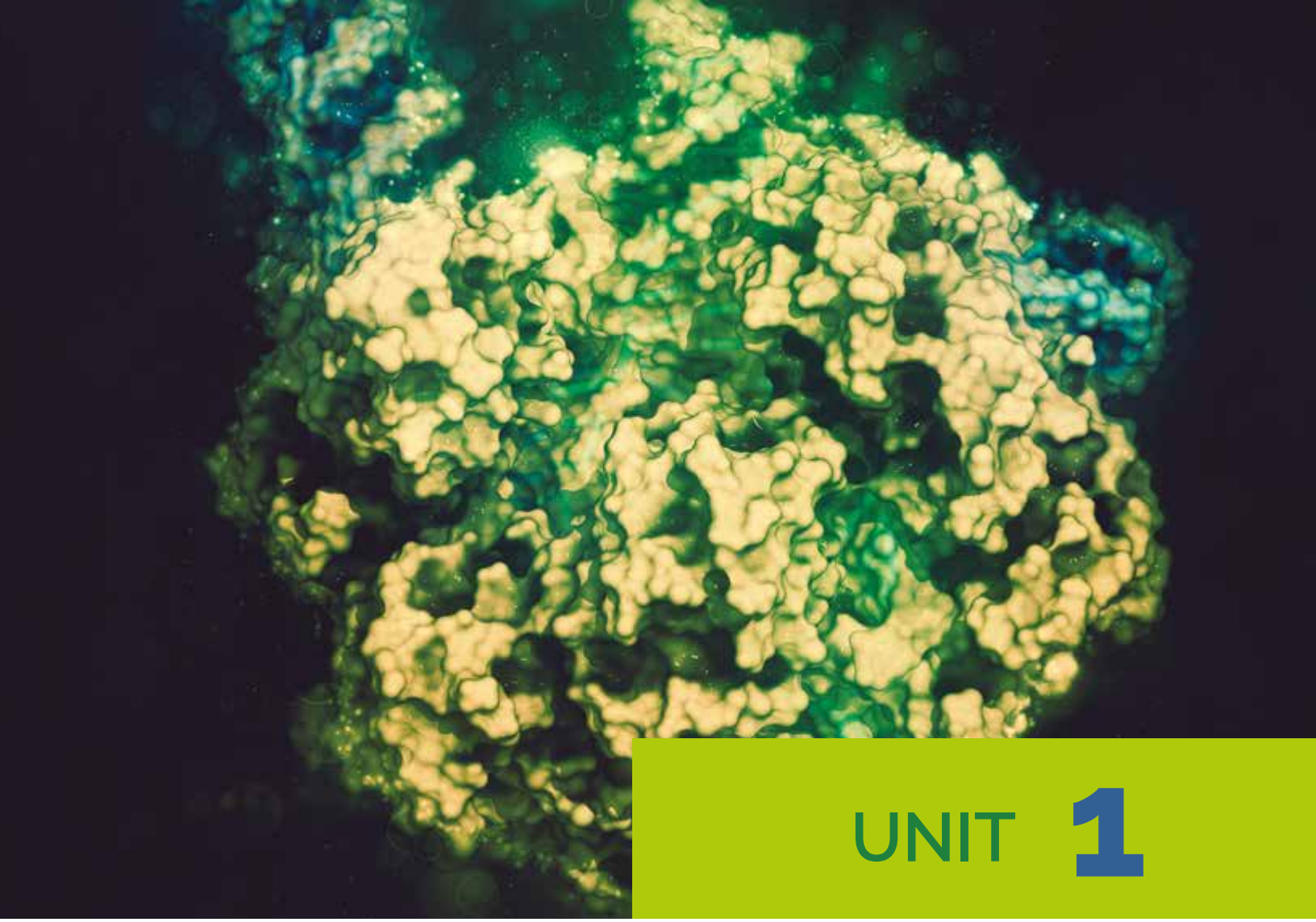


UNIT 1

Defining and Working with Genes

CHAPTER 1

Genes, Genomes, and Genetic Analysis

CHAPTER OUTLINE

ROOTS OF DISCOVERY: The Black Urine Disease
Archibald E. Garrod (1908)
Inborn Errors of Metabolism

ROOTS OF DISCOVERY: One Gene, One Enzyme
George W. Beadle and Edward L. Tatum (1941)
Genetic Control of Biochemical Reactions in Neurospora

Medical Images RM/Jane Hurd.

LEARNING OBJECTIVES & SCIENCE COMPETENCIES

Once you have understood and can apply the principles of genes, genomes, and genetic analysis discussed in this chapter, you will have acquired the following science competencies:

- Given the base sequence of a transcribed strand of protein-coding DNA, specify the sequence of bases in the corresponding region of messenger RNA and the corresponding sequence of amino acids in the protein.
- For a mutation in the DNA in which one base is replaced with another, show how the mRNA and protein will be altered.
- Given a linear metabolic pathway for an essential nutrient, determine which intermediates will restore the ability to grow to mutant strains that are defective for any of the enzymes in the pathway.
- Interpret data specifying which intermediates in a linear metabolic pathway restore the ability of mutants to grow, so as to determine the order in which the enzymes and the intermediates appear in the pathway.
- Given data on the complementation or lack of complementation among all pairs of a set of mutations affecting a biological process, categorize the mutations into groups, each corresponding to a different gene.

Each species of living organism has a unique set of inherited characteristics that makes it different from every other species. Each species has its own developmental plan—often described as a sort of "blueprint" for building the organism—which is encoded in the DNA molecules present in its cells. This developmental plan determines the characteristics that are inherited. Because organisms in the same species share the same developmental plan, organisms that are members of the same species usually resemble one another. For example, it is easy to distinguish a human being from a chimpanzee or a gorilla. A human being habitually stands upright and has long legs, relatively little body hair, a large brain, and a flat face with a prominent nose, jutting chin, distinct lips, and small teeth. All of these traits are inherited—part of our developmental plan—and help set us apart as *Homo sapiens*.

But human beings are by no means identical. Many traits, or observable characteristics, differ from one person to another. In addition to notable differences between males and females, there is a great deal of variation in hair color, eye color, skin color, height, weight, personality traits, and other characteristics. Some human traits (such as sex) are transmitted biologically, others culturally. The color of our eyes results from biological inheritance, but the native language we learned as a child results from cultural inheritance. Many traits are influenced jointly by biological inheritance and environmental factors. For example, weight is determined in part by inheritance but also in part by environment: how much food we eat, its nutritional content, our exercise regimen, and so forth.

Genetics is the study of biologically inherited traits, including traits that are influenced in part by the environment. **Genomics** is the study of all the genes in an organism to understand their molecular organization, function, interaction, and evolutionary history. The fundamental concept of genetics and genomics is as follows:

> Inherited traits are determined by the elements of heredity that are transmitted from parents to offspring in reproduction; these elements of heredity are called **genes**.

The existence of genes and the rules governing their transmission from generation to generation were first articulated by Gregor Mendel in 1866 (described in the chapter titled *Mendelian Genetics: The Principles of Segregation and Assortment*). Mendel's formulation of inheritance was stated in terms of the abstract rules by which hereditary elements (he called them "factors") are transmitted from parents to offspring. His objects of study were garden peas, with variable traits such as pea color and plant height. Mendel thus introduced studying genetics through analysis of the progeny produced from matings. Genetic differences between species were impossible to define, because organisms of different species usually do not mate; if they do mate, they typically produce hybrid progeny that die or are sterile. The approach to the study of genetics through the analysis of the offspring from matings is often referred to as *classical* genetics, or organismic or morphological genetics.

By contrast, *molecular* genetics is the study of the chemical nature of genes and their products. Advances in this area have made it possible to study gene structure and function, as well as differences between species, through the comparison and analysis of the DNA itself. Nevertheless, *there is no fundamental distinction between classical and molecular genetics*. Rather, the two are different and complementary ways of

studying the same thing: the function of the genetic material. This text includes many examples showing how molecular and classical genetics can be used in combination to enhance the power of genetic analysis.

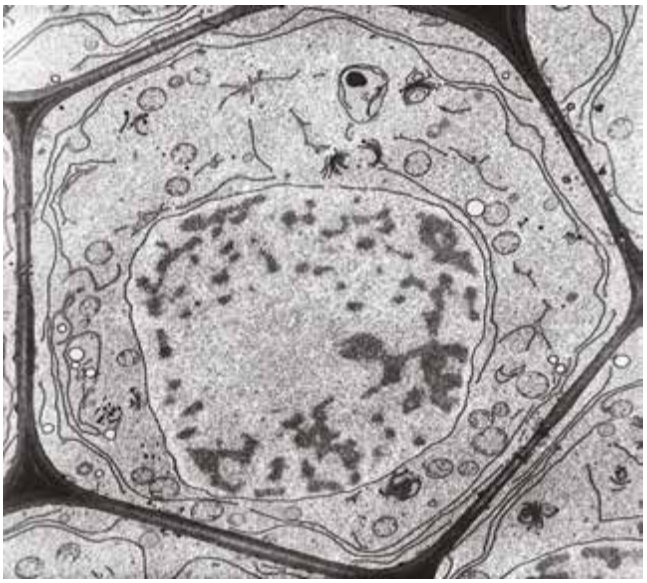
In the rest of this chapter, we will focus on groundbreaking observations and experiments that form the basis for our current understanding of the structure and function of genes. Before we do so, however, we need to consider the four necessary properties of genetic material:

1. The genetic material must encode information
2. That information must be used to direct the functioning of cellular processes.
3. The information contained must be transmissible from cell to cell and from generation to generation.
4. There must be the potential for *mutation* that can result in the physical variation that exists among individuals and between species.

1.1 DNA as the Genetic Material

The foundation of genetics as a molecular science dates back to 1869, just three years after Mendel reported his experiments. It was in that year that Friedrich Miescher discovered a new type of weak acid, abundant in the nuclei of white blood cells. Miescher's weak acid turned out to be the chemical substance we now call **DNA (deoxyribonucleic acid)**. For many years, the biological function of DNA was unknown, and no role in heredity was ascribed to it. This section describes how DNA was eventually isolated and identified as the material of which genes are made.

That the cell nucleus plays a key role in inheritance was recognized in the 1870s when scientists observed that the nuclei of male and female reproductive cells undergo fusion in the process of fertilization. Soon thereafter, chromosomes were first observed inside the nucleus as thread-like objects that become visible in the light microscope when the cell is stained with certain dyes. Chromosomes were found to exhibit a characteristic "splitting" behavior in which each daughter cell formed by cell division receives an identical complement of chromosomes (see the chapter titled *The Chromosomal Basis of Inheritance*). Further evidence for the importance of chromosomes was provided by the observation that, whereas the number of chromosomes in each cell may differ among biological species, the number of chromosomes is nearly always constant within the cells of any particular species. These features of chromosomes were well understood by about 1900, and they made it seem likely that chromosomes were the carriers of the genes.



By the 1920s, several lines of indirect evidence began to suggest a close relationship between chromosomes and DNA. Microscopic studies with special stains showed that DNA is present in chromosomes. Chromosomes also contain various types of proteins, but the amount and kinds of chromosomal proteins differ greatly from one cell type to another, whereas the amount of DNA per cell is constant. Furthermore, nearly all of the DNA present in cells of higher organisms is present in the chromosomes. These arguments for DNA as the genetic material were unconvincing, however, because crude chemical analyses had suggested (erroneously, as it turned out) that DNA lacks the chemical diversity needed to encode the complex set of instructions necessary to build even a simple cell or organism. The favored candidate for the genetic material was protein, because proteins were known to be an exceedingly diverse collection of molecules. Proteins therefore became widely accepted as the genetic material, and DNA was assumed to function merely as the structural framework of the chromosomes. The experiments described in this section showed that, in fact, DNA is the genetic material.

Experimental Proof of the Genetic Function of DNA

In 1928, Frederick Griffith took an important first step in elucidating DNA's role when he demonstrated that genetic material can be transferred from one bacterial cell to another. Griffith was working with two strains of the bacterium *Streptococcus pneumoniae* identified as S and R. When a bacterial cell is grown on solid medium, it undergoes repeated cell divisions to form a visible clump of cells called a **colony**. The S type of *S. pneumoniae* synthesizes a gelatinous capsule composed of complex carbohydrate (polysaccharide). The enveloping capsule makes each

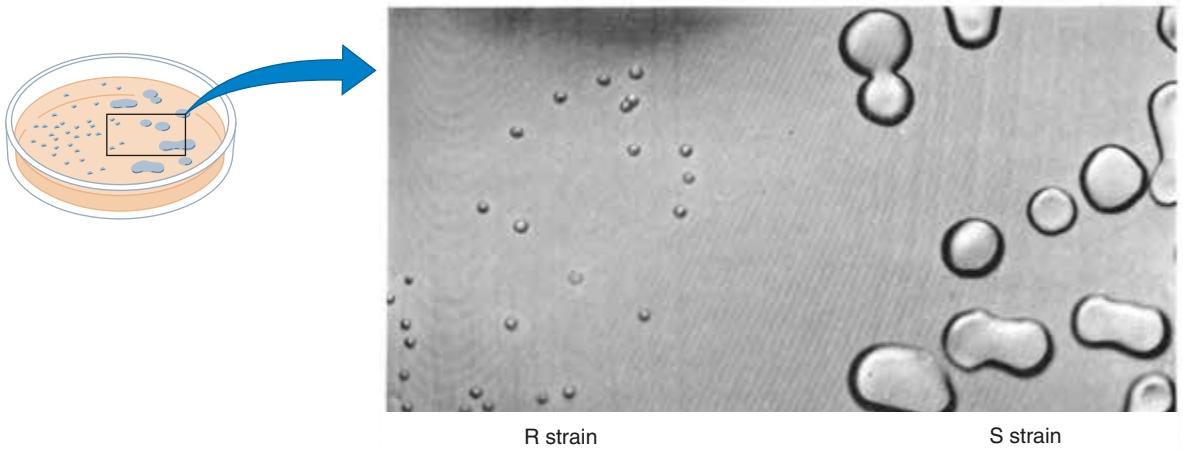


FIGURE 1.1 Colonies of rough (R, the small colonies) and smooth (S, the large colonies) strains of *Streptococcus pneumoniae*. The S colonies are larger because of the gelatinous capsule on the S cells.
Photograph reproduced from Avery, O.T., MacLeod, C.M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine, 79*(2), 137–158. Copyright 1994 by Rockefeller University Press. Reproduced with permission of Rockefeller University Press. doi: 10.1084/jem.79.2.137.

colony large and gives it a glistening or smooth (S) appearance (**FIGURE 1.1**). This capsule also enables the bacterium to cause pneumonia, by protecting it from the defense mechanisms of an infected animal. The R strains of *S. pneumoniae* are unable to synthesize the capsular polysaccharide; they form small colonies that have a rough (R) surface (Figure 1.1). The R strain of the bacterium does not cause pneumonia, because without the capsule the bacteria are inactivated by the immune system of the host. Both types of bacteria "breed true," in the sense that the progeny formed by cell division have the capsular type of the parent, either S or R.

Mice injected with living S cells get pneumonia. In contrast, mice injected with either living R cells or heat-killed S cells remain healthy. Griffith's critical finding was that mice injected with a *mixture* of living R cells and heat-killed S cells contract the disease and often die of pneumonia (**FIGURE 1.2**). Bacteria

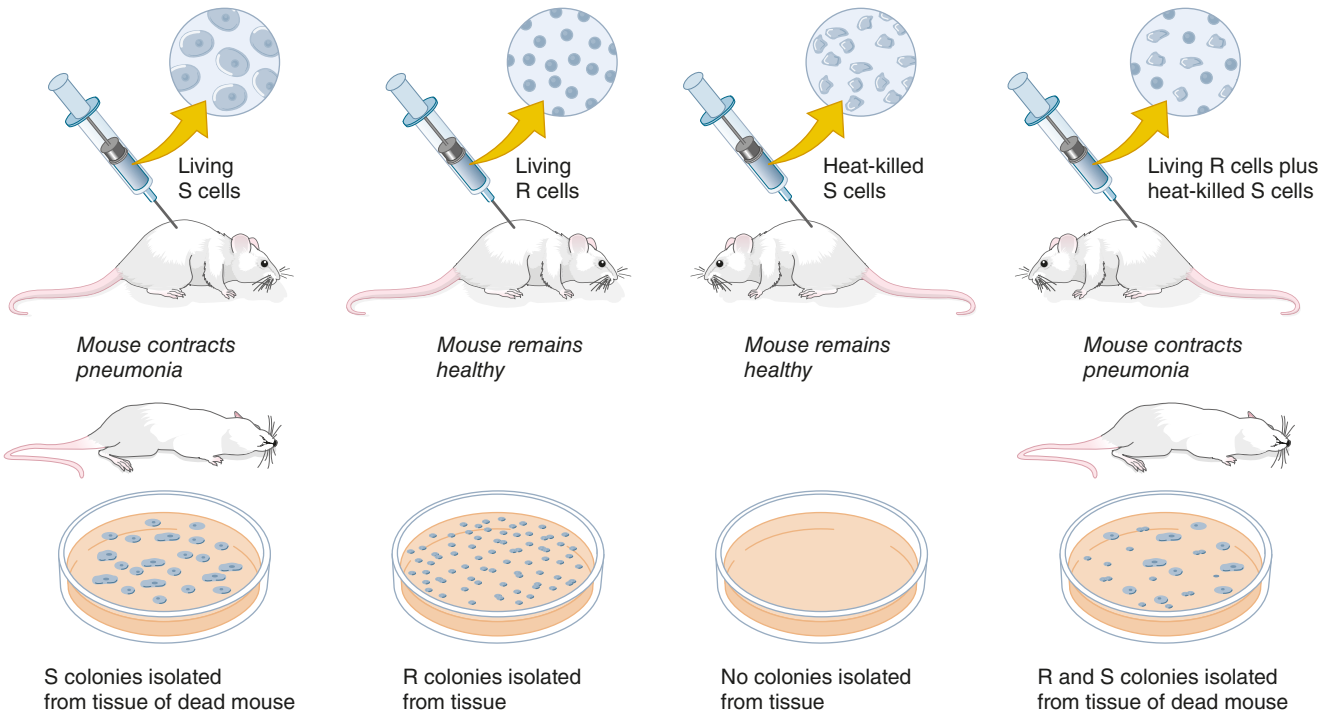


FIGURE 1.2 Griffith's experiment demonstrating bacterial transformation. A mouse remains healthy if injected with either the nonvirulent R strain of *S. pneumoniae* or heat-killed cell fragments of the usually virulent S strain. R cells in the presence of heat-killed S cells are transformed into the virulent S strain, causing pneumonia in the mouse.

isolated from blood samples of these dead mice produce S cultures with a capsule typical of the injected S cells, even though the injected S cells had been killed by heat. Evidently, the injected material from the dead S cells includes a substance that can be transferred to living R cells, which then enables the R cells to synthesize the S-type capsule. In other words, the R bacteria can be changed—or undergo **transformation**—into S bacteria. Furthermore, the ability to synthesize the capsule is inherited by descendants of the transformed bacteria.

Transformation in *Streptococcus* was originally discovered in 1928, but it was not until 1944 that the chemical substance responsible for changing the R cells into S cells was identified. In a milestone experiment, Oswald Avery, Colin MacLeod, and Maclyn McCarty showed that the substance causing the transformation of R cells into S cells was DNA. In doing their experiments, these researchers first needed to develop chemical procedures for isolating almost pure DNA from cells, which had never been done before. When they added DNA isolated from S cells to growing cultures of R cells, they observed transformation—that is, a few S cells were produced. Although the DNA preparations contained traces of protein and RNA (ribonucleic acid, an abundant cellular macromolecule chemically related to DNA), the transforming activity was not altered by treatments that destroyed either protein or RNA. Conversely, treatments that destroyed DNA eliminated the transforming activity (**FIGURE 1.3**). These experiments implied that the substance responsible for genetic transformation was the DNA of the cell—an hence that DNA is the genetic material.

Genetic Role of DNA in Bacteriophages

Another pivotal finding was reported by Alfred Hershey and Martha Chase in 1952. They studied cells of the intestinal bacterium *Escherichia coli* after infection by the virus T2. A virus that attacks bacterial cells is called a **bacteriophage**, a term often shortened to **phage**. *Bacteriophage* means "bacteria-eater." The structure of a bacteriophage T2 particle is illustrated in **FIGURE 1.4**. This particle is exceedingly small, yet it has a complex structure composed of head (which contains the phage DNA), collar, tail, and tail fibers. (The head of a human sperm is about 30–50 times larger in both length and width than the head of T2.) Hershey and Chase were already aware that T2 infection proceeds via the attachment of the tip of a phage particle's tail to the bacterial cell wall, entry of phage material into the cell, multiplication of this material to form a hundred or more progeny phage, and release of the progeny phage by bursting (lysis) of the bacterial host cell. They also knew that T2 particles were composed of DNA and protein in approximately equal amounts.

Because DNA contains phosphorus but no sulfur, whereas most proteins contain sulfur but no phosphorus, it is possible to label DNA and proteins differentially by using radioactive isotopes of the two elements. Hershey and Chase produced particles containing radioactive DNA by infecting *E. coli* cells that had been grown for several generations in a medium that included ^{32}P (a radioactive isotope of phosphorus) and then collecting the phage progeny. Other particles containing labeled proteins were obtained in the same way, by using a medium that included ^{35}S (a radioactive isotope of sulfur).

In the experiments summarized in **FIGURE 1.5**, nonradioactive *E. coli* cells were infected with phage labeled with *either* ^{32}P (part A) *or* ^{35}S (part B) so that the DNA and proteins could be followed on their individual paths. Infected cells were separated from unattached phage particles by centrifugation, resuspended in fresh medium, and then swirled violently in a kitchen blender to shear attached phage material from the cell surfaces. This treatment was found to have no effect on the subsequent course of the infection, which implies that the phage genetic material must enter the infected cells very soon after phage attachment. The kitchen blender turned out to be the critical piece of equipment. Other methods had been tried to tear the phage heads from the bacterial cell surface, but nothing had worked reliably. Hershey later explained, "We tried various grinding arrangements, with results that weren't very encouraging. When Margaret McDonald loaned us her kitchen blender, the experiment promptly succeeded."

After the phage heads were removed by the blender treatment, the infected bacteria were examined. Most of the radioactivity from ^{32}P-labeled phage was found to be associated with the bacteria, whereas only a small fraction of the ^{35}S radioactivity was present in the infected cells. The retention of most of the labeled DNA, in contrast to the loss of most of the labeled protein, implied that a T2 phage transfers most of its DNA, but very little of its protein, to the cell it infects. The critical finding (Figure 1.5) was that about 50 percent of the transferred ^{32}P-labeled DNA, but less than 1 percent of the transferred ^{35}S-labeled protein, was inherited by the *progeny* phage particles. Hershey and Chase interpreted this result to mean that the genetic material in T2 phage is DNA.

The experiments of Avery, MacLeod, and McCarty and those of Hershey and Chase are regarded as classics in the demonstration that genes consist of DNA. At the present time, the equivalent of the transformation experiment is carried out daily in many research laboratories throughout the world, usually with bacteria, yeast, or animal or plant cells grown in culture. These experiments indicate that DNA is the genetic material

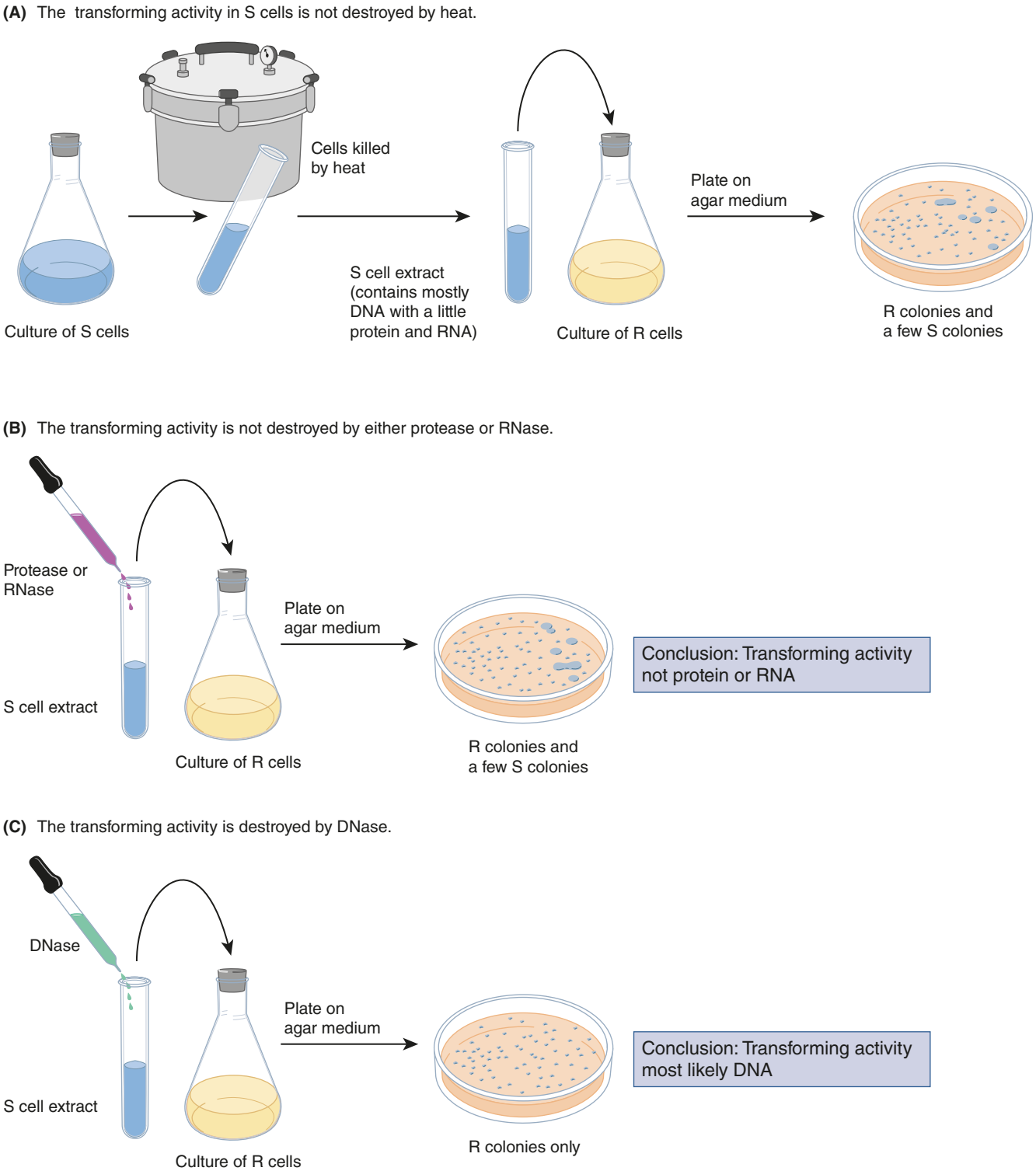


FIGURE 1.3 A diagram of the Avery–MacLeod–McCarty experiment demonstrating that DNA is the active material in bacterial transformation. (A) Purified DNA extracted from heat-killed S cells can convert some living R cells into S cells, but the material may still contain undetectable traces of protein and/or RNA. (B) The transforming activity is not destroyed by either protease or RNase. (C) The transforming activity is destroyed by DNase and so probably consists of DNA.

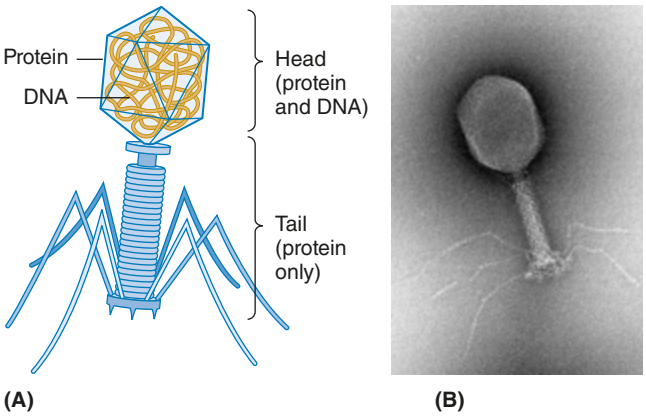


FIGURE 1.4 (A) Drawing of *E. coli* phage T2, showing various components. The DNA is confined to the interior of the head. (B) An electron micrograph of phage T4, a closely related phage. Electron micrograph courtesy of Robert Duda, University of Pittsburgh.

in these organisms as well as in phage T2. Although there are no known exceptions to the generalization that DNA is the genetic material in all cellular organisms and many viruses, in several types of viruses the genetic material consists of RNA. These RNA-containing viruses include the human immunodeficiency virus HIV-1 that causes AIDS (acquired immunodeficiency syndrome).

SUMMING UP

- DNA was identified as a component of chromosomes, but its significance as the genetic material was initially not recognized.
- Frederick Griffith showed that genetic information could be transferred in the process now known as transformation.
- Avery, McLeod, and McCarty demonstrated that the "transforming principle" was DNA.
- Hershey and Chase showed that DNA—but not protein—is transferred to host cells during phage infection.

1.2 DNA Structure and Replication

The inference that DNA is the genetic material left many questions unanswered. How is the DNA in a gene duplicated when a cell divides? How does the DNA in a gene control a hereditary trait? What happens to the DNA when a mutation (a change in the DNA) takes place in a gene?

In the early 1950s, a number of researchers began to try to elucidate the detailed molecular structure of DNA in hopes that the structure alone would suggest answers to these questions. In 1953, James Watson and Francis Crick at Cambridge University proposed the first essentially correct three-dimensional structure of the DNA molecule. The structure was dazzling in its elegance and revolutionary in suggesting how DNA duplicates itself, controls hereditary traits, and undergoes mutation. Even while their tin-and-wire model of the DNA molecule was still incomplete, Crick would visit his favorite pub and exclaim "we have discovered the secret of life."

In the Watson–Crick structure, DNA consists of two long chains of subunits, each twisted around the other to form a double-stranded helix. The double helix is right-handed, which means that as one looks along the barrel, each chain follows a clockwise path as it progresses. You can visualize the right-handed coiling in part A of **FIGURE 1.6** if you imagine yourself looking up into the structure from the bottom. The dark spheres outline the "backbone" of each individual strand, and they coil in a clockwise direction. The subunits of each strand are **nucleotides**, each of which contains any one of four chemical constituents called **bases** attached to a phosphorylated molecule of the five-carbon sugar **deoxyribose**. The four bases in DNA are:

- **Adenine (A)**
- **Guanine (G)**
- **Thymine (T)**
- **Cytosine (C)**

The chemical structures of the nucleotides and bases need not concern us at this time. They are examined in the *DNA Structure and Genetic Variation* chapter.

A key point for our present purposes is that the bases in the double helix are paired as shown in Figure 1.6B:

> At any position on the paired strands of a DNA molecule:
>
> - If one strand has an A, then the partner strand has a T.
> - If one strand has a G, then the partner strand has a C.

The pairing between A and T and between G and C is said to be **complementary**; the complement of A is T, and the complement of G is C. The complementary pairing means that each base along one strand of the DNA is matched with a base in the opposite position on the other strand. Furthermore:

> Nothing restricts the sequence of bases in a single strand, so any sequence could be present along one strand.

This principle explains how only four bases in DNA can code for the huge amount of information needed to make an organism. It is the *sequence* of bases along the DNA that encodes the genetic information, and the sequence is completely unrestricted.

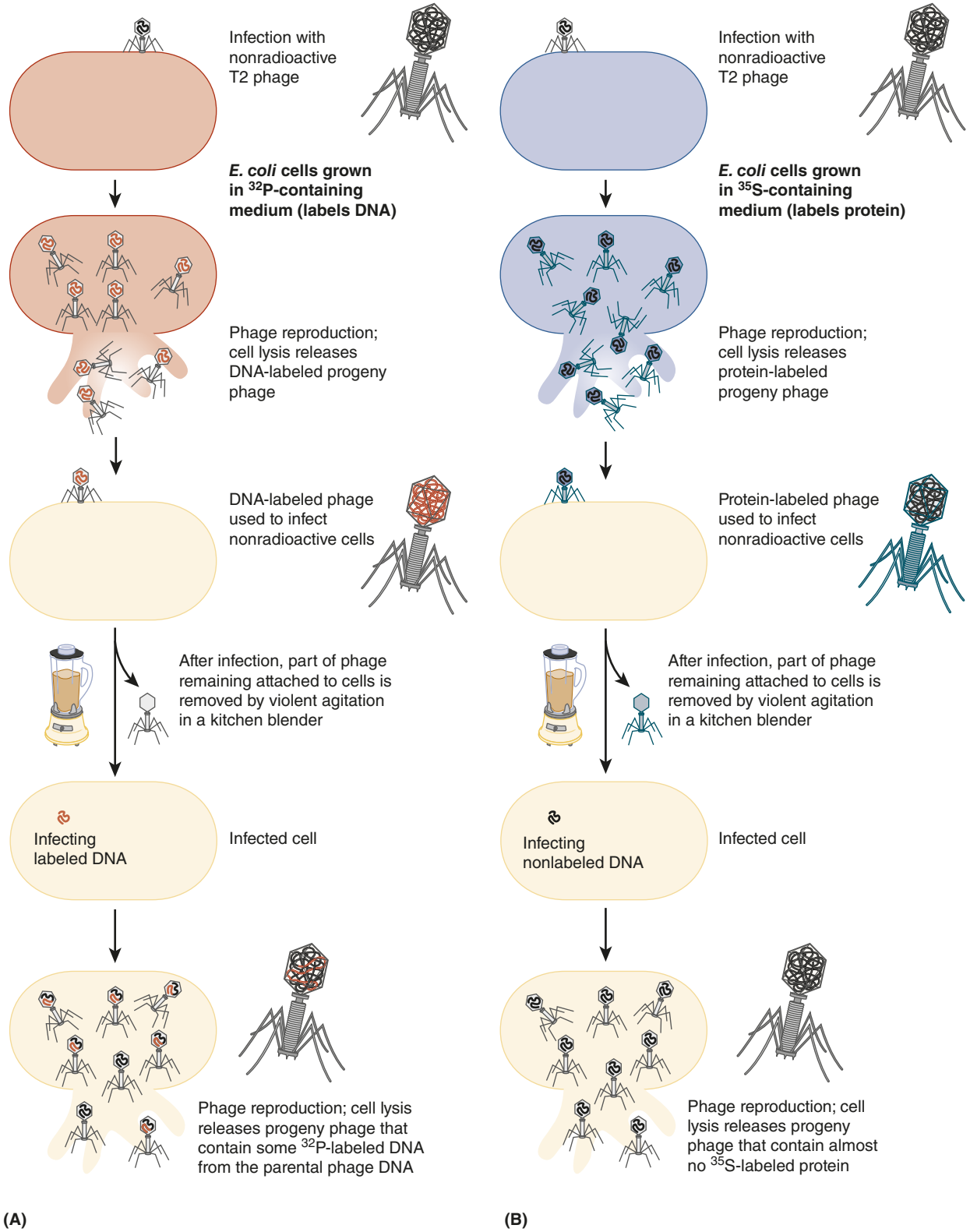


FIGURE 1.5 The Hershey–Chase ("blender") experiment demonstrating that DNA, not protein, is responsible for directing the reproduction of phage T2 in infected *E. coli* cells. (A) Radioactive DNA is transmitted to progeny phage in substantial amounts. (B) Radioactive protein is transmitted to progeny phage in negligible amounts.

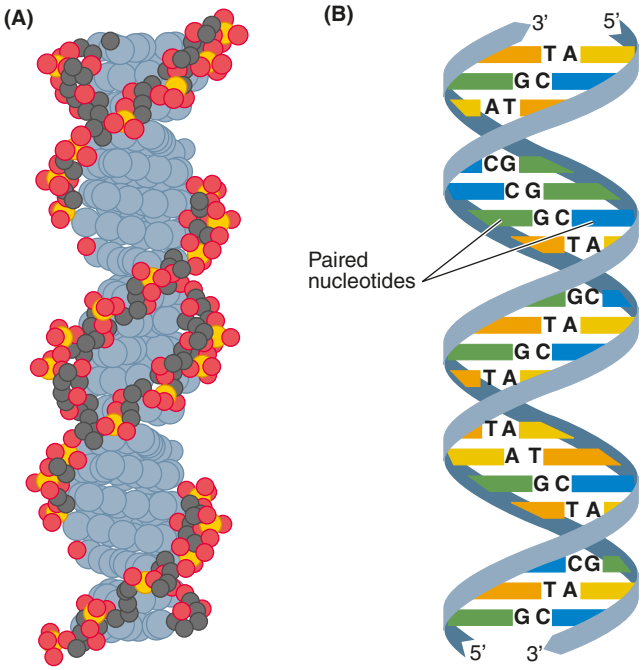


FIGURE 1.6 Molecular structure of the DNA double helix in the standard "B form." (A) A space-filling model, in which each atom is depicted as a sphere. (B) A diagram highlighting the helical strands around the outside of the molecule and the A–T and G–C base pairs inside.

The complementary pairing is also called **Watson–Crick pairing**. In the three-dimensional structure in Figure 1.6A, the base pairs are represented by the lighter spheres filling the interior of the double helix. The base pairs lie almost flat, stacked on top of one another perpendicular to the long axis of the double helix, like pennies in a roll. When discussing a DNA molecule, biologists frequently refer to the individual strands as **single-stranded DNA** and to the double helix as **double-stranded DNA** or **duplex DNA**.

Each DNA strand has a **polarity**, or directionality, like a chain of circus elephants linked trunk to tail. In this analogy, each elephant corresponds to one nucleotide along the DNA strand. The polarity is determined by the direction in which the nucleotides are pointing. The "trunk" end of the strand is called the *3′ end* of the strand, and the "tail" end is called the *5′ end*. In double-stranded DNA, the paired strands are oriented in opposite directions, with the 5′ end of one strand aligned with the 3′ end of the other strand. The molecular basis of the polarity, and the reason for the opposite orientation of the strands in duplex DNA, are explained in the *DNA Structure and Genetic Variation* chapter. In illustrating DNA molecules in this text, we use an arrow-like ribbon to represent the backbone, and we use tabs jutting off the ribbon to represent the nucleotides. The polarity of a DNA strand is indicated by the direction of the arrow-like ribbon. The tail of the arrow represents the 5′ end of the DNA strand, the head the 3′ end.

Beyond the most optimistic hopes, knowledge of the structure of DNA immediately gave clues to its function:

1. The sequence of bases in DNA could be copied by using each of the separate "partner" strands as a pattern for the creation of a new partner strand with a complementary sequence of bases.
2. The DNA could contain genetic information in coded form in the sequence of bases, analogous to letters printed on a strip of paper.
3. Changes in genetic information (mutations) could result from errors in copying in which the base sequence of the DNA became altered.

In the remainder of this chapter, we discuss some of the implications of these clues.

An Overview of DNA Replication

Watson and Crick noted that the structure of DNA itself suggested a mechanism for its replication. "It has not escaped our notice," they wrote, "that the specific base pairing we have postulated immediately suggests a copying mechanism." The copying process in which a single DNA molecule gives rise to two identical molecules is called **replication**. The replication mechanism that Watson and Crick had in mind is illustrated in **FIGURE 1.7**.

As shown in part A of Figure 1.7, the strands of the original (parent) duplex separate, and each individual strand serves as a pattern, or **template**, for the synthesis of a new strand (replica). The replica strands are synthesized by the addition of successive nucleotides in such a way that each base in the replica is complementary (in the Watson–Crick pairing sense) to the base across the way in the template strand (Figure 1.7B). Although the mechanism in Figure 1.7 is simple in principle, it is a complex process that is fraught with geometrical problems and requires a variety of enzymes and other proteins. The details are examined in the *DNA Replication and Sequencing* chapter. The end result of replication is that a single double-stranded molecule becomes replicated into two copies with identical sequences:

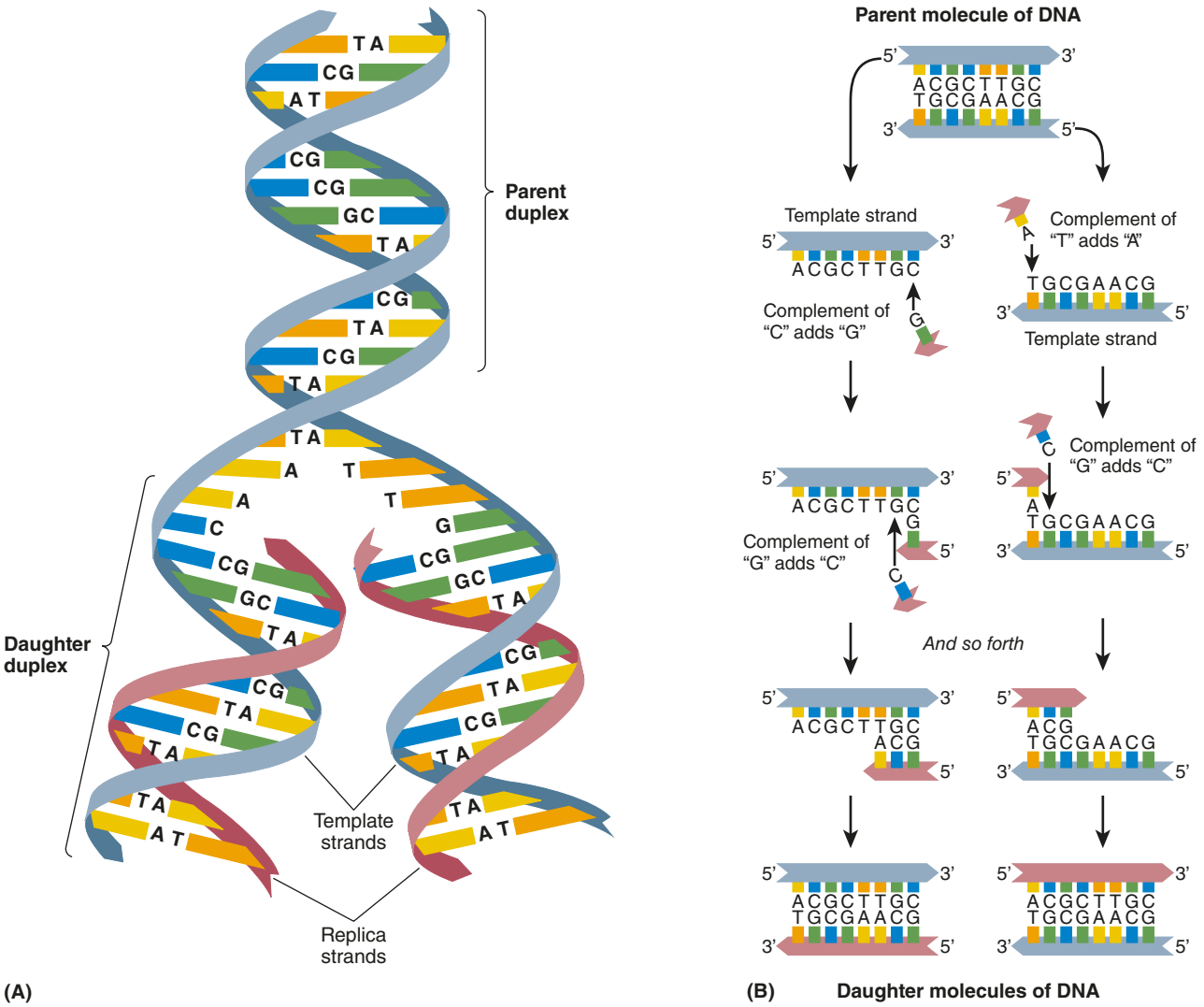


FIGURE 1.7 Replication of DNA. (A) Replication of a DNA duplex as originally envisioned by Watson and Crick. As the parental strands separate, each parental strand serves as a template for the formation of a new daughter strand by means of A–T and G–C base pairing. (B) Greater detail showing how each of the parental strands serves as a template for the production of a complementary daughter strand, which grows in length by the successive addition of single nucleotides to the 3′ end.

Here the bases in the newly synthesized strands are shown in red. In the duplex on the left, the top strand is the template from the parental molecule and the bottom strand is newly synthesized; in the duplex on the right, the bottom strand is the template from the parental molecule and the top strand is newly synthesized. Note in Figure 1.7B that in the synthesis of each new strand, new nucleotides are added only to the 3′ end of the growing chain:

> The obligatory elongation of a DNA strand only at the 3′ end is an essential feature of DNA replication.

This was a point not recognized by Watson and Crick in 1953, but rather one that became clear once the cellular machinery of DNA replication was characterized.

SUMMING UP

- DNA consists of a double helix, with the two strands held together by complementary base pairing.
- The two strands of the double helix have opposite polarities.
- Each strand of a double-helical DNA molecule serves as a template for synthesis of a new one during DNA replication.

1.3 Genes and Proteins

The structure of DNA, by itself, suggested how two of two of the four necessary properties of the genetic material could be explained. First, complementary base pairing

provides a mechanism for replication (Figure 1.7). Second, if the sequence of nucleotides along the DNA is thought of as a string of letters on a sheet of paper, then the genes could be envisioned as distinct words that form sentences and paragraphs that give meaning to the pattern of letters. Based on the structure of DNA alone, however, it is not clear what is coded for. As we will see, a contemporaneous series of experiments showed that what is encoded are proteins, a class of macromolecules that carry out most of the biochemical activities in the cell.

Cells are largely made up of proteins. These include structural proteins that give the cell rigidity and mobility, proteins that form pores in the cell membrane to control the traffic of small molecules into and out of the cell, and receptor proteins that regulate cellular activities in response to molecular signals from the growth medium or from other cells. Proteins are also responsible for most of the metabolic activities of cells. They are essential for the synthesis and breakdown of organic molecules and for generating the chemical energy needed for cellular activities. In 1878, the term **enzyme** was introduced to refer to the biological catalysts that accelerate biochemical reactions in cells. By 1900, thanks largely to the work of the German biochemist Emil Fischer, enzymes were shown to be proteins. As often happens in science, nature's "mistakes" provide clues as to how things work. Such was the case in establishing a relationship between genes and disease, because a "mistake" in a gene (a mutation) can result in a "mistake" (lack of function) in the corresponding protein. This provided a fruitful avenue of research for the study of genetics.

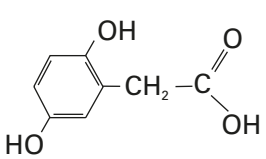
Inborn Errors of Metabolism as a Cause of Hereditary Disease

More than a century ago, the British physician Archibald Garrod realized that certain heritable diseases followed the rules of transmission that Mendel had described for his garden peas. In 1908, Garrod gave a series of lectures in which he proposed a fundamental hypothesis about the relationship between heredity, enzymes, and disease:

> Any hereditary disease in which cellular metabolism is abnormal results from an inherited defect in an enzyme.

Such diseases became known as **inborn errors of metabolism**, a term still in use today.

Garrod studied a number of inborn errors of metabolism in which the patients excreted abnormal substances in their urine. One of these diseases was **alkaptonuria**. In this case, the abnormal substance excreted is **homogentisic acid**:



An early name for homogentisic acid was *alkapton*—hence the name *alkaptonuria* for this disease. Even though alkaptonuria is rare, with an incidence of about one in 200,000 people, it was well known even before Garrod studied it. The disease itself is relatively mild, but it has one striking symptom: The urine of the patient turns black because of the oxidation of homogentisic acid (**FIGURE 1.8**). For this reason, alkaptonuria is also called *black urine disease*. An early case was described in the year 1649:

> The patient was a boy who passed black urine and who, at the age of fourteen years, was submitted to a drastic course of treatment that had for its aim the subduing of the fiery heat of his viscera, which was supposed to bring about the condition in question by charring and blackening his bile. Among the measures prescribed were bleedings, purgation, baths, a cold and watery diet, and drugs galore. None of these had any obvious effect, and eventually the patient, who tired of the futile and superfluous therapy, resolved to let things take their natural course. None of the predicted evils ensued. He married, begat a large family, and lived a long and healthy life, always passing urine black as ink. (Recounted by Garrod, 1908.)

Garrod was primarily interested in the biochemistry of alkaptonuria, but he took note of family studies that indicated that the disease was inherited as though it were due to a defect in a single gene. As to the biochemistry, he deduced that the problem in alkaptonuria was the patients' inability to break down the phenyl ring of six carbons that is present in homogentisic acid. Where does this ring come from? Most animals obtain it from foods in their diet. Garrod proposed that homogentisic acid originates as a

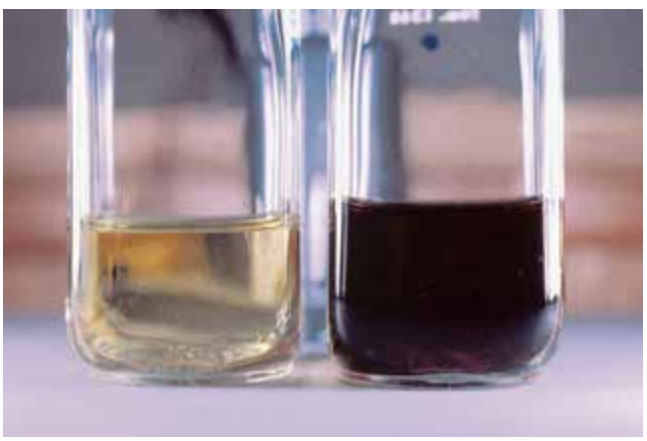


FIGURE 1.8 Urine from a person with alkaptonuria turns black because of the oxidation of the homogentisic acid that it contains. Courtesy of Daniel De Aguiar.

breakdown product of two amino acids, phenylalanine and tyrosine, which also contain a phenyl ring. An **amino acid** is one of the "building blocks" from which proteins are made. Phenylalanine and tyrosine are constituents of normal proteins. The scheme that illustrates the relationship between the molecules is shown in **FIGURE 1.9**. Any such sequence of biochemical reactions is called a **biochemical pathway** or a **metabolic pathway**. Each arrow in the pathway represents a single step depicting the transition from the "input" or **substrate molecule**, shown at the head of the arrow, to the "output" or **product molecule**, shown at the tip. Biochemical pathways are usually oriented either vertically with the arrows pointing down, as in Figure 1.9, or horizontally, with the arrows pointing from left to right. Garrod did not know all of the details of the pathway in Figure 1.9, but he did understand that the key step in the breakdown of homogentisic acid is the breaking open of the phenyl ring and that the phenyl ring in homogentisic acid comes from dietary phenylalanine and tyrosine.

What allows each step in a biochemical pathway to occur? Garrod correctly surmised that each step requires a specific enzyme to catalyze the reaction for the chemical transformation. Persons with an inborn error of metabolism, such as alkaptonuria, have a defect in a single step of a metabolic pathway because they lack a functional enzyme for that step. When an enzyme in a pathway is defective, the pathway is said to have a **block** at that step. One frequent result of a blocked pathway is that the substrate of the defective enzyme accumulates. Observing the accumulation of homogentisic acid in patients with alkaptonuria, Garrod proposed that there must be an enzyme whose function is to open the phenyl ring of homogentisic acid and that this enzyme is missing in these patients. Isolation of the enzyme that opens the phenyl ring of homogentisic acid was not actually achieved until 50 years after Garrod's lectures. In the average person, it is found in cells of the liver, and just as Garrod had predicted, the enzyme is defective in patients with alkaptonuria.

The pathway for the breakdown of phenylalanine and tyrosine, as it is understood today, is shown in **FIGURE 1.10**. In this figure, the emphasis is on the enzymes rather than on the structures of the **metabolites**, or small molecules, on which the enzymes act. Each step in the pathway requires the presence of a particular enzyme that catalyzes that step.

Although Garrod knew only about alkaptonuria, in which the defective enzyme is homogentisic acid 1,2-dioxygenase, we now know the clinical consequences of defects in the other enzymes. Unlike alkaptonuria, which is a relatively benign inherited disease,

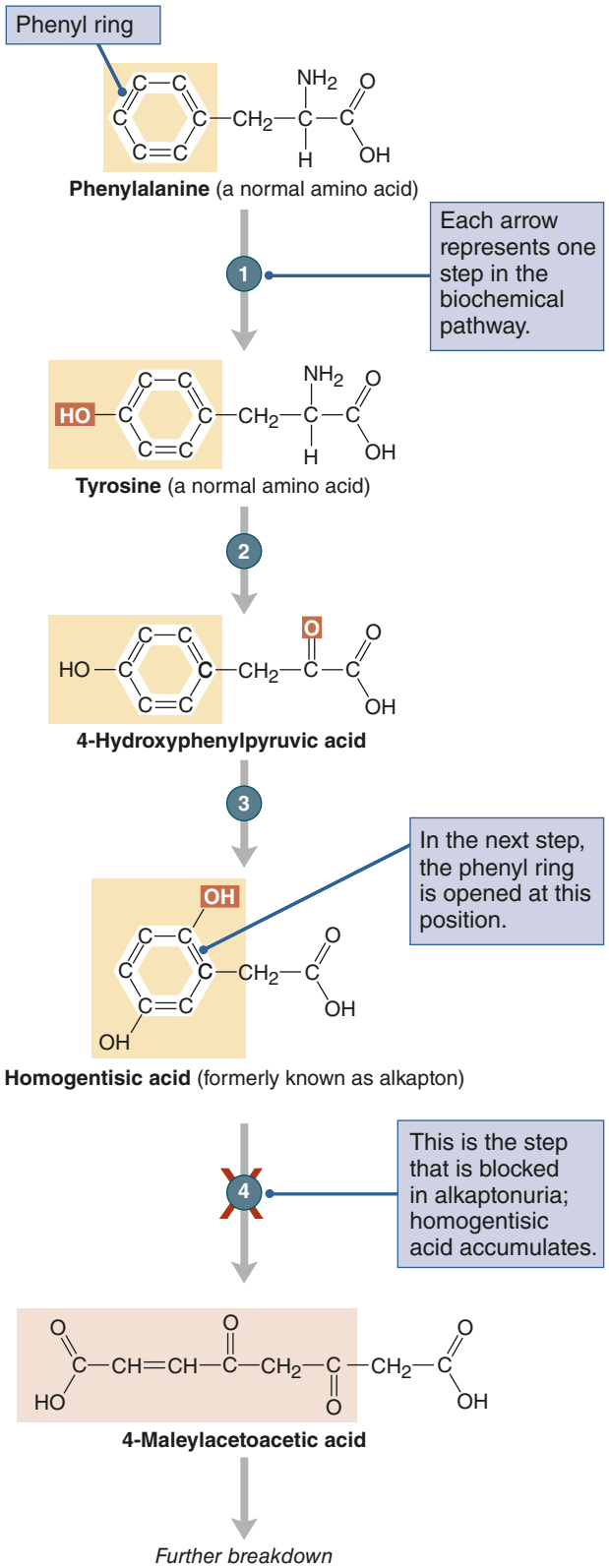


FIGURE 1.9 Metabolic pathway for the breakdown of phenylalanine and tyrosine. Each step in the pathway, represented by an arrow, requires a specific enzyme to catalyze the reaction. The key step in the breakdown of homogentisic acid is the breaking open of the phenyl ring.

ROOTS OF DISCOVERY

The Black Urine Disease

Archibald E. Garrod (1908)
St. Bartholomew's Hospital London, England

Inborn Errors of Metabolism

Although he was a distinguished physician, Garrod's lectures on the relationship between heredity and congenital defects in metabolism had no impact when they were delivered. The important concept that one gene corresponds to one enzyme (the "one gene–one enzyme hypothesis") was developed independently in the 1940s by George W. Beadle and Edward L. Tatum, who used the bread mold Neurospora crassa *as their experimental organism. When Beadle finally became aware of Garrod's hypothesis about inborn errors of metabolism, he was generous in praising it. This excerpt shows Garrod at his best, interweaving history, clinical medicine, heredity, and biochemistry in his account of alkaptonuria. The excerpt also illustrates how the severity of a genetic disease depends on its social context. Garrod writes as though alkaptonuria were a harmless curiosity. This is indeed largely true when the life expectancy is short. With today's longer life span, alkaptonuria patients accumulate the dark pigment in their cartilage and joints and can eventually develop arthritis.*

Garrod's book Inborn Errors of Metabolism, *published in 1903, dealt with a number of inherited metabolic disorders, but two chapters (four and five) were devoted to alkaptonuria. In them, he proposes two critical hypotheses. The first hypothesis relates to the source of the homogentisic acid in the urine of affected individuals.*

Two explanations are possible [for] the fact that alcaptonurics excrete homogentisic acid whereas normal persons do not. Either the alcapton acid is a strictly abnormal product formed by a perverted metabolism of tyrosin and phenylalanin, or it is an intermediate product of normal metabolism, which in alcaptonurics escapes further change . . . It appears to me that the evidence in favour of the theory of an intermediate product far outweighs that which can be brought against it (p. 38) . . .

The second hypothesis is the one that proved to be a cornerstone of modern genetics, connecting Mendel's laws with metabolic traits.

It was pointed out [by others] that the mode of incidence of alkaptonuria finds a ready explanation if the anomaly be regarded as a rare recessive character in the Mendelian sense . . . Of the cases of alkaptonuria, a very large proportion have been in the children of first cousin marriages . . . It is also noteworthy that, if one takes families with five or more children [with both parents normal and at least one child affected with alkaptonuria], the totals work out in strict conformity to Mendel's law, i.e., 57 [normal children]:19 [affected children] in the proportions 3:1 (pp. 23–25).

> ***"We may further conceive that the splitting of the phenyl ring in normal metabolism is the work of a special enzyme and that in congenital alkaptonuria this enzyme is wanting."***

It is completely understandable why Beadle had high praise for Garrod. In fact, the hypotheses in Inborn Errors of Metabolism *were in essence those that Beadle and Tatum tested in their experiments with* Neurospora. *Remarkably, Garrod arrived at them through synthesis of medical and physiological observations (his own and that of others) that were not made in the context of genetic analysis. Indeed, in 1903, that context was in its infancy.*

Source: Harris, H. (1963). Garrod, A. (1903) *Inborn errors of metabolism*. Oxford University Press.

the others are very serious. The condition known as **phenylketonuria (PKU)** results from the absence of (or a defect in) the enzyme **phenylalanine hydroxylase (PAH)**. When this step in the pathway is blocked, phenylalanine accumulates. The excess phenylalanine is broken down into harmful metabolites that cause defects in myelin formation that damage a child's developing nervous system and lead to severe intellectual disability.

However, if PKU is diagnosed in children soon enough after birth, they can be placed on a specially formulated diet that is low in phenylalanine. The child is

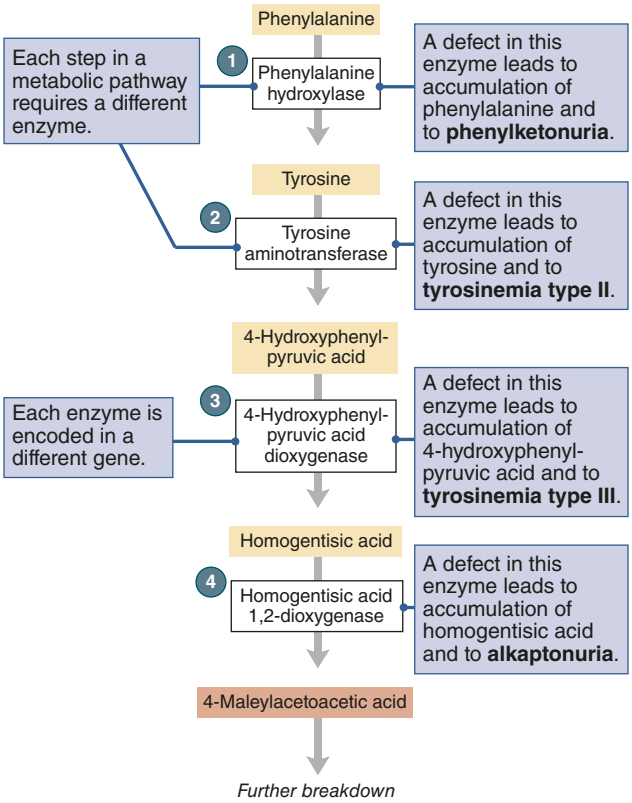


FIGURE 1.10 Inborn errors of metabolism that affect the breakdown of phenylalanine and tyrosine. An inherited disease results when any of the enzymes is missing or defective. Alkaptonuria results from a mutant homogentisic acid 1,2-dioxygenase; phenylketonuria results from a mutant phenylalanine hydroxylase.

allowed only as much phenylalanine as can be used in the synthesis of proteins, thereby ensuring excess phenylalanine does not accumulate. The special diet is very strict. It excludes meat, poultry, fish, eggs, milk and milk products, legumes, nuts, and bakery goods manufactured with regular flour. These foods are replaced by an expensive synthetic formula. With the special diet, however, the detrimental effects of excess phenylalanine on intellectual development can largely be avoided, although in adult women with PKU who are pregnant, the fetus is at risk.

In many countries, including the United States, all newborn babies have their blood tested for chemical signs of PKU. Routine screening is cost-effective because PKU is relatively common. In the United States, the incidence is about 1 in 8000 among Caucasian births. The disease is less common in other ethnic groups.

In the metabolic pathway depicted in Figure 1.10, defects in the breakdown of tyrosine or of 4-hydroxyphenylpyruvic acid lead to types of tyrosinemia. Like PKU, these diseases have severe effects: Type II is associated with skin lesions and intellectual disability, and Type III with severe liver dysfunction.

The genes for the enzymes in the biochemical pathway in Figure 1.10 have all been identified and the nucleotide sequence of the DNA determined. In the following list, and throughout this text, we use the standard typographical convention that *genes* are written in *italic* type, whereas gene products are not printed in italics. This convention is convenient, because it means that the protein product of a gene can be represented with the same symbol as the gene itself, but whereas the gene symbol is in italics, the protein symbol is not. In Figure 1.10, the numbers correspond to the following genes and enzymes:

1. The gene *PAH* on the long arm of chromosome 12 encodes phenylalanine hydroxylase (PAH).
2. The gene *TAT* on the long arm of chromosome 16 encodes tyrosine aminotransferase (TAT).
3. The gene *HPD* on the long arm of chromosome 12 encodes 4-hydroxyphenylpyruvic acid dioxygenase (HPD).
4. The gene *HGD* on the long arm of chromosome 3 encodes homogentisic acid 1,2 dioxygenase (HGD).

SUMMING UP

- Archbold Garrod identified the genetic basis of "inborn errors of metabolism."
- Diseases such as alkaptonuria are the result of the absence of a particular enzyme in a biochemical pathway.
- With some inborn errors of metabolism, such as phenylketonuria, dietary regulation can prevent adverse symptoms from developing.

1.4 Genetic Analysis

The genetic implications of Garrod's research on inborn errors of metabolism were not widely appreciated, most likely because his writings focused primarily on biochemical pathways rather than inheritance. And, of course, since the genes and enzymes were being studied in humans, the potential for classical genetic analysis was limited.

The definitive connection between genes and enzymes came from studies carried out in the 1940s by George W. Beadle and Edward L. Tatum using a filamentous fungus *Neurospora crassa*, commonly called red bread mold—an organism they chose because both genetic and biochemical analyses could be done with ease. In these experiments, Beadle and Tatum identified new mutations, each of which caused a block in the metabolic pathway for the synthesis of some needed

ROOTS OF DISCOVERY

One Gene, One Enzyme

George W. Beadle and Edward L. Tatum (1941)
Stanford University, Stanford, California

Genetic Control of Biochemical Reactions in Neurospora

How do genes control metabolic processes? The suggestion that genes control enzymes was made very early in the history of genetics, most notably by the British physician Archibald Garrod in his 1903 book Inborn Errors of Metabolism. *Nevertheless, the precise relationship between genes and enzymes was still uncertain. Perhaps each enzyme is controlled by more than one gene, or perhaps each gene contributes to the control of several enzymes. The classic experiments of Beadle and Tatum showed that the relationship is usually remarkably simple: One gene codes for one enzyme. Their pioneering experiments united genetics and biochemistry, and for the "one gene, one enzyme" concept, Beadle and Tatum were awarded a Nobel Prize in 1958 (Joshua Lederberg shared the prize for his contributions to microbial genetics). Because we now know that some enzymes contain polypeptide chains encoded by two (or occasionally more) different genes, a more accurate statement of the principle is "one gene, one polypeptide." Beadle and Tatum's experiments also demonstrate the importance of choosing the right organism.* Neurospora *had been introduced as a genetic organism only a few years earlier, and Beadle and Tatum realized that they could take advantage of this organism's ability to grow on a simple medium composed of known substances.*

Beadle and Tatum's work was published in 1941 and can be found at the reference at the end of this feature. In it, they point out the limitations of starting with the physiological basis of a trait (such as black urine disease) and attempting to determine its genetic basis. First, these analyses are limited to traits whose variants are nonlethal. Second, the variants must have visible effects. To get around these problems, Beadle and Tatum turned the problem on its head.

[These limitations] have led us to investigate the general problem of the genetic control of development and metabolic reactions by reversing the ordinary procedure . . . [by setting out] to determine if and how genes control known biochemical reactions . . . If the organism must be able to carry out a certain chemical reaction to survive on a given medium, a mutant unable to do this will obviously be lethal on this medium. . . . [It can be] studied, however, if it will grow on a medium to which has been added the essential product of the genetically blocked reaction. . . .

> ***"These preliminary results appear to us to indicate that the approach may offer considerable promise as a method of learning more about how genes regulate development and function."***

Thus, rather than starting with observed differences in traits among individuals, Beadle and Tatum started by generating mutations (in their case, mutations resulting from X-irradiation of Neurospora *cells) and then identified those that on minimal medium are lethal, but on medium supplemented with the normal product of the mutated gene. This experimental approach ranks among the most important experimental tool of genetic analysis.*

Source: G. W. Beadle and E. L. Tatum, Genetic Control of Biochemical Reactions in Neurospora. *Proc. Natl. Acad. Sci. USA 27* (1941): 499–506.

nutrient, and showed that each of these blocks corresponded to a defective enzyme needed for one step in the pathway. Their research was important not only because it solidified the link between genetics and biochemistry, but also because their experimental approach, now called **genetic analysis**, has been successfully applied to understanding a wide range of biological processes, from the genetic control of the cell cycle and cancer to the genetic influences on development and behavior. For this reason, we will examine the Beadle–Tatum experiments in some detail.

Mutant Genes and Defective Proteins

N. crassa grows in the form of filaments on a great variety of substrates, including a laboratory medium containing only inorganic salts, a sugar, and the vitamin biotin. Such a **minimal medium** contains only the nutrients that are essential for growth of the organism. The *N. crassa* filaments consist of a mass of branched threads separated into interconnected, multinucleate compartments, which allow free interchange of nuclei and cytoplasm. Each nucleus contains a single set of seven chromosomes. Beadle and Tatum recognized that the

ability of *Neurospora* to grow in minimal medium implied that the organism must be able to synthesize all metabolic components other than biotin. If the biosynthetic pathways needed for growth are controlled by genes, then a mutation in a gene responsible for synthesizing an essential nutrient would be expected to render a strain unable to grow unless the strain was provided with the nutrient.

These ideas were tested in the following way. Spores of nonmutant *Neurospora* were irradiated with either x-rays or ultraviolet light to produce mutant strains with various nutritional requirements. (Why these treatments cause mutations is discussed in the *Mutation, Repair, and Recombination* chapter.) The isolation of a set of mutants affecting any biological process—in this case, metabolism—is called a **mutant screen**. In the initial step for identifying mutants, summarized in **FIGURE 1.11**, the irradiated spores (purple) were used in crosses with an untreated strain (green). Ascospores produced by the sexual cycle in fruiting bodies were individually germinated in a **complete medium**—that is, a complex medium enriched with a variety of amino acids, vitamins, and other substances expected to be essential metabolites whose synthesis could be blocked by a mutation. Even those ascospores containing a new mutation affecting synthesis of an essential nutrient would be expected to germinate and grow in complete medium.

To identify which of the irradiated ascospores contained a new mutation affecting the synthesis of an essential nutrient, conidia from each culture were transferred to minimal medium (**FIGURE 1.12A**). The vast majority of cultures could grow on minimal medium; these were discarded because they lacked any new mutation of the desired type. The retained cultures were the small number unable to grow in minimal medium, because they contained a new mutation blocking the synthesis of some essential nutrient.

Samples from a portion of each mutant culture were then transferred to a series of media to determine whether the mutation resulted in a requirement for a vitamin, an amino acid, or some other substance (**FIGURE 1.12B**). In the example illustrated, the mutant strain requires one (or possibly more than one) amino acid, because a mixture of all amino acids added to the minimal medium allows growth. Because the proportion of irradiated cultures with new mutations was very small, only a negligible number of cultures contained two or more new mutations that had occurred simultaneously.

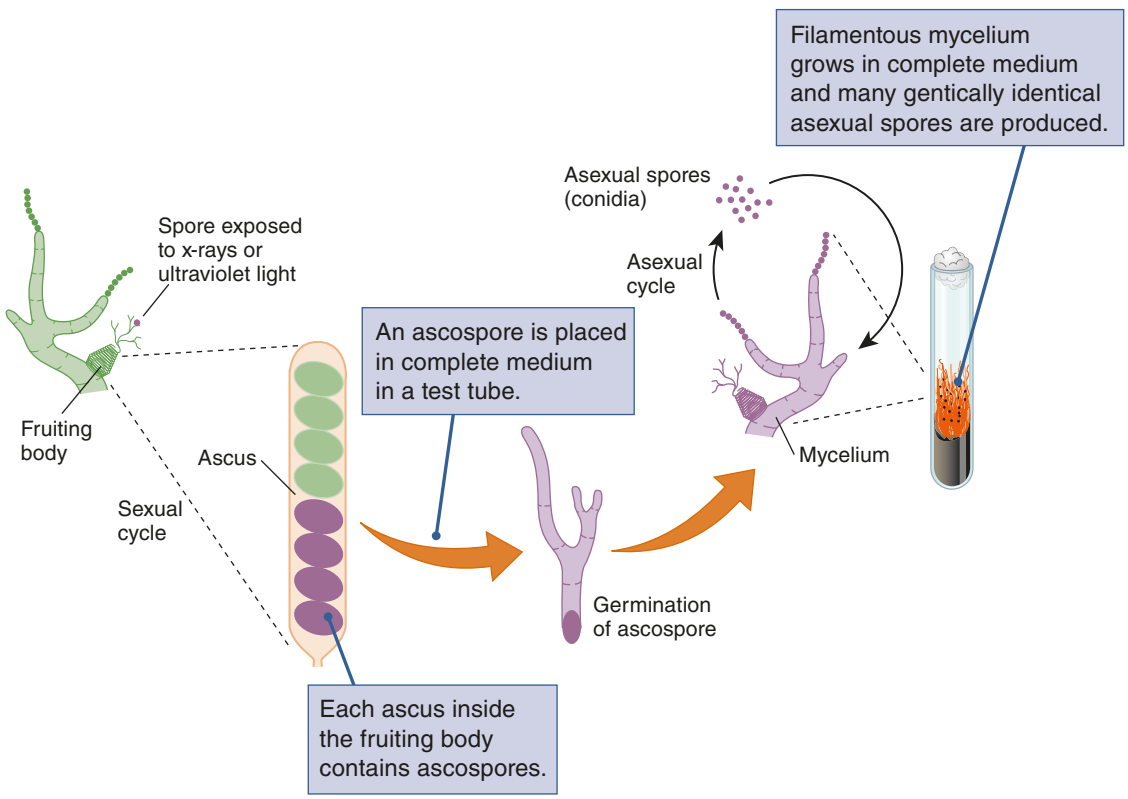


FIGURE 1.11 Beadle and Tatum obtained mutants of the filamentous fungus *Neurospora crassa* by exposing asexual spores to x-rays or ultraviolet light. The treated spores were used to start the sexual cycle in fruiting bodies. After any pair of cells and their nuclei undergo fusion, meiosis takes place almost immediately and results in eight sexual spores (ascospores) included in a single ascus. These are removed individually and cultured in complete medium. Ascospores that carry new nutritional mutants are identified later by their inability to grow in minimal medium.

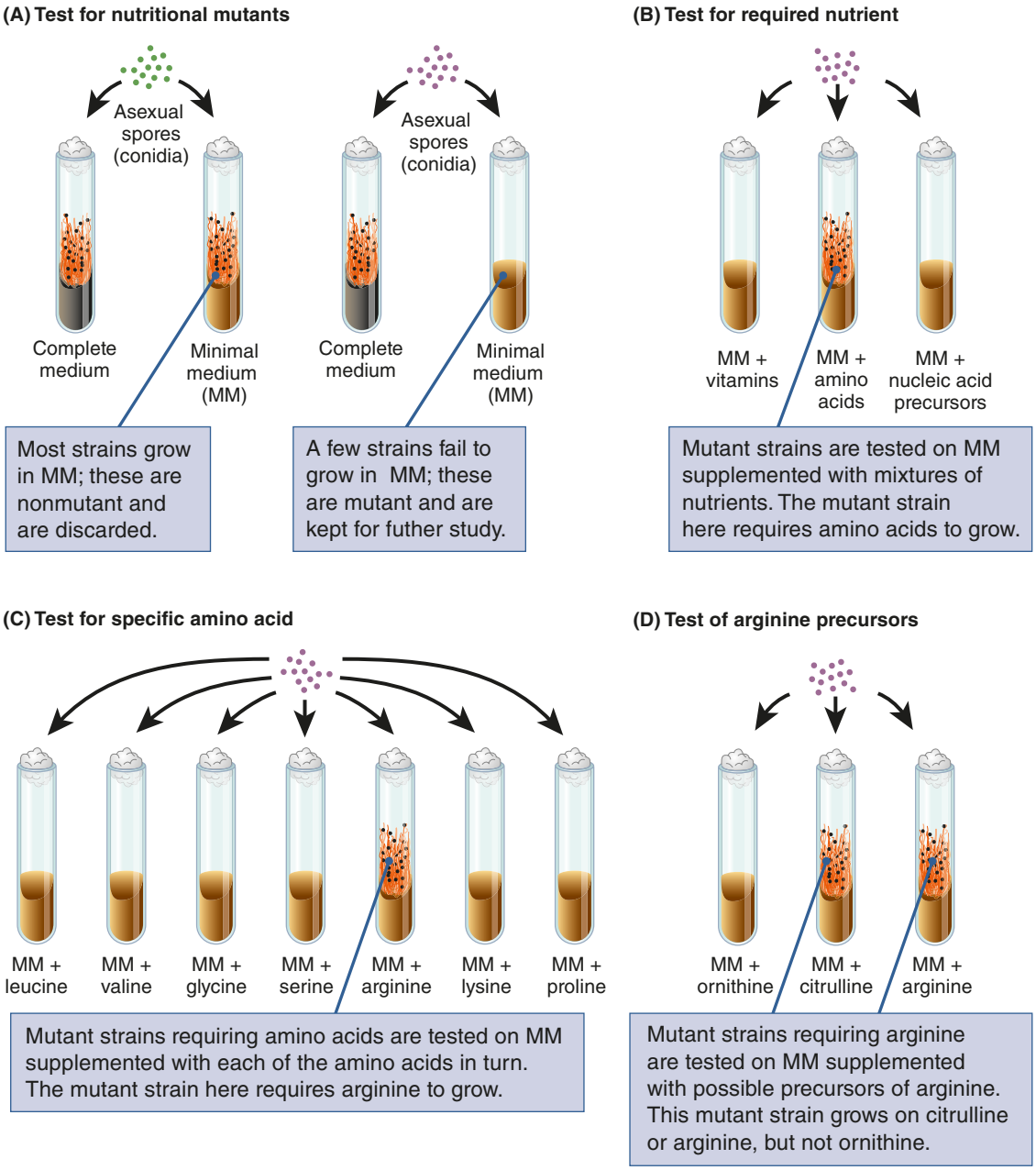


FIGURE 1.12 (A) Mutant spores can grow in complete medium but not in minimal medium. (B) Each new mutant is tested for growth in minimal medium supplemented with a mixture of nutrients. (C) Mutants that can grow on minimal medium supplemented with amino acid are tested with each amino acid individually. (D) Mutants unable to grow in the absence of arginine are tested with likely precursors of arginine.

For nutritional mutants requiring amino acids, further experiments testing each amino acid individually usually revealed that only one additional amino acid was needed in the minimal medium to support growth. In **FIGURE 1.12C**, the mutant strain requires the amino acid arginine. As early as the 1940s, some of the possible intermediates in amino acid biosynthesis had been identified. They were recognized by their chemical resemblance to the amino acid and by their presence at low levels in the cells of organisms. In the case of arginine, two candidates were ornithine and citrulline. All mutants requiring arginine, therefore, were tested in medium supplemented with either ornithine alone or citrulline alone (**FIGURE 1.12D**). One class of arginine-requiring mutants, designated Class I, was able to grow in minimal medium supplemented with either ornithine, citrulline, or arginine. Other mutants, designated Class II, were able to grow in minimal medium supplemented with either citrulline or arginine, but not ornithine. A third class, Class III, was able to grow only in minimal medium supplemented with arginine.

The types of arginine-requiring mutants illustrate the logic of genetic analysis as applied to metabolic pathways. The logic is easiest to see in the context of the metabolic pathway, shown in **FIGURE 1.13**, where arginine is the end product of a linear metabolic pathway starting with some precursor metabolite, and ornithine and citrulline are intermediates in the pathway. The mutants support this structure of the pathway for the following reasons:

- Mutants able to grow in the presence of either ornithine, citrulline, or arginine must have a metabolic block between the precursor metabolite and both of the intermediates.
- Mutants able to grow only in the presence of arginine must have a metabolic block in the pathway between arginine and the most "downstream" of the intermediates.
- Mutants able to grow in the presence of citrulline or arginine but not ornithine imply that ornithine is "upstream" from citrulline in the metabolic pathway.

The structure of the pathway was further confirmed by the observations that Class III mutants accumulate citrulline and Class II mutants accumulate ornithine. Finally, it was shown by direct biochemical experiment

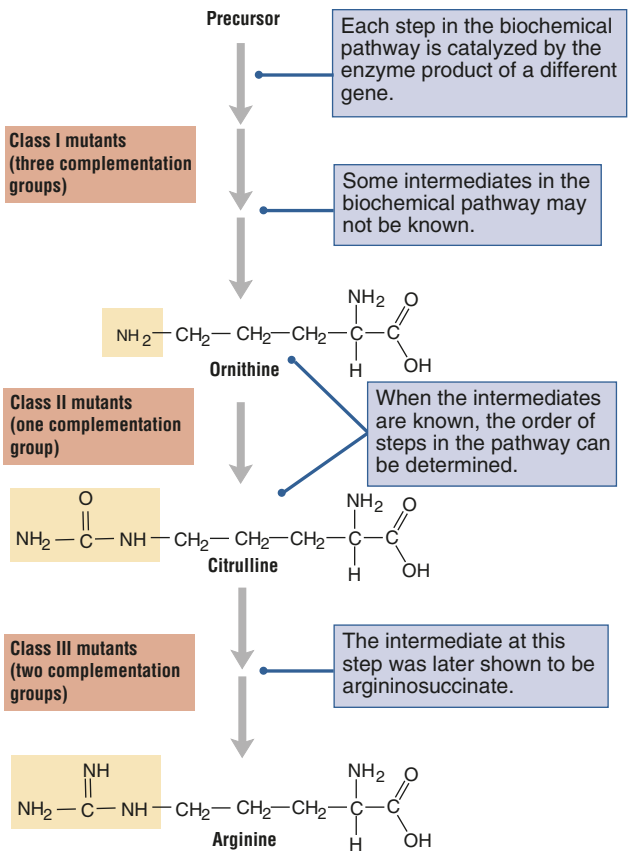


FIGURE 1.13 Metabolic pathway for arginine biosynthesis inferred from genetic analysis of *Neurospora* mutants.

that the inferred enzymes were actually present in nonmutant strains but either absent or nonfunctional in mutant strains.

Complementation Test for Mutations in the Same Gene

Beadle and Tatum were fortunate to study metabolic pathways in a relatively simple organism in which one gene corresponds to one enzyme. In such a situation, genetic analysis of the mutants reveals a great deal more about the metabolic pathway than merely the order of the intermediates. When each mutation is classified according to the particular gene it is in, and when all the mutations in each gene are grouped together, each set of mutations—and therefore each individual gene—corresponds to one enzymatic step in the metabolic pathway. In Figure 1.13, for example, the mutants in Class I each correspond to one of three genes, which implies that there are three steps in the pathway between the precursor and ornithine. Similarly, the mutants in Class III each correspond to one of two genes, which implies that there are two steps in the pathway between citrulline and arginine. However, all of the mutants in Class II have mutations in the same gene, which implies only one step in the pathway between ornithine and citrulline.

Mutations that have defects in the same gene are identified by means of a **complementation test**, in which two mutations are brought together into the same cell. In most multicellular organisms and even some sexual unicellular organisms, the usual way to do this is by means of a mating. When two parents, each carrying one of the two mutations, are crossed, fertilization brings the reproductive cells containing the two mutations together. As the result of ordinary cell division, each cell in the offspring then carries one copy of each mutant gene. In *Neurospora,* this procedure does not work because nuclear fusion is followed almost immediately by the formation of ascospores, each of which has only one set of chromosomes.

Complementation tests are nevertheless possible in *Neurospora* owing to the multinucleate nature of the filaments. Certain strains, including those studied by Beadle and Tatum, have the property that when the filaments from two mutant (or nonmutant) organisms come into physical contact, the filaments fuse and the new filament contains multiple nuclei from each of the participating partners. This sort of hybrid filament, which is called a **heterokaryon**, contains mutant forms of both genes. The word roots of the term *heterokaryon* mean "different nuclei." (A list of the most common word roots used in genetics can be found at the end of this text.)

When a heterokaryon formed from two nutritional mutants is inoculated into minimal medium, it may either grow or fail to grow. If it grows in minimal medium, the mutant genes are said to undergo

complementation, and this result indicates that the mutations are in different genes. Conversely, if the heterokaryon fails to grow in minimal medium, the result indicates **noncomplementation**, and the two mutations are inferred to be in the same gene.

To illustrate this concept, let's consider two hypothetical examples:

1. Heterokaryons are formed between Class II and Class III mutants. In this case, the nuclei containing the Class II mutants are unable to convert ornithine to citrulline, but they do produce the enzyme (or enzymes) necessary to convert citrulline to arginine. The Class III mutants, in contrast, can perform the ornithine-to-citrulline conversion, but they are unable to convert citrulline to arginine. In the same heterokaryon, therefore, both steps can occur, so that growth on minimal medium occurs. These two mutations are said to be **complementary**.
2. Heterokaryons are formed between two independently derived Class II mutations. Neither of these mutations can carry out the ornithine-to-citrulline step; hence no growth on minimal medium will occur. These mutations are **noncomplementary**.

Now let's consider two Class III mutations. Are they complementary? If the citrulline-to-arginine conversion includes only a single biochemical step, then we would expect them to be noncomplementary. Beadle and Tatum, however, discovered that some pairs of mutants did complement each other, whereas others did not. The theory that one gene codes for one enzyme led them to conclude that more than one enzyme is required for this step.

The inferences from complementation or noncomplementation emerge from the logic illustrated in **FIGURE 1.14**. Here the multinucleate filament is shown, and the mutant nuclei are color coded according to which of two different genes (red or purple) is mutant. The red and purple lines represent the proteins encoded in the mutant nuclei, and an asterisk represents a defect in the protein at that position resulting from a mutation in the corresponding gene.

Part A depicts the situation in which the mutant strains have mutations in different genes. In the heterokaryon, the red nuclei produce mutant forms of the red protein and normal forms of the purple protein, whereas the purple nuclei produce mutant forms of the purple protein and normal forms of the red protein. The result is that the red/purple heterokaryon has normal forms of both the red and purple protein. It also has mutant forms of both proteins, but these do not matter. What matters is that the normal forms allow the heterokaryon to grow on minimal medium because all needed nutrients can be synthesized. In other words, the normal purple gene in the red nucleus complements the defective purple gene in the purple nucleus, and vice versa. The logic of complementation is captured in the ancient nursery rhyme, "Jack Sprat could eat no fat / His wife could eat no lean / And so between the two of them / They licked the platter clean," because each partner makes up for the defect in the other.

Part B in Figure 1.14 shows a heterokaryon formed between mutants with defects in the same gene—in this case, purple. Both of the purple nuclei encode a normal form of the red protein, but each purple nucleus encodes a defective purple protein. When the nuclei are together, two different mutant forms of the purple protein are produced, so the biosynthetic pathway that requires the purple protein is still blocked and the heterokaryon is unable to grow in minimal medium. In other words, the mutants 2 and 3 in Figure 1.14 fail to complement, so they are judged to have mutations in the same gene.

The following principle underlies the complementation test:

> **Principle of Complementation:** A complementation test brings two mutant genes together in the same cell or organism. If this cell or organism is nonmutant, the mutations are said to complement each other, and the parental strains must have mutations in different genes. If the cell or organism is mutant, the mutations fail to complement each other, and the parental mutations must be in the same gene. These latter mutations form a **complementation group**.

Analysis of Complementation Data

In the mutant screen for *Neurospora* mutants requiring arginine, Beadle and Tatum found that mutants in different classes (Class I, Class II, and Class III in Figure 1.13) always complemented one another. This result makes sense because the genes in each class encode enzymes that act at different levels between the known intermediates. However, some of the mutants in Class I failed to complement others in Class I, and some in Class III failed to complement others in Class III. These results allow the number of genes in each class to be identified.

To illustrate this aspect of genetic analysis, we will consider six mutant strains in Class III. These strains were taken in pairs to form heterokaryons and their growth on minimal medium assessed. The data are shown in **FIGURE 1.15**. The mutant genes in the six strains are denoted *x1*, *x2*, and so forth, and the data are presented in the form of a matrix in which a plus sign (+) indicates growth in minimal medium (complementation) and a minus sign (−) indicates lack of growth in minimal medium (lack of complementation). The diagonal entries are all minus signs, which reflects the fact that two copies of the identical mutation cannot show complementation.

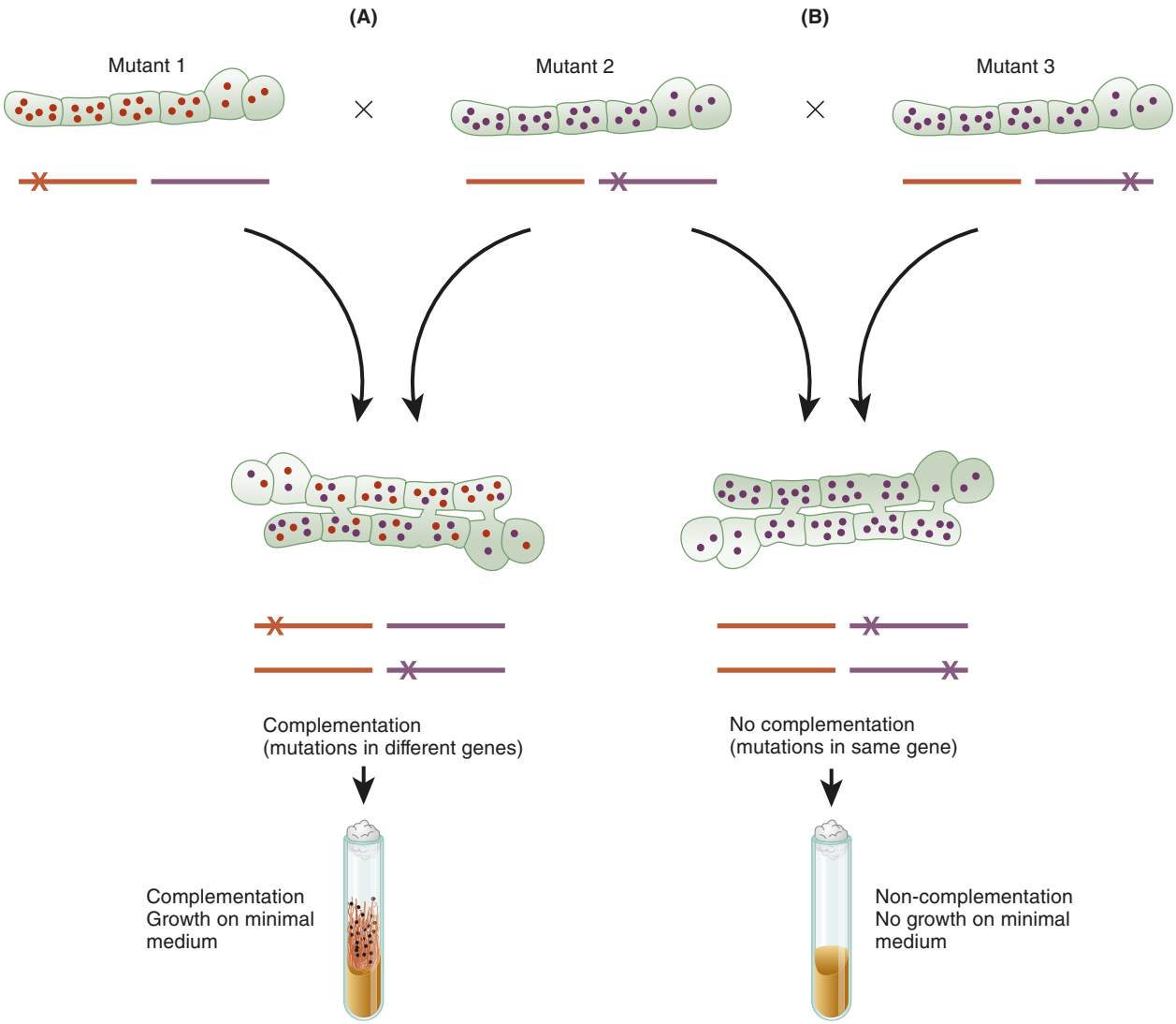


FIGURE 1.14 Molecular interpretation of a complementation test using heterokaryons to determine whether two mutant strains have mutations in different genes (A) or mutations in the same gene (B). In (A), each nucleus contributes a nonmutant form of one or the other polypeptide chain, so the heterokaryon is able to grow in minimal medium. In (B), both nuclei contribute a mutant form of the same polypeptide chain; hence, no nonmutant form of that polypeptide can be synthesized and the heterokaryon is unable to grow in minimal medium.

The pattern of + and − signs in the matrix indicates that mutations *x1* and *x5* fail to complement each other; hence *x1* and *x5* are mutations in the same gene. Likewise, mutations *x2*, *x3*, *x4*, and *x6* fail to complement one another in all pairwise combinations; hence *x2*, *x3*, *x4*, and *x6* are all mutations in the same gene, but a different gene from that represented by *x1* and *x5*.

Each of the groups of noncomplementary mutations is called a *complementation group*. As we have seen, each complementation group defines a gene, so the complementation test actually provides the geneticist's operational definition:

> A *gene* is defined experimentally as a set of mutations that make up a single complementation group. Any pair of mutations within a complementation group fails to complement each other.

The mutations in Figure 1.15 therefore represent two genes, and mutation of any one of them results in the inability of the strain to convert citrulline to arginine. On the assumption that one gene encodes one enzyme, which is largely true for metabolic enzymes in *Neurospora*, the pathway from citrulline to arginine depicted in Figure 1.13 must consist of two steps, with an unknown intermediate in between the steps. This intermediate was later found to be argininosuccinate. Likewise, the Class I mutants define three complementation groups, so there are three enzymatic steps from the precursor to ornithine. These intermediates also were soon identified. Finally, Class II mutations all fail to complement one another, and the finding of only one complementation group means that there is but a single enzymatic step that converts ornithine to citrulline.

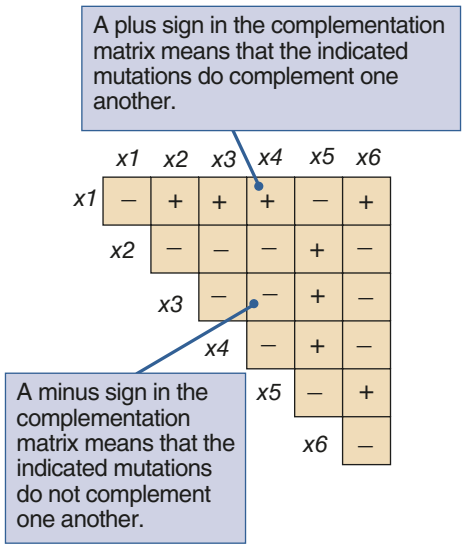


FIGURE 1.15 Interpreting the results of complementation tests. Rows and columns are particular, independently derived mutations. Pluses indicate pairs that complement; minuses are ones that do not. A complementation group is defined as all mutations that do *not* complement one another but *do* complement all other mutations. In this case, *x1* and *x4* constitute one group, while *x2, x3, x4,* and *x6* make up a second group.

Other Applications of Genetic Analysis

The type of genetic analysis pioneered by Beadle and Tatum is immensely powerful for identifying the genetic control of complex biological processes. Their approach lays out a systematic path—a cookbook if you will—for gene discovery, one that we will see used repeatedly throughout this text:

1. Decide which process you want to study, and figure out which characteristics would be displayed by mutant organisms with a disruption in that process.
2. Perform mutant screening for mutants showing these characteristics.
3. Carry out complementation tests to find out how many different genes you have identified.
4. Identify the products of those genes, and determine what they do, how they interact with each other, and in which order they function.

Beadle and Tatum analyzed many metabolic pathways for a wide variety of essential nutrients, but their experiments were especially important in deciphering the pathways of amino acid biosynthesis. Their findings over the span of just a few years were remarkable, in that they deduced much more about the nature of biosynthetic pathways than had been learned from decades of biochemical research. Beadle and Tatum were awarded the 1958 Nobel Prize in Physiology or Medicine for their research, and in the intervening years many more Nobel Prizes in Physiology or Medicine were awarded in which genetic analysis carried out along the lines of Beadle and Tatum played a significant role. Here is a list, with quotations from the official citations of the Nobel Foundation:

- 1958: George Beadle and Edward Tatum "for their discovery that genes act by regulating definite chemical events," shared with Joshua Lederberg "for his discoveries concerning genetic recombination and the organization of the genetic material of bacteria."
- 1965: François Jacob, André Lwoff, and Jacques Monod "for their discoveries concerning genetic control of enzyme and virus synthesis."
- 1995: Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus "for their discoveries concerning the genetic control of early embryonic development."
- 2001: Leland H. Hartwell, Tim Hunt, and Sir Paul Nurse "for their discoveries of key regulators of the cell cycle."
- 2002: Sydney Brenner, H. Robert Horvitz, and John E. Sulston "for their discoveries concerning genetic regulation of organ development and programmed cell death."
- 2007: Mario R. Capecchi, Martin J. Evans, and Oliver Smithies "for their discoveries of principles for introducing specific gene modifications in mice by the use of embryonic stem cells."
- 2009: Elizabeth Blackburn, Carol Greider, and Jack Szostak "for the discovery of how chromosomes are protected by telomeres and the enzyme telomerase."
- 2013: James Rothman, Randy Shekman, and Thomas Sudhof "for their discoveries of machinery regulating vesicle traffic, a major transport system in our cells."
- 2015: Tomas Lindahl, Paul Modrich, and Aziz Sancar "for mechanistic studies of DNA repair."

The Beadle and Tatum experiments established that a defective enzyme results from a mutant gene, but how? For all they knew, genes *were* enzymes. This would have been a logical hypothesis at the time. We now know that the relationship between genes and enzymes is somewhat indirect. With a few exceptions, each enzyme is *encoded* in a particular sequence of nucleotides present in a region of DNA. The DNA region that codes for the enzyme, as well as adjacent regions that regulate when and in which cells the enzyme is produced, make up the "gene" that encodes the enzyme. Next, we turn to the issue of *how* genes code for enzymes and other proteins.

SUMMING UP

- Genetic analysis is an experimental methodology whereby inferences based on the outcomes of genetic crosses can be used to elucidate a wide variety of biological processes.
- Using *Neurospora crassa*, Beadle and Tatum developed the methods of mutant screening.
- Complementation analysis is a powerful means for determining the number of steps in a biochemical pathway.
- These analyses led Beadle and Tatum to the "one gene, one enzyme" hypothesis.

1.5 Gene Expression: The Central Dogma

Watson and Crick were correct in proposing that the genetic information in DNA is contained in the sequence of bases in a manner analogous to letters printed on a strip of paper. In a region of DNA that directs the synthesis of a protein, the genetic code for the protein is contained in only one strand, and it is decoded in a linear order. A typical protein is made up of one or more polypeptide chains; each **polypeptide chain** consists of a linear sequence of amino acids connected end to end. For example, the enzyme PAH consists of four identical polypeptide chains, each 452 amino acids in length. In the decoding of DNA, each successive "code word" in the DNA specifies the next amino acid to be added to the polypeptide chain as it is being made. The amount of DNA required to code for the polypeptide chain of PAH is therefore $452 \times 3 \nabla 1356$ nucleotide pairs. The entire gene is very much longer—about 90,000 nucleotide pairs. Only 1.5 percent of the gene is devoted to coding for the amino acids. The noncoding part includes some sequences that control the activity of the gene, but it is not known how much of the gene is involved in regulation.

There are 20 different amino acids encoded by DNA. Only four bases code for these 20 amino acids, with each "word" in the genetic code consisting of three adjacent bases. For example, the base sequence ATG specifies the amino acid methionine (Met), TCC specifies serine (Ser), ACT specifies threonine (Thr), and GCG specifies alanine (Ala). There are 64 possible three-base combinations but only 20 amino acids because some combinations specify the same amino acid. For example, TCT, TCC, TCA, TCG, AGT, and AGC all code for serine (Ser), and CTT, CTC, CTA, CTG, TTA, and TTG all code for leucine (Leu).

An example of the relationship between the nucleotide sequence in a DNA duplex and the amino acid sequence of the corresponding protein is shown in **FIGURE 1.16**. This particular DNA duplex is the human sequence that codes for the first seven amino acids in the polypeptide chain of PAH. The scheme outlined in

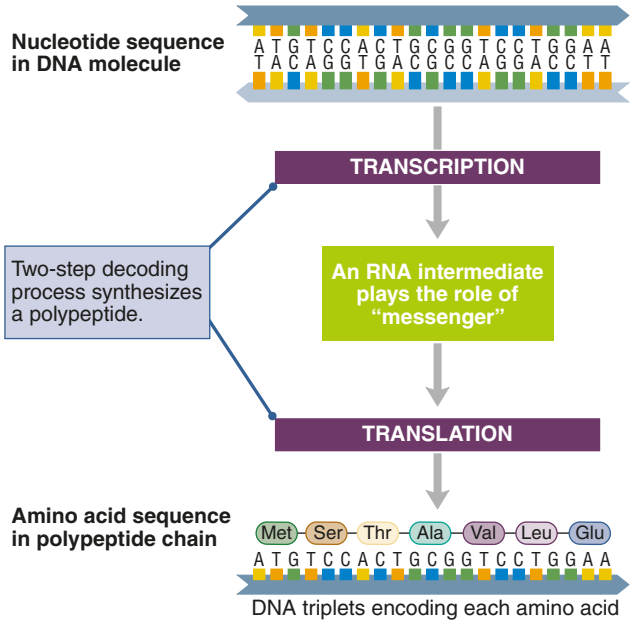


FIGURE 1.16 DNA sequence coding for the first seven amino acids in a polypeptide chain. The DNA sequence specifies the amino acid sequence through a molecule of RNA that serves as an intermediary "messenger." Although the decoding process is indirect, the net result is that each amino acid in the polypeptide chain is specified by a group of three adjacent bases in the DNA. In this example, the polypeptide chain is that of phenylalanine hydroxylase (PAH).

Figure 1.16 indicates that DNA codes for protein not directly but indirectly through the processes of *transcription* and *translation*. The indirect route of information transfer is thus:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

This relationship leads us to the **central dogma** of molecular genetics. The term *dogma* means "set of beliefs"; it dates from the time the idea was put forward first as a theory. Since then the "dogma" has been confirmed experimentally, but the term persists. The main concept in the central dogma, as stated by Francis Crick in 1958, is as follows:

> Once "information" has passed into protein, it *cannot get out again.*

With respect to gene expression in most cells, the central dogma can be illustrated as in **FIGURE 1.17**. The "information" is that which is encoded in DNA, and which specifies the sequences of proteins. Its passage from DNA to protein does not occur directly, but rather through interactions with the intermediary molecule known as **ribonucleic acid (RNA)**. The structure of RNA is similar, but not identical, to the structure of DNA. There is a difference in the sugars in DNA and RNA (RNA contains the sugar **ribose** instead of deoxyribose), RNA is usually single-stranded (not a duplex), and RNA contains the base **uracil (U)** instead of thymine

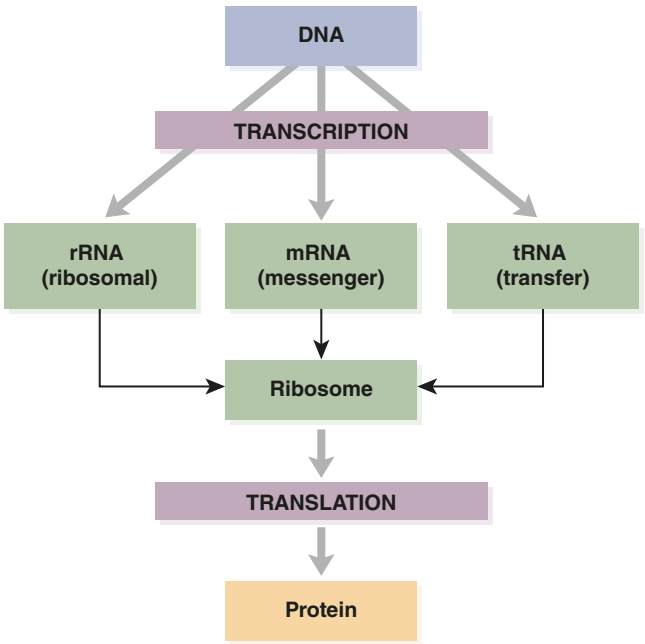


FIGURE 1.17 Gene expression at the molecular level. DNA codes for RNA, and RNA codes for protein. The DNA → RNA step is transcription, and the RNA → protein step is translation.

(T), which is present in DNA. Information flow between DNA and RNA follows the rules of Watson–Crick base pairing, and as we will see elsewhere in this text, in some cases it can occur in both directions (DNA → RNA or RNA → DNA). However, information flow from RNA to proteins requires **translation** of nucleic acid sequences into polypeptide sequences. This process is irreversible. There is no known case in which the sequence of a protein codes or the sequence of a nucleic acid (either DNA or RNA).

Three types of RNA take part in the synthesis of proteins:

- A molecule of **messenger RNA (mRNA)**, which carries the genetic information from DNA and is used as a template for polypeptide synthesis. In most mRNA molecules, there is a high proportion of nucleotides that actually code for amino acids. For example, the mRNA for PAH is 2400 nucleotides in length and codes for a polypeptide of 452 amino acids; in this case, more than 50 percent of the length of the mRNA codes for amino acids.
- Four types of **ribosomal RNA (rRNA)**, which are major constituents of the cellular particles called **ribosomes** on which polypeptide synthesis takes place.
- A set of approximately 45 **transfer RNA (tRNA)** molecules, each of which carries a particular amino acid as well as a three-base recognition region that base-pairs with a group of three adjacent bases in the mRNA. As each tRNA participates in translation, its amino acid becomes the terminal subunit added to the length of the growing polypeptide chain. A tRNA that carries methionine is denoted $tRNA^{Met}$, one that carries serine is denoted $tRNA^{Ser}$, and so forth. (Because there are more than 20 different tRNAs but only 20 amino acids, some amino acids correspond to more than one tRNA.)

The central dogma is the fundamental principle of molecular genetics because it summarizes how the genetic information in DNA becomes expressed in the amino acid sequence in a polypeptide chain:

> The sequence of nucleotides in a gene specifies the sequence of nucleotides in a molecule of messenger RNA; in turn, the sequence of nucleotides in the messenger RNA specifies the sequence of amino acids in the polypeptide chain.

Given a process as conceptually simple as DNA coding for protein, what might account for the additional complexity of RNA intermediaries? One possible explanation is that an RNA intermediate gives another level for control—for example, by degrading the mRNA for an unneeded protein.

Another possible reason may be historical. RNA structure is unique in having both informational content present in its sequence of bases and a complex, folded three-dimensional structure that endows some RNA molecules with catalytic activities. Many scientists believe that in the earliest forms of life, RNA had roles both as genetic information and in catalysis. As evolution proceeded, the informational role was transferred to DNA and the catalytic role to protein. However, RNA became locked into its central location as a go-between in the processes of information transfer and protein synthesis. This hypothesis implies that the participation of RNA in protein synthesis is a relic of the earliest stages of evolution—a "molecular fossil." It is supported by a variety of observations. For example, (1) DNA replication requires an RNA molecule to get started, (2) an RNA molecule is essential in the synthesis of the tips of the chromosomes, and (3) some RNA molecules act to catalyze key reactions in protein synthesis.

Transcription

FIGURE 1.18 depicts the manner in which genetic information is transferred from DNA to RNA. The DNA opens up, and one of the strands is used as a template for the synthesis of a complementary strand of RNA. (How the template strand is chosen is discussed in the *Molecular Biology of Gene Expression* chapter.) The process of making an RNA strand from a DNA template is called **transcription**, and the RNA molecule that is made is the **transcript**. The base sequence in the RNA is complementary (in the Watson–Crick pairing sense)

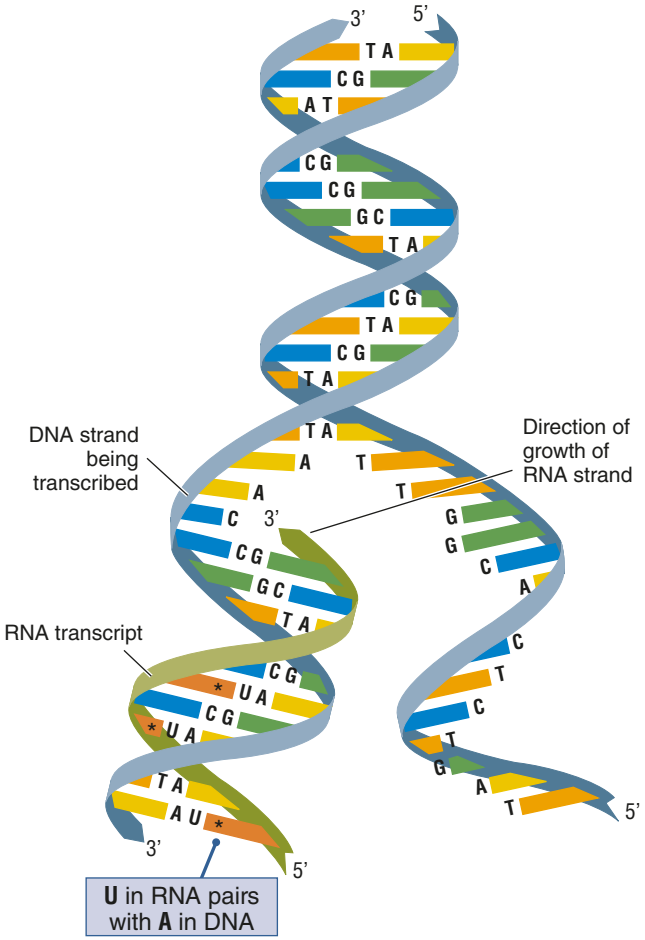


FIGURE 1.18 Transcription is the production of an RNA strand that is complementary in base sequence to a DNA strand. In this example, the DNA strand at the bottom is being transcribed into a strand of RNA. Note that in an RNA molecule, the base U (uracil) plays the role of T (thymine), in that it pairs with A (adenine). Each A–U pair is marked.

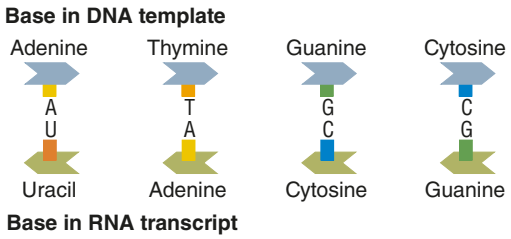


FIGURE 1.19 Pairing between bases in DNA and in RNA. The DNA bases A, T, G, and C pair with the RNA bases U, A, C, and G, respectively.

to that in the DNA template, except that U (which pairs with A) is present in the RNA in place of T. The rules of base pairing between DNA and RNA are summarized in **FIGURE 1.19**.

Each RNA strand has a polarity—a 5′ end and a 3′ end—and, as in the synthesis of DNA, nucleotides are added only to the 3′ end of a growing RNA strand. Hence, the 5′ end of the RNA transcript is synthesized first, and transcription proceeds along the template DNA strand in the 3′-to-5′ direction.

Each gene includes nucleotide sequences that initiate and terminate transcription. The RNA transcript made from any gene begins at the initiation site in the template strand, which is located "upstream" from the amino acid–coding region, and ends at the termination site, which is located "downstream" from the amino acid–coding region. For any gene, the length of the RNA transcript is very much smaller than the length of the DNA in the chromosome. For example, the transcript of the *PAH* gene for phenylalanine hydroxylase is about 90,000 nucleotides in length, but the DNA in chromosome 12 is about 130,000,000 nucleotide pairs. In this case, the length of the *PAH* transcript is less than 0.1 percent of the length of the DNA in the chromosome. A different gene in chromosome 12 would be transcribed from a different region of the DNA molecule in chromosome 12, and perhaps from the opposite strand, but the transcribed region would again be small in comparison with the total length of the DNA in the chromosome.

Translation

The synthesis of a polypeptide under the direction of an mRNA molecule is known as **translation**. Although the sequence of bases in the mRNA codes for the sequence of amino acids in a polypeptide, the molecules that actually do the "translating" are the tRNA molecules. The mRNA molecule is translated in nonoverlapping groups of three bases called **codons**. For each codon in the mRNA that specifies an amino acid, there is one tRNA molecule containing a complementary group of three adjacent bases that can pair with the codon. The correct amino acid is attached to one end of the tRNA, and when the tRNA comes into line, the amino acid to which it is attached becomes the most recent addition to the growing end of the polypeptide chain.

The role of tRNA in translation is illustrated in **FIGURE 1.20** and can be described as follows:

> The mRNA is read codon by codon. Each codon that specifies an amino acid matches with a complementary group of three adjacent bases in a single tRNA molecule. One end of the tRNA is attached to the correct amino acid, so the correct amino acid is brought into line.

The tRNA molecules used in translation do not line up along the mRNA simultaneously, as shown in Figure 1.20. The process of translation takes place on a ribosome, which combines with a single mRNA and moves along it from one end to the other in steps, three nucleotides at a time (codon by codon). As each new codon comes into place, the next tRNA binds with the ribosome. Then the growing end of the polypeptide chain becomes attached to the amino acid on the tRNA. In this way, each tRNA in turn serves temporarily to hold the polypeptide chain as it is being synthesized. As the

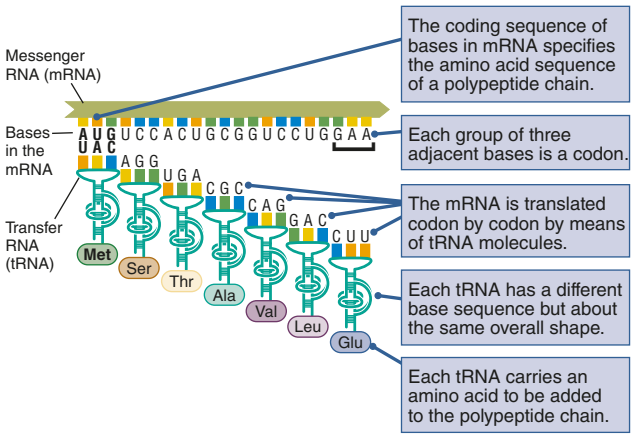


FIGURE 1.20 The role of messenger RNA in translation is to carry the information contained in a sequence of DNA bases to a ribosome, where it is translated into a polypeptide chain. Translation is mediated by transfer RNA (tRNA) molecules, each of which can base-pair with a group of three adjacent bases in the mRNA. Each tRNA also carries an amino acid. As each tRNA, in turn, is brought to the ribosome, the growing polypeptide chain is elongated.

polypeptide chain is transferred from each tRNA to the next in line, the tRNA that previously held the polypeptide is released from the ribosome. The polypeptide chain elongates one amino acid at each step until any one of three particular codons specifying "stop" is encountered.

At this point, synthesis of the chain of amino acids is finished, and the polypeptide chain is released from the ribosome. (This brief description of translation glosses over many of the details that are presented in the *Molecular Biology of Gene Expression* chapter.)

The Genetic Code

Figure 1.20 indicates that the mRNA codon AUG specifies methionine (Met) in the polypeptide chain, UCC specifies Ser (serine), ACU specifies Thr (threonine), and so on. The complete decoding table is called the **genetic code**, and it is shown in **TABLE 1.1**. For any codon, the column on the left corresponds to the first nucleotide in the codon (reading from the 5′ end), the row across the top corresponds to the second nucleotide, and the column on the right corresponds to the third nucleotide. The complete codon is given in the body of the table, along with the amino acid (or translational "stop") that the codon specifies. Each amino acid is designated by its full name and by a three-letter abbreviation as well as a single-letter abbreviation. Both types of abbreviations are used in molecular genetics.

The code in Table 1.1 is the "standard" genetic code used in translation in the cells of nearly all organisms. In the *Molecular Biology of Gene Expression* chapter, we examine general features of the standard genetic code and the minor differences found in the genetic codes of

TABLE 1.1 The standard genetic code

First nucleotide in codon (5′ end)	Second nucleotide in codon: U				Second nucleotide in codon: C				Second nucleotide in codon: A				Second nucleotide in codon: G				Third nucleotide in codon (3′ end)
U	UUU	Phe	F	*Phenylalanine*	UCU	Ser	S	*Serine*	UAU	Tyr	Y	*Tyrosine*	UGU	Cys	C	*Cysteine*	U
	UUC	Phe	F	*Phenylalanine*	UCC	Ser	S	*Serine*	UAC	Tyr	Y	*Tyrosine*	UGC	Cys	C	*Cysteine*	C
	UUA	Leu	L	*Leucine*	UCA	Ser	S	*Serine*	UAA			Termination	UGA			Termination	A
	UUG	Leu	L	*Leucine*	UCG	Ser	S	*Serine*	UAG			Termination	UGG	Trp	W	*Tryptophan*	G
C	CUU	Leu	L	*Leucine*	CCU	Pro	P	*Proline*	CAU	His	H	*Histidine*	CGU	Arg	R	*Arginine*	U
	CUC	Leu	L	*Leucine*	CCC	Pro	P	*Proline*	CAC	His	H	*Histidine*	CGC	Arg	R	*Arginine*	C
	CUA	Leu	L	*Leucine*	CCA	Pro	P	*Proline*	CAA	Gln	Q	*Glutamine*	CGA	Arg	R	*Arginine*	A
	CUG	Leu	L	*Leucine*	CCG	Pro	P	*Proline*	CAG	Gln	Q	*Glutamine*	CGG	Arg	R	*Arginine*	G
A	AUU	Ile	I	*Isoleucine*	ACU	Thr	T	*Threonine*	AAU	Asn	N	*Asparagine*	AGU	Ser	S	*Serine*	U
	AUC	Ile	I	*Isoleucine*	ACC	Thr	T	*Threonine*	AAC	Asn	N	*Asparagine*	AGC	Ser	S	*Serine*	C
	AUA	Ile	I	*Isoleucine*	ACA	Thr	T	*Threonine*	AAA	Lys	K	*Lysine*	AGA	Arg	R	*Arginine*	A
	AUG	Met	M	*Methionine*	ACG	Thr	T	*Threonine*	AAG	Lys	K	*Lysine*	AGG	Arg	R	*Arginine*	G
G	GUU	Val	V	*Valine*	GCU	Ala	A	*Alanine*	GAU	Asp	D	*Aspartic acid*	GGU	Gly	G	*Glycine*	U
	GUC	Val	V	*Valine*	GCC	Ala	A	*Alanine*	GAC	Asp	D	*Aspartic acid*	GGC	Gly	G	*Glycine*	C
	GUA	Val	V	*Valine*	GCA	Ala	A	*Alanine*	GAA	Glu	E	*Glutamic acid*	GGA	Gly	G	*Glycine*	A
	GUG	Val	V	*Valine*	GCG	Ala	A	*Alanine*	GAG	Glu	E	*Glutamic acid*	GGG	Gly	G	*Glycine*	G

Codon — Three-letter and single-letter abbreviations

certain organisms and cellular organelles. At this point, we are interested mainly in understanding how the genetic code is used to translate the codons in mRNA into the amino acids in a polypeptide chain.

In addition to the 61 codons that code only for amino acids, there are four codons that have specialized functions:

- The codon AUG, which specifies Met (methionine), is also the "start" codon for polypeptide synthesis. The positioning of a $tRNA^{Met}$ bound to AUG is one of the first steps in the initiation of polypeptide synthesis, so all polypeptide chains begin with Met. (Many polypeptides have the initial Met cleaved off after translation is complete.) In most organisms, the $tRNA^{Met}$ used for initiation of translation is the same $tRNA^{Met}$ used to specify methionine at internal positions in a polypeptide chain.
- The codons UAA, UAG, and UGA are each a "stop" that specifies the termination of translation and results in release of the completed polypeptide chain from the ribosome. These codons do not have tRNA molecules that recognize them, but are instead recognized by protein factors that terminate translation.

How the genetic code table is used to infer the amino acid sequence of a polypeptide chain can be illustrated by using PAH again—in particular, the DNA sequence coding for amino acids 1 through 7. The DNA sequence is:

```
5′-ATGTCCACTGCGGTCCTGGAA-3′
3′-TACAGGTGACGCCAGGACCTT-5′
```

This region is transcribed into RNA in a left-to-right direction. Because RNA grows by the addition of successive nucleotides to the 3′ end (Figure 1.18), it is the bottom strand that is transcribed. The nucleotide sequence of the RNA is that of the top strand of the DNA, except that U replaces T, so the mRNA for amino acids 1 through 7 is:

```
5′-AUGUCCACUGCGGUCCUGGAA-3′
```

The codons are read from left to right according to the genetic code shown in Table 1.1. Codon AUG codes for Met (methionine), UCC codes for Ser (serine), and so on. Altogether, the amino acid sequence of this region of the polypeptide is:

```
5′-AUG|UCC|ACU|GCG|GUC|CUG|GAA-3′
   Met|Ser|Thr|Ala|Val|Leu|Glu
```

or, in terms of the single-letter abbreviations:

```
5′-AUG|UCC|ACU|GCG|GUC|CUG|GAA-3′
   M  | S | T | A | V | L | E
```

The full decoding operation for this region of the *PAH* gene is shown in **FIGURE 1.21**. In this figure, the initiation codon AUG is highlighted because some patients with PKU have a mutation in this particular codon. As might be expected from the fact that AUG is the initiation codon for polypeptide synthesis, cells in patients with this particular mutation fail to produce any of the PAH polypeptide. Mutation and its consequences are considered next.

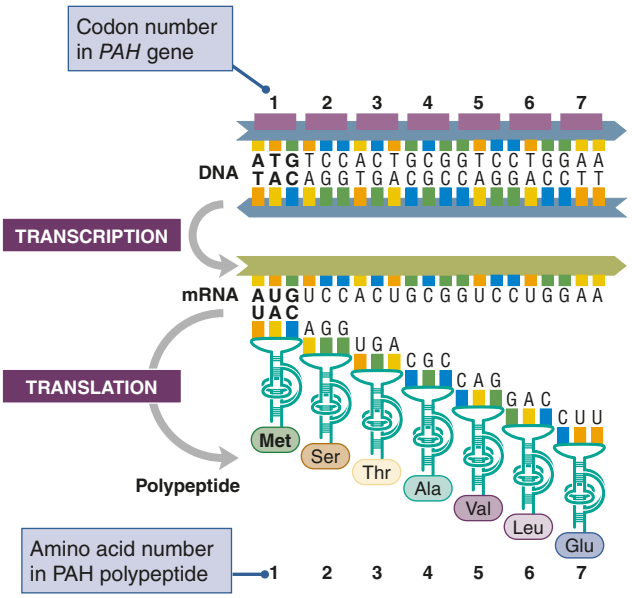


FIGURE 1.21 The central dogma in typical gene expression. The DNA that encodes PAH serves as a template for the production of a messenger RNA, and the mRNA serves to specify the sequence of amino acids in the PAH polypeptide chain through interactions with the ribosome and tRNA molecules. The total number of amino acids in the PAH polypeptide chain is 452, but only the first 7 are shown.

SUMMING UP

- Proteins consist of polypeptides—that is, chains of amino acids whose order is determined by the sequences of bases in the genes encoding them.
- The central dogma of molecular biology states that information, in the form of molecular sequences, passes from nucleic acids to proteins, but not the reverse.
- In gene expression, this process involves transcription of DNA into messenger RNA and translation of mRNA into protein.
- In addition to mRNA, ribosomal RNA and transfer RNA play critical roles in the translation process.
- The standard genetic code consists of 64 three-letter words, or codons, 61 of which encode amino acids.

THE WOMEN IN THE WEDDING photograph are sisters. Both have two copies of the same mutant *PAH* gene. The bride is the younger of the two. She was diagnosed just three days after birth and put on the PKU diet soon after. Her older sister, the maid of honor, was diagnosed too late to begin the diet and is intellectually disabled. The two-year old pictured in the photo at the right is the daughter of the married couple. They planned the pregnancy: Dietary control was strict from conception to delivery to avoid the hazards of excess phenylalanine harming the fetus. Their daughter has passed all developmental milestones with distinction.
Courtesy of Charles R. Scriver, Montreal Children's Hospital Research Institute, McGill University Health Center.

1.6 Mutation and Variation

We now turn to the fourth essential property of the genetic material—namely, its ability to generate variation through mutation. The term **mutation** refers to any heritable change in a gene (or, more generally, in the genetic material) or to the process by which such a change takes place. One type of mutation results in a change in the sequence of bases in DNA. This kind of change may be simple, such as the substitution of one pair of bases in a duplex molecule for a different pair of bases. For example, a C–G pair in a duplex molecule may mutate to T–A, A–T, or G–C. The change in nucleotide sequence may also be more complex, such as the deletion or addition of base pairs. These and other types of mutations are considered in the *Mutation, Repair, and Recombination* chapter.

Geneticists also use the term **mutant**, which refers to the result of a mutation. A mutation yields a mutant gene, which in turn produces a mutant mRNA, a mutant protein, and finally a mutant organism that exhibits the effects of the mutation—for example, an inborn error of metabolism.

DNA from patients from all over the world who have phenylketonuria has been studied to determine which types of mutations are responsible for the inborn error. A large variety of mutant types have been identified. More than 400 different mutations have been described in the gene for PAH. In some cases part of the gene is missing, so the genetic information to make a complete PAH enzyme is absent. In other cases the genetic defect is more subtle, but the result is still either the failure to produce a PAH protein or the production of an inactive PAH protein. In the mutation shown in **FIGURE 1.22**, substitution of a G–C base pair for the normal A–T base pair at the very first position in the coding sequence changes the normal codon AUG (Met) used for the initiation of translation into the codon GUG, which normally specifies valine (Val) and cannot be used as a "start" codon. The result is that translation of the PAH mRNA cannot occur,

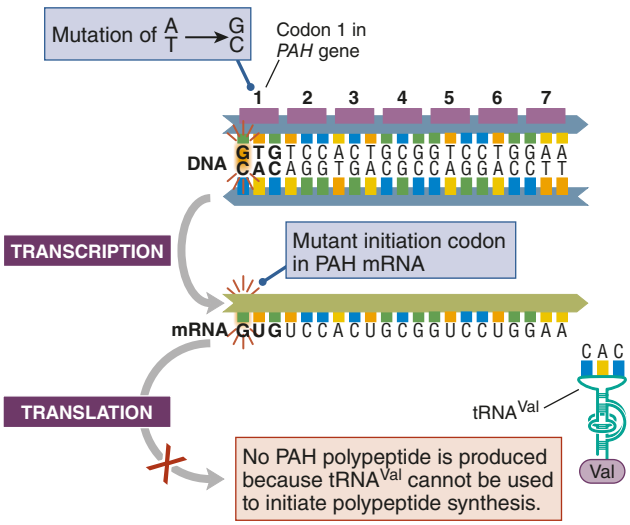


FIGURE 1.22 The M1V mutant in the *PAH* gene. The methionine codon needed for initiation mutates into a codon for valine. Translation cannot be initiated, and no PAH polypeptide is produced.

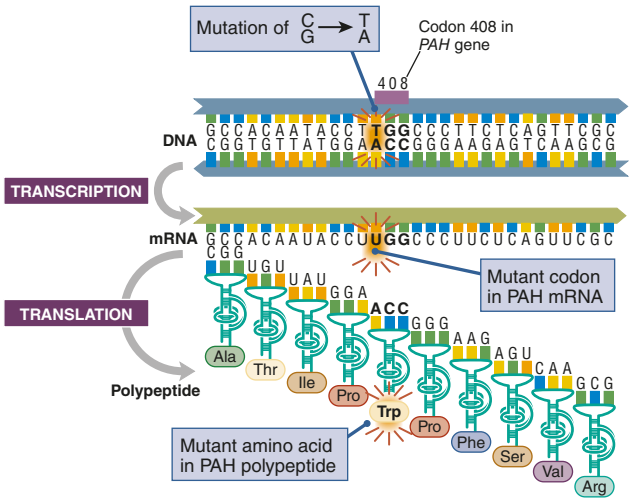


FIGURE 1.23 The R408W mutant in the *PAH* gene. Codon 408 for arginine (R) is mutated into a codon for tryptophan (W). The result is that position 408 in the mutant PAH polypeptide is occupied by tryptophan rather than by arginine. The mutant protein has only a low level of PAH enzyme activity.

so no PAH polypeptide is made. This mutant is designated M1V because the codon for M (methionine) at amino acid position 1 in the PAH polypeptide has been changed to a codon for V (valine). Although the M1V mutant is quite rare worldwide, it is common in some localities, such as Québec Province in Canada.

One PAH mutant that is quite common is designated R408W, which means that codon 408 in the PAH polypeptide chain has been changed from one coding for arginine (R) to one coding for tryptophan (W). This mutation is one of the four most common among European Caucasians with PKU. The molecular basis of the mutant is shown in **FIGURE 1.23**. In this case, the first base pair in codon 408 is changed from C–G base into T–A. As a result, the PAH mRNA has a mutant codon at position 408; specifically, it has UGG instead of CGG. Translation does occur in this mutant because everything else about the mRNA is normal, but the result is that the mutant PAH carries a tryptophan (Trp) instead of an arginine (Arg) at position 408 in the polypeptide chain. The consequence of the seemingly minor change of one amino acid is very drastic. Although the R408W polypeptide is complete, the enzyme has less than 3 percent of the activity of the normal enzyme.

Variation in Populations

Much of the classic work in genetics, from Gregor Mendel onward, involved genetic analysis of visible differences in organismal traits that are the results of particular mutations. Most of the variants used, such as white eyes in *Drosophila*, shrunken kernels in maize, or even inherited diseases like alkaptonuria in humans, are quite rare. Thus, early geneticists tended to think of populations as consisting of a few rare variants among large numbers of "wild type" individuals, suggesting that overall levels of genetic variation were quite low.

This picture later changed dramatically as, based on the one geneone enzyme hypothesis, it became possible to quantify genetic variation—first at the level of protein sequence, and subsequently at the level of DNA sequence. Early work with protein variation in both fruit flies (*Drosophila pseudoobscura*) and humans yielded unexpected findings: When the protein products of several genes were examined, approximately 30 percent of them were found to be **polymorphic**, meaning that there were two or more variants, or **alleles**, for those genes present in the sampled individuals. Similar findings were made for almost every species examined with these methods, leading to the recognition that variation—rather than uniformity—is the norm at the genetic level in most species.

In recent years, the methods we describe in the chapter titled *DNA Replication and Sequencing* have allowed investigators to change their focus from single genes to entire genomes. One outcome of that effort has been the *1000 Genomes Project*, an ongoing international research program seeking to survey common human genetic variation and to identify common human genetic variants associated with disease. To date, the project has sequenced 2504 genomes from 26 worldwide populations, which have been divided into 5 "superpopulations" (**FIGURE 1.24**). More than 80 million polymorphisms have been identified, most of which consist of single nucleotide differences (**single-nucleotide polymorphisms**, or SNPs). These publicly available data provide a global perspective on genetic variation in *Homo sapiens*, and have become an invaluable tool for researchers in medical genetics, evolutionary biology, and anthropology. Most importantly, genomic analysis of humans and other species has shown that the classical view, which suggests genetic variation consists of common wild-type alleles and rare variants, needs to be revised substantially. Much of the genome of a particular species is indeed invariant, but for significant portions of it, variation is the rule rather than the exception.

SUMMING UP

- All new genetic variation originates as a mutation, or change in DNA sequence.
- Different, independently occurring mutations may occur in the same gene.
- Genomic sequencing has provided powerful tools to characterize the extent and nature of genetic variation in many species, including humans.

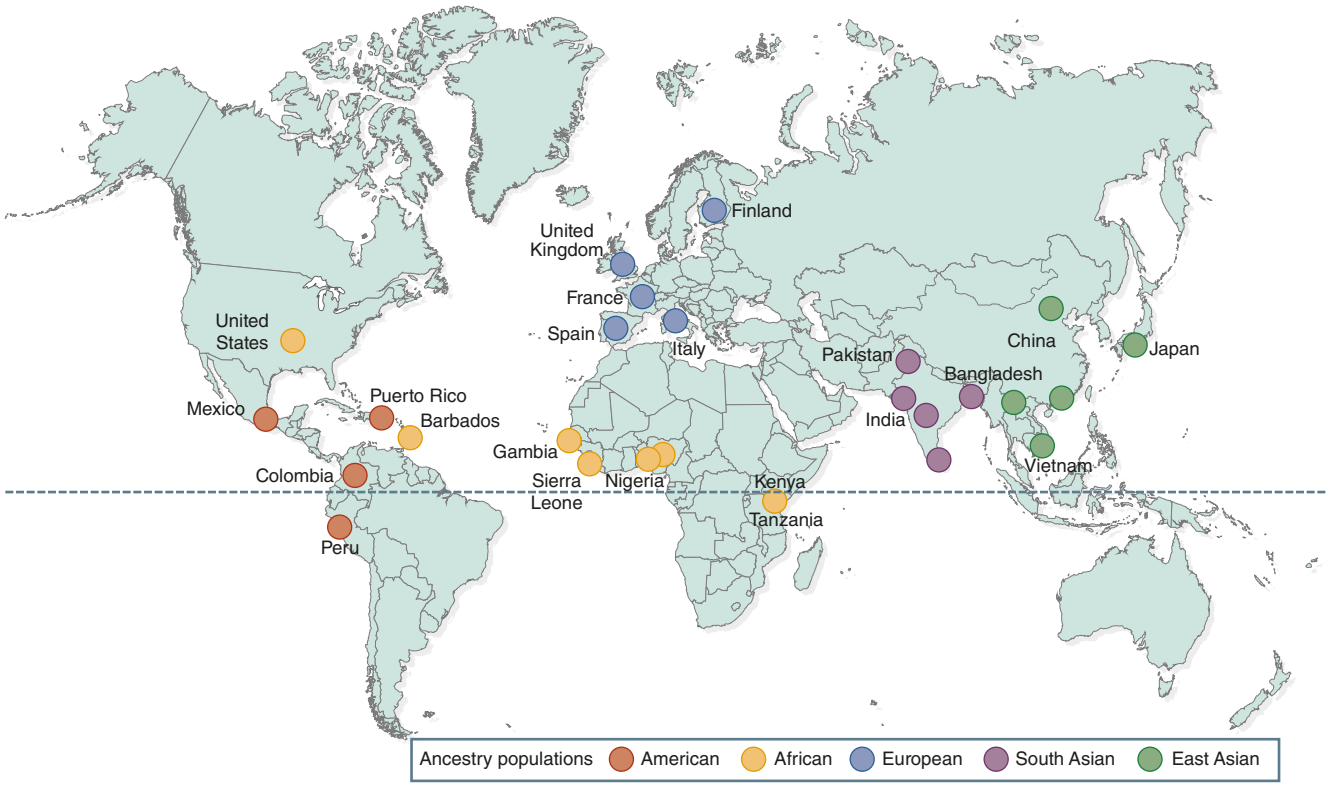


FIGURE 1.24 Geographic distribution of populations sampled in the 1000 Genomes Project. Each dot indicates a study population; colors indicate superpopulations as indicated in the legend.
Data from International Genome Sample Resource. (2016). IGSR and the 1000 Genomes Project. Retrieved from http://www.internationalgenome.org.

1.7 Genes and Environment

Inborn errors of metabolism illustrate the general principle that genes code for proteins and mutant genes code for mutant proteins. In diseases such as PKU, mutant proteins cause such a drastic change in metabolism that a severe genetic defect results. Nevertheless, biology is not necessarily destiny—organisms are also affected by the environment. PKU serves as an example of this principle, because patients who adhere to a diet restricted in the amount of phenylalanine develop intellectual capacities within the normal range. What is true in this example is true in general: Most traits are determined by the interaction of genes and environment.

It is also true that most traits are affected by multiple genes. No one knows how many genes are involved in the development and maturation of the brain and nervous system, but the number must be in the thousands. This number is in addition to the genes that are required in all cells to carry out metabolism and other basic life functions. It is easy to lose sight of the multiplicity of genes when considering extreme examples, such as PKU, in which a single mutation can have such a drastic effect on intellectual development. The situation is the same as that with any complex machine. An airplane can function properly provided that thousands of parts work together in harmony, but just one defective part, if that part affects a vital system, can bring it down. Likewise, the development and functioning of every trait require a large number of genes working in harmony, but in some cases a single mutant gene can have catastrophic consequences.

In short, the relationship between a gene and a trait is not necessarily a simple one. The biochemistry of organisms is a complex branching network in which different enzymes may share substrates, yield the same products, or be responsive to the same regulatory elements. Most visible traits of organisms are the net result of many genes acting together and in combination with environmental factors. PKU affords examples of each of three principles governing these interactions:

1. *One gene can affect more than one trait.* Children with extreme forms of PKU often have blond hair and reduced body pigment. These manifestations occur because the absence of PAH is a metabolic block that prevents conversion of phenylalanine into tyrosine, which is the precursor of the pigment melanin. The relationship between severe intellectual disability and decreased pigmentation in PKU makes sense only if one knows the metabolic connections among phenylalanine, tyrosine, and melanin. If these connections were not known, the traits would seem completely unrelated.

FIGURE 1.25 Among cats with white fur and blue eyes, approximately 40 percent are born deaf. The reason is that pigment cells derived from the neural crest migrate to various tissues including hair follicles, the eyes, and the middle ear, where their function is essential for hearing. Defective pigment cells resulting in white fur and blue eyes can, therefore, lead to deafness, which may be regarded as a pleiotropic effect of white fur and blue eyes.

PKU is not unusual in this regard. Many mutant genes affect multiple traits through their secondary or indirect effects. The various, sometimes seemingly unrelated, effects of a mutant gene are called **pleiotropic effects**, and the phenomenon itself is known as **pleiotropy**. **FIGURE 1.25** shows a cat with white fur and blue eyes, a pattern of pigmentation that is often (about 40 percent of the time) associated with deafness. Hence, deafness can be regarded as a pleiotropic effect of white coat and blue eye color. The developmental basis of this pleiotropy is unknown.

2. *Any trait can be affected by more than one gene.* We discussed this principle earlier in connection with the large number of genes that are required for the normal development and functioning of the brain and nervous system. Among these are genes that affect the function of the blood–brain barrier, which consists of specialized glial cells wrapped tightly around capillary walls in the brain, forming an impediment to the passage of most water-soluble molecules from the blood to the brain. The blood–brain barrier therefore affects the extent to which excess free phenylalanine in the blood can enter the brain itself. Because the effectiveness of the blood–brain barrier differs among individuals, patients with PKU who have very similar levels of blood phenylalanine can have dramatically different levels of cognitive development. This factor also explains in part why adherence to a controlled-phenylalanine diet is critically important in children but less so in adults; the blood–brain barrier is less well developed in children and, therefore, less effective in blocking the excess phenylalanine.

   Multiple genes affect even simpler metabolic traits. Phenylalanine breakdown and excretion serve as a convenient example. The metabolic pathway illustrated in Figure 1.10 indicates four enzymes play roles in this process, but even more enzymes are involved at the stage labeled "further breakdown." Because differences in the activity of any of these enzymes can affect the rate at which phenylalanine is broken down and excreted, all of the enzymes in the pathway are important in determining the amount of excess phenylalanine in the blood of patients with PKU.

3. *Most traits are affected by environmental factors as well as by genes.* Here we come back to the low-phenylalanine diet. Children with PKU are not doomed to severe intellectual deficiency; their capabilities can be brought into the normal range by dietary treatment. PKU serves as an example of what motivates geneticists to try to discover the molecular basis of inherited disease. The hope is that knowing the metabolic basis of a disease will eventually lead to methods for clinical intervention through diet, medication, or other treatments that will ameliorate the severity of the disease.

SUMMING UP

- Most traits are affected by multiple genes.
- One gene can affect multiple traits.
- Most traits are affected by a combination of genes and the environment.

1.8 The Molecular Unity of Life

The pathway for the breakdown and excretion of phenylalanine is by no means unique to human beings. One of the remarkable generalizations to have emerged from molecular genetics is that organisms that are very distinct—for example, plants and animals—share many features in their genetics and biochemistry. These similarities indicate a fundamental "unity of life":

> All creatures on Earth share many features of the genetic apparatus, including genetic information encoded in the sequence of bases in DNA, transcription into RNA, and translation into protein on ribosomes with the use of transfer RNAs. All creatures also share certain characteristics in their biochemistry, including many enzymes and other proteins that are similar in amino acid sequence, three-dimensional structure, and function.

Prokarya, Archaea, and Eukarya

Organisms share a common set of similar genes and proteins because they evolved by descent from a common ancestor. The process of **evolution** takes place when a population of organisms gradually changes in its genetic composition through time. Evolutionary changes in genes and proteins result in differences in metabolism, development, and behavior among organisms, which allows them to become progressively better adapted to their environments. From an evolutionary perspective, the unity of fundamental molecular processes in organisms alive today reflects inheritance from a distant common ancestor in which the molecular mechanisms were already in place.

Not just the unity of life but many other features of living organisms become comprehensible from an evolutionary perspective. For example, the interposition of an RNA intermediate in the basic flow of genetic information from DNA to RNA to protein makes sense if the earliest forms of life used RNA for both genetic information and enzyme catalysis. The importance of the evolutionary perspective in understanding aspects of biology that seem pointless or needlessly complex is summed up in a famous aphorism of the evolutionary biologist Theodosius Dobzhansky: "Nothing in biology makes sense except in the light of evolution."

Biologists distinguish three major domains of organisms:

1. **Prokarya**. This group includes most bacteria and cyanobacteria (formerly called blue-green algae). Cells of these organisms lack a membrane-bounded nucleus and mitochondria, are surrounded by a cell wall, and divide by binary fission.
2. **Archaea**. This group was initially discovered among microorganisms that produce methane gas or that live in extreme environments, such as hot springs or high salt concentrations. They are widely distributed in more normal environments as well. Superficially resembling bacteria, the cells of archaeans show important differences in the manner in which their membrane lipids are chemically linked. The machinery for DNA replication and transcription in archaeans resembles that of eukarya, whereas their metabolism strongly resembles that of bacteria. DNA sequence analysis indicates that about half of the genes found in the kingdom Archaea are unique to this group.
3. **Eukarya**. This group includes all organisms whose cells contain an elaborate network of internal membranes, a membrane-bounded nucleus, and mitochondria. Their DNA is present in the form of linear molecules organized into true chromosomes, and cell division takes place by means of mitosis (discussed in *The Chromosomal Basis of Inheritance* chapter). The eukaryotes include plants and animals as well as fungi and many single-celled organisms, such as amoebae and ciliated protozoa.

Genomes and Proteomes

The totality of DNA in a cell, nucleus, or organelle is called its **genome**. When used with reference to a species of organism (for example, in phrases such as "the human genome"), the term *genome* is defined as the DNA present in a normal reproductive cell.

Modern methods for sequencing DNA (discussed in the *DNA Replication and Sequencing* chapter) are so rapid and efficient that the complete DNA sequence is now known for more than 1000 different species of organisms. These include the genomes of multiple representatives of all of the groups of organisms, including Neandertals (an extinct species closely related to *Homo sapiens*, sequenced from DNA extracted from fossil bones).

TABLE 1.2 shows a small sample of sequenced genomes. Genome size is given in megabases (Mb), or millions of base pairs. The genome of the bacterium *Haemophilus influenzae*, like that of most bacteria, is very compact in that most of it codes for proteins. A high density of genes, relative to the amount of DNA, is also found in the budding yeast, the nematode worm, the fruit fly, and the diminutive flowering plant *Arabidopsis thaliana*. The human genome, by contrast, contains large amounts of noncoding DNA. Comparison with the nematode is illuminating. Whereas the human genome is about 30 times larger than that of the worm, the number of genes is less than 2 times larger. This discrepancy reflects the fact that only about 1.5 percent of the human genome sequence codes for protein. (Approximately 27 percent of the human genome is present in genes, but much of the DNA sequence present in genes does not code for proteins.)

The complete set of proteins encoded in the genome is known as the **proteome**. In less complex genomes, such as the yeast, worm, and fruit fly, the number of proteins in an organism's proteome is approximately the same as the number of genes. However, as we shall see the chapter titled *The Molecular Biology of Gene Expression*, some genes encode two or more proteins through a process called *alternative splicing* in which segments of the original RNA transcript are joined together in a variety of combinations to produce different messenger RNAs. Alternative splicing is especially prevalent in the human genome. At least one third of human genes, and possibly as many as two thirds, undergo alternative splicing; among the genes that undergo alternative splicing, the number of different messenger RNAs per gene

TABLE 1.2 Comparison of genes and proteins

Organism	Genome size, Mb[a] (approximate)	Number of genes (approximate)	Number of distinct proteins in proteome[b] (approximate)	Shared protein families
Haemophilus influenzae (causes bacterial meningitis)	1.9	1700	1400	
Saccharomyces cerevisiae (budding yeast)	13	6000	4400	
Caenorhabditis elegans (soil nematode)	100	20,000	9500	} 3000
Drosophila melanogaster (fruit fly)	120[c]	16,000	8000	} 5000
Mus musculus (laboratory mouse)	2500	25,000	10,000	} 7000[d]
Homo sapiens (human being)	2900[e]	25,000	10,000	} 9900[f]

[a] Millions of base pairs.
[b] Excludes "families" of proteins with similar sequences (and hence related functions).
[c] Excludes 60 Mb of specialized DNA ("heterochromatin") that has a very low content of genes.
[d] Based on similarity with sequences in messenger RNA (mRNA).
[e] For convenience, this estimate is rounded to 3000 Mb elsewhere in this text.
[f] Based on the observation that only about 1 percent of mouse genes lack a similar gene in the human genome, and *vice versa*.

ranges from 2 to 7. Hence, with its seemingly limited repertoire of 25,000 genes, the human genome can create approximately 60,000 to 90,000 different mRNAs. The widespread use of alternative splicing to multiply the coding capacity of genes is one source of human genetic complexity.

Most eukaryotic organisms contain *families* of related proteins that can be grouped according to similarities in their amino acid sequence. These families exist because the evolution of a new gene function is typically preceded by the duplication of an existing gene, followed by changes in nucleotide sequence in one of the copies that gives rise to the new function. The new function is usually similar to the previous one (for example, a change in the substrate specificity of a transporter protein), so that the new protein retains enough similarity in amino acid sequence to the original that their common ancestry can be recognized.

The molecular unity of life can be seen in the similarity of proteins in the proteome among diverse types of organisms. Such comparisons are shown in the right-hand column in Table 1.2. In this tabulation, each family of related proteins is counted only once, to estimate the number of proteins in the proteome that are "distinct" in the sense that their sequences are dissimilar. In yeast, worms, and flies, the number of distinct proteins is approximately 4400, 9500, and 8000, respectively. The brackets in Table 1.2 indicate the number of distinct proteins that share sequence similarity between species. From these comparisons, it appears that most multicellular animals share 5000 to 10,000 proteins that are similar in sequence and function. Approximately 3000 of these are shared with eukaryotes as distantly related as yeast, and approximately 1000 with prokaryotes as distantly related as bacteria. What these comparisons among proteomes imply is that biological systems are based on protein components numbering in the thousands. This is a challenging level of complexity to understand, but the challenge is much less intimidating than it appeared to be at an earlier time when human cells were thought to produce as many as 1 million different proteins.

SUMMING UP

- The processes of DNA replication, transcription, and translation, as well as the genetic code, are nearly universal properties of life.
- Cellular organisms are divided into three domains—prokarya, eukarya, and archaea.
- The genomes of organism vary greatly in total size, but less so in terms of the number of proteins encoded by them.
- The proteome of an organism consists of families of related proteins, many of which are shared among organisms in different domains of life.

CHAPTER SUMMARY

- Inherited traits are affected by genes.
- Genes are composed of the chemical deoxyribonucleic acid (DNA).
- DNA replicates to form copies of itself that are identical (except for rare mutations).
- DNA contains a genetic code specifying which types of enzymes and other proteins are made in cells.
- DNA occasionally mutates, and the mutant forms specify altered proteins that have reduced activity or stability.
- A mutant enzyme is an "inborn error of metabolism" that blocks one step in a biochemical pathway for the metabolism of small molecules.
- Traits are affected by environment as well as by genes.
- The molecular unity of life is seen in comparisons of genomes and proteomes.

REVIEW THE BASICS

- What were the key experiments showing that DNA is the genetic material?
- How did understanding the molecular structure of DNA give clues to its ability to replicate, to code for proteins, and to undergo mutation?
- Why is the pairing of complementary bases a key feature of DNA replication?
- What is the process of transcription, and in which ways does it differ from DNA replication?
- What is the difference between transcription and translation?
- Which three types of RNA participate in protein synthesis, and what is the role of each type of RNA?
- What is the "genetic code," and how is it relevant to the translation of a polypeptide chain from a molecule of messenger RNA?
- What is an inborn error of metabolism? How did this concept serve as a bridge between genetics and biochemistry?
- How does the "central dogma" explain Garrod's discovery that nonfunctional enzymes result from mutant genes?
- Explain why many mutant forms of phenylalanine hydroxylase have a simple amino acid replacement, yet the mutant polypeptide chains are absent or present in very small amounts.
- What is a pleiotropic effect of a gene mutation? Give an example.
- What are some of the major differences in cellular organization among prokarya, archaea, and eukarya?

GUIDE TO PROBLEM SOLVING

PROBLEM 1 In the human gene for the protein huntingtin (so named because it is associated with Huntington disease), the first 21 nucleotides in the amino acid–coding region are:

3′-TACCCACCGTTATAAGAGAGT-5′

What is the sequence of a partner strand?

ANSWER The base pairing between the strands is A with T and C with G, but it is equally important that the strands in a DNA duplex have opposite polarity. The complementary strand is therefore oriented with its 5′-end at the left. The base sequence of the partner strand is:

5′-ATGGGTGGCAATATTCTCTCA-3′

PROBLEM 2 If the DNA duplex for huntingtin in Problem 1 were transcribed from left to right, deduce the base sequence of the RNA in this coding region.

ANSWER To deduce the RNA sequence, we must apply three concepts. First, the base pairing is such so that A, T, C, and G in the DNA template strand are transcribed as U, A, G, and C, respectively, in the RNA. Second, the DNA template strand and RNA transcript have the opposite polarity. Third (and critically for this problem), the RNA strand is always transcribed in the 5′-to-3′ direction. Because we are told that transcription takes place from left to right, we can deduce that the transcribed strand is that given in Problem 1. The RNA transcript therefore has the base sequence

5′-AUGGGUGGCAAUAUUCUCUCA-3′

PROBLEM 3 Given the RNA sequence coding for part of human huntingtin deduced in Problem 2, what is the amino acid sequence in this part of huntingtin?

ANSWER The polypeptide chain is translated in successive groups of three nucleotides (codons), starting at 5′ end of

the coding sequence and moving in the 5′-to-3′ direction. The amino acid corresponding to each codon can be found in the genetic code table. The first seven amino acids in the polypeptide chain are:

```
5′-AUG GGU GGC AAU AUU CUC UCA-3′
   Met Gly Gly Asn Ile Leu Ser
```

PROBLEM 4 Suppose that a mutation in the human huntingtin gene occurs in the DNA sequence shown in Problem 1. In this mutation, the red G is replaced with a T. What is the nucleotide sequence of this region of the DNA duplex (both strands), and that of the messenger RNA, and what is the amino acid sequence of the mutant huntingtin?

ANSWER The DNA, RNA, and polypeptide chain have the sequences as follows, where the differences from the nonmutant gene are in red. The mutation results in stop codon and premature termination of huntingtin synthesis.

```
DNA (transcribed strand)
        3′-TAC CCA CCG TTA TAA GAG ATT-5′
DNA (nontranscribed strand)
        5′-ATG GGT GGC AAT ATT CTC TAA-3′
RNA coding region
        5′-AUG GGU GGC AAU AUU CUC UAA-3′
Polypeptide chain
        Met Gly Gly Asn Ile Leu STOP
```

ANALYSIS AND APPLICATIONS

1.1 Classify each of the following statements as true or false.

- **(a)** Each gene is responsible for only one visible trait.
- **(b)** Every trait is potentially affected by many genes.
- **(c)** The sequence of nucleotides in a gene specifies the sequence of amino acids in a protein encoded by the gene.
- **(d)** There is one-to-one correspondence between the set of codons in the genetic code and the set of amino acids encoded.

1.2 From their examination of the structure of DNA, what were Watson and Crick able to infer about the probable mechanisms of DNA replication, coding capability, and mutation?

1.3 What does it mean to say that each strand of a duplex DNA molecule has a polarity? What does it mean to say that the paired strands in a duplex molecule have opposite polarity?

1.4 What important observation about the S and R strains of *Streptococcus pneumoniae* prompted Avery, MacLeod, and McCarty to study this organism?

1.5 In the transformation experiments of Avery, MacLeod, and McCarty, what was the strongest evidence that the substance responsible for the transformation was DNA rather than protein?

1.6 A chemical called phenol (carbolic acid) destroys proteins but not nucleic acids, and strong alkalis such as sodium hydroxide destroy both proteins and nucleic acids. In the transformation experiments with *Streptococcus pneumoniae*, which result would be expected if the S-strain extract had been treated with phenol? If it had been treated with a strong alkali?

1.7 Which feature of the physical organization of bacteriophage T2 made it suitable for use in the Hershey Chase experiments?

1.8 Like DNA, molecules of RNA contain large amounts of phosphorus. When Hershey and Chase grew their T2 phage in bacterial cells in the presence of radioactive phosphorus, the RNA must also have incorporated the labeled phosphorus, yet the experimental result was not compromised. Why not?

1.9 The DNA extracted from a bacteriophage contains 16 percent A, 16 percent T, 34 percent G, and 34 percent C. What can you conclude about the structure of this DNA molecule?

1.10 The DNA extracted from a bacteriophage consists of 20 percent A, 34 percent T, 35 percent G, and 11 percent C. What is unusual about this DNA? What can you conclude about its structure?

1.11 A region along a DNA strand that is transcribed contains no A. Which base will be missing in the corresponding region of the RNA?

1.12 If one strand of a DNA duplex has the sequence 5′-ATCAG-3′, what is the sequence of the complementary strand? (Write the answer with the 5′ end at the left.)

1.13 Suppose that a double-stranded DNA molecule is separated into its constituent strands, and the strands are purified using a high-speed centrifuge. In one of the strands, the base composition is

25 percent A, 18 percent T, 20 percent G, and 37 percent C. What is the base composition of the other strand?

1.14 Consider a region along one strand of a double-stranded DNA molecule that consists of tandem repeats of the trinucleotide 5′-GTA-3′, so that the sequence in this strand is 5′-GTAGTAGTAGT…-3′. What is the sequence in the other strand? (Write the answer with the 5′ end at the left.)

1.15 If the *template* strand of a DNA duplex has the sequence 5′-TCAG-3′, what is the sequence of the RNA transcript? (Write the answer with the 5′ end at the left.)

1.16 If the *nontemplate* strand of a DNA duplex has the sequence 5′-ATCAG-3′, what is the sequence of the RNA transcript across this region? (Write the answer with the 5′ end at the left).

1.17 A duplex DNA molecule contains a random sequence of the four nucleotides with equal proportions of each. What is the average spacing between consecutive occurrences of the sequence 5′-ATGC-3′? Between consecutive occurrences of the sequence 5′-TACGGC-3′?

1.18 An RNA molecule folds back upon itself to form a "hairpin" structure held together by a region of base pairing. One segment of the molecule in the paired region has the base sequence 5′-UAUCGUAU-3′. What is the base sequence with which this segment is paired?

1.19 A synthetic mRNA molecule consists of the repeating base sequence

5′-AAAAAAAA…-3′

When this molecule is translated in vitro using ribosomes, transfer RNAs, and other necessary constituents from *E. coli*, the result is a polypeptide chain consisting of the repeating amino acid Lys–Lys–Lys–. . . . If you assume that the genetic code is a triplet code, what does this result imply about the codon for lysine (Lys)?

1.20 A synthetic mRNA molecule consisting of the repeating base sequence

5′-UUUUUUUUUUUU…-3′

is terminated by the addition, to the right-hand end, of a single nucleotide bearing A. When translated in vitro, the resulting polypeptide consists of a repeating sequence of phenylalanines terminated by a single leucine. What does this result imply about the codon for leucine?

1.21 With in vitro translation of an RNA into a polypeptide chain, the translation can begin anywhere along the RNA molecule. A synthetic RNA molecule has the sequence

5′-CGCUUACCACAUGUCGCGAAC-3′

How many reading frames are possible if this molecule is translated in vitro? How many reading frames are possible if this molecule is translated in vivo, in which translation starts with the codon AUG?

1.22 You have sequenced both strands of a double-stranded DNA molecule. To evaluate the potential of this molecule for coding amino acids, you conceptually transcribe it into RNA and then conceptually translate the RNA into a polypeptide chain. How many reading frames would you have to examine?

1.23 A synthetic mRNA molecule consists of the repeating base sequence

5′-ACACACACACACAC…-3′

When this molecule is translated in vitro, the result is a polypeptide chain consisting of the alternating amino acids Thr–His–Thr–His–Thr–His–. . . . Why do the amino acids alternate? What does this result imply about the codons for threonine (Thr) and histidine (His)?

1.24 A synthetic mRNA molecule consists of the repeating base sequence

5′-AUCAUCAUCAUCAUC…-3′

When this molecule is translated in vitro, the result is a mixture of three different polypeptide chains. One consists of repeating isoleucines (Ile – Ile – Ile – Ile –. . .), another of repeating serines (Ser – Ser – Ser – Ser –. . .), and the third of repeating histidines (His – His – His –His –. . .). What does this result imply about the manner in which an mRNA is translated?

1.25 How is it possible for a gene with a mutation in the coding region to encode a polypeptide with the same amino acid sequence as the nonmutant gene?

1.26 A polymer has a random sequence consisting of 75 percent G and 25 percent U. Among the amino acids in the polypeptide chains resulting from in vitro translation, what is the expected frequency of Trp? Of Val? Of Phe?

1.27 Results of complementation tests of the six independent recessive mutations *a–f* are shown in the accompanying matrix, where + indicates complementation and − indicates lack of complementation. Classify the mutations into complementation groups, and write the names of the mutations in

each complementation group in the blanks provided below. (Some of the blanks may remain empty.)

	b	*c*	*d*	*e*	*f*
a	+	+	−	−	+
b		+	+	+	+
c			+	+	+
d				−	+
e					+

Mutations in complementation group 1: ___ ___ ___ ___ ___

Mutations in complementation group 2: ___ ___ ___ ___ ___

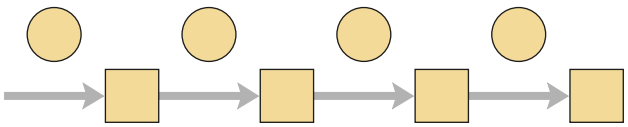
Mutations in complementation group 3: ___ ___ ___ ___ ___

Mutations in complementation group 4: ___ ___ ___ ___ ___

1.28 A mutant screen is carried out in *Neurospora crassa* for mutants that are unable to synthesize an amino acid, that we will symbolize as G. A number of mutants are isolated and classified into four groups according to their ability to grow (+) or not grow (−) in minimal medium supplemented with possible precursors D, E, and F. The data are shown in the accompanying table.

	D	E	F	G
Class 1	+	+	+	+
Class 2	−	−	−	+
Class 3	−	−	+	+
Class 4	+	−	+	+

Complete the diagram shown below. Each arrow indicates one or more biochemical reactions. Within each circle, write the class of mutants (1–4) whose products contribute to the reactions symbolized by the arrow; in the squares, write the name of the amino acid or precursor (D–G) at that position in the pathway.



1.29 The coding sequence in the messenger RNA for amino acids 1 through 10 of human phenylalanine hydroxylase is:

5′-AUGUCCACUGCGGUCCUGGAAAACCCAGGC-3′

(a) What are the first 10 amino acids?
(b) Which sequence would result from a mutant RNA in which the red A was changed to G?
(c) Which sequence would result from a mutant RNA in which the red C was changed to G?
(d) Which sequence would result from a mutant RNA in which the red U was changed to C?
(e) Which sequence would result from a mutant RNA in which the red G was changed to U?

1.30 A "frameshift" mutation is a mutation in which some number of base pairs, other than a multiple of three, is inserted into or deleted from a coding region of DNA. The result is that, at the point of the frameshift mutation, the reading frame of protein translation is shifted with respect to the nonmutant gene. To see the consequence of a frameshift mutation, consider that the coding sequence in the messenger RNA for the first 10 amino acids in human beta hemoglobin (part of the oxygen-carrying protein in the blood) is

5′-AUG GUG CAC CUG ACU CCU GAG GAG AAG UCU…-3′

(a) What is the amino acid sequence in this part of the polypeptide chain?
(b) What would be the consequence of a frameshift mutation resulting in an RNA missing the red U?
(c) What would be the consequence of a frameshift mutation resulting in an RNA with a G inserted immediately in front of the red U?

1.31 A news headline reads "Gene for Schizophrenia Identified." Does this necessarily mean that schizophrenia in an individual can be diagnosed by genetic analysis alone? Explain your answer.

1.32 Using the data in Table 1.2, construct a graph, plotting genome size for the different organisms on the *x*-axis and number of genes on the *y*-axis. Now, consider the case of the Mexican axolotl (*Ambystoma mexicana*), with an estimated genome size of 32 billion base pairs. Based on your graph, can you predict the number of genes that are present in this genome? Explain your answer.

CHALLENGE PROBLEMS

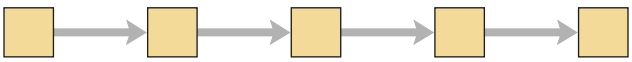
CHALLENGE PROBLEM 1 With regard to the wild-type and mutant RNA molecules described in Problem 1.30, deduce the base sequence in both strands of the corresponding double-stranded DNA for:

(a) The wild-type sequence
(b) The single-base deletion
(c) The single-base insertion

CHALLENGE PROBLEM 2 You are carrying out Beadle Tatum type experiments to analyze a metabolic pathway in *Neurospora*. You know that the precursor in the pathway is a molecule symbolized as P and that the product is a vitamin symbolized as Z. You are sure that the pathway from P to Z is linear and that the molecules W, X, and Y are intermediates. However, there may be other intermediates not yet identified. You obtain 10 independent mutations that cannot grow on minimal medium supplemented with P, but can grow on minimal medium supplemented with Z. The 10 mutants fall into four classes that can grow (+) or cannot grow (−) on minimal medium supplemented with the nutrients W, X, or Y. The data are shown in the accompanying table.

	P	W	X	Y	Z
Class I (mutant 5)	−	−	−	−	+
Class II (mutants 1, 3, 4, 6, 7)	−	+	+	+	+
Class III (mutant 2)	−	+	−	+	+
Class IV (mutants 8, 9, 10)	−	−	−	+	+

Draw a linear metabolic pathway with P on the left and Z on the right, in which each of the intermediates W, X, and Y is shown in the order in which it occurs in the metabolic pathway in the synthesis of Z from P.



CHALLENGE PROBLEM 3 The 10 mutants in Challenge Problem 2 were tested for complementation in all pairwise combinations using heterokaryons. The results are shown in the matrix, in which + indicates the ability of the heterokaryon to grow in minimal medium and − indicates inability to grow in minimal medium.

	2	3	4	5	6	7	8	9	10
Mutant 1	+	+	−	+	+	+	+	+	+
2		+	+	+	+	+	+	+	+
3			+	+	+	−	+	+	+
4				+	+	+	+	+	+
5					+	+	+	+	+
6						+	+	+	+
7							+	+	+
8								+	+
9									−

Assume that each complementation group defines a different gene, and assume further that each gene encodes an enzyme that catalyzes a single step in the metabolic pathway, which converts one molecule of substrate into one molecule of product.

(a) Redraw the metabolic pathway deduced from Challenge Problem 2. Use a right arrow to indicate each enzymatic step in the pathway, and label each arrow with the mutant number 1–10 that blocks the enzymatic step. In some cases, you will not be able to specify the order in which the enzymes occur in the pathway, so you may write them in any order you wish.
(b) In the metabolic pathway that you have deduced from the data, how many *unknown* intermediates are there between the precursor P and the vitamin Z?

FOR FURTHER READING

The 1000 Genomes Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://doi.org/10.1038/nature15393

An overview of the 1000 Genomes Project.

Avery, O. T. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus type III. *Journal of Experimental Medicine*, *79*(2), 137–158. http://doi.org/10.1084/jem.79.2.137

Considered by most to be the definitive evidence that DNA is the genetic material.

Beadle, G. W., & Tatum, E. L. (1941). Genetic control of biochemical reactions in *Neurospora*. *Proceedings of the National Academy of Sciences of the United States of America*, *27*(11), 499–506. http://www.ncbi.nlm.nih.gov/pubmed/16588492

The origin of the one gene–one enzyme hypothesis.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. http://doi.org/10.1038/227561a0

Crick's clearest statement of the central dogma, in terms of information flow from DNA to protein but not the reverse.

Griffith, F. (1928). The significance of pneumococcal types. *Journal of Hygiene*, *27*(2), 113. http://doi.org/10.1017/S0022172400031879

The first demonstration that there is a "transforming principle."

Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *Journal of General Physiology*, *36*(1), 39–56. http://doi.org/10.1085/jgp.36.1.39

Further confirmation that DNA, as opposed to protein, is the genetic material.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738. http://www.ncbi.nlm.nih.gov/pubmed/13054692

The original publication describing the double-helical structure of DNA.

Watson, J. D., & Crick, F. H. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*. http://doi.org/10.1038/171964b0

A less well-known but nevertheless very important paper in which Watson and Crick first suggest a basis for the genetic code.