so that they are only expressed under specific conditions.

expression vector A cloning vector that allows the expression, either translation or just transcription, of the insert.

Alternatively, the goal may be to study the function of a cloned promoter of interest in order to understand the normal regulation of a gene. In this case, rather than using the actual gene, we can use an easily detected **reporter gene** under control of the promoter of interest.

reporter gene A sequence that is attached to another gene, which codes for a peptide that is easily identified or measured.

The type of reporter gene that is most appropriate depends on whether we are interested in quantitating the efficiency of the promoter (and, for example, determining the effects of mutations in it or the activities of transcription factors that bind to it) or determining its tissue-specific pattern of expression. **FIGURE 2.8** summarizes a common system for assaying promoter activity. A cloning vector is created that has a eukaryotic promoter linked to the coding region of *luciferase*, a gene that encodes the enzyme responsible for bioluminescence in the firefly. In general, a transcription termination signal is added to ensure the proper generation of the messenger RNA (mRNA). The hybrid vector is introduced into target cells, and the cells are grown and subjected to any appropriate experimental treatments. The level of luciferase activity is measured by addition of its substrate luciferin. Luciferase activity results in light emission that can be measured at 562 nanometers (nm) and is directly proportional to the amount of enzyme that was made, which in turn depends upon the activity of the promoter.

Some very striking reporters are now available for visualizing gene expression. The *lacZ* gene, described in the blue-white selection strategy before, also serves as a very useful reporter gene. **FIGURE 2.9** shows what

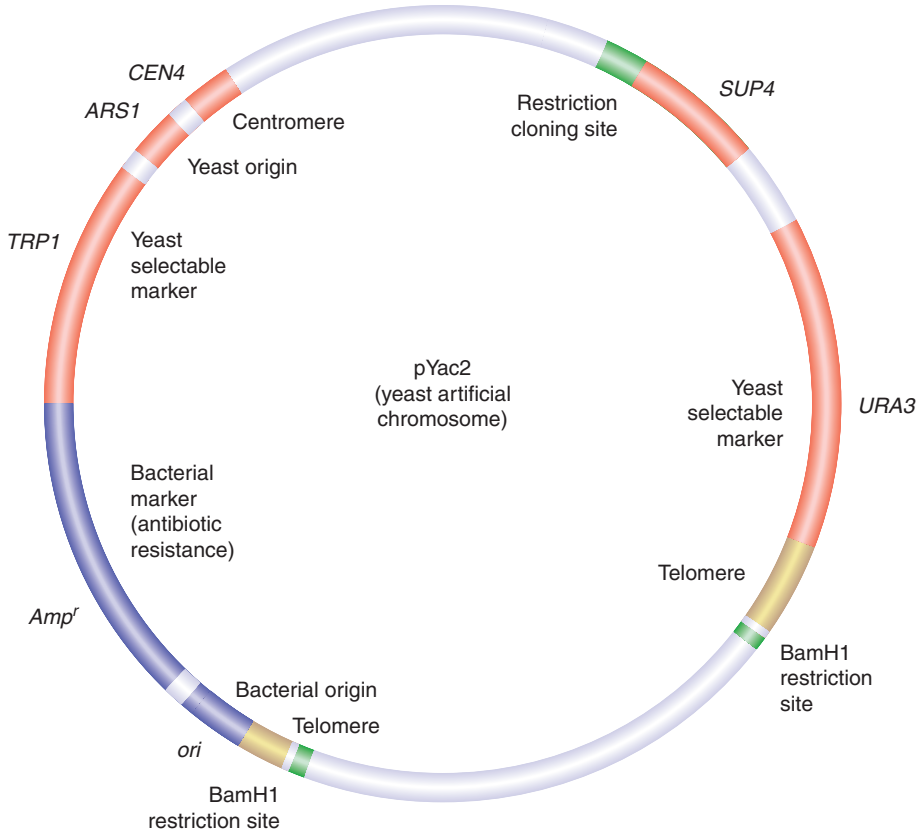


FIGURE 2.7 pYac2 is a cloning vector with features that allow replication and selection in both bacteria and yeast. Bacterial features (described in blue) include an origin of replication and antibiotic-resistance gene. Yeast features (described in orange and yellow) include an origin, a centromere, two selectable markers, and telomeres.

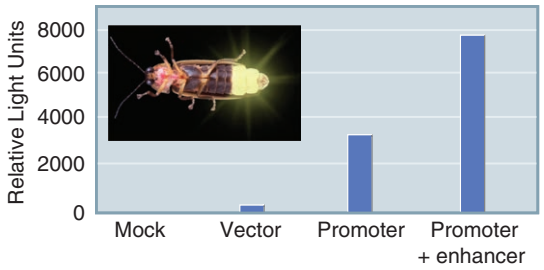


FIGURE 2.8 Luciferase (derived from fireflies such as the one shown here) is a popular reporter. The graph shows the results from mammalian cells transfected with a luciferase vector driven by a minimal promoter or the promoter plus a putative enhancer. The levels of luciferase activity correlate with the activities of the promoters.

happens when the *lacZ* gene is placed under the control of a tissue-specific promoter. The tissues in which this promoter is normally active can be visualized by providing the X-gal substrate to stain the embryo.

One of the most popular reporters that can be used to visualize patterns of gene expression is green fluorescent protein (GFP), which is obtained from jellyfish. GFP is a naturally fluorescent protein that, when excited with one wavelength of light, emits fluorescence in another wavelength. In addition to the original GFP, numerous variants that fluoresce in different colors, such as yellow (YFP), cyan (CFP), and blue (BFP), have been developed. GFP and

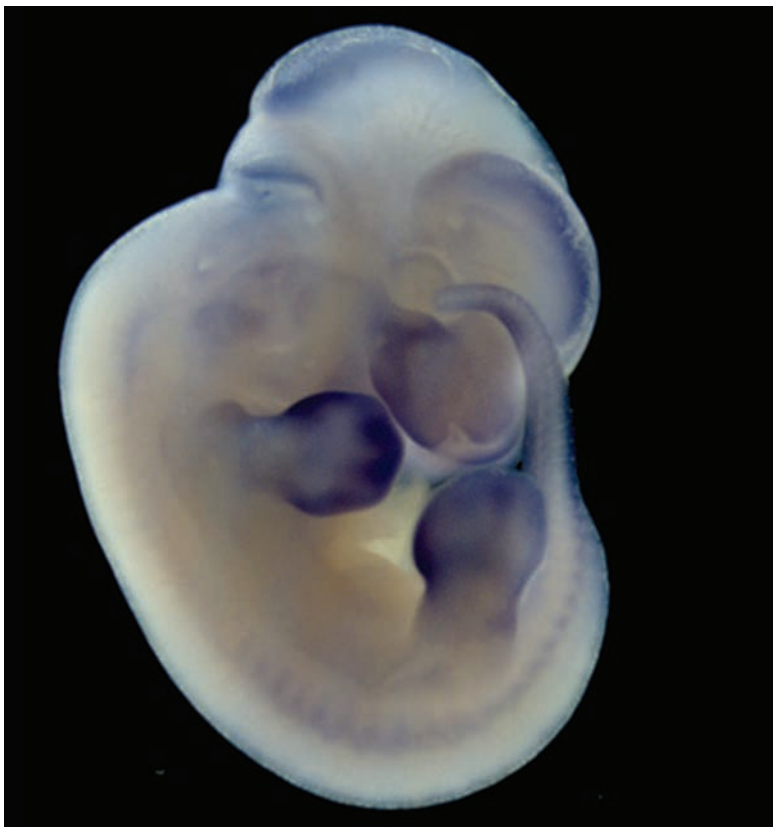


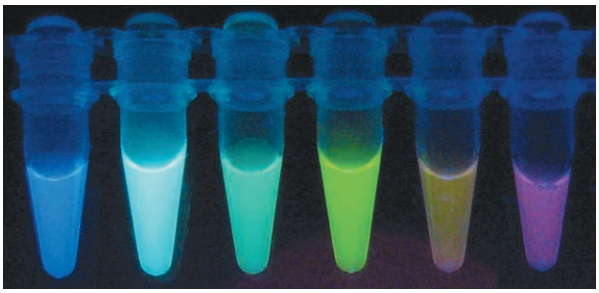
FIGURE 2.9 Expression of a *lacZ* gene can be followed in the mouse by staining for β-galactosidase (in blue). In this example, *lacZ* was expressed under the control of a promoter of a mouse gene that is expressed in the developing heart, limbs, and other tissues. The corresponding tissues can be visualized by blue staining.

Reproduced from Abbasi, A. A., Paparidis, Z., Malik, S., Bangs, F., Schmidt, A., Koch, S., Lopez-Rios, J., and Grzeschik, K. H. (2010). Human intronic enhancers control distinct sub-domains of Gli3 expression during mouse CNS and limb development. *BMC Dev. Biol.* **10**(1), 44.

its variants can be used as reporter genes on their own, or they can be used to generate *fusion proteins*, in which a protein of interest is fused to GFP and can thus be visualized in living tissues, as is shown in the example in **FIGURE 2.10**.

Vectors are introduced into different species in a variety of different ways. Bacteria and simple eukaryotes like yeast can be transformed easily, using chemical treatments that permeabilize the cell membranes, as discussed previously. Many types of cells cannot be transformed so easily, though, and other methods must be used, as summarized in **FIGURE 2.11**. Some types of cloning vectors use natural methods of infection to pass the DNA into the cell, such as a viral vector that uses the viral infective process to enter the cell. *Liposomes* are small spheres made from artificial membranes, which can contain DNA or other biological materials. Liposomes can fuse with plasma membranes and release their contents into the cell. *Microinjection* uses a very fine needle to puncture the cell membrane. A solution containing DNA can be introduced either into the cytoplasm or directly into the nucleus in the case where the nucleus is large enough to be chosen as a target (such as an egg). The thick cell walls of plants are an impediment to many transfer methods, and the "gene gun" was invented as a means for overcoming this obstacle. A gene gun shoots very small particles into the cell by propelling them through the wall at high velocity. The particles can consist of gold or nanospheres coated with DNA. This method has been adapted for use with a variety of species, including mammalian cells.

(a)



(b)

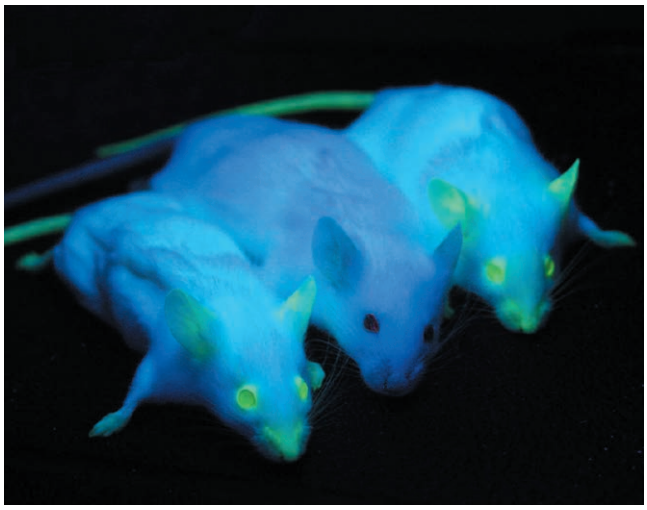


FIGURE 2.10 (a) Since the discovery of GFP, derivatives that fluoresce in different colors have been engineered. (b) A live transgenic mouse expressing human rhodopsin (a protein expressed in the retina of the eye) fused to GFP.

(a) Courtesy of Joachim Goedhart, Molecular Cytology, SILS, University of Amsterdam.

(b) Reproduced from Moen, I., Jevne, C., Wang, J., Kalland, K. H., Chekenya, M., Akslen, L. A., Sleire, L., Enger, P. Ø., Reed, R. K., Øyan, A. M., and Stuhr, L. E. (2012). Gene expression in tumor cells and stroma in dsRed 4T1 tumors in eGFP-expressing mice with and without enhanced oxygenation. *BMC Cancer* **12**(1), 21.

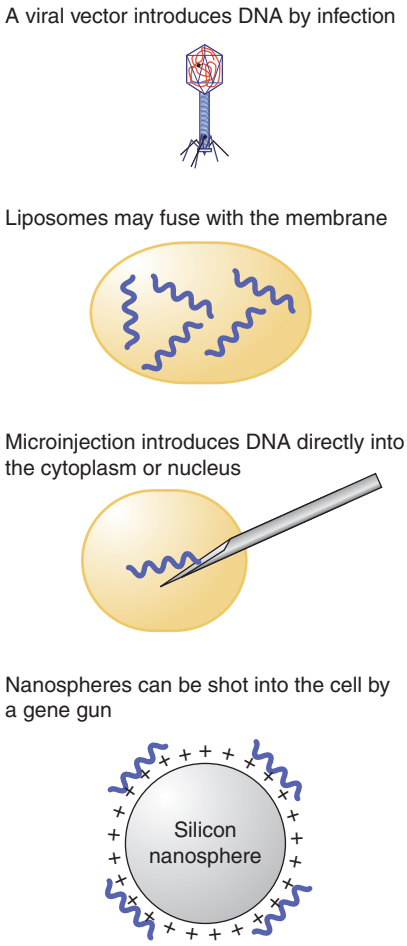


FIGURE 2.11 DNA can be released into target cells by methods that pass it across the membrane naturally, such as by means of a viral vector (in the same way as a viral infection) or by encapsulating it in a liposome (which fuses with the membrane). Alternatively, it can be passed manually, by microinjection, or by coating it on the exterior of nanoparticles that are shot into the cell by a gene gun that punctures the membrane at very high velocity.

KEY CONCEPTS

- Cloning vectors may be bacterial plasmids, phages, cosmids, or YACs.
- Shuttle vectors can be propagated in more than one type of host cell.
- Expression vectors contain promoters that allow transcription of any cloned gene.
- Reporter genes can be used to measure promoter activity or tissue-specific expression.
- Numerous methods exist to introduce DNA into different target cells.

CONCEPT AND REASONING CHECK

Why must a GFP fusion reporter be ligated in the same reading frame as the protein of interest?

2.5 Nucleic Acid Detection

There are a number of different ways to detect DNA and RNA. The classical method relies on the ability of nucleic acids to absorb light with a wavelength of 260 nm. The amount of light absorbed is proportional to the amount of nucleic acid present. There is a slight difference in the amount of absorption by single-stranded as compared to double-stranded nucleic acids, but not DNA versus RNA. Protein contamination can affect the outcome, but because proteins absorb maximally at 280 nm, tables have been published of 260/280 ratios that allow quantitation of the amount of nucleic acid present.

DNA and RNA can be nonspecifically stained with ethidium bromide (EtBr) to make visualization more sensitive. EtBr is an organic tricyclic compound that binds strongly to double-stranded DNA (and RNA) by intercalating into the double helix between the stacked base pairs. It binds to DNA; as a result, it is a strong mutagen, and care must be taken when using it. EtBr fluoresces when exposed to ultraviolet (UV) light, which increases the sensitivity. SYBR green is a safer alternative DNA stain.

There are also methods that allow detection of *specific* sequences of nucleic acids. The ability to identify a specific sequence relies on hybridization of a **probe** with a known sequence to a target. The probe will detect and bind to a sequence to which it is **complementary**. The percentage of match does not have to be perfect, but as the match percentage decreases, the stability of the nucleic acid hybrid decreases. G-C base pairs are more stable than A-T base pairs, so base composition (usually referred to as % G-C) is an important variable. The second set of variables that affects hybrid stability is extrinsic; it includes the buffer conditions (concentration and composition) and the temperature at which hybridization occurs. This is called the **stringency** under which the hybridization is carried out.

probe A labeled nucleic acid used to identify a complementary sequence.

complementary Base pairs that match up in the pairing reactions in double-stranded nucleic acids (A with T in DNA or with U in RNA, and C with G).

stringency A measure of the exactness of complementarity required between two nucleic acid strands to allow them to hybridize. Stringency is related to buffer ionic strength and reaction temperature.

The probe functions as a single-stranded molecule (if it is double-stranded, it must be melted). The target may be single-stranded or double-stranded. If the target is double-stranded, it also must be melted to single strands to begin the hybridization process. The reaction can take place in solution (e.g., during sequencing or PCR) or can be performed when the target has been bound to a solid membrane support, such as a nitrocellulose filter. On a solid support, if the target is DNA, this is called a Southern blot; if it is RNA, it is a Northern blot (the probe is usually DNA in both cases).

For this exercise we will use a Southern blot from an experiment in which we have restricted a large DNA fragment into smaller fragments

and subcloned the individual fragments, as described earlier. Starting with the clones on the plate from Figure 2.5, we isolate plasmid DNA from each white clone and restrict the DNA with the same restriction enzymes that we used to clone the fragments. The DNA fragments are separated on an agarose gel and blotted onto nitrocellulose.

In order to increase sensitivity, the probe must be labeled. The earliest method for doing this was radiolabeling (e.g., with ^{32}P, ^{33}P or ^{3}H), but a number of methods of alternative labeling without radioactivity have since been developed. Probes can be labeled in several different ways. One is *end labeling*, in which a strand of DNA (which has no 5′ phosphate) is labeled using a kinase and phosphate label with ^{32}P or other label. Alternatively, a probe can be generated by *nick-translation* or *random priming* with labeled nucleotides using the Klenow DNA polymerase fragment or during a PCR reaction (discussed later in this chapter).

In performing nucleic acid hybridization studies, standard procedures are typically used that allow hybridization over a large range of G-C content, with salt concentration and hybridization temperatures varying over a range depending upon the required stringency. The actual hybridization between a labeled probe and a target DNA bound to a membrane usually takes place in a sealed container in a buffer that contains a set of molecules to reduce background hybridization of probe to the filter. The hybridization reaction is stochastic and depends upon the abundance of each different sequence. The more copies of a sequence, the greater the chance of a given probe molecule encountering its complementary sequence.

After hybridization, the filter is washed to remove all the probe that is not specifically bound to a complementary sequence of nucleic acid. Again, the stringency can be varied as needed to promote detection of perfectly matched targets (high stringency) or to tolerate mismatches (low stringency). Higher-stringency conditions include higher temperature (closer to the melting temperature of the probe) and lower concentration of cations. (Lower salt concentrations result in less shielding of the negative phosphate groups of the DNA backbone, which in turn inhibits strand annealing.)

Finally, the location in which the probe has hybridized to the target(s) must be detected. In the case of a radiolabeled probe, the washed filter is subjected to **autoradiography**. The dried filter is placed against a sheet of X-ray film, a *phosphorimaging* screen (a solid-state liquid scintillation device), or similar. **FIGURE 2.12** shows a illustration of the result. A fragment of a particular size in several samples has blackened the X-ray film. The film can be aligned to the filter to determine which band corresponds to the probe.

autoradiography A method of capturing an image of radioactive materials on film or nuclear emulsion.

A simple modification of the autoradiography procedure called *in situ* **hybridization** allows one to peer into a cell and determine the location, at a microscopic level, of specific nucleic acid sequences. We simply modify a few steps in the preceding process to perform the hybridization between our probe, usually labeled with ^{3}H, and complementary nucleic acids in an intact cell or tissue. The goal is to determine exactly where the target is located. The cell or tissue slice is mounted on a microscope slide. Following hybridization, a photographic emulsion instead of film is applied to the slide, covering it. The emulsion, when developed, is transparent to visible light so that it is possible to see the exact location in the cell where the grains in the emulsion blackened by the radioactivity are located. Development time can be weeks to months because ^{3}H has less energetic radiation and its longer half-life results in lower activity.

hybridization The pairing of complementary nucleic acid strands from different sources to give a hybrid.

As noted earlier, there are nonradioactive alternatives to these procedures that use either colorimetric or fluorescence labeling. The digoxygenin-labeled

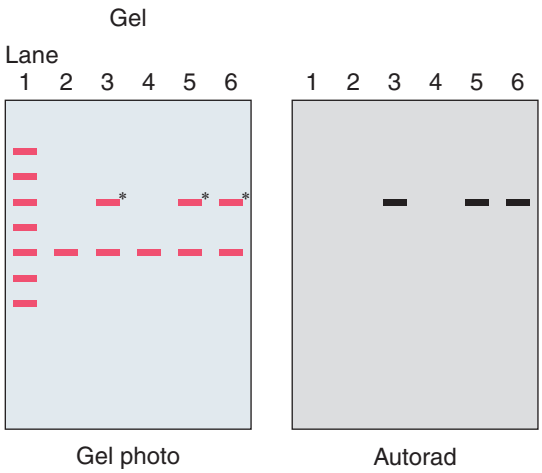


FIGURE 2.12 Illustration of an autoradiogram of a gel prepared from the colonies described in Figure 2.5. The gel was blotted onto nitrocellulose and probed with a radioactive gene fragment. Lane 1 contains a set of standard DNA size markers. Lane 2 is the original vector cleaved with EcoR1. Lanes 3–6 each contain plasmid DNA from one of the white clones from Figure 2.4 that was restricted with EcoR1. A illustration of the photograph of the gel is on the left; the radioactive bands are marked with an asterisk.

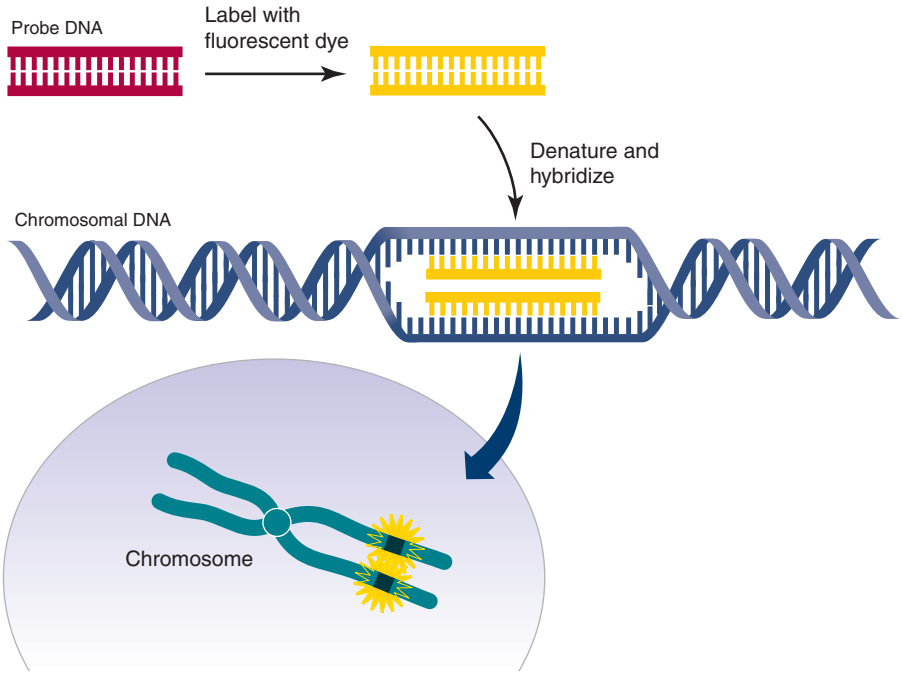


FIGURE 2.13 Fluorescence *in situ* hybridization (FISH).

Based on illustration by Darryl Leja, National Human Genome Research Institute (www.genome.gov).

probe is a commonly used colorimetric procedure. Probe bound to the target is localized with an antidigoxygenin antibody coupled to alkaline phosphatase, an enzyme that acts on a colorless substrate to develop color. The advantage is the time required to see the results. It is typically a single day, but sensitivity is usually less than with radioactivity. Fluorescence *in situ* hybridization (FISH) is another very common nonradioactive procedure that uses a fluorescently labeled probe. This method is illustrated in **FIGURE 2.13**. Multiple fluorophores in different colors are available—more than a dozen—but ratios of different probe color combinations can be used to create additional colors.

This is still less quantitative than traditional scintillation counting but is a highly versatile method that allows detection of multiple sequences in the same sample.

KEY CONCEPT

- Hybridization of labeled nucleic acid to complementary sequences can identify specific nucleic acids in a mixture.

CONCEPT AND REASONING CHECK

Why is the % G-C content an important variable in hybridization reaction?

2.6 DNA Separation Techniques

With a few exceptions, the individual pieces of DNA (chromosomes) making up a living organism's genome are on the order of megabases in length, making them too physically large to be manipulated easily in the laboratory. Individual genes or chromosomal regions of interest, by contrast, are often quite small and readily manageable, on the order of hundreds or a few thousand base pairs in length. A necessary first step, therefore, in many experimental processes investigating a specific gene or region is to break the large original chromosomal DNA molecule down into smaller manageable pieces and then begin isolation and selection of the particular relevant fragment or fragments of interest. This breakage can be done by mechanical shearing of chromosomes in a process that produces breakages randomly to produce a uniform size distribution of assorted molecules. This approach is useful if a randomness in breakpoints is required, such as to create a library of short DNA molecules that "tile," or partially overlap, each other while together representing a much larger genomic region, such as an entire chromosome or genome. Alternatively, restriction endonucleases may be employed to cut large DNA molecules into defined shorter segments in a way that is reproducible. This reproducibility is frequently useful, in that a DNA section of interest can be identified in part by its size. Consider a hypothetical gene *genX* on a bacterial chromosome, with the entire gene lying between two EcoRI sites spaced 2.3 kb apart. Digestion of the bacterial DNA with EcoRI will yield a range of small DNA molecules, but *genX* will always occur on the same 2.3-kb fragment. Depending on the size and complexity of the starting genome, several other DNA segments of similar size may be produced, or in a simple enough system, this 2.3-kb size may be unique to the *genX* fragment. In this latter case, detection or visualization of a 2.3-kb fragment is enough to definitively identify the presence of *genX*. Many of the earliest laboratory techniques developed in working with DNA relate to separating and concentrating DNA molecules based on size expressly to take advantage of these concepts. An ability to separate DNA molecules based on size allows for taking a complex mixture of many fragment sizes and selecting a much smaller, less complex subset of interest for further study.

The simplest method for separation and visualization of DNA molecules based on size is gel electrophoresis. In neutral agarose gel electrophoresis, the most basic type of gel, this is done by preparing a small slab of gel in an electrically conductive, mildly basic buffer. While similar to the gelatins used

to make dessert dishes, this type of gel is made from agarose, a polysaccharide that is derived from seaweed and has very uniform molecular sizes. Preparation of agarose gels of a specific percentage of agarose by mass (usually in the range of 0.8–3%) creates, in effect, a molecular sieve, with a "mesh" pore size being determined by the percentage of agarose (higher percentages yielding smaller pores). The gel is poured in a molten state into a rectangular container, with discrete wells forming near one end of the product. After cooling and solidifying, the slab is submerged in the same conductive, mildly alkaline, buffer, and samples of mixed DNA fragments are placed in the preformed wells. A direct current (DC) electric current is then applied to the gel, with the positive charge being at the opposite end of the gel from the wells. The alkalinity of the solution ensures that the DNA molecules have a uniform negative charge from their backbone phosphates, and the DNA fragments begin to be drawn electrostatically toward the positive electrode. Shorter DNA fragments are able to move through the agarose pores with less resistance than longer fragments, and so over time the smallest DNA molecules move the farthest from the wells and the largest move the least. All fragments of a given size will move at about the same rate, effectively concentrating any population of equal-sized molecules into a discrete band at the same distance from the well. Addition of a DNA-binding fluorescent dye, such as EtBr or SYBR green, to the gel stains these DNA bands such that they can be directly seen by eye when the gel is exposed to fluorescence-exciting light. In practice, a standard sample consisting of a set of DNA molecules of a known size is run in one of the wells, with sizes of bands in other wells estimated in comparison to the standard, as shown in **FIGURE 2.14**.

DNA molecules of roughly 50–10,000 bp can be quickly separated, identified, and sized to within about 10% accuracy by this simple method, which remains a common laboratory technique. DNA can be separated not only by size but also by shape. Supercoiled DNA, which is compact compared to relaxed or linear DNA, migrates more rapidly on a gel, and the more supercoiling, the faster the migration, as seen in **FIGURE 2.15**.

Variations on this method primarily relate to changing the gel matrix from agarose to other molecules, such as synthetic polyacrylamides, which can have even more precisely controlled pore sizes. These can offer finer-size resolution of DNA molecules from roughly 10 to 1,500 bp in size. Both resolution and sensitivity are further improved by making these types of gels as thin as possible, normally requiring they be formed between glass plates for mechanical strength. When chemical denaturants such as urea are added to the buffer system, the DNA molecules are forced to unfold (losing any secondary structures) and take on hydrodynamic properties related only to molecule length. This approach can clearly resolve DNA molecules differing in length by only a single nucleotide. Denaturing polyacrylamide electrophoresis is a key component of classical DNA sequencing techniques whereby the separation and detection of a series of single nucleotide length-difference DNA products allows for the reading of the underlying order of nucleotide bases.

The next level of refinement to this technique is to place the gel matrix in a very fine capillary, which can be even thinner than a glass plate–supported gel and thus still further improve on sensitivity and resolution capacity. Unlike a glass-supported slab gel, where multiple lanes can be run side by side, a capillary can handle only one sample at a time; however, a capillary can be run clean of sample and reused, making it ideal for system automation and high-throughput applications. Instruments with multiple parallel capillaries allow for parallel analysis of multiple samples to further increase

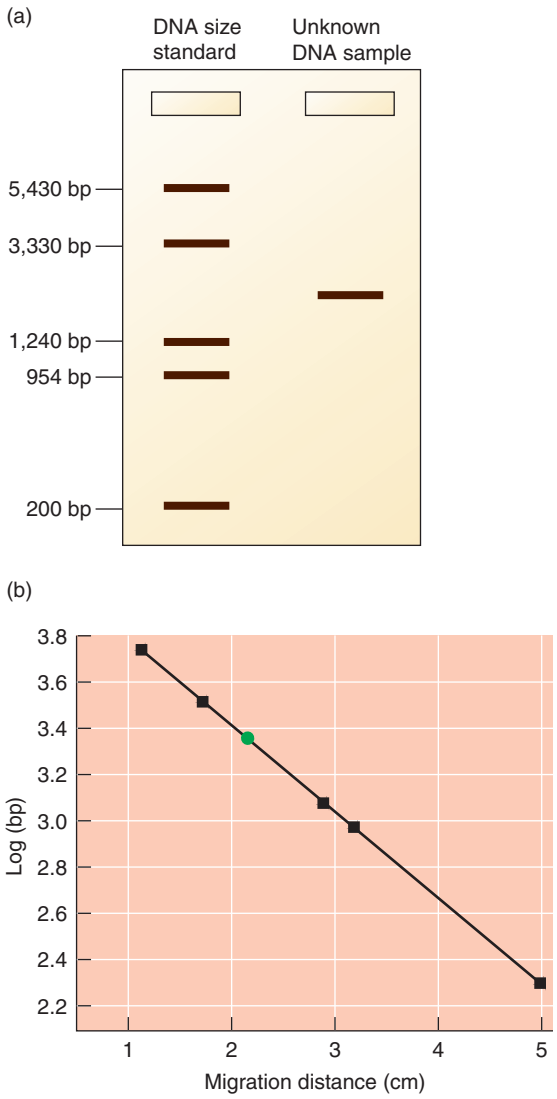


FIGURE 2.14 DNA sizes can be determined by gel electrophoresis. A DNA size standard and a DNA of unknown size are run in two lanes of a gel, depicted schematically. The migration of the DNAs of known size in the standard is graphed to create a standard curve (migration distance in cm vs. log bp). The point shown in green is for the DNA of unknown size.

Based on illustration by Michael Blaber, Florida State University.

throughput. Technologies of this form mark the apparent apex of chain termination–based sequencing methods and are now largely replaced with next-generation sequencing methods.

Further miniaturization of capillaries onto the surfaces of inert "chips" with etched-in microfluidic reservoirs, valves, pumps, and mixing chambers can be employed to create entire "lab-on-a-chip" disposable nucleic acid sample analysis cartridges. These cartridges can process, separate, perform size analysis, and quantitate DNA or RNA in a small input sample. Frequently, these devices are controlled and have data output processed by a computer, which in turn will manipulate the data output in order to present it as a traditional stained agarose or polyacrylamide gel—in effect, bringing the technology full circle.

Another method for separating DNA molecules from other contaminating biomolecules—or, in some cases, for fractionation of specific small DNA

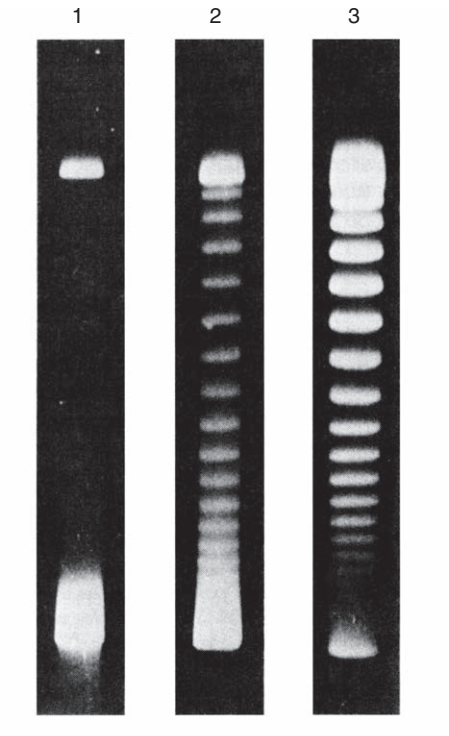


FIGURE 2.15 Supercoiled DNAs separated by agarose gel electrophoresis. Lane 1 contains untreated negatively supercoiled DNA (lower band). Lanes 2 and 3 contain the same DNA that was treated with a type 1 topoisomerase for 5 or 30 minutes, respectively. The topoisomerase makes a single strand break in the DNA and relaxes negative supercoils in single steps (one supercoil relaxed per strand broken and reformed).

Reproduced from Keller, W. (1975). *Proc. Natl. Acad. Sci. USA* **72**, 2550–2554. Photo courtesy of Walter Keller, University of Basel.

molecules from other kinds of DNA—is through the use of gradients, as depicted in **FIGURE 2.16**. The most frequent implementation of this is *isopycnic banding*, which is based on the fact that specific DNA molecules have unique densities based on their G-C content. Under the influence of extreme g-forces, such as through ultracentrifugation, a high-concentration solution of a salt (such as cesium chloride, CsCl) will form a stable density gradient from low density (near the top of the tube/center of rotor) to high density (near the bottom of the tube or outside of the rotor). When placed on top of this gradient (or even mixed uniformly within the gradient) and subjected to continued centrifugation, individual DNA molecules will migrate to a position in the gradient where their density matches that of the surrounding medium. Individual DNA bands can then be either visualized (e.g., through the incorporation of DNA-binding fluorescent dyes in the gradient matrix and exposure to fluorescence excitation) or recovered by careful puncture of the centrifuge tube and fractional collection of the tube contents. This method can also be used to separate double-stranded from single-stranded molecules and RNA from DNA molecules, again based solely on density differences.

Choice of the gradient matrix material, its concentration, and the centrifugation conditions can influence the total density range separated by the process, with very narrow ranges being used to fractionate one particular type of DNA molecule from others and wider ranges being used to separate types of DNA in general from other biomolecules. Historically, one of the

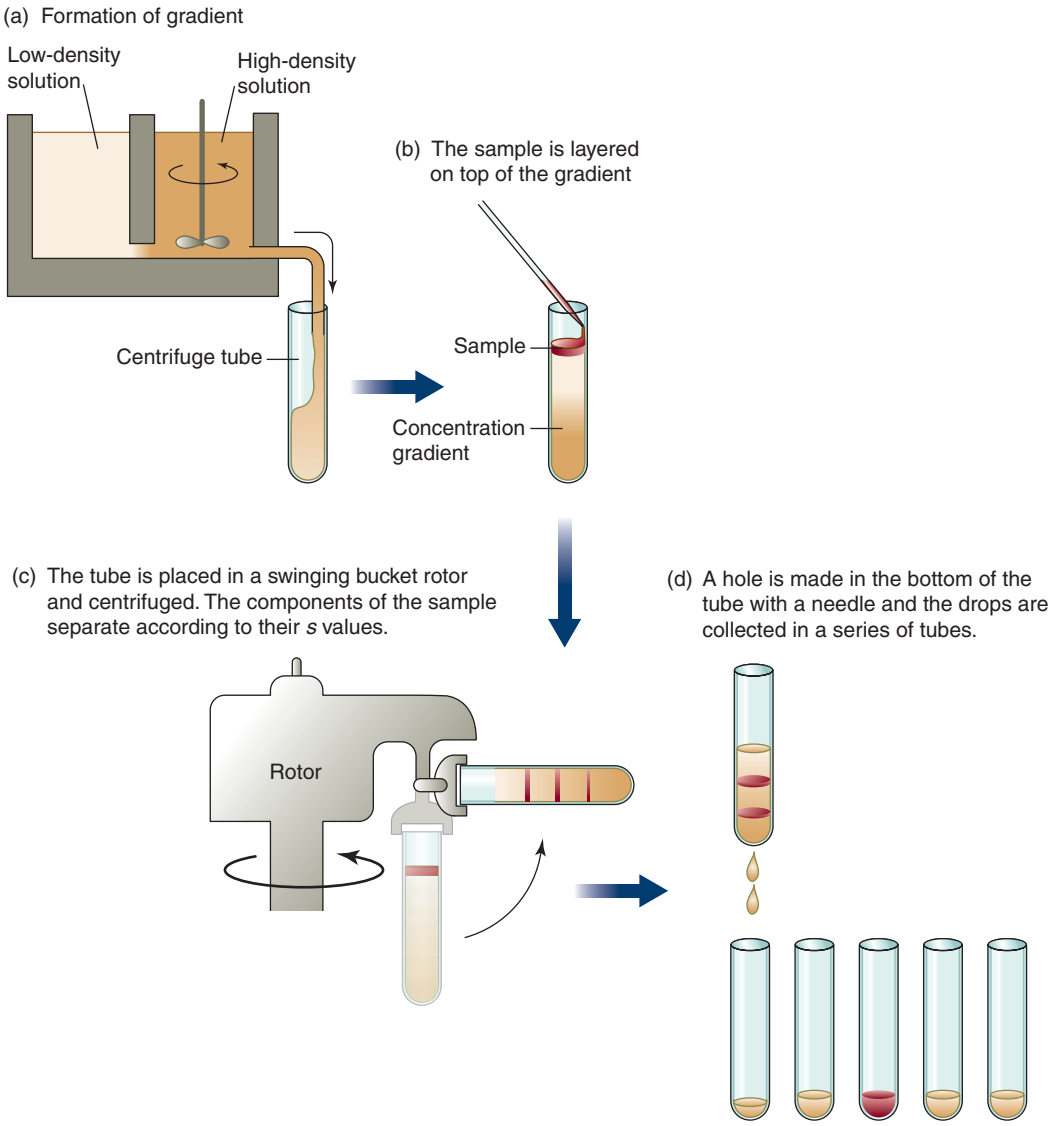


FIGURE 2.16 Gradient centrifugation separates samples based on their density.

best-known uses of this technique was in the Meselson–Stahl experiment of 1958 (introduced in the *Genes Are DNA and Encode RNAs and Polypeptides* chapter), in which the stepwise density changes in the DNA genomes of bacteria shifted from growth in "heavy" nitrogen (^{15}N) to "regular" nitrogen (^{14}N) were observed. The method's capacity to differentially band DNA with pure ^{15}N, half ^{15}N/half ^{14}N, and pure ^{14}N conclusively demonstrated the semiconservative nature of DNA replication. Today, the method is most frequently employed as a large-scale preparative purification technique, with wider density ranges to purify DNA types as a group away from proteins and RNAs.

KEY CONCEPTS

- Gel electrophoresis separates DNA fragments by size, using an electric current to cause the DNA to migrate toward a positive charge.
- DNA can also be isolated using density gradient centrifugation.

CONCEPT AND REASONING CHECK

Why are DNA fragments of different G-C contents able to be separated by CsCl gradients?

2.7 DNA Sequencing

For many years, the dominant approach for DNA sequencing was based on variations of *dideoxy sequencing*, a method developed by Frederick Sanger and colleagues in 1977 (Sanger, along with Walter Gilbert, shared the 1980 Nobel Prize in Chemistry for developing DNA sequencing). This method requires many identical copies of the DNA, an oligonucleotide **primer** that is complementary to a short stretch of the DNA, DNA polymerase, deoxynucleotides (dNTPs: dATP, dCTP, dGTP, and dTTP), and **dideoxynucleotides (ddNTPs)**. Dideoxynucleotides are modified nucleotides that can be incorporated into the growing DNA strand but lack the 3′ hydroxyl group needed to attach the next nucleotide. Thus, their incorporation terminates the synthesis reaction. The ddNTPs are added at much lower concentrations than the normal nucleotides so that they are incorporated at a low rate, randomly, and often only after synthesis has proceeded normally for a strand length of up to several hundred nucleotides.

primer A short sequence that is paired with one strand of DNA and provides a 3′-OH end at which an DNA polymerase starts synthesis of a DNA chain.

dideoxynucleotide (ddNTP) A chain-terminating nucleotide that lacks a 3′-OH group and therefore is not a substrate for DNA polymerization; used in DNA sequencing.

Originally, four separate reactions were necessary, with a single different ddNTP added to each one. The reason for this was that the strands were labeled with radioisotopes and could not be distinguished from each other on the basis of the label. Thus, the reactions were loaded into adjacent lanes on a denaturing acrylamide gel and separated by electrophoresis at a resolution that distinguished between strands differing by a length of one nucleotide. The gel was transferred to a solid support, dried, and exposed to a film. The results were read from top to bottom, with a band appearing in the ddATP lane indicating that the strand terminated with an adenine, the next band appearing in the ddTTP lane indicating that the next base was a thymine, and so on.

A major improvement of the classic method was to use a different fluorescent label for each ddNTP. This allows a single reaction to be run that is read as the strands are hit with a laser and pass by an optical sensor, with a computer detecting which ddNTP terminated each sequential fragment. This became more efficient with the replacement of large polyacrylamide slabs with the thin, long, glass capillary tubes filled with gel, as described earlier. These tubes dissipate heat more rapidly, allowing the electrophoresis to be run at a higher voltage, greatly reducing the time required for separation. A schematic illustrating this process is shown in **FIGURE 2.17**. These modifications, with their resulting automation and increased throughput, ushered in the era of whole-genome sequencing.

This was the process used to sequence the first set of genomes, including the human genome. It was relatively slow and very expensive. The determination of the human genome sequence took several years and cost several billion dollars to complete.

The next generation of sequencing technologies that followed sought to eliminate the need for time-consuming gel separation and reliance on human labor. Modifications of procedures and new instrumentation beginning in about 2005 aided in the automation and scaling up of the procedure. This

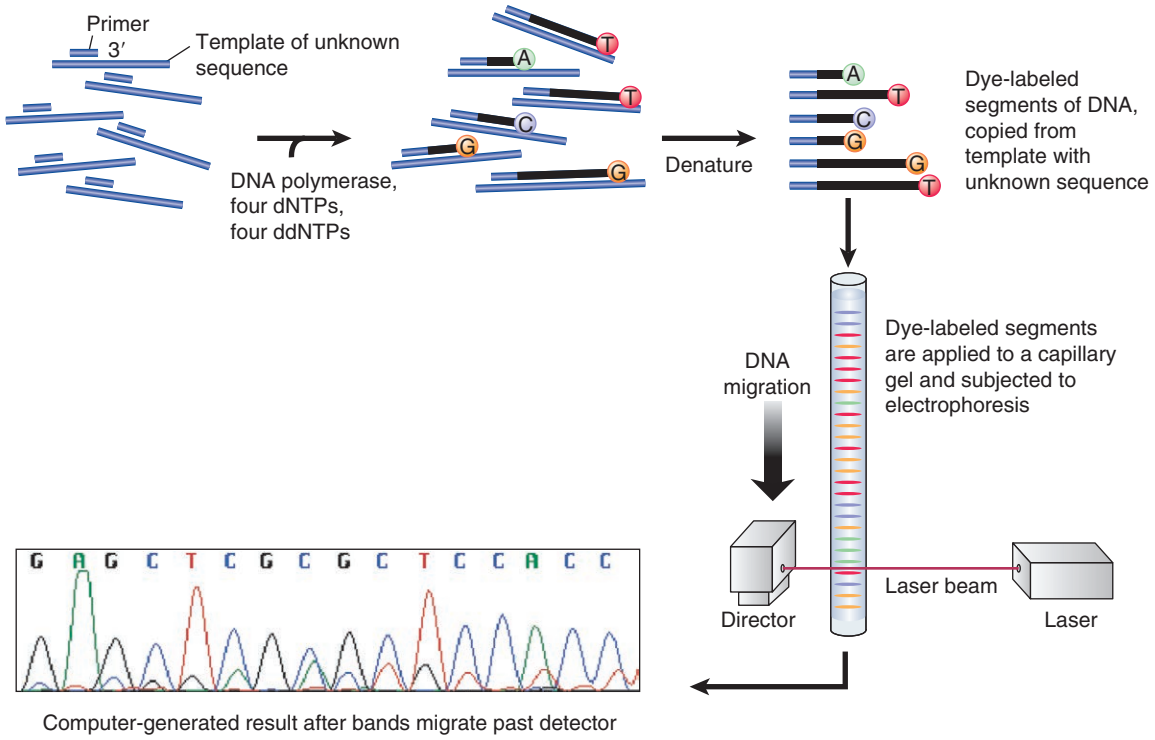


FIGURE 2.17 ddNTP sequencing using fluorescent tags.

Inset photo courtesy of Jan Kieleczawa.

was called next-generation sequencing (NGS), and is now known as second-generation or *massively parallel sequencing*. This still requires amplification of the starting material, which is first randomly fragmented and then amplified. Individual amplified fragments (typically very short, a few hundred bp) are anchored to a solid support and read out one base, in one set of fragments, at a time, in a massively parallel array. These modifications resulted in sequencing on a very large scale at a much lower cost per kb of DNA than the original first-generation methods.

This technology, sometimes called sequencing-by-synthesis (SBS) or wash-and-scan sequencing, relies on the detection and identification of each nucleotide as it is added to a growing strand. In one such application, the primer is tethered to a glass surface and the complementary DNA to be sequenced anneals to the primer. Sequencing proceeds by adding polymerase and fluorescently labeled nucleotides individually, washing away any unused dNTPs. After illuminating with a laser, the nucleotide that has been incorporated into the DNA strand can be detected. Other versions use nucleotides with reversible termination, so that only one nucleotide can be incorporated at a time even if there is a stretch of homopolymeric DNA (such as a run of adenines). Still another version, called pyrosequencing, detects the release of pyrophosphate from the newly added base. These second-generation systems utilize amplification of material to produce massively parallel analysis runs, but the drawback is that there are typically short read lengths (a few hundred bases). The data then requires computation to stitch them together into *contigs* (contiguous sequences). Currently, second-generation methods can generate about a billion independent reads in a day or two—a terabase of sequence!

Now, more than 40 years after the introduction of dideoxy sequences, technology has moved from this second generation to a set of third-generation systems, or *real-time, single-molecule sequencing*. Third-generation

sequencing is a collection of methods that avoids the problems of amplification by direct sequencing of the DNA using single-molecule-sequencing (SMS) templates fixed to a surface for sequencing. Nanopore sequencing is a third-generation method that is progressing rapidly. This approach detects individual nucleotides based on patterns of ion flow as a DNA sequence is run through a silicone nanopore. Tiny transistors are used to control a current passing through the pore. As a nucleotide passes through, it disturbs the current in a manner unique to its chemical structure. Reads approaching a megabase have been achieved with this technology, which also can detect covalent modifications in the DNA, such as methylation. Other methods under development include examination by electron microscopy and single-base synthesizing.

KEY CONCEPTS

- Chain-termination sequencing uses dideoxynucleotides to terminate DNA synthesis at particular nucleotides.
- Fluorescently tagged ddNTPs and capillary gel electrophoresis allow automated, high-throughput DNA sequencing.
- Second-generation (massively parallel) sequencing increases automation and decreases time and cost of sequencing; read lengths are short but are performed on billions of templates simultaneously.
- Third-generation (real-time, single-molecule) sequencing techniques are advancing rapidly and do not require the amplification steps needed in first- and second-generation methods.

CONCEPT AND REASONING CHECK

What is the role of dideoxynucleotides in first- and second-generation DNA sequencing?

2.8 PCR and RT-PCR

Few advances in the life sciences have had the broad-reaching and even paradigm-shifting impact of the **polymerase chain reaction (PCR)**. Although evidence exists that the underlying core principles of the method were understood and in fact used in practice by a few isolated people prior to 1983, credit for independent conceptualization of the mature technology and foresight of its applications goes to Kary Mullis, who was awarded the 1993 Nobel Prize in Chemistry for his insight.

polymerase chain reaction (PCR) A process for the amplification of a defined nucleic acid section through repeated thermal cycles of denaturation, annealing, and polymerase extension.

The underlying concepts are simple and based on the knowledge that DNA polymerases require a template strand with an annealed primer containing a 3′ hydroxyl to commence strand extension. The steps of PCR are illustrated in **FIGURE 2.18**. Although in the context of normal cellular DNA replication (see the *DNA Replication* chapter) this primer is in the form of a short RNA molecule provided by DNA primase, it can equally well be provided in the form of a short, single-stranded, synthetic DNA oligonucleotide having a defined sequence complementary to the 3′ end of any known sequence of interest. Heating of the double-stranded target sequence of interest (known as the "template molecule," or just "template," for short) to near 100°C in appropriate buffer causes thermal denaturation as the template strands melt apart from each other (Figure 2.18a and b). Rapid cooling to the annealing temperature (or T_m) of the primer/template pair and a vast molar excess of the short, kinetically active synthetic primer ensures that a primer molecule finds and appropriately anneals to its complementary target sequence more rapidly

T_m The theoretical melting temperature of a duplex nucleic acid segment into separate strands. T_m is dependent on parameters that include sequence composition, duplex length, and buffer ionic strength.

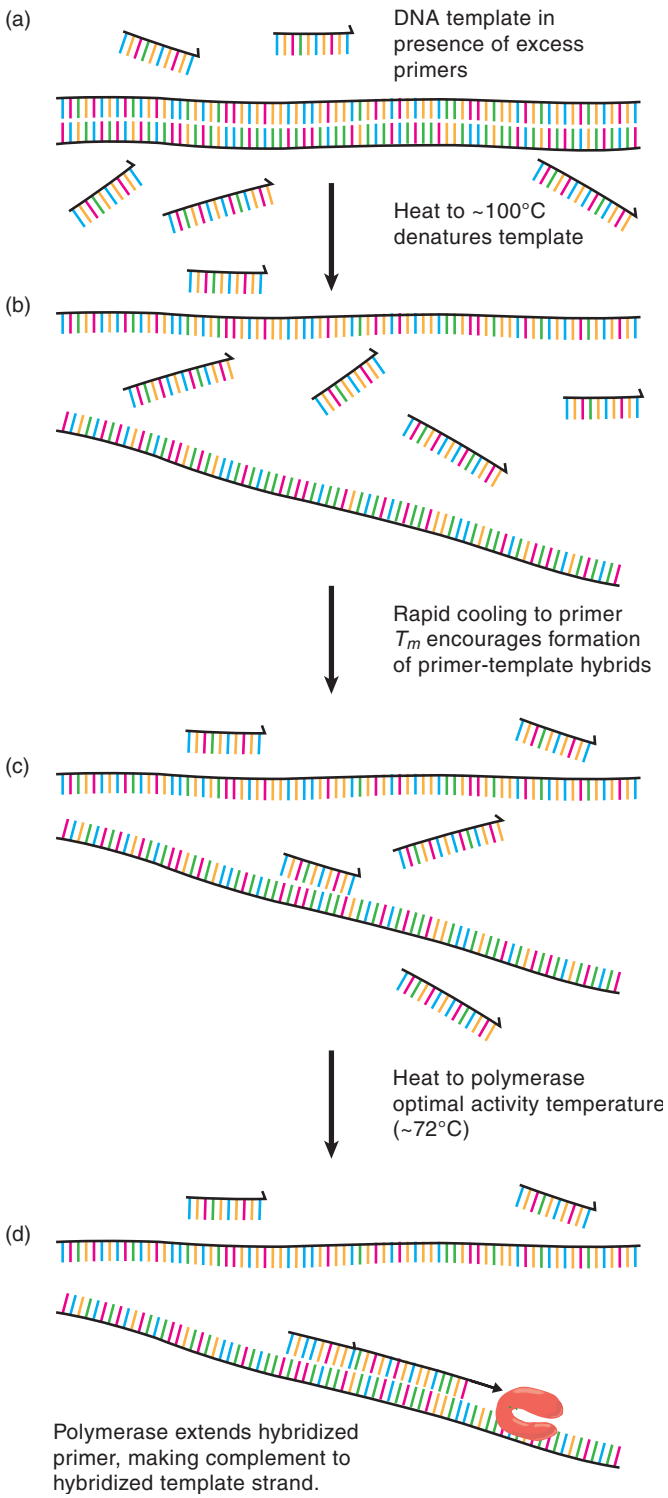


FIGURE 2.18 Denaturation (a) and rapid cooling (b) of a DNA template molecule in the presence of excess primer allows the primer to hybridize to any complementary sequence region of the template (c). This provides a substrate for polymerase action and primer extension (d), creating a complementary copy of one template strand downstream from the primer.

than the original opposing strand can do so (Figure 2.18c). If presented to a polymerase, this annealed primer presents a defined location from which to commence primer extension (Figure 2.18d). In general, this extension will occur until either the polymerase is forced off the template or it reaches the 5′ end of the template molecule and effectively runs out of template to copy.

The ingenuity of PCR arises from simultaneously incorporating a nearby second primer of opposing polarity (i.e., complementary to the opposite strand the first primer anneals to) and then subjecting the mixture of template, two primers (at high concentrations), thermostable DNA polymerase, and dNTP containing polymerase buffer to repeated cycles of thermal denaturation, annealing, and primer extension. Consider just the first cycle of the process: denaturation and annealing occur as described previously, but with both primers, creating the situation depicted in **FIGURE 2.19**. If polymerase

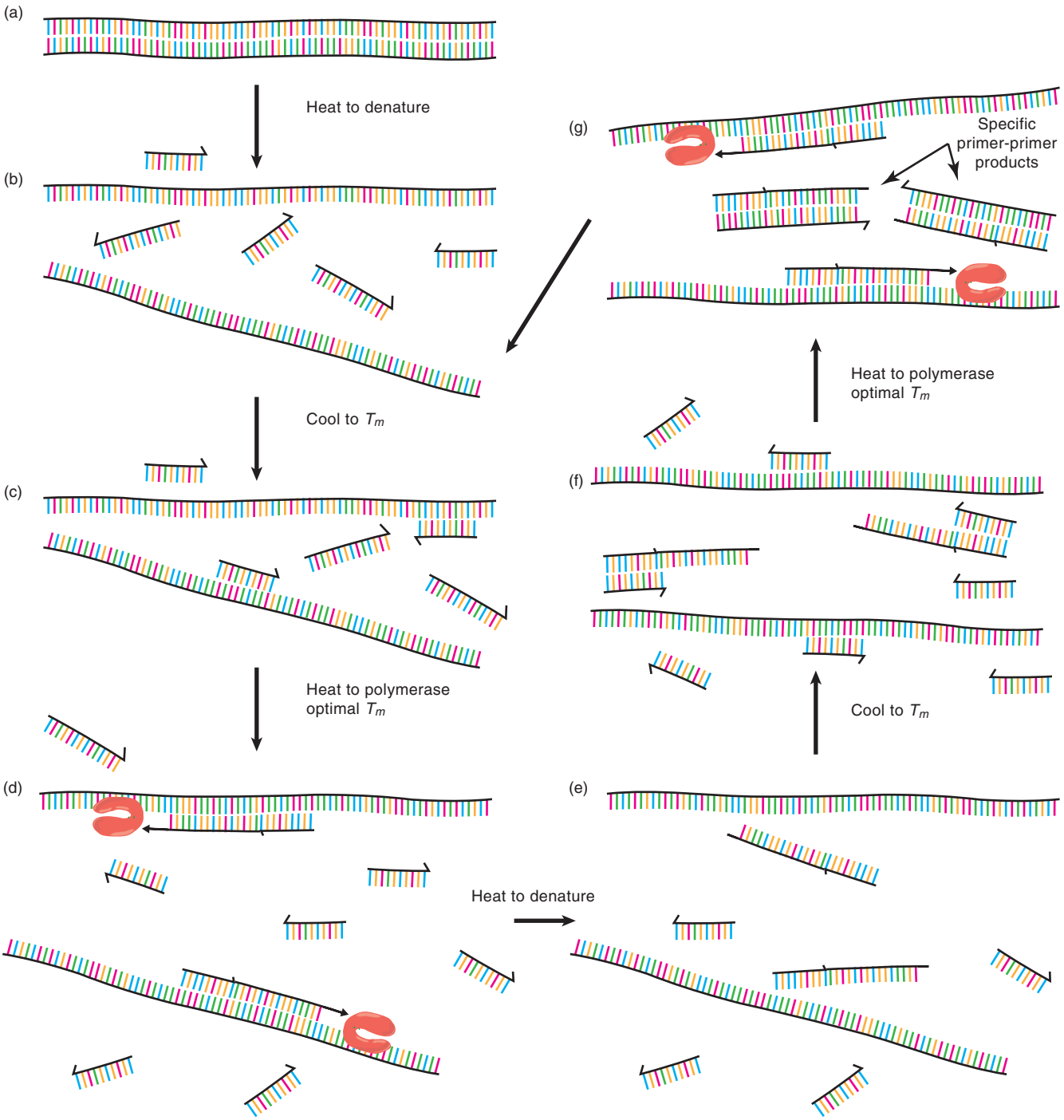


FIGURE 2.19 Thermally driven cycles of primer extension, where primers of opposite polarity have nearby priming sites on each of the two template strands, which leads to the exponential production of the short, primer-to-primer–defined sequence (the amplicon).

extension is allowed to proceed for a short period of time (on the order of 1 minute per 1,000 bp), each of the primers will be extended out and past the location of the other, thus creating a new complementary annealing site for the opposing primer. Raising the temperature back to denaturation stops the primer elongation process and displaces the polymerases and newly created strands. As the system is cooled once more to the annealing temperature, each of the newly formed, short, single DNA strands serves as an annealing site for its opposite polarity primer. In this second thermal cycle, extension of the primers proceeds only as far as the template exists—that is, the 5′ end of the opposing primer sequence. The process has now made both strands of the short, defined, precisely primer-to-primer DNA sequence. Repeating the thermal steps of denaturation, annealing, and primer extension lead to an exponential (2^N, where N is number of thermal cycles) increase in the number of this defined product, allowing for phenomenal levels of "sequence amplification." Close consideration of the process reveals that while this also creates uncertain-length products from the extension of each primer off the original template molecule with each cycle, these products accrue in a linear fashion and are quickly vastly outnumbered by the primer-to-primer defined product (known as the **amplicon**). In fact, within 40 thermal cycles of an idealized PCR reaction, a single-template DNA molecule generates approximately 10^{12} amplicons—more than enough to go from an invisible target to a clearly visible, fluorescent dye–stained product.

amplicon The precise, primer-to-primer, double-stranded nucleic acid product of a PCR or RT-PCR reaction.

Perhaps not surprisingly, there are many technical complexities underlying this deceptively simple description. Primer design must take into account issues such as DNA secondary structures, uniqueness of sequence, and similarity of T_m between primers. Use of a thermostable polymerase (i.e., one that is not inactivated by the high temperatures used in the denaturation steps) is an essential concept identified by Mullis and coworkers. Taq, the first thermostable polymerase to be purified (along with a number of subsequent related enzymes) was obtained from *Thermus aquaticus*, a prokaryote first found in extremely hot geothermal pools in Yellowstone National Park. Different sources have provided thermostable enzymes with differing properties (such as exonuclease activities for increased accuracy). Buffer composition (including agents such as dimethyl sulfoxide [DMSO] to help reduce secondary structural barriers to effective amplification and inclusion of divalent cations such as Mg^{2+} at sufficient concentration to not be depleted by chelation to nucleotides) often needs some optimization for effective reactions. In general, the PCR process works best when the primers are within short distances of each other (100–500 bp), but well-optimized reactions have been successful at distances into the tens of kilobases. "Hot start" techniques—frequently through covalent modification of the polymerase—can be employed to ensure that no inappropriate primer annealing and extension can occur prior to the first denaturation step, thereby avoiding the production of incorrect products. Generally, somewhere around 40 thermal cycles marks an effective limit for a PCR reaction with good kinetics in the presence of appropriate template, as depletion of dNTPs into amplicons effectively occurs around this point, and a "plateau phase" occurs wherein no more product is made. Conversely, if the appropriate template was not present in the reaction, proceeding beyond 40 cycles primarily increases the likelihood of production of rare, incorrect products.Pairing PCR with a preliminary reverse transcription step (either random-primed or using one of the PCR primers to direct activity of the RNA-dependent DNA polymerase [reverse transcriptase]) allows for RNA templates to be converted to complementary DNA (cDNA) and then subject

to regular PCR, in a variation known as **reverse transcription PCR (RT-PCR)**. In general, the subsequent discussion uses the term *PCR* to refer to both PCR and RT-PCR.

reverse transcription PCR (RT-PCR) A technique for the detection and quantification of expression of a gene by reverse transcription and amplification of RNAs.

Detection of PCR products can be done in a number of ways. Postreaction "endpoint techniques" include gel electrophoresis and DNA-specific dye staining. Long a staple of molecular biological techniques, this is a simple but effective technique to rapidly visualize both that an amplicon was produced and that it is of an expected size. If the particular application requires exact, to-the-nucleotide product sizing, capillary electrophoresis can be used instead. Hybridization of PCR products to microarrays or suspension bead arrays can be used to detect specific amplicons when more than one product sequence may come out of an assay. These, in turn, use a variety of methods for amplicon labeling, including chemiluminescence, fluorescence, and electrochemical techniques. Alternatively, **real-time PCR** methodologies employ some way of directly detecting the ongoing production of amplicons in the reaction vessel, most commonly through monitoring a direct or indirect fluorescence change linked to amplicon production by optical methods. These methods allow the reaction vessel to stay sealed throughout the process. In contrast to endpoint methods where final amplicon concentration bears little relationship to starting template concentration, real-time methods show good correlations between the thermocycle number at which clear signals are measurable (usually referred to as the **threshold cycle**, or C_T) and the starting template concentration. Thus, real-time methods are effective template quantification approaches. As a result, these methods are often referred to as **quantitative PCR (qPCR)** methods.

real-time PCR A PCR technique with continuous monitoring of product formation as synthesis proceeds, usually through fluorometric methods.

threshold cycle (C_T) The thermocycle number in a real-time PCR or RT-PCR reaction at which the product signal rises above a specified cutoff value to indicate amplicon production is occurring.

quantitative PCR (qPCR) A PCR reaction used to amplify and simultaneously quantify an amplicon.

Conceptually, the simplest method for real-time PCR detection is based on the use of dyes that selectively bind and become fluorescent in the presence of double-stranded DNA, such as SYBR green. Production of a PCR product during thermocycling leads to an exponential increase in the amount of double-stranded product present at the annealing and extension thermal steps of each cycle. The real-time instrument monitors fluorescence in each reaction tube during these thermal steps of each cycle and calculates the change in fluorescence per cycle to generate a sigmoidal amplification curve. A cutoff threshold value placed approximately midrange in the exponential phase of this curve is used for calculating the C_T of each sample and can be used for quantitation if appropriate controls are present.

A potential issue with this approach is that the reporter dyes are not sequence specific, so any spurious products produced by the reaction can lead to false-positive signals. In practice, this is usually controlled by performance of a melt-point analysis at the end of regular thermocycling. The reaction is cooled to the annealing temperature and then the temperature is slowly raised while fluorescence is constantly monitored. Specific amplicons will have a characteristic melt point at which fluorescence is lost, whereas nonspecific amplicons will demonstrate a broad range of melt points, giving a gradual loss in sample fluorescence.

A number of alternative approaches use probe-based fluorescence reporters, which avoid this potential nonspecific signal. Probe-based approaches work through the application of a process called **fluorescence resonant energy transfer (FRET)**. In simple terms, FRET occurs when two fluorophores are in close proximity and the emission wavelength of one (the reporter) matches the excitation wavelength of the other (the quencher). Photons emitted at the reporter dye-emission wavelength are effectively captured by the nearby quencher dye and reemitted at the quencher-emission

fluorescence resonant energy transfer (FRET) A process whereby the emission from an excited fluorophore is captured and reemitted at a longer wavelength by a nearby second fluorophore whose excitation spectrum matches the emission frequency of the first fluorophore.

wavelength. In the simplest form of this approach, two short oligonucleotide probes with homology to adjoining sequences within the expected amplicon are included in the assay reaction; one probe carries the reporter dye and the other, the quencher. If specific PCR product is formed in the reaction, then at each annealing step these two probes can anneal to the single-stranded product and thereby place the reporter and quencher molecules close to each other. Illumination of the reaction with the excitation wavelength of the reporter dye will lead to FRET and fluorescence at the quencher dye's characteristic emission frequency. By contrast, if the homologous template for the probe molecules is not present (i.e., the expected PCR product), the two dyes will not be colocalized and excitation of the reporter dye will lead to fluorescence at its emission frequency. This is illustrated in **FIGURE 2.20**. As with the DNA-binding dye approach, the real-time instrument monitors the quencher emission wavelength during each cycle and generates a similar sigmoidal amplification curve. Multiple alternative ways of exploiting FRET for this process exist, including 5′ fluorogenic nuclease assays, molecular beacons, and molecular scorpions. Although the details of these differ, the underlying concept is similar and all generate data in a similar fashion.

The applications of the PCR process are incredibly diverse. The simple appearance or nonappearance of an amplicon in a properly controlled reaction can be taken as evidence for the presence or absence, respectively, of the assay target template. This leads to medical applications such as the detection of infectious disease agents at sensitivities, specificities, and speeds much greater than alternative methods. Although the two primer sites must be of known sequence, the internal section may be any sequence of a general length; this fact leads directly to applications where a PCR product for a region known to vary between species (or even between individuals) can be produced and subject to sequence analysis to identify the species (or individual identity, in the latter case) of the sample template. Coupled with single-molecule sensitivity, this has provided criminal forensics with tools powerful enough to identify individuals from residual DNA on crime scene traces as simple as cigarette butts, smudged fingerprints, or a single hair. Evolutionary biologists have made use of PCR to amplify DNA from well-preserved samples, such as insects in amber millions of years old, with subsequent sequencing and phylogenetic analysis yielding fascinating results on the continuity and evolution of life on Earth. Quantitative real-time approaches have applications in medicine (e.g., monitoring viral loads in transplant patients), research (such as examining transcriptional activation of a specific target gene in a single cell), or environmental monitoring (for instance, water purification quality control).

In general, PCR reactions are run with carefully optimized T_m values that maximize sensitivity and amplification kinetics while ensuring that primers will anneal only to their exact hybridization matches. Lowering the T_m of a PCR reaction—in effect, relaxing the reaction stringency and allowing primers to anneal to not-quite-perfect hybridization partners—has useful applications as well, such as in searching a sample for an unknown sequence suspected to be similar to a known one. This technique has been successfully employed for the discovery of new virus species when primers matching a similar virus species are employed. Similarly, during a PCR-directed cloning of a gene or region of interest, use of planned mismatches in the primer sequence and slightly lowered T_ms can be used to introduce wanted mutations in a process called *site-directed mutagenesis*. Differential detection of single nucleotide polymorphisms (SNPs) (see the chapter titled *The Content of the Genome*), which can be directly indicative of particular genotypes or serve as surrogate linked markers for nearby genetic targets of interest, can be

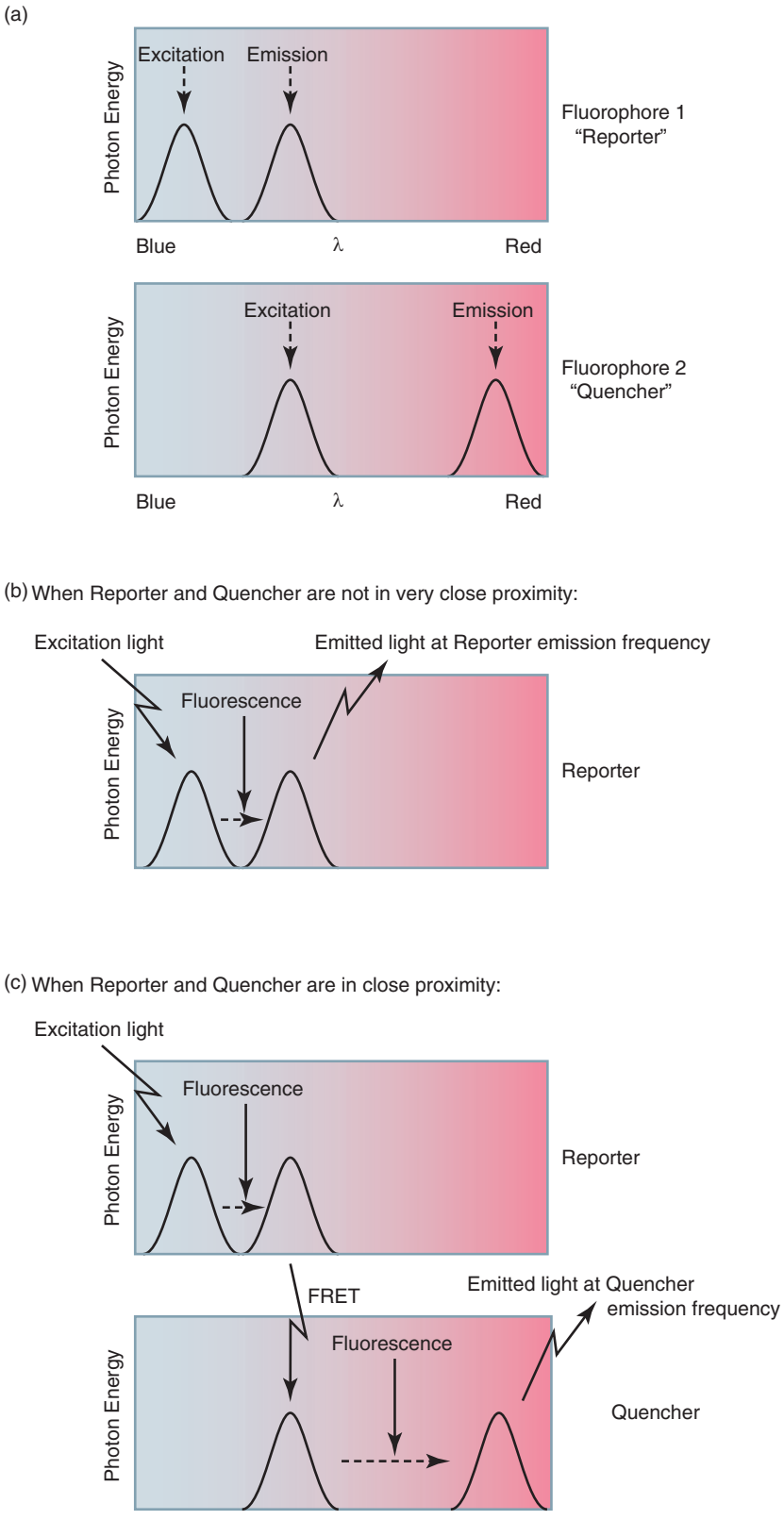


FIGURE 2.20 FRET occurs only when the reporter and quencher fluorophores are very close to each other, leading to the detection of light at the quencher emission frequency when the reporter is stimulated by light of its excitation frequency. If the reporter and quencher are not colocalized, stimulation of the reporter instead leads to detection of light at the reporter's emission frequency. By placing the reporter and quencher fluorophores on single-stranded nucleic acid probes complementary to the expected amplicon, different variations on this method can be designed such that the occurrence of FRET can be used to monitor the production of sequence-specific amplicons.

done through design of PCR primers with a 3′ terminal nucleotide specific to the expected polymorphism. At the optimal T_m, this final crucial nucleotide can hybridize and provide a 3′ hydroxyl to the waiting polymerase only if the matching SNP occurs in a process known by several names, including amplification refractory mutation selection (ARMS) or allele-specific PCR (ASPE).

The PCR process described thus far has been restricted to amplification of a single target per reaction, or "simplex" PCR. Although this is the most common application, it is possible to combine multiple, independent PCR reactions into a single reaction, allowing for an experiment to query a single minute specimen for the presence, absence, or possibly the amount of multiple unrelated sequences. This *multiplex PCR* is particularly useful in forensics applications and medical diagnostic situations but entails rapidly increasing levels of complexity in ensuring that multiple primer sets do not have unwanted interactions that lead to undesired false products. At best, multiplexing tends to result in loss of some sensitivity for each individual PCR due to effective competition between them for limited polymerase and nucleotides.

A final point of interest to many students with regard to PCR is its consideration from a philosophical perspective. In practice, performance of this now incredibly pervasive method requires the use of a thermostable polymerase. As noted earlier, these polymerases (of which there are a number of varieties) primarily derive from bacterial DNA polymerases originally identified in extremophiles living in boiling hot springs and deep-sea volcanic thermal vents. Few people would have been likely to suspect that studying deep-sea thermal vent microbes would be of such direct importance to so many other aspects of science, including ones with impact on their daily lives. These unexpected links between topics serve to highlight the importance of basic research on all manner of subjects; critical discoveries can come from the least expected avenues of research.

KEY CONCEPTS

- PCR permits the exponential amplification of a desired sequence, using primers that anneal to the sequence of interest.
- RT-PCR uses reverse transcriptase to convert RNA to DNA for use in a PCR reaction.
- Real-time, or quantitative, PCR detects the products of PCR amplification during their synthesis and is more sensitive and quantitative than conventional PCR.
- PCR depends on the use of thermostable DNA polymerases that can withstand multiple cycles of template denaturation.

CONCEPT AND REASONING CHECK

Why does PCR depend on a thermostable DNA polymerase to function?

2.9 Blotting Methods

After nucleic acids are separated by size in a gel matrix, they can be detected using dyes that are nonsequence specific, or specific sequences can be detected using a method generically referred to as *blotting*. Although slower and more involved than direct visualization by fluorescent dye staining, blotting techniques have two major advantages: they have a greatly increased sensitivity relative to dye staining, and they allow for the specific detection of defined sequences of interest among many similarly sized bands on a gel.

The method was first developed for application to DNA agarose gels and was introduced earlier in the chapter. The method is referred to as **Southern blotting** after its inventor, Dr. Edwin Southern. A schematic of this process is shown in **FIGURE 2.21**. A regular agarose gel is made and run (and, if desired, stained), as described previously. Following this, the gel is soaked in alkali buffer to denature the DNA and then placed in contact with a sheet of porous membrane (commonly nitrocellulose or nylon). A buffer is drawn through the gel and then the membrane, using capillary action or a gentle vacuum pressure. This slow flow of buffer, in turn, draws each nucleic acid band in the gel out of the gel matrix and onto the membrane surface. Nucleic acids bind to the membrane, which in many cases is positively charged to increase efficiency of DNA binding. This in effect creates a "contact print" of the order and position of all nucleic acid bands as resolved by size in the gel.

Southern blotting A process for the transfer of DNA bands separated by gel electrophoresis from the gel matrix to a solid support membrane for subsequent probing and detection.

Following transfer, the nucleic acids are permanently fixed to the membrane either through drying or through exposure to UV light, which can create physical cross-links between the membrane and the nucleic acids

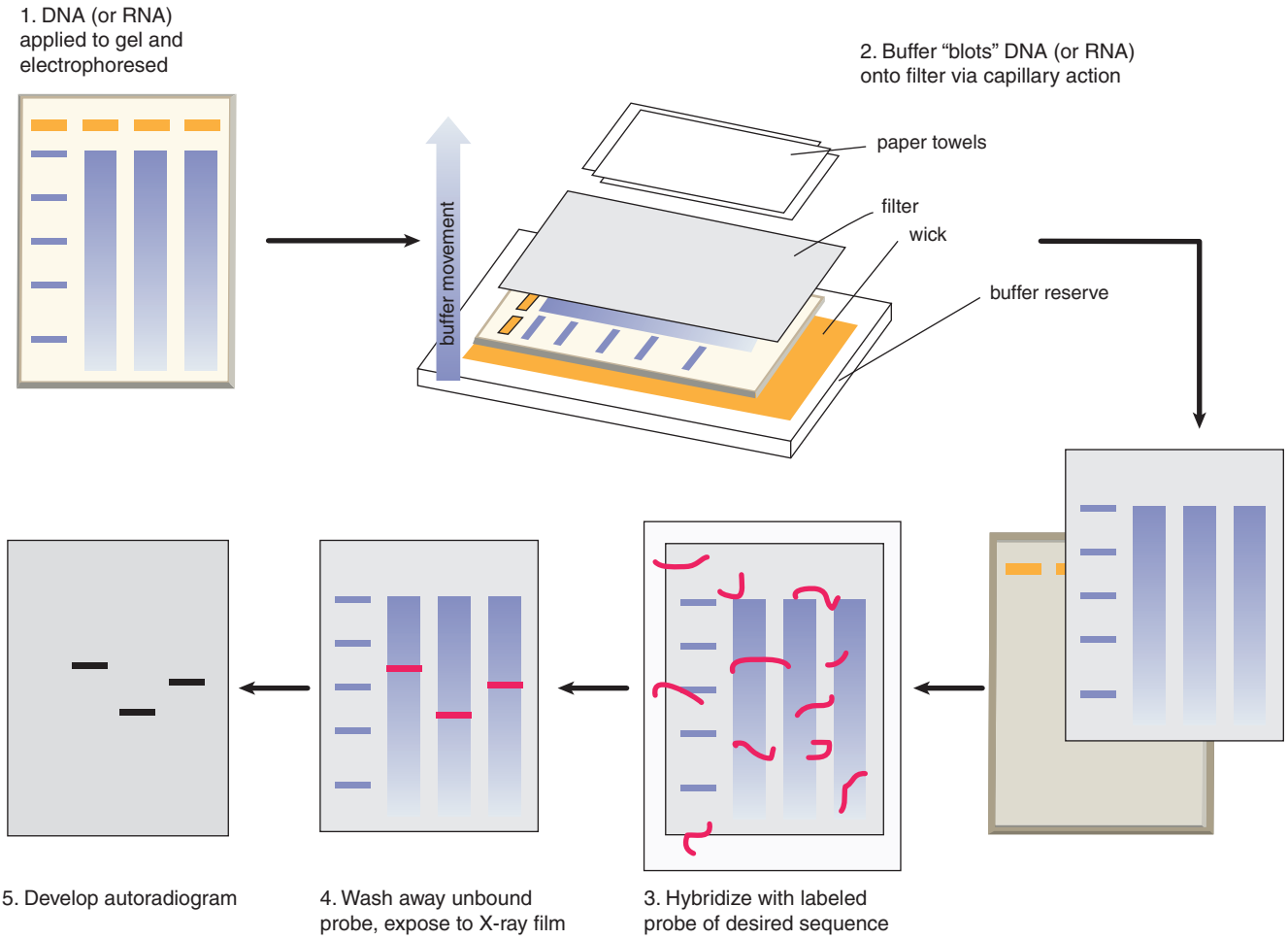


FIGURE 2.21 To perform a Southern blot, DNA digested with restriction enzymes is electrophoresed to separate fragments by size. Double-stranded DNA is denatured in an alkali solution either before or during blotting. In this example, the gel is placed on a wick (such as a sponge) in a container of transfer buffer and a membrane (nylon or nitrocellulose) is placed on top of the gel. Absorbent materials such as paper towels are placed on top. Buffer is drawn from the reservoir through the gel by capillary action, transferring the DNA to the membrane. Alternatively, an apparatus that uses a vacuum to pull the buffer through the gel and membrane is employed. The membrane is then incubated with a labeled probe (usually DNA). The unbound probe is washed away, and the bound probe is detected by autoradiography, phosphorimaging or other appropriate detection methods. In northern blotting, RNA, rather than DNA, is run on a gel.

(primarily pyrimidines). The blot is now ready for blocking, where it is immersed in a warmed, low-salt buffer containing materials that will bind to and block areas of the blot that may bind organic compounds nonspecifically. Then, a probe consisting of a labeled complement of the target sequence of interest is incubated with the blocked membrane. Following this hybridization step, the membrane is washed to remove nonspecifically associated probe molecules and then visualized, using whatever method (autoradiographic, chemiluminescent, etc.) is appropriate for the type of label used on the probe.

A benefit of Southern blotting is that it is a quantitative method.

Numerous variations on the Southern blot approach exist. In northern blotting, RNA molecules are separated by size, usually on formaldehyde or other denaturing gel, which eliminates RNA secondary structures. This allows measurement of actual RNA sizes and, like Southern blotting, provides a similarly quantitative method for detection of any type of RNA. If mRNA is the target of interest, it is possible to separate mRNA from other classes of RNA in the cell. mRNA (and some noncoding RNA) differs from other RNAs in that it is polyadenylated (it has a string of adenine residues added to the 3′ end) (see the *RNA Splicing and Processing* chapter). Poly(A)+ mRNA can be enriched by, for example, use of an oligo(dT) column, in which oligomers of oligo(dT) are immobilized on a solid support and used to capture mRNA from the total RNA in a sample. This is illustrated in **FIGURE 2.22**.

A conceptually similar process for proteins based on protein-separation gels and blotting to membrane is known as western blotting. This method is depicted in **FIGURE 2.23**. There are some key differences between the procedures for blotting proteins compared to nucleic acids. First, protein-separation gels typically contain a detergent such as sodium dodecyl sulfate (SDS), which both serves to unfold the proteins so that they will migrate according to size rather than shape and also provides a uniform negative charge to all proteins so that they will migrate toward the positive pole of the gel. (In the absence of SDS, each protein has a specific individual charge at a given pH; it is possible to separate proteins based on these charges, rather than size, in a technique called *isoelectric focusing*.)

Once the proteins are separated on the gel, they are transferred to a nitrocellulose membrane using an electric current to effect the transfer, rather than the capillary or vacuum methods used for nucleic acids. The most significant

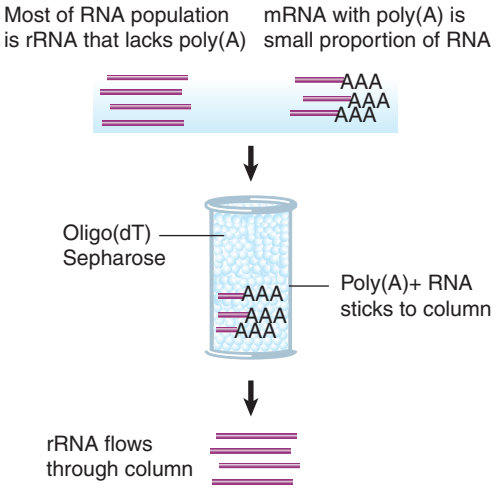


FIGURE 2.22 Poly(A)+ RNA can be separated from other RNAs by fractionation on an oligo(dT) column.

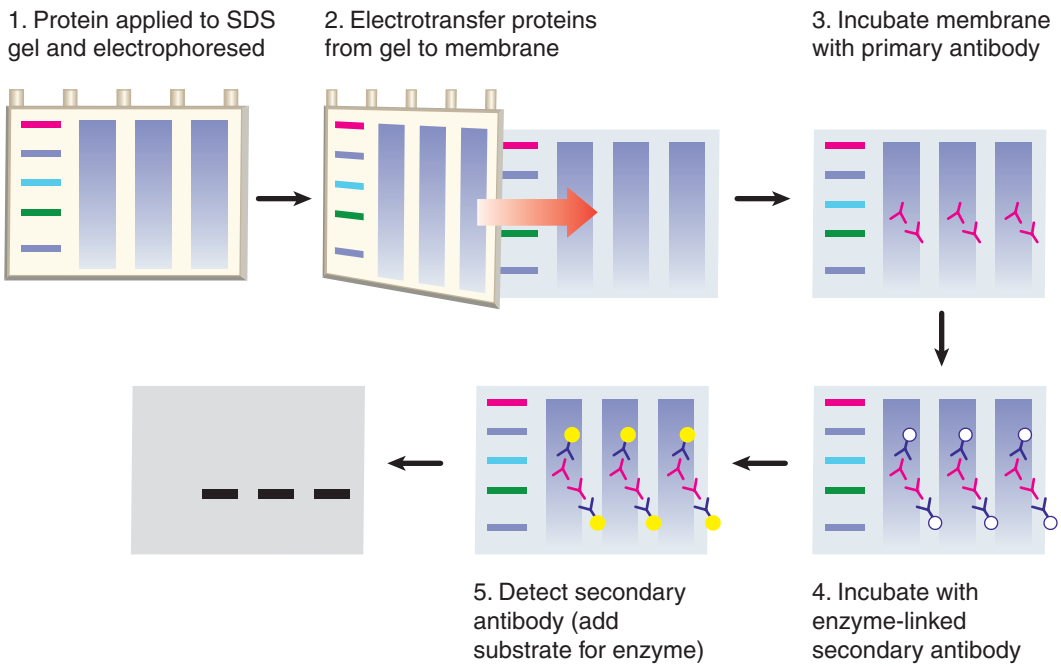


FIGURE 2.23 In a western blot, proteins are separated by size on an SDS gel, transferred to a nitrocellulose membrane, and detected using an antibody. The primary antibody detects the protein, and the enzyme-linked secondary antibody detects the primary antibody. The secondary antibody is detected in this example via addition of a chemiluminescent substrate, which results in emission of light that can be detected on X-ray film.

difference in Western blotting is the method of detecting proteins on the membrane. Complementary base pairing cannot be used to detect a protein, so western blots use antibodies to recognize the protein of interest. The antibody can either recognize the protein itself, if such an antibody is available, or can recognize an **epitope tag** that has been fused to the protein sequence. An epitope tag is a short peptide sequence that is recognized by a commercially available antibody; the DNA encoding the tag can be cloned in-frame to a gene of interest, resulting in a product containing the epitope (typically at the N- or C-terminus of the protein). Sequences for the most commonly used epitope tags (such as the HA, FLAG, and myc tags) are often available in expression vectors for ease of fusion.

epitope tag A polypeptide that has been added to a protein that allows its identification by an antibody.

The antibody that recognizes the target on the membrane is known as the *primary antibody*. The final stage of western blotting is detection of the primary antibody with a *secondary antibody*, which is the antibody that can be visualized. Secondary antibodies are raised in a different species than the primary antibody used and recognize the constant region of the primary antibody (e.g., a "goat anti-rabbit" antibody is raised in goats and will recognize a primary antibody raised in a rabbit) (see the *Homologous, Somatic, and Site-Specific Recombination* chapter for a review of antibody structure). The secondary antibody is typically linked to a moiety that allows its visualization—for example, a fluorescent dye or an enzyme such as alkaline phosphatase or horseradish peroxidase. These enzymes serve as visualization tools because they can convert added substrates to a colored product (*colorimetric detection*) or can release light as a reaction product (*chemiluminescent detection*). Use of primary and secondary antibodies (rather than linking a visualizer to the primary antibody) increases the sensitivity of western blotting. The result is semiquantitative detection of the protein of interest.

Continuing in the same vein of nomenclature, techniques used to identify interactions between DNA and proteins (through protein gel separation and blotting followed by probing with a DNA) are called Southwestern blotting. When an RNA probe is used, the technique is called orthwestern blotting.

KEY CONCEPTS

- Southern blotting involves the transfer of DNA from a gel to a membrane, followed by detection of specific sequences by hybridization with a labeled probe.
- Northern blotting is similar to Southern blotting but involves the transfer of RNA from a gel to a membrane.
- Western blotting entails separation of proteins on an SDS gel, transfer to a nitrocellulose membrane, and detection of proteins of interest using antibodies.

CONCEPT AND REASONING CHECK

Why does a Southern blot require blocking before a probe is added?

2.10 DNA Microarrays

The immediate technical progression from Southern and northern blotting was the development of the microarray. Instead of having the unknown sample on the membrane and the probe in solution, this effectively reverses the two. These originated in the form of "slot-blots" or "dot-blots," where a researcher would spot individual DNA sequences of interest directly onto a hybridization membrane, in an ordered pattern, with each spot consisting of a different, single-known sequence. Drying of the membrane immobilized these spots, creating a premade blotting array. In use, the researcher would then take a nucleic acid sample of interest, such as total cellular DNA, fragment it, and then uniformly label this DNA in a random, sequence-independent method. This labeled mix of sample DNA could then be used, exactly as in a Southern blot, as a probe to hybridize to the premade blot. Labeled DNA sequences homologous to any of the array spots would hybridize and be visualized in the known, fixed location of that spot(s). By knowing the physical location of each specific probe spot, the pattern of hybridized versus nonhybridized spots could be read out to indicate the presence or absence of each of the corresponding known sequences in the unknown sample.

Technological improvements to this approach followed rapidly through miniaturization of the size and physical density of the immobilized spots, going from membranes with 30–100 spots to glass microscope slides with up to 1,000 spots. Silicon chip substrates have up to a million or more individual spots in an area about the size of a postage stamp.

To visualize distinct spots in such a high-density array, automated optical microscopy is used to detect fluorescence and quantify each hybridization signal. In parallel with the increased total number of spots per array, the length of each unique probe became shorter, allowing for each spot in the array to be specific to a smaller target area—in effect, giving greater "resolution" on a molecular scale.

Microarrays have been used in numerous applications. One of the first and most common applications was in gene expression profiling, where a total mRNA sample from a specimen of interest (such as tissue in a disease state or under a particular environmental challenge) is collected and converted en masse to cDNA by reverse transcription, simultaneously incorporating a label. This labeled cDNA is then hybridized to an array where the immobilized spots consist of complementary strands to a number of known mRNAs from the target organism (if designed correctly, the sequences can allow the researcher to distinguish splicing variants). Hybridization, washing, and visualization allow for the detection of those spots that have bound their complementary labeled cDNA and thus the readout of those genes being expressed in the original sample. This process is depicted in **FIGURE 2.24**. The method is quantitative, meaning that the observed signal on each spot should correspond to the original level of its particular mRNA. In practice, critical quality controls and standardization, as well as sophisticated data analysis methods, are needed but not always included, with the result that there have been issues with reproducibility and validity of some published microarray data. More recently, second-generation sequencing of RNA (RNA-Seq) has become the primary method for this type of analysis. RNA-Seq still begins with cDNA library synthesis, but the cDNA is subjected to deep sequencing rather than being hybridized to a microarray.

A second major application is in genotyping. Analysis of many genomes has led to the identification of large numbers of **single nucleotide polymorphisms (SNPs)**, which are single nucleotide substitutions at a specific genetic locus (see the chapter titled *The Content of the Genome*). Individual SNPs occur at known frequencies, which often differ between populations. The most straightforward examples are where the SNP creates a missense mutation within a gene of interest, such as one involved in metabolism of a drug. People carrying one allele of the SNP may clear a drug from circulation at a very different rate from those with an alternative allele, and thus determination of a patient's allele at this SNP may be an important consideration in choosing an appropriate drug dosage. An example of this that has come all the way from theory into everyday use is *CYP450* SNP genotyping to determine appropriate dosage of the anticoagulant warfarin. Another is in SNP genotyping of the *K-Ras* oncogene in some types of cancer patients in order to determine whether epidermal growth factor receptor (EGFR)–inhibitory drugs will be of therapeutic value. Other (most) SNPs are of no direct biological consequence but can be valuable genetic markers if found to be closely associated to a particular allele of interest—that is, if in genetic terms it is closely linked. Millions of SNPs have been mapped in the human genome, and arrays containing up to nearly 2 million markers (about half of which are SNPs) are available that can be probed with a subject's DNA and allow for the genotype at each of these to be simultaneously determined, with concurrent determination of what the linked genetic alleles are. These arrays (as well as smaller targeted arrays for a variety of species, or for particular applications like pharmacogenomics) are still in common use, and a number of genotyping chips are approved by the Food and Drug Administration (FDA) for clinical use, like the *CYP450* chip described earlier. However, the improvements in cost, speed, and accuracy of sequencing-based genotyping may lead these methods to ultimately supplant microarrays in this area as well.

single nucleotide polymorphism (SNP) A polymorphism or variation in sequence between individuals caused by a change in a single nucleotide. This is responsible for much of the genetic variation between individuals.

The third major application of DNA microarrays has been in *array comparative genome hybridization* (*array-CGH*). This is a technique that

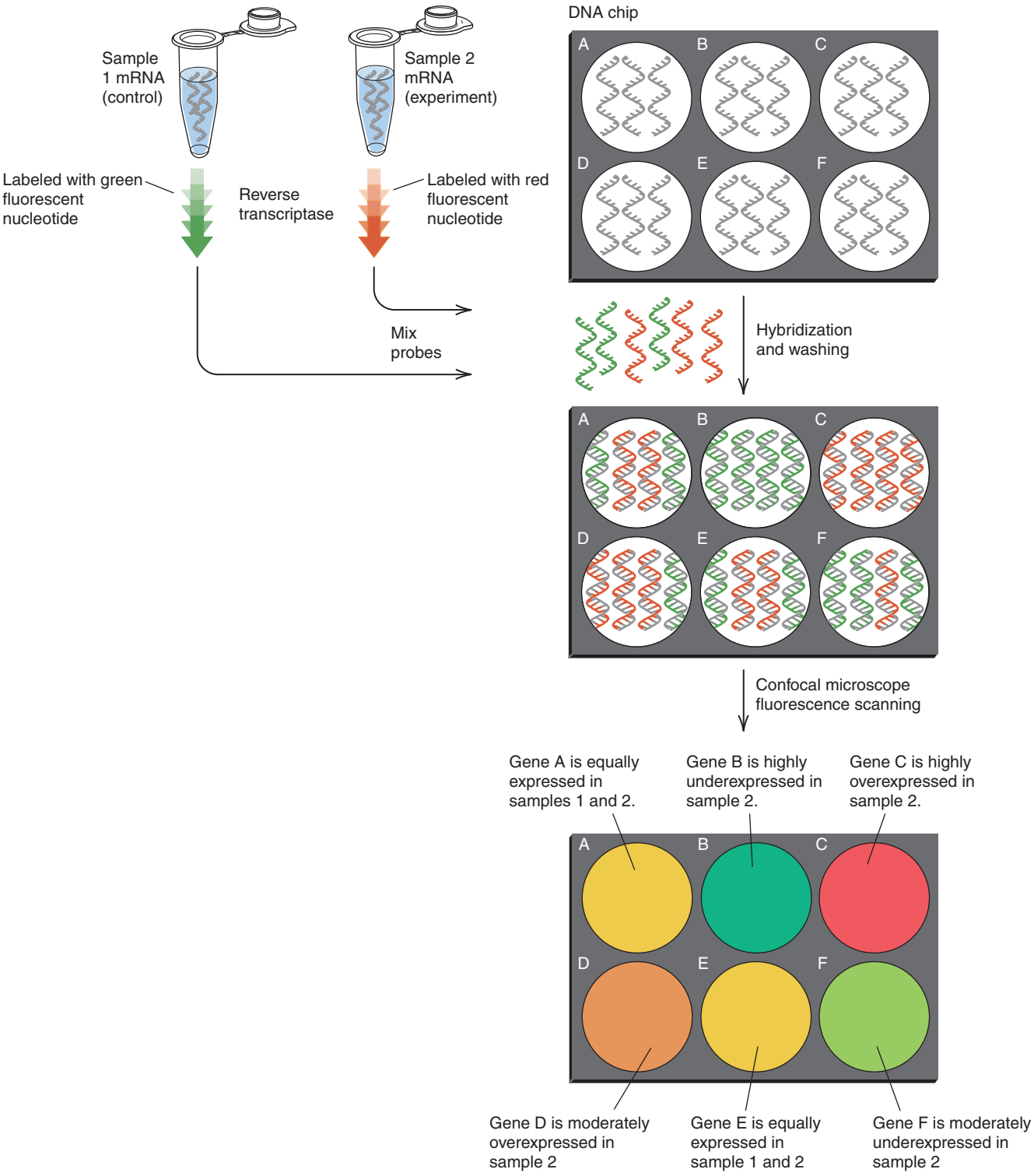


FIGURE 2.24 Gene expression arrays are used to detect the levels of all the expressed genes in an experimental sample. mRNAs are isolated from control and experimental cells or tissues and reverse transcribed in the presence of fluorescently labeled nucleotides (or primers), resulting in labeled cDNAs with different fluorophores (red and green strands) for each sample. Competitive hybridization of the red and green cDNAs to the microarray is proportional to the relative abundance of each mRNA in the two samples. The relative levels of red and green fluorescence are measured by microscopic scanning and displayed as a single color. Red or orange indicates increased expression in the red (experimental) sample, green or yellow-green indicates lower expression, and yellow indicates equal levels of expression in the control and experiment.

has been used for the detection and localization of chromosomal abnormalities that change the copy number of a given sequence—that is, deletions or duplications, known as copy number variants (CNVs). In this technique, the array chip (known as a **tiling array**) is spotted with an organism's genomic sequences that together represent the entire genome; the higher the density of the array, the smaller genetic region each spot represents and thus the higher resolution the assay can provide. Two DNA samples (one from normal control tissue and one from the tissue of interest) are each labeled with a different fluorophore, such that one sample, for example, is green and the other is red (similar to the mRNA labeling described earlier for the expression arrays). These two differentially labeled specimens are mixed at exactly equal ratios for total DNA and then hybridized to the chip. Regions of DNA that occur equally in the two samples will hybridize equally to their complementary array spots, giving a "mixed" color signal. By comparison, any DNA regions that occur more in one sample than the other will outcompete and thus show a stronger color on its complementary probe spot than the deficient sample will. Computer-assisted image analysis can read out and quantitate small color changes on each array spot and thus detect hemizygous loss of even very small regions in a test sample. This technique has been adopted in diagnostic settings for the detection of chromosomal copy number changes associated with a range of hereditary diseases. Once again, however, modern sequencing techniques have largely replaced microarrays for detecting CNVs in research settings, and are likely to be optimized for many different clinical/diagnostic applications as well.

tiling array An array of immobilized nucleic acid sequences, which together represent the entire genome of an organism. The shorter each array sequence is, the larger the total required number of spots is, but the greater the genetic resolution of the array.

In addition to the original chiplike solid-phase arrays, lower-density arrays for focused applications have been made in microbead-based formats. In these approaches, each microscopic bead has a distinct optical signal or code, and its surface can be coated with the target DNA sequence. Different bead codes can be mixed and matched into a single sample of labeled sample DNA or cDNA and then sorted, detected, and quantitated by optical and/or flow sorting methods. Although of much lower density than chip-type arrays, bead arrays can be modified and adapted much more readily to suit a particular focused biological question, and in practice show faster three-dimensional hybridization kinetics than chips, which effectively have two-dimensional kinetics.

KEY CONCEPTS

- DNA microarrays comprise known DNA sequences spotted or synthesized on a small chip.
- Development of DNA microarrays allowed the first genome-wide analyses in three key areas: transcription analysis with labeled cDNA from experimental samples, genome-wide genotyping of single nucleotide polymorphisms, and array-CGH to detect CNVs.
- Next-generation sequencing methods are replacing microarrays in many applications, but many microarrays remain in FDA-approved diagnostic or clinical use.

CONCEPT AND REASONING CHECK

How can microarray technology tell which genes are active in a given cell type? How would this differ from an RNA-Seq approach?

2.11 Chromatin Immunoprecipitation

Most of the methods discussed thus far in this chapter are *in vitro* methods that allow the detection or manipulation of nucleic acids or proteins that have been isolated from cells (or produced synthetically). Many other powerful molecular techniques have been developed, though, that allow either direct visualization of the *in vivo* behavior of macromolecules (such as imaging of GFP fusions in live cells) or that allow researchers to take a "snapshot" of the *in vivo* localization or interactions of macromolecules at a particular condition or point in time.

Throughout this book, we will discuss numerous proteins that function by interacting directly with DNA, such as chromatin proteins, or the factors that perform replication, repair, and transcription. While much of our understanding of these processes is derived from *in vitro* reconstitution experiments, it is critical to map the dynamics of protein–DNA interactions in living cells in order to fully understand these complex functions. The powerful technique of **chromatin immunoprecipitation (ChIP)** was developed to capture such interactions. *Chromatin* refers to the native state of eukaryotic DNA *in vivo*, in which it is packaged extensively with proteins (this is discussed in the *Chromatin* chapter). ChIP allows researchers to detect the presence of any protein of interest at a specific DNA sequence *in vivo*.

chromatin immunoprecipitation (ChIP) A method for detecting *in vivo* protein–DNA interactions that entails isolating proteins with an antibody and identifying DNA sequences that are associated with these proteins.

FIGURE 2.25 shows the process of ChIP. This method depends on the use of an antibody to detect the protein of interest. As was discussed earlier for western blots, this antibody can be against the protein itself or against an epitope-tagged target.

The first step in ChIP is typically the cross linking of the cell (or tissue or organism) of interest by fixing it with formaldehyde. This serves two purposes: (1) it kills the cell and arrests all ongoing processes at the time of fixation, providing the snapshot of cellular activity, and (2) it covalently links any protein and DNA that are in very close proximity, thus preserving protein–DNA interactions through the subsequent analysis. ChIP can be performed on cells or tissues under different experimental conditions (such as different phases of the cell cycle or after specific treatments) to look for changes in protein–DNA interactions under different conditions.

After cross linking, the chromatin is then isolated from the fixed material and cleaved into small chromatin fragments, usually 200–500 bp each. This can be achieved by sonication, which uses high-intensity sound waves to nonspecifically shear the chromatin. Nucleases (either sequence specific or nonsequence specific) can be used to fragment the DNA. These small chromatin fragments are then incubated with the antibody against the protein target of interest.

These antibodies can then be used to immunoprecipitate the protein by pulling the antibodies out of the solution using heavy beads coated with a protein (such as Protein A) that binds to the antibodies.

After washing away unbound material, the remaining material contains the protein of interest still cross linked to any DNA it was associated with *in vivo*. This is sometimes called a "guilt by association" assay because the DNA target is isolated only due to its interaction with the protein of interest. The final stages of ChIP entail reversal of the cross links so that the DNA can be purified and analyzed, most commonly by next-generation sequencing, to map the genome-wide binding ("ChIP-Seq").

In addition to revealing the presence of a specific protein at a given DNA sequence (such as a transcription factor bound to the promoter of a gene of interest), highly specialized antibodies can provide even more detailed

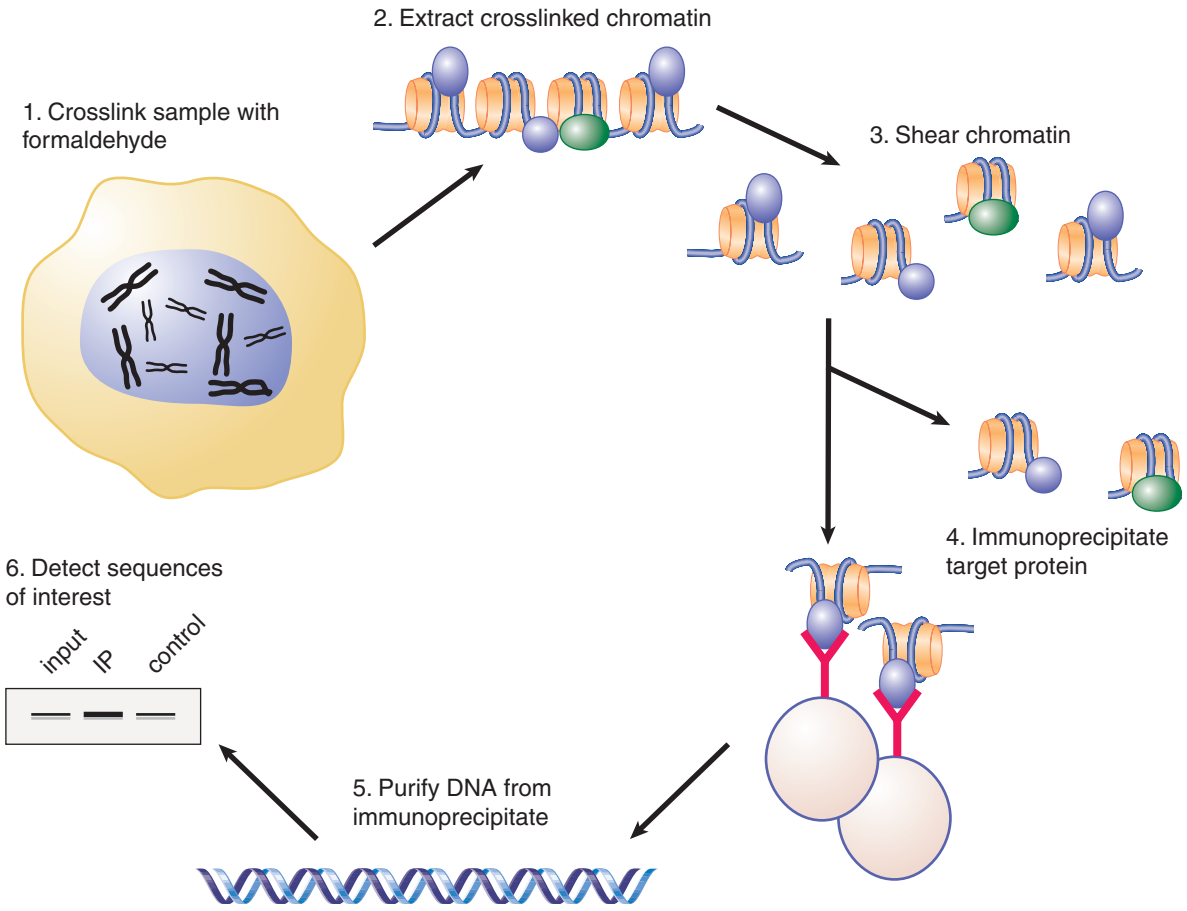


FIGURE 2.25 Chromatin immunoprecipitation detects protein–DNA interactions in the native chromatin context *in vivo*. Proteins and DNA are cross linked, chromatin is broken into small fragments, and an antibody is used to immunoprecipitate the protein of interest. Associated DNA is then purified and analyzed by one of several ways. If only a limited set of particular targets are of interest, they can be detected by PCR (as shown). For genome-wide analysis, the DNA is labeled and applied to a tiling array (ChIP on chip), or, most commonly today, the immunoprecipitated DNA is subjected to deep sequencing (ChIP-Seq).

information. For example, antibodies that distinguish between different post-translational modifications of the same protein can allow mapping of RNA polymerase II engaged in initiation versus polymerase II that has entered the elongation phase, because polymerase II is differentially phosphorylated in these two states (see the *Eukaryotic Transcription* chapter). ChIP has been used extensively to map nucleosome positions and to detect specific histone modifications across genomes as well.

KEY CONCEPTS

- ChIP allows detection of specific protein–DNA interactions *in vivo*.
- ChIP can be performed with antibodies that distinguish different covalently modified forms of target proteins.
- ChIP-Seq allows mapping of all the protein-binding sites for a given protein across the entire genome.

CONCEPT AND REASONING CHECK

Does ChIP of formaldehyde-treated cells identify only sequences that were in direct physical contact with the immunoprecipitated protein? Why or why not?

2.12 Gene Knockouts, Transgenics, and Genome Editing

transgenic An organism created by introducing exogenous DNA into the germline. The DNA may be inserted into the genome or exist in an extrachromosomal structure.

An organism that gains new genetic information from the addition of foreign DNA is described as **transgenic**. For simple organisms, such as bacteria or yeast, it has long been easy to generate transgenics by transformation with DNA constructs containing sequences of interest. Transgenesis in multicellular organisms has historically been much more challenging.

Transgenic mice were initially created by directly injecting DNA into eggs, as shown in **FIGURE 2.26**. In this method, plasmids carrying the gene of interest are injected into the nucleus of the oocyte or into the pronucleus of the fertilized egg. The egg is implanted into a pseudopregnant mouse (a mouse that has mated with a vasectomized male to trigger a receptive state). After birth, the recipient mouse can be examined to see whether it has gained the foreign DNA and, if so, whether it is expressed. This low-efficiency method typically results in about 15% of the implanted mice carrying the transfected sequence, usually with 1–150 tandem copies of the plasmid integrated into a single chromosomal site. The levels of gene expression from *transgenes* introduced in this way is highly variable, both due to copy number and the site of integration. A gene may be highly expressed if it integrates within an active chromatin domain, but not if it integrates in or near a silenced region of the chromosome.

knockout A genetically engineered organism that has had a gene disabled by a targeted mutation.

Subsequently, these transgenesis methods were elaborated to more versatile approaches for studying the functions of genes. Transgenesis allowed DNA to be added to cells or animals, but in order to understand the function of a gene, it is more useful to be able to remove or alter the gene or its function and observe the resulting phenotype. Changes that eliminate gene function (including complete gene deletion) are usually referred to as **knockouts**, whereas replacement of a gene with an alternative version is called a **knock-in**.

knock-in A genetically engineered organism that has had a gene sequence replaced by a different sequence.

In yeast, producing knockouts is a simple process in which DNA encoding a selectable marker flanked by short regions of homology to a

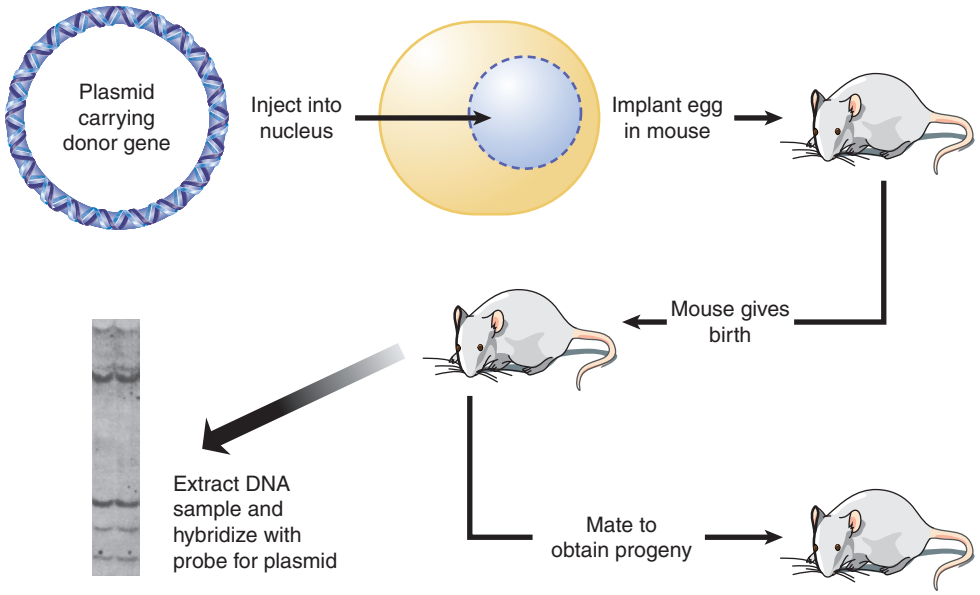


FIGURE 2.26 Transfection can introduce DNA directly into the germline of animals.

Photo reproduced from Chambon, P. (1981). *Sci. Am.* **244**, 60–71. Used with permission of Pierre Chambon, Institute of Genetics and Molecular and Cellular Biology, College of France.

target gene is transformed into the yeast. As little as 40 bp or so of homology will result in extremely efficient replacement of the target gene by the introduced marker gene via homologous recombination using the short regions of homology, and this has been used to systematically knock out all genes in yeast.

In many organisms and cultured cells, for a long time there was no good method for deleting endogenous genes. As an alternative, researchers developed a variety of **knockdown** approaches, which reduce the amount of a gene product (RNA or protein) produced, even while the endogenous gene is intact. There are a number of different knockdown methods; the one used most extensively in mammalian tissue culture is the use of RNA interference (RNAi) to selectively target specific mRNAs for destruction. (RNAi is described in the *Regulatory RNA* chapter.) Briefly, introduction of double-stranded RNA into most eukaryotic cells triggers a response in which these RNAs are cleaved by a nuclease called Dicer into 21-bp dsRNA fragments, unwound into single strands, and then used by another enzyme, RISC (RNA-induced silencing complex), to find and anneal to mRNAs containing complementary sequence. When a complementary mRNA is found, it is either cleaved and destroyed or its translation is blocked. In practice, this means that the mRNA for a given gene can be targeted for silencing by introduction of a dsRNA designed to anneal to the target of interest.

knockdown A genetically engineered organism that has had a gene downregulated by introducing a silencing vector to reduce the expression (usually translation) of the gene.

In some multicellular organisms, direct gene deletion is possible, but the process is more complicated than in organisms like yeast. A big advance came when it became possible to delete or replace genes by homologous recombination in mice, as had long been possible in yeast. In mammals, the target is usually the genome of an embryonic stem (ES) cell, which is then used to generate a mouse with the knockout. ES cells are derived from the mouse blastocyst (an early stage of development, which precedes implantation of the egg in the uterus). **FIGURE 2.27** illustrates the general approach.

ES cells are transfected with DNA containing homology to the target gene of interest that also carries a selectable marker (like drug resistance) so it is possible to select ES cells that have obtained an integrated transgene. These selected ES cells are then injected into a recipient blastocyst; at this stage, introduced ES cells are able to participate in normal development of the blastocyst. The blastocyst is implanted into a foster mother and in due course develops into a *chimeric* mouse. Some of the tissues of the chimeric mice are derived from the cells of the recipient blastocyst; other tissues are derived from the injected ES cells. The proportions of tissues in the adult mouse that are derived from cells in the recipient blastocyst and from injected ES cells vary widely in individual progeny; a visible marker such as coat-color gene is used so that areas of tissue representing each type of cell can be easily seen. In Figure 2.27, the ES cells were obtained from a donor mouse with dark fur, while the foster mother has light fur.

To determine whether the ES cells contributed to the germline, the chimeric mouse is crossed with a mouse that lacks the donor trait (in this example, dark fur). Any progeny that have the trait (all dark fur) must be derived from germ cells that have descended from the injected ES cells.

When donor DNA is introduced into the ES cell, it can insert into the genome by either nonhomologous or homologous recombination. Homologous recombination is relatively rare, probably representing less than 1% of all recombination events and thus occurring at a frequency of about 10^{-7}. Thus, it is crucial to use selective techniques to identify those ES cells in which homologous recombination has occurred.

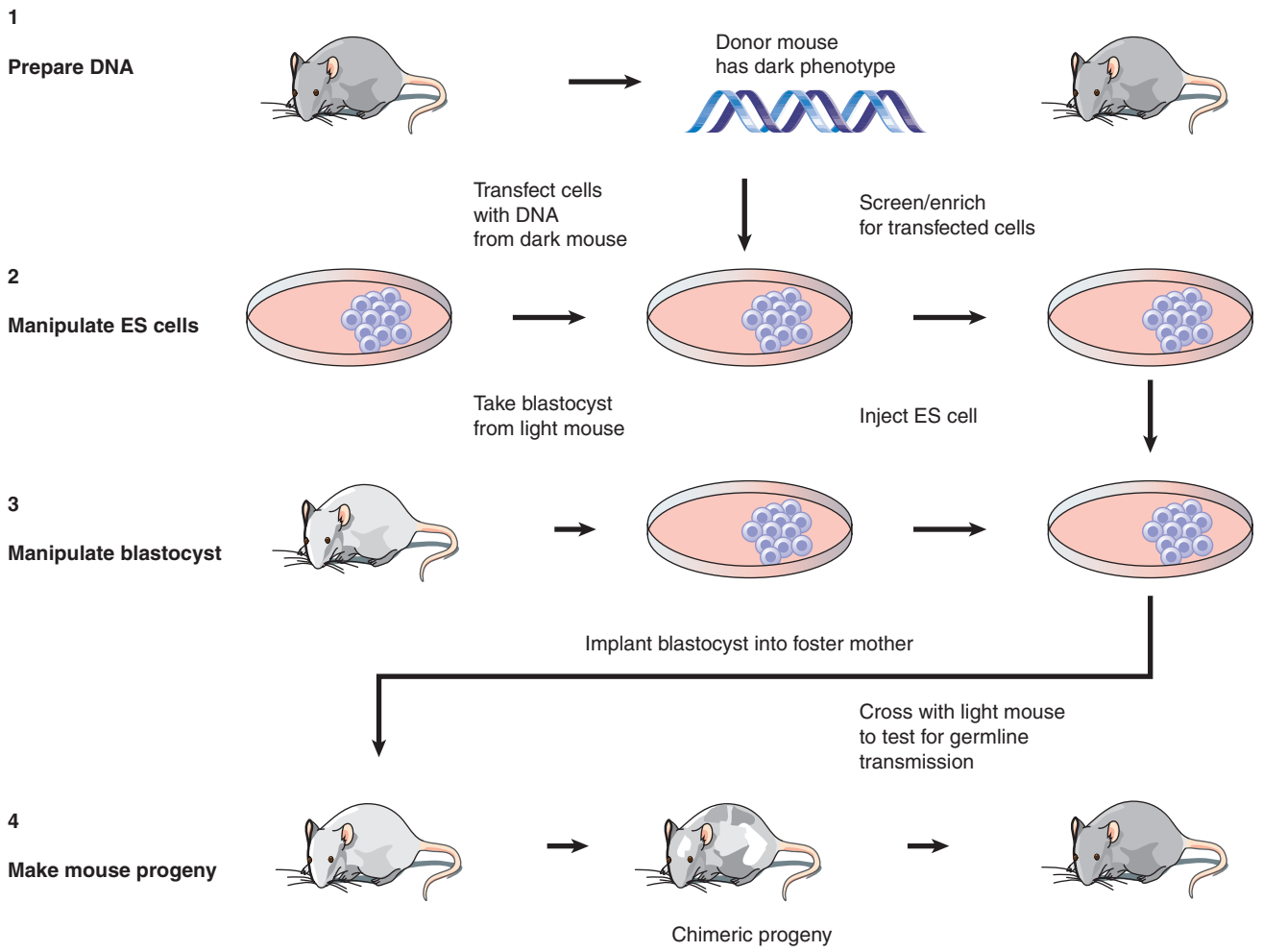


FIGURE 2.27 ES cells can be used to generate mouse chimeras, which breed true for the transfected DNA when the ES cell contributes to the germline.

FIGURE 2.28 illustrates the use of a basic knockout construct to disrupt endogenous genes. The construct contains two different markers that are designed to allow nonhomologous and homologous recombination events in the ES cells to be distinguished. The donor DNA is homologous to a target gene but has two key modifications. First, the gene is inactivated by interrupting or replacing an exon with a gene encoding a selectable marker (such as the *neo*^{R} gene that confers resistance to the drug G418). Second, a *counterselectable* marker (a gene that can be selected *against*) is added on one side of the gene outside the region of homology. In this example, the herpes virus *TK* gene is used.

When this knockout construct is introduced into an ES cell, homologous and nonhomologous recombinations will result in different outcomes. Nonhomologous recombination inserts the entire construct, including the flanking *TK* gene. These cells are resistant to neomycin, and they also express thymidine kinase, which makes them *sensitive* to the drug ganciclovir (thymidine kinase phosphorylates ganciclovir, which converts it to a toxic compound). In contrast, homologous recombination involves two exchanges within the sequence of the donor gene, resulting in the loss of the flanking *TK* gene. Cells in which homologous recombination has occurred therefore gain neomycin resistance in the same way as cells that have nonhomologous recombination, but they do *not* have thymidine kinase activity, and so are

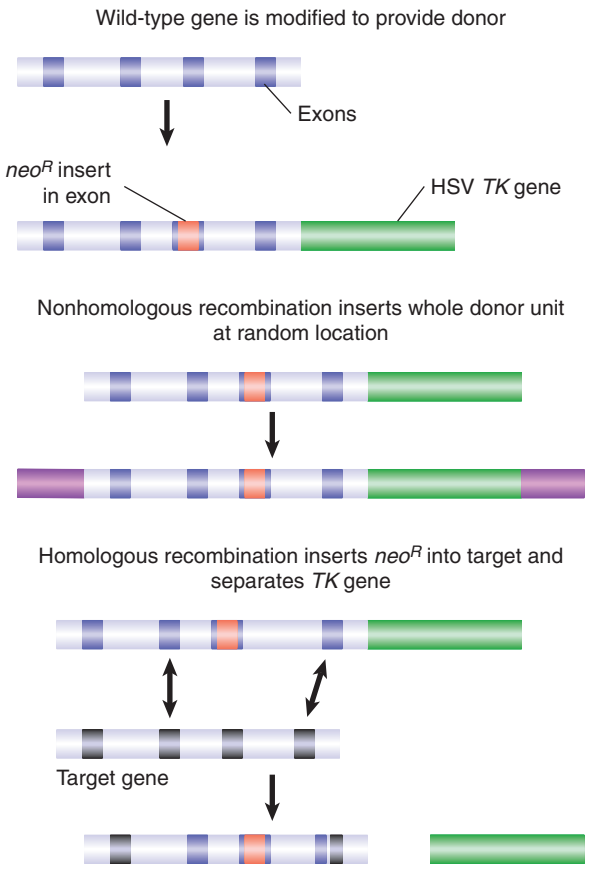


FIGURE 2.28 A transgene containing *neo*^{R} within an exon and *TK* downstream can be selected by resistance to G418 and loss of *TK* activity.

resistant to ganciclovir. Thus, plating the cells in the presence of neomycin plus ganciclovir specifically selects those in which homologous recombination has replaced the endogenous gene with the donor gene.

The presence of the *neo*^{R} gene in an exon of the donor gene disrupts gene expression and thereby creates a null allele. A particular target gene can therefore be knocked out by this means; once a mouse with one null allele has been obtained, it can be bred to generate the homozygote. This technique has long been used to study gene function in the mouse.

A major extension of the ability to manipulate a target genome was made possible by using the phage Cre/*lox* system to engineer site-specific recombination in a eukaryotic cell. The Cre enzyme catalyzes a site-specific recombination reaction between two *lox* sites, which are identical 34-bp sequences (see the *Homologous, Somatic, and Site-Specific Recombination* chapter). **FIGURE 2.29** shows that the consequence of the reaction is to excise the stretch of DNA between the two *lox* sites.

The Cre/*lox* system requires no additional components: it works when the Cre enzyme is produced in any cell that has a pair of *lox* sites. **FIGURE 2.30** shows that by placing the *cre* gene under the control of a regulated promoter, excision can be targeted to a particular cell type or induced by specific conditions. The procedure starts with two mice; one has the *cre* gene controlled by a promoter that is activated only under certain conditions (such as tissue-specific expression or drug-inducible). The other mouse has a target sequence flanked by *lox* sites. Progeny of the cross of these two mice have both elements of the system, and the system is turned on by the regulated

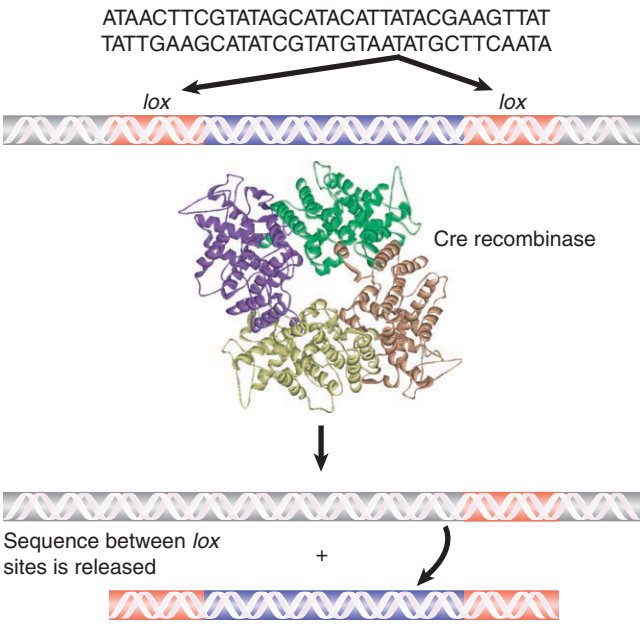


FIGURE 2.29 The Cre recombinase catalyzes a site-specific recombination between two identical *lox* sites, releasing the DNA between them.

Structure from Ennifar, E. et al. (2003). Protein Data Bank: 1OUQ. *Nucleic Acids Res*. **31**, 5449–5460.

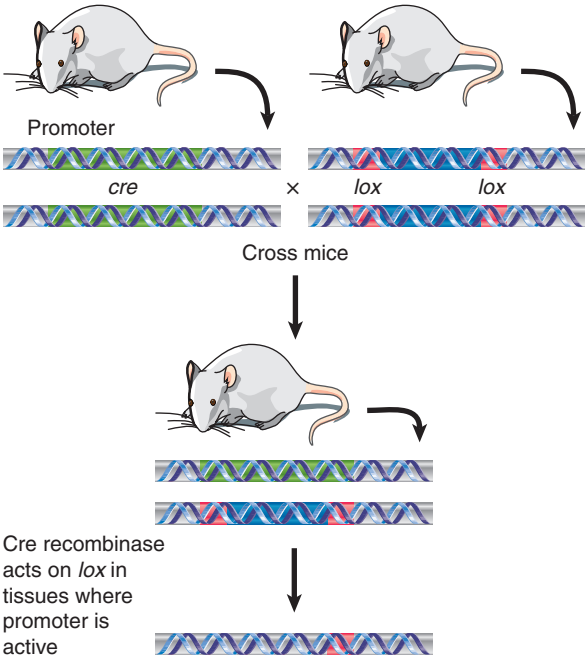


FIGURE 2.30 By placing the gene encoding the Cre recombinase under the control of a regulated promoter, it is possible to activate the excision system only in specific cells or in response to specific inducers. One mouse with a promoter-*cre* construct is crossed with another that has a target sequence flanked by *lox* sites. The progeny have both constructs, and excision of the target sequence can be triggered by activating the promoter.

activation of the *cre* gene. This allows the sequence between the *lox* sites to be excised in a controlled way.

The Cre/*lox* system has been used to create *reversible knockouts* by flanking the *neo*^{R} gene (or other selectable marker) in a knockout construct with *lox* sites. After the knockout has been made, the target gene can be

reactivated by using Cre to excise the *neo*^{R} gene in some particular circumstance (such as in a specific tissue).

FIGURE 2.31 shows a modification of this procedure that allows a knock-in to be created. In this case, the construct contains an altered version of the target gene that is used to replace the endogenous gene in the usual way. Then, when the inserted gene is reactivated by excising the *neo*^{R} sequence, we have in effect replaced the original gene with a different version.

A useful variant of this method is to introduce a wild-type copy of the gene of interest in which the gene itself (or one of its exons) is flanked by *lox* sites. This results in a phenotypically normal animal that can be crossed to a mouse containing *cre* under control of a tissue-specific or otherwise regulated promoter. The offspring of this cross are *conditional knockouts*, in which the function of the gene is lost only in cells that express the Cre recombinase. This is particularly useful for studying essential genes, such as those required for embryonic development; genes in this class would be lethal in homozygous embryos and thus are very difficult to study. We are now in the midst of a new era of genome manipulation resulting from technologies that allow direct editing of target sequences in the genome *in vivo*. Several methods have been developed based on endonucleases that can be targeted very specifically to genomic sites. The double-strand breaks created by these nucleases then utilize the cell's own repair machinery (homologous recombination or nonhomologous end-joining) (see the *Repair Systems* chapter) to generate sequence alterations. These changes can include gene mutation, deletion, insertion, or even precise gene editing or correction based on a provided donor template.

The specificity and outcomes of these techniques depend on the specific targeting of endonucleases to only the site(s) of interest. Four general classes of nucleases have been used:

- *Zinc finger nucleases (ZFNs)* take advantage of the fact that zinc finger DNA binding domains (discussed in the *Eukaryotic Transcription*

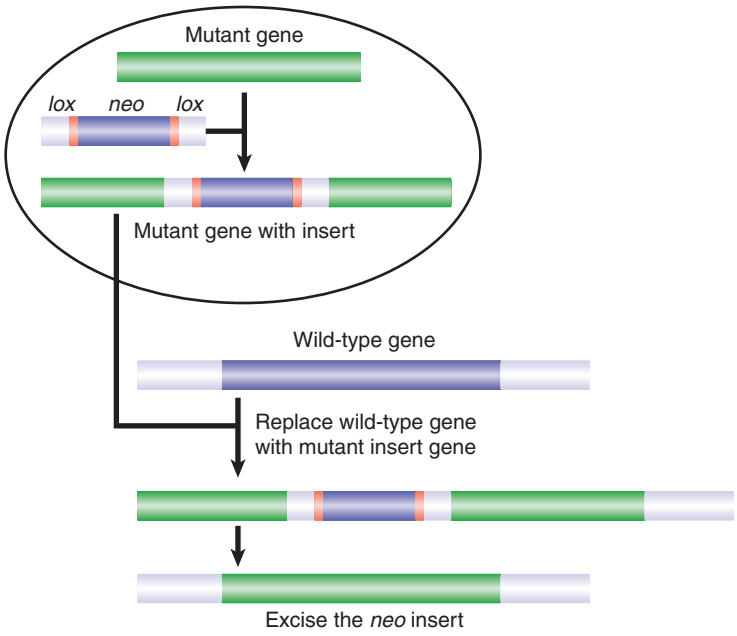


FIGURE 2.31 An endogenous gene is replaced in the same way as when a knockout is made (see Figure 2.28), but the *neo*^{R} gene is flanked by *lox* sites. After gene replacement, the *neo*^{R} gene can be removed by expressing Cre recombinase, leaving an active insert.

Regulation chapter) are modular domains that each recognize a 3-bp sequence and can be strung together into multifinger domains to recognize longer sequences. The ZF portion is fused to the endonuclease domain of the FokI restriction enzyme to create the ZFN, which then dimerizes to make a double-strand break (DSB) at the desired site.

- *Transcription activator-like effector nucleases (TALENs)* utilize a modular 33–35 amino acid DNA binding repeat derived from the TALE proteins of *Xanthomonas* bacterial plant pathogens. Each TALE repeat recognizes a single base pair (determined by two variable amino acids within the 33–35 amino acid repeat), so multiple TALE repeats can be strung together to recognize virtually any sequence (with the only requirement that there be a T at the 5′ end of the target). The TALE array is fused to FokI to provide the cleavage.
- *Meganucleases* (despite their name, the smallest of the editing nucleases and thus the easiest to deliver) are derived from naturally occurring "homing endonucleases," a family of nucleases encoded within introns or as self-splicing inteins. They naturally recognize long, usually asymmetric, sites of up to 40 bp that occur only 1–2 times in a genome. (The large target sites are the origin of the name.) Meganucleases can be engineered or selected to recognize novel sequences, but because they lack the modular nature of ZFNs and TALENs, this is often challenging.
- *Prokaryotic clustered regularly interspaced short palindromic repeats (CRISPR)* and *CRISPR-associated (Cas)* nucleases are the most recent—and most exciting—gene-editing tool to be developed. They are based on the CRISPR-Cas RNA-guided nucleases that form the basis of a bacterial adaptive immune response against viruses and plasmids; the Cas9 nuclease is the most used; thus, the system is generally known as CRISPR-Cas9. The native CRISPR-Cas system is described in detail in the *Regulatory RNA* chapter.

Briefly, the native system involves integration of invading nucleic acids into CRISPR loci, where they are transcribed into CRISPR RNAs (crRNAs). These then form a complex with a *trans*-activating crRNA and Cas proteins. The crRNA then targets cleavage of complementary DNA sequences. To adapt this system for gene editing, the two RNAs are fused into a single-guide RNA (sgRNA), and changes to a portion of this sequence can be used to define desired targets. This is an enormous advantage over the other technologies, which must engineer novel proteins for every desired target sequence. The same Cas9 protein can simply be delivered with a gRNA (or several!) designed against the site of interest. Cas9 proteins do strictly require a short (3 bp) *protospacer-adjacent motif* (PAM) 3′ to the target site, which can limit some target sequences. Some efforts have focused on developing Cas9 proteins with different PAM specificities to expand this repertoire, as well as developing Cas9 variants with increased specificity to reduce off-target cleavage.

Genome editing with CRISPR-Cas9 methods has been applied in many species ranging from prokaryotes, to fungi, to complex plants and animals. In addition, researchers are continually expanding the repertoire of applications for CRISPR-Cas9; for example, nuclease-deactivated Cas9 (dCas9) can be fused to transcriptional activators or repressors, or to epigenetic modifiers, in order to specifically target regulatory proteins anywhere in the genome to effect regulation without concomitant DNA cleavage. dCas9 (as well as ZFs and TALEs) has also been fused to fluorescent proteins for live-cell imaging of genomic regions.

With these kinds of techniques, we are able to investigate the functions and regulatory features of genes in whole animals. Gene-editing techniques have already begun to show promise as a gene therapy tool to treat human genetic disorders and other diseases, though at this time the technology is outpacing the ability to thoughtfully address all of the safety and ethical issues surrounding human genome editing—especially in the germline. Currently, several clinical trials testing the safety and efficacy of CRISPR-based approaches have begun, including trials to modify the CCR5 receptor (used by HIV to enter cells) in HIV-infected patients, and use of CRISPR-edited autologous T cells in patients with certain cancers. Someday, we may expect routinely to replace or repair defective genes in the genome in a targeted manner, but for now the risks and ethical concerns still temper the potential benefits of editing the human genome.

KEY CONCEPTS

- ES cells that are injected into a mouse blastocyst generate descendant cells that become part of a chimeric adult mouse.
- When the ES cells contribute to the germline, the next generation of mice may be derived from the ES cell.
- Genes can be added to the mouse germline by transfecting them into ES cells before the cells are added to the blastocyst.
- An endogenous gene can be replaced by a transfected gene using homologous recombination.
- The occurrence of successful homologous recombination can be detected by using two selectable markers, one of which is incorporated with the integrated gene, the other of which is lost when recombination occurs.
- The Cre/*lox* system is widely used to make conditional knockouts and knock-ins.
- New tools like the CRISPR/Cas9 system allow direct genome editing in a variety of living cells.

CONCEPT AND REASONING CHECK

What is the advantage of being able to create a conditional knockout in a selected tissue or at a particular time in development?

2.13 Summary

DNA can be manipulated and propagated using the techniques of cloning. These include digestion by restriction endonucleases, which cut DNA at specific sequences, and insertion into cloning vectors, which permit DNA to be maintained and amplified in host cells such as bacteria. Cloning vectors can have specialized functions as well, such as allowing expression of the product of a gene of interest or fusion of a promoter of interest to an easily assayed reporter gene.

DNA (and RNA) can be detected nonspecifically by the use of dyes that bind independently of sequence. Specific nucleic acid sequences can be detected using base complementarity. Specific primers can be used to detect and amplify particular DNA targets via PCR. RNA can be reverse transcribed into DNA to be used in PCR; this is known as RT-PCR. Labeled probes can be used to detect DNA or RNA on Southern or northern blots, respectively. Proteins are detected on western blots using antibodies.

DNA microarrays are solid supports (usually silicon chips or glass slides) on which DNA sequences corresponding to ORFs, or complete genomic sequences are arrayed. Microarrays have been used to detect gene expression, for SNP genotyping, and to detect changes in DNA copy number, as well as

many other applications. Many of these applications are now being accomplished using modern sequencing technology.

Sequencing technology is advancing rapidly. The original cost to determine the human genome sequence was ~$1 billion. By the beginning of 2012, multiple individuals had their sequence determined. The original goal of the second-generation sequencing methodologies was a $1,000 genome, a target that is now here. Third-generation technologies are in development.

Protein–DNA interactions can be detected *in vivo* using ChIP. The DNA obtained in a ChIP experiment can be sequenced to map all localization sites for a given protein in the genome.

New sequences of DNA may be introduced into a cultured cell by transfection or into an animal egg by microinjection. The foreign sequences may become integrated into the genome, often as large tandem arrays. The array appears to be inherited as a unit in a cultured cell. The sites of integration appear to be random. A transgenic animal arises when the integration event occurs into a genome that enters the germ cell lineage. Often a transgene responds to tissue and temporal regulation in a manner that resembles the endogenous gene. Under conditions that promote homologous recombination, an inactive sequence can be used to replace a functional gene, thus creating a knockout, or deletion, of the target locus. Extensions of this technique can be used to make conditional knockouts, where the activity of the gene can be turned on or off (such as by Cre-dependent recombination), and knock-ins, where a donor gene specifically replaces a target gene. Transgenic mice can be obtained by injecting recipient blastocysts with ES cells that carry transfected DNA. Knockdowns, mostly commonly achieved using RNA interference, have been used to eliminate gene products in cell types for which knockout technologies were not available. New genome editing technologies based on targeted endonucleases have dramatically expanded our capacity to make changes to genomes *in vivo*.

Chapter Questions

1. Restriction enzymes that generate sticky ends:
   A. make a single cut in the DNA backbone.
   B. make two cuts in the DNA backbone in the same strand.
   C. make two staggered cuts in the DNA backbone.
   D. make two cuts in the DNA backbone directly across the double helix from each other.

2. Recombinant plasmids are usually introduced into bacterial cells by:
   A. gene gun.
   B. transformation of chemically treated cells.
   C. transfection using liposomes.
   D. microinjection.

3. The best vector for cloning very large (>1 Mb) fragments of DNA is:
   A. plasmid.
   B. bacteriophage.
   C. cosmid.
   D. YAC.

4. Reporter genes are:
   A. genes that produce products that are easy to detect and/or quantify.
   B. genes that are essential in all species.
   C. genes that never occur in nature.
   D. all of the above.

5. A knockout construct does not usually contain:
   A. regions homologous to an endogenous gene.
   B. a selectable marker.
   C. a marker for counterselection.
   D. a reporter gene.
6. Put the following events in the generation of a knockout mouse in the correct order:
   A. Breed heterozygotes to generate homozygous null offspring.
   B. Inject knockout construct into ES cells.
   C. Breed chimeric offspring to obtain heterozygotes from ES-derived germ cells.
   D. Implant blastocyst into foster mother and obtain offspring.
   E. Make knockout construct to target gene of interest.
   F. Introduce ES cells to blastocyst.
   G. Select for ES cells that have integrated knockout construct via homologous recombination.

Key Terms

amplicon
autoradiography
chromatin immunoprecipitation (ChIP)
cloning
cloning vector
complementary
cosmid
dideoxynucleotide (ddNTP)
endonuclease
epitope tag
exonuclease
expression vector
fluorescence resonant energy transfer (FRET)
hybridization
insert
knockdown
knock-in
knockout
ligate
multiple cloning site
nuclease
phosphatase
polymerase chain reaction (PCR)
primer
probe
quantitative PCR (qPCR)
real-time PCR
recombinant DNA
reporter gene
restriction endonuclease
reverse transcription PCR (RT-PCR)
shuttle vector
single nucleotide polymorphism (SNP)
Southern blotting
stringency
subclone
threshold cycle (C_T)
tiling array
T_m
transformation
transgene
vector
yeast artificial chromosome (YAC)

Further Reading

Hrdlickova, R., Toloue, M., and Tian, B. (2017). RNA-seq methods for transcriptome analysis. *WIREs RNA* **8**, e1364. doi:10.1002/wrna.1364.

Jiang, F., and Doudna, J. A. (2017). CRISPR–Cas9 structures and mechanisms. *Annu. Rev. Biophys.* **46**(1), 505–529.

Marzancola, M. G., Sedighi, A., and Li, P. C. H. (2016). DNA microarray-based diagnostics. In: Li, P., Sedighi, A., and Wang, L. (eds). *Microarray Technology. Methods in Molecular Biology*, vol. **1368**. Humana Press, New York, NY.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterson, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353.

Wang, H., La Russa, M., and Qi, L. S. (2016). CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.* **85**(1), 227–264.