CHAPTER 2

DNA Structure and Genetic Variation

CHAPTER OUTLINE

LEARNING OBJECTIVES & SCIENCE COMPETENCIES

Application of the principles of DNA structure and genetic variation examined in this chapter will enable you to solve the following types of problems:

- Explain the structure of DNA and how that structure facilitates its manipulation.
- Produce a restriction map of a DNA fragment based on gel patterns of fragments produced by cleaving it with multiple restriction enzymes.
- Use data from analysis of DNA markers to identify individuals and relatedness.
- For a given sequence of DNA, select the sequences of primer oligonucleotides that would allow any specific fragment of the molecule to be amplified in the polymerase chain reaction.

Prior to the mid-1970s, classical and molecular genetics, while addressing the same questions, were often treated as separate disciplines. Since then, studies in genetics have undergone a revolution based on the use of increasingly sophisticated ways to isolate and identify specific fragments of DNA, which have substantially merged the two approaches. The culmination of these developments was large-scale genomic sequencing—the ability to determine the correct sequence of the base pairs that make up the DNA in an entire genome and to identify the sequences associated with genes. Because many of the laboratory organisms used in genetics experiments have relatively small genomes, these sequences were completed first. The techniques used to sequence these simpler genomes were then scaled up to sequence much larger genomes, including the human genome. This has greatly expanded our ability to investigate even the most complex of traits at the molecular, cellular, and organismal levels.

2.1 Genetic Differences Among Individuals

The human genome in a reproductive cell consists of approximately 3 billion base pairs organized into 23 distinct chromosomes (each chromosome contains a single molecule of duplex DNA). A typical chromosome can contain several hundred to several thousand genes, arranged in linear order along the DNA molecule present in the chromosome. The sequences that make up the protein-coding part of these genes actually account for only about 1.3 percent of the entire genome. The other 98.7 percent of the sequences do not code for proteins. Some encode RNAs that are not mRNAs, but rather are molecules that contribute to a wide variety of cellular functions. Others are noncoding sequences, with many of these being relatively short sequences that are found in hundreds of thousands of copies scattered throughout the genome. Still other noncoding sequences are decayed remnants of genes called **pseudogenes**. And still others are noncoding sequences whose functions are the subject of much current investigation. As might be expected, identifying the protein-coding genes against the large background of noncoding DNA in the human genome is a challenge in itself.

Geneticists often speak of the nucleotide sequence of "the" human genome because corresponding DNA sequences from any two individuals are identical at approximately 99.9 percent of their nucleotide sites. These shared sequences are our evolutionary legacy: They contain the genetic information that makes us human beings. In reality, however, there are many different human genomes. Geneticists have the most interest in the 0.1 percent of the human DNA sequence—3 million base pairs—that differs from one genome to the next. Most of this variation is "normal," but these differences also include the mutations that are responsible for genetic diseases such as phenylketonuria (PKU) and other inborn errors of metabolism, as well as the mutations that increase individuals' risk of developing more complex diseases such as heart disease, breast cancer, and diabetes.

Fortunately, only a small proportion of all differences in DNA sequence are associated with disease. Some of the others are associated with inherited differences in height, weight, hair color, eye color, facial features, and other traits. Most of the genetic differences between people are completely harmless. Many have no detectable effects on appearance or health. Such differences can be studied only through direct examination of the DNA itself. These differences are nevertheless important, because they serve as genetic markers.

DNA Markers as Landmarks in Chromosomes

In genetics, a **genetic marker** is any difference in DNA, no matter how it is detected, whose pattern of transmission from generation to generation can be tracked. Each individual who carries the marker also carries a length of chromosome on either side of it, so that the marker *marks* a particular region of the genome. Any difference in DNA sequence between

two individuals can serve as a genetic marker. Although genetic markers are often harmless in themselves, they allow the positions of disease genes to be located along the chromosomes and their DNA to be isolated, identified, and studied.

Genetic markers that are detected by direct analysis of the DNA are often called **DNA markers**. DNA markers are important in genetics because they serve as landmarks in long DNA molecules, such as those found in chromosomes, which allow researchers to track genetic differences among individuals. In this sense, they are like signposts along a highway. Using DNA markers as landmarks, geneticists can identify the positions of normal genes, mutant genes, breaks in chromosomes, and other features important in genetic analysis.

A hypothetical example is given in **FIGURE 2.1**, involving the *PAH* gene encoding phenylalanine hydroxylase. In this figure, we see examples of two copies of the segment of chromosome 12, one of which codes for the normal protein and the other of which codes for a protein associated with PKU. These two segments also differ with respect to a noncoding nucleotide. This nucleotide could be a genetic marker, in that

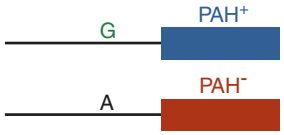


FIGURE 2.1 This schematic shows two DNA molecules that contain both *PAH* variants and a DNA marker—in this case, an A or a G at the indicated position. In this hypothetical case, the A allele is associated with the disease-associated variant and, therefore, could be used as an indicator of the latter's presence in an individual's genome.

while the G residue associated with the PKU variant is *not* the difference that causes PKU, its presence on a particular chromosome could be used as a marker to indicate that variant's presence.

Of course, the problem we face is how to start from the total DNA of an individual and isolate a particular gene of interest, so that we can identify genetic differences between individuals. In this chapter, we will examine some of the principal ways in which DNA is manipulated to achieve this feat, whether or not these differences result in observable differences. An overview of the steps involved is shown in **FIGURE 2.2**.

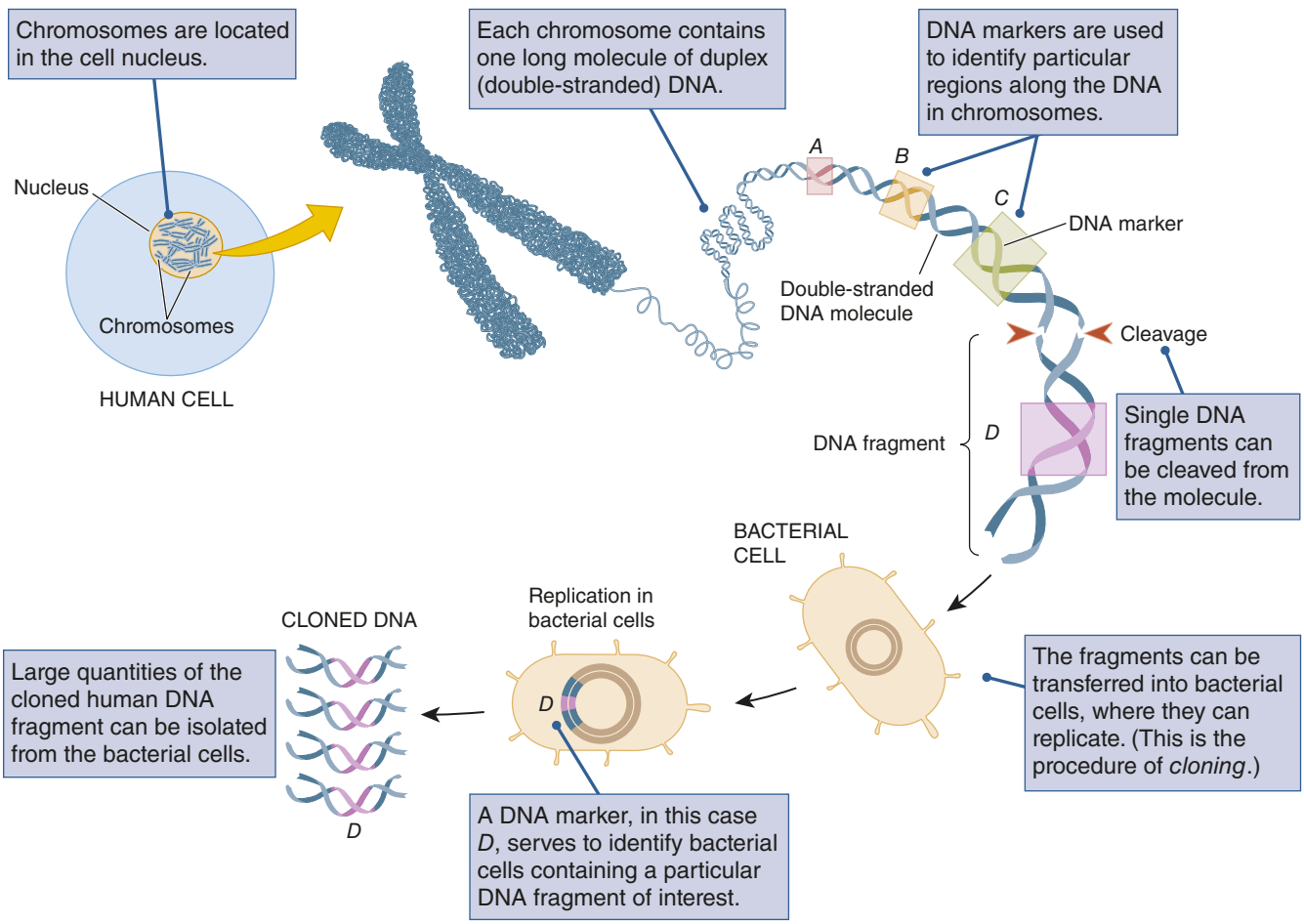


FIGURE 2.2 DNA markers serve as landmarks that identify physical positions along a DNA molecule, such as DNA from a chromosome. A DNA marker can also be used to identify bacterial cells into which a particular fragment of DNA has been introduced. The procedure of DNA cloning is not quite as simple as indicated here; it is discussed further in the *Manipulating Genes and Genomes* chapter.

Use of these methods broadens the scope of genetics, making it possible to carry out genetic analysis in *any* organism. As a consequence, detailed genetic analysis is no longer restricted to human beings, domesticated animals, cultivated plants, and the relatively small number of model organisms favorable for genetic studies. Direct study of DNA eliminates the need for prior identification of genetic differences between individuals; it even eliminates the need for controlled crosses. The methods of molecular analysis discussed in this chapter have transformed genetics and are the principal techniques used in almost every modern genetics laboratory; furthermore, having a basic understanding of them makes it possible to appreciate the overall unity of classical and molecular genetics.

SUMMING UP

- Every human genome is unique, but only a small fraction of the genome differs from individual to individual.
- Most of the variation that does exist has no obvious physical effect. Variants associated with hereditary diseases are rare.
- Protein-coding genes represent only a small fraction of the DNA in a human genome.
- Studying DNA markers greatly extends the power of genetic analysis.

2.2 The Terminology of Genetic Analysis

To discuss genetic analysis at any level, we must first introduce some key terms that provide the essential vocabulary of genetics. These terms can be understood with reference to **FIGURE 2.3**. In the *Genes, Genomes, and Genetic Analysis* chapter, we defined a **gene** as an element of heredity, transmitted from parents to offspring in reproduction, that influences one or more hereditary traits. Chemically, a gene is a sequence of nucleotides along a DNA molecule. In a population of organisms, not all copies of a gene may have exactly the same nucleotide sequence. For example, whereas one form of a gene may have the codon GCA at a certain position, another form of the same gene may have the codon GCG. Both codons specify alanine. Hence, the two forms of the gene encode the same sequence of amino acids, yet differ in DNA sequence. The alternative forms of a gene are called **alleles** of the gene. Different alleles may also code for different amino acid sequences, sometimes with drastic effects. Recall the example of the *PAH* gene for phenyalanine hydroxylase in the *Genes, Genomes, and Genetic Analysis* chapter, in which a change in codon 408 from CGG (arginine) to TGG (tryptophan) results in an inactive enzyme that becomes expressed as the inborn error of metabolism phenylketonuria.

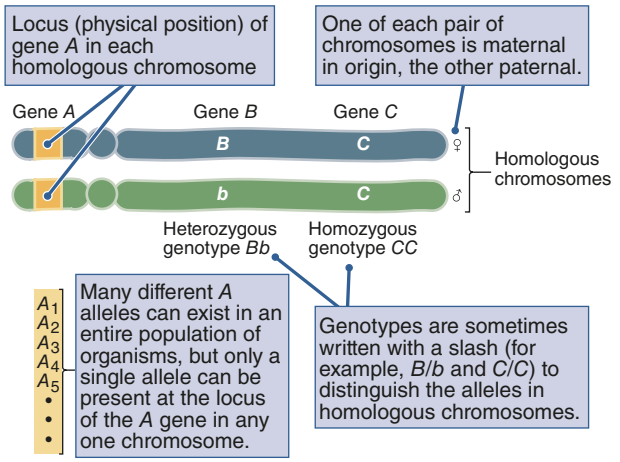


FIGURE 2.3 Key concepts and terms used in modern genetics. Note that a single gene can have any number of alleles in the population as a whole, but no more than two alleles can be present in any one individual.

Within a cell, genes are arranged in linear order along microscopic thread-like bodies called **chromosomes**, which we examine in detail in the chapters titled *The Chromosomal Basis of Inheritance* and *Human Karyotypes and Chromosome Behavior*. Each human reproductive cell contains one complete set of 23 chromosomes containing 3×10^9 base pairs of DNA. A typical chromosome contains several hundred to several thousand genes. In humans, the average is approximately 1000 genes per chromosome. Each chromosome contains a single molecule of duplex DNA along its length, complexed with proteins and very tightly coiled. The DNA in the average human chromosome, when fully extended, has relative dimensions comparable to those of a wet spaghetti noodle 25 miles long; when the DNA is coiled in the form of a chromosome, its physical compaction is comparable to that of the same noodle coiled and packed into an 18-foot canoe.

The physical position of a gene along a chromosome is called the **locus** of the gene. In most higher organisms, including human beings, each cell other than a sperm or egg contains two copies of each type of chromosome—one inherited from the mother and one inherited from the father. Each member of such a pair of chromosomes is said to be **homologous** to the other. (The chromosomes that determine sex are an important exception, which we will ignore for now.) At any locus, therefore, each individual carries two alleles, because one allele is present at a corresponding position in each of the homologous maternal and paternal chromosomes (Figure 2.3).

The genetic constitution of an individual is called its **genotype**. For a particular gene, if the two alleles at the locus in an individual are indistinguishable from each other, then the genotype of the individual is said

to be **homozygous** for the allele that is present. If the two alleles at the locus are different from each other, then the genotype of the individual is said to be **heterozygous** for the alleles that are present. Typographically, genes are indicated in italics, and alleles are typically distinguished by uppercase or lowercase letters (*A* versus *a*), subscripts (A_1 versus A_2), superscripts (a^+ versus a^-), or sometimes just + and −. Using these symbols, homozygous genes would be portrayed by any of these formulas: *AA*, *aa*, A_1A_1, A_2A_2, a^+a^+, a^-a^-, +/+, or −/−. As in the last two examples, the slash is sometimes used to separate alleles present in homologous chromosomes to avoid ambiguity. Heterozygous genes would be portrayed by any of the formulas *Aa*, A_1A_2, a^+a^-, or +/−. In Figure 2.3, the genotype *Bb* is heterozygous because the *B* and *b* alleles are distinguishable (which is why they are assigned different symbols), whereas the genotype *CC* is homozygous. These genotypes could also be written as *B*/*b* and *C*/*C*, respectively.

Whereas the alleles that are present in an individual constitute its genotype, the physical or biochemical expression of the genotype is called the **phenotype**. To put it as simply as possible, the distinction is that the genotype of an individual is what is on the *inside* (the alleles in the DNA), whereas the phenotype is what is on the *outside* (the observable traits, including biochemical traits, behavioral traits, and so forth). The distinction between genotype and phenotype is critically important because there usually is not a one-to-one correspondence between genes and traits. Most complex traits—such as hair color, skin color, height, weight, behavior, life span, and susceptibility to disease—are influenced by many genes. Most traits are also influenced more or less strongly by environment. Thus, depending on the environment, the same genotype can result in different phenotypes. Compare, for example, two people with a genetic risk for lung cancer: If one smokes and the other does not, the smoker is much more likely to develop the disease. Environmental effects also imply that the same phenotype can result from more than one genotype. Smoking again provides an example, because most smokers who are not genetically at risk can also develop lung cancer.

Harmful mutations represent only a small fraction of the total allelic variation in a species, and most such alleles are very uncommon. In contrast, in many cases, different alleles at a locus can have no evident effects on phenotypes; often multiple alleles of such loci are found at relatively high frequencies (greater than 5 percent). Such cases are called **polymorphisms**; the term *polymorphism* literally means "multiple forms." At the sequence level, these variants are referred to as **DNA polymorphisms**.

SUMMING UP

- The genome of a typical human cell contains 23 pairs of homologous chromosomes, each of which carries an average of 1000 genes.
- Variants of particular genes are known as alleles.
- Individuals with two identical alleles of a particular gene are homozygotes; those with two different alleles are heterozygotes.
- The combination of alleles carried by an individual is the genotype.
- The physical manifestation of the genotype is the phenotype.
- Many loci in the human genome are polymorphic.

2.3 The Molecular Structure of DNA

Modern experimental methods for the manipulation and analysis of DNA grew out of a detailed understanding of its molecular structure and replication. Therefore, to understand these methods, one needs to know something about the molecular structure of DNA. As we saw in the *Genes, Genomes, and Genetic Analysis* chapter, DNA is a helix consisting of two paired, complementary strands, each composed of an ordered string of *nucleotides*, with each nucleotide bearing one of the bases A (adenine), T (thymine), G (guanine), or C (cytosine). Watson–Crick base pairing between A and T and between G and C in the complementary strands holds the strands together. The complementary strands also hold the key to replication, because each strand can serve as a template for the synthesis of a new complementary strand. We will now take a closer look at DNA structure and at the key features of its replication.

Polynucleotide Chains

In terms of biochemistry, a DNA strand is a polymer—a large molecule built from repeating units. The units in DNA are composed of 2′-deoxyribose (a five-carbon sugar), phosphoric acid, and the four nitrogen-containing bases denoted A, T, G, and C. The chemical structures of the bases are shown in **FIGURE 2.4**. Note that two of the bases have a double-ring structure; these are called **purines**. The other two bases have a single-ring structure; these are called **pyrimidines**.

- The purine bases are adenine (A) and guanine (G).
- The pyrimidine bases are thymine (T) and cytosine (C).

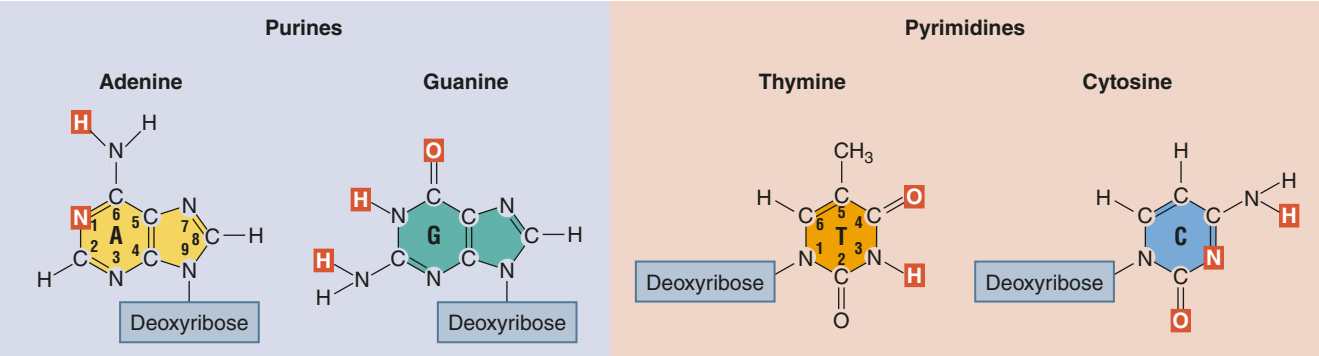


FIGURE 2.4 Chemical structures of the four nitrogen-containing bases in DNA: adenine, thymine, guanine, and cytosine. The nitrogen atom linked to the deoxyribose sugar is indicated. The atoms shown in red participate in hydrogen bonding between the DNA base pairs.

In DNA, each base is chemically linked to one molecule of the sugar deoxyribose, forming a compound called a **nucleoside**. When a phosphate group is also attached to the sugar, the nucleoside becomes a **nucleotide** (**FIGURE 2.5**). Thus, a nucleotide is a nucleoside plus a phosphate. In the conventional numbering of the carbon atoms in the sugar in Figure 2.3, the carbon atom to which the base is attached is the 1′ carbon. (The atoms in the sugar are given primed numbers to distinguish them from atoms in the bases.) The nomenclature of the nucleoside and nucleotide derivatives of the DNA bases is summarized in **TABLE 2.1**. Most of these terms are not needed in this text; they are included because they are likely to be encountered in further reading.

In nucleic acids, such as DNA and RNA, the nucleotides join together to form a **polynucleotide chain**, in which the phosphate attached to the 5′ carbon of one sugar is linked to the hydroxyl group attached to the 3′ carbon of the growing chain (**FIGURE 2.6**). The chemical bonds by which the sugar components of adjacent nucleotides are linked through the phosphate groups are called **phosphodiester bonds**. The 3′–5′–3′–5′ orientation of these linkages continues throughout the chain, which typically consists of millions of nucleotides. Note that the terminal groups of each polynucleotide chain are a **5′-phosphate (5′-P) group** at one end and a **3′-hydroxyl (3′-OH) group** at the other end. The asymmetry of the ends of a DNA strand implies that each strand has a **polarity** determined by which end bears the 5′ phosphate and which end bears the 3′ hydroxyl.

A few years before Watson and Crick proposed their essentially correct three-dimensional structure of DNA as a double helix, Erwin Chargaff developed a chemical technique to measure the amount of each base present in DNA. As we describe this technique, we denote the molar concentration of any base by the symbol for the

TABLE 2.1 DNA nomenclature

Base	Nucleoside	Nucleotide
Adenine (A)	Deoxyadenosine	Deoxyadenosine-5′ monophosphate (dAMP) diphosphate (dADP) triphosphate (dATP)
Guanine (G)	Deoxyguanosine	Deoxyguanosine-5′ monophosphate (dGMP) diphosphate (dGDP) triphosphate (dGTP)
Thymine (T)	Deoxythymidine	Deoxythymidine-5′ monophosphate (dTMP) diphosphate (dTDP) triphosphate (dTTP)
Cytosine (C)	Deoxycytidine	Deoxycytidine-5′ monophosphate (dCMP) diphosphate (dCDP) triphosphate (dCTP)

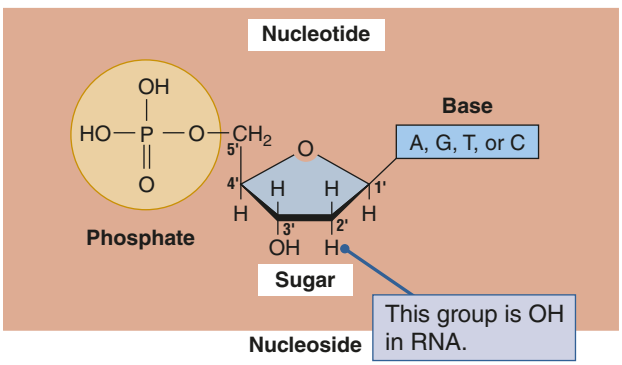
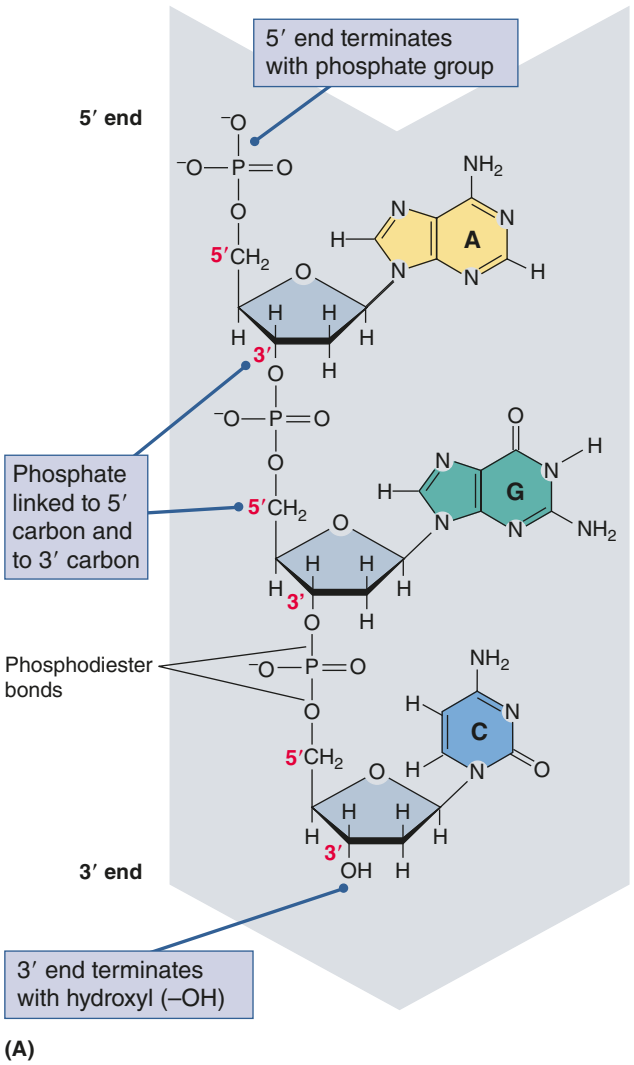
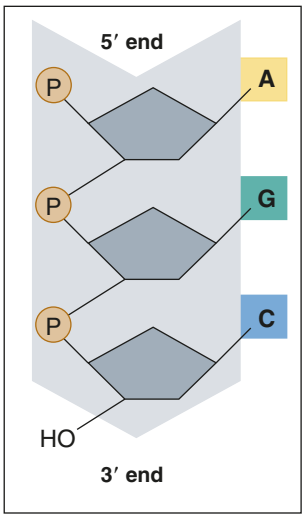


FIGURE 2.5 A typical nucleotide, showing the three major components (phosphate, sugar, and base) that are the difference between DNA and RNA. A nucleoside consists of the sugar and base only. Nucleotides are monophosphates (with one phosphate group). Nucleoside diphosphates contain two phosphate groups, and nucleoside triphosphates contain three.



(A)



(B)

FIGURE 2.6 Three nucleotides at the 5′ end of a single polynucleotide strand. (A) The chemical structure of the sugar–phosphate linkages, showing the 5′-to-3′ orientation of the strand (the red numbers are those assigned to the carbon atoms). (B) A common schematic way to depict a polynucleotide strand.

base enclosed in square brackets; for example, [A] denotes the molar concentration of adenine. Chargaff used his technique to measure the [A], [T], [G], and [C] content of the DNA from a variety of sources. He found that the **base composition** of the DNA, defined as the **percent G + C**, differs among species but is constant in all cells of an organism and within a species. Data on the base composition of DNA from a variety of organisms are given in **TABLE 2.2**.

Chargaff also observed certain regular relationships among the molar concentrations of the different bases. These relationships are now called **Chargaff's rules**:

- The amount of adenine equals that of thymine: [A] = [T].
- The amount of guanine equals that of cytosine: [G] = [C].
- The amount of the purine bases equals that of the pyrimidine bases: [A] + [G] = [T] + [C].

Although the chemical basis of these observations was not known at the time, one of the appealing features of the Watson–Crick structure of paired complementary strands was that it explained Chargaff's rules. Because A is always paired with T in double-stranded DNA, it must follow that [A] = [T]. Similarly, because G is paired with C, we know that [G] = [C]. The third rule follows by addition of the other two: [A] + [T] = [G] + [C]. In the next section, we examine the molecular basis of base pairing in more detail.

The Double Helix

In addition to Chargaff's rules and the basic chemistry of nucleotides, Watson and Crick drew on structural observations made by Rosalind Franklin and Maurice Wilkins. These scientists used **x-ray crystallography** to examine the three-dimensional structure of DNA. In this procedure, crystals of large molecules are exposed to beams of x-rays. The shape of the crystals causes the x-rays to scatter, or diffract, in a manner that is determined by the structure of the crystal. The pattern of diffraction is then recorded on photographic film, and the crystallographer can use it to make inferences about the crystal's structure.

TABLE 2.2 Base composition of DNA from different organisms

Organism	Base (and percentage of total bases) Adenine	Thymine	Guanine	Cytosine	Base composition (percent G + C)
Bacteriophage T7	26.0	26.0	24.0	24.0	48.0
Bacteria					
Clostridium perfringens	36.9	36.3	14.0	12.8	26.8
Streptococcus pneumoniae	30.2	29.5	21.6	18.7	40.3
Escherichia coli	24.7	23.6	26.0	25.7	51.7
Sarcina lutea	13.4	12.4	37.1	37.1	74.2
Fungi					
Saccharomyces cerevisiae	31.7	32.6	18.3	17.4	35.7
Neurospora crassa	23.0	22.3	27.1	27.6	54.7
Higher plants					
Wheat	27.3	27.2	22.7	22.8*	45.5
Maize	26.8	27.2	22.8	23.2*	46.0
Animals					
Drosophila melanogaster	30.8	29.4	19.6	20.2	39.8
Pig	29.4	29.6	20.5	20.5	41.0
Salmon	29.7	29.1	20.8	20.4	41.2
Human being	29.8	31.8	20.2	18.2	38.4

* Includes one-fourth 5-methylcytosine, a modified form of cytosine found in most plants more complex than algae and in many animals.

Perhaps the most famous such image ever made was Franklin's Photo 51 (**FIGURE 2.7**). From it, crystallographers could make the following inferences:

- The crystal has a helical structure.
- The diameter of the helix is 20 angstroms (Å).
- The helix has two levels of repetitive structure, or periodicities, along its length. One of these occurs every 3.4 Å, and the other occurs every 34 Å.

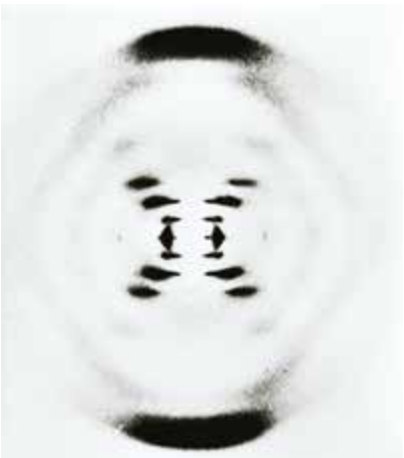


FIGURE 2.7 Rosalind Franklin's "Photo 51," the x-ray diffraction pattern obtained from crystallized DNA. The cross-like nature of the image indicates the helical structure of DNA.

Watson and Crick used all of these data when building their model of the double helix.

Base Pairing and Base Stacking

In the three-dimensional structure of the DNA molecule proposed in 1953 by Watson and Crick, the molecule consists of two polynucleotide chains twisted around each other to form a double-stranded helix in which adenine and thymine, and guanine and cytosine, are paired in opposite strands (**FIGURE 2.8**). In the standard structure, which is called the **B form of DNA**, each chain makes one complete turn every 34 Å. The helix is right-handed, which means that as one looks down the barrel, each chain follows a clockwise path as it progresses. The bases are spaced at 3.4 Å, so there are 10 bases per helical turn in each strand and 10 base pairs per turn of the double helix. These correspond to the two periodicities inferred from the x-ray crystallographic data.

The strands feature **base pairing**, in which each base is paired with a complementary base in the other strand by hydrogen bonds. (A **hydrogen bond** is a weak bond in which two participating atoms share a hydrogen atom between them.) The hydrogen bonds provide one type of force holding the strands together. In Watson–Crick base pairing, adenine (A) pairs with thymine (T), and guanine (G) pairs with cytosine (C). The hydrogen bonds that form in the adenine–thymine base pair and in the guanine–cytosine pair are

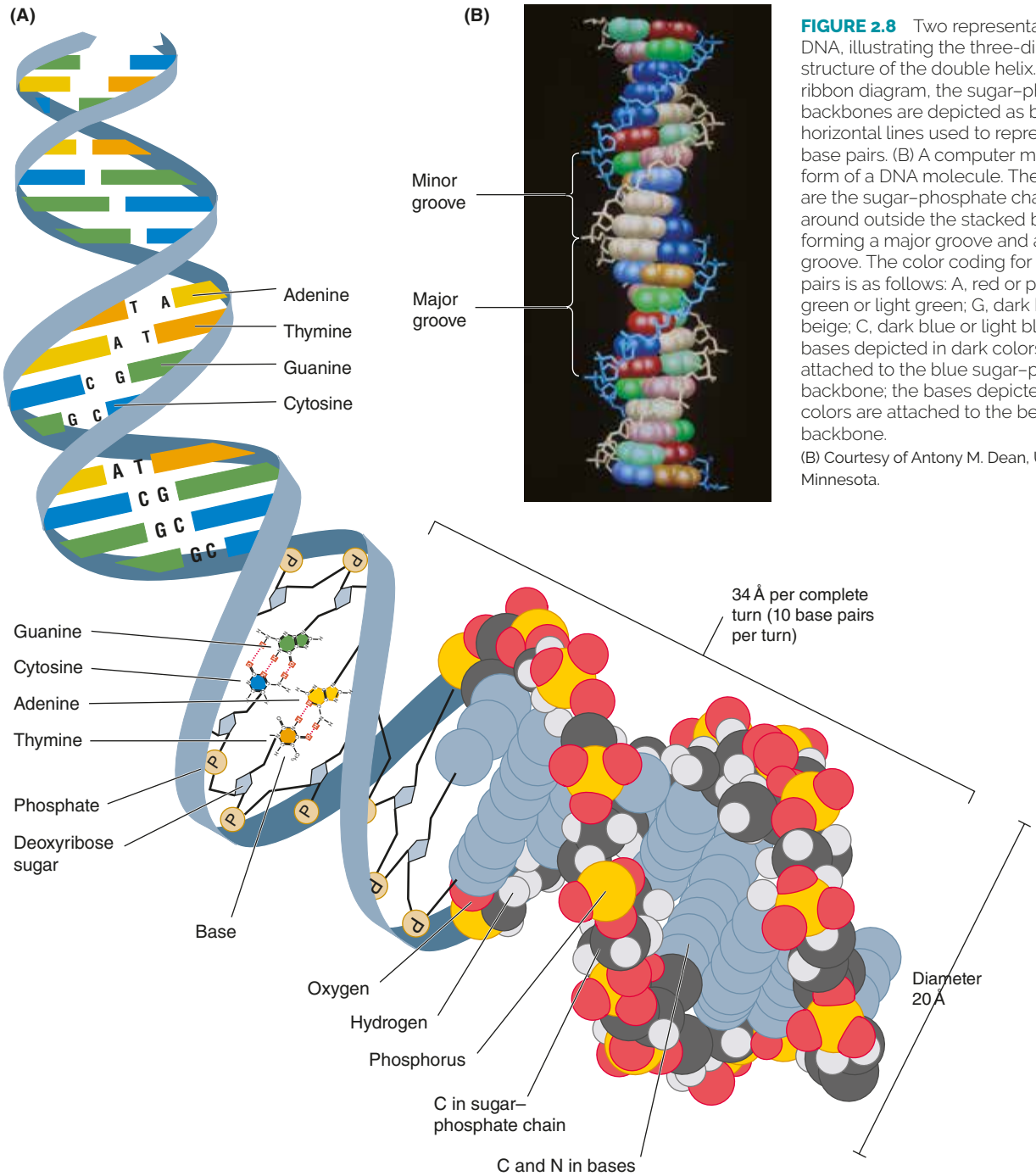


FIGURE 2.8 Two representations of DNA, illustrating the three-dimensional structure of the double helix. (A) In a ribbon diagram, the sugar–phosphate backbones are depicted as bands, with horizontal lines used to represent the base pairs. (B) A computer model of the B form of a DNA molecule. The stick figures are the sugar–phosphate chains winding around outside the stacked base pairs, forming a major groove and a minor groove. The color coding for the base pairs is as follows: A, red or pink; T, dark green or light green; G, dark brown or beige; C, dark blue or light blue. The bases depicted in dark colors are those attached to the blue sugar–phosphate backbone; the bases depicted in light colors are attached to the beige backbone.
(B) Courtesy of Antony M. Dean, University of Minnesota.

illustrated in **FIGURE 2.9**. Note that an A–T pair (Figure 2.9A and B) has two hydrogen bonds and that a G–C pair (Figure 2.9C and D) has three hydrogen bonds. This means that the hydrogen bonding between G and C is stronger in the sense that it requires more energy to break; for example, the amount of heat required to separate the paired strands in a DNA duplex increases with the percentage of G + C. Because nothing restricts the sequence of bases in a single strand, any sequence could be present along one strand. This explains Chargaff's observation that DNA from different organisms may differ in base composition. However, because the strands in duplex DNA are complementary, Chargaff's rules of [A] = [T] and [G] = [C] are true whatever the base composition.

In the B form of DNA, the paired bases are stacked on top of one another like pennies in a roll. The upper and lower faces of each nitrogenous base are relatively flat and nonpolar (uncharged). These surfaces are said to be *hydrophobic* because they bind poorly to water molecules, which are very polar. (The polarity refers to the asymmetrical distribution of charge across the V-shaped water molecule; the oxygen atom at the base of the V tends to be quite negative, whereas the hydrogen atoms at the tips

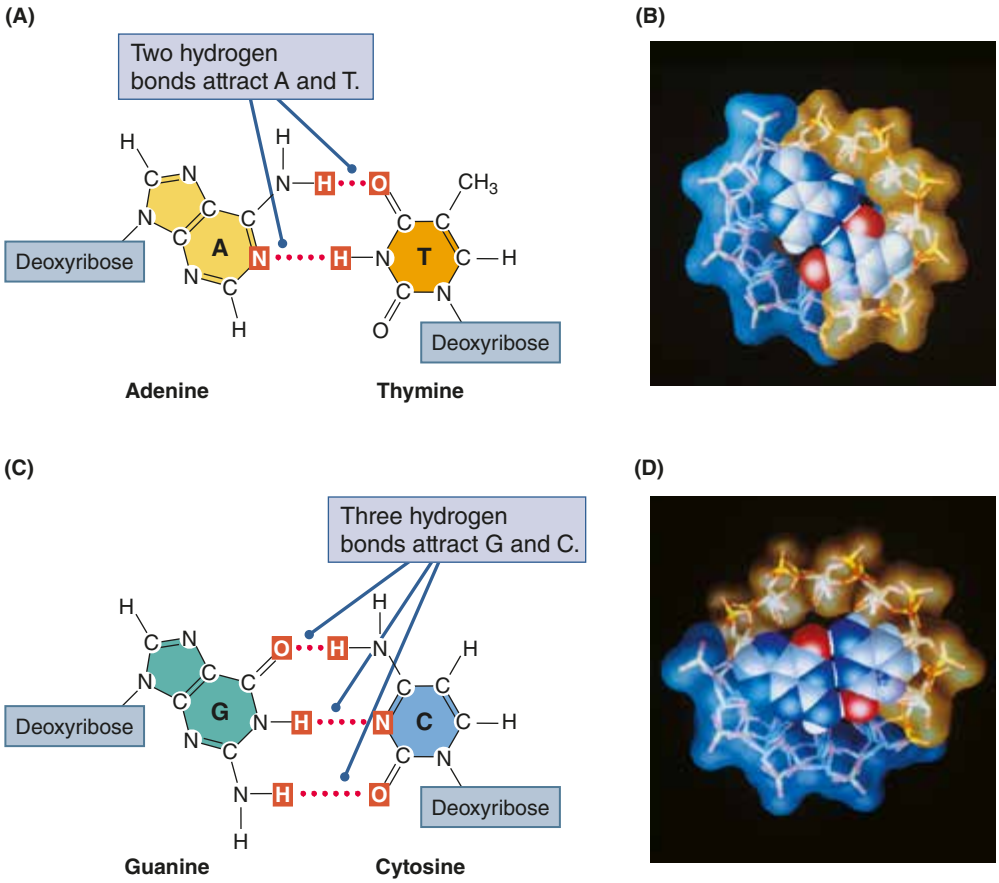


FIGURE 2.9 Normal base pairs in DNA. On the left, the hydrogen bonds (dotted lines) with the joined atoms are shown in red. (A and B) A–T base pairing. (C and D) G–C base pairing. In the space-filling models (B and D), the colors are as follows: C, gray; N, blue; O, red; and H (shown in the bases only), white. Each hydrogen bond is depicted as a white disk squeezed between the atoms sharing the hydrogen. The stick figures on the outside represent the backbones winding around the stacked base pairs.
Source: (B, D) Space-filling models courtesy of Antony M. Dean, University of Minnesota.

are quite positive). Owing to their repulsion of water molecules, the paired nitrogenous bases tend to stack on top of one another in such a way as to exclude the maximum amount of water from the interior of the double helix. This feature of double-stranded DNA is known as **base stacking**. A double-stranded DNA molecule therefore has a hydrophobic core composed of stacked bases, and it is the energy of base stacking that provides double-stranded DNA with much of its chemical stability (base pairing, by itself, would be insufficient to stabilize the molecule under physiological conditions).

When discussing a DNA molecule, molecular biologists frequently refer to the individual strands as single strands or as single-stranded DNA; they refer to the double helix as double-stranded DNA or a *duplex* molecule. The two grooves spiraling along outside of the double helix are not symmetrical; one groove, called the **major groove**, is larger than the other, which is called the **minor groove**. Proteins that interact with double-stranded DNA often have regions that make contact with the base pairs by fitting into the major groove, into the minor groove, or into both grooves (Figure 2.8B).

Antiparallel Strands

Each backbone in a double helix consists of deoxyribose sugars alternating with phosphate groups that link the 3′ carbon atom of one sugar to the 5′ carbon of the next in line (Figure 2.5). The two polynucleotide strands of the double helix have opposite polarity, in the sense that the 5′ end of one strand is paired with the 3′ end of the other strand. Strands with such an arrangement are said to be **antiparallel**. One implication of the presence of antiparallel strands in duplex DNA is that in each pair of bases, one base is attached to a sugar that lies above the plane of pairing, and the other base is attached to a sugar that lies below the plane of pairing. Another implication is that each terminus of the double helix possesses one 5′-P group (on one strand) and one 3′-OH group (on the other strand), as shown in **FIGURE 2.10**.

The diagram of the DNA duplex in Figure 2.8 is static and, therefore, somewhat misleading. DNA is actually a dynamic molecule, constantly in motion. In some regions, the strands can separate briefly and then come together again in the same conformation or in a different one. The right-handed double helix in Figure 2.8 is the

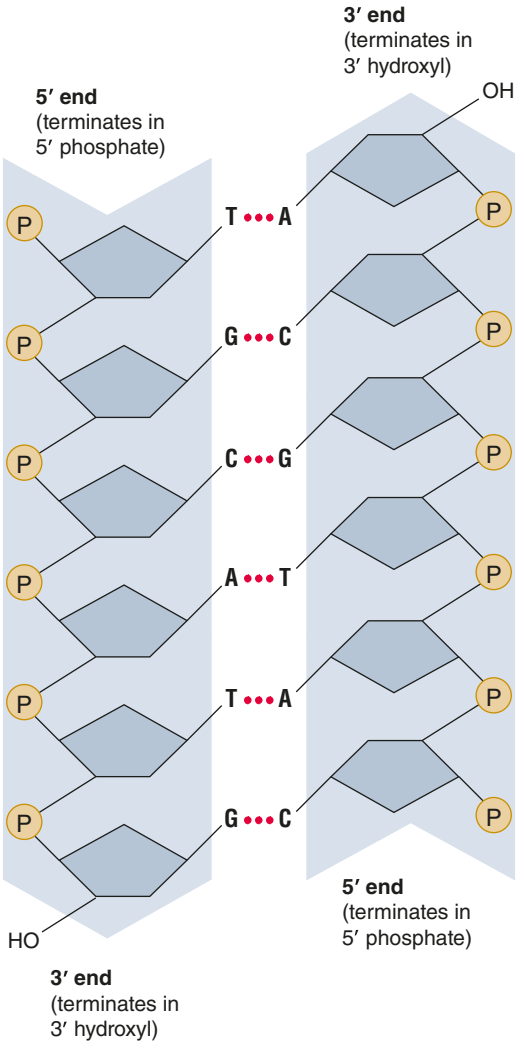


FIGURE 2.10 A segment of a DNA molecule, showing the antiparallel orientation of the complementary strands. The shaded blue arrows indicate the 5′-to-3′ direction of each strand. The phosphates (P) join the 3′ carbon atom of one deoxyribose to the 5′ carbon atom of the adjacent deoxyribose.

standard B form, but depending on conditions, DNA can actually form more than 20 slightly different variants of a right-handed helix, and some regions can even form helices in which the strands twist to the left (called the *Z form of DNA*). If complementary stretches of nucleotides appear in the same strand, then a single strand, separated from its partner, can fold back upon itself like a hairpin. Even triple helices consisting of three strands can form in regions of DNA that contain suitable base sequences.

DNA Structure as Related to Function

In the structure of the DNA molecule, we can see how the four essential requirements of a genetic material are met.

1. *Any genetic material must be able to be replicated accurately, so that the information it contains will be precisely replicated and inherited by daughter cells.* The basis for exact duplication of a DNA molecule is the pairing of A with T and of G with C in the two polynucleotide chains. Unwinding and separation of the strands, with each free strand being copied, results in the formation of two identical double helices.
2. *Genetic material must carry encoded information.* Quite clearly, the order of the bases in DNA provides the basis for such a code—specifically, a genetic code in which groups of three bases specify amino acids. Because the four bases in a DNA molecule can be arranged in any sequence, and because the sequence can vary from one part of the molecule to another and from organism to organism, DNA can contain a great many unique regions, each of which can be a distinct gene.
3. *Genetic material must have the capacity to direct the organization and metabolic activities of the cell.* As we saw in the *Genes, Genomes, and Genetic Analysis* chapter, genes can direct the synthesis of a protein molecule—a polymer composed of repeating units of amino acids. The sequence of amino acids in the protein determines its chemical and physical properties. A gene is expressed when its protein product is synthesized, and one requirement of the genetic material is that it direct the sequence in which amino acid units are added to the end of a growing protein molecule.
4. *Genetic material must be capable of undergoing occasional mutations in which the information it carries is altered.* If these mutations are to be heritable, the mutant molecules must also be capable of being replicated as faithfully as the parental molecule. This feature is necessary to account for the evolution of diverse organisms through the slow accumulation of favorable mutations. Watson and Crick suggested that heritable mutations might be possible in DNA by rare mispairing of the bases, with the result that an incorrect nucleotide becomes incorporated into a replicating DNA strand.

SUMMING UP

- DNA is a helical molecule, consisting of two antiparallel strands.
- The two strands of DNA are complementary: A pairs with T, and G pairs with C.
- The helix is stabilized by the hydrogen bonds that form between complementary bases, as well as by the stacking interactions that occur between adjacent base pairs.
- The structure of DNA is consistent with the four necessary properties of genetic material.

ROOTS OF DISCOVERY

The Double Helix

James D. Watson and Francis H. C. Crick (1953)
Cavendish Laboratory Cambridge, England

A Structure for Deoxyribose Nucleic Acid

This is one of the watershed papers of twentieth-century biology. Watson and Crick benefited tremendously from knowing that their structure was consistent with the unpublished structural studies of Maurice Wilkins and Rosalind Franklin. The same issue of Nature *that included the Watson and Crick paper also included, back to back, a paper from the Wilkins group and one from the Franklin group detailing their data and the consistency of their data with the proposed structure. It has been said that Franklin was poised a mere two half-steps from making the discovery herself, alone. In any event, Watson and Crick and Wilkins were awarded the 1962 Nobel Prize for their discovery of DNA structure. Rosalind Franklin, tragically, died of cancer in 1958 at the age of 38.*

Watson and Crick's 1953 paper was concise, occupying only a single printed page in the journal Nature. *It describes the basic molecular model—the helical structure, the antiparallel strands, and the physical positioning of the bases and the phosphate groups. It then describes base pairing and its significance.*

Only specific pairs of bases can bond together. These pairs are adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine). . . . The sequence of bases on a single chain does not appear to be restricted in any way. However, . . . it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined. . . . It has not escaped our notice that the specific pairing we have postulated immediately suggests a plausible copying mechanism for the genetic material.

"If only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined."

This last sentence is one of the most widely quoted lines in all of biology. In a subsequent, more detailed paper, Watson and Crick went a step further and considered how information could be encoded in DNA, given their proposed molecular structure.

The phosphate–sugar backbone of our model is completely regular, but any sequence of the pairs of bases can fit into the structure. It follows that in a long molecule many different permutations are possible, and it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information.

While Watson and Crick's first paper is considered iconic, it is in the second one that the biological and genetic significance of the double helix is most fully explored.

Source: J.D. Watson and F.H.C. Crick, Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171 (1953): 737–738.

2.4 The Separation and Identification of Genomic DNA Fragments

The following sections show how an understanding of DNA structure and replication has been put to practical use in the development of procedures for the separation and identification of particular DNA fragments. These methods are used primarily either to identify DNA markers or to aid in the isolation of particular DNA fragments that are of genetic interest. For example, consider a pedigree of familial breast cancer in which a particular DNA fragment serves as a marker for a region of chromosome that also includes the gene, an allele of which is responsible for the increased risk. In this case, the ability to identify the marker genotype for each woman in the pedigree is critically important for predicting her risk of breast cancer. To take another example, suppose one is testing the hypothesis that an allelic variant associated with a hereditary disease is present in a particular DNA fragment. In this situation it is important to be able to pinpoint this fragment using genetic markers so as to isolate the fragment from affected individuals, verify whether the hypothesis is true, and identify the nature of the mutation.

Most procedures for the separation and identification of DNA fragments can be grouped into two general categories:

- Those that identify a specific DNA fragment present in genomic DNA by making use of the fact that complementary single-stranded DNA sequences can, under the proper conditions, form a duplex molecule. These procedures rely on *nucleic acid hybridization*.
- Those that use prior knowledge of the sequence at the ends of a DNA fragment to specifically and repeatedly replicate this one fragment from genomic DNA. These procedures rely on selective DNA replication (*amplification*) by means of the *polymerase chain reaction*.

The major difference between these approaches is that the first (relying on nucleic acid hybridization) identifies fragments that are present in the genomic DNA itself, whereas the second (relying on DNA amplification) identifies experimentally manufactured *replicas* of fragments whose original templates (but not the replicas) were present in the genomic DNA. This difference has practical implications:

- Hybridization methods require a greater amount of genomic DNA for the experimental procedures, but relatively large fragments can be identified, and no prior knowledge of the DNA sequence is necessary.
- Amplification methods require extremely small amounts of genomic DNA for the experimental procedures, but the amplification is usually restricted to relatively small fragments, and some prior knowledge of DNA sequence is necessary.

Historically, hybridization methods were the most important means for detecting specific DNA fragments. However, with the development of the polymerase chain reaction and automated DNA sequencing technologies in the late 1980s, amplification-based methods came to predominate these investigation. In the following section, we will briefly describe hybridization methods, after which we will focus in more depth on amplification-based methods.

Restriction Enzymes and Site-Specific DNA Cleavage

In methods that use nucleic acid hybridization to identify particular fragments present in genomic DNA, the first step is usually cutting the genomic DNA into specific fragments of experimentally manageable size. When genomic DNA is isolated from cells, it is typically fragmented into pieces with an average length of about 50,000 bases, or 50 kilobases (kb). It is possible to fragment them further, but the breakage is by and large random. Thus, only a small fraction of all of the fragments will contain a *particular* DNA sequence, and those fragments will vary in size (**FIGURE 2.11**). This creates a "needle-in-a-haystack" problem: How can that sequence of interest be separated from all of the rest?

One of the most important discoveries in the history of molecular genetics (for which Werner Arbers, Daniel Nathans, and Hamilton Smith shared the Nobel Prize in 1978) was that of **restriction endonucleases** or **restriction enzymes**. These enzymes, which are found in bacteria, can cleave DNA molecules at positions at which specific short sequences of DNA (typically 4–6 base pairs in length) occur. Restriction enzymes function in nature to protect bacteria by selectively degrading the genomes of bacteriophages that attack them. For example, the restriction enzyme *Bam*H1 (technically known as a *type II restriction endonuclease*) recognizes the sequence

5′-GGATCC-3′
3′-CCTAGG-5′

and cleaves each strand between the two G-bearing nucleotides, as shown in **FIGURE 2.12**.

FIGURE 2.13 shows nine of the several hundred restriction enzymes that are known. Most restriction enzymes are named after the species in which they were found. *Bam*HI, for example, was isolated from the bacterium *Bacillus amyloliquefaciens* strain H, and it is the first (I) restriction enzyme isolated from this organism. Because the first three letters in the name of each restriction enzyme stand for the bacterial species of origin, these letters are printed in italics; the rest of the symbols in the name are not italicized.

Most restriction enzymes recognize only one short base sequence, usually four or six nucleotide pairs. The enzyme binds with the DNA at these sites and makes a break in each strand of the DNA molecule, producing free 3′-OH and 5′-P groups at each position.

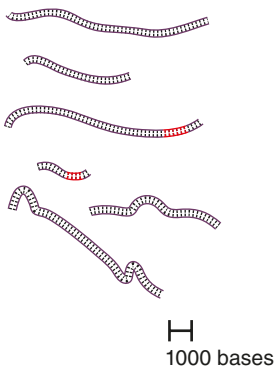


FIGURE 2.11 Illustration of random fragments of DNA, some of which carry a sequence of interest (shown in red).

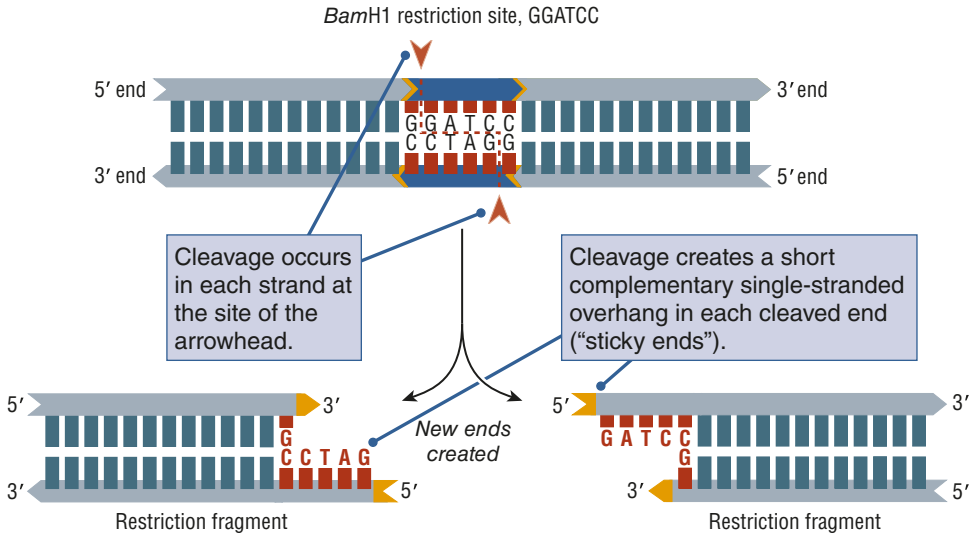


FIGURE 2.12 The mechanism of DNA cleavage by the restriction enzyme *Bam*HI. Wherever the duplex contains a *Bam*HI restriction site, the enzyme makes a single cut in the backbone of each DNA strand. Each cut creates a new 3′ end and a new 5′ end, separating the duplex into two fragments. In the case of *Bam*HI, the cuts are staggered cuts, so the resulting ends terminate in single-stranded regions, each four nucleotides in length.

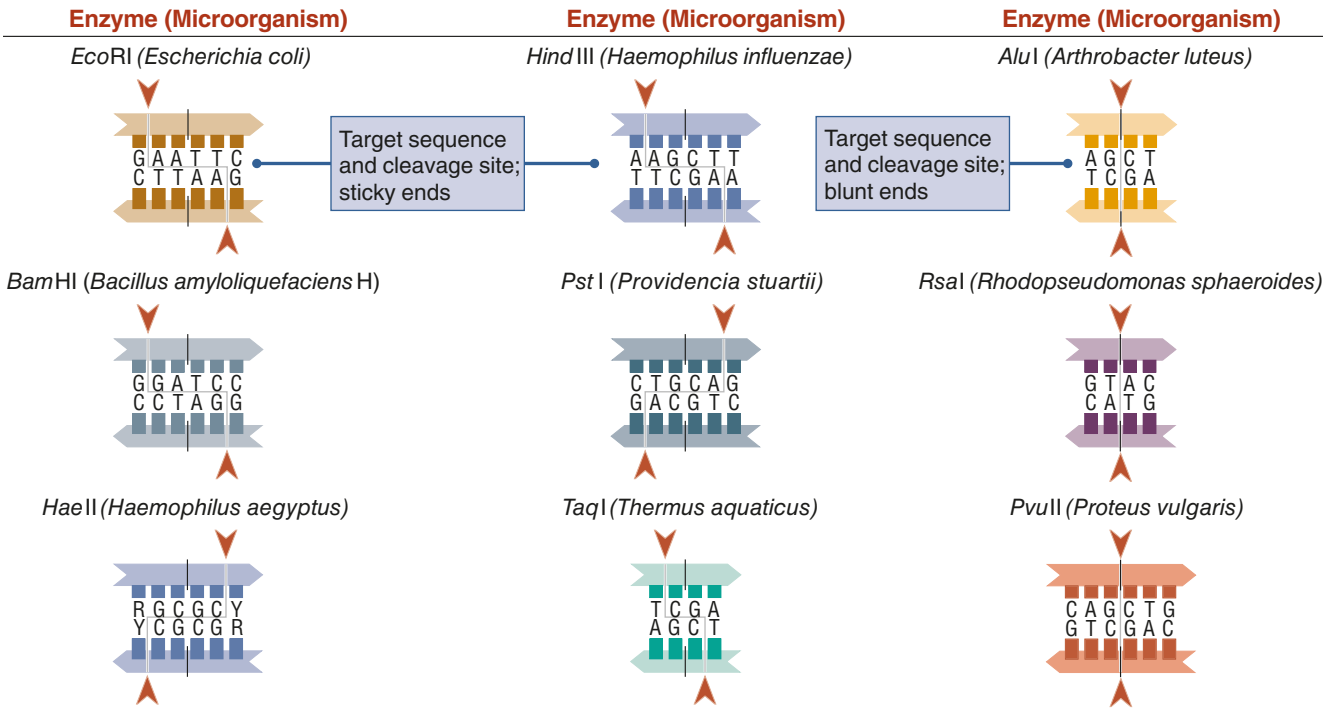
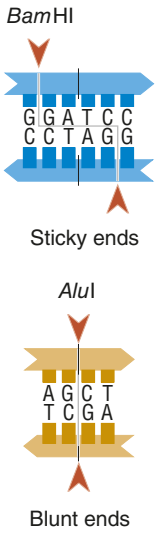


FIGURE 2.13 Recognition sites for various restriction enzymes. The vertical dashed line indicates the axis of symmetry in each sequence. Red arrows indicate the sites of cutting. Enzymes in the first two columns produce sticky ends; those in the third column produce blunt ends. The enzyme *Taq*I yields cohesive ends consisting of two nucleotides, whereas the cohesive ends produced by the other enzymes contain four nucleotides. R and Y refer to any complementary purines and pyrimidines, respectively.

The nucleotide sequence recognized for cleavage by a restriction enzyme is called the **restriction site** of the enzyme. Some restriction enzymes cleave their restriction site asymmetrically (at different sites in the two DNA strands), but other restriction enzymes cleave the site symmetrically (at the same site in both strands). The former leave **sticky ends** (also referred to as cohesive ends) because each end of the cleaved

site has a small, single-stranded overhang that is complementary in base sequence to the other end (Figure 2.12). In contrast, enzymes that have symmetrical cleavage sites yield DNA fragments that have **blunt ends**. In virtually all cases, the restriction site of a restriction enzyme reads the same on both strands, provided that the opposite polarity of the strands is taken into account; for example, each strand in the restriction site of *Bam*HI reads 5′-GGATCC-3′ (Figure 2.10). A DNA sequence with this type of symmetry is called a **palindrome**. (In ordinary English, a palindrome is a word or phrase that reads the same forward and backward, such as "madam.")



Restriction enzymes have the following important characteristics:

- Most restriction enzymes recognize a single restriction site.
- The restriction site is recognized without regard to the source of the DNA.
- Because most restriction enzymes recognize a unique restriction-site sequence, the number of cuts in the DNA from a particular organism is determined by the number of restriction sites present.

The DNA fragment produced by a pair of adjacent cuts in a DNA molecule is called a **restriction fragment**. A large DNA molecule typically will be cut into many restriction fragments of different sizes. For example, an *E. coli* DNA molecule, which contains 4.6×10^6 base pairs, is cut into several hundred to several thousand fragments, and mammalian genomic DNA is cut into more than a million fragments. Most importantly, these fragments are not generated randomly; instead, their ends are determined by the presence of restriction sites in the DNA being digested. Therefore, in digested DNA, a sequence of interest should always occur in fragments with identical ends. This has three important implications:

1. Restriction sites can be used as genetic markers if sequence variation within particular sites exists in a population of individuals.
2. If a particular restriction fragment can be isolated, then the sequences contained within it can be further characterized.
3. Two fragments of DNA, generated by digestion with an enzyme like *Bam*H1, will have complementary ends that can (at least in theory) base-pair with each other. In fact, this base pairing can be readily accomplished, and it forms the basis for cloning of DNA shown in Figure 2.4 and described in the *Manipulating Genes and Genomes* chapter.

Gel Electrophoresis

So if we now have genomic DNA that has been digested, how do we separate the thousands of different fragments? They can be separated by size using the fact that DNA is negatively charged and moves in response to an electric field. If the terminals of an electrical power source are connected to the opposite ends of a horizontal tube containing a DNA solution, then the DNA molecules will move toward the positive end of the tube at a rate that depends on the electric field strength and on the shape and size of the molecules. The movement of charged molecules in an electric field is called *electrophoresis*.

The type of electrophoresis most commonly used in genetics is **gel electrophoresis**. An experimental arrangement for gel electrophoresis of DNA is shown in **FIGURE 2.14A**. A thin slab of a gel, usually agarose or acrylamide, is prepared containing small slots (called *wells*) into which samples are placed. An electric field is applied, and the negatively charged DNA molecules penetrate and move through the gel toward the anode (the positively charged electrode). A gel is a complex molecular network that contains narrow, tortuous passages, so smaller DNA molecules pass through more easily; hence the rate of movement increases as the size of the DNA fragment decreases. **FIGURE 2.14B** shows the result of electrophoresis of a set of double-stranded DNA molecules in an agarose gel. Each discrete region containing DNA is called a **band**. The bands can be visualized under ultraviolet light after soaking the gel in the dye *ethidium bromide*, the molecules of which intercalate into duplex DNA and render it fluorescent. In Figure 2.14B, each band in the gel results from the fact that all DNA fragments of a given size have migrated to the same position in the gel. To produce a visible band, a minimum of about 5×10^{-9} grams of DNA is required, which for a fragment of size 3 kb works out to approximately 10^9 molecules. The point is that a very large number of copies of any particular DNA fragment must be present to yield a visible band in an electrophoresis gel.

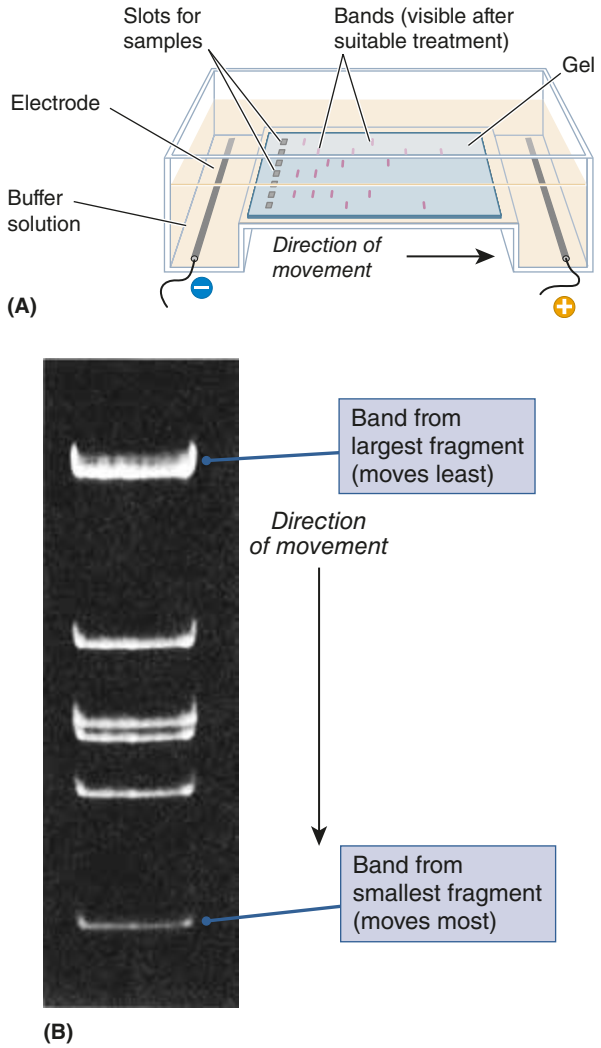


FIGURE 2.14 Gel electrophoresis of DNA. (A) Liquid gel is allowed to harden with an appropriately shaped mold in place to form slots for the samples. After electrophoresis, the DNA fragments, which are located at various positions in the gel, are made visible by immersing the gel in a solution containing a reagent that binds to or reacts with DNA. (B) After staining with a fluorescent dye (ethidium bromide), the separated fragments in a sample appear as bands when the gel is exposed to ultraviolet light.

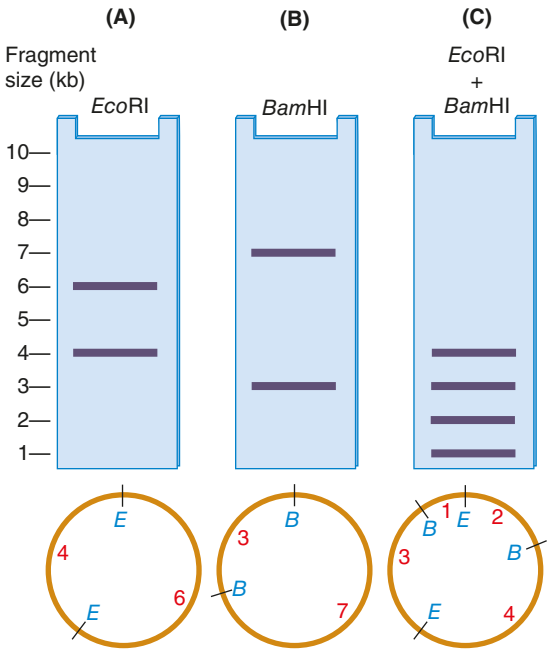


FIGURE 2.15 Gel diagrams showing the sizes of restriction fragments produced by digestion of a 10-kb circular molecule of double-stranded DNA with (A) *Eco*RI, (B) *Bam*HI, and (C) both enzymes together. Beneath each diagram is a restriction map of the circular DNA showing the locations of the restriction sites. The restriction map in (C) takes into account those in (A) and (B) as well as the fragment sizes produced by digestion with both enzymes.

Because of the sequence specificity of cleavage, *a particular restriction enzyme produces a unique set of fragments for a particular DNA molecule*. Another enzyme will produce a different set of fragments from the same DNA molecule. In **FIGURE 2.15**, this principle is illustrated for the digestion of a circular molecule of double-stranded DNA with a length of 10 kb. When digested with the restriction enzyme *Eco*RI (Figure 2.15A), the circular molecule yields bands of 4 kb and 6 kb. This pattern would result from *Eco*RI restriction sites located in the circle at the relative positions shown beneath the gel. The circle is oriented arbitrarily, with one of the *Eco*RI (*E*) sites at the top. Similarly, digestion of the circle with the enzyme *Bam*HI (Figure 2.15B) results in bands of 3 kb and 7 kb, which implies that the circle contains *Bam*HI sites at the positions indicated in the diagram beneath the gel. Again the circle is oriented arbitrarily, this time with one of the *Bam*HI (B) sites located at the top. A diagram showing sites of cleavage of one or more restriction sites along a DNA molecule is called a **restriction map**.

When both *Eco*RI and *Bam*HI are used together, the resulting DNA fragments reveal where the *Eco*RI sites and the *Bam*HI sites are located relative to each other. In this case, digestion with both enzymes yields bands of 1 kb, 2 kb, 3 kb, and 4 kb (Figure 2.15C). The restriction map shown beneath the gel indicates where the two types of restriction sites must be located to yield these band sizes. This restriction map can be obtained by superimposing that in part B over that in part A, and rotating the result until the distances between adjacent pairs of restriction sites equal 1, 2, 3, and 4 kb (not necessarily in that order). In this case, one need only rotate the restriction map in part B a distance of 2 kb to the right. Note that, in the restriction-enzyme digest in Figure 2.13C, the 4-kb fragment is not the same 4-kb fragment as observed in part A, and the 3-kb fragment is not the same 3-kb fragment as observed in part B. This discordance arises because each restriction enzyme cleaves the fragments produced by the other. The orientation of the restriction map in Figure 2.13C is arbitrary. It can be flipped over or rotated by any amount in any direction, and it will still be the same restriction map.

Construction of such a map facilitates two outcomes. First, the individual fragments can be isolated

from the gel, inserted into self-replicating molecules such as bacteriophage, plasmids, or even small artificial chromosomes (Figure 2.4), and introduced into bacterial cells (DNA cloning, a topic we will cover in the *Manipulating Genes and Genomes* chapter). Second, and most important with respect to DNA markers, we have taken the first step toward identifying particular positions in DNA—those at which the enzymes cleave. The question thus becomes how we can apply restriction analysis to complex genomes.

Nucleic Acid Hybridization

Most genomes are sufficiently large and complex that digestion with a restriction enzyme produces many bands that are the same or similar in size. Identifying a particular DNA fragment within a background of many other fragments of similar size presents a needle-in-a-haystack problem. Suppose, for example, that we are interested in a particular 3.0-kb *Bam*HI fragment from the human genome that serves as a marker indicating the presence of a genetic risk factor for breast cancer among women in a particular pedigree. On the basis of size alone, this fragment of 3.0 kb is indistinguishable from fragments ranging in size from about 2.9 to 3.1 kb. How many fragments in this size range are expected? When human genomic DNA is cleaved with *Bam*HI, the average length of a restriction fragment is $4^6 = 4096$ base pairs, and the expected total number of *Bam*HI fragments is about 730,000; in the size range 2.9–3.1 kb, the expected number of fragments is about 17,000. Thus, even though we know that the fragment of interest is 3 kb in length, it is only one of 17,000 fragments that are so similar in size that the sought-after fragment cannot be distinguished from the others by length alone. In fact, as seen in **FIGURE 2.16**, digested genomic DNA, when visualized, appears as a smear rather than as a set of discrete bands.

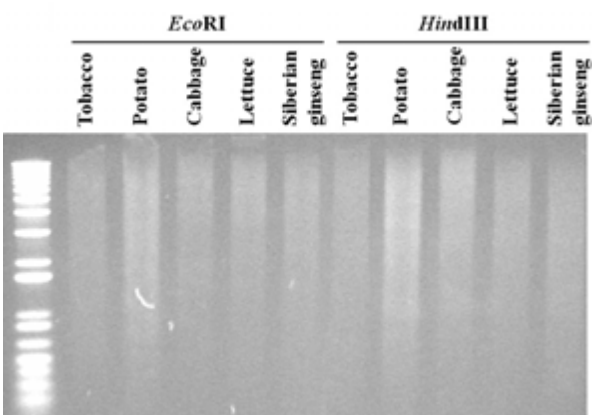


FIGURE 2.16 An example of genomic DNA from various plants digested with the enzymes EcoR1 and HindIII.
Reproduced from Kang, T. J., & Yang, M. S. Rapid and reliable extraction of genomic DNA from various wild-type and transgenic plants. *BMC Biotechnol.* 2004;4:20. doi:10.1186/1472-6750-4-20. Creative Commons license available: https://creativecommons.org/licenses/by/2.0/

This identification task is actually more difficult than finding a needle in a haystack because haystacks are usually *dry*. A more accurate analogy would be looking for a needle in a haystack that had been pitched into a swimming pool full of water. This analogy is more relevant because gels, even though they contain a supporting matrix to make them semisolid, are primarily composed of water, and each DNA molecule within a gel is surrounded entirely by water. Clearly, we need some method by which *specific* fragments can be either visualized or purified.

At this point, we need to return to the structure of DNA (Figure 2.8), examine how the two strands in a double helix can be separated to form single strands, and see how, under the proper conditions, two single strands that are complementary or nearly complementary in sequence can come back together to form a different double helix. The separation of the strands is called **denaturation**, and the coming together of complementary strands is called **nucleic acid hybridization** or **renaturation**. (The term *hybridization* is appropriate because the two strands that anneal to form a DNA duplex may not be exactly the same strands that were paired prior to denaturation.) The practical applications of nucleic acid hybridization are many:

- A small part of a DNA fragment can be hybridized with a much larger DNA fragment. This principle is used in identifying specific DNA fragments in a complex mixture, such as the 3-kb *Bam*HI marker for breast cancer that we have been considering. Applications of this type include tracking of genetic markers in pedigrees and isolation of fragments containing a particular mutant gene.
- A DNA fragment from one gene can be hybridized with similar fragments from other genes in the same genome; this principle is used to identify different members of *families* of genes that are similar, but not identical, in sequence and that have related functions.
- A DNA fragment from one species can be hybridized with similar sequences from other species. This allows the isolation of genes that have the same or related functions in multiple species. This method is used to study aspects of molecular evolution, such as how differences in sequence are correlated with differences in function, and the patterns and rates of change in gene sequences as they evolve.

As we saw in Section 2.3, The Molecular Structure of DNA, the double-stranded helical structure of DNA is maintained by base stacking and by hydrogen bonding between the complementary base pairs. When solutions containing DNA fragments are raised to temperatures in the range 85°C–100°C, or to the high pH of strong alkaline solutions, the paired strands begin

to separate. Unwinding of the helix happens in less than a few minutes (the time depends on the length of the molecule). A common way to detect denaturation is by measuring the capacity of DNA in solution to absorb ultraviolet light of wavelength 260 nm, because the absorption at 260 nm (A_{260}) of a solution of single-stranded molecules is 37 percent higher than the absorption of the double-stranded molecules at the same concentration. As shown in **FIGURE 2.17**, the progress of denaturation can be followed by slowly heating a solution of double-stranded DNA and recording the value of A_{260} at various temperatures. The temperature required for denaturation increases with G + C content, not only because G–C base pairs have three hydrogen bonds and A–T base pairs have two such bonds, but also because consecutive G–C base pairs have stronger base stacking.

For denatured DNA strands to be able to undergo renaturation, two requirements must be met:

1. The salt concentration must be high (greater than 0.25 M) to neutralize the negative charges of the phosphate groups, which would otherwise cause the complementary strands to repel each other.
2. The temperature must be high enough to disrupt hydrogen bonds that form at random between short sequences of bases within the same strand, but not so high that stable base pairs between the complementary strands are disrupted.

The initial phase of renaturation is a slow process because the rate is limited by the chance that a region of two complementary strands will come together at random to form a short sequence of correct base pairs. This initial pairing step is followed by a rapid pairing of the remaining complementary bases and rewinding of the helix. Rewinding is accomplished in a matter of seconds, and its rate is independent of DNA concentration because the complementary strands have already found each other.

Importantly, if two DNA molecules from different sources contain identical sequences and are allowed to reanneal together, then some of the renatured DNA will consist of **heteroduplex** DNA—that is, molecules containing base-paired strands that originated from different sources (**FIGURE 2.18**). If one DNA source is genomic DNA, and the other is a DNA **probe**, consisting of part of our sequence of interest that is suitably labeled (for example, with radioactive ^{32}P), then all of the heteroduplexes created will be radioactively labeled. Geneticists say that probe DNA hybridizes with DNA fragments containing sequences that are *similar*, rather than *complementary* to the probe.

This process can be used to address the problem of visualizing particular fragments in a conceptually simple way. If denatured DNA is immobilized on a membrane (typically made of nitrocellulose or nylon

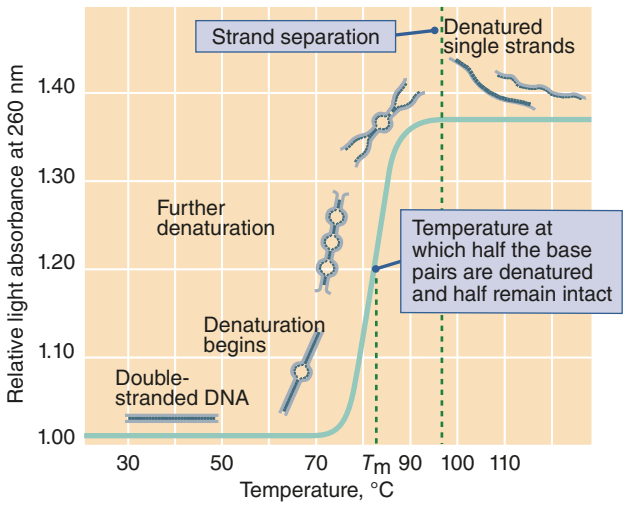


FIGURE 2.17 The mechanism of denaturation of DNA by heat. The temperature at which 50 percent of the base pairs are denatured is the *melting temperature*, symbolized as T_m.

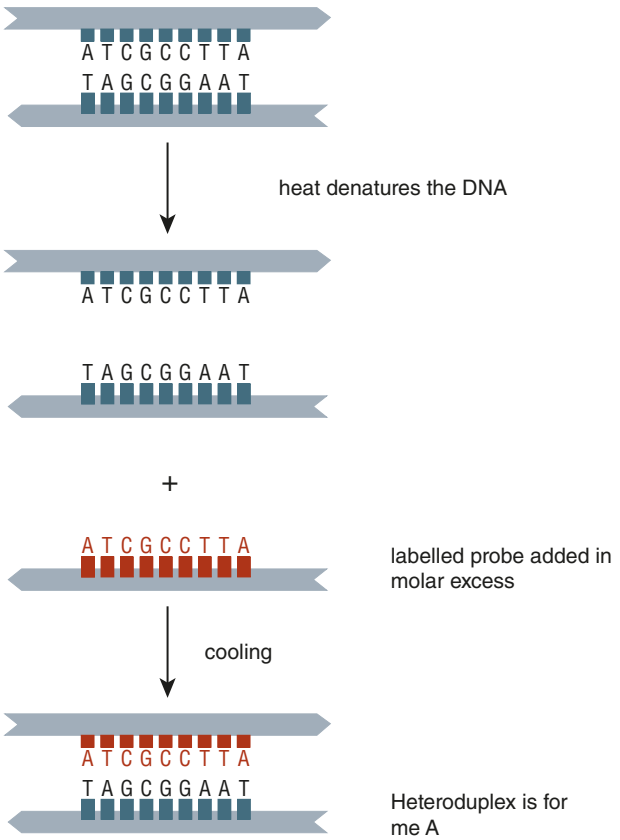


FIGURE 2.18 DNA detection by nucleic acid hybridization. Duplex genomic DNA is denatured with heat, after which it is combined with a labeled probe sequence homologous to a specific sequence in the genomic DNA. The mixture is cooled, allowing formation of heteroduplex molecules containing one strand of the genomic DNA base-paired with the labeled probe sequence.

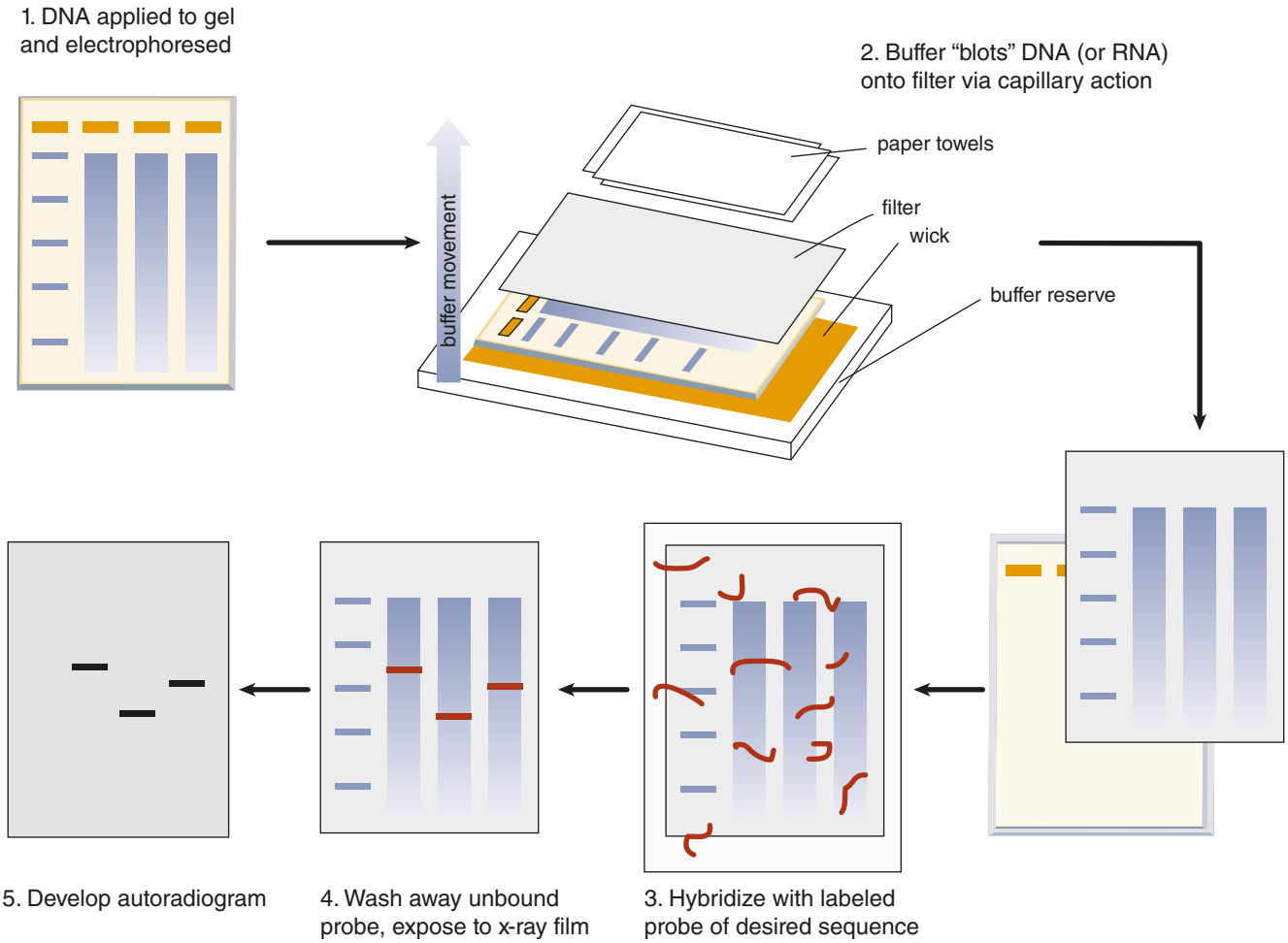


FIGURE 2.19 The Southern blot: an experimental procedure for identifying the position of a specific DNA fragment in a gel.

membrane), then that membrane can be incubated in a solution containing a radioactively labeled probe, consisting of copies of the sequence of interest. This process, which is shown in **FIGURE 2.19**, is referred to as a **Southern blot** (named after E. M. Southern, who developed the technique). In this process, digested genomic DNA is transferred from an agarose gel to a membrane. DNA on the membrane is then denatured, after which it is incubated with a denatured probe under conditions that allow renaturation and heteroduplex formation. The pattern of hybridization can then be visualized by exposure to x-ray film.

Returning to our original problem, that of testing DNA for a marker of a breast cancer risk allele in a pedigree, we could use this technique to analyze DNA from all of the individuals from whom we could obtain genomic DNA. In fact, this process is extremely sensitive: Under typical conditions, a band can be observed on the film that contains only 5×10^{-12} grams of DNA—a thousand times less DNA than the amount required to produce a visible band in the gel itself.

SUMMING UP

- Restriction enzymes cleave DNA at specific recognition sites, generating specific fragments.
- These fragments can be analyzed using gel electrophoresis.
- Restriction sites can be mapped on DNA molecules, and restriction fragments can be isolated for further study.
- DNA can be denatured and reannealed.
- Restriction analysis of genomic DNA can be performed by hybridizing electrophoretically separated DNA with labeled specific probes.

2.5 Amplification of Specific DNA for Detection and Purification

Despite its sensitivity, Southern blotting has its limitations. It requires a relatively large amount of genomic DNA, and in some applications (for example, analysis of

trace amounts of DNA at a crime scene or from ancient remains), that kind of sample is not available. In addition, while it allows a particular DNA fragment to be identified when present in a complex mixture of fragments, it does not enable the fragment to be separated from the others and purified. Obtaining the fragment in purified form requires cloning, which is straightforward but time consuming. (Cloning methods are discussed in the *Manipulating Genes and Genomes* chapter.) However, if the fragment of interest is not too long, and if the nucleotide sequence at each end is known, then it becomes possible to obtain large quantities of the fragment merely by selective replication. This process is called **amplification**. Once this is accomplished, the amplified fragment can be analyzed directly. In fact, these procedures have largely supplanted hybridization methods in genetic screening protocols.

How would one know the nucleotide sequences at the ends of interest? Let us return to our example of the 3.0-kb *Bam*HI fragment that serves to mark a risk allele for breast cancer in certain pedigrees. Suppose that this fragment is cloned and sequenced from one affected individual, and it is found that, relative to the sequence in unaffected individuals, the *Bam*HI fragment is missing a region of 500 base pairs. At this point, the sequences at the ends of the fragment are known, and we can also infer that amplification of genomic DNA from individuals with the risk factor will yield a band of 3.0 kb, whereas amplification from genomic DNA of noncarriers will yield a band of 3.5 kb. This difference allows every person in the pedigree to be diagnosed as a carrier or noncarrier merely by means of DNA amplification. To understand how amplification works, it is first necessary to examine a few key features of DNA replication.

Constraints on DNA Replication: Primers and 5′-to-3′ Strand Elongation

As was the case with restriction analysis, amplification utilizes a biological process—in this case, DNA replication. In doing so, it uses the enzyme that forms the sugar–phosphate bond (the phosphodiester bond) between adjacent deoxynucleotides in a DNA chain, called a **DNA polymerase**. A variety of DNA polymerases have been purified, and for amplification of a DNA fragment, the DNA synthesis is carried out in vitro by combining purified cellular components in a test tube under precisely defined conditions. (*In vitro*, literally "in glass," implies the absence of living cells.)

For DNA polymerase to catalyze synthesis of a new DNA strand, preexisting single-stranded DNA must be present. Each single-stranded DNA molecule present in the reaction mix can serve as a template upon which a new partner strand is created by the DNA polymerase. For DNA replication to take place, the 5′-triphosphates of the four deoxynucleoside triphosphates must also be present. This requirement is rather obvious, because these are the precursors from which new DNA strands are created. The triphosphates needed are the compounds denoted in Table 2.1 as dATP, dGTP, dTTP, and dCTP, which contain the bases adenine, guanine, thymine, and cytosine, respectively. Details of the structures of dCTP and dGTP are shown in **FIGURE 2.20**, in which the phosphate groups cleaved off during DNA synthesis are indicated. DNA synthesis requires all four nucleoside 5′-triphosphates and does not take place if any of them are omitted.

One feature of all DNA polymerases is that a DNA polymerase can only *elongate* a DNA strand. It is not possible for DNA polymerase to *initiate* synthesis of a new strand, even when a template molecule is present. An important implication of this principle is that DNA synthesis requires a preexisting segment of nucleic acid that is hydrogen-bonded to the template strand; this segment is called a **primer**. Because the primer molecule can be very short, it is an **oligonucleotide**, which literally means "few nucleotides." As we shall see in the *DNA Replication and Sequencing* chapter, in living cells the primer is a short segment of RNA; in contrast, in DNA amplification in vitro, the primer employed is usually DNA.

It is the 3′ end of the primer that is essential, because DNA synthesis proceeds only by addition of successive nucleotides to the 3′ end of the growing strand. In other words, chain elongation always takes place in the 5′-to-3′ direction (5′ → 3′). The reason for the 5′ → 3′ direction of chain elongation is illustrated in

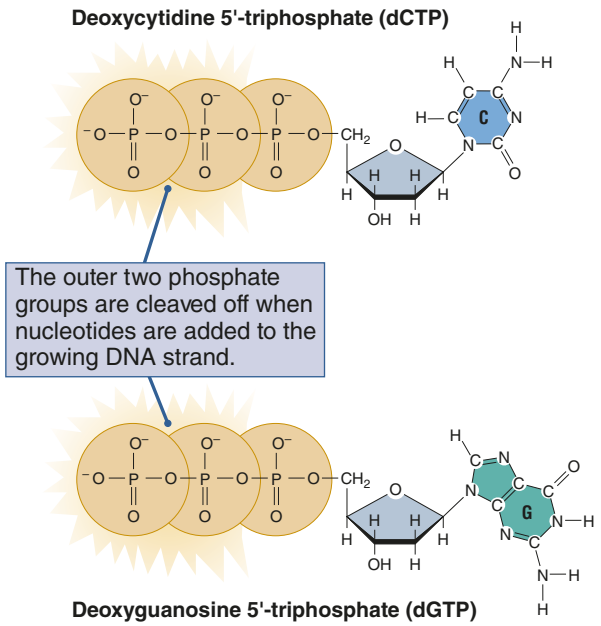


FIGURE 2.20 Two deoxynucleoside triphosphates used in DNA synthesis. The outer two phosphate groups are removed during synthesis.

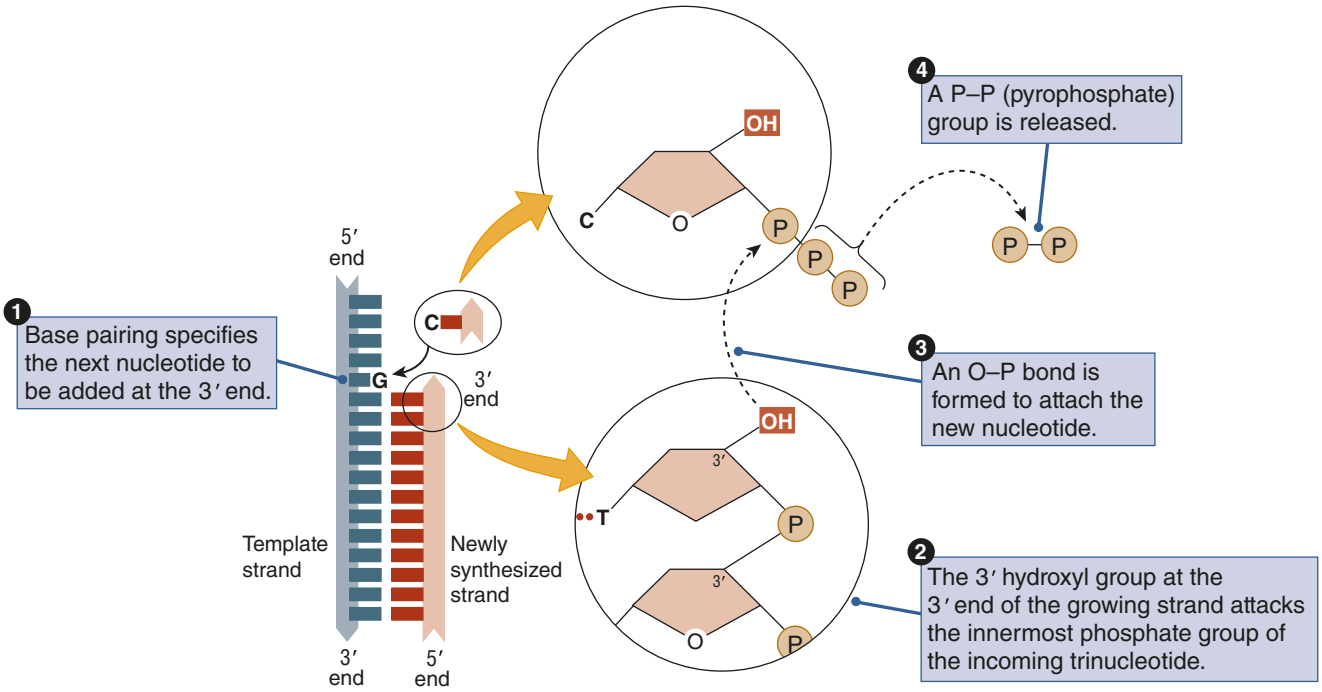


FIGURE 2.21 Addition of nucleotides to the 3′-OH terminus of a growing strand. The recognition step is shown as the formation of hydrogen bonds between the G and the C. The chemical reaction is that the 3′-OH group of the 3′ end of the growing chain attacks the innermost phosphate group of the incoming trinucleotide.

FIGURE 2.21: The reaction catalyzed by DNA polymerase is the formation of a phosphodiester bond between the free 3′-OH group of the chain being extended and the innermost phosphorus atom of the nucleoside triphosphate being incorporated at the 3′ end. Recognition of the appropriate incoming nucleoside triphosphate in replication depends on base pairing with the opposite nucleotide in the template strand. DNA polymerase will usually catalyze the polymerization reaction that incorporates the new nucleotide at the primer terminus only when the correct base pair is present. The same DNA polymerase is used to add each of the four deoxynucleoside phosphates to the 3′-OH terminus of the growing strand.

The Polymerase Chain Reaction

The requirement for an oligonucleotide primer, and the constraint that chain elongation must always occur in the 5′ → 3′ direction, make it possible to obtain large quantities of a particular DNA sequence by selective amplification in vitro. The method for selective amplification is called the **polymerase chain reaction (PCR)**. For its invention, Californian Kary B. Mullis was awarded a Nobel Prize in 1993. PCR amplification uses DNA polymerase and a *pair* of short, synthetic oligonucleotide primers, usually 18–22 nucleotides in length, that are complementary in sequence to the ends of the DNA sequence to be amplified. **FIGURE 2.22** gives an example in which the primer oligonucleotides (green) are 9-mers. These are too short for most practical purposes, but they will serve for illustration. The original duplex molecule (part A) is shown in blue. This duplex is mixed with a vast excess of primer molecules, DNA polymerase, and all four nucleoside triphosphates. When the temperature is raised, the strands of the duplex denature and become separated. When the temperature is lowered again to allow renaturation, the primers, because they are in great excess, hybridize (or *anneal*) to the separated template strands (part B). Note that the primer sequences are different from each other but complementary to sequences present in opposite strands of the original DNA duplex and flanking the region to be amplified. The primers are oriented with their 3′ ends pointing in the direction of the region to be amplified, because each DNA strand elongates only at the 3′ end. After the primers have annealed, each is elongated by DNA polymerase using the original strand as a template, and the newly synthesized DNA strands (red) grow toward each other as synthesis proceeds (part C). Note that a region of duplex DNA present in the original reaction mix can be PCR-amplified only if the region is flanked by the primer oligonucleotides.

To start a second cycle of PCR amplification, the temperature is raised again to denature the duplex DNA. Upon lowering of the temperature, the original parental strands anneal with the primers and are replicated as shown in Figure 2.22B and C. The daughter strands produced in the first round of amplification also anneal with primers and are replicated, as shown in part D. In this case, although the daughter duplex molecules are identical in sequence to the original parental molecule, they consist entirely of primer oligonucleotides and nonparental DNA that was synthesized in

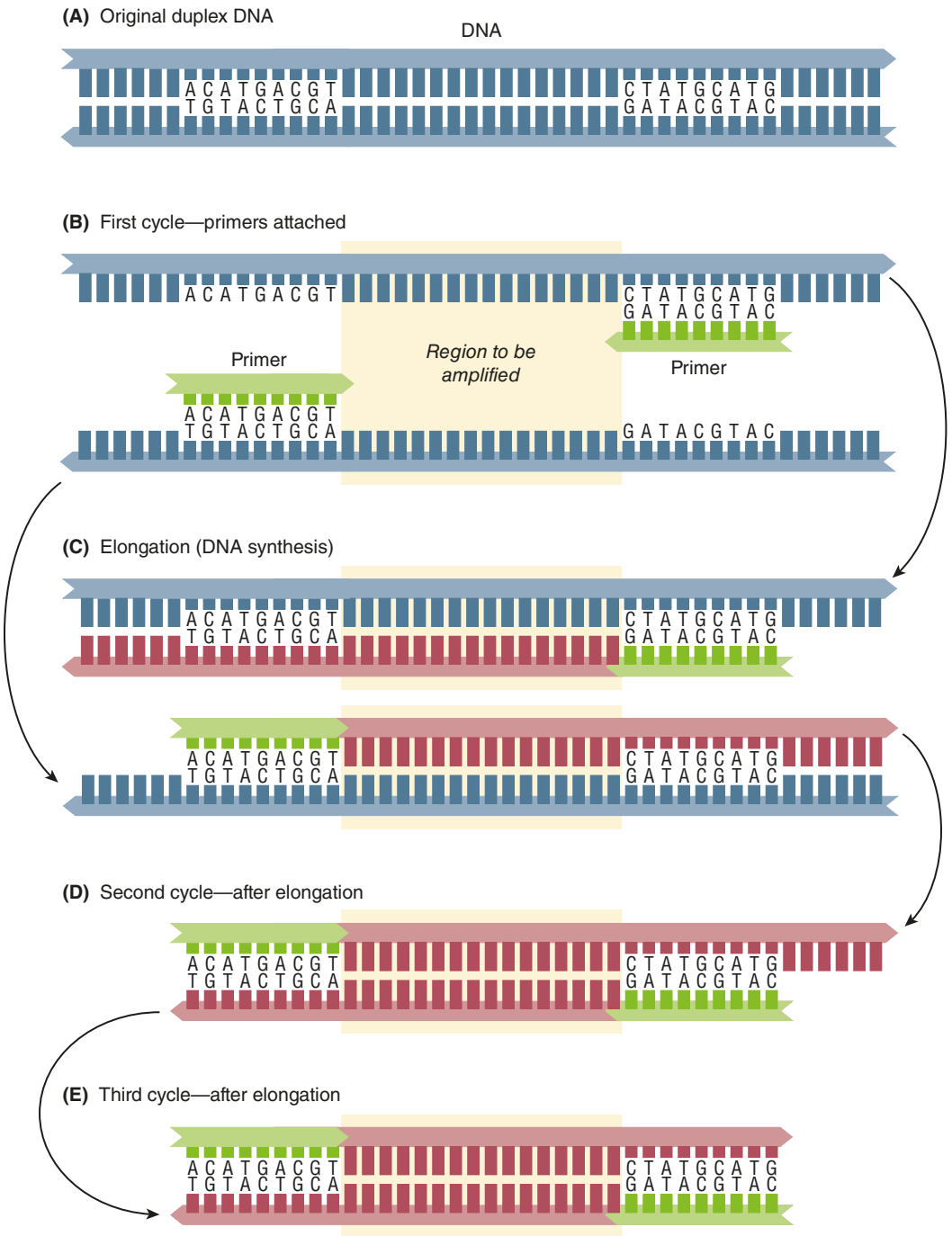


FIGURE 2.22 Role of primer sequences in PCR amplification. (A) Target DNA duplex (blue), showing sequences chosen as the primer-binding sites flanking the region to be amplified. (B) Primer (green) bound to denatured strands of target DNA. (C) First round of amplification. Newly synthesized DNA is shown in pink. Note that each primer is extended *beyond* the other primer site. (D) Second round of amplification (only one strand shown); in this round, the newly synthesized strand terminates at the opposite primer site. (E) Third round of amplification (only one strand shown); in this round, both strands are truncated at the primer sites. Primer sequences are normally at least twice as long as shown here.

either the first or the second cycle of PCR. As successive cycles of denaturation, primer annealing, and elongation occur, the original parental strands are diluted out by the proliferation of new daughter strands until eventually almost every molecule produced in the PCR has the structure shown in part E.

The power of PCR amplification derives from the fact that the number of copies of the template strand increases in exponential progression: 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, and so forth, doubling with each cycle of replication. Starting with a mixture containing as little as one molecule of the fragment of interest, repeated rounds of DNA replication increase the number of amplified molecules exponentially. For example, starting with a single molecule, 25 rounds of DNA replication will result in $2^{25} = 3.4 \times 10^7$ molecules. This number of molecules of the amplified fragment is so much greater than that of the other unamplified molecules in the original mixture that the amplified DNA can often be used without further purification. For example, a single 3-kb fragment in *E. coli* accounts for only 0.06 percent of the DNA in this organism. However, if this single fragment were replicated through 25 rounds of replication, then 99.995 percent of the resulting mixture would consist of the amplified sequence. A 3-kb fragment of human DNA constitutes only 0.0001 percent of

the total genome size. Amplification of a 3-kb fragment of human DNA to 99.995 percent purity would require approximately 34 cycles of PCR.

FIGURE 2.23 provides an overview of the polymerase chain reaction. The DNA sequence to be amplified is again shown in blue and the oligonucleotide primers in green. The oligonucleotides anneal to the ends of the sequence to be amplified and become the substrates for chain elongation by DNA polymerase. In the first cycle of PCR amplification, the DNA is denatured to separate the strands. The denaturation temperature is usually around 95°C. The temperature is then decreased to allow annealing in the presence of a vast excess of the primer oligonucleotides. The annealing temperature is typically in the range of 50°C–60°C, depending largely on the G + C content of the

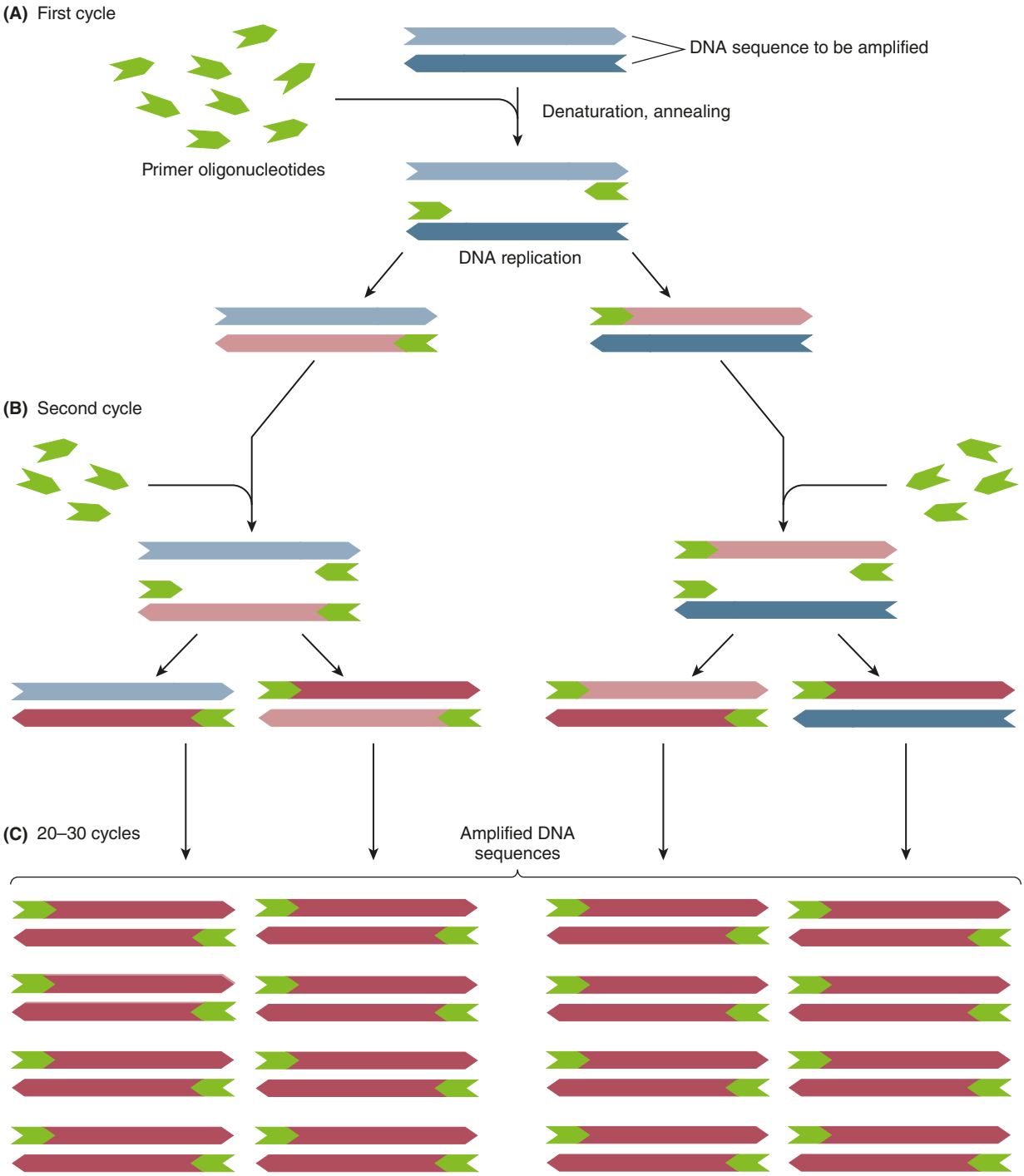


FIGURE 2.23 Polymerase chain reaction (PCR) for amplification of particular DNA sequences. Oligonucleotide primers (green) that are complementary to the ends of the target sequence (blue) are used in repeated rounds of denaturation, annealing, and DNA replication. Newly replicated DNA is shown in pink. The number of copies of the target sequence doubles in each round of replication, eventually overwhelming any other sequences that may be present.

oligonucleotide primers. To complete the cycle, the temperature is raised slightly, to about 70°C, for the elongation of each primer. The steps of denaturation, renaturation, and replication are repeated from 20 to 30 times, and in each cycle the number of molecules of the amplified sequence is doubled.

Implementation of PCR with conventional DNA polymerases is not practical, because at the high temperature necessary for denaturation, the polymerase is itself irreversibly unfolded and becomes inactive. However, DNA polymerase isolated from certain types of organisms is heat stable because those organisms normally live in hot springs at temperatures well above 90°C, such as are found in Yellowstone National Park. Such organisms are said to be **thermophiles**. The most widely used heat-stable DNA polymerase is called *Taq* polymerase, because it was originally isolated from the thermophilic bacterium *Thermus aquaticus*.

PCR amplification is very useful for generating large quantities of a specific DNA sequence. The principal limitation of the technique is that the DNA sequences at the ends of the region to be amplified must be known so that primer oligonucleotides can be synthesized. In addition, sequences longer than about 5000 base pairs cannot be replicated efficiently by conventional PCR procedures, although some modifications of PCR allow longer fragments to be amplified. Conversely, many applications require amplification of relatively small fragments. The major advantage of PCR amplification is that it requires only trace amounts of template DNA. Theoretically, only one template molecule is required, but in practice the amplification of a single molecule may fail because the molecule may, by chance, be broken or damaged. In practice, amplification is usually reliable with as few as 10–100 template molecules, which makes PCR amplification 10,000–100,000 times more sensitive than detection via nucleic acid hybridization.

With PCR, genotyping individuals is greatly simplified. Rather than having to go through all the steps involved in Southern blotting, a small amount of genomic DNA can be used in combination with appropriate primers to amplify the sequence of interest. Returning to our example of a breast cancer risk marker, **FIGURE 2.24** illustrates what we would observe if we genotyped two individuals, one homozygous and one heterozygous for the allele of interest. Furthermore, the exquisite sensitivity of PCR amplification has led to its use in DNA typing for criminal cases in which a minuscule amount of biological material has been left behind by the perpetrator (skin cells on a cigarette butt or hair-root cells on a single hair can yield enough template DNA for amplification).

In research, PCR is widely used in the study of independent mutations in a gene whose sequence is known so as to identify the molecular basis of each mutation, to study DNA sequence variations among alternative forms of a gene that may be present in natural populations, or to examine differences among genes with the same function in different species. The PCR procedure has also come into widespread use in clinical laboratories for diagnosis. To take just one very important example, the presence of the human immunodeficiency virus (HIV), which causes acquired immunodeficiency syndrome (AIDS), can be detected in trace quantities in blood banks via PCR by using primers complementary to sequences in the viral genetic material. These and other applications of PCR are facilitated by the fact that the procedure lends itself to automation by the use of mechanical robots to set up and run the reactions.

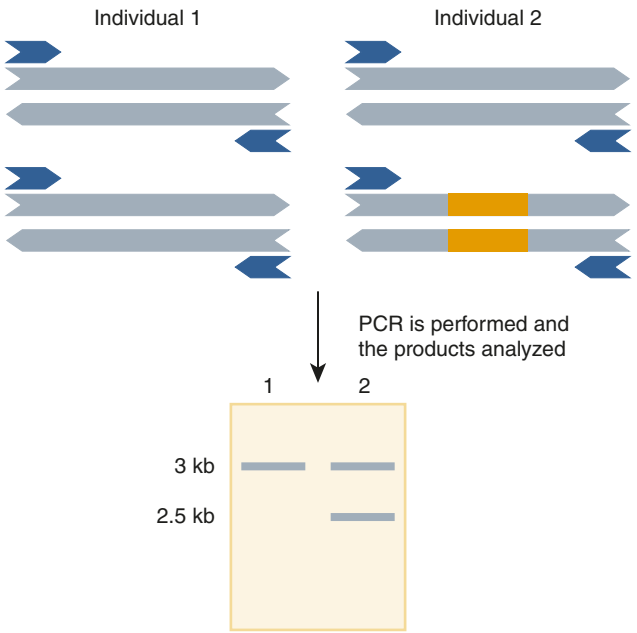


FIGURE 2.24 Determining a DNA marker genotype based on PCR. Shown are the genotypes of two individuals. Individual 1 is homozygous; individual 2 is heterozygous for two alleles that differ in length, due to a deletion in the second allele (shown in orange). The PCR products can be directly analyzed without resorting to blotting, resulting in one band from individual 1 and two bands from individual 2.

SUMMING UP

- Specific DNA fragments contained in a complex mixture like genomic DNA can be selectively amplified with the polymerase chain reaction (PCR).
- PCR requires the presence of template DNA, nucleoside triphosphates, DNA polymerase, and two oligonucleotide primers that flank the region to be amplified.
- PCR requires repeated rounds of DNA synthesis, denaturation, and primer reannealing. This is accomplished using a thermocycler and a heat-stable DNA polymerase.
- Use of PCR greatly simplifies the genotyping of DNA markers, and it can be used for many other applications, including pathogen detection and forensic DNA analysis.

2.6 Types of DNA Markers Present in Genomic DNA

Genetic variation, in the form of polymorphisms, exists in most natural populations of organisms. The methods of DNA manipulation examined in Sections 2.4 and 2.5 can be used in a variety of combinations to detect differences among individuals. Anyone who reads the literature in modern genetics will encounter a bewildering variety of acronyms referring to different ways in which genetic polymorphisms are detected. The different approaches are in use because no single method is ideal for all applications, each method has its own advantages and limitations, and new methods are continually being developed. In this section, we examine some of the principal methods for detecting DNA polymorphisms among individuals.

Restriction Fragment Length Polymorphisms

The simplest kind of genetic polymorphism is a **single-nucleotide polymorphism (SNP)** (pronounced "snip"). This is simply a position in the genome where, within a population, two or more different bases are found in different genomes. We now know that SNPs are ubiquitous: The 1000 Genomes Project has cataloged more than 80 million of them in the human genome. However, prior to the development of modern high-throughput genotyping, detection of SNPs and other forms of sequence variation was difficult. Restriction enzymes proved to be valuable tools in elucidating their presence.

We have already seen that restriction enzymes can be used to map the occurrence of particular sequences (the recognition sites) in DNA molecules. If a recognition site contains a DNA polymorphism, then digestion can be used as a means of detecting it. An example is shown in **FIGURE 2.25**, in which a T–A nucleotide pair appears in some molecules and a C–G pair in others. In this example, the polymorphic nucleotide site is included in a cleavage site for the restriction enzyme *Eco*RI (5′-GAATTC-3′). The two nearest flanking *Eco*RI sites are also shown. In this kind of situation, DNA molecules with T–A nucleotide pairs will be cleaved at both flanking sites as well as at the middle site, yielding two *Eco*RI restriction fragments. In contrast, DNA molecules with C–G nucleotide pairs will be cleaved at both flanking sites but not at the middle site (because the presence of C–G destroys the *Eco*RI restriction site), so they will yield only one larger restriction fragment. A SNP that eliminates a restriction site is known as a **restriction fragment length polymorphism (RFLP)** (pronounced either as "riflip" or by spelling it out).

In the early days of analysis of genomic DNA, prior to the development of automated DNA sequencing (see the *DNA Replication and Sequencing* chapter), RFLP analysis was an extremely important tool for mapping chromosomes and detecting DNA polymorphisms, since such polymorphisms could be detected by Southern blotting. With the advent of PCR, however, the process could be greatly simplified. As shown in **FIGURE 2.26**, primers that flank the RFLP site can be used to amplify the DNA containing it, and then digestion and simple gel electrophoresis can be used to analyze the product. Given that a fragment that has been amplified can also be sequenced, comparison of products from different alleles can then be extended to include all positions in the fragment.

Single-Nucleotide Polymorphisms

RFLP analysis can detect only those allelic variants that affect the pattern of digestion of the enzyme(s) used, which are only a small fraction of the vast number of allelic differences that exist in genomic DNA. If, instead of digesting the fragments, their sequences are determined completely, then SNPs can be detected. A SNP is

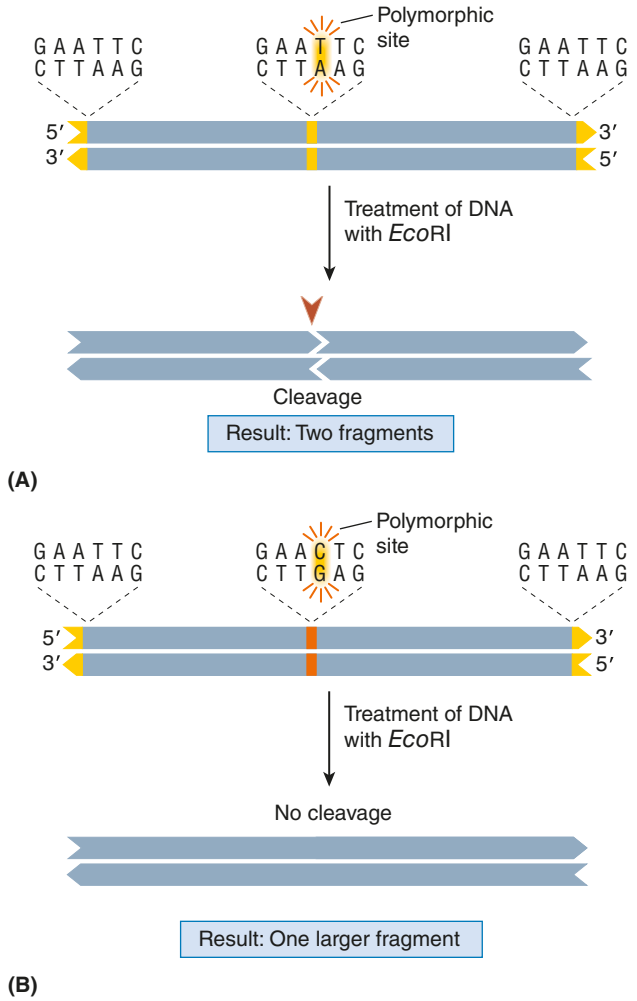


FIGURE 2.25 A difference in the DNA sequence of two molecules can be detected if the difference eliminates a restriction site. (A) This molecule contains three restriction sites for *Eco*RI, including one at each end. It is cleaved into two fragments by the enzyme. (B) This molecule has an altered *Eco*RI site in the middle, in which 5′-GAATTC-3′ becomes 5′-GAACTC-3′. The altered site cannot be cleaved by *Eco*RI, so treatment of this molecule with *Eco*RI results in one larger fragment.

a particular nucleotide site in DNA that differs among individuals with respect to the identity of the nucleotide pair that occupies the site. For example, some DNA molecules may have a T–A base pair at a particular nucleotide site, whereas other DNA molecules in the same population may have a C–G base pair at the same site. This difference constitutes a SNP. The SNP defines two alleles for which there could be three genotypes among individuals in the population: homozygous with T–A at the corresponding site in both homologous chromosomes, homozygous with C–G at the corresponding site in both homologous chromosomes, or heterozygous with T–A in one chromosome and C–G in the homologous chromosome. In the human genome, any two randomly chosen DNA molecules are likely to differ at one SNP site about every 1000 bp in noncoding DNA and at one SNP site about every 3000 bp in protein-coding DNA.

Note, in the discussion of a SNP, the stipulation is that DNA molecules must differ at the nucleotide site "often." This provision excludes rare genetic variations of the sort found in less than 1 percent of the DNA molecules in a population. Extremely rare genetic variants are not generally as useful in genetic analysis as the more common variants.

SNPs are the most common form of genetic differences among people. Approximately 10 million SNPs

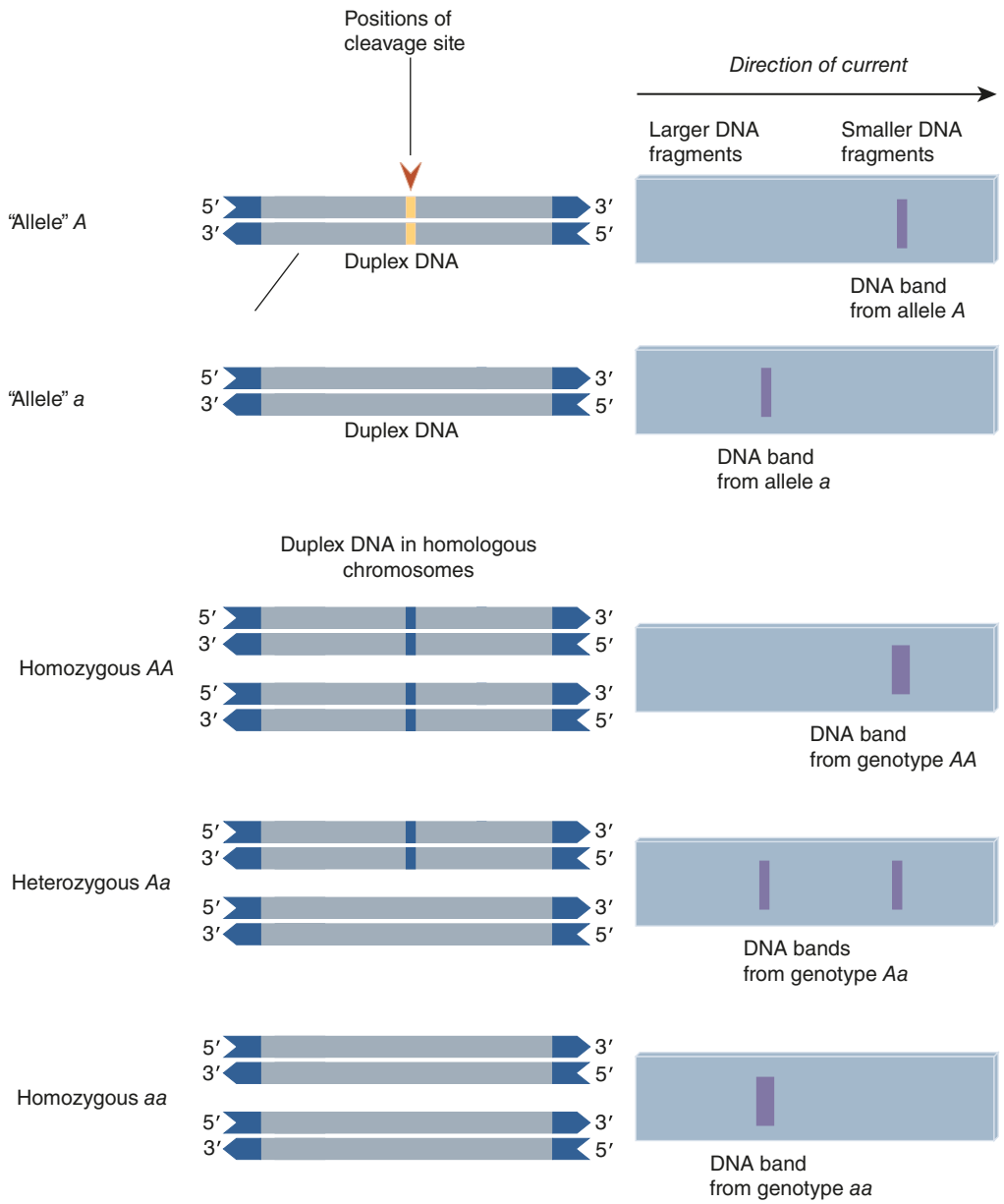


FIGURE 2.26 In a restriction fragment length polymorphism (RFLP), alleles may differ in the presence or absence of a cleavage site in the DNA. This example shows a PCR fragment containing two single-base-pair alleles. Primer sequences are shown in blue. The *a* allele lacks a restriction site that is present in the DNA of the *A* allele. The difference in fragment length can be detected by digestion of PCR products containing the cleavage site. RFLP alleles are codominant, which means that DNA from the heterozygous *Aa* genotype yields each of the single bands observed in DNA from homozygous *AA* and *aa* genotypes.

have been identified that are relatively common in the human population, and 300,000–600,000 of these are typically used in a search for SNPs that might be associated with complex diseases such as diabetes or high blood pressure (see the *Genetic Linkage and Chromosome Mapping* and *The Genetic Basis of Complex Traits* chapters).

Identifying the particular nucleotide present at each of a million SNPs is made possible through the use of **DNA microarrays** containing millions of infinitesimal spots on a glass slide about the size of a postage stamp. Each tiny spot contains a unique DNA oligonucleotide sequence present in millions of copies synthesized by microchemistry when the microarray is manufactured. Each oligonucleotide sequence is designed to hybridize specifically with small fragments of genomic DNA that include one or the other of the nucleotide pairs present in a SNP. The microarrays also include numerous controls for each hybridization. The controls consist of oligonucleotides containing deliberate mismatches that are intended to guard against being misled by particular nucleotide sequences that are particularly "sticky" and hybridize too readily with genomic fragments and other sequences that form structures that hybridize poorly or not at all. Such microarrays, sometimes called **SNP chips**, enable the SNP genotype of an individual to be determined with nearly 100 percent accuracy.

The principles behind oligonucleotide hybridization are illustrated in **FIGURE 2.27**. In the figure, the length of each oligonucleotide is seven nucleotides; in practice, this is too short, as typical SNP chips consist of oligonucleotides at least 25 nucleotides in length. Part A shows the two types of DNA duplexes that might form a SNP. In this example, some chromosomes carry a DNA molecule with a T–A base pair at the position shown in red, whereas the DNA molecular in other chromosomes has a C–G base pair at the corresponding position. Short fragments of genomic DNA are labeled with a fluorescent tag, and then the single strands are hybridized with a SNP chip containing the complementary oligonucleotides as well as the numerous controls. The duplex containing the T–A will hybridize only with the two oligonucleotides on the left, and that containing the C–G will hybridize only with the two oligonucleotides on the right.

After the hybridization takes place, the SNP chip is examined with fluorescence microscopy to detect the spots that fluoresce due to the tag on the genomic DNA. The possible patterns are shown in part B. Genomic DNA from an individual whose chromosomes contain

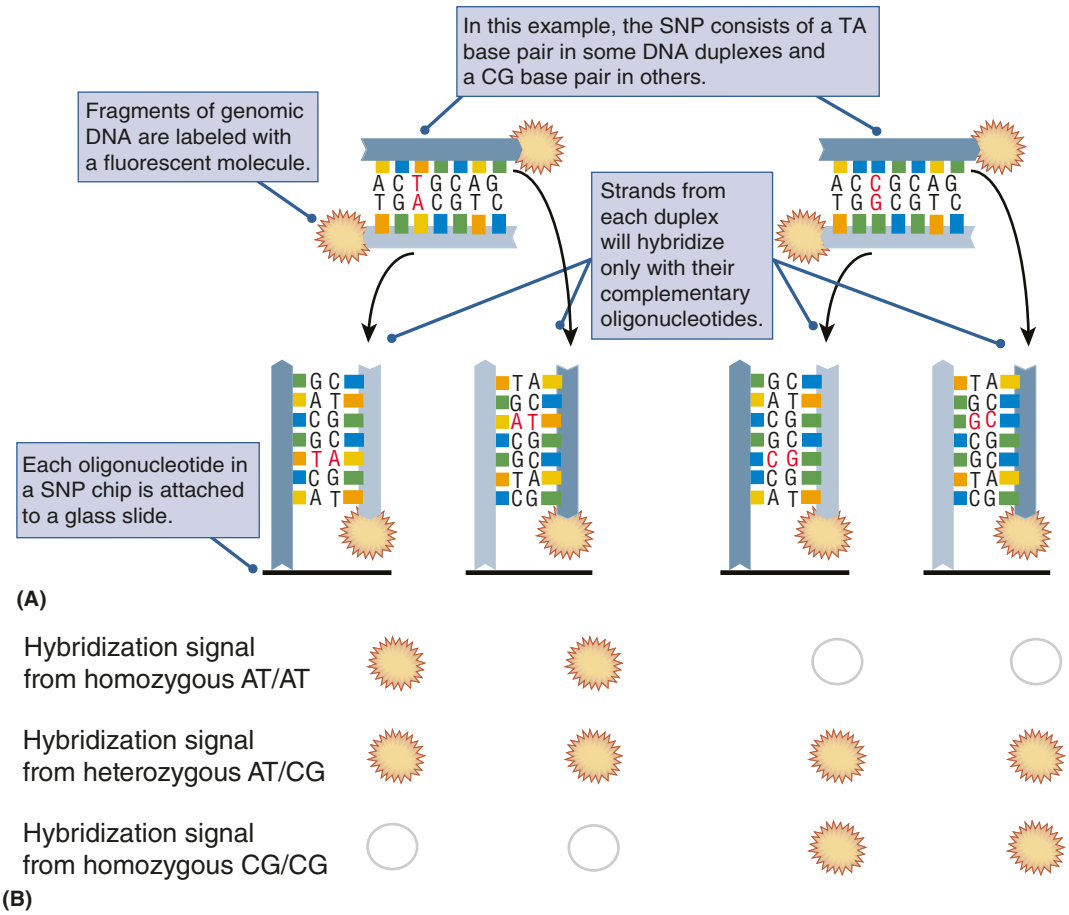


FIGURE 2.27 (A) Oligonucleotides attached to a glass slide in a SNP chip can be used to identify duplex DNA molecules containing alternative base pairs for a SNP—in this example, a T–A base pair versus a C–G base pair. (B) The SNP genotype of an individual can be determined by hybridization because DNA samples from genotypes that are homozygous TA/TA, homozygous CG/CG, or heterozygous TA/CG all give different patterns of fluorescence.

two copies of the TA form of the duplex (homozygous TA/TA) will cause the two leftmost spots to fluoresce, but the two rightmost spots will remain unlabeled. Similarly, genomic DNA from a homozygous CG/CG individual will cause the two rightmost spots to fluoresce, but not the two spots on the left. Finally, genomic DNA from a heterozygous TA/CG individual will cause fluorescence of all four spots, because the TA duplex labels the two leftmost spots and the CG duplex labels the two rightmost spots.

Use of SNP chips or other available technologies for high-throughput genotyping of millions of SNPs in thousands of individuals allows genetic risk factors for disease to be identified. A typical study compares the genotypes of patients with particular diseases with the genotypes of healthy people matched with the patients on such factors as sex, age, and ethnic group. Comparing the SNP genotypes among these groups often reveals which SNPs in the genome mark the location of the genetic risk factors, as will be explained in greater detail in Section 2.7.

Tandem Repeat Polymorphisms

An important type of DNA polymorphism results from differences in the number of copies of a DNA sequence that may be repeated many times in tandem at a particular site in a chromosome. Any particular chromosome may have any number of copies of the tandem repeat, typically ranging from ten to a few hundred. **FIGURE 2.28** illustrates DNA molecules that differ in

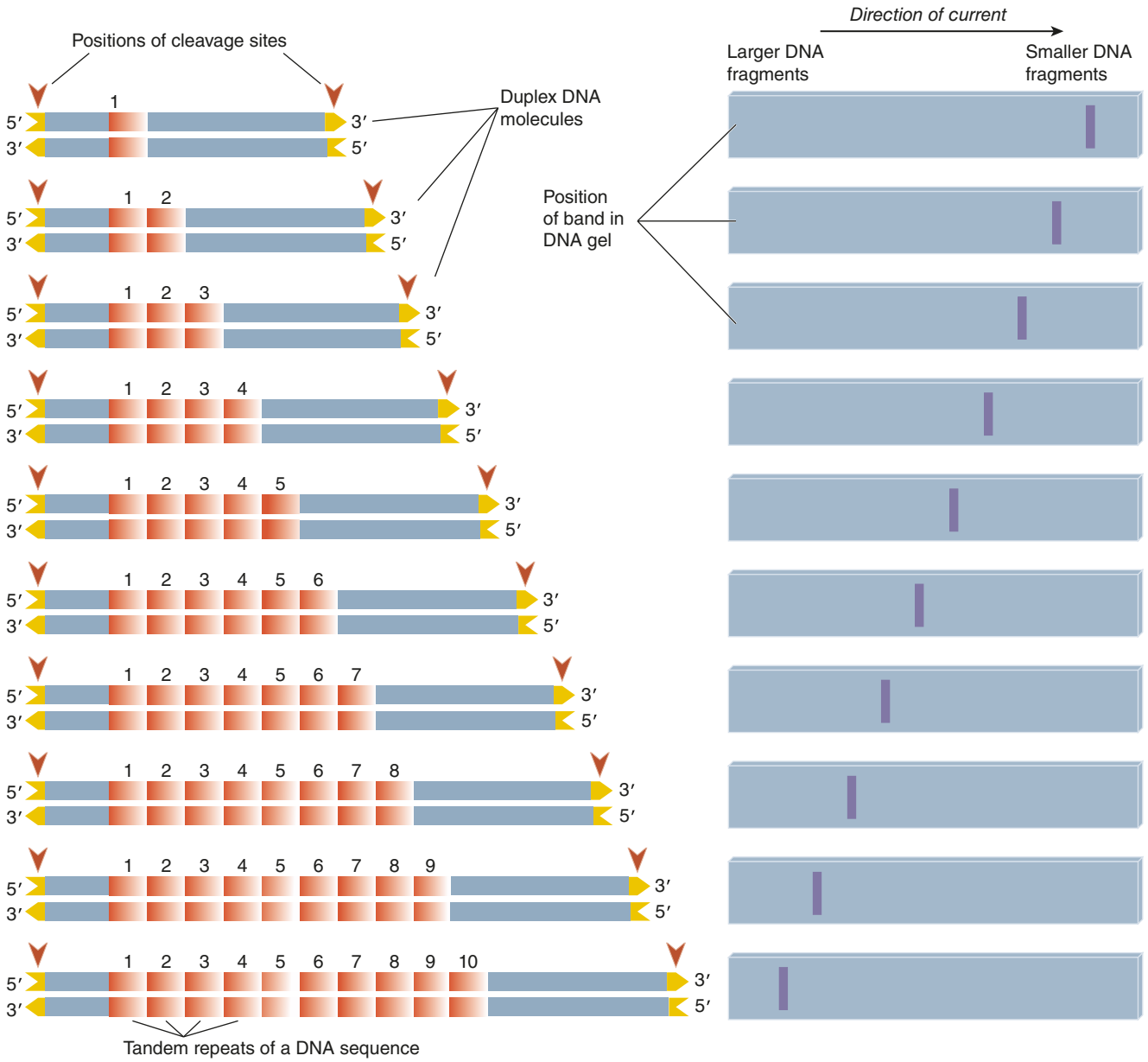


FIGURE 2.28 A genetic polymorphism in which the alleles in a population differ in the number of copies of a DNA sequence (typically 2–60 bp) that is repeated in tandem along the chromosome. This example shows alleles in which the repeat number varies from 1 to 10. Amplification using primers flanking the repeat yields a unique fragment length for each allele.

THE CUTTING EDGE: High-Throughput SNP Genotyping

Naturally occurring enzymes are often used for critical steps in molecular genetic analysis. Using restriction enzymes and DNA polymerase for DNA marker analysis and PCR are two examples, and the tremendous diversity of species (especially microbial) on Earth continues to be a source of new tools.

SNP detection is a good example. The basic SNP chip approach, as shown in Figure 2.27, depends on the availability of sufficient genomic DNA to be analyzed; it also requires accurate hybridization of genomic DNA fragments to oligonucleotides on the chip. Both of these limitations have been addressed by the latest techniques.

To increase the quantity of genomic DNA, the technique of PCR-free whole-genome amplification has been developed. PCR by itself has two problems. First, it requires thermal cycling for the denaturation, renaturation, and elongation steps. Second, the *Taq* polymerase (the one most commonly used in PCR) is prone to errors: An incorrect base is inserted every 800 bases or so. To get around these problems, scientists have developed a technique for amplification of whole genomes called **multiple displacement amplification**. As illustrated in **FIGURE A**, random six-base sequences (hexamers) are used as primers. The DNA polymerase comes from the bacteriophage φ29, an enzyme that replicates DNA with high fidelity. Furthermore, when it encounters the 5′ end of a strand that is base-paired to its template, the DNA polymerase displaces that strand; the resulting single strand can then hybridize with another of the hexamer primer sequences present in the reaction, thereby serving as the template for further replication. By this means, total genomic DNA can be amplified from a single cell, and the process occurs without thermal cycling (no denaturation and renaturation is necessary).

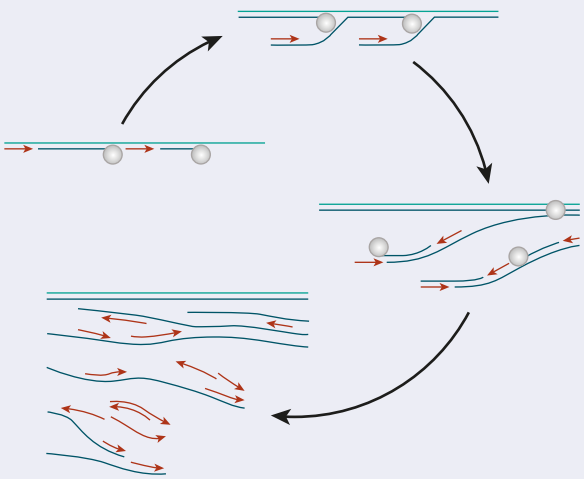


FIGURE A

Once this phase is complete, the amplified DNA is sheared into small fragments, denatured, and hybridized to oligonucleotides (**FIGURE B**). Multiple copies of different oligonucleotides have been attached to microbeads, with their 3′ ends free, and those beads have been embedded in a solid matrix. However, the SNP to be assayed is *not* part of the sequence included in the oligonucleotide; rather, the SNP is the *next* base in the sequence after the oligonucleotide's 3′ end. DNA polymerase is added, along with modified nucleoside triphosphates that can be detected fluorescently, under conditions where one base will be added to the oligonucleotide probe. The identity of that base depends on the SNP allele in the hybridized genomic DNA, and it can be determined by scanning the fluorescence on the chip. In the case illustrated in Figure B the individual being assayed is homozygous for the A allele of SNP 1, heterozygous for the A and G alleles of SNP 2, and homozygous for the G allele of SNP 3.

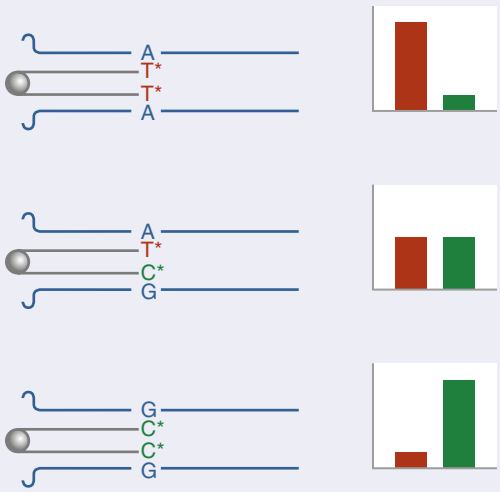


FIGURE B

These kinds of cutting-edge methods make SNP genotyping fast and efficient with even only a very small amount of starting material (such as the saliva samples used by the personal genomics company 23andMe), and they can generate enormous amounts of data. In fact, as many as 48 samples can be genotyped for 300,000 or more SNPs in as little as three days. These methods do require some sophisticated instrumentation (much of it can be done robotically), but they remain fundamentally grounded in the biological processes of base pairing and DNA replication.

the number of tandem repeats. In this case, the number of repeats varies from 1 to 10.

When any of the DNA molecules is cleaved with a restriction endonuclease that cleaves at sites flanking the tandem repeat, the size of the resulting restriction fragment is determined by the number of repeats that it contains. A molecule of duplex DNA containing the repeats can also be amplified by means of the polymerase chain reaction, using primers flanking the tandem repeat. Whether obtained by endonuclease digestion or PCR, the resulting DNA fragments increase in size according to the number of repeats they contain. Therefore, as shown at the right in Figure 2.28, two DNA molecules that differ in the number of copies of the tandem repeat can be distinguished because each will produce a different-sized DNA fragment that can be separated by means of amplification of that fragment.

The acronym **SSR** stands for **simple sequence repeat** (these sequences are also known as **microsatellite loci**). An example of an SSR is the repeating sequence 5′-…ACACACAC…-3′. SSRs are important in genetic analysis for two reasons:

- SSRs are abundant in the genomes of most eukaryotes.
- SSRs are often highly polymorphic.

To illustrate their abundance in the human genome, the most common SSRs are tabulated in **TABLE 2.3**. There are many more dinucleotide repeats than trinucleotide repeats, but there are great differences in abundance within each class. The most common dinucleotide repeat is 5′-…ACACACAC…-3′, which is present at more than 80,000 locations throughout the human genome. On average, the human genome has one SSR in every 2 kb of human DNA, or about 1.5 million SSRs altogether.

Not only do SSRs tend to be polymorphic, but most polymorphisms tend to have a large number of alleles. Each "allele" corresponds to a DNA molecule with a different number of copies of the repeating unit, as illustrated in Figure 2.28. In other words, a polymorphism in the number of tandem repeats usually has **multiple alleles** in the population. Even with multiple alleles, however, any particular chromosome must carry only one of the alleles defined by the number of tandem repeats, and any individual genotype can carry at most two different alleles. Nevertheless, a large number of alleles implies an even larger number of genotypes. For example, even with only 10 alleles in a population, there could be 10 different homozygous genotypes and 45 different heterozygous genotypes.

More generally, with n alleles there are a total of $n(n + 1)/2$ possible genotypes, of which n are homozygous and $n(n - 1)/2$ are heterozygous. Consequently, if an SSR has a relatively large number of alleles in a population, but no one allele is exceptionally common, then each of the many genotypes in the population will have a relatively low frequency. If the genotypes at 6–8 highly polymorphic loci are considered simultaneously, then each possible multiple-locus genotype is exceedingly rare. The very low frequency of any multiple-locus genotype is what gives tandem-repeat polymorphisms their utility in **DNA typing** (sometimes called **DNA fingerprinting**) for individual identification and for assessing the degree of genetic relatedness between individuals. In forensic DNA profiling, for example, genotypes are determined for 13 SSR loci that are distributed throughout the human genome; the probability that two individuals (other than unrelated twins) will have identical genotypes at all 13 loci is less than one in a billion. The use of DNA typing in criminal investigation is exemplified in Problem 2.24 and Challenge Problem 2 at the end of this chapter and discussed further in the *Genes in Populations* chapter.

Technological advances have done much to facilitate DNA typing. To this point, we have considered only electrophoretic analyses performed manually. In fact, SSR genotype is routinely performed using automated systems, in which multiple loci can be genotyped in a single reaction. In this process (**FIGURE 2.29**), PCR is performed using multiple primers simultaneously, and the primers are labeled with chemical tags that can be detected fluorescently. The resulting products are then separated electrophoretically in a gel contained in a capillary tube. As the molecules emerge from the tube, they are detected by a

TABLE 2.3 Some simple sequence repeats in the human genome

SSR repeat unit	Number of SSRs in the human genome
5′-AC-3′	80,330
5′-AT-3′	56,260
5′-AG-3′	23,780
5′-GC-3′	290
5′-AAT-3′	11,890
5′-AAC-3′	7,540
5′-AGG-3′	4,350
5′-AAG-3′	4,060
5′-ATG-3′	2,030
5′-CGG-3′	1,740
5′-ACC-3′	1,160
5′-AGC-3′	870
5′-ACT-3′	580

Data from International Human Genome Sequencing Consortium, *Nature*. 2001; 409: 860–921.

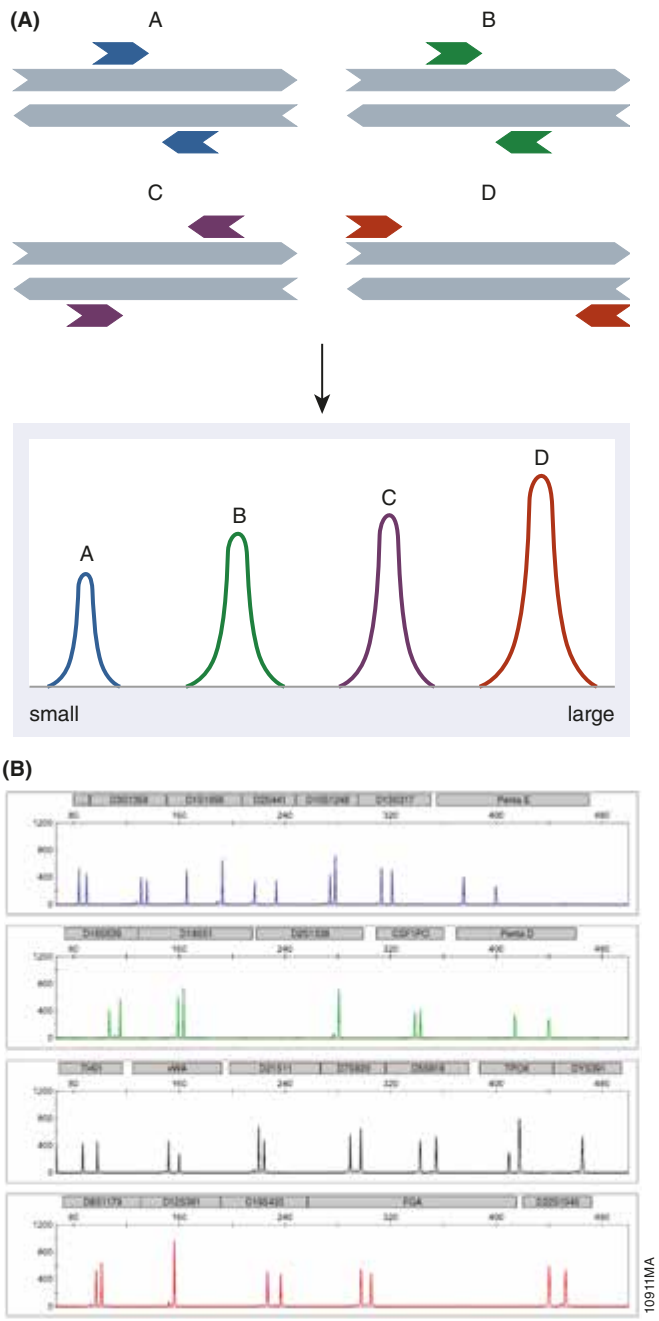


FIGURE 2.29 SSR genotyping with automated multiplexed PCR. (A) Schematic diagram showing four loci being amplified, each with primers carrying different fluorescent labels (blue, green, purple, and red) and producing different-sized fragments. (B) An example of SSR genotyping results. In this case, 24 different sets of SSR locus primers were used in combination with four different fluorescent markers. Shaded boxes indicate the names of the loci. If an individual is homozygous (in this case, D2S1338 and D12S391), a single peak is observed. If an individual is heterozygous, two peaks are seen. (B) Courtesy of Promega Corporation. Retrieved from http://www.promega.com/~/media/images/resources/figures/10900-10999/10911ma_800px.jpg?h=650&la=en&w=800.

photodetector and their sizes recorded. Figure 2.29B shows the results of such an analysis. In this case, 24 SSR loci were genotyped from a starting sample of 1.2 picograms of human genomic DNA—comparable to the amount of DNA that can be obtained from a single cheek swab.

Copy-Number Variation

In addition to the small-scale variation in copy number represented by such genomic features as tandem repeats of short sequences (SSRs), a substantial portion of the human genome can be duplicated or deleted in much larger but still submicroscopic chunks ranging from 1 kb to 1 Mb (Mb stands for *megabase pairs*, or 1 million base pairs). This type of variation is known as **copy-number variation (CNV)**. The extra or missing copies of the genome in CNVs can be detected by means of hybridization with oligonucleotides in DNA microarrays. Because each spot on the microarray consists of millions of identical copies of a particular oligonucleotide sequence, the number of these sequences that undergo hybridization depends on how many copies of the complementary sequence are present in genomic DNA. A typical region is present in two copies (one inherited from the mother and the other from the father). If an individual has an extra copy of the region, the ratio of hybridization and therefore fluorescence intensity will be 3 : 2; by comparison, if an individual has a missing copy, the ratio of hybridization and therefore fluorescence intensity will be 1 : 2. These differences can readily be detected with DNA microarrays. Moreover, because CNVs are relatively large, a microarray typically will include many different oligonucleotides that are complementary to sequences at intervals across the CNV; hence, the CNV will result in an increase or decrease in signal intensity of all the oligonucleotides included in the CNV. Current SNP chips also include about 1 million oligonucleotide probes designed to detect known CNVs.

CNVs, by definition, exceed 1 kb in size, but many are much larger. In one study of approximately 300 individuals with ancestry tracing to Africa, Europe, or Asia, approximately 1500 CNVs were discovered by hybridization with microarrays. These averaged 200–300 kb in length. In the aggregate, the CNVs included 300–450 million base pairs, or 10–15 percent of the nucleotides in the entire genome. Many of the CNVs were located in regions near known mutant genes associated with hereditary diseases. CNVs in alpha and beta hemoglobin genes are known to be associated with resistance to malaria, and CNVs in the HIV-1 receptor gene *CCL3* are associated with resistance to AIDS. CNVs have been reported to be risk factors for complex diseases such as Alzheimer's disease, autism, and schizophrenia.

2.7 Applications of DNA Markers

Why are geneticists interested in DNA markers and DNA polymorphisms? Their interest can be justified on any number of grounds. In this section, we consider the reasons most often cited.

Genetic Markers, Genetic Mapping, and "Disease Genes"

Perhaps the key goal in studying DNA polymorphisms in human genetics is to identify the chromosomal location of

genes having mutant alleles associated with hereditary diseases. In the context of disorders caused by the interaction of multiple genetic and environmental factors, such as heart disease, cancer, diabetes, depression, and so forth, it is important to think of a harmful allele as a risk factor for the disease that increases the probability of occurrence of the disease, rather than as the sole causative agent. This point needs to be emphasized, especially because genetic risk factors are often called **disease genes**.

For example, a major "disease gene" for breast cancer in women is the gene *BRCA1*. For women who carry a mutant allele of *BRCA1*, the lifetime risk of breast cancer is about 60 percent. By comparison, among women who are not carriers, the lifetime risk of breast cancer is about 12 percent. Hence, many women without the genetic risk factor do develop breast cancer. Indeed, *BRCA1* mutations are found in only 16 percent of affected women who have a family history of breast cancer.

The importance of a genetic risk factor can be expressed quantitatively as the **relative risk**, which equals the ratio of affected persons to nonaffected persons among those individuals who carry the risk factor divided by the ratio of affected persons to nonaffected persons among those individuals who do not carry it. In the case of *BRCA1*, for example, the relative risk is (0.60/0.40 = 1.5) divided by (0.12/0.88 = 0.136), which equals 11.

The utility of DNA polymorphisms in locating and identifying disease genes results from **genetic linkage**, or the tendency for genes that are sufficiently close together in a chromosome to be inherited together. To see how this works, refer back to Figure 2.1. In that hypothetical example, the DNA marker is not actually in the *PAH* gene, but rather one marker allele is *associated* with the disease allele as a result of genetic linkage. Genetic linkage is discussed in detail in the *Genetic Linkage and Chromosome Mapping* chapter, but the key concepts are summarized in **FIGURE 2.30**, which shows the location of many DNA polymorphisms along a chromosome that also carries a genetic risk factor denoted *D* (for disease gene). Each DNA polymorphism serves as a genetic marker for its own location in the chromosome. The importance of genetic linkage is that alleles of DNA markers that are sufficiently close to the disease gene will tend to be inherited together with the disease gene in pedigrees—and the closer the markers, the stronger this association. Hence, the initial approach to the identification of a disease gene is to find DNA markers that are genetically linked with the disease gene so as to identify its chromosomal location, a procedure known as **genetic mapping**. Once the chromosomal position is known, other methods can be used to pinpoint the disease gene itself and to study its functions.

If genetic linkage seems a roundabout way to identify disease genes, consider the alternative. The human genome contains approximately 30,000 genes. If genetic linkage did not exist, then we would have to examine 30,000 DNA polymorphisms, one in each gene, to identify a disease gene. But the human genome has only 23 pairs of chromosomes, and because of genetic linkage and the power of genetic mapping, it actually requires only a few hundred DNA polymorphisms to identify the chromosome and approximate location of a genetic risk factor.

Other Uses for DNA Markers

DNA polymorphisms are widely used in all aspects of modern genetics because they provide a large number of easily accessed genetic markers for genetic mapping and other purposes. Some other uses of DNA polymorphisms are discussed in this section.

Individual identification. We have already mentioned that DNA polymorphisms have application as a means of DNA typing (DNA fingerprinting) to identify different individuals in a population. DNA typing in other organisms is used to determine individual animals in endangered species and to identify the degree of genetic relatedness among individual organisms

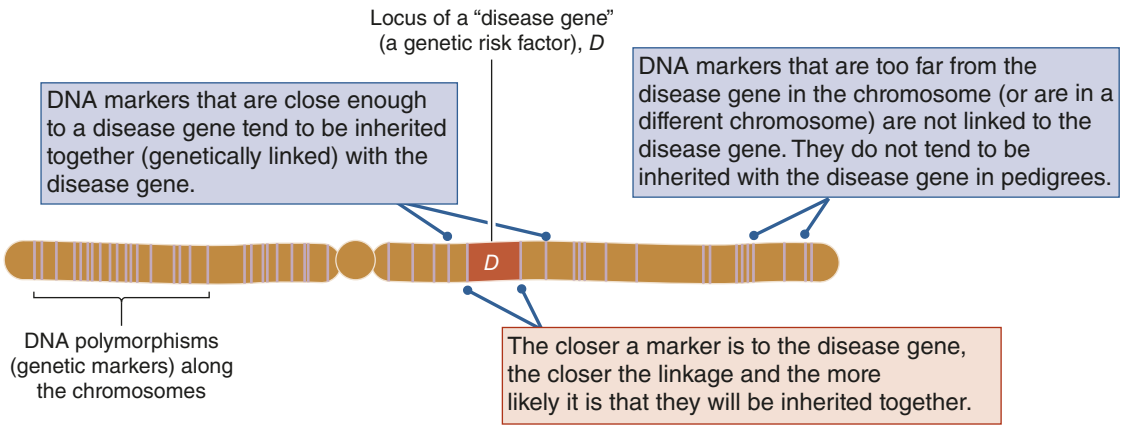


FIGURE 2.30 Concepts in genetic localization of genetic risk factors for disease. Polymorphic DNA markers (indicated by the vertical lines) that are close to a genetic risk factor (*D*) in the chromosome tend to be inherited together with the disease itself. The genomic location of the risk factor is determined by examining the known genomic locations of the DNA polymorphisms that are linked with it.

that live in packs or herds. For example, DNA typing in wild horses has shown that the wild stallion in charge of a harem of mares actually sires fewer than one-third of the foals.

Epidemiology and food safety science. DNA typing has important applications in tracking the spread of viral and bacterial epidemic diseases, as well as in identifying the source of contamination in contaminated foods.

Human population history. DNA polymorphisms provide important information that enables anthropologists to reconstruct the evolutionary origin, global expansion, and diversification of the human population. We will consider this issue in depth in the *Molecular and Human Evolutionary Genetics* chapter.

Improvement of domesticated plants and animals. Plant and animal breeders have turned to DNA polymorphisms as genetic markers in pedigree studies to identify, by genetic mapping, genes that are associated with favorable traits. These genes may then be incorporated into currently used varieties of plants and breeds of animals.

History of domestication. Plant and animal breeders also study genetic polymorphisms to identify the wild ancestors of cultivated plants and domesticated animals, as well as to infer the practices of artificial selection that led to genetic changes in these species during domestication.

DNA polymorphisms as ecological indicators. DNA polymorphisms are being evaluated as biological indicators of genetic diversity in key indicator species present in biological communities exposed to chemical, biological, or physical stress. They are also used to monitor genetic diversity in endangered species and species bred in captivity.

Evolutionary genetics. DNA polymorphisms are studied in an effort to describe the patterns in which different types of genetic variation occur throughout the genome, to infer the evolutionary mechanisms by which genetic variation is maintained, and to illuminate the processes by which genetic polymorphisms within species become transformed into genetic differences between species.

Population studies. Population geneticists employ DNA polymorphisms to assess the level of genetic variation in diverse populations of organisms that differ in genetic organization (prokaryotes, eukaryotes, organelles), population size, breeding structure, or life-history characteristics. They use genetic polymorphisms within subpopulations of a species as indicators of population history, patterns of migration, and so forth.

Evolutionary relationships among species. Differences in DNA sequences between species serve as the basis of *molecular phylogenetics*, in which the sequences are analyzed to determine the ancestral history (phylogeny) of the species and to trace the origin of morphological, behavioral, and other types of adaptations that have arisen in the course of evolution.

SUMMING UP

- Analysis of RFLPs was an important tool for initial mapping of the human genome.
- Single-nucleotide polymorphisms occur, on average, in one of every 1000 bases, so they provide a wealth of potential genetic markers.
- DNA microarray technology is used to simultaneously genotype individuals at hundreds of thousands of SNP loci.
- Simple sequence repeats are also ubiquitous in the human genome and are widely used in DNA forensics.
- Use of DNA markers for genetic analysis has been applied to areas ranging from epidemiology to evolutionary genetics.

CHAPTER SUMMARY

- A DNA strand is a polymer of A, T, G, and C deoxyribonucleotides joined in the 3′-to-5′ direction by phosphodiester bonds.
- The two DNA strands in a duplex are held together by hydrogen bonding between the A–T and G–C base pairs and by base stacking of the paired bases.
- Each type of restriction endonuclease enzyme cleaves double-stranded DNA at a particular sequence of bases that is usually four or six nucleotides in length.
- The DNA fragments produced by a restriction enzyme can be separated by electrophoresis, isolated, sequenced, and manipulated in other ways.
- Separated strands of DNA or RNA that are complementary in nucleotide sequence can come together (hybridize) spontaneously to form duplexes.
- DNA replication takes place only by elongation of the growing strand in the 5′-to-3′ direction through the addition of successive nucleotides to the 3′ end.
- In the polymerase chain reaction, short oligonucleotide primers are used in successive

cycles of DNA replication to amplify selectively a particular region of a DNA duplex.

- Genetic markers in DNA provide a large number of easily accessed sites in the genome that can be used to identify the chromosomal locations of disease genes, for DNA typing in individual identification, for the genetic improvement of cultivated plants and domesticated animals, and for many other applications.

REVIEW THE BASICS

- Define and give an example of each of the following genetic terms: locus, allele, genotype, heterozygous, homozygous, phenotype.
- What is a DNA marker? Explain how harmless DNA markers can serve as aids in identifying disease genes through genetic mapping.
- Which four bases are commonly found in the nucleotides in DNA? Which form base pairs?
- Which chemical groups are present at the extreme 3′ and 5′ ends of a single polynucleotide strand?
- What does it mean to say that a single strand of DNA strand has a polarity? What does it mean to say that the DNA strands in a duplex molecule are antiparallel?
- What are restriction enzymes, and why are they important in the study of particular DNA fragments? What does it mean to say that most restriction sites are palindromes?
- Describe how a Southern blot is carried out. What is it used for? What is the role of the probe?
- How does the polymerase chain reaction work? What is it used for? Which information about the target sequence must be known in advance? What is the role of the oligonucleotide primers?
- What is a DNA microarray? How is one used for SNP genotyping? What are the possible sources of error in the process?
- What is a DNA marker? Explain how harmless DNA markers can serve as aids in identifying disease genes through genetic mapping.

GUIDE TO PROBLEM SOLVING

PROBLEM 1 A geneticist plans to use the polymerase chain reaction (PCR) to amplify part of the DNA sequence shown below, using oligonucleotide primers that hybridize in the regions marked in red. (These are illustrative only and too short to be used in practice.) Specify the sequence of the primers that should be used, including the polarity, and deduce the sequence of the amplified DNA fragment.

5′-GCGAAACGATCCTCATCCTGTCTCTTGATCAGAGCTTGATCCCCTG-3′
3′-CGCTTTGCTAGGAGTAGGACAGAGAACTAGTCTCGAACTAGGGGAC-5′

ANSWER The primers should be able to pair with the chosen primer sites and must be oriented with their 3′ ends facing each other. Thus the "forward" primer (the one that is elongated in the left-to-right direction) should have the sequence 5′-ACGAT-3′, and the "reverse" primer (the one that is elongated in the right-to-left direction) should have the sequence 3′-ACTAG-5′. Because the convention for writing nucleic acid sequences is to put 5′ end on the left, the reverse primer is 5′-GATCA-3′.

PROBLEM 2 The genetic material of the bacteriophage M13 is a single-stranded DNA molecule of 6407 nucleotides, whose complete sequence can be found in publicly accessible databases such as GenBank. Upon M13 DNA being introduced into a bacterial cell, a complementary DNA strand is synthesized by bacterial enzymes, resulting in a double-stranded replicative form of the phage genome. How would you determine whether both strands of the replicative form are transcribed within an infected cell?

ANSWER To determine whether both DNA strands are templates for RNA synthesis, one could isolate the replicative form of the phage DNA, separate the two strands, and test each strand separately for the ability to hybridize with mRNA found in M13-infected cells. Alternatively (and less definitively), one could examine the genome sequence of M13 in GenBank and search both DNA strands for open reading frames that could code for proteins. As it happens, only one strand of M13 contains such open reading frames, and hence only one DNA strand is likely to be transcribed.

PROBLEM 3 A plasmid of 6200 base pairs includes at least three complete copies of a viral gene arranged in tandem. Cleavage of the plasmid with the restriction enzyme *Hind*III results in fragments of 900, 1300, and 4000 bp. The tandem copies of the gene are contained within only one of the *Hind*III fragments. The gene encodes a protein of 405 amino acids. Which *Hind*III fragment is likely to contain the gene copies?

ANSWER To encode a protein of 405 amino acids requires at least 1215 bp. Three tandem copies of the gene would therefore require 3645 bp, and four copies 4860 bp. The smallest *Hind*III fragment is not long enough to contain

even one copy of the gene. The 1300-bp fragment could contain one copy but not two or more. Only the largest fragment could include three copies of the gene, and it could not include more than three copies.

PROBLEM 4 Factor V is a protein that is involved in the blood clotting cascade. Men of European ancestry who are heterozygous for a particular rare allele of the *F5* gene that encodes it have a 52 percent chance of developing venous thromboembolism, a blood clot forming in a vein in the leg. In the population as a whole, 12 percent of all similar men develop this problem. What is the risk factor associated with this particular genotype?

ANSWER If 52 percent of men with the risk-associated genotype develop the problem, then 100 percent − 52 percent, or 48 percent, do not. Thus, we can calculate the ratio of men with this genotype who do and do not develop the condition as 52/48, or 1.08. Next, we need to be able to characterize the risk of venous thromboembolism occurring regardless of genotype. We do the same calculation for the population as a whole, where 12 percent of men develop the condition and 88 percent do not. The corresponding ratio is thus 12/88, or 0.136. Finally, we can determine the increased risk associated with F5 genotype as

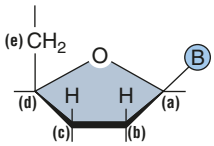
$$(52/48)/(12/88) = 1.08/.136 = 7.9$$

Thus, the risk of developing venous thromboembolism for a European man heterozygous for the risk-associated *F5* allele is 7.9 times higher than it is for the European male population as a whole.

ANALYSIS AND APPLICATIONS

2.1 Which chemical groups are present at the 3′ and 5′ ends of a single-polynucleotide molecule?

2.2 In the deoxyribonucleotide shown below, which carbon atom carries the phosphate group and which carries the 3′ hydroxyl group?



2.3 Many restriction enzymes produce restriction fragments that have "sticky ends." What does this mean?

2.4 Which of the following sequences are palindromes and which are not? Explain your answer.

(a) 5′-CCGG-3′
(b) 5′-TTTT-3′
(c) 5′-GCTAGC-3′
(d) 5′-CCGCTC-3′
(e) 5′-AAGGTT-3′

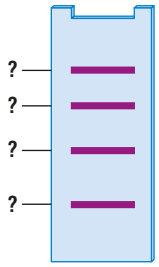
2.5 The list below gives half of each of a set of palindromic restriction sites. Replace the Ns to complete the sequence of each restriction site.

(a) 5′-AGNN-3′
(b) 5′-ATGNNN-3′
(c) 5′-ATTNNN-3′
(d) 5′-NNNAGC-3′

2.6 Apart from nucleotide sequence, what is different about the ends of restriction fragments produced by the following restriction enzymes? (The downward arrow represents the site of cleavage in each strand.)

(a) *Sca*I (5′-AGT↓ACT-3′)
(b) *Nhe*I (5′-G↓CTAGC-3′)
(c) *Cfo*I (5′-GCG↓C-3′)

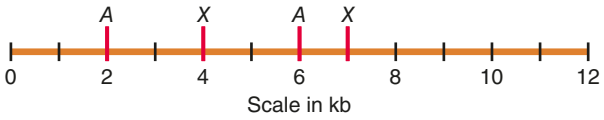
2.7 A solution contains double-stranded DNA fragments of size 4 kb, 8 kb, 10 kb, and 13 kb that are separated in an electrophoresis gel. In the accompanying diagram of the gel, match the fragments sizes with the correct bands.

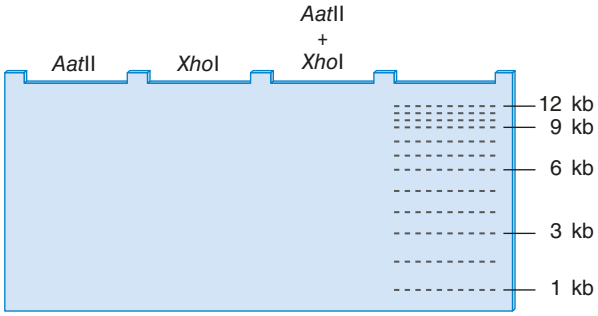


2.8 The linear DNA fragment shown here has cleavage sites for *Aat*II (*A*) and *Xho*I (*X*). In the accompanying diagram of an electrophoresis gel, indicate the positions at which bands would be found after digestion with:

(a) *Aat*II alone.
(b) *Xho*I alone.
(c) *Aat*II and *Xho*I together.

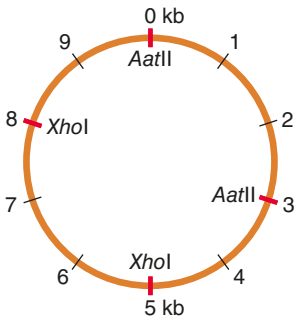
The dashed lines on the right indicate the positions to which bands of 1–12 kb would migrate.



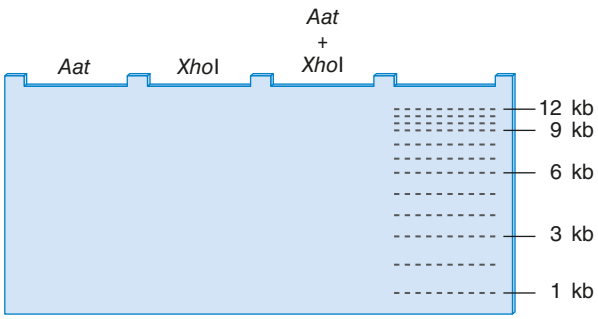


2.9 The circular DNA molecule shown here has cleavage sites for *Aat*II and *Xho*I. In the accompanying diagram of an electrophoresis gel, indicate the positions at which bands would be found after digestion with:

(a) *Aat*II alone.
(b) *Xho*I alone.
(c) *Aat*II and *Xho*I together.

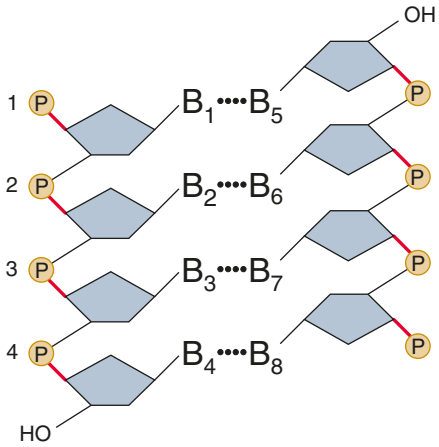


The dashed lines indicate the positions to which bands of 1–12 kb would migrate.



2.10 Consider the accompanying diagram of a region of duplex DNA, in which the Bs represent bases in Watson–Crick pairs. Specify as precisely as possible the identity of:

(a) B_5, assuming that B_1 = A.
(b) B_6, assuming that B_2 = C.
(c) B_7, assuming that B_3 = any purine.
(d) B_8, assuming that B_4 = A or T.



2.11 In the precursor nucleotides of the DNA duplex diagrammed in Problem 2.10, with which base was each of the phosphate groups 1–4 associated with prior to its incorporation into the polynucleotide strand?

2.12 In a random sequence consisting of equal proportions of all four nucleotides, what is the probability that a particular short sequence of nucleotides matches a restriction site for:

(a) A restriction enzyme with a 4-base cleavage site?
(b) A restriction enzyme with a 6-base cleavage site?
(c) A restriction enzyme with an 8-base cleavage site?

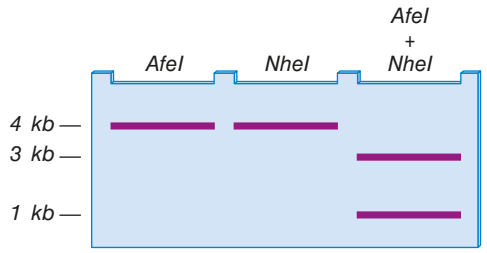
2.13 In a random sequence consisting of equal proportions of all four nucleotides, what is the average distance between restriction sites for:

(a) A restriction enzyme with a 4-base cleavage site?
(b) A restriction enzyme with a 6-base cleavage site?
(c) A restriction enzyme with an 8-base cleavage site?

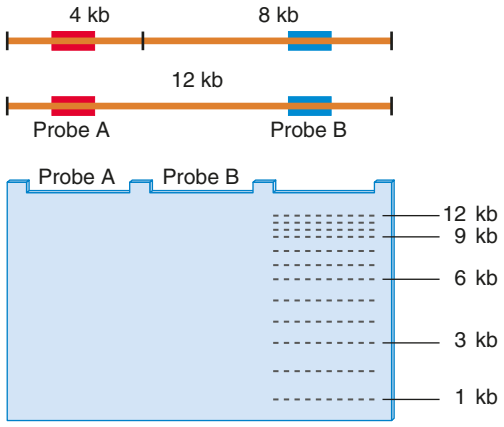
2.14 If *Escherichia coli* DNA were essentially a random sequence of 4.6×10^6 bp with equal proportions of all four nucleotides (this is an oversimplification), approximately how many restriction fragments would be expected from cleavage with:

(a) A "4-cutter" restriction enzyme?
(b) A "6-cutter" restriction enzyme?
(c) An "8-cutter" restriction enzyme?

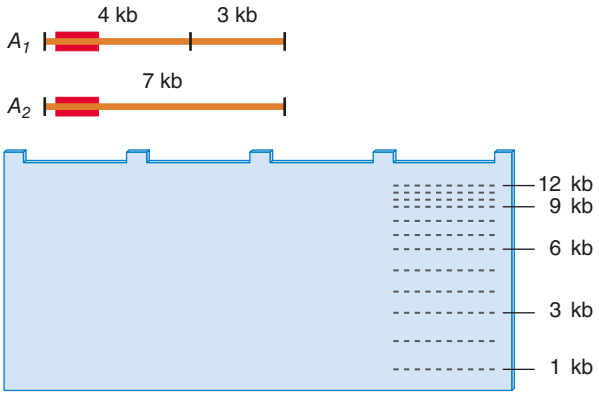
2.15 A circular DNA molecule is cleaved with *Afe*I, *Nhe*I, or both restriction enzymes together. The accompanying diagram shows the resulting electrophoresis gel, with the band sizes indicated. Draw a diagram of the circular DNA showing the relative positions of the *Afe*I and *Nhe*I sites.



2.16 In the diagrams of DNA fragments shown here, the tick marks indicate the positions of restriction sites for a particular restriction enzyme. A mixture of the two types of molecules is digested and analyzed with a Southern blot using either probe A or probe B, which hybridize to the fragments at the positions shown by the rectangles. In the accompanying gel diagram, indicate the bands that would result from each of these probes. (The scale on the right shows the expected positions of fragments from 1–12 kb.)



2.17 In the accompanying diagram, the tick marks indicate the positions of restriction sites in two alternative DNA fragments that can be present at a locus in a human chromosome. An RFLP analysis is carried out, using probe DNA that hybridizes with the fragments at the position shown by the rectangle. With respect to this RFLP, how many genotypes are possible? (Use the symbol A_1 to refer to the allele that yields the upper DNA fragment, and A_2 to refer to the allele that yields the lower DNA fragment.) In the accompanying gel diagram, indicate the genotypes across the top and the phenotype (band position or positions) expected of each genotype. (The scale on the right shows the expected positions of fragments from 1 to 12 kb.)



2.18 If pentamers were long enough to serve as specific oligonucleotide primers for PCR (in practice, they are too short), which DNA fragment would be amplified using the "forward" primer 5′-AATGC-3′ and the "reverse" primer 3′-GCATG-5′ acting on the double-stranded DNA molecule shown here?

```
5′-GATTACCGGTAAATGCCGGATTAACCCGGGTTATCAGGCCACGTACAACTGGAGTCC-3′
3′-CTAATGGCCATTTACGGCCTAATTGGGCCCAATAGTCCGGTGCATGTTGACCTCAGG-5′
```

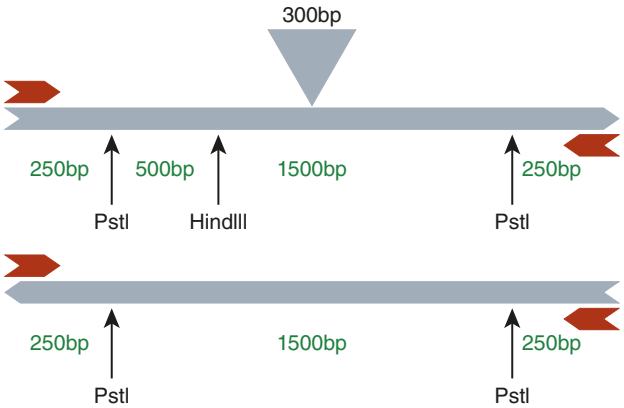
2.19 Would the primer pairs 3′-AATGC-5′ and 5′-GCATG-3′ amplify the same fragment described in Problem 2.18? Explain your answer.

2.20 Suppose that a fragment of human DNA of length 3 kb is to be amplified by PCR. The total genome size is 3×10^6 kb, which equals 3×10^9 base pairs.

(a) Prior to amplification, what fraction of the total DNA does the target sequence constitute?
(b) What fraction does it constitute after 10 cycles of PCR?
(c) After 20 cycles of PCR?
(d) After 30 cycles of PCR?

2.21 The two allelic fragments of DNA shown below are amplified by PCR, and the products digested with both *Pst*I and *Hind*3.

(a) What will the total sizes of the two PCR products be?
(b) Diagram the gel pattern you would expect to see from the doubly digested DNA.

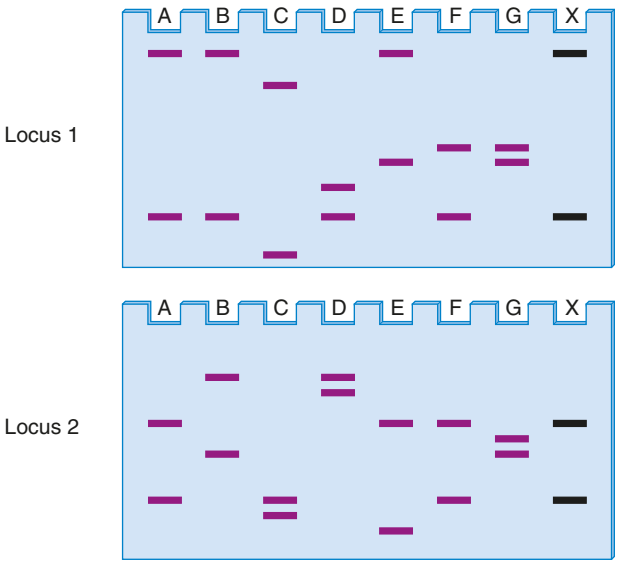


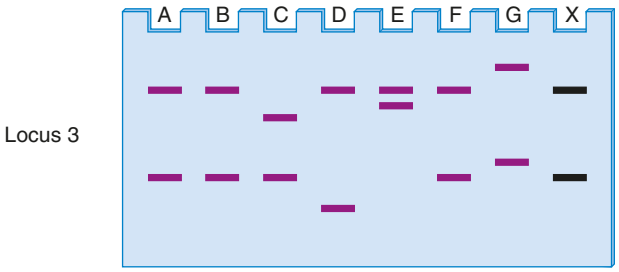
2.22 Below are diagrammed six allelic DNA fragments, showing a number of SNP polymorphisms. Three of these fragments also carry an allele of a gene associated with an inherited disorder (indicated by D). As a clinical geneticist, you would like to use these SNP markers to determine whether an individual is at increased risk of getting the disease. Which of these SNPs would assist you in that determination? Which would not? Justify your answers.

A	G	C	D	G
A	A	C	D	T
A	G	C	D	G
A	A	T	d	T
A	A	T	d	T
T	G	T	d	T

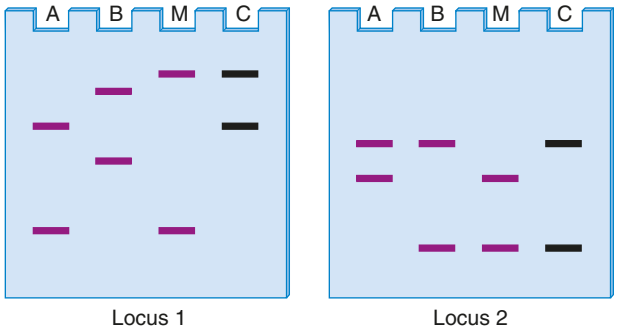
2.23 Psoriasis is a skin disease that has been shown to have a significant genetic component, although, as with breast cancer, multiple genes contribute to its development. An allele of one such gene, known as *PSORS1*, has been located on chromosome 6, as has a SNP marker (rs10484554, a C/T polymorphism); the T allele of this SNP is associated with the risk allele of *PSORS1*. In a population study, 22.4 percent of heterozygous (genotype CT) individuals contracted psoriasis, while in a control population of individuals with genotype CC, the incidence of psoriasis was 11.4 percent. What is the relative risk associated with this genotype?

2.24 A cigarette butt found at the scene of a robbery is found to have a sufficient number of epithelial cells stuck to the paper that the DNA can be extracted and analyzed by DNA typing. Shown here are the results of typing for three probes (locus 1–locus 3) of the evidence (X) and cells from seven suspects (A–G). Which of the suspects can be excluded? Which cannot be excluded? Can you identify the perpetrator? Explain your reasoning.

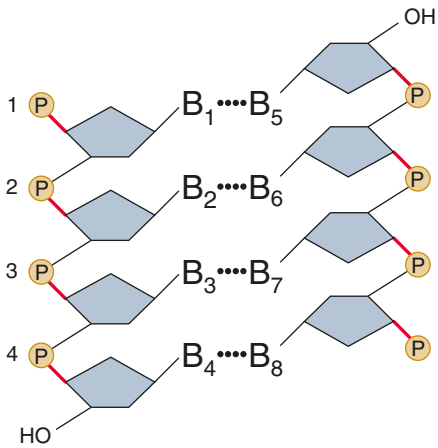




2.25 A woman is uncertain which of two men is the father of her child. DNA typing is carried out on blood from the child (C), the mother (M), and the two males (A and B), using probes for a highly polymorphic DNA marker on two different chromosomes ("locus 1" and "locus 2"). The result is shown in the accompanying diagram. Does either or both of the tested loci rule out any of the males as being the possible father? Explain your reasoning.



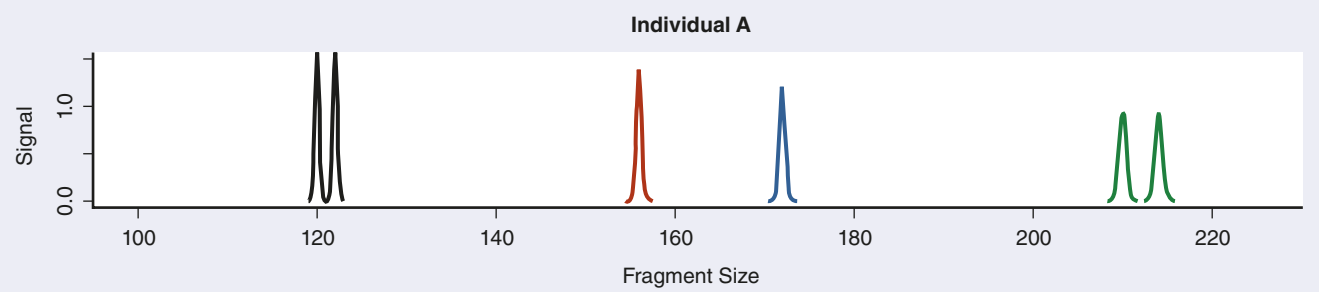
2.26 Snake venom diesterase cleaves the chemical bonds shown in red in the accompanying diagram, leaving mononucleotides that are phosphorylated in the 3′ position. If the phosphates numbered 2 and 4 are radioactive, which mononucleotides will be radioactive after cleavage with snake venom diesterase?

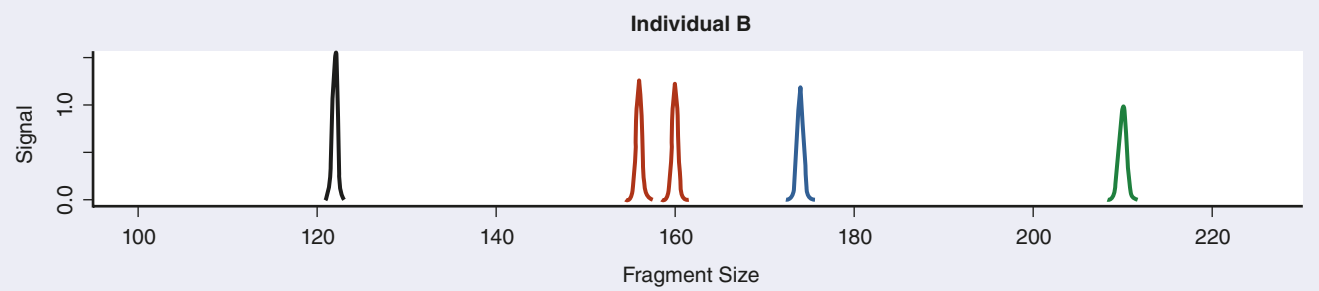


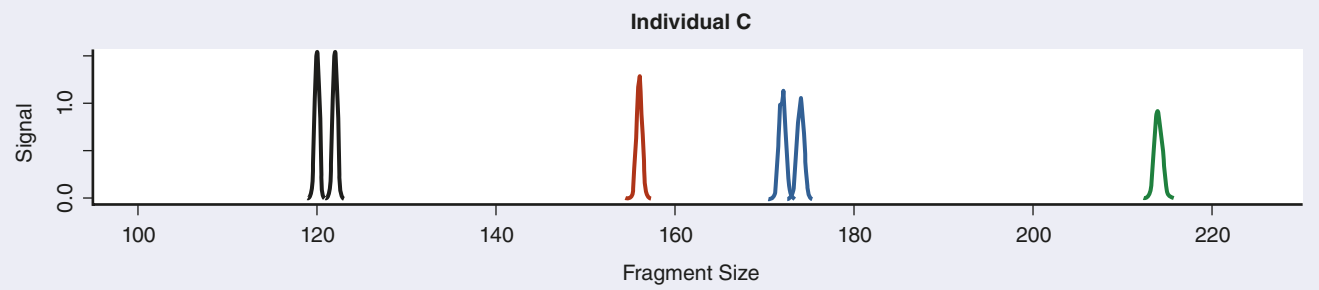
2.27 A DNA sample collected from the scene of a crime is believed to be from the perpetrator. It is genotyped for four SSR loci. Four suspects are genotyped as well. All five genotypes are shown below as fluorometric traces, with the colors corresponding to different loci.

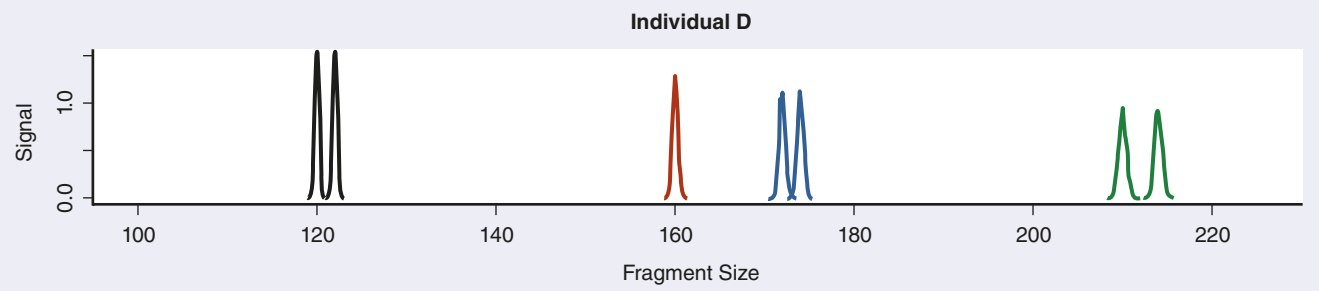
(a) Which individuals can be excluded as suspects and why?

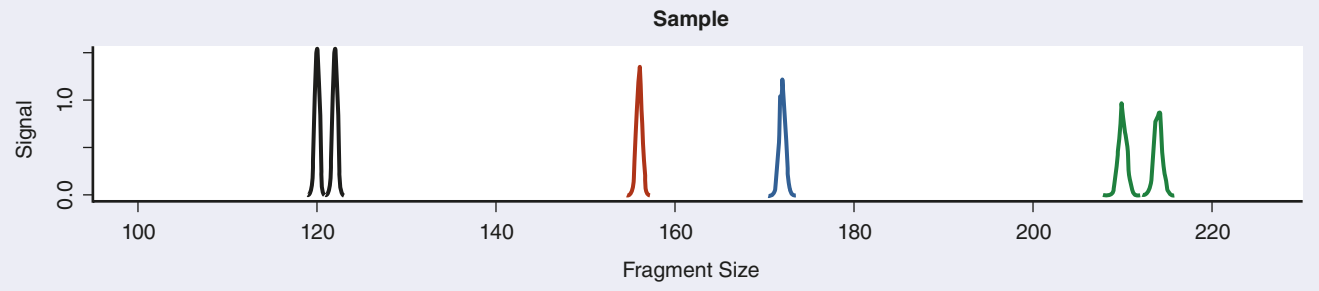
(b) Which of the individuals could be the perpetrator? Do you think the evidence is conclusive? Why or why not?







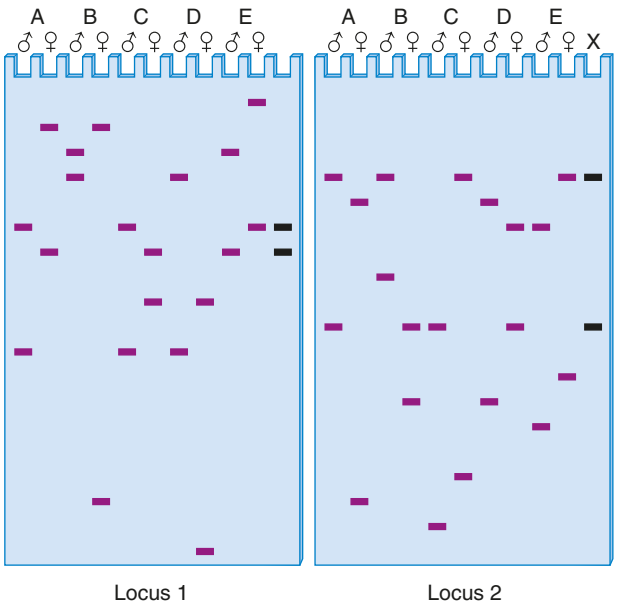




CHALLENGE PROBLEMS

CHALLENGE PROBLEM 1 The genome of *Drosophila melanogaster* is 180×10^6 bp, and a fragment of size 1.8 kb is to be amplified by PCR. How many cycles of PCR are necessary for the amplified target sequence to constitute at least 99 percent of the total DNA?

CHALLENGE PROBLEM 2 The body of a young victim of murder is found in an advanced state of decomposition and cannot be identified. Police suspect the victim is one of five persons reported by their parents as missing. DNA typing is carried out on tissues from the victim (X) and on the five sets of parents (A–E), using probes for a highly polymorphic DNA marker on two different chromosomes ("locus 1" and "locus 2"). The result is shown in the accompanying diagram. How do you interpret the fact that genomic DNA from each individual yields two bands? Can you identify the parents of the victim? Explain your reasoning.



CHALLENGE PROBLEM 3 The snake venom diesterase enzyme described in Problem 2.26 was originally used in a procedure called "nearest neighbor" analysis. In this procedure, a DNA strand is synthesized in the presence of all four nucleoside triphosphates, one of which carries a radioactive phosphate in the (innermost) position. Then the DNA is digested to completion with snake venom diesterase, and the resulting mononucleotides are separated and assayed for radioactivity. Examine the diagram in Problem 2.26 and explain:

(a) How this procedure reveals the "nearest neighbors" of the radioactive nucleotide.
(b) Whether the "nearest neighbor" is on the 5′ or the 3′ side of the labeled nucleotide.

FOR FURTHER READING

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, *32*(3), 314–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6247908

A classic paper describing how RFLP mapping could be used in human genetic analysis.

Kelly, T. J., & Smith, H. O. (1970). A restriction enzyme from *Haemophilus influenzae*: II. Base sequence of the recognition site. *Journal of Molecular Biology*, *51*(2), 393–409. http://doi.org/10.1016/0022-2836(70)90150-6

One of the first reports of the recognition sequence of a restriction enzyme.

Roewer, L. (2013). DNA fingerprinting in forensics: Past, present, future. *Investigative Genetics*, *4*(1), 22. http://doi org/10.1186/2041-2223-4-22

A recent overview of the status of forensic DNA testing.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, *98*(3), 503–517. http://doi .org/10.1016/S0022-2836(75)80083-0

The original Southern blotting procedure.

Steemers, F. J., Chang, W., Lee, G., Barker, D. L., Shen, R., & Gunderson, K. L. (2006). Whole-genome genotyping with the single-base extension assay. *Nature Methods*, *3*(1), 31–33. http://doi.org/10.1038/nmeth842

A description of the basis for whole-genome SNP genotyping.