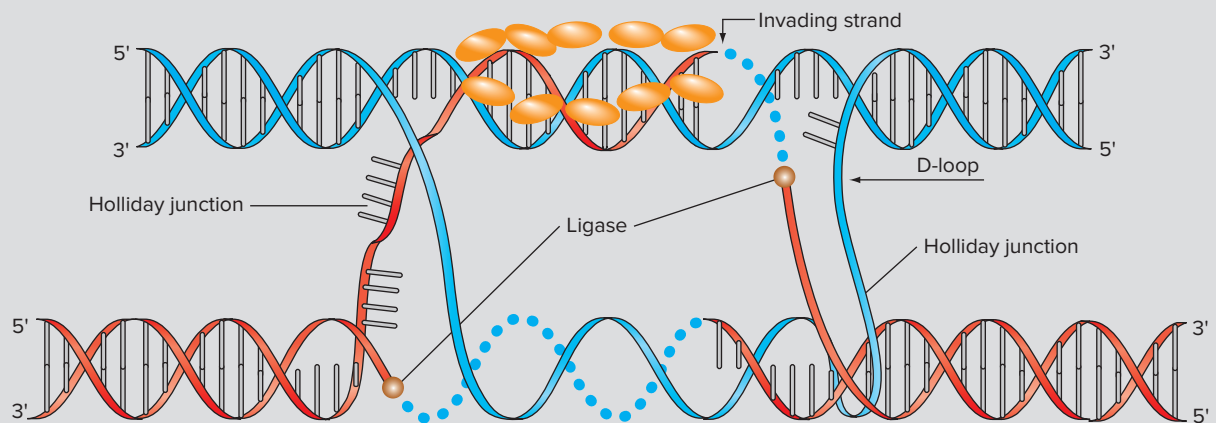
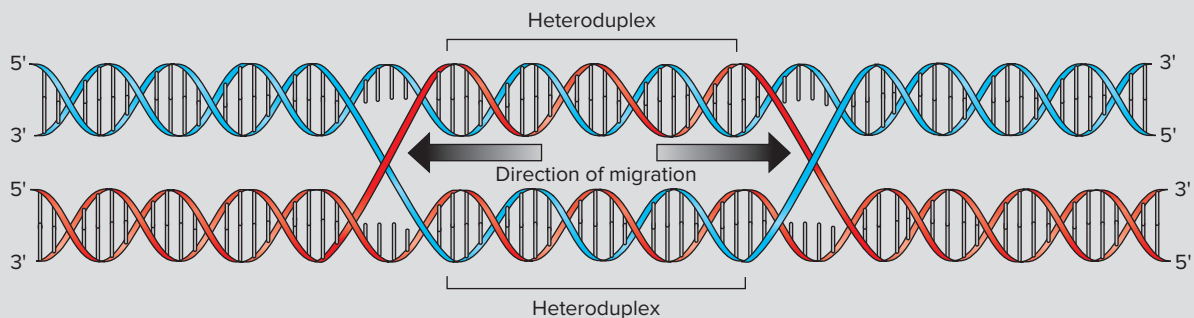


### Crossover Pathway

**Step 4 Formation of a double Holliday junction.** New DNA added to the invading 3' tail (*blue dots* at the top) enlarges the D-loop until the single-stranded bases on the displaced strand can form complementary base pairs with the 3' tail on the non-sister chromatid. New DNA added to this latter tail (*blue dots* at the bottom) re-creates the DNA duplex on the bottom chromatid. At each side of the original break, the 3' end of the newly synthesized DNA becomes adjacent to a 5' end left after resection, and **DNA ligase** forms phosphodiester bonds to rejoin DNA strands without the loss or gain of nucleotides. The resulting X-shaped structures are called **Holliday junctions** after Robin Holliday, the scientist who first proposed their existence as a key intermediate in recombination.



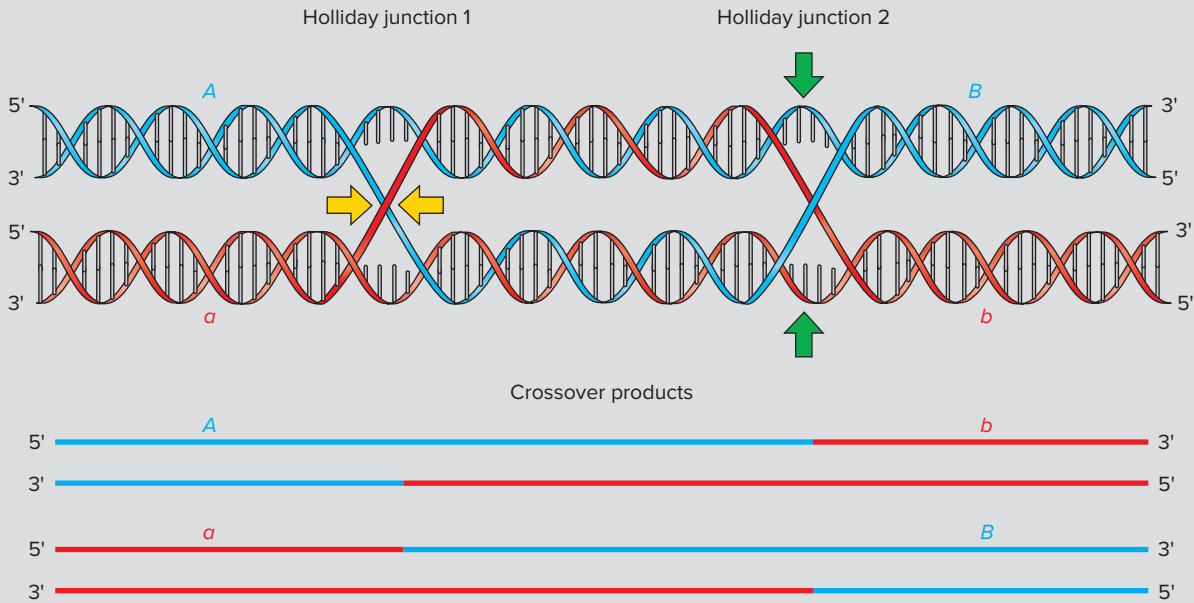
**Step 5 Branch migration.** The two invading strands tend to *zip up* by base pairing with the complementary strands of the parental double helices they invade. The DNA double helices unwind in front of this double zipping action, moving in the direction of the arrows in the figure, and two newly created heteroduplex molecules rewind behind it. Branch migration thus lengthens the heteroduplex region of both DNA molecules from tens of base pairs to hundreds or thousands.



(continued)

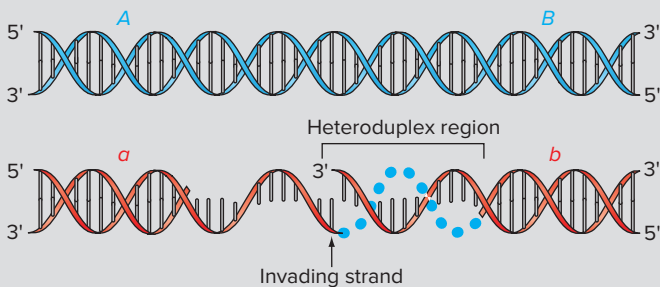
**FEATURE FIGURE 6.27 (Continued)**

**Step 6 Resolution of the double Holliday junction intermediate.** The two interlocked nonsister chromatids must disengage. Separation is achieved by breakage of two DNA strands at each Holliday junction by an enzyme called **resolvase** (*not shown*); the strands are subsequently rejoined by DNA ligase (*not shown*). Different *blue* and *red* strands are cleaved at each junction; one junction is cleaved at the strands indicated by *yellow* arrows, and the strands indicated by the *green* arrows are cleaved at the other junction. At each junction, the strands are cut and rejoined so that *red* DNA connects to *blue* DNA and *vice versa*. Crossing-over results because each of the four strands is cut once and rejoined. Note in the diagram at the *bottom* that both of the recombinant chromatids have short heteroduplex regions.

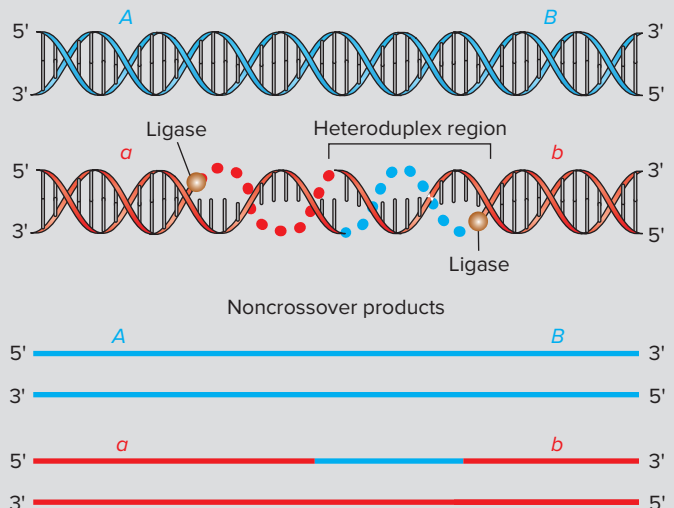


**Noncrossover Pathway**

**Step 4' Strand displacement and annealing.** Just as in Step 4 of the Crossover Pathway, the invading strand (*arrow*) is first extended by DNA synthesis (*blue dots*) using the nonsister chromatid (*blue*) as a template. But next, an **anticrossover helicase** enzyme (*not shown*) disentangles the invading strand and the nonsister chromatid to yield the intermediates diagrammed.



**Step 5' DNA synthesis and ligation.** The remaining gap in the double-stranded DNA sequence is filled by DNA synthesis (*red dots*), and DNA ligase forms phosphodiester bonds to rejoin the DNA strands. The result is no crossover, but a heteroduplex region nonetheless remains in one chromatid.



The molecular intermediate formed at the conclusion of Step 3 on Fig. 6.27 may have two alternative fates. One pathway, depicted in Steps 4 through 6, results in crossing-over. The second pathway, shown in Steps 4' and 5', does not yield a crossover, but one of the resultant chromatids has a heteroduplex region.

### The crossover pathway

The strand displaced by strand invasion in Step 3 now forms a second heteroduplex with the other 3' single-stranded tail (Fig. 6.27, Step 4). DNA synthesis to extend the two 3' tails replaces the DNA that was degraded by the exonuclease, and **DNA ligase** reseals the DNA backbones (Fig. 6.27, Step 4). The result is that the two nonsister chromatids are interlocked at two **Holliday junctions** (Fig. 6.27, Step 5). The Holliday junctions move away from each other and thereby enlarge the heteroduplex between them—a process called **branch migration** (Fig. 6.27, Step 5).

Now, the two nonsister chromatids must be separated. The two chromatids disengage by the cutting and joining of two strands of DNA at each Holliday junction. As shown in Fig. 6.27 (Step 6), crossing-over (and recombination of flanking alleles) results when a different pair of DNA strands is cut and rejoined by **resolvase** and ligase enzymes at each junction. Because resolvase almost always cuts all four DNA strands, resolution of the double Holliday junctions usually results in crossing-over.

### The noncrossover pathway

Recombination initiated by Spo11 can also result in no crossing-over through the action of an enzyme called **anticrossover helicase**. The helicase helps disentangle the invading strand from the nonsister chromatid, thus interrupting Holliday junction formation (Fig. 6.27, Step 4'). Note that although the end result of this pathway is no crossing-over (Step 5'), one of the resultant chromatids nonetheless contains a heteroduplex region.

### Controlling where and when recombination occurs

Only cells undergoing meiosis express the Spo11 protein, which is responsible for a rate of meiotic recombination several orders of magnitude higher than that found in mitotically dividing cells. In yeast and humans, where meiotic double-strand breaks have been mapped, it is clear that Spo11 has a preference for cleavage of some genomic sequences over others, resulting in *hotspots* for crossing-over (recall Fig. 5.17).

Unlike meiotic cells, mitotic cells do not usually initiate recombination as part of the normal cell-cycle program; instead, recombination in mitotic cells is a consequence of environmental damage to the DNA. As you will see in Chapter 7, X-rays and ultraviolet light, for example, can cause either double-strand breaks or single-strand nicks.

The cell's enzymatic machinery works to repair the damaged DNA site, and recombination is a side effect of this process.

### A summary: Evidence for the current molecular model of homologous recombination

The double-strand-break repair model of meiotic recombination was proposed in 1983, well before the direct observation of any recombination intermediates. Scientists have now seen—at the molecular level—the formation of double-strand breaks, the resection of those breaks to produce 3' single-strand tails, and double Holliday junction structures. The double-strand-break repair model has become established because it explains much of the data obtained from genetic and molecular studies as well as the six properties of recombination deduced from genetic experiments:

1. Homologs physically break, exchange parts, and rejoin. The Meselson-Weigle experiment with phage lambda provided key evidence for this key aspect of recombination (review Fig. 6.26).
2. Crossing-over occurs between nonsister chromatids after DNA replication. When yeast dihybrid for linked genes sporulate, the appearance of T tetrads and the rarity of NPDs make sense only if recombination happens at the four-strand, as opposed to the two-strand, stage (review Fig. 6.25).
3. Breakage and repair generate reciprocal products of recombination. Yeast and *Neurospora* tetrads are almost always NPD, PD, or T because the reciprocal recombinants are found in the same ascus.
4. Recombination events can occur anywhere along the DNA molecule. If enough progeny are counted, crossing-over can be observed between any pair of genes in a variety of different experimental organisms.
5. Precision in the exchange—no gain or loss of nucleotide pairs—prevents mutations from occurring during the process. Geneticists originally deduced the precision of crossing-over from observing that recombination usually does not cause mutations; today, we know this to be true from DNA sequence analysis.
6. **Gene conversion**—the physical change of one allele in a heterozygote into the other—sometimes occurs as a result of a recombination event. In the next section, you will see how gene conversion is explained by the formation of heteroduplexes during recombination events.

### DNA Repair of Heteroduplexes Can Result in Gene Conversion

Yeast tetrad analysis allows us to see, just as Mendel predicted, that alleles segregate equally into gametes. Diploids heterozygous at a particular locus produce two

spores containing one allele, and two spores containing the other allele (2:2 segregation)—most of the time. The opportunity to examine all four products of a single meiosis together in an ascus allowed the discovery that rarely, tetrads exhibit 3:1 or 1:3 segregation patterns, thereby breaking Mendel's first law. These rare tetrads are a consequence of heteroduplex formation during recombination; the phenomenon that produces these tetrads is called *gene conversion*.

### Mismatched bases in heteroduplexes

The molecular model for recombination includes formation of a heteroduplex region, which occurs because the two strands of a recombinant DNA molecule do not break and rejoin at the same location on the double helix. In addition, through branch migration, the heteroduplex region can be expanded to hundreds or even thousands of base pairs. The name *heteroduplex* applies not only because the two DNA strands came from different nonsister chromatids, but also because the base pairing of the strands may produce mismatches in which one or a few bases are not complementary. If the heteroduplex region is within a gene and the maternal and paternal alleles of the gene are different, gene conversion may result.

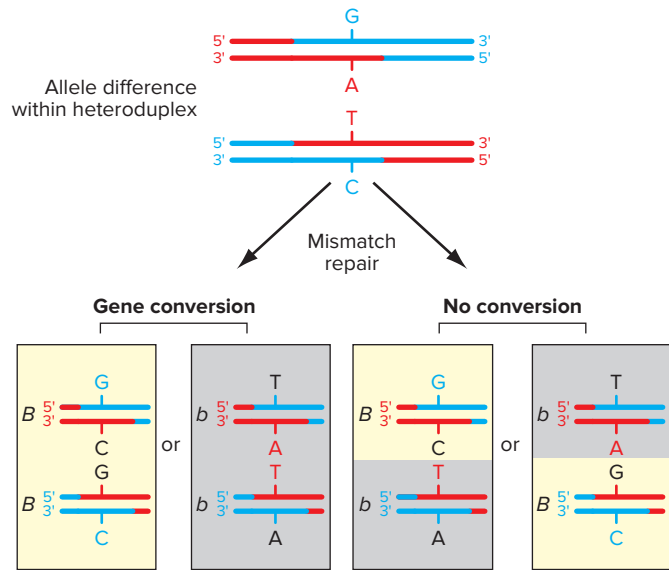
### Gene conversion through mismatch repair

Mismatched heteroduplex molecules do not persist for long. The same DNA repair enzymes that operate to correct mismatches during replication (to be discussed in Chapter 7) also correct heteroduplexes during recombination. The outcome of the repair enzymes' work depends on which strands they alter. For example, the G–A mismatch in Fig. 6.28 can become either G–C or T–A, and the T–C mismatch may be repaired to either G–C or T–A (*italics* indicate the altered base). Therefore, four possible repair outcomes exist for the two mismatches generated at a heteroduplex. Two of those outcomes—those in which both heteroduplexes are repaired to generate the same base pair—may result in gene conversion. Suppose that, as in Fig. 6.28, the base pair difference within a heteroduplex is the molecular difference between two alleles *B* and *b*. One nonsister chromatid started out with *B* and the other with *b*. The result of gene conversion is that both chromatids end up with the same allele—both are either *B* or *b*.

### Gene conversion in yeast and *Neurospora* asci

Gene conversion is noticeable in yeast and *Neurospora* because all of the products of a single meiosis stay together in an ascus. Gene conversion may be detected as an unusual ascus that is neither PD, NPD, nor T. Recall from Fig. 5.22

**Figure 6.28** How gene conversion occurs. Alleles *B* and *b* differ by a single base pair; where *B* is G–C (yellow), *b* is T–A (gray). If gene *B* is within the heteroduplex region after a recombination event, repair of mismatched bases may convert *B* to *b* or vice versa. Gene conversion results when the bases changed by DNA repair (black) both originated from the same chromatid. Note that the blue and red lines are single DNA strands.

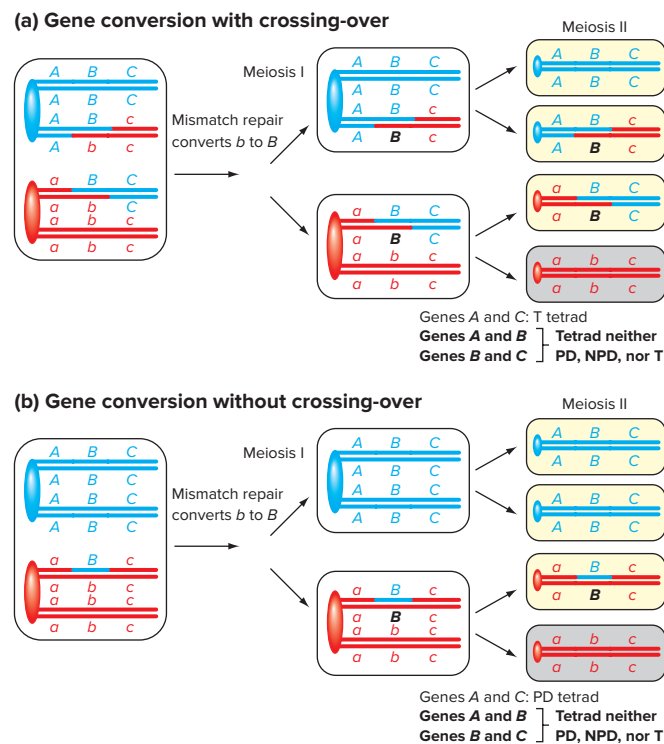


that an *A B / a b* diploid yeast cell, where *A* and *B* are linked genes, could produce tetrads of any one of the three types. A key feature common to all three of these tetrad types is that the ratio of *A:a* or *B:b* alleles is always 2:2. However, a rare conversion of *b* to *B* results in a tetrad that is neither PD, NPD, nor T, because the ratio of *B:b* alleles is 3:1 (Fig. 6.29).

The idea that gene conversion is due to heteroduplex formation during a recombination event is supported by the observation that gene conversion is often associated with crossing-over of flanking alleles. For example, suppose during meiosis in an *A B C / a b c* diploid yeast, a recombination event occurs between gene *A* and gene *C* such that gene *B* is within the heteroduplex region (Fig. 6.29). Resolution of Holliday junctions on either side of gene *B* results in crossing-over—recombination between alleles of the flanking genes *A* and *C*. Subsequent DNA repair of the heteroduplex regions containing gene *B* can result in gene conversion, producing a tetrad that displays 3:1 segregation of *B:b* or *b:B* (Fig. 6.29a).

You should note that heteroduplexes resulting from recombination events that enter the noncrossover pathway can also generate tetrads with 3:1 segregation patterns. In such cases, gene conversion occurs but is not accompanied by recombination of the alleles of the flanking genes (Fig. 6.29b).

**Figure 6.29 Detection of gene conversion in yeast tetrads.** Throughout this figure, the *blue* and *red* lines represent single DNA strands. **(a)** Recombination during meiosis in an  $A B C / a b c$  diploid yeast cell generates heteroduplex regions in which each DNA strand has different alleles of gene  $B$ . Conversion of  $b$  to  $B$  by mismatch repair (*black Bs*) results in an unusual tetrad with a 3:1 ratio of  $B:b$  alleles instead of 2:2. In this case, the recombination event resulted in crossing-over and thus recombination of the alleles of the flanking genes  $A$  and  $C$ . The tetrad is T with respect to  $A$  and  $C$ . **(b)** Here, the recombination event is resolved by the noncrossover pathway. Because crossing-over does not occur, the resulting tetrad is PD with respect to genes  $A$  and  $C$ . However, mismatch repair of the heteroduplex region converts  $b$  into  $B$ , so this tetrad also shows a 3:1 ratio of  $B:b$ .



### essential concepts

- In tetrad analysis, the existence of Ts and the very low number of NPDs observed establishes that recombination occurs after chromosome replication, when each pair of homologs contains four chromatids. T and NPD tetrads exhibit equal numbers of both classes of recombinants, indicating reciprocal exchange.
- The exchange of chromosome parts during recombination involves the breakage and rejoining of DNA molecules.
- At the molecular level, crossing-over during meiotic prophase entails the formation of *heteroduplex* DNAs between two *Holliday junctions* and resolution of the junctions by endonucleases and DNA ligase.

- Recombination events result in crossing-over only part of the time because helicases can disentangle the chromatids before Holliday junctions form.
- *Gene conversion*, a process whereby one allele in a heterozygote is physically changed into the other, provides evidence for heteroduplex formation during recombination events.

## 6.6 Site-Specific Recombination

### learning objectives

1. Diagram the possible outcomes of site-specific recombination.
2. List the components that would have to be introduced to import site-specific recombination into a newly discovered organism.
3. Contrast the functions of Spo11 and Cas9, two enzymes that catalyze the formation of double-strand breaks.

Homologous recombination, as discussed in the previous section, begins with preexisting DNA molecules, breaks them apart, and then rejoins them to create new sequences of DNA. Natural selection then tests these new DNA molecules for their ability to help the organisms in which they are found to survive and reproduce in a changing environment. The more types of DNA molecules that are created in a population of organisms, the greater is the possibility that the population will continue in future generations. It is thus not surprising that homologous recombination can occur nearly at random at any of a very large number of sites in a genome, likely between any two adjacent pairs of nucleotides. In this way, homologous recombination helps to produce an enormous diversity in chromosome base sequences upon which natural selection can act.

### Recombinase Enzymes Catalyze Recombination Between Specific DNA Sequences

In contrast with this type of nearly random homologous recombination, some organisms find it useful to have systems of **site-specific recombination** that promote the breakage and rejoining of DNA molecules at particular

DNA sequences. Site-specific recombination is crossing-over that occurs only between two specific DNA target sites that are usually less than 200 base pairs long. Site-specific recombination is much simpler at the molecular level than is the homologous recombination discussed in the previous section. In particular, in most systems of site-specific recombination, a single protein logically called a **recombinase** is sufficient to catalyze all the breakage and joining steps of the process. If you are curious, **Fig. 6.30** depicts the mode of action of one class of such recombinases.

The organisms that take advantage of site-specific recombination include certain kinds of bacteriophages that use this process for the **integration** (incorporation) of their small, circular genome into the chromosome of the host bacterium (**Fig. 6.31a**). In this way, the bacteriophage DNA “hitchhikes” along with the bacterial chromosome: When the host DNA replicates, so does the integrated bacteriophage genome.

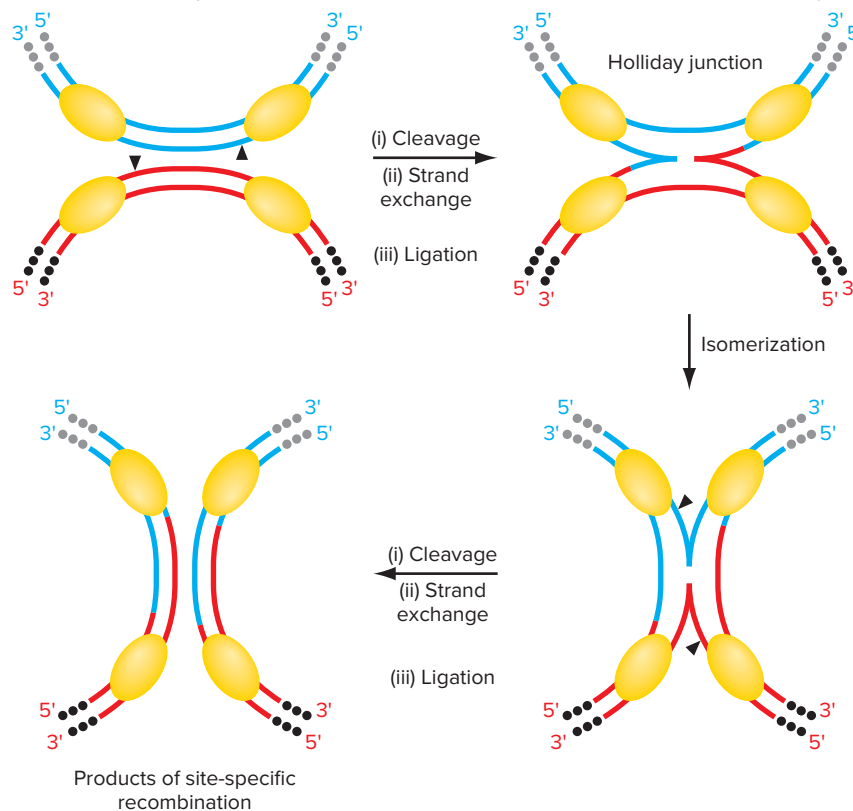
Site-specific recombination is also important for the reverse process of **excision**, in which the DNA integrated

between two target sites in a single chromosome is removed to create two independent DNA molecules (**Fig. 6.31b**). If a bacteriophage genome was previously integrated into the host chromosome, excision is crucial to allow the bacteriophage genome to extricate itself and then to become incorporated in the virus particle.

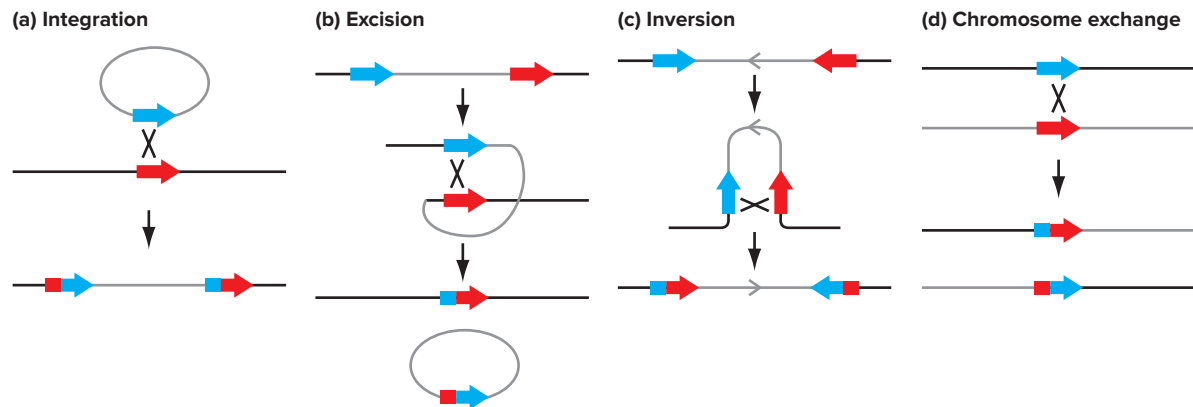
A third potential outcome of site-specific recombination systems is the **inversion** of a segment of DNA that is located between the two target sites (**Fig. 6.31c**). As you can imagine, such inversion could constitute a molecular switch between two configurations of the same chromosome. The in-between segment is oriented in one direction in one state and in the other direction in the other state.

A final mode of site-specific recombination can occur if the target site is found at the same position on each of two homologous chromosomes. Action of the recombinase on these target sites will result in the reshuffling of regions on nonsister chromatids, an outcome that leads to recombinant chromosomes (**Fig. 6.31d**). To our knowledge, this situation is not normally encountered in organisms that naturally use site-specific recombination. However, geneticists

**Figure 6.30 One site-specific recombination mechanism.** The Cre and Flp enzymes discussed in the text function as shown. The red and blue target DNA sequences are identical to each other but are represented in different colors for clarity. These targets are embedded in different DNA molecules (black and gray dots). The subunits of the recombinase tetramer are yellow ovals; this enzyme catalyzes all steps of the reaction. Black triangles are sites where recombinase cleaves single-stranded DNA. Note that resolution of the Holliday junction intermediate involves cleavage of the blue and red DNA strands that were not cleaved initially.



**Figure 6.31 Possible outcomes of site-specific recombination.** The blue and red arrows represent different identical target sites; the arrows can point in either of two directions because the target sites are asymmetric. The single black and gray lines in which the target sites are embedded are double-stranded DNA.



can create organisms with this arrangement of target sites that also make the recombinase protein. This technique is particularly useful in causing mitotic crossing-over to occur with high frequency at these defined locations.

### Scientists can exploit the Flp/FRT and Cre/loxP site-specific recombination systems to turn genes on and off

Site-specific recombination is a property of only certain organisms, and its use in those organisms is usually restricted to a very specific process such as bacteriophage integration or excision. If site-specific recombination is not a general phenomenon like homologous recombination, why are we telling you about it? The answer is that geneticists can now export site-specific recombination to a wide variety of species, and these researchers have found such recombination to be incredibly useful. By adding target sequences to genomes, the geneticists can control precisely where in a genome recombination will take place. And by regulating the production of the recombinase enzyme, researchers can determine at what time and in what tissues the site-specific recombination occurs.

The later chapters of this book discuss two such systems of site-specific recombination: Flp recombinase/FRT sites (Flp/FRT), normally used for the replication of small circular DNAs (plasmids) in yeast cells; and Cre recombinase/loxP sites (Cre/loxP), needed for several stages in the life cycle of a type of bacteriophage called P1.

These feats of genetic engineering have several purposes. Using site-specific recombination, researchers can turn on or off the expression of a specific gene within an organism at a specific time or in a specific tissue. In addition, because site-specific recombination can occur with high efficiency in nearly all cell types, geneticists can use

this method to induce mitotic recombination and thus reliably create clones of homozygous mutant cells within a heterozygous organism. By performing these manipulations, scientists can now ask important questions about the roles of particular genes in biological processes such as the development of a multicellular organism from a single cell, the fertilized egg; Chapters 18 and 19 will describe these issues in detail.

### CRISPR-Cas9-induced recombination is a powerful tool for manipulating genomes

One important limitation in importing site-specific recombination to new organisms is that the target sequences need to be introduced into genomes, but in most cases researchers cannot direct those target sites into a preselected genomic region. Instead, the target sites become incorporated into random positions, and the scientists then search for a strain with the target site in the most advantageous location.

Remarkable methodologies developed very recently now allow researchers to alter genomes precisely in almost any way imaginable. One particularly exciting technology is based on small RNAs called *CRISPRs* and an enzyme called Cas9 that is produced in a few bacterial species. It is premature to describe this method in great detail so early in this book, but for the time being it is sufficient to tell you that a CRISPR can direct Cas9 to any specific DNA sequence in a complex genome. The importance is that Cas9 is an enzyme that produces double-strand breaks in DNA. As we saw in a previous section, the formation of a double-strand break (by Spo11) initiates the process of homologous recombination; in other words, double-strand breaks are *recombinogenic*.

Because CRISPR/Cas9 causes double-strand breaks at a specific genomic location determined by the sequence of the CRISPR RNA, researchers can now induce recombination to occur at high frequency at any specific location of the genome. As will be seen in Chapter 18, this recombination allows scientists to alter the sequence of DNA near the breakpoint in any desired way. The potential significance of this newfound ability to alter genomes is staggering. Just to cite one example, such pinpoint genome editing may allow for *gene therapy* in which mutant alleles in the genomes of the somatic cells of a person suffering from a genetic disease such as cystic fibrosis could be changed to wild-type alleles.

### essential concepts

- *Site-specific recombination* is crossing-over between two short DNA target sites catalyzed by a *recombinase* enzyme.
- Researchers can import target sites and the corresponding recombinase gene into an organism's genome to promote site-specific recombination at a particular genomic location, at a particular time, and in a particular tissue.
- The *CRISPR/Cas9* system can induce double-strand breaks at almost any position in the genome. The fact that these double-strand breaks are recombinogenic allows scientists to edit genomes in the vicinity of the breakage.

### WHAT'S NEXT

The Watson-Crick model for the structure of DNA, the single most important biological discovery of the twentieth century, clarified how the genetic material fulfills its primary functions of carrying and accurately reproducing information: Each DNA molecule carries one of a vast number of potential arrangements of the four nucleotide building blocks (A, T, G, and C). The model also suggested how base complementarity could provide a mechanism for faithful DNA replication. We have further seen how the structure of DNA enables the recombining of genetic information from maternal and paternal chromosomes.

Unlike its ability to carry information, DNA's capacities for replication and recombination are not solely properties of the DNA molecule itself. Rather they depend on the cell's complex enzymatic machinery. But even though they rely on the complicated orchestration of many

different proteins, replication and recombination both occur with extremely high fidelity—normally not a single base pair is gained or lost. Occasionally, however, errors do occur, providing the genetic basis of evolution.

DNA copying or recombination errors that occur within genes sometimes produce dramatic changes in phenotype. How do mutations in genes arise? And how did we come to understand that different alleles of genes produce their phenotypic effects through the proteins that they specify?

We begin to answer these questions in Chapter 7. We first describe the molecular processes that lead to mutation. Next, you will see that scientists used mutations to determine what a gene actually is—a linear sequence of base pairs in DNA, and what a gene does—it encodes the information for producing a protein.

### SOLVED PROBLEMS

5' TAAGCGTAACCCGCTAA  
3' ATTTCGATTGGGCGATT

CGTATGCGAAC  
GCATACGCTTG

GGTCCTATTAACGTGCGTACAC 3'  
CCCAGGATAATTGCACGCATGTG 5'

- I. Imagine that the double-stranded DNA molecule shown was broken at the sites indicated by spaces in the sequence, and that before the breaks were repaired, the 11 base pair DNA fragment between the breaks was reversed (*inverted*). What would be the base sequence of the repaired molecule? Explain your reasoning.

### Answer

To answer this question, you need to keep in mind the polarity of the DNA strands involved.

The top strand has the polarity left to right of 5' to 3'. The reversed region must be rejoined with the same polarity. Label the polarity of the strands within the inverting region. **To have a 5'-to-3' polarity maintained on the top strand, the**



fragment that is reversed must also be flipped over, so the strand that was formerly on the bottom is now on top.

```

5' TAAGCGTAACCCGCTAAGTTCGCATACGGGGTCTATTAACGTGCGTACAC 3'
ATTCGCATTGGGCGATTCAAGCGTATGCCCCAGGATAATTGCACGCATGTG
3'                                     5'

```

- II. A new virus has recently been discovered that infects human lymphocytes. The virus can be grown in the laboratory using cultured lymphocytes as host cells. Design an experiment using a radioactive label that would tell you if the virus contains DNA or RNA.

### Answer

Use your knowledge of the differences between DNA and RNA to answer this question. RNA contains uracil instead of the thymine found in DNA. You could set up one culture in which you add radioactive uracil to the medium and a second one in which you add radioactive thymine to the culture. After the viruses have infected cells and produced more new viruses, collect the newly synthesized virus. Determine which culture produced radioactive viruses. If the virus contains RNA, the collected virus grown in medium containing radioactive uracil will be radioactive, but the virus grown in radioactive thymine will not be radioactive. If the virus contains DNA, the collected virus from the culture containing radioactive thymine will be radioactive, but the virus from the radioactive uracil culture will not.

(You might also consider using radioactively labeled ribose or deoxyribose to differentiate between an RNA- and DNA-containing virus. Technically this does not work as well because the radioactive sugars are processed by cells before they become incorporated into nucleic acid, thereby obscuring the results.)

- III. If you expose a culture of human cells (for example, HeLa cells) to  $^3\text{H}$ -thymidine during S phase, how would the radioactivity be distributed over a pair of homologous chromosomes at metaphase? Would the radioactivity be in (a) one chromatid of one homolog, (b) both chromatids of one homolog, (c) one chromatid each of both homologs, (d) both chromatids of both homologs, or (e) some other pattern? Choose the correct answer and explain your reasoning.

### Answer

This problem requires application of your knowledge of the molecular structure and replication of DNA and how it relates to chromatids and homologs. DNA replication occurs during S phase, so the  $^3\text{H}$ -thymidine would be incorporated into the new DNA strands. A chromatid is a replicated DNA molecule, and each new DNA molecule contains one new strand of DNA (semiconservative replication). The radioactivity would be in both chromatids of both homologs (answer d).

## PROBLEMS

### Vocabulary

1. For each of the terms in the left column, choose the best matching phrase in the right column.

- |                        |   |                                 |  |
|------------------------|---|---------------------------------|--|
| a. transformation      | 1. the strand that is synthesized discontinuously during replication        | h. Okazaki fragments            | 8. two nitrogenous bases that can pair via hydrogen bonds                                |
| b. bacteriophage       | 2. the sugar within the nucleotide subunits of DNA                          | i. purine                       | 9. catalyzes site-specific recombination   |
| c. pyrimidine          | 3. a nitrogenous base containing a double ring                              | j. topoisomerases               | 10. a nitrogenous base containing a single ring  |
| d. deoxyribose         | 4. noncovalent bonds that hold the two strands of the double helix together | k. semiconservative replication | 11. a short sequence of bases where unwinding of the double helix for replication begins |
| e. hydrogen bonds      | 5. Meselson and Stahl experiment  | l. lagging strand               | 12. a virus that infects bacteria  |
| f. complementary bases | 6. Griffith experiment  | m. telomeres                    | 13. short DNA fragments formed by discontinuous replication of one of the strands        |
| g. origin              | 7. structures at ends of eukaryotic chromosomes                             | n. recombinase                  | 14. enzymes involved in controlling DNA supercoiling                                     |

## Section 6.1

- Griffith, in his 1928 experiments, demonstrated that bacterial strains could be genetically transformed. The evidence that DNA was the *transforming principle* responsible for this phenomenon came later. What was the key experiment that Avery, MacCleod, and McCarty performed to prove that DNA was responsible for the genetic change from rough cells into smooth cells?
- During bacterial transformation, DNA that enters a cell is not an intact chromosome; instead it consists of randomly generated fragments of chromosomal DNA. In a transformation where the donor DNA was from a bacterial strain that was  $a^+ b^+ c^+$  and the recipient was  $a b c$ , 55% of the cells that became  $a^+$  were also transformed to  $c^+$ . But only 2% of the  $a^+$  cells were  $b^+$ . Is gene  $b$  or  $c$  closer to gene  $a$ ?
- Nitrogen and carbon are more abundant in proteins than sulfur. Why did Hershey and Chase use radioactive sulfur instead of nitrogen and carbon to label the protein portion of their bacteriophages in their experiments to determine whether parental protein or parental DNA is necessary for progeny phage production?

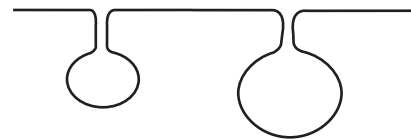
## Section 6.2

- If 30% of the bases in human DNA are A, (a) what percentage are C? (b) What percentage are T? (c) What percentage are G?
- Which of the following statements are true about double-stranded DNA?
  - $A + C = T + G$
  - $A + G = C + T$
  - $A + T = G + C$
  - $A/G = C/T$
  - $A/G = T/C$
  - $(C + A) / (G + T) = 1$
- Imagine you have three test tubes containing identical solutions of purified, double-stranded human DNA. You expose the DNA in tube 1 to an agent that breaks the sugar-phosphate (phosphodiester) bonds. You expose the DNA in tube 2 to an agent that breaks the bonds that attach the bases to the sugars. You expose the DNA in tube 3 to an agent that breaks the hydrogen bonds. After treatment, how would the structures of the molecules in the three tubes differ?
- What information about the structure of DNA was obtained from X-ray crystallographic data?
- A portion of one DNA strand of the human gene responsible for cystic fibrosis is

5'.....ATAGCAGAGCACCATCTG.....3'

Write the sequence of the corresponding region of the other DNA strand of this gene, noting the polarity. What do the dots before and after the given sequence represent?

- When a double-stranded DNA molecule is exposed to high temperature, the two strands separate, and the molecule loses its helical form. We say the DNA has been *denatured*. (Denaturation also occurs when DNA is exposed to acid or alkaline solutions.)
  - Regions of the DNA that contain many A–T base pairs are the first to become denatured as the temperature of a DNA solution is raised. Thinking about the chemical structure of the DNA molecule, why do you think the A–T-rich regions denature first?
  - If the temperature is lowered, the original DNA strands can *reanneal*, or *renature*. In addition to the full double-stranded molecules, some molecules of the type shown here are seen when the molecules are examined under the electron microscope. How can you explain these structures?



- A particular virus with DNA as its genetic material has the following proportions of nucleotides: 20% A, 35% T, 25% G, and 20% C. How can you explain this result?

## Section 6.3

- The underlying structure of DNA is very simple, consisting of only four possible building blocks.
  - How is it possible for DNA to carry complex genetic information if its structure is so simple?
  - What are these building blocks? Can each block be subdivided into smaller units, and if so, what are they? What kinds of chemical bonds link the building blocks?
  - How does the underlying structure of RNA differ from that of DNA?
- An RNA virus that infects plant cells is copied into a DNA molecule after it enters the plant cell. What would be the sequence of bases in the first strand of DNA made complementary to the section of viral RNA shown here?

5' CCCUUGGAACUACAAAGCCGAGAUUAA 3'

- Bacterial transformation and bacteriophage labeling experiments proved that DNA was the hereditary

material in bacteria and in DNA-containing viruses. Some viruses do not contain DNA but have RNA inside the phage particle. An example is the tobacco mosaic virus (TMV) that infects tobacco plants, causing lesions in the leaves.

Two different variants of TMV exist that have different forms of a particular protein in the virus particle that can be distinguished. It is possible to reconstitute TMV *in vitro* (in the test tube) by mixing purified proteins and RNA. The reconstituted virus can then be used to infect the host plant cells and produce a new generation of viruses. Design an experiment to show that RNA, rather than protein, acts as the hereditary material in TMV.

15. The CAP protein is shown bound to DNA in Fig. 6.15. CAP binds a specific sequence of base pairs in DNA (N = any base):

```
5' TGTGANNNNNNNTCAC A 3'
3' AACTNNNNNNNAGTGT 5'
```

- In a long double-stranded DNA molecule with random base sequence and an equal number of A–T and G–C base pairs, how many different kinds of DNA sequences could be bound by CAP?
- In the same DNA molecule, how frequently would a CAP binding site of any type be present? Of a particular type?
- CAP protein binds DNA as dimer; two identical CAP protein subunits bound to each other bind DNA. Can you detect a special feature of the DNA site that CAP binds that suggests that two identical protein subunits bind the DNA? (*Hint*: Try reading the sequence in the 5′-to-3′ direction on each strand.)
- CAP protein binds to the major groove of DNA. Do you expect that DNA helicase is required for CAP to bind DNA?

#### Section 6.4

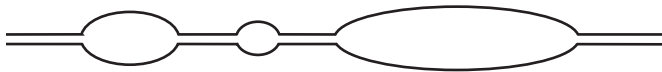
- In Meselson and Stahl's density shift experiments (diagrammed in Fig. 6.20), describe the results you would expect in each of the following situations:
  - Conservative replication after two rounds of DNA synthesis on  $^{14}\text{N}$ .
  - Semiconservative replication after three rounds of DNA synthesis on  $^{14}\text{N}$ .
  - Dispersive replication after three rounds of DNA synthesis on  $^{14}\text{N}$ .
  - Conservative replication after three rounds of DNA synthesis on  $^{14}\text{N}$ .
- When Meselson and Stahl grew *E. coli* in  $^{15}\text{N}$  medium for many generations and then transferred the cells to  $^{14}\text{N}$  medium for one generation, they found that the bacterial DNA banded at a density intermediate between that of pure  $^{15}\text{N}$  DNA and pure  $^{14}\text{N}$  DNA following equilibrium density centrifugation. When they allowed the bacteria to replicate one additional time in  $^{14}\text{N}$  medium, they observed that half of the DNA remained at the intermediate density, while the other half banded at the density of pure  $^{14}\text{N}$  DNA. What would they have seen after an additional generation of growth in  $^{14}\text{N}$  medium? After two additional generations?
- If you expose human tissue culture cells (for example, HeLa cells) to  $^3\text{H}$ -thymidine just as they enter S phase, then wash this material off the cells and let them go through a second S phase before looking at the chromosomes, how would you expect the  $^3\text{H}$  to be distributed over a pair of homologous chromosomes? (Ignore the effect recombination could have on this outcome.) Would the radioactivity be in (a) one chromatid of one homolog, (b) both chromatids of one homolog, (c) one chromatid each of both homologs, (d) both chromatids of both homologs, or (e) some other pattern? Choose the correct answer and explain your reasoning. (This problem extends the analysis begun in Solved Problem III.)
- Draw a replication bubble with both replication forks and label the origin of replication, the leading strands, lagging strands, and the 5′ and 3′ ends of all strands shown in your diagram.
- Do any strands of nucleic acid exist in nature in which part of the strand is DNA and part is RNA? If so, describe when such strands of nucleic acid are synthesized. Is the RNA component at the 5′ end or at the 3′ end?
  - RNA primers in Okazaki fragments are usually very short, less than 10 nucleotides and sometimes as short as 2 nucleotides in length. What does this fact tell you about the *processivity* of the primase enzyme—that is, the relative ability of the enzyme to continue polymerization as opposed to dissociating from the template and from the molecule being synthesized? Which enzyme is likely to have a greater processivity, primase or DNA polymerase III?
- As Fig. 6.21 shows, DNA polymerase cleaves the high-energy bonds between phosphate groups in nucleotide triphosphates (nucleotides in which three phosphate groups are attached to the 5′ carbon atom of the deoxyribose sugar). The enzyme uses this energy to catalyze the formation of a phosphodiester bond when incorporating new nucleotides into the growing chain.
  - How does this information explain why DNA chains grow during replication in the 5′-to-3′ direction?
  - The action of the enzyme DNA ligase in joining Okazaki fragments together is shown in Fig. 6.23.

Remember that these fragments are connected only after the RNA primers at their ends have been removed. Given this information, infer the type of chemical bond whose formation is catalyzed by DNA ligase and whether or not a source of energy will be required to promote this reaction. Explain why DNA ligase and not DNA polymerase is required to join Okazaki fragments.

22. The bases of one of the strands of DNA in a region where DNA replication begins are shown at the end of this problem. What is the sequence of the primer that is synthesized complementary to the bases in bold? (Indicate the 5' and 3' ends of the sequence.)

5' AGGCCTCG**AATTCGTATAG**CTTTCAGAAA 3'

23. Replicating structures in DNA can be observed in the electron microscope. Regions being replicated appear as bubbles.
- Assuming bidirectional replication, how many origins of replication are active in this DNA molecule?
  - How many replication forks are present?
  - Assuming that all replication forks move at the same speed, which origin of replication was activated last?



24. Indicate the role of each of the following in DNA replication: (a) topoisomerase, (b) helicase, (c) primase, and (d) ligase.
25. Draw a diagram of replication that is occurring at the end of a double-stranded linear chromosome. Show the leading and lagging strands with their primers. (Indicate the 5' and 3' ends of the strands.) What difficulty is encountered in producing copies of both DNA strands at the end of a chromosome?
26. Figure 6.18 depicts Watson and Crick's initial proposal for how the double-helical structure of DNA accounts for DNA replication. Based on our current knowledge, this figure contains a serious error due to oversimplification. Identify the problem with this figure.
27. Researchers have discovered that during replication of the circular DNA chromosome of the animal virus SV40, the two newly completed daughter double helices are intertwined. What would have to happen for the circles to come apart?
28. A *DNA synthesizer* is a machine that uses automated chemical synthesis to generate short, single strands

of DNA of any given sequence. You have used the machine to synthesize the following three DNA molecules:

(DNA 1) 5' CTACTACGGATCGGG 3'

(DNA 2) 5' CCAGTCCCAGATCCGT 3'

(DNA 3) 5' AGTAGCCAGTGGGGAAAAACCCCACTGG 3'

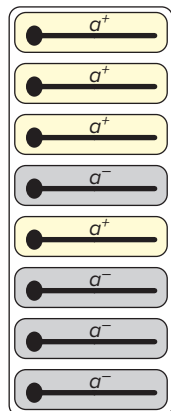
Now you add the DNA molecules either singly or in combination to reaction tubes containing DNA polymerase, dATP, dCTP, dGTP, and dTTP in a buffered solution that allows DNA polymerase to function. For each of the reaction tubes, indicate whether DNA polymerase will synthesize any new DNA molecules, and if so, write the sequence(s) of any such DNAs.

- DNA 1 plus DNA 3
- DNA 2 plus DNA 3
- DNA 1 plus DNA 2
- DNA 3 only

### Section 6.5

29. Bacterial cells were coinfecting with two types of bacteriophage lambda: One carried the  $c^+$  allele and the other the  $c$  allele. After the cells lysed, progeny bacteriophage were collected. When a single such progeny bacteriophage was used to infect a new bacterial cell, it was observed in rare cases that some of the resulting phage progeny were  $c^+$  and others were  $c$ . Explain this result.
30. A yeast strain with a mutant  $spo11^-$  allele has been isolated. The mutant allele is nonfunctional; it makes no Spo11 protein. What do you suppose is the phenotype of this mutant strain?
31. Imagine that you have done a cross between two strains of yeast, one of which has the genotype  $A B C$  and the other  $a b c$ , where the letters refer to three rather closely linked genes in the order given. You examine many tetrads resulting from this cross, and you find two that do not contain the expected two  $B$  and two  $b$  spores. In tetrad I, the spores are  $A B C$ ,  $A B c$ ,  $a b c$ , and  $a b C$ . In tetrad II, the spores are  $A B C$ ,  $A b c$ ,  $a b C$ , and  $a b c$ . How have these unusual tetrads arisen?
32. The *Neurospora* octad shown came from a cross between  $a^+$  and  $a^-$  strains.
- Is this an MI or an MII octad or neither? Explain.
  - Diagram the production of this octad.
  - Is it possible to observe evidence of heteroduplex formation in a *Neurospora* ascus even if gene

conversion did not occur during formation of the octad? Explain.



33. From a cross between  $e^+ f^+ g^+$  and  $e^- f^- g^-$  strains of *Neurospora*, recombination between these linked genes resulted in a few octads containing the following ordered set of spores:

$e^+ f^+ g^+$   
 $e^+ f^+ g^+$   
 $e^+ f^- g^+$   
 $e^+ f^- g^+$   
 $e^- f^- g^-$   
 $e^- f^- g^-$   
 $e^- f^- g^-$   
 $e^- f^- g^-$

- Where was recombination initiated?
  - Did crossing-over occur between genes  $e$  and  $g$ ? Explain.
  - Why do you end up with  $2 f^+ : 6 f^-$  but  $4 e^+ : 4 e^-$  and  $4 g^+ : 4 g^-$ ?
  - Could you characterize these unusual octads as MI or MII for any of the three genes involved? Explain.
34. In Step 6 of Fig. 6.27, the resolvase enzyme almost always cuts all four strands of DNA in the double Holliday junction intermediate: both *blue* strands and both *red* strands. Another way of stating this fact is that the enzyme cuts the DNA at Holliday junctions 1 and 2 in different ways, represented by the *yellow arrows* at junction 1 and the *green arrows* at junction 2 in the figure. But rarely, the resolvase enzyme instead cuts the DNA at both Holliday junctions in the same way (*yellow arrows* at both junctions or *green arrows* at both junctions). In other words, at both junctions, the same *red* strand and the same *blue* strand are cut. What would be the outcome of this rare resolvase enzyme behavior?

## Section 6.6

35. Figure 6.31 shows four potential outcomes of site-specific recombination that depend on the relative arrangement of the target sites for the recombinase enzyme. Could homologous recombination (in the absence of specific target sites and recombinase) also cause all of the same kinds of outcomes? If so, what is so different about how geneticists would use homologous recombination and site-specific recombination?
36. Each of the substrates for site-specific recombination listed (a–f) is also the product of site-specific recombination that occurred at a different one of these substrates. Match each substrate (a–f) with its product (a–f).
- a linear DNA with two target sites in the same orientation
  - a linear DNA with two target sites in opposite orientations
  - a circular DNA with two target sites in the same orientation
  - a circular DNA with two target sites in opposite orientations
  - a circular DNA with one target site and a linear DNA with one target site
  - two circular DNAs each with one target site
37. Problem 52 in Chapter 5 discussed the use of mitotic recombination to study the function of a gene called *smc* during mouse development. The idea was that mitotic recombination in a cell of an  $smc^+/smc^-$  heterozygote could produce a clone of tissue derived from a daughter cell that was homozygous for the  $smc^-$  mutation. You could recognize the cells in this clone by the absence of green fluorescence from GFP.
- Mitotic recombination is a rare event, making it difficult for researchers to find  $smc^+/smc^-$  clones to study. Explain why scientists might want to subject the same mice to X-rays to increase the frequency at which the desired clones could be found. At what stage of mouse development would the researchers expose the animals to X-rays?
  - An even more efficient way to induce mitotic recombination is to construct mice that include the Flp/FRT system. (Flp is a recombinase enzyme in yeast cells that promotes site-specific recombination between two identical copies of a particular 34 base pair long DNA sequence called an FRT site.) Assuming that these mice have a *transgene* that can specify the Flp recombinase protein, where would you put the FRT sites relative to the *smc* gene and to the *GFP* gene?
  - In part (b), what (in general terms) would you have to do with the Flp-encoding gene to make sure that

mitotic recombination does not happen in every cell in the mouse? (This precaution is necessary because a mouse which has many *smc<sup>-</sup>/smc<sup>-</sup>* homozygous cells could die as an embryo, before it reaches adulthood.)

38. Suppose that you could inject a wild-type mouse zygote with a specific CRISPR RNA and the Cas9 enzyme. The RNA directs the Cas9 enzyme to make a double-strand break within a gene that you think may be responsible for a heritable disease. Diagram in rough form how you might inject at the same time another nucleic acid molecule (here, a double-stranded DNA) to exploit homologous recombination so that you could convert the wild-type allele of the gene to a specific mutant allele.
39.  $\Phi$ C31 is a type of bacteriophage that infects *Streptomyces* bacteria. One gene in the bacteriophage genome specifies a recombinase called  $\Phi$ C31 integrase that works through a mechanism slightly different from that of the recombinase shown in Fig. 6.30. Most importantly, the two target DNA sequences are different from each other. One called *attP* is 39 base pairs and is found on the circular bacteriophage chromosome, while the other—*attB*—is 34 base pairs long and is located on the much larger circular bacterial chromosome. Excepting two base pairs roughly in the middle of both targets that are identical and at which recombination takes place, the DNA sequences of *attP* and *attB* are completely different from each other.
- Diagram the reaction that  $\Phi$ C31 integrase performs. How could this reaction be important for the life cycle of the bacteriophage?
  - Using the diagram you just drew, explain why  $\Phi$ C31 integrase cannot reverse the reaction.
  - Now consider how you might exploit this site-specific recombination to place genes from another species (a *transgene*) into the genome of an experimental organism like *Drosophila*. Assume you can make any DNA sequences you want and that you can introduce these DNA sequences into fruit fly germ-line cells by injection. Why is the irreversibility of the  $\Phi$ C31 integrase-mediated reaction valuable for placing the transgene into the *Drosophila* genome?
- d. Bacteriophage  $\Phi$ C31 must eventually reverse this reaction. Why? How do you think the bacteriophage can achieve this reversal?
40. Cre is a recombinase enzyme encoded by a gene in bacteriophage P1. The Cre enzyme promotes site-specific recombination between two copies of a 34 bp long DNA sequence called loxP that is derived from the same bacteriophage.
- Researchers use the Cre/loxP site-specific recombination system in order to make mice homozygous for a deletion of any particular gene and only in a specific tissue. You will see in Chapter 18 that scientists first make mice where a pair of loxP sites are configured in a particular way with respect to the gene to be deleted.
- Draw a diagram that shows the configuration of the loxP sites that would enable deletion of a gene by site-specific recombination.
  - What else would the researchers need to introduce into the mouse genome in order to generate mice where only one tissue is homozygous for the gene deletion?
  - Why do you think that scientists would want to generate mice like this?
  - Unlike the DNA rearrangements at *attP* and *attB* sites catalyzed by  $\Phi$ C31 integrase (described in Problem 39), the DNA rearrangements caused by Cre recombinase are reversible. Why?
  - Why does the reversibility of Cre/loxP-mediated recombination not interfere with the use of the Cre/loxP system to generate mouse tissues with deletions?
41. Like Cre/loxP recombination, site-specific recombination mediated by the Flp/FRT system is reversible. Why doesn't this fact interfere with the experiment described in Problem 37?

chapter **7**

# Anatomy and Function of a Gene: Dissection Through Mutation



*A scale played on a piano keyboard and a gene on a chromosome are both a series of simple, linear elements (keys or nucleotide pairs) that produce information. A wrong note or an altered nucleotide pair calls attention to the structure of the musical scale or the gene.*

© Ingram Publishing RF

## chapter outline

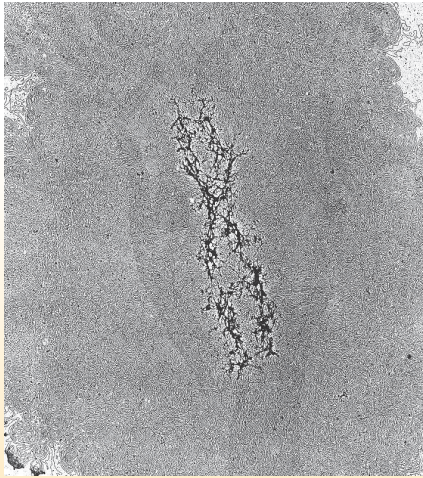
- 7.1 Mutations: Primary Tools of Genetic Analysis
- 7.2 Molecular Mechanisms That Alter DNA Sequence
- 7.3 DNA Repair Mechanisms
- 7.4 What Mutations Tell Us About Gene Structure
- 7.5 What Mutations Tell Us About Gene Function
- 7.6 A Comprehensive Example: Mutations That Affect Vision

**HUMAN CHROMOSOME 3 CONSISTS** of approximately 220 million base pairs and carries about 1600 genes (**Fig. 7.1**). Somewhere on the long arm of the chromosome resides the gene for rhodopsin, a light-sensitive protein active in the rod cells of our retinas. The rhodopsin gene determines perception of low-intensity light. People who carry the normal, wild-type allele of the gene see well in a dimly lit room and on the road at night. One simple change—a mutation—in the rhodopsin gene, however, diminishes light perception just enough to lead to night blindness. Other alterations in the gene cause the destruction of rod cells, resulting in total blindness. Medical researchers have so far identified more than 30 mutations in the rhodopsin gene that affect vision in different ways.

The case of the rhodopsin gene illustrates some very basic questions. Which of the 220 million base pairs on chromosome 3 make up the rhodopsin gene? How are the base pairs that constitute this gene arranged along the chromosome? How can a single gene sustain so many mutations that lead to such divergent phenotypic effects? In this chapter, we describe the ingenious experiments performed by geneticists during the 1950s and 1960s as they examined the relationships among mutations, genes, chromosomes, and phenotypes in an effort to understand, at the molecular level, what genes are and how they function.

We can recognize three main themes from the elegant work of these investigators. The first is that mutations are heritable changes in base sequence that can affect phenotype. The second is that physically, a gene is usually a specific protein-encoding segment of DNA in a discrete region of a chromosome. (We now know that some genes encode various kinds of RNA that do not get translated into protein.) Third, a gene is not simply a bead on a string, changeable only as a whole and only in one way, as some had thought. Rather, genes are divisible, and each gene's subunits—the individual nucleotide pairs of DNA—can mutate independently and can recombine with each other.

Knowledge of what genes are and how they work deepens our understanding of Mendelian genetics by providing a biochemical explanation for how genotype



**Figure 7.1** The DNA of each human chromosome contains hundreds to thousands of genes. The DNA of this human chromosome has been spread out and magnified 50,000 $\times$ . No topological signs reveal where along the DNA the genes reside. The darker, chromosome-shaped structure in the middle is a scaffold of proteins to which the DNA is attached.

© Dr. Don Fawcett/J.R. Paulson & U.K. Laemmli/Science Source

influences phenotype. One mutation in the rhodopsin gene, for example, causes the substitution of one particular amino acid for another in the construction of the rhodopsin protein. This single substitution changes the three-dimensional structure of rhodopsin and thus the protein's ability to absorb photons, ultimately altering a person's ability to perceive light.

## 7.1 Mutations: Primary Tools of Genetic Analysis

### Learning Objectives

1. Distinguish between the effects of mutation in somatic and germ-line cells.
2. Describe four types of point mutations: transitions, transversions, deletions, and insertions.
3. Summarize the factors associated with differences in mutation rate.
4. Explain how the fluctuation test and replica plating have shown that mutations arise randomly and spontaneously.

We saw in Chapter 3 that genes with one common allele are monomorphic, while genes with several common alleles in natural populations are polymorphic. The term *wild-type allele* has a clear definition for monomorphic genes, where the allele found on the large majority of chromosomes in the population under consideration is wild type. In the case of polymorphic genes, the definition is less straightforward. Some geneticists consider all alleles with a frequency of greater than 1% to be wild type, while others describe the many alleles present at appreciable frequencies in the population as *common variants* and reserve *wild-type allele* for use only in connection with monomorphic genes.

### Mutations Are Changes in DNA Base Sequences

A mutation that changes a wild-type allele of a gene (regardless of the definition) to a different allele is called a **forward**

**mutation**. The resulting novel mutant allele can be either recessive or dominant to the original wild-type allele. Geneticists often diagram forward mutations as  $A^+ \rightarrow a$  when the mutation is recessive to the wild-type allele, and as  $b^+ \rightarrow B$  when the mutation is dominant to the wild-type. Mutations can also cause a novel mutant allele to revert back to wild type ( $a \rightarrow A^+$ , or  $B \rightarrow b^+$ ) in a process known as **reverse mutation**, or **reversion**. In this chapter, we designate wild-type alleles, whether recessive or dominant to mutant alleles, with a plus sign (+).

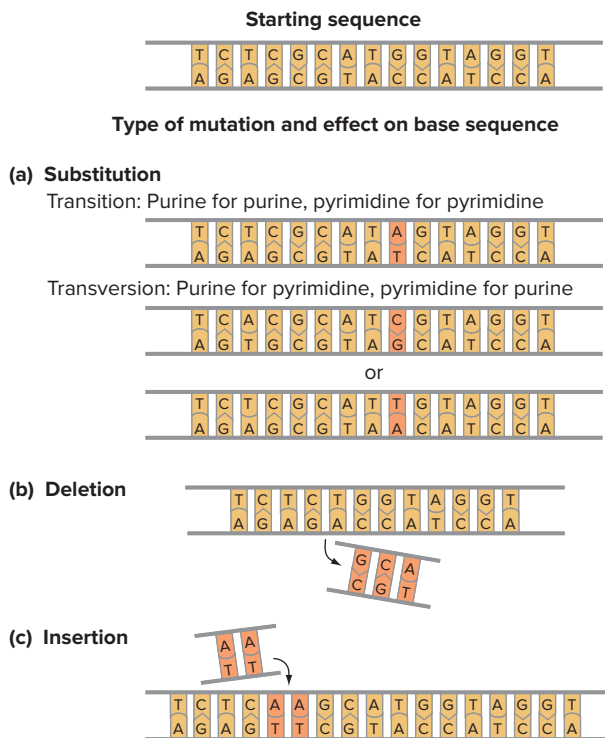
Mendel originally defined genes by the visible phenotypic effects—yellow or green, round or wrinkled—of their alternative alleles. In fact, the only way he knew that genes existed at all was because alternative alleles for seven particular pea genes had arisen through forward mutations. Mutations can occur in somatic cells or in germ-line cells. The mutations in Mendel's pea plants were heritable because they occurred in the germ-line cells of the plants and were thus transmitted through gametes. Close to a century later, knowledge of DNA structure clarified that such mutations are heritable changes in DNA base sequence. DNA thus carries the potential for genetic change in the same place it carries genetic information—the sequence of its bases.

### Mutations May Be Classified by How They Change DNA

A **substitution** occurs when a base at a certain position in one strand of the DNA molecule is replaced by one of the other three bases (**Fig. 7.2a**); after DNA replication, a new base pair will appear in the daughter double helix. Substitutions can be subdivided into *transitions*, in which one purine (A or G) replaces the other purine or one pyrimidine



**Figure 7.2** Point mutations classified by their effect on DNA.



(C or T) replaces the other; and *transversions*, in which a purine changes to a pyrimidine, or *vice versa*.

Other types of mutations rearrange DNA sequences rather than just change the identity of a base pair. A **deletion** occurs when a block of one or more nucleotide pairs is lost from a DNA molecule; an **insertion** is just the reverse—the addition of one or more nucleotide pairs (Figs. 7.2b and c). Deletions and insertions can be as small as a single base pair or as large as megabases (that is, millions of base pairs). Large deletions and insertions are only some of the complex mutations that can reorganize genomes by changing either the order of genes along a chromosome, the number of genes in the genome, or even the number of chromosomes in an organism. We discuss all such chromosomal rearrangements, which affect many genes at a time, in Chapter 13. Here, we focus on **point mutations** (transitions, transversions, and small deletions or insertions) that affect one or just a few base pairs in the DNA and thus alter only one gene at a time.

Only a small fraction of the mutations in a genome actually alter the nucleotide sequences of genes in a way that affects gene function. By changing one allele to another, these mutations modify the structure or amount of a gene’s protein product, and the modification in protein structure or amount can influence phenotype. Other mutations either alter genes in ways that do not affect their

function or change the DNA between genes. We discuss mutations without observable phenotypic consequences in Chapter 11; such mutations are invaluable for mapping genes and tracking differences between individuals. In the remainder of this chapter, we focus on those mutations that have an impact on gene function and thereby influence phenotype.

### Spontaneous Mutations Occur at a Very Low Rate

Mutations that modify gene function happen so infrequently that geneticists must examine a very large number of individuals from a formerly homogeneous population to detect the new phenotypes that reflect these mutations. In one ongoing study, dedicated investigators have monitored the coat colors of millions of specially bred mice and discovered that on average, a given gene mutates to a recessive allele in roughly 11 out of every 1 million gametes (Fig. 7.3). Studies of several other multicellular, eukaryotic organisms have yielded similar results: an average spontaneous rate of  $2\text{--}12 \times 10^{-6}$  mutations per gene per gamete.

Looking at the mutation rate from a different perspective, you could ask how many mutations might exist in the

**Figure 7.3** Rates of spontaneous mutation. (a) Wild-type (left) and mutant mouse coat colors: albino (middle), and brown (right). (b) Mutation rates from wild type to recessive mutant alleles for five coat color genes. Mice from highly inbred wild-type strains were mated with homozygotes for recessive coat color alleles. Progeny with mutant coat colors indicated the presence of recessive mutations in gametes produced by the inbred mice. a (left): © imageBROKER/SuperStock RF; (middle, right): © Charles River Laboratories



Locus <sup>a</sup>	Number of gametes tested	Number of mutations	Mutation rate ( $\times 10^{-6}$ )
$a^-$ (albino)	67,395	3	44.5
$b^-$ (brown)	919,699	3	3.3
$c^-$ (nonagouti)	150,391	5	33.2
$d^-$ (dilute)	839,447	10	11.9
$ln^-$ (leaden)	243,444	4	16.4
	2,220,376	25	11.2 (average)

<sup>a</sup> Mutation is from wild type to the recessive allele shown.

genes of an individual. To find out, you would simply multiply the rate of  $2\text{--}12 \times 10^{-6}$  mutations per gene per gamete times 27,000, the current estimate of the number of genes in the human genome, to obtain an answer of between 0.05 and 0.30 mutations per haploid genome. This very rough calculation would mean that, on average, one new mutation affecting phenotype could arise in every 3 to 20 human gametes.

**Different genes, different mutation rates**

Although the average mutation rate per gene per gamete is  $2\text{--}12 \times 10^{-6}$ , this number masks considerable variation in the mutation rates for different genes. Experiments with many organisms show that mutation rates range from less than  $10^{-9}$  to more than  $10^{-3}$  per gene per gamete. Variation in the mutation rate of different genes within the same organism reflects differences in gene size (larger genes are larger targets that sustain more mutations) as well as differences in the susceptibility of particular genes to the various mechanisms that cause mutations.

**Higher mutation rates in multicellular organisms than in bacteria**

Estimates of the average mutation rates in bacteria range from  $10^{-8}$  to  $10^{-7}$  mutations per gene per cell division. Although the units here are slightly different than those used for multicellular eukaryotes (because bacteria do not produce gametes), the average rate of mutation in multicellular eukaryotes still appears to be considerably higher than that in bacteria. The main reason is that numerous cell divisions take place between the formation of a zygote and meiosis, so mutations that appear in a gamete may have actually occurred many cell generations before the gamete formed. In other words, more chances exist for mutations to accumulate.

Some scientists speculate that the diploid genomes of multicellular organisms allow them to tolerate relatively high rates of mutation in their gametes because a zygote would have to receive recessive mutations in the same gene from both gametes for any deleterious effects to occur. In contrast, a bacterium would be affected by just a single mutation that disrupted its only copy of the gene.

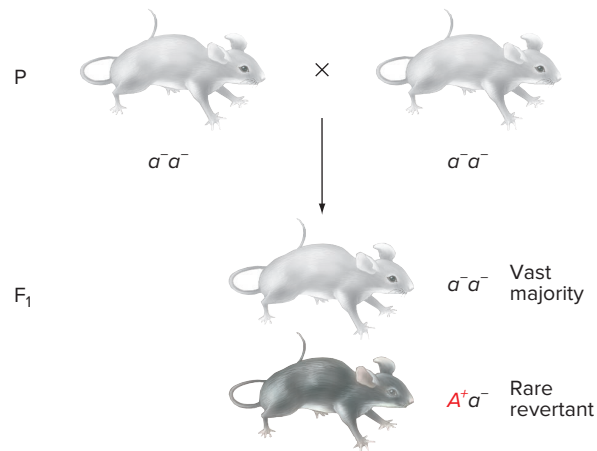
**Gene function: Easy to disrupt, hard to restore**

In the mouse coat color study, when researchers allowed brother and sister mice homozygous for a recessive mutant allele of one of the five mutant coat color genes to mate with each other, they could estimate the rate of

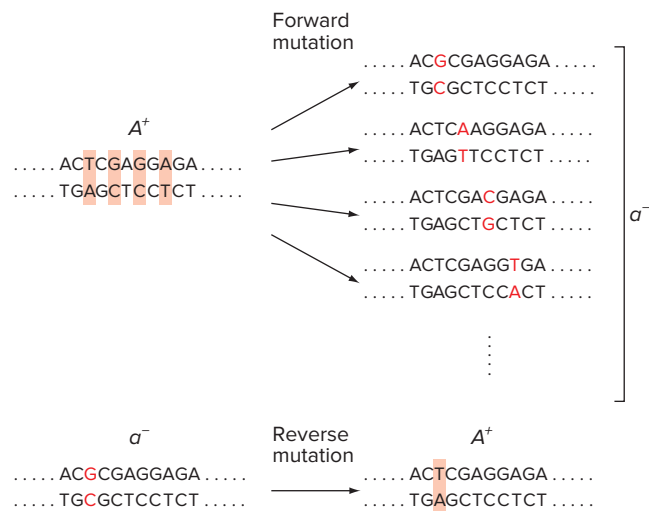
reversion by examining the F<sub>1</sub> offspring (Fig. 7.4a). Any progeny expressing the dominant wild-type phenotype for a particular coat color, of necessity, carried a gene that had sustained a reverse mutation. Calculations based on observations of several million F<sub>1</sub> progeny revealed a reverse mutation rate ranging from 0 to  $2.5 \times 10^{-6}$  per gene per gamete; the rate of reversion varied somewhat from gene to gene. In this study, then, the rate of reversion was significantly lower than the rate of forward mutation, most likely because while many ways exist to disrupt gene function, there are at most a few ways to restore function once it has been disrupted (Fig. 7.4b). The conclusion that the rate of reversion is significantly

**Figure 7.4 Detecting revertants.** (a) Rare revertants of  $a^-$  mutations that are recessive to wild-type alleles ( $A^+$ ) are detected as wild-type grey ( $A^+ a^-$ ) progeny of albino ( $a^- a^-$ ) mice. (b) The rate of forward mutation is usually much higher than the rate of reversion. Many different mutations can disrupt a gene's function, while at best only a few mutations can restore function to a previously inactivated gene.

**(a) Rare reverse mutation of the albino gene**



**(b) Forward mutation rate is higher than reverse mutation rate**



lower than the rate of forward mutation holds true for most types of mutations. In one extreme example, deletions of more than a few nucleotide pairs can never revert because DNA information that has disappeared from the genome cannot reappear spontaneously.

### Higher mutation rate in human sperm than in human eggs

New technologies that will be explained in detail in later chapters have enabled researchers to determine the DNA sequence of the entire genome of thousands of people. By comparing the genome sequences of parents and their children, scientists have measured the human mutation rate with great precision. They found that the average value is about one mutation per hundred million base pairs (bp) per gamete (or  $1 \times 10^{-8}$ ). Because the haploid human genome is about  $3 \times 10^9$  bp, each gamete contains on average about 30 mutations, and each child contains about 60 mutations—that is, 60 base pairs that are different than those in either of their parents' genomes. You should note that this number includes all DNA changes, only very few of which influence phenotype.

Interestingly, most of these 60-odd new mutations in each human are obtained from the sperm rather than the egg. Advances in genome sequencing technology have recently made it possible to sequence the haploid genome contained in a single sperm. (See the Fast Forward Box *Crossovers Mapped in Chromosomes of Human Sperm* in Chapter 5.) By comparing the genome sequences of more than 100 individual sperm from the same person, the per-bp mutation rate was found to be  $2\text{--}4 \times 10^{-8}$ , which indicates that most of the new mutations seen in children come from the sperm rather than the egg.

The idea that sperm carry more mutations than oocytes makes sense. The reason is that more rounds of cell divisions are needed to produce human sperm than human eggs, presenting more opportunities for mutations to occur. Recall from Chapter 4 that human females are born with essentially all of the primary oocytes they will ever produce. It has been estimated that the germ-line cells of a female zygote need to undergo only 24 rounds of mitotic cell divisions to produce all of these oocytes. Male germ-line cells, on the other hand, undergo mitosis continually throughout life. Starting from a male zygote, the number of cell divisions to generate a sperm at age 13 is estimated to be 36. After that, about 23 rounds of mitotic divisions occur per year in the male germ line, meaning that at age 20, the cell lineage producing a given sperm has undergone 200 divisions; at age 30, 430; and at age 45, 770. Therefore, in humans, most new mutations found in the progeny come from the sperm rather than from the egg. Moreover, the older the father, the more mutations are likely to be found in his sperm.

## Spontaneous Mutations Arise from Random Events

Because spontaneous mutations affecting a gene occur so infrequently, it is difficult to study the events that produce them. To overcome this problem, researchers turned to bacteria as the experimental organisms of choice. It is easy to grow many millions of individuals and then search rapidly through enormous populations to find the few that carry a novel mutation. In one study, investigators spread wild-type bacteria on the surface of agar containing sufficient nutrients for growth as well as a large amount of a bacteria-killing substance, such as an antibiotic or a bacteriophage. Although most of the bacterial cells died, a few showed resistance to the bactericidal substance and continued to grow and divide. The descendants of a single resistant bacterium, produced by many rounds of binary fission, formed a mound of genetically identical cells called a **colony**.

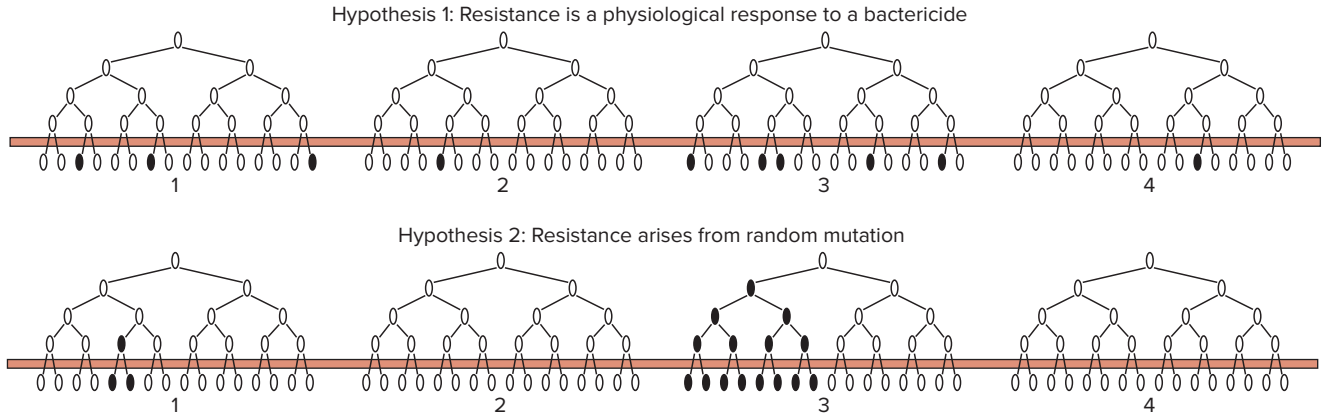
The few bactericide-resistant colonies that appeared presented a puzzle. Had the cells in the colonies somehow altered their internal biochemistry to produce a life-saving response to the antibiotic or bacteriophage? Or did they carry heritable mutations conferring resistance to the bactericide? And if they did carry mutations, did those mutations arise by chance from random spontaneous events that take place continuously, even in the absence of a bactericidal substance, or did they only arise in response to environmental signals (in this case, the addition of the bactericide)?

In 1943, Salvador Luria and Max Delbrück devised an experiment to examine the origin of bacterial resistance (**Fig. 7.5**). According to their reasoning, if bacteriophage-resistant colonies arise in direct response to infection by bacteriophages, separate suspensions of bacteria containing equal numbers of cells will generate similar, small numbers of resistant colonies when spread in separate petri plates on nutrient agar suffused with phages. By contrast, if resistance arises from mutations that occur spontaneously even when the phages are not present, then different liquid cultures, when spread on separate petri plates, will generate very different numbers of resistant colonies. The reason is that the mutation conferring resistance can, in theory, arise at any time during the growth of the culture. If the mutation occurs early, the cell in which it happens will produce many mutant progeny prior to petri plating; if it happens later, far fewer mutant progeny will be present when the time for plating arrives. After plating, these numerical differences will show up as fluctuations in the numbers of resistant colonies growing in the different petri plates.

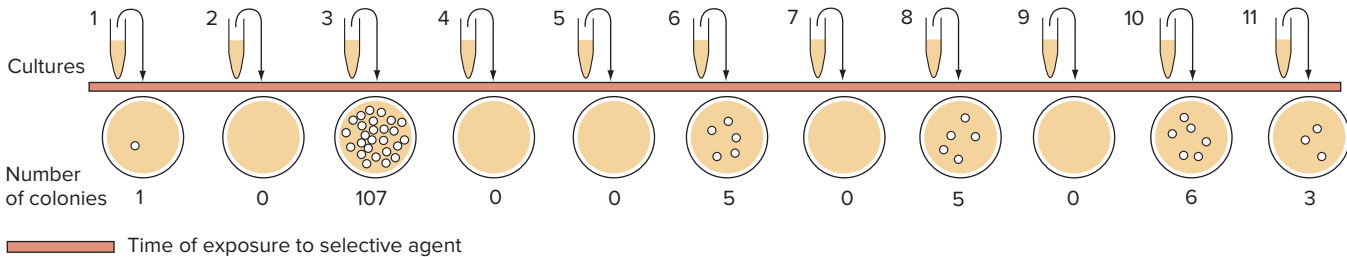
The results of this **fluctuation test** were clear: Most plates supported zero to a few resistant colonies, but a few harbored hundreds of resistant colonies. From this

**Figure 7.5** The Luria-Delbrück fluctuation experiment. (a) Hypothesis 1: If resistance arises only after exposure to a bactericide, all bacterial cultures of equal size should produce roughly the same number of resistant colonies. Hypothesis 2: If random mutations conferring resistance arise before exposure to bactericide, the number of resistant colonies in different cultures should vary (fluctuate) widely. (b) Actual results showing large fluctuations suggest that mutations in bacteria occur as spontaneous mistakes independent of exposure to a selective agent.

(a) Two hypotheses for the origin of bactericide resistance



(b) Fluctuation test results



observation of a substantial fluctuation in the number of resistant colonies in different petri plates, Luria and Delbrück concluded that bacterial resistance arises from mutations that exist before exposure to bacteriophages. After exposure, however, the bactericide in the petri plate becomes a selective agent that kills off nonresistant cells, allowing only the preexisting resistant ones to survive. **Figure 7.6** illustrates how researchers used another technique, known as **replica plating**, to demonstrate even more directly that the mutations conferring bacterial resistance occur before the cells encounter the bactericide that selects for their resistance.

These key experiments showed that bacterial resistance to phages and other bactericides is the result of mutations, and these mutations do not arise in particular genes as a directed response to environmental change. Instead, mutations occur spontaneously as a result of random processes that can happen at any time and hit the genome at any place. Once such random changes occur, however, they usually remain stable. If the resistant mutants of the Luria-Delbrück experiment, for example, were grown for many generations in medium that did not

contain bacteriophages, they would nevertheless remain resistant to this bactericidal virus.

We next describe some of the many kinds of random events that can cause mutations. We also discuss how cells cope with these events and minimize mutation creation.

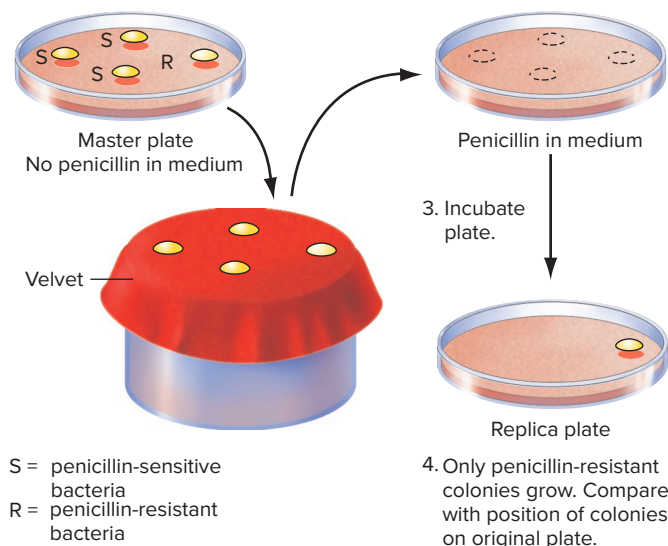
**essential concepts**

- *Mutations* are heritable alterations in the base sequence of DNA.
- *Point mutations* change one or a few base pairs; they include *substitutions* (*transitions* and *transversions*) and small *insertions* and *deletions*.
- Spontaneous mutation rates are low and vary among different genes and organisms.
- The more cells divide, the more likely it is that mutations will accumulate in their genomes.
- Results of the *fluctuation test* and *replica plating* experiments showed that resistance mutations arise randomly in bacterial cells prior to bactericide exposure.

**Figure 7.6** Replica plating verifies that bacterial resistance is the result of preexisting mutations. (a) Pressing a *master plate* onto a velvet surface transfers some cells from each bacterial colony onto the velvet. Pressing a *replica plate* onto the velvet then transfers some cells from each colony onto the replica plate. Investigators track which colonies on the master plate can grow on the replica plate (here, only penicillin-resistant ones). (b) Colonies on a master plate without penicillin are transferred sequentially to three replica plates with penicillin. Resistant colonies grow in the same positions on all three replicas, showing that some colonies on the master plate had multiple resistant cells even before exposure to the antibiotic.

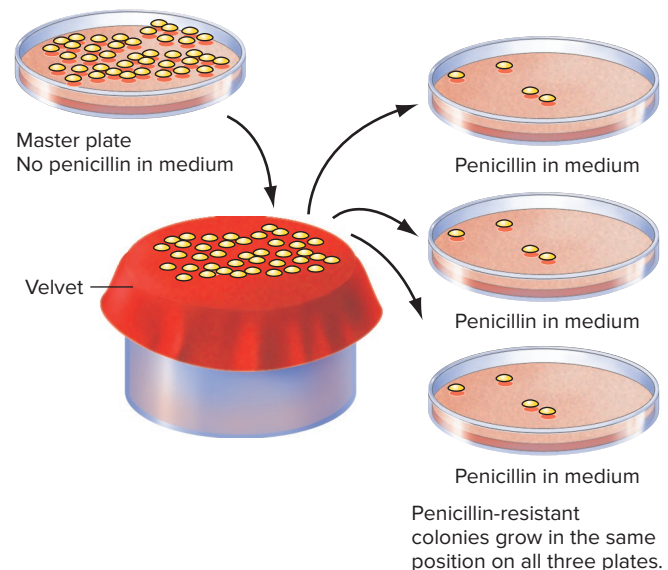
**(a) The replica plating technique**

1. Invert master plate; pressing against velvet surface leaves an imprint of colonies. Save plate.
2. Invert second plate (replica plate); pressing against velvet surface picks up colony imprint.
3. Incubate plate.
4. Only penicillin-resistant colonies grow. Compare with position of colonies on original plate.



**(b) Mutations occur prior to penicillin exposure**

- $10^7$  colonies of penicillin-sensitive bacteria
- Make three replica plates. Incubate to allow penicillin-resistant colonies to grow.



## 7.2 Molecular Mechanisms That Alter DNA Sequence

### learning objectives

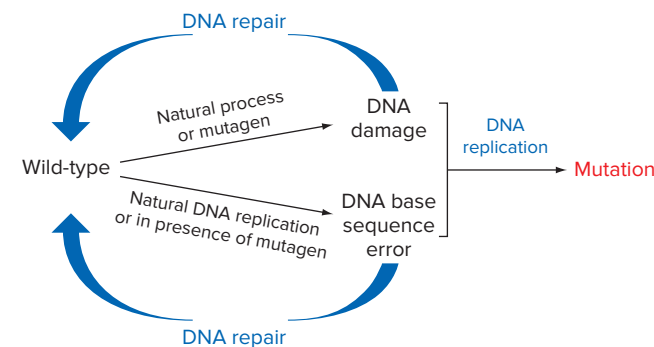
1. Outline natural processes that can produce mutations by damaging DNA.
2. Explain how errors in DNA replication can cause mutations.
3. Define *mutagen* and describe how mutagens are used in genetic research.
4. Describe how the Ames test can detect potential carcinogens.

The creation of a heritable mutation is the outcome of several competing processes: mutation, repair, and replication (Fig. 7.7). First, of course, a random event must occur to alter the DNA. Two different kinds of events initiate DNA changes: Either DNA can be damaged by chemical reactions or irradiation, or alternatively, mistakes can happen when DNA is copied during replication.

When DNA changes first occur, they are not yet actual mutations but only potential mutations. The reason is that most of them are quickly repaired by a variety of enzymatic systems within cells. These DNA repair machines are engaged in a continual race with DNA replication (Fig. 7.7). If repair of damaged DNA or misincorporated nucleotides occurs before the next round of DNA replication, then the sequence is corrected and no mutation will result. However, if the repair enzymes do not correct the problem before the next round of DNA replication, the mutation becomes established permanently in both strands of the double helix and a heritable mutation is the outcome.

In this section, we describe some of the most important mechanisms that can change DNA sequences. The subsequent

**Figure 7.7** Point mutations result when DNA replication wins the race with DNA repair. An alteration that occurs in DNA is heritable only if DNA repair fails to reverse the change before the next round of DNA replication.



section of this chapter will discuss various biochemical pathways that biological systems have evolved to minimize the mutagenic consequences of these DNA alterations.

## Natural Processes Cause Spontaneous Mutations Through DNA Damage

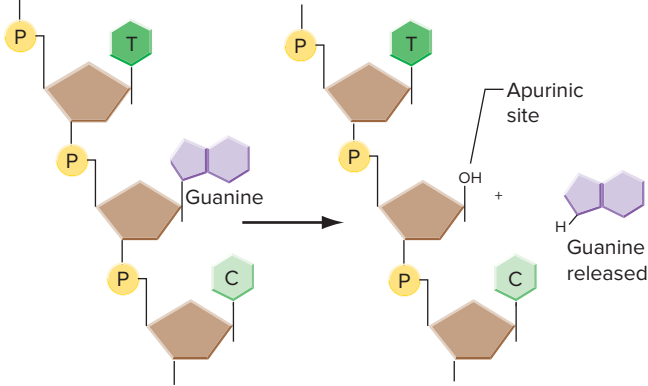
Chemical and physical assaults on DNA are quite frequent. Geneticists estimate, for example, that the hydrolysis of a purine base, A or G, from the deoxyribose-phosphate

backbone occurs 1000 times an hour in every human cell. This kind of DNA alteration is called **depurination** (Fig. 7.8a). Because the resulting *apurinic site* cannot specify a complementary base, the DNA replication process introduces a random base opposite the apurinic site, causing a mutation in the newly synthesized complementary strand three-quarters of the time.

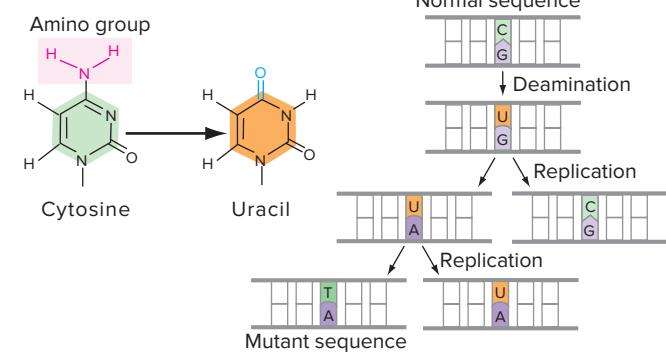
Another naturally occurring process that may modify DNA's information content is **deamination**: the removal of an amino ( $-NH_2$ ) group. Deamination can change cytosine (C) to uracil (U), the nitrogenous base

**Figure 7.8** How natural processes can change the information stored in DNA. (a) In depurination, the hydrolysis of A or G bases leaves a DNA strand with an unspecified base. (b) In deamination, the removal of an amino group from C initiates a process that causes a transition after DNA replication. (c) X-rays break the sugar-phosphate backbone and thereby split a DNA molecule into smaller pieces, which may be ligated back together improperly. (d) Ultraviolet (UV) radiation causes adjacent Ts to form dimers, which can disrupt the readout of genetic information. (e) Irradiation causes the formation of *free radicals* (such as oxygen molecules with an unpaired electron) that can alter individual bases. Here, the pairing of the altered base GO with A creates a transversion that changes a G–C base pair to T–A.

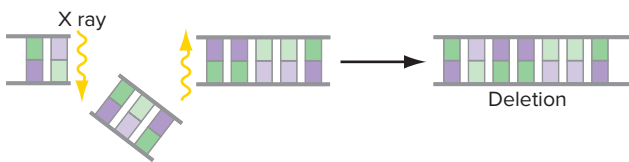
### (a) Depurination



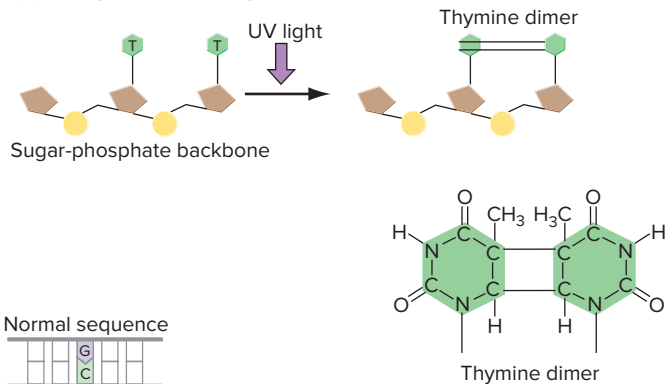
### (b) Deamination



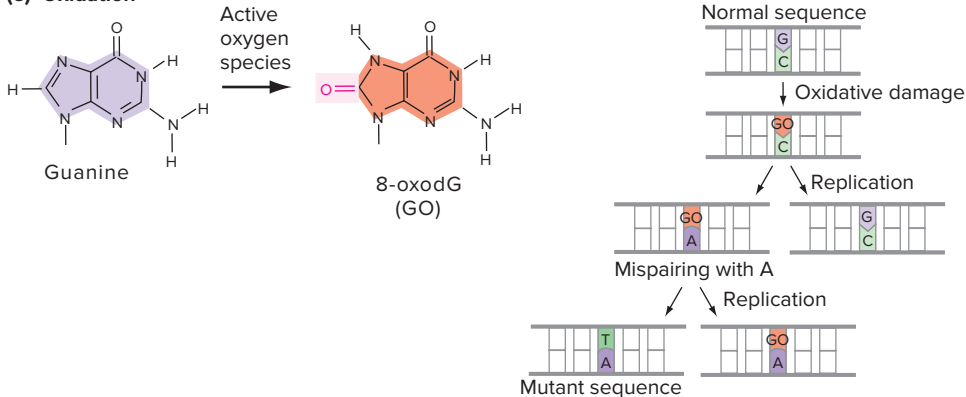
### (c) X-rays break the DNA backbone



### (d) UV light produces thymine dimers



### (e) Oxidation



found in RNA but not in DNA. Because U pairs with A rather than G, deamination followed by replication may alter a C–G base pair to a T–A pair in future generations of DNA molecules (Fig. 7.8b); such a C–G to T–A change is a transition mutation.

Other assaults include naturally occurring radiation such as cosmic rays and X-rays, which break the sugar-phosphate backbone (Fig. 7.8c); ultraviolet light, which causes adjacent thymine residues to become chemically linked into **thymine dimers** (Fig. 7.8d); and oxidative damage to any of the four bases (Fig. 7.8e). If not repaired before DNA replication, all of these changes alter the information content of the DNA molecule permanently.

### Mistakes in DNA Replication Also Cause Spontaneous Mutations

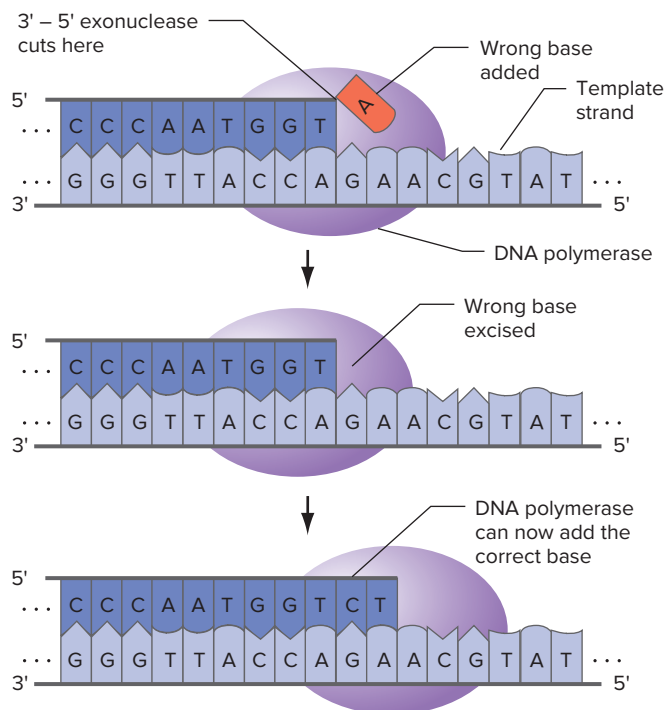
If the cellular machinery for some reason incorporates an incorrect base during replication, for instance, a C opposite an A instead of the expected T, then during the next replication cycle, one of the daughter DNAs will have the normal A–T base pair, while the other will have a mutant G–C. Careful measurements of the fidelity of replication *in vivo*, in both bacteria and human cells, show that such errors are exceedingly rare, occurring less than once in every  $10^9$  base pairs. That rate is equivalent to typing this entire book 1000 times while making only one typing error. Considering the complexities of helix unwinding, base pairing, and polymerization, this level of accuracy is amazing. How do cells avoid most DNA replication errors, and what kinds of mistakes occur nonetheless when DNA is copied?

#### The proofreading function of DNA polymerase

The replication machinery minimizes errors through successive stages of correction. In the test tube, DNA polymerases replicate DNA with an error rate of about one mistake in every  $10^6$  bases copied. This rate is about 1000-fold worse than that achieved by the cell. Even so, it is impressively low and is attained only because polymerase molecules provide, along with their polymerization function, a proofreading/editing function in the form of a nuclease that becomes active whenever the polymerase makes a mistake. This nuclease portion of the polymerase molecule, called the *3'-to-5' exonuclease*, recognizes a mispaired base and excises it, allowing the polymerase to copy the nucleotide correctly on the next try (Fig. 7.9). Without its nuclease portion, DNA polymerase would have an error rate of one mistake in every  $10^4$  bases copied, so this editing function improves the fidelity of replication 100-fold.

DNA polymerase *in vivo* is part of a replication system including many other proteins that improve on the

**Figure 7.9 DNA polymerase's proofreading function.** If DNA polymerase mistakenly adds an incorrect nucleotide at the 3' end of the strand it is synthesizing, the enzyme's 3'-to-5' exonuclease activity removes this nucleotide, giving the enzyme a second chance to add the correct nucleotide.



error rate collectively another 10-fold, bringing it to within about 100-fold of the fidelity attained by the cell. The 100-fold higher accuracy of the cell depends on a backup system called *methyl-directed mismatch repair* that notices and corrects residual errors in the newly replicated DNA. We present the details of this repair system later in the chapter when we describe the various ways in which cells attempt to correct mutations once they occur.

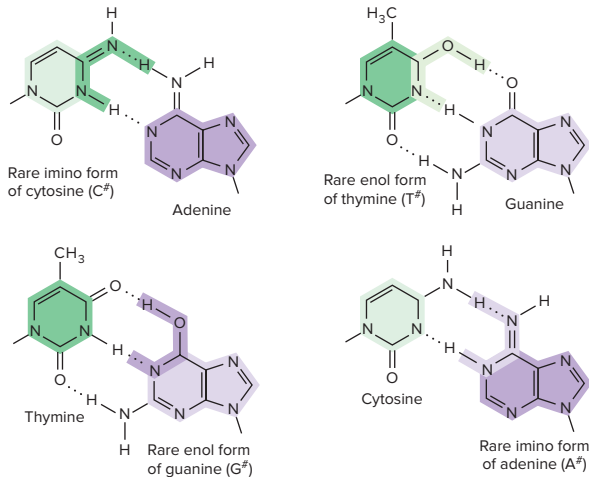
#### Base tautomerization

One reason why DNA polymerase may make mistakes is the **tautomerization** of bases. Each of the four bases has two **tautomers**, similar chemical forms that interconvert continually. The equilibrium between the tautomers is such that each base is almost always in the form in which A pairs with T and G pairs with C. However, if by chance a base in the template strand is in its rare tautomeric form when DNA polymerase arrives, the wrong base will be incorporated into the newly synthesized chain because the rare tautomers pair differently than do the normal forms (Fig. 7.10a). If the misincorporated nucleotide is not corrected by mismatch repair before the next round of replication, a point mutation results (Fig. 7.10b).

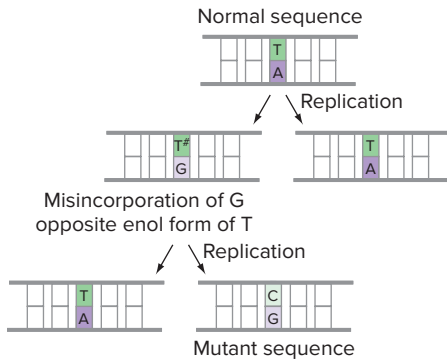
**Figure 7.10** How base tautomerization causes mutation.

(a) Rare tautomeric forms of the four bases have different pairing abilities than the usual base forms. (b) In its rare enol form, T causes DNA polymerase to insert a G in the complementary strand. If the mismatched T:G base pair is not repaired to T:A before the next round of replication, a T:A-to-C:G transition mutation is established in both strands of one daughter DNA molecule.

(a) Rare tautomeric forms of bases have altered base pairing ability.



(b) Tautomerization causes single base pair mutations.



### Unstable trinucleotide repeats

In 1992, molecular geneticists discovered a completely unexpected type of mutation in humans: the excessive amplification of a CGG base triplet normally repeated only a few to 50 times in succession. If, for example, a normal allele of a gene carries five consecutive repetitions of the base triplet CGG (that is, CGGCGGCGGCGGCGG on one strand), an abnormal allele could carry 200 repeats in a row. Repeats of several other trinucleotides—CAG, CTG, GCC, and GAA—can also be unstable, such that the number of repeats often increases or decreases in different somatic cells of a single individual. Instability can also occur during gamete production, resulting in changes in repeat number from one generation to the next.

**Unstable trinucleotide repeats** have now been found within about 20 different human genes, all associated with

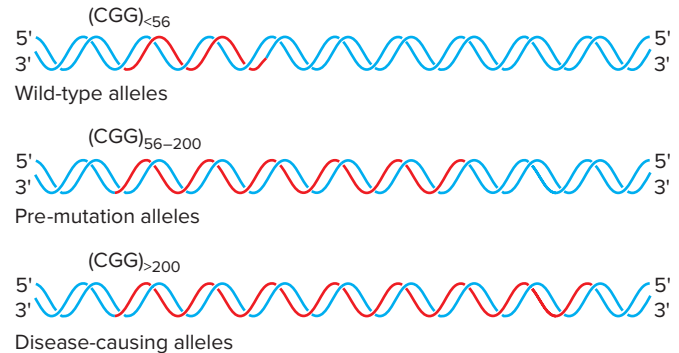
neurodegenerative diseases. In all cases, an expansion of the repeats beyond a certain number results in a disease-causing allele. The Fast Forward Box entitled *Trinucleotide Repeat Diseases: Huntington Disease and Fragile X Syndrome* explains that trinucleotide repeat diseases can be subdivided into two main groups according to the location of the repeats relative to the the part of the gene that specifies the protein product. One is exemplified by fragile X syndrome, the most common form of intellectual disability in boys; the other group is represented by Huntington disease, a neurological disorder discussed in Chapter 2.

A general feature of both groups of trinucleotide repeat diseases is that the more repeats at a particular location, the higher the probability that expansion and contraction will occur. Because larger repeat numbers mean more instability, some alleles with intermediate numbers of trinucleotide repeats behave as so-called **pre-mutation alleles** (Fig. 17.11a). For example, in fragile X syndrome, individuals with pre-mutation alleles have a normal phenotype, but the expanded repeat number means that such pre-mutation alleles are highly likely to expand or contract during replication. Carriers of pre-mutation alleles thus have a high probability of giving new disease alleles (with an expanded number of repeats) to their children (Fig. 7.11b).

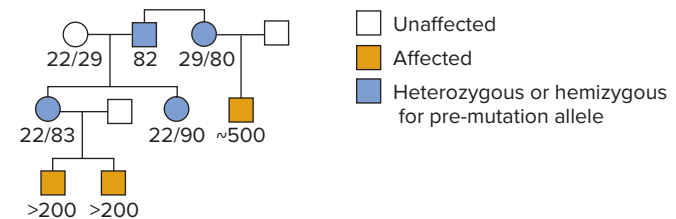
**Figure 7.11** Inheritance of fragile X syndrome.

(a) Wild-type, pre-mutation, and disease-causing alleles for fragile X syndrome differ in the number of CGG trinucleotide repeats. Disease alleles are nonfunctional. Pre-mutation alleles provide normal gene function, but they have a high probability of triplet repeat expansion during DNA replication in female germ-line cells. (b) Normal females heterozygous for pre-mutation alleles are likely to produce gametes with expanded triplet repeat numbers.

(a) Effect of (CGG) repeat number



(b) A fragile X pedigree





## FAST FORWARD



Sprinters: © Robert Michael/Corbis RF

## Trinucleotide Repeat Diseases: Huntington Disease and Fragile X Syndrome

The approximately 20 known neurodegenerative diseases caused by genes with unstable trinucleotide (triplet) repeats fall into two categories: **polyQ diseases** and **non-polyQ diseases** (where Q is the symbol for the amino acid glutamine). In polyQ disease genes, the repeated triplet is always CAG, while in non-polyQ disease genes, the trinucleotide repeat may be either CGG, CTG, GCC, or GAA. The two types of triplet repeat diseases are distinguished by the effect of the repeat sequence on gene function. In polyQ diseases, a disease allele with too many triplet repeats encodes an abnormal protein. Non-polyQ disease alleles encode either no protein or decreased protein amounts. The differences in the two classes of triplet repeat diseases are illustrated by the best-known example of each: Huntington disease, a polyQ disease; and fragile X syndrome, a non-polyQ disease.

Huntington disease affects about 1 in 10,000 people worldwide. The symptoms usually start at about 40 years of age and include muscle coordination difficulties, cognitive decline, and psychiatric problems. You saw in Chapter 2 that Huntington disease is inherited through autosomal dominant mutant alleles (*HD*). While normal *HD*<sup>+</sup> alleles have between 6 and 28 CAG repeats, *HD* disease alleles have an expanded repeat region that has 36 or more CAG repeats. The run of CAGs in the *HD* gene are in the *open reading frame*, or *ORF*, that contains the actual instructions to build a protein from its constituent *amino acids*.

Each CAG specifies that the amino acid glutamine (Q) should be added to the HD protein, so the normal protein has 6 to 28 Q amino acids in a row in its so-called *polyQ region* (Fig. A). An *HD* allele with 36 or more repeats encodes a mutant HD protein with an expanded polyQ region that is toxic to nerve cells. (Proteins encoded by pre-mutation alleles function normally but the alleles have an unstable repeat number.) Scientists do not yet understand the normal function of HD in nerve cells or the reason why the mutant HD protein is toxic.

PolyQ disease alleles like *HD* are called *gain-of-function* mutants because they specify proteins whose functions are qualitatively different from those of the corresponding wild-type protein. Typical for many gain-of-function mutants, polyQ disease alleles show dominant inheritance because the mutant polyQ proteins are toxic even in the presence of the normal proteins.

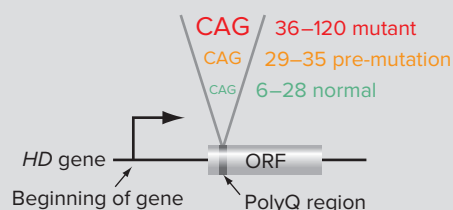
Non-polyQ diseases are exemplified by fragile X syndrome, a leading cause of inherited intellectual disability, affecting about 1 in 4000 males and 1 in 8000 females. The disease is caused by expansion of a CGG repeat region in an X-linked gene called *FMR-1* (for *fragile X mental retardation-1*).

The CGGs of *FMR-1* are located in a region of the gene outside of the ORF called the *5' UTR* (Fig. B). Normal *FMR-1*<sup>+</sup> genes have between 6 and 55 CGG repeats; expansion of the repeat number to 200 CGGs or more results in an *FMR-1* disease allele that cannot produce the *FMR-1* protein. Without *FMR-1* protein, nerve cells cannot properly form connections called *synapses*.

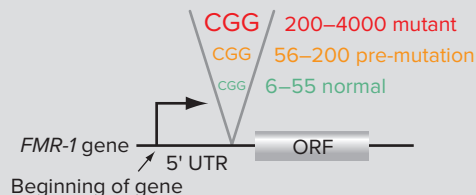
A feature common to all non-polyQ diseases is that the triplets are located in a part of the gene outside of the ORF. The

**Figure A Huntington disease: a polyQ repeat disease.**

The *HD* gene has a run of CAG repeats that specify glutamines (Qs) in the open reading frame (ORF). *HD* disease alleles direct the synthesis of a mutant, toxic HD protein with an expanded polyQ region. Pre-mutation alleles with an intermediate number of CAGs produce a normally functioning protein, but the allele is unstable.

**PolyQ disease: Huntington disease****Figure B Fragile X syndrome: a non-polyQ repeat disease.**

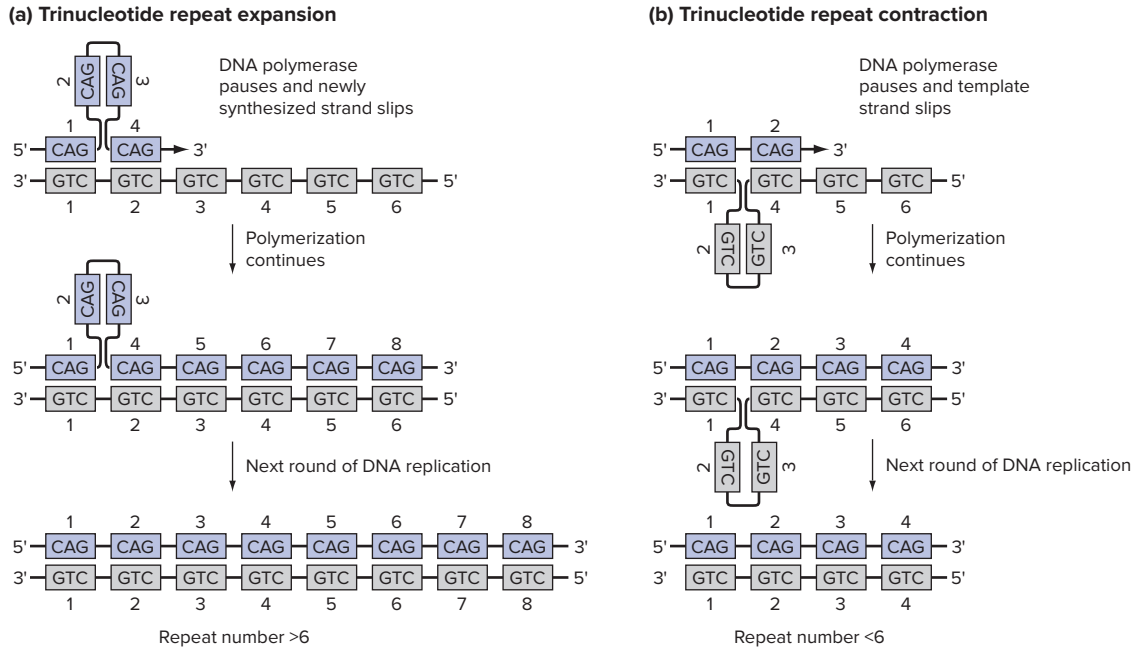
The *FMR-1* gene has a run of CGG repeats in the 5' untranslated region (5' UTR) outside the ORF. *FMR-1* disease alleles have an expanded repeat number, and this prevents synthesis of the gene's protein product. Pre-mutation alleles with an intermediate number of CGGs make normal amounts of protein, but these alleles are unstable.

**Non-polyQ disease: fragile X syndrome**

expanded repeats in non-polyQ disease genes generally prevent protein production, and so non-polyQ disease genes are *loss-of-function* alleles. Because females heterozygous for the disease allele have at least some disease symptoms most of the time, fragile X syndrome shows X-linked dominant inheritance with incomplete penetrance and variable expressivity. Other non-polyQ disease alleles may show either dominant or recessive inheritance patterns, depending on whether two doses or one dose of the normal gene product is required to avoid disease symptoms.

The triplet repeat diseases illustrate two fundamental principles regarding mutations. First, mutations may affect either the nature of the gene product (polyQ diseases) or the amount of the gene product (non-polyQ diseases). Second, certain DNA sequences can mutate at surprisingly high frequencies in special circumstances, as seen by pre-mutation alleles for either Huntington disease or fragile X syndrome. These two principles will be important themes in subsequent chapters.

**Figure 7.12** Expansion of trinucleotide repeats by slipped mispairing during DNA replication. **(a)** Pausing of DNA polymerase at repeat sequences during DNA replication allows slippage of the newly synthesized DNA strand (*blue*) relative to the template strand (*gray*). Because of the repeats, the slipped strand can still pair with the template, and DNA polymerization can continue. Another round of DNA replication will establish the additional repeats in double-stranded DNA. **(b)** Similarly, slippage of the template strand relative to the newly synthesized DNA strand can result in the deletion of repeats.



Researchers do not understand well a curious feature of trinucleotide repeat diseases: Pre-mutation alleles of particular genes tend to expand either in the male or female germ lines, but not both. For example, in Fig. 7.11b you can see that the alleles causing fragile X syndrome were inherited from mothers with pre-mutation alleles, but the repeat number does not expand appreciably in the sperm produced by a father with a pre-mutation allele. Strangely, for Huntington disease the situation is the opposite: Disease alleles almost always originate in the male, but not in the female, germ line.

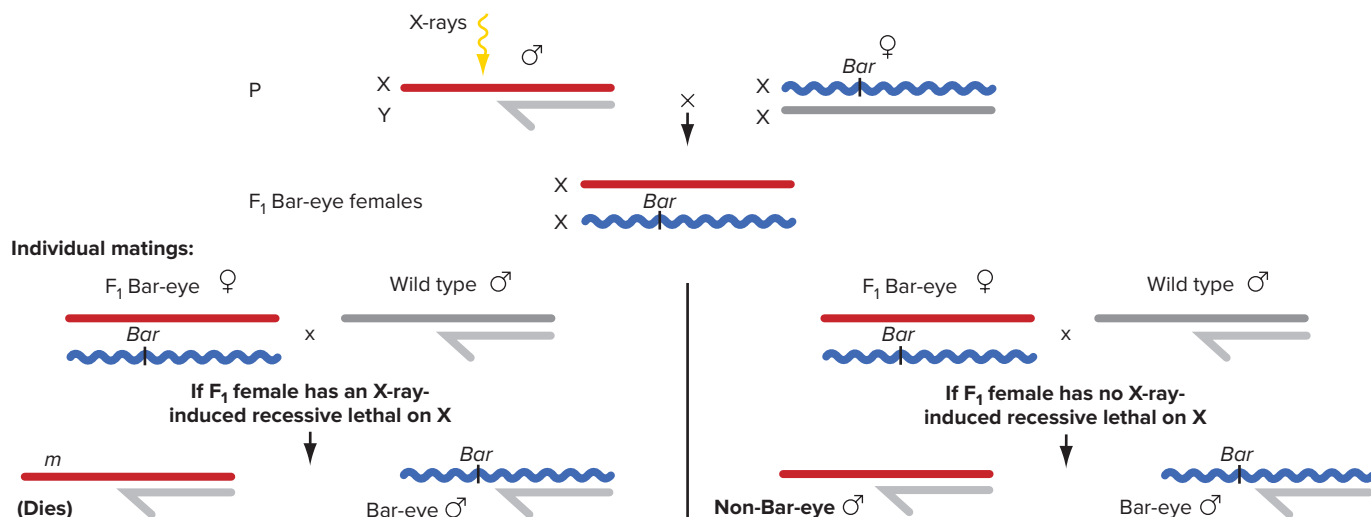
A variety of biochemical mechanisms could be responsible for trinucleotide repeat expansion and contraction. One particularly well-characterized mechanism is **slipped mispairing** during DNA replication. DNA polymerase often pauses as it replicates through repeat regions, which allows one DNA strand (either the newly synthesized strand or the template strand) to slip relative to the other one (**Fig. 7.12**). Because the sequence contains repeats, the slipped strand and the other strand can pair out of register, forming a loop. After another round of DNA replication, this slipped mispairing can result in expansion or contraction of trinucleotide repeat number in both DNA strands.

## Mutagens Induce Mutations

Mutations make genetic analysis possible, but most mutations appear spontaneously at such a low rate that researchers have looked for controlled ways to increase their occurrence. H. J. Muller, an original member of Thomas Hunt Morgan's *Drosophila* group, first showed that exposure to a dose of X-rays higher than the naturally occurring level increases the mutation rate in fruit flies (**Fig. 7.13**).

Muller exposed male *Drosophila* to increasingly large doses of X-rays and then mated these males with females that had one X chromosome containing an easy-to-recognize dominant mutation causing Bar eyes. This X chromosome (called a *Balancer*) also carried chromosomal rearrangements known as *inversions* that prevented it from crossing-over with other X chromosomes. (Chapter 13 explains the details of this phenomenon.) Some of the F<sub>1</sub> daughters of this mating were heterozygotes carrying a mutagenized X from their father and a *Bar*-marked X from their mother. If X-rays induced a recessive lethal mutation anywhere on the paternally derived X chromosome, then these F<sub>1</sub> females would be

**Figure 7.13** Exposure to X-rays increases the mutation rate in *Drosophila*. F<sub>1</sub> females are constructed that have an irradiated paternal X chromosome (red line) and a *Bar*-marked *Balancer* maternal X chromosome (wavy blue line). These two chromosomes cannot recombine because the *Balancer* chromosome prevents crossing-over. Single F<sub>1</sub> females, each with a single X-ray-exposed X chromosome from their father, are then individually mated with wild-type males. If the paternal X chromosome in any one F<sub>1</sub> female has an X-ray-induced recessive lethal mutation (*m*), she can produce only Bar-eye sons (left). If the X chromosome has no such mutation, this F<sub>1</sub> female will produce both Bar-eye and non-Bar-eye sons (right).



unable to produce non-Bar-eye sons. Thus, simply by noting the presence or absence of non-Bar-eye sons, Muller could establish whether a mutation had occurred in any of the more than 1000 genes on the X chromosome that are essential to *Drosophila* viability. He concluded that the greater the X-ray dose, the greater the frequency of recessive lethal mutations.

Any physical or chemical agent that raises the frequency of mutations above the spontaneous rate is called a **mutagen**. Researchers use many different mutagens to produce mutations for study. With the Watson-Crick model of DNA structure as a guide, they can understand the action of most mutagens at the molecular level. The X-rays used by Muller to induce mutations on the X chromosome, for example, can break the sugar-phosphate backbones of DNA strands, sometimes at the same position on the two strands of the double helix. Multiple double-strand breaks produce DNA fragmentation, and the improper stitching back together of the fragments can cause small deletions (review Fig. 7.8c) or large deletions and other rearrangements that will be discussed in Chapter 13.

Another molecular mechanism of mutagenesis involves mutagens known as **base analogs**, which are so similar in chemical structure to the normal nitrogenous bases that the replication machinery can incorporate them into DNA (Fig. 7.14a). Because a base analog may have tautomeric forms with pairing properties different from

those of the base it replaces, the analog can cause base substitutions on the complementary strand synthesized in the next round of DNA replication.

Other chemical mutagens generate substitutions by directly altering a base's chemical structure and properties (Fig. 7.14b). Again, the effects of these changes become fixed in the genome when the altered base causes incorporation of an incorrect complementary base during a subsequent round of replication.

Yet another class of chemical mutagens consists of compounds known as **intercalators**: flat, planar molecules that can sandwich themselves between successive base pairs and disrupt the machinery for replication, generating deletions or insertions of a single base pair (Fig. 7.14c). The intercalator *proflavin* is often used in genetic research precisely for this reason.

## Many Mutagens Are Carcinogens

Although only mutations that occur in the germ line can be passed on to the next generation, mutations in somatic cells can still have an impact on the well-being and survival of individuals. Somatic mutations in genes that help regulate the cell cycle may, for example, lead to cancer. For this reason, many mutagens act as *carcinogens* (cancer-causing agents).

**Figure 7.14 How mutagens alter DNA.** (a) Base analogs incorporated into DNA may pair aberrantly, allowing the addition of incorrect nucleotides to the opposite strand during replication. (b) Some mutagens alter the structure of bases such that they pair inappropriately in the next round of replication. (c) Intercalating agents are roughly the same size and shape as a base pair of the double helix. Their incorporation into DNA produces insertions or deletions of single base pairs.

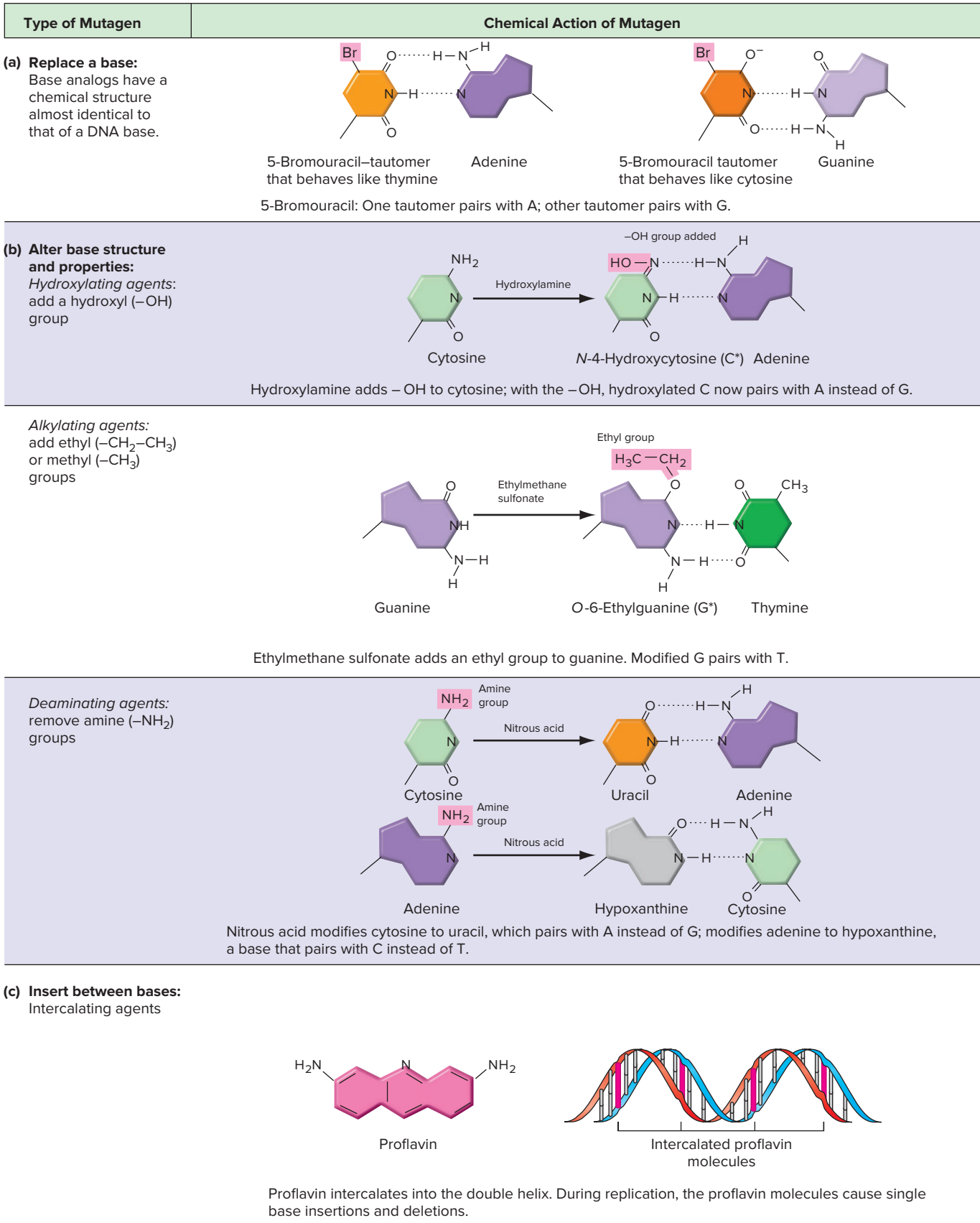
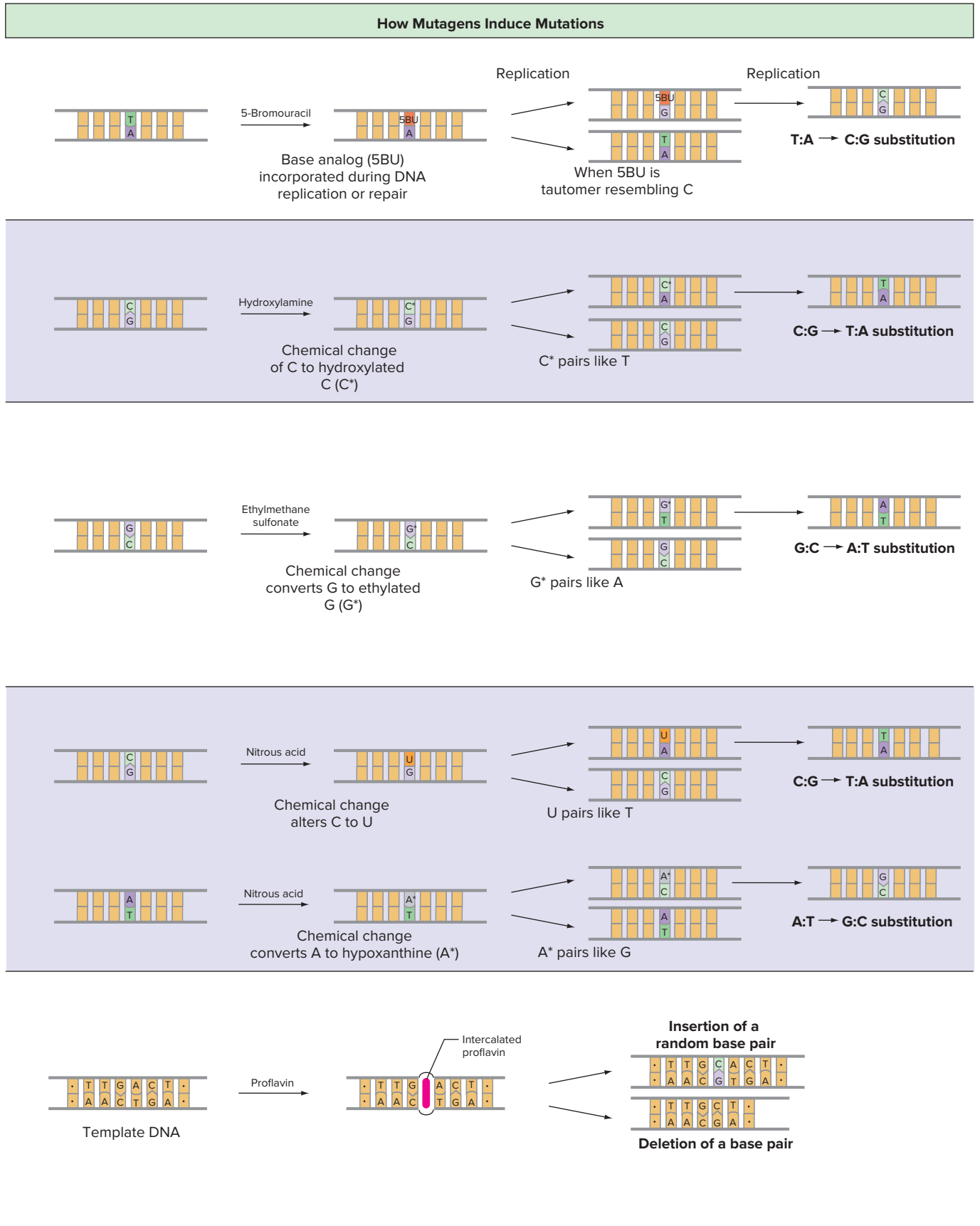
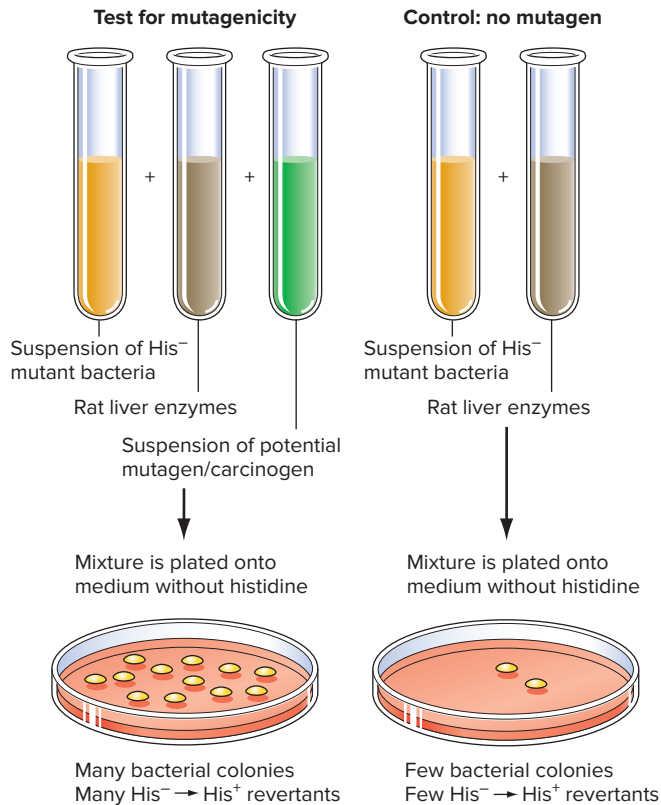


Figure 7.14 How mutagens alter DNA. (Continued)



**Figure 7.15 The Ames test identifies potential carcinogens.** Investigators mix a compound to be tested with cells of a His<sup>-</sup> strain of *Salmonella typhimurium* and with a solution of rat liver enzymes (which can sometimes convert a harmless compound into a mutagen). Only His<sup>+</sup> revertants grow on a petri plate without histidine. If this plate (*bottom left*) has more His<sup>+</sup> revertants than a control plate (also without histidine) containing unexposed cells (*bottom right*), the compound is considered mutagenic and a potential carcinogen. The rare revertants on the control plate represent the rate of spontaneous mutation.



The U.S. Food and Drug Administration tries to identify potential carcinogens by using the **Ames test** to screen for chemicals that cause mutations in bacterial cells (Fig. 7.15). This test asks whether a particular chemical can induce Histidine<sup>+</sup> (His<sup>+</sup>) revertants of a special Histidine<sup>-</sup> (His<sup>-</sup>) mutant strain of the bacterium *Salmonella typhimurium*. The His<sup>+</sup> revertants can synthesize all the histidine they need from simple compounds in their environment, whereas the original His<sup>-</sup> mutants cannot make histidine, so they can survive only if histidine is supplied.

The advantage of the Ames test is that only revertants can grow on petri plates that do not contain histidine, so it is possible to examine large numbers of cells from an originally His<sup>-</sup> culture to find the rare His<sup>+</sup> revertants induced by the chemical in question. To increase the sensitivity of mutation detection, the His<sup>-</sup> strain used in the Ames test system contains a second mutation that inactivates a DNA repair system (to be described in the next section) and

thereby prevents the ready repair of mutations caused by the potential mutagen. The bacteria also carry a third mutation causing defects in the cell wall that allows tested chemicals easier access to the cell interior.

Because most agents that cause mutations in bacteria should also damage the DNA of higher eukaryotic organisms, any mutagen that increases the rate of mutation in bacteria might be expected to cause cancer in people and other mammals. Mammals, however, have complicated metabolic processes capable of inactivating hazardous chemicals. On the other hand, other biochemical events in mammals can create a mutagenic substance from nonhazardous chemicals. To simulate the action of mammalian metabolism, toxicologists often add a solution of rat liver enzymes to the chemical under analysis by the Ames test (Fig. 7.15). Because this simulation is not perfect, Food and Drug Administration agents ultimately assess whether bacterial mutagens identified by the Ames test can cause cancer in rodents by including the agents in test animals' diets.

### essential concepts

- Certain natural agents can induce *spontaneous mutations*. These agents include radiations (such as X-rays and UV light) and chemical reactions (such as deamination and oxidation) that damage DNA.
- DNA replication errors are another source of spontaneous mutations. Many of these errors result from base tautomerization or expansions/contractions of trinucleotide repeats.
- *Mutagens* are agents that raise the mutation frequency above the spontaneous rate. In research, mutagens can help generate mutations of interest for further study.

## 7.3 DNA Repair Mechanisms

### learning objectives

1. List mechanisms by which cells can repair DNA with altered or damaged nucleotides.
2. Contrast the outcomes of homologous recombination and nonhomologous end-joining mechanisms for the repair of double-strand breaks.
3. Explain how methyl-directed mismatch repair can distinguish which strand to repair when replication errors occur.
4. State why cells use certain DNA repair systems only as a last resort.
5. Describe the potential consequences for human health of mutations in genes that specify DNA repair factors.

Recall from Fig. 7.7 that if new DNA damage is repaired before DNA replication occurs, no mutation becomes established in the chromosomes. Cells have in fact evolved a variety of enzymatic systems that locate and repair damaged DNA and thereby dramatically minimize the occurrence of mutations. The combination of these repair systems must be extremely efficient, because the rates of spontaneous mutation observed for almost all genes are very low.

### Some DNA Base Damage Can Be Reversed

Cells contain various enzyme systems that can reverse certain kinds of nucleotide alterations quickly and directly. For example, if methyl or ethyl groups are mistakenly added to guanine (as in Fig. 7.14b), *alkyltransferase* enzymes can remove them so as to re-create the original base.

In a second example, the enzyme *photolyase* recognizes the thymine dimers produced by exposure to ultraviolet light (review Fig. 7.8d) and reverses the damage by splitting the chemical linkage between the thymines. Interestingly, the photolyase enzyme works only in the presence of visible light. In carrying out its DNA repair tasks, it associates with a small molecule called a *chromophore* that absorbs light in the visible range of the spectrum; the enzyme then uses the energy captured by the chromophore to split thymine dimers. Because it does not function in the dark, the photolyase mechanism is called *light repair*, or *photorepair*.

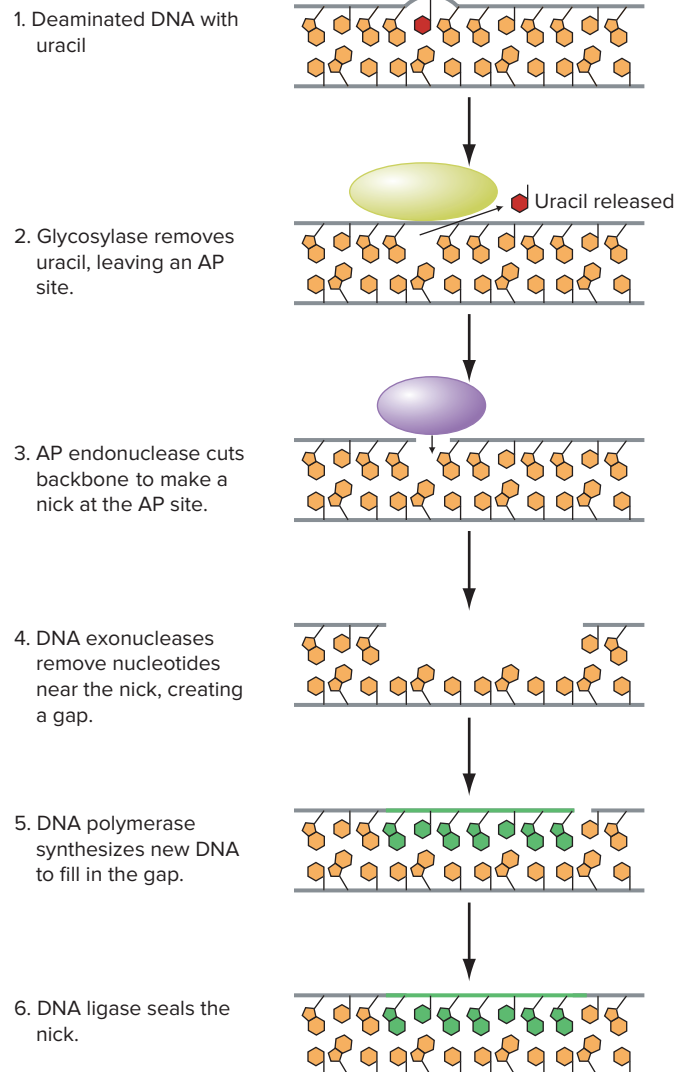
### Damaged Bases Can Be Removed and Replaced

Many repair systems use a general strategy of homology-dependent repair in which they first remove a small region from the DNA strand that contains the altered nucleotide, and then use the other strand as a template to resynthesize the region removed. This strategy makes use of one of the great advantages of the double-helical structure: If one strand sustains damage, cells can use complementary base pairing with the undamaged strand to re-create the original sequence.

#### Base excision repair

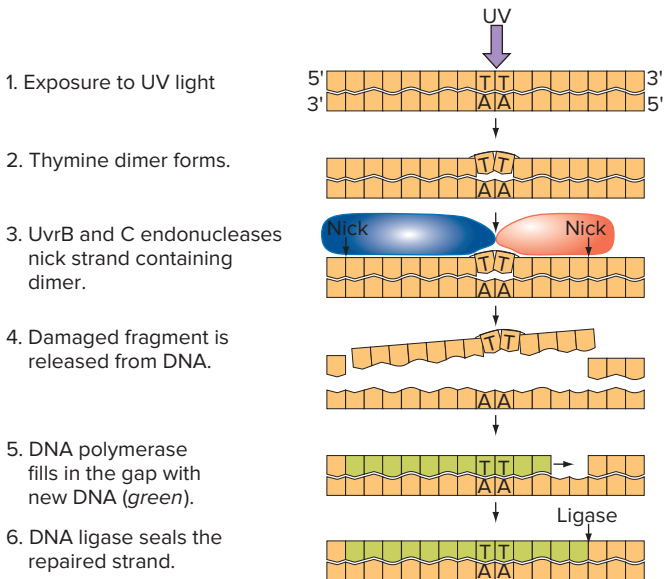
In this type of homology-dependent repair mechanism, enzymes called *DNA glycosylases* cleave an altered nitrogenous base from the sugar of its nucleotide, releasing the base and creating an apurinic or apyrimidinic (AP) site in the DNA chain (Fig. 7.16). Different glycosylase enzymes cleave specific damaged bases. Base excision repair is particularly important in the removal of uracil from DNA

**Figure 7.16** Base excision repair removes damaged bases. Glycosylase enzymes (light green oval) remove aberrant bases [like uracil (red) formed by the deamination of cytosine], leaving an AP site. AP endonuclease (purple oval) cuts the sugar-phosphate backbone, creating a nick. Exonucleases extend the nick into a gap, which is filled in with the correct information (dark green) by DNA polymerase. DNA ligase reseals the corrected strand.



(recall that uracil often results from the natural deamination of cytosine; review Fig. 7.8b). In this repair process, after the enzyme uracil-DNA glycosylase has removed uracil from its sugar, leaving an AP site, the enzyme AP endonuclease makes a nick in the DNA backbone at the AP site. Other enzymes (known as *DNA exonucleases*) attack the nick and remove nucleotides from their vicinity to create a gap in the previously damaged strand. DNA polymerase fills in the gap by copying the undamaged strand, restoring the original nucleotide in the process. Finally, DNA ligase seals up the backbone of the newly repaired DNA strand.

**Figure 7.17 Nucleotide excision repair corrects damaged nucleotides.** A complex of the UvrA and UvrB proteins (*not shown*) scans DNA for distortions caused by DNA damage, such as thymine dimers. At the damaged site, UvrA dissociates from UvrB, allowing UvrB (*red*) to associate with UvrC (*blue*). These enzymes nick the DNA exactly four nucleotides to one side of the damage and seven nucleotides to the other side, releasing a small fragment of single-stranded DNA. DNA polymerases then resynthesize the missing information (*green*), and DNA ligase reseals the now-corrected strand.



### Nucleotide excision repair

This pathway (**Fig. 7.17**) removes alterations that base excision cannot repair because the cell lacks a DNA glycosylase that recognizes the problem base(s). Nucleotide excision repair depends on enzyme complexes containing more than one protein molecule. In *E. coli*, these complexes are made of two out of three possible proteins: UvrA, UvrB, and UvrC. One of the complexes (UvrA + UvrB) patrols the DNA for irregularities, detecting lesions that disrupt Watson-Crick base pairing and thus distort the double helix (such as thymine dimers that have not been corrected by photorepair). A second complex (UvrB + UvrC) cuts the damaged strand in two places that flank the damage. This double-cutting excises a short region of the damaged strand and leaves a gap that will be filled in by DNA polymerase and sealed with DNA ligase.

## Two Important Mechanisms Can Repair Double-Strand Breaks

We have seen previously that X-rays can cause double-strand breaks, in which both strands of the double helix are broken at nearby sites (review Fig. 7.8c). Double-strand breaks represent a particularly dangerous kind of DNA lesion because if not repaired properly, such

chromosomal breakages can lead not only to point mutations, but also to large deletions and other kinds of chromosomal rearrangements.

It is therefore not surprising that organisms have evolved at least two different ways of repairing double-strand breaks. One of these mechanisms, **homologous recombination (HR)**, uses complementary base pairing to repair breaks accurately with no loss or gain of nucleotides. The second pathway, called **nonhomologous end-joining (NHEJ)**, can bring together even DNA ends that were not previously adjacent to each other, and a few base pairs can be lost or added improperly in the process.

Both systems for the repair of double-strand breaks have great practical significance because they are fundamental to new, effective strategies for *genome editing* (altering an organism's genome in specific ways). We will describe these exciting methods for modifying genomes in Chapter 18, but it will be helpful for you to gain here some idea about how these repair mechanisms work.

### Double-strand break repair via homologous recombination (HR)

You will recall that the first step of meiotic recombination is the formation of a double-strand break, and that through strand invasion, cells undergoing recombination eventually repair this double-strand break using the homologous chromosome as a template (review Fig. 6.27). Mitotic cells can employ much of the same enzymatic machinery for homologous recombination to repair double-strand breaks caused by X-ray exposure.

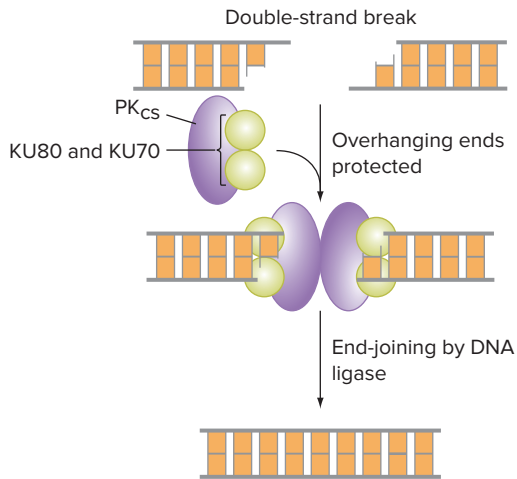
The HR system can use either a homologous chromosome, or more often a sister chromatid, as the template for repair. If the homologous chromosome serves as the template, repair of the break results in mitotic recombination. However, finding a homolog is inefficient, so repair through recombination usually occurs instead between sister chromatids during the  $G_2$  phase of the cell cycle (that is, after the chromosomes have replicated). In this case, repair of the break does not produce mitotic recombination because the broken chromatid and the template chromatid base pair sequences are identical.

### Repair of double-strand breaks by nonhomologous end-joining (NHEJ)

The NHEJ mechanism is an alternative to HR that is especially important for the repair of double-strand breaks formed during the  $G_1$  phase of the cell cycle (that is, before a sister chromatid is available to serve as a template for homologous recombination). The proteins participating in NHEJ bind to DNA ends at the site of the breakage and protect the ends from nucleases. The NHEJ proteins also bridge the two ends, allowing them to be stitched together by the DNA ligase enzyme (**Fig. 7.18**).



**Figure 7.18** Repair of double-strand breaks by nonhomologous end-joining. The proteins KU70, KU80, and PKcs (in mammalian cells) bind to DNA ends, protect them from degradation, and bring them together so that DNA ligase can repair the phosphodiester backbone.



You should note that because NHEJ does not involve DNA homology, it can join together any DNA ends (other than telomeres, which are protected against this pathway), even if those ends were not adjacent to each other in the genome originally. Thus, if the genome suffers more than one double-strand break, NHEJ can potentially join the wrong ends together, causing chromosome rearrangements such as inversions or large deletions.

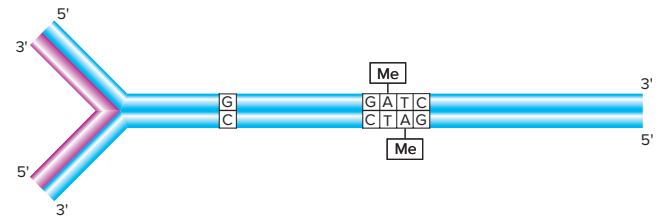
Another property of NHEJ is that although this mechanism is usually accurate, it can sometimes result in small changes to the DNA sequences where the broken ends are joined together. During the NHEJ process, DNA exonucleases and DNA polymerases can act at the broken ends, removing or adding a few base pairs to them before DNA ligase seals them together. Errors due to NHEJ are relatively infrequent, but they do occur and will become of considerable significance when we discuss genome editing techniques in Chapter 18 later in the book.

## Mismatch Repair Corrects Errors in DNA Replication

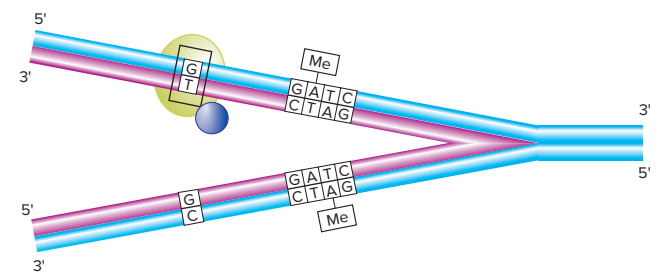
DNA polymerase is remarkably accurate in copying DNA, but the DNA replication system still makes about 100 times more mistakes than most cells can tolerate. A backup repair system called **methyl-directed mismatch repair** corrects almost all of these errors (Fig. 7.19). Because mismatch repair is active only after DNA replication, this system needs to solve a difficult problem. Suppose that a G–C pair has been copied to produce two daughter molecules, one of which has the correct G–C base pair, the other an incorrect G–T. The mismatch repair system can easily recognize the incorrectly matched G–T base pair because the improper

**Figure 7.19** In bacteria, methyl-directed mismatch repair corrects mistakes in replication. Parental strands are in light blue and newly synthesized strands are magenta. The MutS protein is green, MutL is dark blue, and MutH is orange. See text for details.

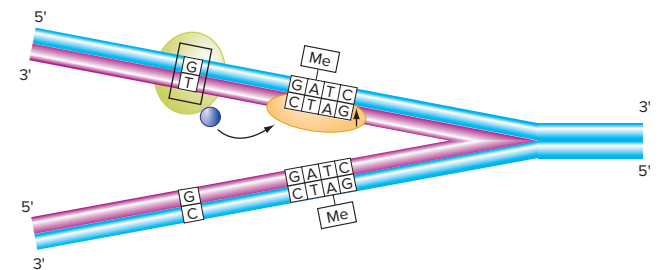
(a) Parental strands are marked with methyl groups.



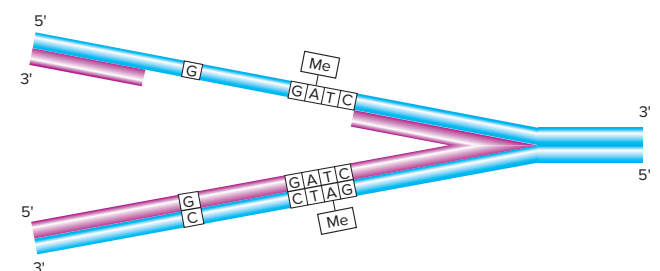
(b) MutS and MutL recognize mismatch in replicated DNA.



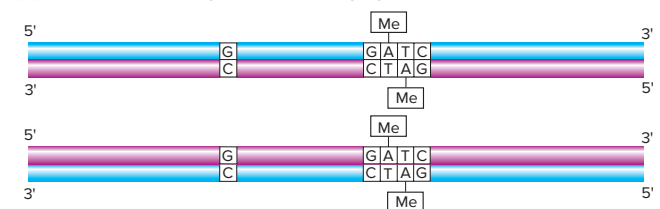
(c) MutL recruits MutH to GATC; MutH makes a nick (short arrow) in strand opposite methyl tag.



(d) DNA exonucleases (not shown) excise DNA from unmethylated new strand.



(e) Repair and methylation of newly synthesized DNA strand.



base pairing distorts the double helix, resulting in abnormal bulges and hollows. But how does the system know whether to correct the pair to a G–C or to an A–T?

Bacteria solve this problem by placing a distinguishing mark on the parental DNA strands at specific places: Everywhere the sequence GATC occurs, the enzyme adenine methylase puts a methyl group on the A (Fig. 7.19a). Shortly after replication, the old template strand bears the methyl mark, while the new daughter strand—which contains the wrong nucleotide—is as yet unmarked (Fig. 7.19b). A pair of proteins in *E. coli*, called MutL and MutS, detect and bind to the mismatched nucleotides. MutL and MutS direct another protein, MutH, to nick the newly synthesized strand of DNA at a position across from the nearest methylated GATC; MutH can discriminate the newly synthesized strand because its GATC is not methylated (Fig. 7.19c). DNA exonucleases then remove all the nucleotides between the nick and a position just beyond the mismatch, leaving a gap on the new, unmethylated strand (Fig. 7.19d). DNA polymerase can now resynthesize the information using the old, methylated strand as a template, and DNA ligase then seals up the repaired strand. With the completion of replication and repair, enzymes mark the new strand with methyl groups so that its parental origin will be evident in the next round of replication (Fig. 7.19e).

Eukaryotic cells also have a mismatch correction system, but we do not yet know how this system distinguishes templates from newly replicated strands. Unlike prokaryotes, GATCs in eukaryotes are not tagged with methyl groups, and eukaryotes do not seem to have a protein closely related to MutH. One potentially interesting clue is that the MutS and MutL proteins in eukaryotes associate with DNA replication factors; perhaps these interactions might help MutS and MutL identify the strand to be repaired.

## Error-Prone Repair Systems Serve as Last Resorts

The repair systems just described are very accurate in repairing DNA damage because they can either replace damaged nucleotides with a complementary copy of the undamaged strand or ligate breaks back together. However, cells sometimes become exposed to levels or types of mutagens that they cannot handle with these high-fidelity repair systems. Strong doses of UV light, for example, might make more thymine dimers than the cell can mend. Any unrepaired damage has severe consequences for cell division; in particular, the DNA polymerases normally used in replication will stall at such lesions, so the cells cannot proliferate. These cells can initiate emergency responses that may allow them to overcome these problems and thus survive and divide, but their ability to proceed in such circumstances comes at the expense of introducing new mutations into the genome.

## Using sloppy DNA polymerases

One type of emergency repair in bacteria, called the **SOS system** (after the Morse code distress signal), relies on error-prone (or *sloppy*) DNA polymerases. These sloppy DNA polymerases are not available for normal DNA replication; they are produced only in the presence of DNA damage. The damage-induced, error-prone DNA polymerases are attracted to replication forks that have become stalled at sites of unrepaired, damaged nucleotides. There the enzymes add random nucleotides to the strand being synthesized opposite the damaged bases.

The SOS polymerase enzymes thus allow the cell with damaged DNA to divide into two daughter cells, but because at each position the sloppy polymerases restore the proper nucleotide only one-quarter of the time, the genomes of these daughter cells carry new mutations. In bacteria, the mutagenic effect of many mutagens either depends on, or is enhanced by, the SOS system.

## Sloppy repair of double-strand breaks

Another kind of emergency repair system, **microhomology-mediated end-joining (MMEJ)**, deals with dangerous double-strand breaks that have not been corrected by homologous recombination or NHEJ. The mechanism of MMEJ is similar to that of NHEJ (previously shown in Fig. 7.18), except in MMEJ the broken DNA ends are cut back on either side of the break (resected) by enzymes. The resection exposes small single-stranded regions of complementary DNA sequence (*microhomology*) on either side of the break that help in bringing the ends together.

Because nucleotides are removed at the sites of the double-stranded breaks during resection, MMEJ results in deletions of tens to hundreds of base pairs in the rejoined DNA. These deletions are longer than the small deletions of a few base pairs that sometimes result from NHEJ.

## Mutations in Genes Encoding DNA Repair Proteins Impact Human Health

DNA repair mechanisms appear in some form in virtually all species. For example, humans have six proteins whose amino acids are about 25% identical with those of the *E. coli* mismatch repair protein MutS. DNA repair systems are thus very old and must have evolved soon after life emerged ~3.5 billion years ago. Some scientists think DNA repair became essential when plants started to deposit oxygen into the atmosphere, because oxygen helps form free radicals that can damage DNA.

The many known human hereditary diseases associated with the defective repair of DNA damage reveal how crucial these mechanisms are for survival. In one example, the cells of patients with *xeroderma pigmentosum* lack the ability to conduct nucleotide excision repair; these people

**Figure 7.20** Skin lesions in a xeroderma pigmentosum patient. This heritable disease is caused by the lack of a critical enzyme in the nucleotide excision repair system.

© Barcroft Media/Getty Images



are homozygous for mutations in any one of seven genes encoding enzymes that normally function in this repair system. As a result, the thymine dimers caused by ultraviolet light cannot be removed efficiently. Unless these people avoid all exposure to sunlight, their skin cells begin to accumulate mutations that eventually lead to skin cancer (Fig. 7.20).

In another example, researchers have recently learned that hereditary forms of colorectal cancer in humans are associated with mutations in human genes that are closely related to the *E. coli* genes encoding the mismatch-repair proteins MutS and MutL. In yet another example, the breast cancer genes *BRCA1* and *BRCA2* (mutation of either of which is associated with a high risk of breast cancer in women) encode proteins that function in double-strand break repair via homologous recombination. Chapter 20 discusses the fascinating connections between DNA repair and cancer in more detail.

### DNA Repair Cannot Be 100% Efficient

“The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music.” In these two sentences, the eminent medical scientist and

self-appointed “biology watcher” Lewis Thomas acknowledges that changes in DNA are behind the phenotypic variations that are the raw material on which natural selection has acted for billions of years to drive evolution.

As Dr. Thomas’ poetic line suggests, the necessity for mutation is fundamental: Without mutations, life would have died out long ago because it could not have responded to changes in the environment. DNA repair processes must therefore walk a fine line. They must be efficient enough to protect genomes from the huge number of assaults on DNA that are always occurring, but the propagation of life requires some mutations to be transmitted to future generations.

#### essential concepts

- Cells have many different enzyme systems that minimize mutations by repairing DNA damage or replication errors.
- Double-strand breaks, which are particularly dangerous to genomes, can be repaired through *homologous recombination (HR)* or *nonhomologous end-joining (NHEJ)*.
- Correction of DNA replication errors requires *mismatch repair* systems to choose the correct strand to change. Bacteria accomplish this task by marking parental strands with methyl groups.
- If normal repair mechanisms are overwhelmed by too much DNA damage, cells can then mobilize *error-prone DNA repair* systems.
- Mutations in genes specifying proteins that participate in DNA repair often lead to human diseases, including cancer.
- Mutations are the raw material of evolution. Although many mutations are harmful, rare mutations may confer a selective advantage.

## 7.4 What Mutations Tell Us About Gene Structure

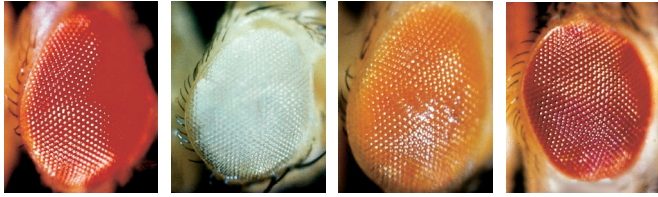
#### learning objectives

1. Describe complementation testing and how its results distinguish mutations in a single gene from mutations in different genes.
2. Explain how Benzer’s experimental results revealed that the *rII* region in bacteriophage T4 contains two genes, each composed of many nucleotide pairs.
3. Discuss how Benzer used deletions to map mutations in the *rII* region.

The science of genetics depends absolutely on mutations, because we can track genes in crosses only through the phenotypic effects of their mutant variants. In the 1950s

**Figure 7.21** *Drosophila* eye color mutations produce a variety of phenotypes. Flies carrying different X-linked eye color mutations. From the left: ruby, white, and apricot; a wild-type eye is at the far right.

(all): © Science Source



and 1960s, scientists realized they could also use mutations to learn how DNA sequences along a chromosome constitute individual genes. These investigators wanted to collect a large series of mutations in a single gene and analyze how these mutations were arranged with respect to each other. For this approach to be successful, they had to establish that various mutations were, in fact, in the same gene. This was not a trivial exercise, as illustrated by the following situation.

Early *Drosophila* geneticists identified a large number of X-linked recessive mutations affecting the normally red wild-type eye color (Fig. 7.21). The first of these to be discovered produced the famous white eyes studied by Morgan's group. Other mutations caused a whole palette of hues to appear in the eyes: darkened shades such as garnet and ruby; bright colors such as vermilion, cherry, and coral; and lighter pigmentations known as apricot, buff, and carnation. This wide variety of eye colors posed a puzzle: Were the mutations that caused them multiple alleles of a single gene, or did they affect more than one gene?

### Complementation Testing Reveals Whether Two Mutations Are in a Single Gene or in Different Genes

Researchers commonly define a gene as a functional unit that directs the appearance of a molecular product that, in turn, contributes to a particular phenotype. They can use this definition to determine whether two mutations are in the same gene or in different genes.

If two homologous chromosomes in an individual each carry a mutation recessive to wild type, that individual will have a normal phenotype if the mutations are in different genes. Such a result is called **complementation**. The normal phenotype occurs because almost all recessive mutations disrupt a gene's function. The dominant wild-type alleles on each of the two homologs can make up for, or *complement*, the defect in the other chromosome by generating enough of both gene products to yield a normal phenotype (Fig. 7.22a, left).

In contrast, if the recessive mutations on the two homologous chromosomes are in the same gene, no wild-type allele of that gene exists in the individual, and neither

mutated copy of the gene will be able to perform the normal function. As a result, no complementation will occur and no normal gene product will be made, so a mutant phenotype will appear (Fig. 7.22a, right). Ironically, a collection of mutations that do *not* complement each other is known as a **complementation group**. Geneticists often use *complementation group* as a synonym for *gene* because the mutations in a complementation group all affect the same unit of function, and thus, the same gene.

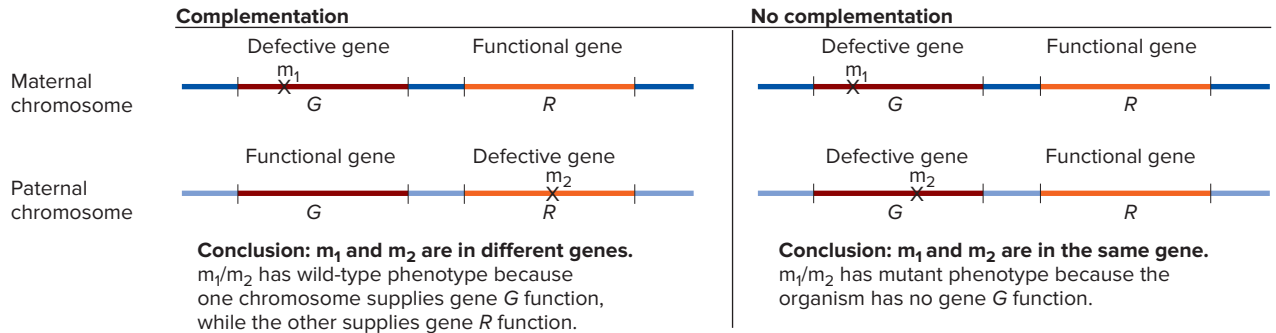
A simple test based on the idea of a gene as a unit of function can determine whether or not two recessive mutations are alleles of the same gene. You simply examine the phenotype of a heterozygous individual in which one homolog of a particular chromosome carries one of the mutations and the other homolog carries the other mutation. If the phenotype is wild-type, the mutations cannot be in the same gene. This technique is known as a **complementation test**. For example, because a female fruit fly simultaneously heterozygous for *garnet* and *ruby* (*garnet ruby*<sup>+</sup>/*garnet*<sup>+</sup> *ruby*) has wild-type brick-red eyes, it is possible to conclude that the mutations causing garnet and ruby colors complement each other and are therefore in different genes.

Complementation testing has, in fact, shown that garnet, ruby, vermilion, and carnation pigmentation are caused by mutations in separate genes. But chromosomes carrying mutations yielding white, cherry, coral, apricot, and buff phenotypes fail to complement each other. These mutations therefore constitute different alleles of a single gene. *Drosophila* geneticists named this gene the *white*, or *w*, gene after the first mutation observed; they designate the wild-type allele as *w*<sup>+</sup> and the various mutations as *w*<sup>1</sup> (the original white-eyed mutation discovered by T. H. Morgan, often simply designated as *w*), *w*<sup>cherry</sup>, *w*<sup>coral</sup>, *w*<sup>apricot</sup>, and *w*<sup>buff</sup>. As an example, the eyes of a *w*<sup>1</sup> / *w*<sup>apricot</sup> female are a dilute apricot color; because the phenotype of this heterozygote is not wild-type, the two mutations are allelic. Figure 7.22b illustrates how researchers collate data from many complementation tests in a **complementation table**. Such a table helps visualize the relationships among a large group of mutants.

In *Drosophila*, mutations in the *w* gene map very close together in the same region of the X chromosome, while mutations in other sex-linked eye color genes lie elsewhere on the chromosome (Fig. 7.22c). This result suggests that genes are not disjointed entities with parts spread out from one end of a chromosome to another; each gene, in fact, occupies only a relatively small, discrete area of a chromosome. Studies defining genes at the molecular level have shown that most genes consist of 1000–20,000 contiguous base pairs (bp). In humans, among the shortest genes are the roughly 500 base pair–long genes that govern the production of histone proteins, while the longest gene so far identified is the *Duchenne muscular dystrophy* (*DMD*) gene, which has a length of more than 2 million nucleotide

**Figure 7.22 Complementation testing of *Drosophila* eye color mutations. (a)** A heterozygote has one mutation ( $m_1$ ) on one chromosome and a different mutation ( $m_2$ ) on its homolog. If the mutations are in different genes, the heterozygote will be wild type; the mutations complement each other (*left*). If both mutations affect the same gene, the phenotype will be mutant; the mutations do not complement each other (*right*). Complementation testing makes sense only when both mutations are recessive to wild type. **(b)** This complementation table reveals five complementation groups (five different genes) for eye color. A *plus* (+) indicates mutant combinations with wild-type eye color; these mutations complement and are thus in different genes. Several mutations fail to complement (-) and are thus alleles of one gene, *white*. **(c)** Recombination mapping shows that mutations in different genes are often far apart, while different mutations in the same gene are very close together.

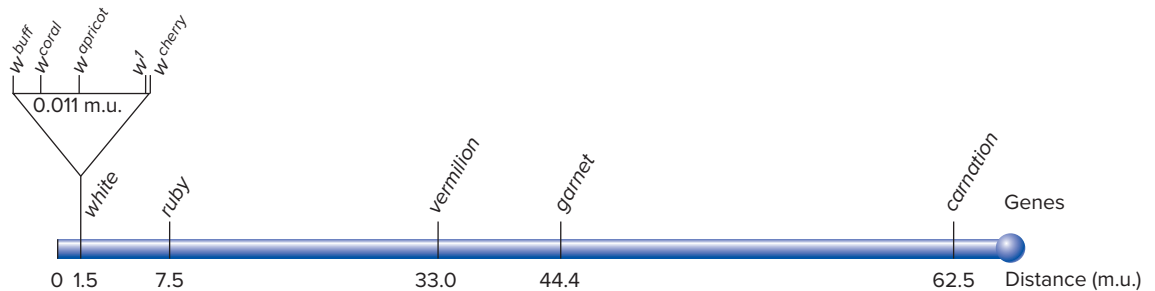
**(a) Complementation testing**



**(b) A complementation table: X-linked eye color mutations in *Drosophila***

Mutation	<i>white</i>	<i>garnet</i>	<i>ruby</i>	<i>vermillion</i>	<i>cherry</i>	<i>coral</i>	<i>apricot</i>	<i>buff</i>	<i>carnation</i>
<i>white</i>	-	+	+	+	-	-	-	-	+
<i>garnet</i>		-	+	+	+	+	+	+	+
<i>ruby</i>			-	+	+	+	+	+	+
<i>vermillion</i>				-	+	+	+	+	+
<i>cherry</i>					-	-	-	-	+
<i>coral</i>						-	-	-	+
<i>apricot</i>							-	-	+
<i>buff</i>								-	+
<i>carnation</i>									-

**(c) Genetic map: X-linked eye color mutations in *Drosophila***



pairs. All known human genes fall somewhere between these extremes. To put these figures in perspective, an average human chromosome is approximately 130 million base pairs in length.

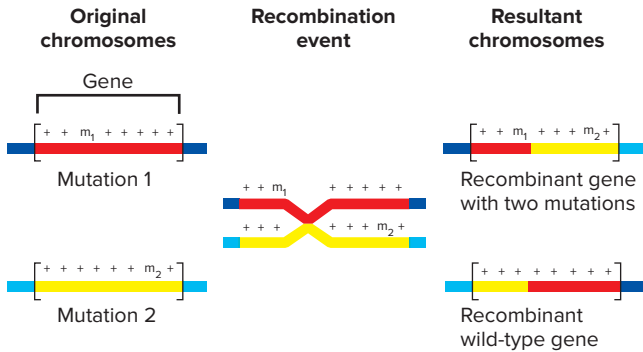
**A Gene Is a Set of Nucleotide Pairs That Can Mutate Independently and Recombine with Each Other**

Although complementation testing makes it possible to distinguish mutations in different genes from mutations in the same gene, it does not clarify how the structure of a

gene can accommodate different mutations and how these different mutations can alter phenotype in different ways. Does each mutation change the whole gene at a single stroke in a particular way, or does it change only a specific part of a gene, while other mutations alter other parts?

In the late 1950s, the American geneticist Seymour Benzer used recombination analysis to show that two different mutations that did not complement each other and were therefore known to be in the same gene can in fact change different parts of that gene. He reasoned that if a gene is composed of separately mutable subunits, then it should be possible for recombination to occur within a gene, between these subunits. Therefore, crossovers

**Figure 7.23** How recombination within a gene could generate a wild-type allele. Suppose a gene, indicated by the region between brackets, is composed of many sites that can mutate independently. Recombination between mutations  $m_1$  and  $m_2$  at different sites in the same gene produces a wild-type allele and a reciprocal allele containing both mutations.



between homologous chromosomes carrying different mutations known to be in the same gene could in theory generate a wild-type allele (Fig. 7.23).

Because mutations affecting a single gene are likely to lie very close together, it is necessary to examine a very large number of progeny to observe even one crossover event between them. The resolution of the experimental system must thus be extremely high, allowing rapid detection of rare genetic events. For his experimental organism, Benzer chose bacteriophage T4, a virus that infects *E. coli* cells (Fig. 7.24a.1). Because each T4 phage that infects a bacterium generates 100 to 1000 progeny in less than an hour, Benzer could easily produce enough rare recombinants for his analysis (Fig. 7.24a.2). Moreover, by exploiting a peculiarity of certain T4 mutations, he devised conditions that allowed only recombinant phages, and not parental phages, to proliferate.

### The experimental system: $rII^-$ mutations of bacteriophage T4

Even though bacteriophages are too small to be seen without the aid of an electron microscope, a simple technique makes it possible to detect their presence with the unaided eye (Fig. 7.24a.3). To do this, researchers mix a population of bacteriophage particles with a much larger number of bacteria in molten agar and then pour this mixture onto a petri plate that already contains a bottom layer of nutrient agar. Uninfected bacterial cells grow throughout the top layer, forming an opalescent lawn of living bacteria. However, if a single phage infects a single bacterial cell somewhere on this lawn, the cell produces and releases progeny viral particles that infect adjacent bacteria, which, in turn, produce and release yet more phage progeny. With each release of virus particles, the bacterial host cell dies. The agar in the top layer prevents the phage particles from diffusing very far. Thus, several cycles of

phage infection, replication, and release produce a circular cleared area in the lawn, called a **plaque**, devoid of living bacterial cells. The process of mixing phages with bacteria to produce a lawn and plaques on a petri plate is called *plating* phages.

Most plaques contain from 1 million to 10 million descendants of the single bacteriophage that originally infected a cell in that position on the petri plate. Sequential dilution of phage-containing solutions makes it possible to measure the number of phages in a particular plaque and arrive at a countable number of viral particles (Fig. 7.24a.4).

When Benzer first looked for genetic traits associated with bacteriophage T4, he found mutants that, when added to a lawn of *E. coli* B strain bacteria, produced larger plaques with sharper, more clearly rounded edges than those produced by the wild-type bacteriophages (Fig. 7.24b). Because these changes in plaque morphology result from the abnormally rapid lysis of the host bacteria, Benzer named the mutations *r* for *rapid lysis*. Many *r* mutations map to a region of the T4 chromosome known as the *rII* region; these are called  $rII^-$  mutations.

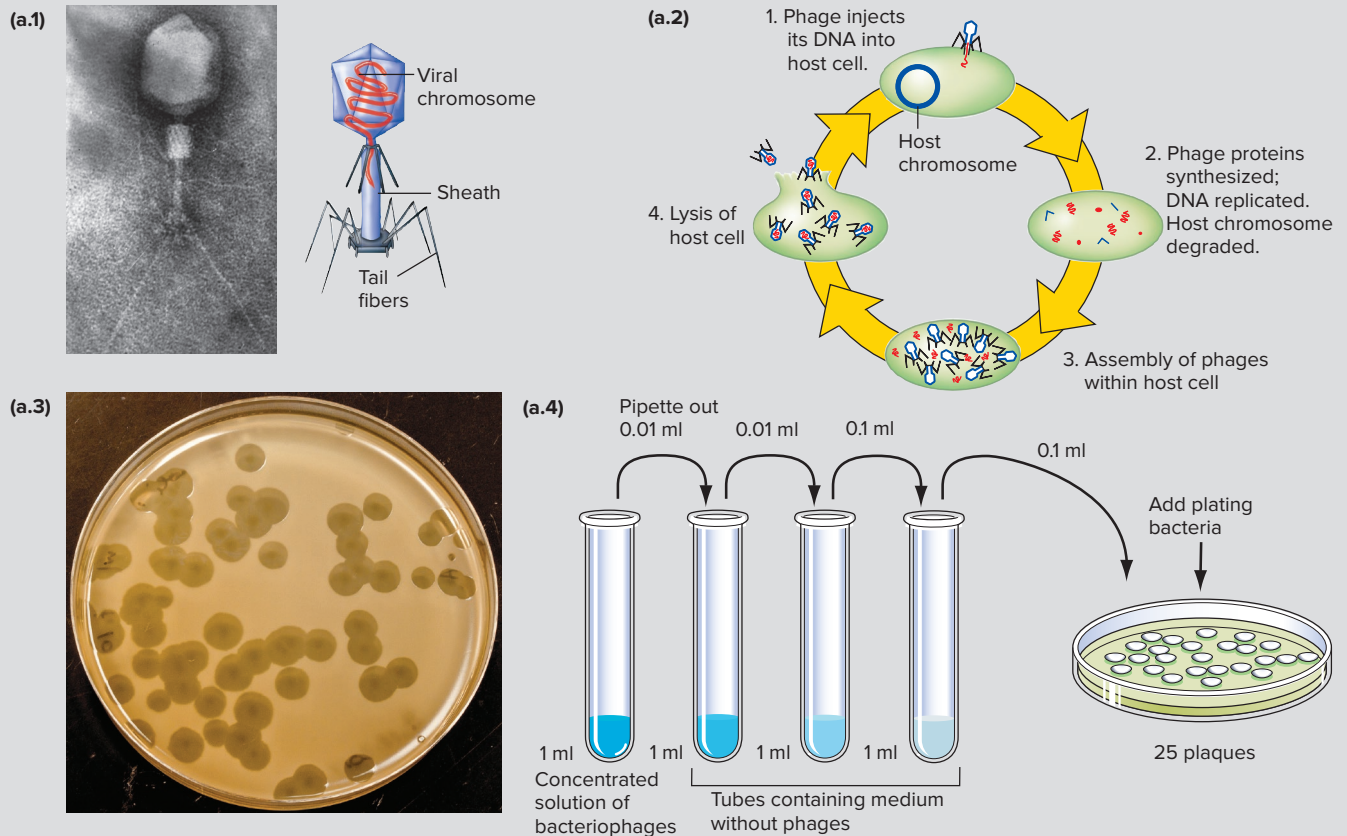
An additional property of  $rII^-$  mutations makes them ideal for the genetic **fine structure mapping** (the mapping of mutations within a gene) undertaken by Benzer. Wild-type  $rII^+$  bacteriophages form plaques of normal shape and size on cells of both the *E. coli* B strain and a strain known as *E. coli* K( $\lambda$ ). The  $rII^-$  mutants, however, have an altered host range; they cannot form plaques with *E. coli* K( $\lambda$ ) cells, although as we have seen, they produce large, unusually distinct plaques with *E. coli* B cells (Fig. 7.24b). The reason that  $rII^-$  mutants are unable to infect cells of the K( $\lambda$ ) strain was not clear to Benzer, but this property allowed him to develop an extremely simple and effective test for  $rII^+$  gene function, as well as an ingenious way to detect rare *intragenic* (within the same gene) recombination events.

### The *rII* region has two genes

Before he could check whether two mutations in the same gene could recombine, Benzer had to be sure he was really looking at two mutations in a single gene. To verify this, he performed customized complementation tests tailored to two significant characteristics of bacteriophage T4: They are **monoploid** (that is, each phage carries a single T4 chromosome, so the phages have one copy of each of their genes), and they can replicate only in a host bacterium. Because T4 phages are monoploid, Benzer needed to ensure that two different T4 chromosomes entered the same bacterial cell in order to test for complementation between the mutations. In his complementation tests, he simultaneously infected *E. coli* K( $\lambda$ ) cells with two types of T4 chromosomes—one carried one  $rII^-$  mutation, the other carried a different  $rII^-$  mutation—and then looked for cell lysis

## FEATURE FIGURE 7.24

### How Benzer Analyzed the *rII* Genes of Bacteriophage T4



a.1: © Science Source; a.3: © McGraw-Hill Education/Lisa Burgess

#### (a) Working with bacteriophage T4

1. Bacteriophage T4 (at a magnification of approximately 100,000 $\times$ ) and in an artist's rendering. The viral chromosome is contained within a protein head. Other proteinaceous parts of the phage particle include the tail fibers, which help the phage attach to host cells, and the sheath, a conduit for injecting the phage chromosome into the host cell.

2. The lytic cycle of bacteriophage T4. A single phage particle infects a host cell; the phage DNA replicates and directs the synthesis of viral protein components using the machinery of the host cell; the new DNA and protein components assemble into new bacteriophage particles. Eventual lysis of the host cell releases up to 1000 progeny bacteriophages into the environment.

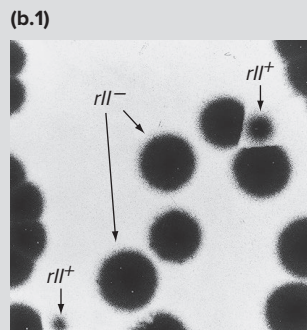
3. Clear plaques of bacteriophages in a lawn of bacterial cells. A mixture of bacteriophages and a large number of bacteria in molten agar are poured into a petri plate. Uninfected bacterial cells grow, producing an opalescent *lawn*. A bacterium infected by a single bacteriophage will lyse and release progeny bacteriophages, which infect adjacent bacteria. Several cycles of infection result in a *plaque*, a circular cleared area containing millions of genetically identical bacteriophages.

4. Counting bacteriophages by *serial dilution*. A small sample of a concentrated solution of bacteriophages is transferred to a test tube containing fresh medium, and a small sample of this dilution is transferred to another tube of fresh medium. Successive repeats of this process increase the degree of dilution. A sample of the final dilution, when mixed with bacteria in molten agar, yields a countable number of plaques from which it is possible to extrapolate the number of bacteriophages in the starting solution. The original 1 ml of solution in this illustration contained roughly  $2.5 \times 10^7$  bacteriophages.

#### (b) Phenotypic properties of *rII*<sup>-</sup> mutants of bacteriophage T4

1. *rII*<sup>-</sup> mutants, when plated on *E. coli* B cells, produce plaques that are larger and more distinct (with sharper edges) than plaques formed by *rII*<sup>+</sup> wild-type phages.

2. *rII*<sup>-</sup> mutants are particularly useful for finding rare recombination events because they have an altered host range. In contrast to *rII*<sup>+</sup> wild-type phages, *rII*<sup>-</sup> mutants cannot form plaques in lawns of *E. coli* strain K( $\lambda$ )



b.1: © Seymour Benzer

**(b.2)**

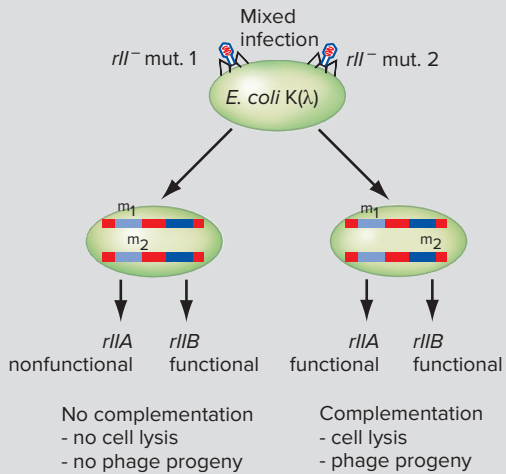
T4 strain	<i>E. coli</i> strain B	<i>E. coli</i> strain K( $\lambda$ )
<i>rII</i> <sup>-</sup>	Large, distinct	No plaques
<i>rII</i> <sup>+</sup>	Small, fuzzy	Small, fuzzy

(Continued)

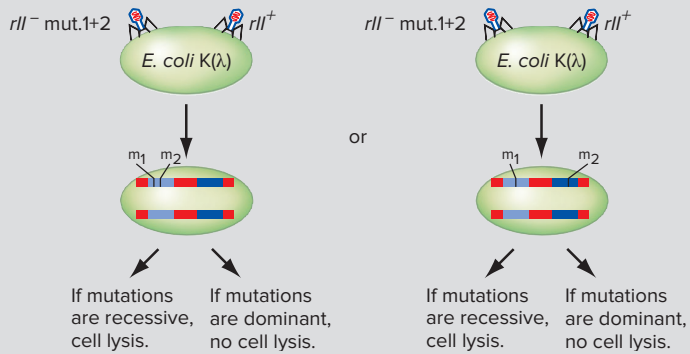
## FEATURE FIGURE 7.24 (Continued)

host bacteria. In a lysate with millions of  $rII^-$  phages, even a single  $rII^+$  recombinant phage can be identified because it can form a plaque on *E. coli* K( $\lambda$ ).

### (c.1) Complementation test (*trans* configuration)



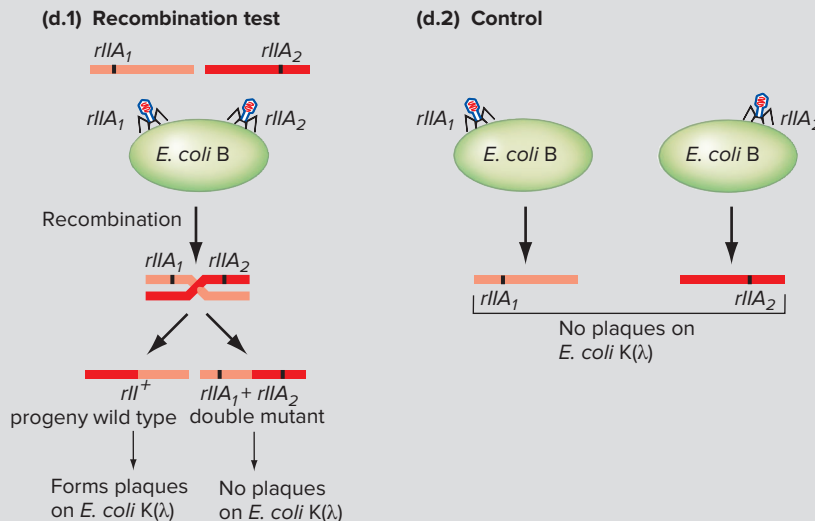
### (c.2) Control (*cis* configuration)



### (c) A customized complementation test between $rII^-$ mutants of bacteriophage T4

1. *E. coli* K( $\lambda$ ) cells are infected simultaneously with an excess of two different  $rII^-$  mutants ( $m_1$  and  $m_2$ ). Inside the cell, the two mutations will be in *trans*; that is, they lie on different chromosomes. If the two mutations are in the same gene, they will affect the same function and cannot complement each other, so no progeny phages will be produced. If the two mutations are in different genes ( $rIIA$  and  $rIIB$ ), they will complement each other, leading to progeny phage production and cell lysis.

2. An important control for this complementation test is the simultaneous infection of *E. coli* K( $\lambda$ ) with a wild-type T4 and a T4 strain in which  $m_1$  and  $m_2$  pairs that fail to complement have been recombined onto the same chromosome—the two mutations will be in *cis*. Release of phage progeny shows that both mutations are recessive to wild type and that the mutations do not interact in a way that prevents the cells from producing progeny phages. Complementation tests are meaningful only if the two mutations tested are both recessive to wild type.



### (d) Detecting recombination between two mutations in the same gene

1. *E. coli* B cells are infected with a large excess of two different  $rIIA^-$  mutants ( $rIIA_1$  and  $rIIA_2$ ). If no recombination between the two  $rIIA^-$  mutations takes place, all progeny phages will be  $rII^-$ . If recombination between the two mutations occurs, one of the products will be an  $rII^+$  recombinant, while the reciprocal product will be a double mutant containing both  $rIIA_1$  and  $rIIA_2$ . When the phage progeny subsequently infect *E. coli* K( $\lambda$ ) bacteria, only  $rII^+$  recombinants will be able to form plaques.

2. As a control, *E. coli* B cells are infected with a large amount of only one kind of mutant ( $rIIA_1$  or  $rIIA_2$ ). The only  $rII^+$  phages that can result are revertants of that mutation. Such revertants turn out to be extremely rare and can be ignored in most recombination experiments. Even if the two  $rIIA^-$  mutations are in adjacent base pairs, the number of  $rII^+$  recombinants obtained is more than 100 times higher than the number of  $rII^+$  revertants the cells infected by a single mutant can produce.



(Fig. 7.24c.1). To ensure that the two kinds of phages would infect almost every bacterial cell, he added many more phages of each type than there were bacteria.

When tested by Benzer's method, if the two  $rII^-$  mutations were in different genes, they would complement each other: Each of the mutant T4 chromosomes would supply one wild-type  $rII^+$  gene function, making up for the lack of that function in the other chromosome and resulting in lysis. On the other hand, if the two  $rII^-$  mutations were in the same gene, they would fail to complement: No plaques would appear because neither mutant chromosome would be able to supply the missing function.

Tests of many different pairs of  $rII^-$  mutations showed that they fall into two complementation groups:  $rIIA$  and  $rIIB$ . However, Benzer had to satisfy one final experimental requirement: For the complementation test to be meaningful, he had to make sure that pairs of  $rII^-$  mutations that failed to complement were each recessive to wild type and also did not interact with each other to produce an  $rII^-$  phenotype dominant to wild type. He checked these points by a control experiment in which he recombined pairs of  $rIIA^-$  or  $rIIB^-$  mutations onto the same chromosome (as described in the next section) and then simultaneously infected *E. coli* K( $\lambda$ ) with these double  $rII^-$  mutants and with wild-type phages (Fig. 7.24c.2). If the mutations were recessive and did not interact with each other, the cells would lyse, in which case the complementation test would be interpretable.

The significant distinction between the actual complementation test and the control experiment is in the placement of the two  $rII^-$  mutations. In the complementation test, one  $rII^-$  mutation is on one chromosome, while the other  $rII^-$  mutation is on the other chromosome (Fig. 7.24c.1); two mutations arranged in this way are said to be in the **trans configuration**. In the control experiment (Fig. 7.24c.2), the two mutations are on the same chromosome, in the so-called **cis configuration**. The complete test, including the complementation test and the control experiment, is known as a *cis-trans* test. In the complete experiment, two mutations that do not produce lysis in *trans* but do so when in *cis* are in the same complementation group. Benzer called any complementation group identified by the *cis-trans* test a **cistron**, and some geneticists still use the term *cistron* as a synonym for *gene*.

With the knowledge that the *rII* locus consists of two genes ( $rIIA$  and  $rIIB$ ), Benzer could look for two mutations in the same gene and then see if they ever recombine to produce wild-type progeny.

### Recombination between different mutations in a single gene

When Benzer infected *E. coli* B strain bacteria with a mixture of phages carrying different mutations in the same gene ( $rIIA_1$  and  $rIIA_2$ , for example), he did observe the

appearance of rare  $rII^+$  progeny (Fig. 7.24d.1). He knew these wild-type progeny resulted from recombination and not from reverse mutations because the frequencies of the  $rII^+$  phage particles he observed, even if rare, were much higher than the frequencies of  $rII^+$  revertants seen among progeny produced by infecting B strain bacteria with either mutant alone (Fig. 7.24d.2).

These experiments were possible only because Benzer devised a **selection** for rare  $rII^+$  recombinants. In a selection, conditions are such that the only survivors are the rare individuals you seek to identify. Benzer's selection condition for identifying rare  $rII^+$  recombinant progeny was plating for plaques on *E. coli* K( $\lambda$ ). Benzer could assay a phage lysate containing tens of thousands of phage progeny on a single petri plate containing a lawn of *E. coli* K( $\lambda$ ). Because none of the  $rII^-$  phage in the lysate could form plaques, even a single  $rII^+$  recombinant among them could be identified as a plaque.

On the basis of his observations with the *rII* genes, Benzer drew three conclusions about gene structure and function: (1) A gene consists of different parts that can each mutate; (2) recombination can occur between different mutable sites in the same gene; and (3) a gene performs its normal function only if all of its components are wild type. From what we now know about the molecular structure of DNA, this all makes perfect sense: The different mutable units are the base pairs that constitute the gene.

## A Gene Is a Discrete Linear Set of Nucleotide Pairs

How are the multiple nucleotide pairs that make up a gene arranged—in a continuous row, or dispersed in precise patterns around the genome? And do the various mutations that affect gene function alter many different nucleotides, or only a small subset within each gene?

### Using deletions to map mutations approximately

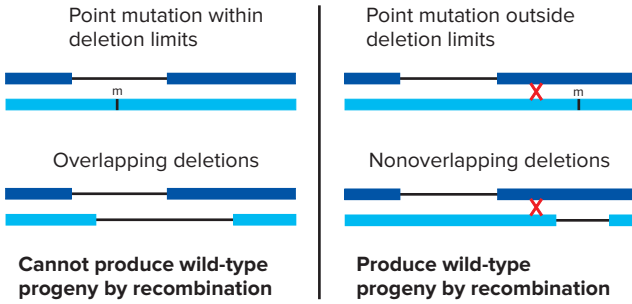
To answer these questions about the arrangement of nucleotides in a gene, Benzer eventually obtained thousands of spontaneous and mutagen-induced  $rII^-$  mutations that he needed to map with respect to each other.

To map the location of a thousand mutants through comparisons of all possible two-point crosses, Benzer would have had to set up a million ( $10^3 \times 10^3$ ) matings. But by taking advantage of bacteriophage strains with large deletions, he could obtain the same information with far fewer crosses.

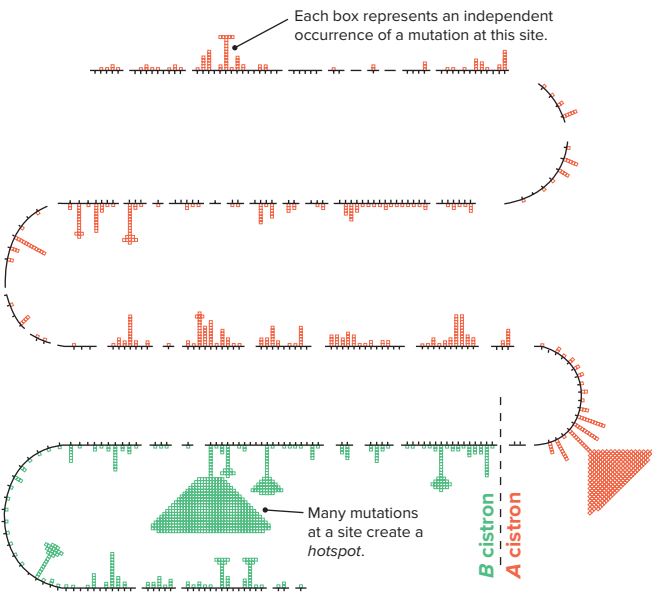
These large deletions are mutations that remove many contiguous nucleotide pairs along a DNA molecule. In crosses between bacteriophages carrying a mutation and bacteriophages carrying deletions of the corresponding region, no wild-type recombinant progeny can arise, because

**Figure 7.25** Fine structure mapping of the bacteriophage T4 *rII* genes. (a) A phage cross between a point mutation and a deletion removing the DNA at the position of the mutation cannot yield wild-type recombinants. The same is true if two different deletion mutations overlap each other. (b) Large deletions divide the *rII* locus into regions; finer deletions divide each region into subsections. Point mutations, such as 271 (in red at bottom), map to region 3 if they do not recombine with deletions PT1, PB242, or A105 but do recombine with deletion 638 (top). Point mutations can be mapped to subsections of region 3 using other deletions (middle). Recombination tests map point mutations in the same subregion (bottom). Point mutations 201 and 155 cannot recombine to yield wild-type recombinants because they affect the same nucleotide pair. (c) Benzer's fine structure map. Hotspots are locations with many independent mutations that cannot recombine with each other.

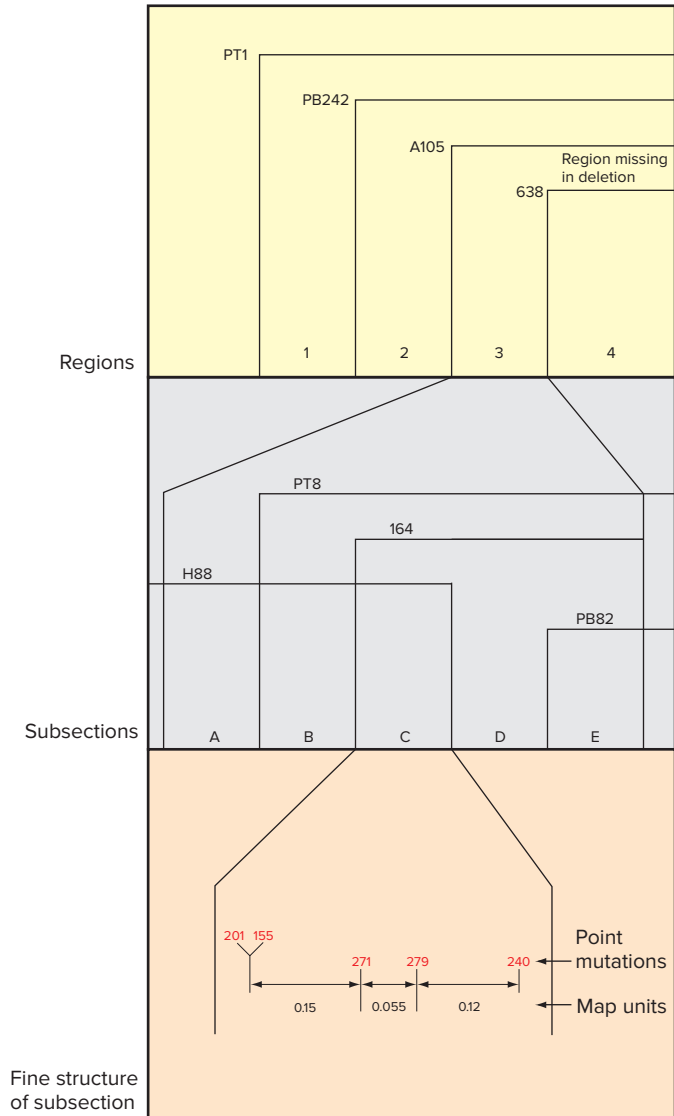
(a) Using deletions for rapid mapping



(c) Fine structure of the *rII* region



(b) Portion of the *rIIA* deletion map at increasing resolutions



neither chromosome carries the proper information at the location of the mutation. However, if the mutation lies outside the region deleted from the homologous chromosome, wild-type progeny can appear (Fig. 7.25a). This is true whether the mutation is a point mutation affecting one or a few nucleotides, or is itself a large deletion. Crosses between any uncharacterized mutation and a known deletion thus immediately reveal whether the mutation resides in the

region deleted from the other phage chromosome, providing a rapid way to find the general location of a mutation.

Using a series of overlapping deletions, Benzer divided the *rII* region into a series of relatively small regions, or intervals. He could then assign any point mutation to an interval by observing whether the point mutation recombined to give *rII<sup>+</sup>* progeny when crossed with the series of deletions (Fig. 7.25b).

Benzer mapped 1612 spontaneous point mutations and several deletions in the *rII* locus of bacteriophage T4 through recombination analysis. He first used recombination to determine the relationship between the deletions. He then found the approximate location of individual point mutations by observing which deletions could recombine with each point mutant to yield wild-type progeny.

### Determining RF between *rII*<sup>-</sup> mutations for precise mapping

Benzer next performed recombination tests to measure the genetic distance between pairs of point mutations he had found by deletion mapping to lie in the same small region of the chromosome. The distance between any two *rII*<sup>-</sup> mutants could be measured simply by counting the number of *rII*<sup>+</sup> and total phages in an aliquot of lysate from a phage cross. The RF (in map units) is simply the number of *rII*<sup>+</sup> recombinants [plaques on *E. coli* K(λ)] divided by the total number of phages (plaques on *E. coli* B), multiplied by 2 to account for the *rII*<sup>-</sup> double mutant reciprocal recombinants that must exist but cannot be detected easily:

$$\text{RF} = \frac{2[\text{number of plaques on } E. \text{ coli K}(\lambda)]}{(\text{number of plaques on } E. \text{ coli B})}$$

Benzer combined the results from deletion mapping and RF calculations to produce a map of the *fine structure* of the *rII* region (Fig. 7.25c). Note that all of the point mutations in the *rIIA* complementation group mapped to one side of the *rII* region, and all of the *rIIB* point mutations mapped to the other side.

### How DNA nucleotides are organized into genes

Benzer knew that the genetic distances between all mapped genes in the T4 genome add up to about 1500 m.u. He also knew that the T4 chromosome constitutes about 169,000 bp of DNA, so he could calculate that for bacteriophage T4, 1 m.u. corresponds to  $169,000/1500 = 113$  bp. The lowest RF that he measured between any pair of *rII*<sup>-</sup> mutants was 0.02 m.u., which would represent about 2 bp. Benzer thus inferred that a mutation can arise from the change of even a single nucleotide pair, and that recombination can occur between adjacent nucleotide pairs. From the observation that mutations within the *rII* region form a self-consistent, linear recombination map, he concluded that a gene is composed of a continuous linear sequence of nucleotide pairs within the DNA. And from observations that the positions of mutations in the *rIIA* gene did not overlap those of the *rIIB* gene, he determined that the nucleotide sequences composing those two genes are separate and distinct. A *gene* is thus a linear set of nucleotide pairs, located within a discrete region of a chromosome, that serves as a unit of function.

### Hotspots of mutation

Some sites within a gene mutate spontaneously more often than others and as a result are known as **mutation hotspots** (Fig. 7.25c). The existence of hotspots suggests that certain nucleotides can be altered more readily than others. Treatment with mutagens also turns up hotspots, but because mutagens have specificities for particular nucleotides, the highly mutable sites that turn up with various mutagens are often at different positions in a gene than the hotspots resulting from spontaneous mutation.

Nucleotides are the same chemically whether they lie within a gene or in the DNA between genes. Furthermore, as Benzer's experiments imply, the molecular machinery responsible for mutation and recombination does not discriminate between those nucleotides that are *intragenic* (within a gene) and those that are *intergenic* (between genes). The main distinction between DNA within and DNA outside a gene is that the array of nucleotides composing a gene has evolved a function that determines phenotype. Next, we describe how geneticists discovered what that function is.

#### essential concepts

- A *complementation test* determines whether two different recessive mutations occur in the same gene or in different genes.
- At the DNA level, a gene is a linear sequence of nucleotide pairs in a discrete region of a chromosome that confers a specific unit of function.
- Recombination can occur between any two nucleotide pairs, whether they are within the same gene or not.

## 7.5 What Mutations Tell Us About Gene Function

#### learning objectives

1. Explain how the analysis of arginine auxotrophs implied that a single gene corresponds to a single enzyme.
2. Describe how missense mutations were used to show that genes determine the amino acid sequences of proteins.
3. Differentiate between primary, secondary, tertiary, and quaternary structures of proteins.

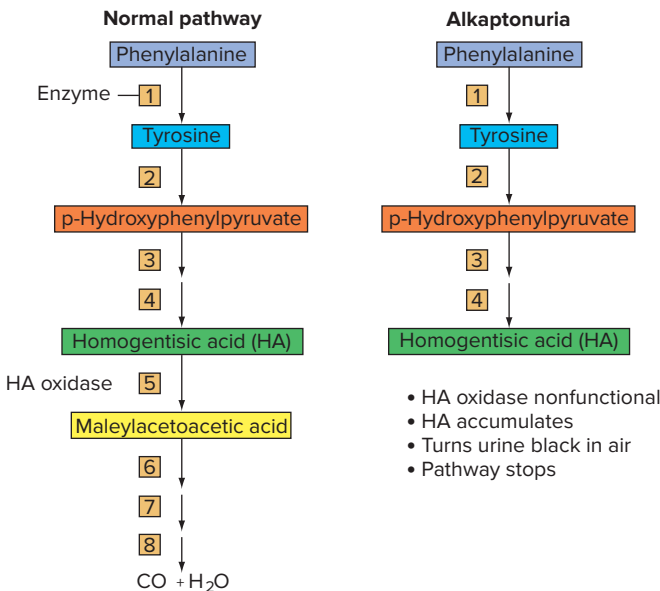
Mendel's experiments established that an individual gene can control a visible characteristic, but his laws do not explain how genes actually govern the appearance of traits.

Investigators working in the first half of the twentieth century studied carefully the biochemical changes caused by mutations in an effort to understand the genotype–phenotype connection.

In one of the first of these studies, conducted in 1902, the British physician Dr. Archibald Garrod showed that a human genetic disorder known as *alkaptonuria* is determined by the recessive allele of an autosomal gene. Garrod analyzed family pedigrees and performed biochemical analyses on family members with and without the trait. The urine of people with alkaptonuria turns black on exposure to air. Garrod found that a substance known as *homogentisic acid*, which blackens upon contact with oxygen, accumulates in the urine of alkaptonuria patients. Alkaptonurians excrete all of the homogentisic acid they ingest, while people without the condition excrete no homogentisic acid in their urine even after ingesting the substance.

From these observations, Garrod concluded that people with alkaptonuria are incapable of metabolizing homogentisic acid to the breakdown products generated by normal individuals (**Fig. 7.26**). Because many biochemical reactions within the cells of organisms are catalyzed by enzymes, Garrod hypothesized that lack of the enzyme that breaks down homogentisic acid is the cause of alkaptonuria. In the absence of this enzyme, homogentisic acid accumulates and causes the urine to turn black on contact with oxygen. He called this condition an *inborn error of metabolism*.

**Figure 7.26 Alkaptonuria: An inborn error of metabolism.** The biochemical pathway in humans that degrades phenylalanine and tyrosine via homogentisic acid (HA). In alkaptonuria patients, the enzyme HA hydroxylase is not functional, so it does not catalyze the conversion of HA to maleylacetoacetic acid. As a result, HA, which oxidizes to a black compound, accumulates in the urine.



Garrod studied several other inborn errors of metabolism and suggested that all arose from mutations that prevented a particular gene from producing an enzyme required for a specific biochemical reaction. In today's terminology, the wild-type allele of the gene would allow production of functional enzyme (in the case of alkaptonuria, the enzyme is homogentisic acid oxidase), whereas the mutant allele would not. Because the single wild-type allele in heterozygotes generates sufficient enzyme to prevent the accumulation of homogentisic acid and thus the condition of alkaptonuria, the mutant allele is recessive.

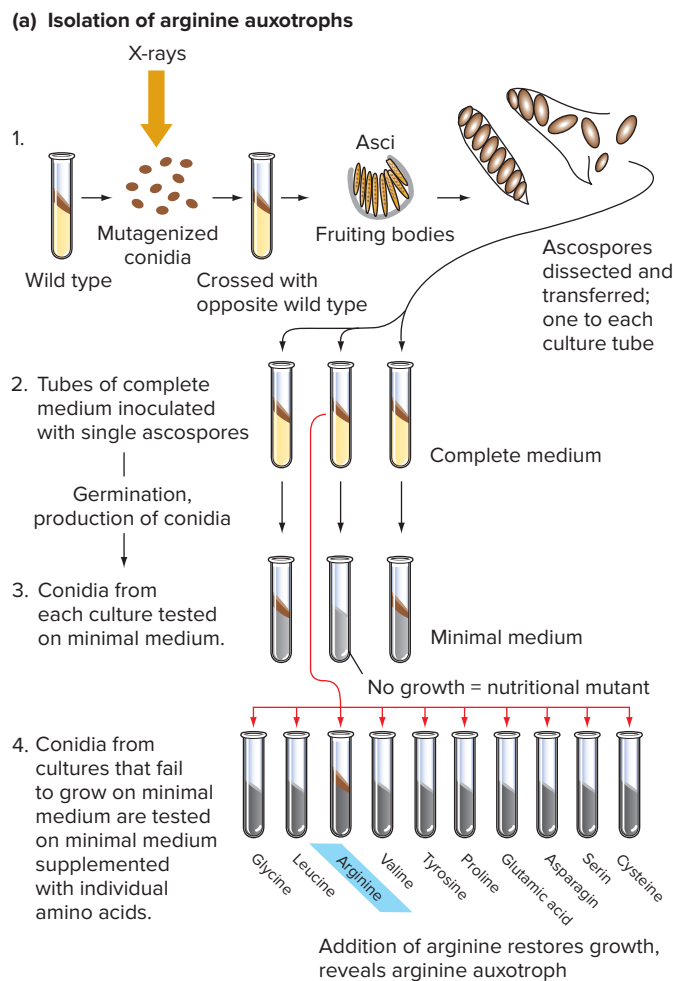
## A Gene Contains the Information for Producing a Specific Enzyme: The One Gene, One Enzyme Hypothesis

In the 1940s, George Beadle and Edward Tatum carried out a series of experiments on the bread mold *Neurospora crassa* (whose life cycle was described in Chapter 5) that demonstrated a direct relationship between genes and the enzymes that catalyze specific biochemical reactions. Their strategy was simple. They first isolated a number of mutations that disrupted the synthesis of the amino acid arginine, a compound needed for *Neurospora* growth. They next hypothesized that different mutations blocked different steps in a particular **biochemical pathway**: the orderly series of reactions that allows *Neurospora* to obtain simple molecules from the environment and convert them step-by-step into successively more complicated molecules culminating in the end product arginine.

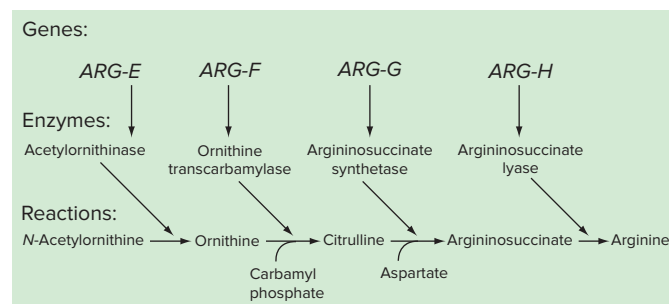
### Experimental evidence for *one gene, one enzyme*

**Figure 7.27a** illustrates the experiments Beadle and Tatum performed to test their hypothesis. They first obtained a set of mutagen-induced mutations that prevented *Neurospora* from synthesizing arginine. Cells with any one of these mutations were unable to make arginine and could therefore grow on a minimal medium containing salt and sugar only if it had been supplemented with arginine. A nutritional mutant microorganism that requires supplementation with substances not needed by wild-type strains is known as an **auxotroph**. The cells just mentioned were arginine auxotrophs. (In contrast, a cell that does not require the addition of a substance is a **prototroph** for that factor. In a more general meaning, *prototroph* refers to a wild-type cell that can grow on minimal medium alone.)

Recombination analyses located the auxotrophic arginine-blocking mutations in four distinct regions of the genome, and complementation tests showed that each of the four regions correlated with a different complementation group. On the basis of these results, Beadle and Tatum concluded that at least four genes support the biochemical

**(b) Growth response if nutrient is added to minimal medium**

Mutant strain	Supplements				
	Nothing	Ornithine	Citrulline	Arginino-succinate	Arginine
Wildtype: Arg <sup>+</sup>	+	+	+	+	+
ARG-E <sup>-</sup>	-	+	+	+	+
ARG-F <sup>-</sup>	-	-	+	+	+
ARG-G <sup>-</sup>	-	-	-	+	+
ARG-H <sup>-</sup>	-	-	-	-	+

**(c) Inferred biochemical pathway**

**Figure 7.27** Experimental support for the *one gene, one enzyme* hypothesis. **(a)** Beadle and Tatum mated an X-ray-mutagenized strain of *Neurospora* with another strain, and they isolated haploid ascospores that grew on complete medium. Cultures that failed to grow on minimal medium were nutritional mutants. Nutritional mutants that could grow on minimal medium plus arginine were Arg<sup>-</sup> auxotrophs. **(b)** The ability of wild-type and mutant strains to grow on minimal medium supplemented with intermediates in the arginine pathway. **(c)** Each of the four ARG genes specifies an enzyme needed to convert one intermediate to the next in the pathway.

pathway for arginine synthesis. They named the four genes *ARG-E*, *ARG-F*, *ARG-G*, and *ARG-H*.

They next asked whether any of the mutant *Neurospora* strains could grow in minimal medium supplemented with any of three known intermediates (ornithine, citrulline, and argininosuccinate) in the biochemical pathway leading to arginine, instead of with arginine itself. This test would identify *Neurospora* mutants able to convert the intermediate compound into arginine. Beadle and Tatum compiled a table describing which arginine auxotrophic mutants were able to grow on minimal medium supplemented with each of the intermediates (**Fig. 7.27b**).

### Interpretation of results: Genes encode enzymes

On the basis of these results, Beadle and Tatum proposed a model of how *Neurospora* cells synthesize arginine (**Fig. 7.27c**). In the linear progression of biochemical reactions by which a cell constructs arginine from the

constituents of minimal medium, each intermediate is both the product of one step and the substrate for the next. Each reaction in the precisely ordered sequence is catalyzed by a specific enzyme, and the presence of each enzyme depends on one of the four ARG genes.

A mutation in one gene blocks the pathway at a particular step because the cell lacks the corresponding enzyme and thus cannot make arginine on its own. Supplementing the medium with any intermediate that occurs beyond the blocked reaction restores growth to the mutant because the organism has all the enzymes required to convert the intermediate to arginine. Supplementation with an intermediate that occurs before the missing enzyme does not work because the mutant cell cannot convert the intermediate into arginine.

Each mutation abolishes the cell's ability to make an enzyme capable of catalyzing a certain reaction. By inference, then, each gene controls the synthesis or activity of an enzyme, or as stated by Beadle and Tatum: one gene, one enzyme. Of course, the gene and the enzyme are not the same thing; rather, the sequence of nucleotides in a

gene contains information that somehow encodes the structure of an enzyme molecule.

Although the analysis of the arginine pathway studied by Beadle and Tatum was straightforward, studies of biochemical pathways are not always so easy to interpret. Some biochemical pathways are not linear progressions of stepwise reactions. For example, a branching pathway occurs if different enzymes act on the same intermediate to convert it into two different end products. If the cell requires both of these end products for growth, a mutation in a gene encoding any of the enzymes required to synthesize the intermediate would make the cell dependent on supplementation with both end products. A second possibility is that a cell might employ either of two independent, parallel pathways to synthesize a needed end product. In such a case, a mutation in a gene encoding an enzyme in one of the pathways would be without effect. Only a cell with mutations affecting both pathways would display an aberrant phenotype.

Even with nonlinear progressions such as these, careful genetic analysis can reveal the nature of the biochemical pathway on the basis of Beadle and Tatum's insight that genes specify proteins.

### Genes Specify the Identity and Order of Amino Acids in Polypeptide Chains

Although the one gene, one enzyme hypothesis was a crucial advance in understanding how genes influence phenotype, it is an oversimplification. Not all genes govern the construction of enzymes active in biochemical pathways. Enzymes are only one class of the molecules known as *proteins*, and cells contain many other kinds of proteins. Among the other types are proteins that provide shape and rigidity to a cell, proteins that transport molecules in and out of cells, proteins that help fold DNA into chromosomes, and proteins that act as hormonal messengers. Genes direct the synthesis of all proteins, enzymes and nonenzymes alike. Moreover, as we see next, genes actually determine the construction of *polypeptides*, and because some proteins are composed of more than one type of polypeptide, more than one gene determines the construction of such proteins.

#### Proteins: Linear polymers of amino acids linked by peptide bonds

**Proteins** are polymers composed of building blocks known as **amino acids**. Cells use mainly 20 different amino acids to synthesize the proteins they need. All of these amino acids have certain basic features, encapsulated by the formula

$\text{NH}_2\text{-CHR-COOH}$  (**Fig. 7.28a**). The  $\text{-COOH}$  component, also known as *carboxylic acid*, is, as the name implies, acidic; the  $\text{-NH}_2$  component, also known as an *amino group*, is basic. The R refers to side chains that distinguish each of the amino acids (**Fig. 7.28b**). An R group can be as simple as a hydrogen atom (in the amino acid glycine) or as complex as a benzene ring (in phenylalanine). Some side chains are relatively neutral and nonreactive, others are acidic, and still others are basic.

In addition to the 20 common amino acids, two rare ones can be incorporated into proteins in specific circumstances (**Fig. 7.28c**). A very few proteins (only 25 in humans) are known to contain selenocysteine. Pyrrolysine is present only in the proteins of certain prokaryotic organisms.

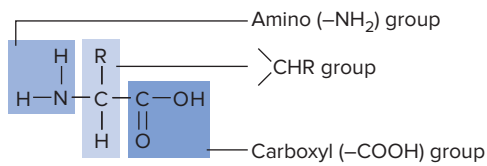
During protein synthesis, a cell's protein-building machinery links amino acids by constructing covalent **peptide bonds** that join the  $\text{-COOH}$  group of one amino acid to the  $\text{-NH}_2$  group of the next (**Fig. 7.28d**). A pair of amino acids connected in this fashion is a *dipeptide*; several amino acids linked together constitute an *oligopeptide*. The amino acid chains that make up proteins contain hundreds to thousands of amino acids joined by peptide bonds and are known as **polypeptides**. Proteins are thus linear polymers of amino acids. Like the chains of nucleotides in DNA, polypeptides have a chemical polarity. The end of a polypeptide synthesized first is called the **N terminus** because it contains a free amino group that is not connected to any other amino acid. The other end of the polypeptide chain is the **C terminus** because it contains a free carboxylic acid group.

#### Mutations can alter amino acid sequences

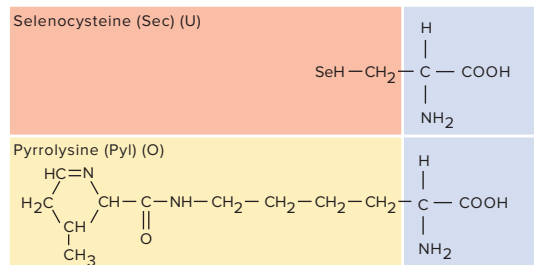
Each protein is composed of a unique sequence of amino acids. The chemical properties that enable structural proteins to give a cell its shape, or allow enzymes to catalyze specific reactions, are a direct consequence of the identity, number, and linear order of amino acids in the protein.

If genes specify proteins, then at least some mutations could be changes in a gene that alter the normal sequence of amino acids in the protein specified by that gene. In the mid-1950s, Vernon Ingram began to establish what kinds of changes particular mutations cause in the corresponding protein. Using techniques that had just been developed for determining the sequence of amino acids in a protein, he compared the amino acid sequence of the normal adult form of hemoglobin (HbA) with that of hemoglobin in the bloodstream of people homozygous for the mutation that causes sickle-cell anemia (HbS). Remarkably, he found only a single amino acid difference

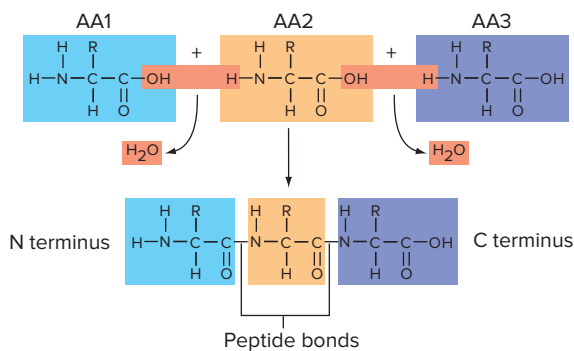
**(a) Generic amino acid structure**



**(c) Rare amino acids**

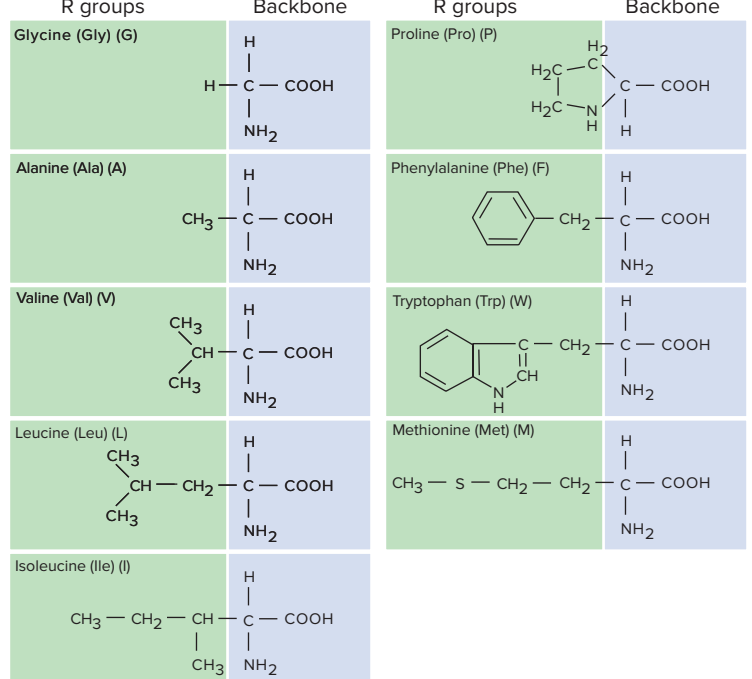


**(d) Peptide bond formation**

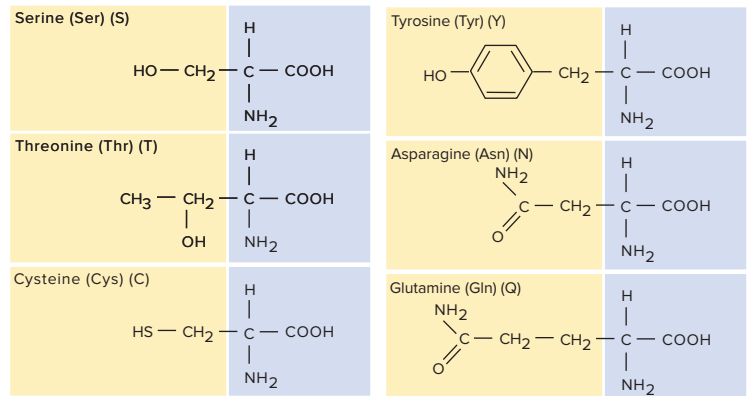


**Figure 7.28** Proteins are chains of amino acids linked by peptide bonds. **(a)** Amino acids contain a basic amino group ( $-\text{NH}_2$ ), an acidic carboxylic acid group ( $-\text{COOH}$ ), and one of 22 different side chains (R). **(b)** The 20 amino acids commonly found in proteins, arranged according to the properties of their R groups. **(c)** Selenocysteine and pyrrolysine are amino acids found only in a few proteins or in specific organisms. **(d)** One molecule of water is lost when a covalent amide linkage (a peptide bond) is formed between the  $-\text{COOH}$  of one amino acid and the  $-\text{NH}_2$  of the next amino acid. Polypeptides such as the tripeptide shown here have polarity; they extend from an N terminus (with a free amino group) to a C terminus (with a free carboxylic acid group).

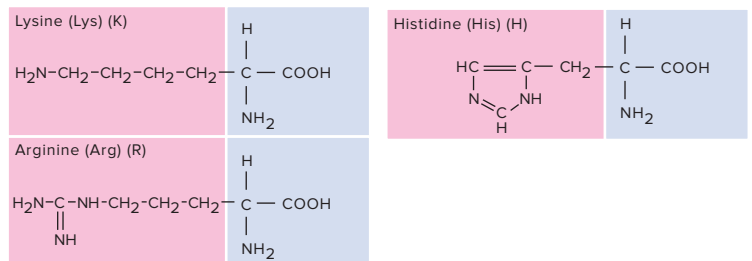
**(b) Amino acids with nonpolar R groups**



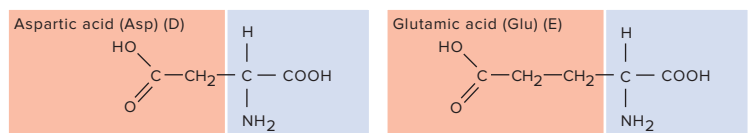
**Amino acids with uncharged polar R groups**



**Amino acids with basic R groups**

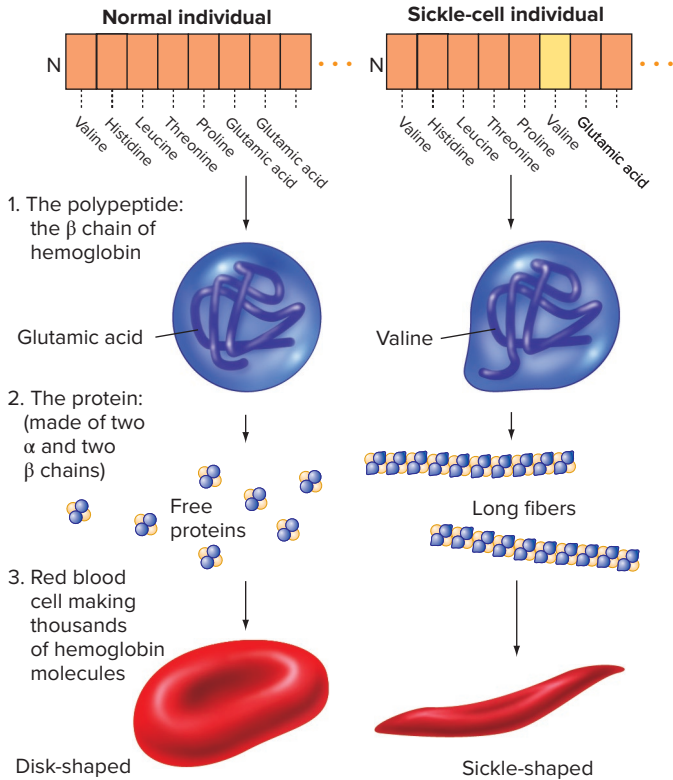


**Amino acids with acidic R groups**

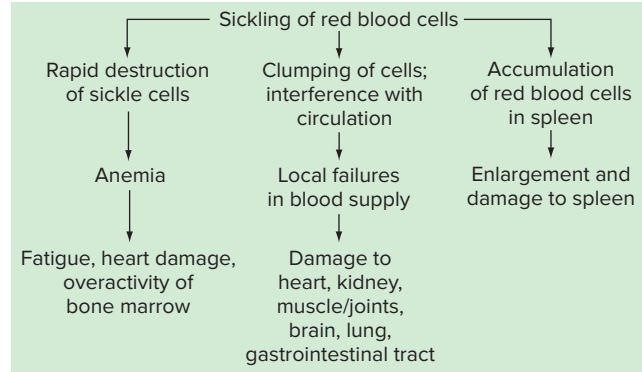


**Figure 7.29** The molecular basis of sickle-cell and other anemias. (a) Substitution of glutamic acid by valine at the sixth amino acid from the N terminus affects the three-dimensional structure of the  $\beta$  chain of hemoglobin. Hemoglobins incorporating the mutant  $\beta$  chain form aggregates that cause red blood cells to sickle. (b) Red blood cell sickling has many phenotypic effects. (c) Other mutations in the  $\beta$  chain gene also cause anemias.

**(a) From mutation to phenotype**



**(b) Sickle-cell anemia is pleiotropic**



**(c)  $\beta$  chain substitutions/variants**

	Amino acid position														
	1	2	3	...	6	7	...	26	...	63	...	125	...	146	
Normal (HbA)	Val	His	Leu		Glu	Glu		Glu		His		Val		Glu	His
HbS	Val	His	Leu		Val	Glu		Glu		His		Val		Glu	His
HbC	Val	His	Leu		Lys	Glu		Glu		His		Val		Glu	His
HbG San Jose	Val	His	Leu		Glu	Gly		Glu		His		Val		Glu	His
HbE	Val	His	Leu		Glu	Glu		Lys		His		Val		Glu	His
HbM Saskatoon	Val	His	Leu		Glu	Glu		Glu		Tyr		Val		Glu	His
Hb Zurich	Val	His	Leu		Glu	Glu		Glu		Arg		Val		Glu	His
HbM Milwaukee 1	Val	His	Leu		Glu	Glu		Glu		His		Glu		Glu	His
HbD $\beta$ Punjab	Val	His	Leu		Glu	Glu		Glu		His		Val		Gln	His

between the wild-type and mutant proteins (Fig. 7.29a). Hemoglobin consists of two types of polypeptides: a so-called  $\alpha$  (alpha) chain and a  $\beta$  (beta) chain. The sixth amino acid from the N terminus of the  $\beta$  chain was glutamic acid in normal individuals but valine in sickle-cell patients.

Ingram thus established that a mutation substituting one amino acid for another had the power to change the structure and function of hemoglobin and thereby alter the phenotype from normal to sickle-cell anemia (Fig. 7.29b). We now know that the glutamic acid-to-valine change affects the solubility of hemoglobin within the red blood cell. At low concentrations of oxygen, the less soluble sickle-cell form of hemoglobin aggregates into long chains that deform the red blood cell (Fig. 7.29a).

Because people suffering from a variety of inherited anemias also have defective hemoglobin molecules, Ingram and other geneticists were able to determine how a large number of different mutations affect the amino acid sequence of hemoglobin (Fig. 7.29c). Most of the altered hemoglobins have a change in only one amino acid. In various patients with anemia, the alteration is generally in

different amino acids, but occasionally, two independent mutations result in different substitutions for the same amino acid. Geneticists use the term **missense mutation** to describe a genetic alteration that causes the substitution of one amino acid for another.

**A Protein’s Amino Acid Sequence Dictates Its Three-Dimensional Structure**

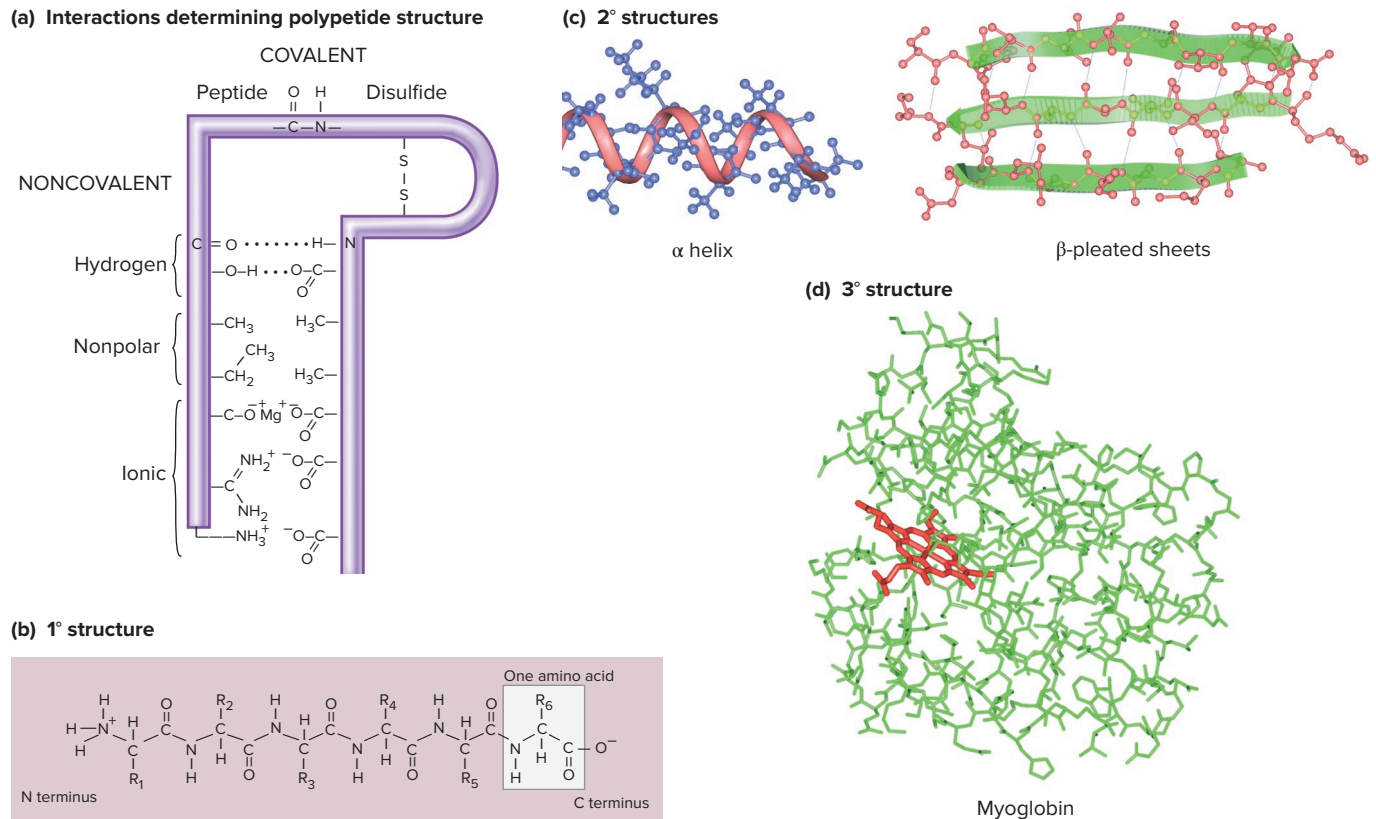
Despite the uniform nature of protein construction—a string of amino acids joined by peptide bonds—each polypeptide folds into a unique three-dimensional shape. Biochemists often distinguish between four levels of protein structure: *primary*, *secondary*, *tertiary*, and *quaternary*. The first three of these apply to any one polypeptide chain, while the quaternary level describes associations between multiple polypeptides within a protein complex.

**Primary, secondary, and tertiary protein structures**

The linear sequence of amino acids within a polypeptide is its **primary structure**. Each unique primary structure



**Figure 7.30 Levels of polypeptide structure.** (a) Covalent and noncovalent interactions determine the structure of a polypeptide. (b) A polypeptide's primary (1°) structure is its amino acid sequence. (c) Localized regions form secondary (2°) structures such as a helix and  $\beta$ -pleated sheets. (d) The tertiary (3°) structure is the complete three-dimensional arrangement of a polypeptide. In this portrait of myoglobin, the iron-containing heme group, which carries oxygen, is red, while the polypeptide itself is green.



places constraints on how a chain can arrange itself in three-dimensional space. Because the R groups distinguishing the 22 amino acids have dissimilar chemical properties, some amino acids form hydrogen bonds or electrostatic bonds when brought into proximity with other amino acids. Nonpolar amino acids, for example, may become associated with each other by interactions that hide them from water in localized hydrophobic regions. As another example, two cysteine amino acids can form covalent disulfide bridges ( $-S-S-$ ) through the oxidation of their  $-SH$  groups.

All of these interactions (**Fig. 7.30a**) help stabilize the polypeptide in a specific three-dimensional conformation. The primary structure (**Fig. 7.30b**) determines three-dimensional shape by generating **secondary structure**: localized regions with a characteristic geometry (**Fig. 7.30c**). Primary structure is also responsible for other folds and twists that together with the secondary structure produce the ultimate three-dimensional **tertiary structure** of the entire polypeptide (**Fig. 7.30d**). Normal tertiary structure—the way a long chain of amino acids naturally folds in three-dimensional space under physiological

conditions—is known as a polypeptide's **native configuration**. Various forces, including hydrogen bonds, electrostatic bonds, hydrophobic interactions, and disulfide bridges help stabilize the native configuration.

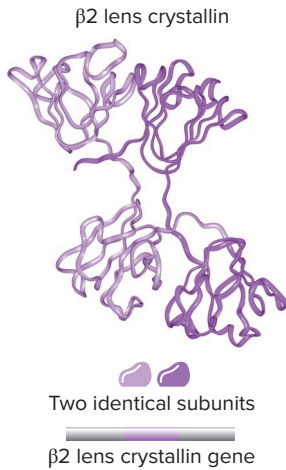
It is worth repeating that primary structure—the sequence of amino acids in a polypeptide—directly determines secondary and tertiary structures. The information required for the chain to fold into its native configuration is inherent in its linear sequence of amino acids.

In one example of this principle, many proteins unfold, or become **denatured**, when exposed to urea and mercaptoethanol or to increasing heat or pH. These treatments disrupt the interactions that normally stabilize the secondary and tertiary structures. When conditions return to normal, some proteins spontaneously refold into their native configuration without help from other agents. No other information beyond the primary structure is needed to achieve the proper three-dimensional shape of such proteins.

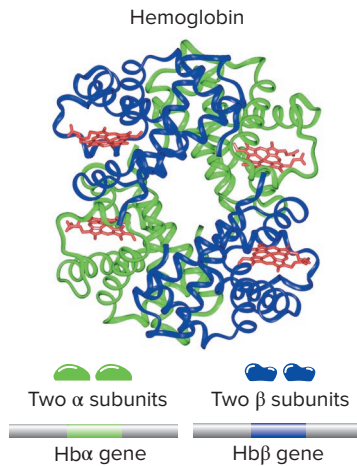
You should know that some proteins are unable after denaturation to refold by themselves into their correct tertiary structure. The proper folding of these proteins requires other proteins called **chaperones** that help stabilize

**Figure 7.31 Multimeric proteins.** (a)  $\beta 2$  lens crystallin contains two copies of one kind of subunit; the two subunits are the product of a single gene. The peptide backbones of the two subunits are shown in different shades of purple. (b) Hemoglobin is composed of two different kinds of subunits, each encoded by a different gene. (c) Three distinct protein receptors for the immune-system molecules called interleukins (ILs; purple). All contain a common gamma ( $\gamma$ ) chain (yellow), plus other receptor-specific polypeptides (green). A mutant  $\gamma$  chain blocks the function of all three receptors, leading to XSCID. (d) One  $\alpha$ -tubulin (red) and one  $\beta$ -tubulin (blue) polypeptide associate to form a tubulin dimer. Many tubulin dimers form a single microtubule. The mitotic spindle is an assembly of many microtubules.

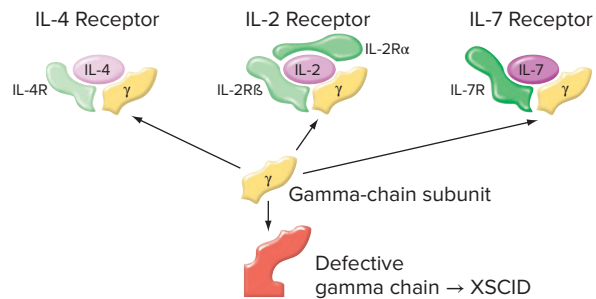
**(a) A multimer with identical subunits**



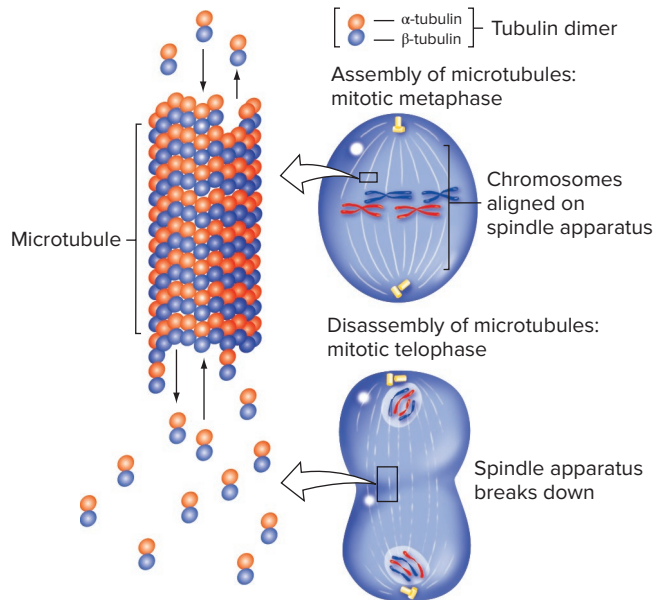
**(b) A multimer with nonidentical subunits**



**(c) One polypeptide in different proteins**



**(d) Microtubules: large assemblies of subunits**



the native configuration. Because elevated temperatures cause protein unfolding, many chaperones are *heat shock proteins* that are made when organisms are exposed to high temperatures. These heat shock proteins protect cells from damage due to protein misfolding under high-temperature conditions. But even for proteins that need chaperones to achieve their native configurations, the amino acid sequence of the protein dictates the final three-dimensional structure.

### Quaternary structure: Multimeric proteins

Certain proteins, such as the rhodopsin that promotes black-and-white vision, consist of a single polypeptide. Many others, however, such as the lens crystallin protein,

which provides rigidity and transparency to the lenses of our eyes, or the hemoglobin molecule described earlier, are composed of two or more polypeptide chains that associate in a specific way (Fig. 7.31a and b). The individual polypeptides in an aggregate are known as **subunits**, and the complex of subunits is often referred to as a **multimer**. The three-dimensional configuration of subunits in a multimer is a complex protein's **quaternary structure**.

The same forces that stabilize the native form of a polypeptide (that is, hydrogen bonds, electrostatic bonds, hydrophobic interactions, and disulfide bridges) also contribute to the maintenance of quaternary structure. As Fig. 7.31a shows, in some multimers, the two or more interacting subunits are identical polypeptides. These identical chains are specified by one gene. In other multimers, by

contrast, more than one kind of polypeptide makes up the protein (Fig. 7.31b). The different polypeptides in these multimers are specified by different genes.

Alterations in just one kind of subunit, caused by a mutation in a single gene, can affect the function of a multimer. The adult hemoglobin molecule, for example, consists of two  $\alpha$  and two  $\beta$  subunits, with each type of subunit determined by a different gene—one for the  $\alpha$  chain and one for the  $\beta$  chain. A mutation in the Hb $\beta$  gene resulting in an amino acid switch at position 6 in the  $\beta$  chain causes sickle-cell anemia.

Similarly, if several multimeric proteins share a common subunit, a single mutation in the gene encoding that subunit may affect all the proteins simultaneously. An example is an X-linked mutation in mice and humans that incapacitates several different proteins all known as *interleukin (IL) receptors*. Because all of these receptors are essential to the normal function of immune system cells that fight infection and generate immunity, this one mutation causes the life-threatening condition known as *X-linked severe combined immune deficiency (XSCID)*; Fig. 7.31c).

The polypeptides of complex proteins can assemble into extremely large structures capable of changing with the needs of the cell. For example, the microtubules that make up the spindle during mitosis are gigantic assemblages of mainly two polypeptides:  $\alpha$ -tubulin and  $\beta$ -tubulin (Fig. 7.31d). The cell can organize these subunits into very long hollow tubes that grow or shrink as needed at different stages of the cell cycle.

### One gene, one polypeptide

Because more than one gene governs the production of some multimeric proteins and because not all proteins are enzymes, the *one gene, one enzyme* hypothesis is not broad enough to define gene function. A more accurate statement is *one gene, one polypeptide*: Each gene governs the construction of a particular polypeptide. As you will see in Chapter 8, even this reformulation does not encompass the function of all genes, as some genes in all organisms do not determine the construction of proteins; instead, they specify RNAs that are not translated into polypeptides.

Knowledge about the connection between genes and polypeptides enabled geneticists to analyze how different mutations in a single gene can produce different phenotypes. If each amino acid has a specific effect on the three-dimensional structure of a protein, then changing amino acids at different positions in a polypeptide chain can alter protein function in different ways. For example, most enzymes have an active site that carries out the enzymatic task, while other parts of the protein support the shape and position of that site. Mutations that change the identity of amino acids at the active site may have more serious consequences than those affecting amino acids outside the active

site. Some kinds of amino acid substitutions, such as replacement of an amino acid having a basic side chain with an amino acid having an acidic side chain, would be more likely to compromise protein function than would substitutions that retain the chemical characteristics of the original amino acid.

Some mutations do not affect the amino acid composition of a protein but still generate an abnormal phenotype. As will be discussed in Chapter 8, such mutations change the amount of normal polypeptide produced by disrupting the biochemical processes responsible for decoding a gene into a polypeptide.

### essential concepts

- Most genes specify the linear sequence of amino acids in a *polypeptide*; this sequence determines the polypeptide's three-dimensional structure and thus its function.
- A *missense mutation* changes the identity of a single amino acid in a polypeptide.
- *Multimeric proteins* include two or more polypeptides (*subunits*). If these subunits are different, they must be encoded by different genes.

## 7.6 A Comprehensive Example: Mutations That Affect Vision

### learning objectives

1. Describe the functions of the four photoreceptor proteins in human vision.
2. Outline how the genes encoding the photoreceptors evolved through duplication and divergence of an ancestral gene.
3. Explain how mutations in the photoreceptor genes result in different vision defects.

Researchers first described anomalies of color perception in humans close to 200 years ago. Since that time, they have discovered a large number of mutations that modify human vision. By examining the phenotype associated with each mutation and then looking directly at the DNA alterations inherited with the mutation, they have learned a great deal about the genes influencing human visual perception and the function of the proteins they specify.

Using human subjects for vision studies has several advantages. First, people can recognize and describe variations in the way they see, from trivial differences in what

the color red looks like, to not seeing any difference between red and green, to not seeing any color at all. Second, the highly developed science of *psychophysics* provides sensitive, noninvasive tests for accurately defining and comparing phenotypes. Finally, because inherited variations in the visual system rarely affect one's life span or ability to reproduce, mutations generating many of the new alleles that change visual perception remain in a population over time.

### Cells of the Retina Contain Light-Sensitive Proteins

People perceive light through nerve cells (neurons) in the retina at the back of the eye (Fig. 7.32a). These neurons are of two types: *rods* and *cones*. The rods, which make up 95% of all light-receiving neurons, are stimulated by weak light over a range of wavelengths. At higher light intensities, the rods become saturated and no longer send meaningful information to the brain. This is when the cones take over, processing wavelengths of bright light that enable us to see color.

The cones come in three forms—one specializes in the reception of red light, a second in the reception of green, and a third in the reception of blue. For each photoreceptor cell, the act of reception consists of absorbing photons from light of a particular wavelength, transducing information about the number and energy of those photons to electrical signals, and transmitting the signals via the optic nerve to the brain. The brain integrates the information from the three types of cones and enables humans to discriminate more than 1 million colors.

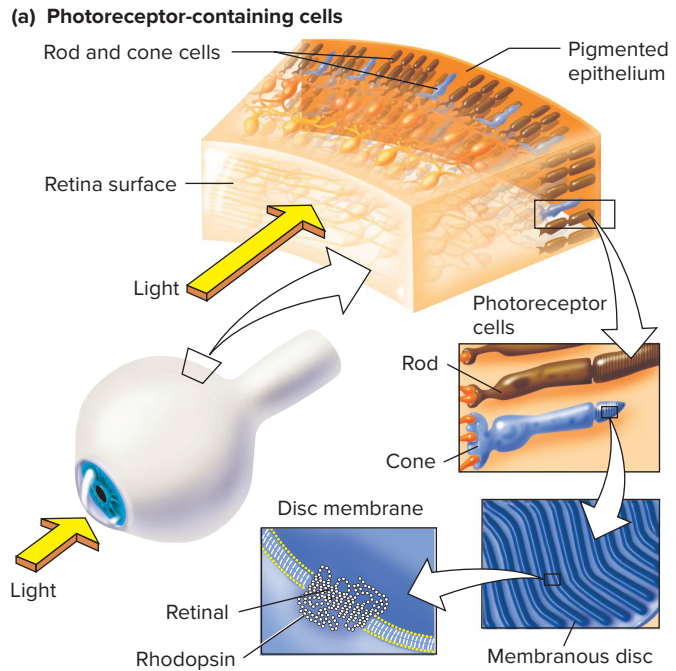
### Four related proteins with different light sensitivities

The protein that receives photons and triggers the processing of information in rod cells is *rhodopsin*. It consists of a single polypeptide chain containing 348 amino acids that snakes back and forth across the cell membrane (Fig. 7.32b). One lysine within the chain associates with *retinal*, a carotenoid pigment molecule that actually absorbs photons. The amino acids in the vicinity of the retinal constitute rhodopsin's active site; by positioning the retinal in a particular way, those amino acids determine its response to light. Each rod cell contains approximately 100 million molecules of rhodopsin in its specialized membrane. As you learned at the beginning of this chapter, the gene governing the production of rhodopsin is on chromosome 3.

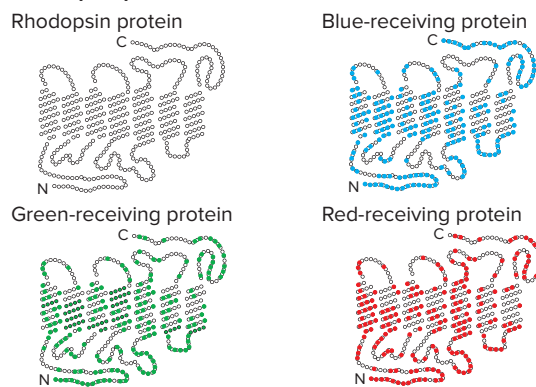
The protein that receives and initiates the processing of photons in the blue cones is a relative of rhodopsin, also consisting of a single polypeptide chain containing 348 amino acids and also encompassing one molecule of retinal. Slightly less than half of the 348 amino acids in the

**Figure 7.32** The cellular and molecular basis of vision.

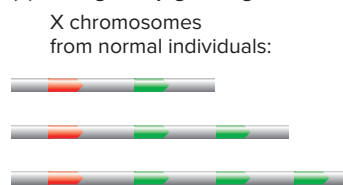
(a) Rod and cone cells in the retina carry membrane-bound photoreceptors. (b) The photoreceptor in rod cells is rhodopsin. The blue, green, and red receptor proteins in cone cells are related to rhodopsin. The colored dots are amino acids that differ between rhodopsin and the diagrammed protein. (c) One red photoreceptor gene and one to three green photoreceptor genes are clustered on the X chromosome. (d) The genes for rhodopsin and the three color receptors probably evolved from a primordial photoreceptor gene through three gene duplication events followed by divergence of the duplicated copies.



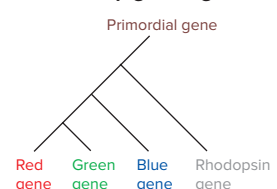
**(b) Photoreceptor proteins**



**(c) Red/green pigment genes**



**(d) Evolution of visual pigment genes**



blue-receiving protein are the same as those found in rhodopsin; the rest are different and account for the specialized light-receiving ability of the protein (Fig. 7.32b). The gene for the blue protein is on chromosome 7.

Similarly related to rhodopsin are the red- and green-receiving proteins in the red and green cones. These are also single polypeptides associated with retinal and embedded in the cell membrane, although they are both slightly larger at 364 amino acids in length (Fig. 7.32b). Like the blue protein, the red and green proteins differ from rhodopsin in nearly half of their amino acids; they differ from each other in only 15 of their 364 amino acids. Even these small differences, however, are sufficient to differentiate the spectral sensitivities of red and green cone cells. The genes for the red and green proteins both reside on the X chromosome in a tandem head-to-tail arrangement. Most people have one red gene and one to three green genes on their X chromosomes (Fig. 7.32c).

### Evolution of the rhodopsin gene family

The similarities in structure and function among rhodopsin and the three rhodopsin-related photoreceptor proteins suggest that the genes encoding these polypeptides arose by a series of gene duplication events in which the duplicated copies subsequently diverged through the accumulation of mutations. Many of the mutations that promoted the ability to see color must have provided selective advantages to their bearers.

Biologists can infer the evolutionary history of these duplications from the relatedness of the genes and protein products. The red and green genes are the most similar, differing by fewer than five nucleotides out of every hundred. This fact suggests they diverged from each other only in the relatively recent evolutionary past. The less pronounced amino acid similarity of the red or green proteins with the blue protein, and the even lower relatedness between rhodopsin and any color photoreceptor, reflect earlier duplication and divergence events (Fig. 7.32d).

### How Mutations in the Rhodopsin Gene Family Affect the Way We See

Mutations in the genes encoding rhodopsin and the three color photoreceptor proteins can alter vision through many different mechanisms. These mutations range from point mutations that change the identity of a single amino acid in a single protein to larger aberrations that can increase or decrease the number of photoreceptor genes.

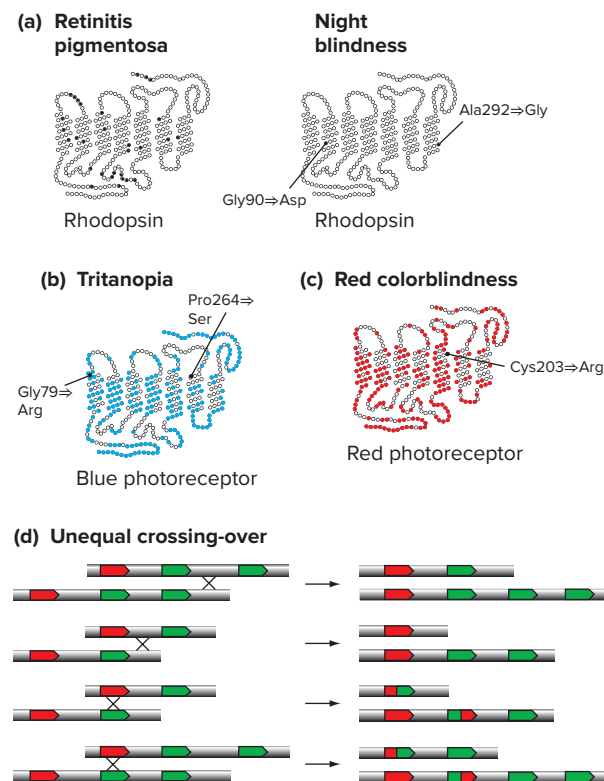
#### Mutations in the rhodopsin gene

At least 29 different single nucleotide substitutions in the rhodopsin gene cause an autosomal dominant vision disorder known as *retinitis pigmentosa* that begins with an

early loss of rod function, followed by a slow progressive degeneration of the peripheral retina. Figure 7.33a shows the location of the amino acids affected by these mutations. These amino acid changes result in abnormal rhodopsin proteins that either do not fold properly or, once folded, are unstable. Although normal rhodopsin is an essential structural element of rod cell membranes, these nonfunctional mutant proteins are retained in the body of the cell, where they remain unavailable for insertion into the membrane. Rod cells that cannot incorporate enough rhodopsin into their membranes eventually die. Depending on how many rod cells die, partial or complete blindness ensues.

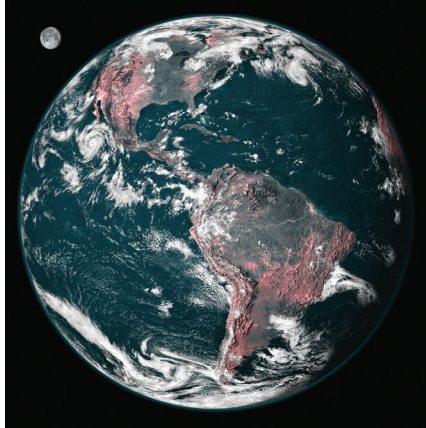
Other mutations in the rhodopsin gene cause the far less serious condition of night blindness (Fig. 7.33a). These mutations change the protein's amino acid sequence so that the threshold of stimulation required to trigger the vision cascade increases. With the changes, very dim light is no longer enough to initiate vision.

**Figure 7.33** How mutations modulate light and color perception. (a) Amino acid substitutions (black dots) that disrupt rhodopsin's three-dimensional structure result in retinitis pigmentosa. Other substitutions diminishing rhodopsin's sensitivity to light cause night blindness. (b) Substitutions in the blue pigment can produce tritanopia (blue color blindness). (c) Red color blindness can result from particular mutations that destabilize the red photoreceptor. (d) Unequal crossing-over between the red and green genes can change gene number and create genes that specify hybrid photoreceptor proteins.



**Figure 7.34** How the world looks to a person with **tritanopia**. Compare with Fig. 4.22.

Color deficit simulation courtesy of Vischeck ([www.vischeck.com](http://www.vischeck.com)). Source image courtesy of NASA



### Mutations in the cone cell pigment genes

Vision problems caused by mutations in the cone cell pigment genes are less severe than those caused by similar defects in the rod cell rhodopsin gene. Most likely, this difference occurs because the rods make up 95% of a person's light-receiving neurons, while the cones constitute only about 5%. Some mutations in the blue gene on chromosome 7 cause *tritanopia*, a defect in the ability to discriminate between colors that differ only in the amount of blue light they contain (Figs. 7.33b and 7.34). Mutations in the red gene on the X chromosome can modify or abolish red protein function and as a result, the red cone cells' sensitivity to light. For example, a change at position 203 in the red-receiving protein from cysteine to arginine disrupts one of the disulfide bonds required to support the protein's tertiary structure (see Fig. 7.33c). Without that bond, the protein cannot stably maintain its native configuration, and a person with the mutation has red color blindness.

### Unequal crossing-over between the red and green genes

People with normal color vision have a single red gene; some of these normal individuals also have a single adjacent green gene, while others have two or even three green genes. The red and green genes are 96% identical in DNA sequence; the different green genes, 99.9% identical.

Their proximity and high degree of homology make these genes unusually prone to an error in meiotic recombination called **unequal crossing-over**. When homologous chromosomes associate during meiosis, two closely related DNA sequences that are adjacent to each other, like the red and green photoreceptor genes, can pair with each other incorrectly. If recombination takes place between the mispaired sequences, photoreceptor genes may be deleted, added, or changed.

A variety of unequal recombination events produce DNA containing no red gene, no green gene, various combinations of green genes, or hybrid red-green genes (see Fig. 7.33d). These different DNA combinations account for the large majority of the known aberrations in red-green color perception, with the remaining abnormalities stemming from point mutations, as described earlier. Because the accurate perception of red and green depends on the differing ratios of red and green light processed, people with no red or no green gene perceive red and green as the same color (see Fig. 4.22).

#### essential concepts

- The vision pigments in humans consist of the protein rhodopsin in rods plus the blue-, red-, and green-sensitive photoreceptors in cones.
- The four genes of the rhodopsin family evolved from an ancestral photoreceptor gene by successive rounds of gene duplication and divergence.
- Mutations in the rhodopsin gene may disrupt rod function, leading to blindness. Mutations in cone cell photoreceptor genes are responsible for various forms of color blindness.

### WHAT'S NEXT

Careful studies of mutations showed that genes are linear arrays of mutable elements that direct the assembly of amino acids in a polypeptide. The mutable elements are the nucleotide building blocks of DNA.

Biologists call the parallel between the sequence of nucleotides in a gene and the order of amino acids in a

polypeptide **colinearity**. In Chapter 8, we explain how colinearity arises from base pairing, a genetic code, specific enzymes, and macromolecular assemblies like ribosomes that guide the flow of information from DNA through RNA to protein.


**SOLVED PROBLEMS**

I. Imagine that 10 independently isolated recessive lethal mutations ( $l^1, l^2, l^3$ , etc.) map to chromosome 7 in mice. You perform complementation testing by mating all pairwise combinations of heterozygotes bearing these lethal mutations, and you score the absence of complementation by examining pregnant females for dead fetuses. A + in the chart means that the two lethals complemented, and dead embryos were not found. A – indicates that dead embryos were found, at the rate of about one in four conceptions. (The crosses between heterozygous mice would be expected to yield the homozygous recessive showing the lethal phenotype in 1/4 of the embryos.) The lethal mutation in the parental heterozygotes for each cross are listed across the top and down the left side of the chart (that is,  $l^1$  indicates a heterozygote in which one chromosome bears the  $l^1$  mutation and the homologous chromosome is wild type).

	$l^1$	$l^2$	$l^3$	$l^4$	$l^5$	$l^6$	$l^7$	$l^8$	$l^9$	$l^{10}$
$l^1$	–	+	+	+	+	–	–	+	+	+
$l^2$		–	+	+	+	+	+	+	+	–
$l^3$			–	–	–	+	+	–	–	+
$l^4$				–	–	+	+	–	–	+
$l^5$					–	+	+	–	–	+
$l^6$						–	–	+	+	+
$l^7$							–	+	+	+
$l^8$								–	–	+
$l^9$									–	+
$l^{10}$										–

How many genes do the 10 lethal mutations represent? What are the complementation groups?

### Answer

This problem involves the application of the complementation concept to a set of data. There are two ways to analyze these results. You can focus on the mutations that do complement each other, conclude that they are in different genes, and begin to create a list of mutations in separate genes. Alternatively, you can focus on mutations that do not complement each other and therefore are alleles of the same genes. The latter approach is more efficient when several mutations are involved. For example,  $l^1$  does not complement  $l^6$  and  $l^7$ . These three alleles are in one complementation group.  $l^2$  does not complement  $l^{10}$ ; they are in a second complementation group.  $l^3$  does not complement  $l^4, l^5, l^8$ , or  $l^9$ , so they form a third complementation group. **Three complementation groups exist.** (Note also that for each mutant, the cross between individuals carrying the

same alleles resulted in no complementation because homozygotes for the recessive lethal mutation were generated.) **The three complementation groups consist of (1)  $l^1, l^6, l^7$ ; (2)  $l^2, l^{10}$ ; and (3)  $l^3, l^4, l^5, l^8, l^9$ .**

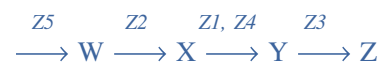
- II. W, X, and Y are the intermediates (in that order) in a biochemical pathway whose product is Z.  $Z^-$  mutants are found in five different complementation groups.  $Z1$  mutants will grow on Y or Z, but not W or X.  $Z2$  mutants will grow on X, Y, or Z.  $Z3$  mutants will only grow on Z.  $Z4$  mutants will grow on Y or Z. Finally,  $Z5$  mutants will grow on W, X, Y, or Z.
- Order the five complementation groups in terms of the steps they block.
  - What does this genetic information reveal about the nature of the enzyme that carries out the conversion of X to Y?

### Answer

This problem requires that you understand complementation and the connection between genes and enzymes in a biochemical pathway.

- A biochemical pathway represents an ordered set of reactions that must occur to produce a product. This problem gives the order of intermediates in a pathway for producing product Z. The lack of any enzyme along the way will cause the phenotype of  $Z^-$ , but the block can occur at different places along the pathway. If the mutant grows when given an intermediate compound, the enzymatic (and hence gene) defect must be before production of that intermediate compound.

The  $Z1$  mutants that grow on Y or Z (but not on W or X) must have a defect in the enzyme that produces Y.  $Z2$  mutants have a defect prior to X;  $Z3$  mutants have a defect prior to Z;  $Z4$  mutants have a defect prior to Y;  $Z5$  mutants have a defect prior to W. **The five complementation groups can be placed in order of activity within the biochemical pathway as follows:**



- Mutants  $Z1$  and  $Z4$  affect the same step, but because they are in different complementation groups, we know they are in different genes. **Mutations  $Z1$  and  $Z4$  are probably in genes that encode subunits of a multisubunit enzyme that carries out the conversion of X to Y.** Alternatively, a currently unknown additional intermediate step between X and Y could exist.

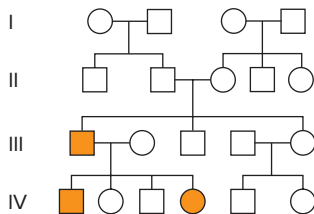

**PROBLEMS**
**Vocabulary**

1. The following is a list of mutational changes. For each of the specific mutations described, indicate which of the terms in the right-hand column applies, either as a description of the mutation or as a possible cause. More than one term from the right column can apply to each statement in the left column.

- |  |                       |
|--|-----------------------|
| 1. an A–T base pair in the wild-type gene is changed to a G–C pair     | a. transition         |
| 2. an A–T base pair is changed to a T–A pair                           | b. base substitution  |
| 3. the sequence AAGCTTATCG is changed to AAGCTATCG                     | c. transversion       |
| 4. the sequence CAGCAGCAGCAGCAGCAG is changed to CAGCAGCAGCAGCAGCAGCAG | d. deletion           |
| 5. the sequence AACGTTATCG is changed to AATGTTATCG                    | e. insertion          |
| 6. the sequence AACGTCACACACATCG is changed to AACGTCACATCG            | f. deamination        |
| 7. the sequence AAGCTTATCG is changed to AAGCTTATCG                    | g. X-ray irradiation  |
|  | h. intercalator       |
|  | i. slipped mispairing |

**Section 7.1**

2. What explanations can account for the following pedigree of a very rare trait? Be as specific as possible. How might you be able to distinguish between these explanations?



3. The DNA sequence of one strand of a gene from three independently isolated mutants is given here (5' ends are at left). Using this information, what is the sequence of the wild-type gene in this region?

mutant 1	ACCGTAATCGACTGGTAAACTTTGCGCG
mutant 2	ACCGTAGTCGACCGGTAACCTTTGCGCG
mutant 3	ACCGTAGTCGACTGGTAACTTTGCGCG

4. Among mammals, measurements of the rate of generation of autosomal recessive mutations have been made almost exclusively in mice, while many measurements of the rate of generation of dominant mutations have been made both in mice and in humans. What do you think is the reason for this difference?

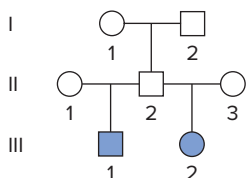
5. Over a period of several years, a large hospital kept track of the number of births of babies displaying the trait achondroplasia. Achondroplasia is a very rare autosomal dominant condition resulting in dwarfism with abnormal body proportions. After 120,000 births, it was noted that 27 babies had been born with achondroplasia. One physician was interested in determining how many of these dwarf babies resulted from new mutations and whether the apparent mutation rate in this geographical area was higher than normal. He looked up the families of the 27 dwarf births and discovered that four of the dwarf babies had a dwarf parent. What is the apparent mutation rate of the achondroplasia gene in this population? Is it unusually high or low?

6. Suppose you wanted to study genes controlling the structure of bacterial cell surfaces. You decide to start by isolating bacterial mutants resistant to infection by a bacteriophage that binds to the cell surface. The selection procedure is simple: Spread cells from a culture of sensitive bacteria on a petri plate, expose them to a high concentration of phages, and pick the bacterial colonies that grow. To set up the selection you could (1) spread cells from a single liquid culture of sensitive bacteria on many different plates and pick every resistant colony; *or* (2) start many different cultures, each grown from a single colony of sensitive bacteria, spread one plate from each culture, and then pick a single mutant from each plate. Which method would ensure that you are isolating many independent mutations?

7. In a genetics lab, Kim and Maria infected a sample from an *E. coli* culture with a particular virulent bacteriophage. They noticed that most of the cells were lysed, but a few survived. The survival rate in their sample was about  $1 \times 10^{-4}$ . Kim was sure the bacteriophage induced the resistance in the cells, while Maria thought that resistant mutants probably already existed in the sample of cells they used. Earlier, for a different experiment, they had spread a dilute suspension of *E. coli* onto solid medium in a large petri dish, and, after seeing that about  $10^5$  colonies were growing up, they had replica-plated that plate onto three other plates. Kim and Maria decide to use these plates to test their theories. They pipette a suspension of the bacteriophage onto each of the three replica plates. What should they see if Kim is right? What should they see if Maria is right?

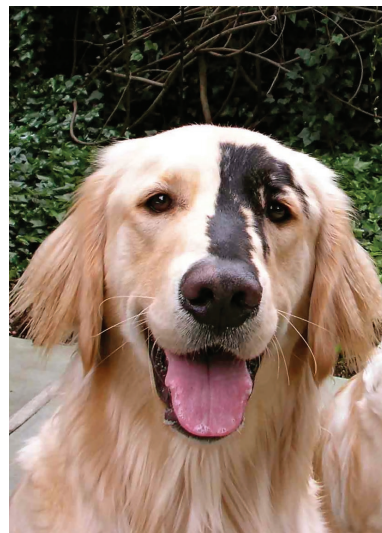


8. The results of the fluctuation test (Fig. 7.5) were interpreted to mean that different numbers of mutant bacteria preexisted in each of the 11 culture tubes because the mutations arose spontaneously at different times during the growth of each culture. However, another possibility is that the differences in the number of colonies on the plates are simply due to differences in the ability of the petri plates to support the growth of colonies. For example, perhaps the selective agent or the nutrients in the media were not evenly distributed in the molten agar poured into the petri dishes. What experiment could you do to determine whether or not differences in the petri plates were a factor in the experiment?
9. The following pedigree shows the inheritance of a completely penetrant, dominant trait called *amelogenesis imperfecta* that affects the structure and integrity of the teeth. DNA analysis of blood obtained from affected individuals III-1 and III-2 shows the presence of the same disease-causing mutation in one of the two copies of an autosomal gene called *ENAM* that is not seen in DNA from the blood of any of the parents in generation II. Explain this result, citing Fig. 4.19 and Fig. 7.5. Do you think this type of inheritance pattern is rare or common?



10. Autism is a neurological disorder thought to be caused by mutant alleles of one or more genes. Scientists had been wondering why the number of children diagnosed as autistic increased dramatically in a decade, from 1 in 500 in 2002 to 1 in 88 in 2012. Researchers now think that they might have found at least part of the answer: Men are fathering children at later and later ages. A paper published in the journal *Nature* in 2012 showed a correlation between paternal age and the incidence of autism; the age of the mother was not a factor. How does this observation provide a possible explanation for the apparent increase in the rate of autism?
11. Like the yellow Labrador retrievers featured in Chapter 3, golden retrievers are usually solid yellow. The golden retriever shown has an extremely rare black marking on its face. Explain the genetic basis for the appearance of this dog. Consider only the

*E* and *B* genes listed in Table 3.3 (see also Figs. 3.12 and 3.13). (Also note that the black marking is not a mask.)



© Sally MacBurney

## Section 7.2

12. Remember that *Balancer* chromosomes prevent the recovery of recombinant chromosomes between the *Balancer* and its normal homolog. Why was the *Balancer X* chromosome crucial to the design of Muller's experiment (Fig. 7.13)? (*Hint*: The best way to answer this question is to consider what the experimental results would have been without the *Balancer*.)
13. Figure 7.14 shows examples of base substitutions induced by the mutagens 5-bromouracil, hydroxylamine, ethylmethane sulfonate, and nitrous acid. Which of these mutagens cause transitions, and which cause transversions?
14. Figure 7.14a shows the mutagen 5-bromouracil (5-BU), which can resemble either T or C depending on its tautomeric state. The figure first shows 5-BU incorporated into DNA as the T-like tautomer, but then it changes its state to the C-like tautomer during the next round of DNA replication. The result was a T:A→C:G substitution. Suppose that the tautomeric states of 5-BU during the two rounds of replication were reversed. What kind of mutation would result?
15. So-called *two-way mutagens* can induce both a particular mutation and (when added subsequently to cells whose chromosomes carry this mutation) a reversion of the mutation that restores the original DNA sequence. In contrast, *one-way mutagens* can induce mutations but not exact reversions of these

mutations. Based on Fig. 7.14, which of the following mutagens can be classified as one-way and which as two-way?

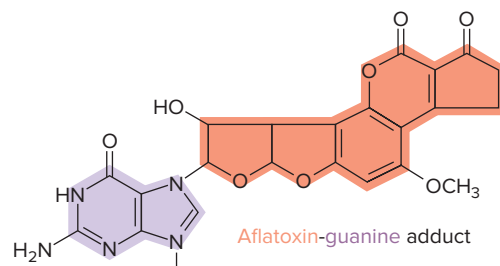
- 5-bromouracil
  - hydroxylamine
  - ethylmethane sulfonate
  - nitrous acid
  - proflavin
16. In 1967, J. B. Jenkins treated wild-type male *Drosophila* with the mutagen ethylmethane sulfonate (EMS) and mated them with females homozygous for a recessive mutation called *dumpy* that causes shortened wings. He found some F<sub>1</sub> progeny with two wild-type wings, some with two short wings, and some with one short wing and one wild-type wing. In a second cross, when he mated single F<sub>1</sub> flies with two short wings to *dumpy* homozygotes, he found, surprisingly, that only a fraction of these matings produced all short-winged progeny.
- Explain these results in light of the mechanism of action of EMS shown in Fig. 7.14.
  - Should the short-winged progeny of the second cross have one or two short wings? Why?
17. When a particular mutagen identified by the Ames test is injected into mice, it causes the appearance of many tumors, showing that this substance is carcinogenic. When cells from these tumors are injected into other mice not exposed to the mutagen, almost all of the new mice develop tumors. However, when mice carrying mutagen-induced tumors are mated to unexposed mice, virtually all of the progeny are tumor free. Why can the tumor be transferred horizontally (by injecting cells) but not vertically (from one generation to the next)?
18. When the His<sup>-</sup> *Salmonella* strain used in the Ames test is exposed to substance X, no His<sup>+</sup> revertants are seen. If, however, rat liver supernatant is added to the cells along with substance X, revertants do occur. Is substance X a potential carcinogen for human cells? Explain.
19. The Ames test uses the reversion rate (His<sup>-</sup> to His<sup>+</sup>) to test compounds for mutagenicity.
- Is it possible that a known mutagen, like proflavin, would be unable to revert a particular His<sup>-</sup> mutant used in the Ames test? How do you think that the Ames test is designed to deal with this issue?
  - Can you think of a way to use forward mutation (His<sup>+</sup> to His<sup>-</sup>) to test a compound for mutagenicity? (*Hint*: Consider using the replica plating technique in Fig. 7.6.)
  - Given that the rate of forward mutation is so much higher than the rate of reversion, why does the Ames test use the reversion rate to test for mutagenicity?

Consult the Fast Forward Box *Trinucleotide Repeat Disease: Huntington Disease and Fragile X Syndrome* in considering the following two problems.

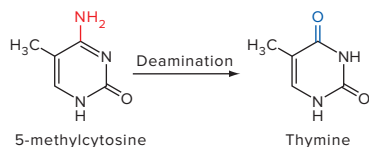
20. The mutant *FMR-1* allele that causes fragile X syndrome is considered to be X-linked dominant with incomplete penetrance and variable expressivity. Why do most females heterozygous for one mutant and one normal allele have at least some symptoms of the disease?
21. The physicist Stephen Hawking, famous for his theories about black holes, has lived past the age of 70 with amyotrophic lateral sclerosis (ALS), a paralyzing neurodegenerative disease that is usually fatal at a much younger age. Recently, geneticists discovered that a major cause of ALS is the unusual expansion of a hexanucleotide repeat (5'-GGGGCC-3') that lies within a gene called *C9ORF72*, at a location outside of the gene's open reading frame (ORF). A single expanded allele is sufficient to cause ALS, but the reason the disease allele is dominant remains unclear. Some experimental results support the theory that the allele makes a toxic RNA containing the expanded repeat. If this theory is correct, in what ways is the mutant ALS-causing allele similar to the mutant allele that causes Huntington disease? In what ways is it similar to the mutant allele that causes fragile X syndrome?

### Section 7.3

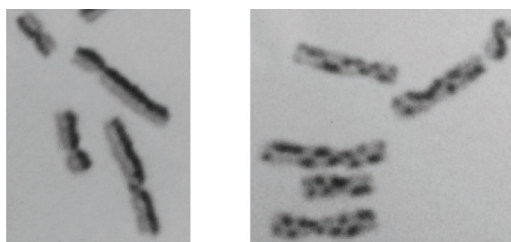
22. Aflatoxin B<sub>1</sub> is a highly mutagenic and carcinogenic compound produced by certain fungi that infect crops such as peanuts. Aflatoxin is a large, bulky molecule that chemically bonds to the base guanine (G) to form the aflatoxin-guanine *adduct* that is pictured below. (In the figure, the aflatoxin is *orange*, and the guanine base is *purple*.) This adduct distorts the DNA double helix and blocks replication.
- What type(s) of DNA repair system is (are) most likely to be involved in repairing the damage caused by exposure of DNA to aflatoxin B<sub>1</sub>?
  - Recent evidence suggests that the adduct of guanine and aflatoxin B<sub>1</sub> can attack the bond that connects it to deoxyribose; this liberates the adducted base, forming an apurinic site. How does this new information change your answer to part (a)?



23. In human DNA, 70% of cytosine residues that are followed by guanine (so-called *CpG dinucleotides*, where *p* indicates the phosphate in the phosphodiester bond between these two nucleotides) are methylated to form 5-methylcytosine. As shown in the following figure, if 5-methylcytosine should undergo spontaneous deamination, it becomes thymine.



- Methylated CpG dinucleotides are hotspots for point mutations in human DNA. Can you propose a hypothesis that explains why?
  - Making the simplifying assumptions that human DNA has an equal number of C–G and A–T base pairs, and that the human DNA sequence is random, how frequently in the human genome would you expect to find the base sequence CpG?
  - It turns out that, even after taking into account the actual GC content of human DNA (~42%), the frequency of CpG in human DNA is much lower than predicted by the calculation in part (b). Explain why this might be the case.
24. Bromodeoxyuridine (BrdU) is a synthetic nucleoside that can be incorporated into newly synthesized DNA in place of thymine (T). Researchers can add BrdU to medium in which cells are growing, and then they can detect its presence in chromosomes by staining with an antibody specific to BrdU.
- The following figure at the *left* shows metaphase chromosomes isolated from somatic cells grown for some time in the presence of BrdU. Darker staining indicates more BrdU. For how many cell generations have these cells been growing in the presence of BrdU?



Normal chromosomes Harlequin chromosomes  
 © Maureen Sanz, Molloy College, Rockville Centre, NY

- So-called *harlequin chromosomes* (similar to those at the *right* of the figure) can be isolated from cells grown in exactly the same way as those at the left, if the cells are exposed to X-rays a few hours before their chromosomes are prepared for analysis.

The X-rays induce double-strand breaks in the DNA. What kind of repair process are you witnessing in these harlequin chromosomes?

- At what point in the cell cycle did the repair process described in part (b) take place? Explain.
- Bloom syndrome* is an autosomal recessive disease in humans characterized by short stature, distinctive facial features, and an increased risk of cancer. The preceding figure at the *right* actually shows cells of Bloom syndrome patients grown in BrdU for the proper number of cell generations; almost all of these cells display harlequin chromosomes even though the cells were not subjected to X-rays. What do you think is the function of the protein specified by the wild-type allele of the Bloom syndrome gene (called *BLM*<sup>+</sup>)?

### Section 7.4

25. Albinism in animals is caused by recessive mutations in one of several autosomal genes required for synthesis of melanin, a chemical precursor for many skin and eye pigments. Albino animals are often confused with so-called *leucistic* animals that are white due to recessive mutations in a gene required in a different pathway, for example a pathway for development of the cells that produce all skin pigments. Suppose you have two white hummingbirds—a male and female—and they have mated. Assuming that all relevant mutations are rare, autosomal, and recessive to wild-type alleles, what would you expect their progeny to look like under the following conditions:
- They are both albinos.
  - They are both leucistic.
  - One is albino and the other is leucistic.
26. a. In Figure 7.22b, what can you say about the phenotype(s) of the progeny indicated by a +? Explain.  
 b. What about the phenotypes of the progeny indicated by a – in the same figure? Explain.
27. Imagine that you caught a female albino mouse in your kitchen and decided to keep it for a pet. A few months later, while vacationing in Guam, you caught a male albino mouse and decided to take it home for some interesting genetic experiments. You wonder whether the two mice are both albino due to mutations in the same gene. What could you do to find out the answer to this question? Assume that both mutations are recessive.
28. Plant breeders studying genes influencing leaf shape in the plant *Arabidopsis thaliana* identified six independent recessive mutations that resulted in plants that had unusual leaves with serrated rather than smooth edges. The investigators started to perform complementation tests with these mutants, but some

of the tests could not be completed because of an accident in the greenhouse. The results of the complementation tests that could be finished are shown in the table that follows.

	1	2	3	4	5	6
1	-	+	-		+	
2		-				-
3			-	-		
4				-		
5					-	+
6						-

- Exactly what experiment was done to fill in individual boxes in the table with a plus (+) or a minus (-)? What does + represent? What does - represent? Why are some boxes in the table filled in green?
  - Assuming no complications, what do you expect for the results of the complementation tests that were not performed? That is, complete the table by placing a + or a - in each of the blank boxes.
  - How many genes are represented among this collection of mutants? Which mutations are in which genes?
29. In humans, albinism is normally inherited in an autosomal recessive fashion. Figure 3.24c in Chapter 3 shows a pedigree in which two albino parents have several children, none of whom is an albino.
- Interpret this pedigree in terms of a complementation test.
  - It is very rare to find examples of human pedigrees such as Fig. 3.24c that in effect represent a complementation test. The reason is that most genetic conditions in humans are rare, so it is highly unlikely that unrelated people with the same condition would mate. In the absence of complementation testing, what kinds of experiments could be done to determine whether a particular human disease phenotype can be caused by mutations at more than one gene?
  - Complementation testing requires that the two mutations to be tested are both recessive to wild type. Suppose that two dominant mutations cause similar phenotypes. How could you establish whether these mutations affected the same gene or different genes?
30. a. Seymour Benzer's fine structure analysis of the *rII* region of bacteriophage T4 depended in large part on deletion analysis as shown in Fig. 7.25. But to perform such deletion analysis, Benzer had to know which *rII*<sup>-</sup> bacteriophage strains were deletions and which were point mutations.
- How do you think he was able to distinguish *rII*<sup>-</sup> deletions from point mutations?
- b. Figure 7.25c shows Benzer's fine structure map of point mutations in the *rII* region. A key feature of this map is the existence of *hotspots*, which Benzer interpreted as nucleotide pairs that were particularly susceptible to mutation. How could Benzer say that all of the independent mutations in a hotspot were due to mutations of the same nucleotide pair?
31. a. You have a test tube containing 5 ml of a solution of bacteriophages, and you would like to estimate the number of bacteriophages in the tube. Assuming the tube actually contains a total of 15 billion bacteriophages, design a serial dilution experiment that would allow you to estimate this number. Ideally, the final plaque-containing plates you count should contain more than 10 and fewer than 1000 plaques.
- b. When you count bacteriophages by the serial dilution method as in part (a), you are assuming a *plating efficiency* of 100%; that is, the number of plaques on the petri plate represents exactly the number of bacteriophages you mixed with the plating bacteria. Is there any way to test the possibility that only a certain percentage of bacteriophage particles can form plaques (so that the plating efficiency would be less than 100%)? Why is it fair to assume that any plaques are initiated by one rather than multiple bacteriophage particles?
32. You found five T4 *rII*<sup>-</sup> mutants that will not grow on *E. coli* K(λ). You mixed together all possible combinations of two mutants (as indicated in the following chart), added the mixtures to *E. coli* K(λ), and scored for the ability of the mixtures to grow and make plaques (indicated as a + in the chart).
- |   | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | - | + | + | - | + |
| 2 |   | - | - | + | - |
| 3 |   |   | - | + | - |
| 4 |   |   |   | - | + |
| 5 |   |   |   |   | - |
- How many genes were identified by this analysis?
  - Which mutants belong to the same complementation groups?
33. The *rosy* (*ry*) gene of *Drosophila* encodes an enzyme called xanthine dehydrogenase. Flies homozygous for *ry* mutations exhibit a rosy eye color. Heterozygous females were made that had *ry*<sup>41</sup> *Sb* on one homolog and *Ly ry*<sup>564</sup> on the other homolog, where *ry*<sup>41</sup> and *ry*<sup>564</sup> are two independently isolated alleles of *ry*. *Ly* [*Lyra* (narrow) wings] and *Sb* [*Stubble* (short) bristles] are dominant mutant alleles of genes to the left

and right of *ry*, respectively. These females are now mated to males homozygous for *ry*<sup>41</sup>. Out of 100,000 progeny, 8 have wild-type eyes, *Lyra* wings, and *Stubble* bristles, while the remainder have rosy eyes.

a. What is the order of these two *ry* mutations relative to the flanking genes *Ly* and *Sb*?

b. What is the genetic distance separating *ry*<sup>41</sup> and *ry*<sup>564</sup>?

34. Nine *rII*<sup>-</sup> mutants of bacteriophage T4 were used in pairwise infections of *E. coli* K(λ) hosts. Six of the mutations in these phages are point mutations; the other three are deletions. The ability of the doubly infected cells to produce progeny phages in large numbers is scored in the following chart.

	1	2	3	4	5	6	7	8	9
1	-	-	+	+	-	-	-	+	+
2		-	+	+	-	-	-	+	+
3			-	-	+	-	+	-	-
4				-	+	-	+	-	-
5					-	-	-	+	+
6						-	-	-	-
7							-	+	+
8								-	-
9									-

The same nine mutants were then used in pairwise infections of *E. coli* B hosts. The production of progeny phages that can subsequently lyse *E. coli* K(λ) hosts is now scored. In the table, 0 means the progeny do not produce any plaques on *E. coli* K(λ) cells; - means that only a very few progeny phages produce plaques; and + means that many progeny produce plaques (more than 10 times as many as in the - cases).

	1	2	3	4	5	6	7	8	9
1	-	+	+	+	+	-	-	+	+
2		-	+	+	+	+	-	+	+
3			0	-	+	0	+	+	-
4				-	+	-	+	+	+
5					-	+	-	+	+
6						0	0	-	+
7							0	+	+
8								-	+
9									-

a. Which of the mutants are the three deletions? What criteria did you use to reach your conclusion?

b. If you know that mutation 9 is in the *rIIB* gene, draw the best genetic map possible to explain the data, including the positions of all point mutations and the extent of the three deletions.

c. One uncertainty should remain in your answer to part (b). How could you resolve this uncertainty?

35. In a haploid yeast strain, eight recessive mutations were found that resulted in a requirement for the

amino acid lysine. All the mutations were found to revert at a frequency of about  $1 \times 10^{-6}$  except mutations 5 and 6, which did not revert. Matings were made between *a* and  $\alpha$  cells carrying these mutations. The ability of the resultant diploid strains to grow on minimal medium in the absence of lysine is shown in the following chart (+ means growth and - means no growth.)

	1	2	3	4	5	6	7	8
1	-	+	+	+	+	-	+	-
2	+	-	+	+	+	+	+	+
3	+	+	-	-	-	-	-	+
4	+	+	-	-	-	-	-	+
5	+	+	-	-	-	-	-	+
6	-	+	-	-	-	-	-	-
7	+	+	-	-	-	-	-	+
8	-	+	+	+	+	-	+	-

a. How many complementation groups were revealed by these data? Which point mutations are found within which complementation groups?

The same diploid strains are now induced to undergo sporulation. The vast majority of resultant spores are auxotrophic; that is, they cannot form colonies when plated on minimal medium (without lysine). However, particular diploids can produce rare spores that do form colonies when plated on minimal medium (prototrophic spores). The following table shows whether (+) or not (-) any prototrophic spores are formed upon sporulation of the various diploid cells.

	1	2	3	4	5	6	7	8
1	-	+	+	+	+	-	+	+
2	+	-	+	+	+	+	+	+
3	+	+	-	+	-	+	+	+
4	+	+	+	-	-	-	+	+
5	+	+	-	-	-	-	+	+
6	-	+	+	-	-	-	+	+
7	+	+	+	+	+	+	-	+
8	+	+	+	+	+	+	+	-

b. When prototrophic spores occur during sporulation of the diploids just discussed, what ratio of auxotrophic to prototrophic spores would you generally expect to see in any tetrad containing such a prototrophic spore? Explain the ratio you expect.

c. Using the data from all parts of this question, draw the best map of the eight lysine auxotrophic mutations under study. Show the extent of any deletions involved, and indicate the boundaries of the various complementation groups.

36. In Problem 24, you learned that Bloom syndrome is an autosomal recessive disease characterized by the high frequency of harlequin chromosomes (as

detected after growth in BrdU). These chromosomes are caused by high levels of chromosome breakage followed by repair through homologous recombination. In some patients, every cell has many harlequin chromosomes. In other patients, the majority of cells have many harlequin chromosomes, but about 10% of the cells surprisingly have none.

- What kinds of events produce the 10% of the cells in certain Bloom syndrome patients with no harlequin chromosomes? (*Hint:* Think about recombination.) What does the existence of these cells lacking harlequin chromosomes say about the alleles of the Bloom syndrome gene carried by these patients?
- In what way do Bloom syndrome patients of both classes reflect the results of complementation tests?
- Why does it make sense that the events you described in part (a) might occur in Bloom syndrome patients?
- What is different about the events you described in part (a) from the events that give rise to harlequin chromosomes?
- Could the events you described in part (a) occur during G<sub>1</sub> of the cell cycle? During G<sub>2</sub>?
- The events that give rise to the cells without harlequin chromosomes are very rare, occurring in less than one in a million cell divisions even in Bloom syndrome patients. Surprisingly, however, roughly 10% of the cells in certain patients lack harlequin chromosomes. How can these two statements be true simultaneously?

**Section 7.5**

37. The pathway for arginine biosynthesis in *Neurospora crassa* involves several enzymes that produce a series of intermediates.

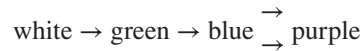


N-acetylornithine → ornithine → citrulline → argininosuccinate → arginine

- If you did a cross between *ARG-E*<sup>-</sup> and *ARG-H*<sup>-</sup> *Neurospora* strains, what would be the distribution of Arg<sup>+</sup> and Arg<sup>-</sup> spores within parental ditype and nonparental ditype asci? Give the spore types in the order in which they would appear in the ascus.
  - For each of the spores in your answer to part (a), what nutrients could you supply in the media to get spore growth?
38. In corn snakes, the wild-type color is brown. One autosomal recessive mutation causes the snake to be orange, and another causes the snake to be black. An orange snake was crossed to a black one, and the F<sub>1</sub> offspring were all brown. Assume that all relevant genes are unlinked.

- Indicate what phenotypes and ratios you would expect in the F<sub>2</sub> generation of this cross if there is one pigment pathway, with orange and black being different intermediates on the way to brown.
- Indicate what phenotypes and ratios you would expect in the F<sub>2</sub> generation if orange pigment is a product of one pathway, black pigment is the product of another pathway, and brown is the effect of mixing the two pigments in the skin of the snake.

39. In a certain species of flowering plants with a diploid genome, four enzymes are involved in the generation of flower color. The genes encoding these four enzymes are on different chromosomes. The biochemical pathway involved is as follows; the figure shows that either of two different enzymes is sufficient to convert a blue pigment into a purple pigment.



A true-breeding green-flowered plant is mated with a true-breeding blue-flowered plant. All of the plants in the resultant F<sub>1</sub> generation have purple flowers. F<sub>1</sub> plants are allowed to self-fertilize, yielding an F<sub>2</sub> generation. Show genotypes for P, F<sub>1</sub>, and F<sub>2</sub> plants, and indicate which genes specify which biochemical steps. Determine the fraction of F<sub>2</sub> plants with the following phenotypes: white flowers, green flowers, blue flowers, and purple flowers. Assume the green-flowered parent is mutant in only a single step of the pathway.

40. The intermediates A, B, C, D, E, and F all occur in the same biochemical pathway. G is the product of the pathway, and mutants 1 through 7 are all G<sup>-</sup>, meaning that they cannot produce substance G. The following table shows which intermediates will promote growth in each of the mutants. Arrange the intermediates in order of their occurrence in the pathway, and indicate the step in the pathway at which each mutant strain is blocked. A + in the table indicates that the strain will grow if given that substance, an O means lack of growth.

Mutant	Supplements						
	A	B	C	D	E	F	G
1	+	+	+	+	+	O	+
2	O	O	O	O	O	O	+
3	O	+	+	O	+	O	+
4	O	+	O	O	+	O	+
5	+	+	+	O	+	O	+
6	+	+	+	+	+	+	+
7	O	O	O	O	+	O	+

41. In each of the following cross schemes, two true-breeding plant strains are crossed to make F<sub>1</sub> plants, all of which have purple flowers. The F<sub>1</sub> plants are

then self-fertilized to produce F<sub>2</sub> progeny as shown here.

Cross	Parents	F <sub>1</sub>	F <sub>2</sub>
1	blue × white	all purple	9 purple: 4 white: 3 blue
2	white × white	all purple	9 purple: 7 white
3	red × blue	all purple	9 purple: 3 red: 3 blue: 1 white
4	purple × purple	all purple	15 purple: 1 white

- For each cross, explain the inheritance of flower color.
  - For each cross, show a possible biochemical pathway that could explain the data.
  - Which of these crosses is compatible with an underlying biochemical pathway involving only a single step that is catalyzed by an enzyme with two dissimilar subunits, both of which are required for enzyme activity?
  - For each of the four crosses, what would you expect in the F<sub>1</sub> and F<sub>2</sub> generations if all relevant genes were tightly linked?
42. The pathways for the biosynthesis of the amino acids glutamine (Gln) and proline (Pro) involve one or more common intermediates. Auxotrophic yeast mutants numbered 1–7 are isolated that require either glutamine or proline or both amino acids for their growth, as shown in the following table (+ means growth; – no growth). These mutants are also tested for their ability to grow on the intermediates A–E. What is the order of these intermediates in the glutamine and proline pathways, and at which point in the pathways is each mutant blocked?

Mutant	A	B	C	D	E	Gln	Pro	Gln + Pro
1	+	–	–	–	+	–	+	+
2	–	–	–	–	–	–	+	+
3	–	–	+	–	–	–	–	+
4	–	–	–	–	–	+	–	+
5	–	–	+	+	–	–	–	+
6	+	–	–	–	–	–	+	+
7	–	+	–	–	–	+	–	+

43. The following complementing *E. coli* mutants were tested for growth on four known precursors of thymine, A–D.

Mutant	Precursor/product				Thymine
	A	B	C	D	
9	+	–	+	–	+
10	–	–	+	–	+
14	+	+	+	–	+
18	+	+	+	+	+
21	–	–	–	–	+

- Show a simple linear biosynthetic pathway of the four precursors and the end product, thymine. Indicate which step is blocked by each of the five mutations.
- What precursor would accumulate in the following double mutants: 9 and 10? 10 and 14?

44. In 1952, an article in the *British Medical Journal* reported interesting differences in the behavior of blood plasma obtained from several people who suffered from X-linked recessive hemophilia. When mixed together, the cell-free blood plasma from certain combinations of individuals could form clots in the test tube. For example, the following table shows whether clots could form (+) or not (–) in various combinations of plasma from four people with hemophilia:

1 and 1	–	2 and 3	+
1 and 2	–	2 and 4	+
1 and 3	+	3 and 3	–
1 and 4	+	3 and 4	–
2 and 2	–	4 and 4	–

What do these data tell you about the inheritance of hemophilia in these individuals? Do these data allow you to exclude any models for the biochemical pathway governing blood clotting?

45. Mutations in an autosomal gene in humans cause a form of hemophilia called von Willebrand disease (vWD). This gene specifies a blood plasma protein cleverly called von Willebrand factor (vWF). vWF stabilizes factor VIII, a blood plasma protein specified by the wild-type hemophilia A gene. Factor VIII is needed to form blood clots. Thus, factor VIII is rapidly destroyed in the absence of vWF.

Which of the following might successfully be employed in the treatment of bleeding episodes in hemophiliac patients? Would the treatments work immediately or only after some delay needed for protein synthesis? Would the treatments have only a short-term or a prolonged effect? Assume that all mutations are null (that is, the mutations result in the complete absence of the protein encoded by the gene) and that the plasma is cell-free.

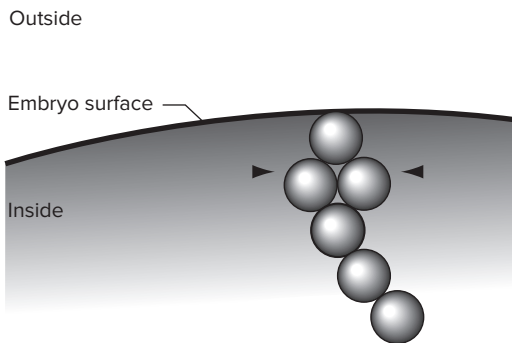
- transfusion of plasma from normal blood into a vWD patient
- transfusion of plasma from a vWD patient into a different vWD patient
- transfusion of plasma from a hemophilia A patient into a vWD patient
- transfusion of plasma from normal blood into a hemophilia A patient
- transfusion of plasma from a vWD patient into a hemophilia A patient

- f. transfusion of plasma from a hemophilia A patient into a different hemophilia A patient
  - g. injection of purified vWF into a vWD patient
  - h. injection of purified vWF into a hemophilia A patient
  - i. injection of purified factor VIII into a vWD patient
  - j. injection of purified factor VIII into a hemophilia A patient
46. Antibodies were made that recognize six proteins that are part of a complex inside the *Caenorhabditis elegans* one-cell embryo. The mother produces proteins that are believed to assemble stepwise into a structure in the egg, beginning at the embryo's inner surface. The antibodies were used to detect the protein location in embryos produced by mutant mothers [who are homozygous recessive for the gene(s) encoding each protein]. The *C. elegans* mothers are self-fertilizing hermaphrodites, so no wild-type copy of a gene will be introduced during fertilization.

In the following table, \* means the protein was present and at the embryo surface, - means that the protein was not present, and + means that the protein was present but not at the embryo surface. Assume all mutations prevent production of the corresponding protein.

Mutant in gene for protein	Protein production and location					
	A	B	C	D	E	F
A	-	+	*	+	*	+
B	*	-	*	*	*	*
C	*	+	-	+	*	+
D	*	+	*	-	*	+
E	+	+	+	+	-	+
F	*	+	*	*	*	-

Complete the following figure, which shows the construction of the hypothetical protein complex, by writing the letter of the proper protein in each circle. The two proteins marked with arrowheads can assemble into the complex independently of each other, but both are needed for the addition of subsequent proteins to the complex.



47. Adult hemoglobin is a multimeric protein with four polypeptides, two of which are  $\alpha$ -globin and two of which are  $\beta$ -globin.
- a. How many genes are needed to define the structure of the hemoglobin protein?
  - b. If a person is heterozygous for wild-type alleles and alleles that would yield amino acid substitution variants for both  $\alpha$ -globin and  $\beta$ -globin, how many different kinds of hemoglobin protein would be found in the person's red blood cells and in what proportion? Assume all alleles are expressed at the same level.
48. Each complementation group (*ARG-E*, *ARG-F*, *ARG-G*, and *ARG-H*) in Fig. 7.27b can grow on a unique subset of supplements. Why were these four subsets the only ones observed? For example, why were no complementation groups observed that behaved like the four hypothetical ones shown in the following table? (Symbols as in Fig. 7.27b: + means growth, - means no growth.)

Hypothetical mutant strain	Supplements				
	Nothing	Ornithine	Citrulline	Arginino-succinate	Arginine
Wildtype: Arg <sup>+</sup>	+	+	+	+	+
<i>ARG-I</i> <sup>-</sup>	-	+	-	+	+
<i>ARG-J</i> <sup>-</sup>	-	-	+	-	+
<i>ARG-K</i> <sup>-</sup>	-	+	-	-	+
<i>ARG-L</i> <sup>-</sup>	-	+	+	-	+

### Section 7.6

49. In addition to the predominant adult hemoglobin, HbA, which contains two  $\alpha$ -globin chains and two  $\beta$ -globin chains ( $\alpha_2\beta_2$ ), there is a minor hemoglobin, HbA<sub>2</sub>, composed of two  $\alpha$  and two  $\delta$  chains ( $\alpha_2\delta_2$ ). The  $\beta$ - and  $\delta$ -globin genes are arranged in tandem and are highly homologous. Draw the chromosomes that would result from an event of unequal crossing-over between the  $\beta$  and  $\delta$  genes.
50. Most mammals, including New World primates such as marmosets (a kind of monkey), are *dichromats*: They have only two kinds of rhodopsin-related color receptors. Old World primates such as humans and gorillas are *trichromats* with three kinds of color receptors. Primates diverged from other mammals roughly 65 million years ago (Myr), while Old World and New World primates diverged from each other roughly 35 Myr.
- a. Using this information, define on Fig. 7.32d the time spans of any events that can be dated.
  - b. Some New World monkeys have an autosomal color receptor gene and a single X-linked color receptor gene. The X-linked gene has three alleles, each of which specifies a photoreceptor that



responds to light of a different wavelength (all three wavelengths are different from that recognized by the autosomal color receptor). How is color vision inherited in these monkeys?

- c. About 95% of all light-receiving neurons in humans and other mammals are rod cells containing rhodopsin, a pigment that responds to low-level light of many wavelengths. The remaining 5% of light-receiving neurons are cone cells with pigments that respond to light of specific wavelengths of high intensity. What do these facts suggest about the lifestyle of the earliest mammals?
51. Humans are normally *trichromats*; we have three different types of retinal cones, each containing either a red, green, or blue rhodopsin-like photoreceptor protein. The reason is that most humans have genes for red and green photoreceptors on the X chromosome, and a blue photoreceptor gene on an autosome. Our brain integrates the information from each type of cone, making it possible for us to see about one million colors.

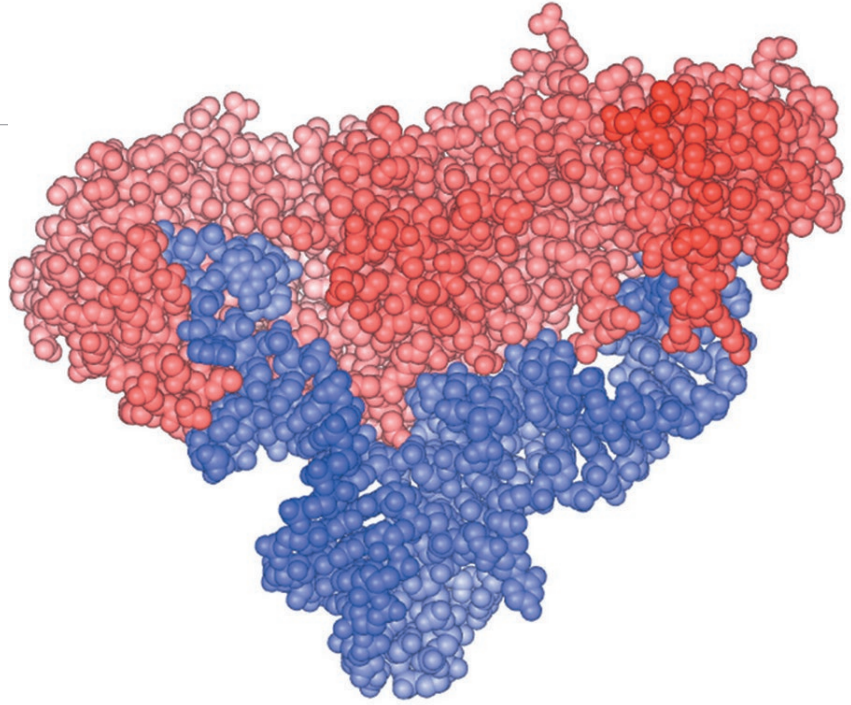
Some scientists think that rare people may be *tetrachromats*, that is, they have four different kinds of

cones. Such people, if they exist, could potentially detect 100 million colors! For parts (a) and (b), assume that each X chromosome has one red and one green photoreceptor protein gene. For all parts, assume that mutant alleles can produce photoreceptors with altered spectral sensitivities.

- a. Explain why scientists expect that many more females than males would be tetrachromats.
- b. In X-linked, red/green color blindness, mutation of either the red or green photoreceptor gene results in a rhodopsin-like protein with altered spectral sensitivity. The mutant photoreceptor is sensitive to wavelengths in between the normal red and green photoreceptors. Why do scientists think that a woman with a son who is red/green color-blind is more likely to be a tetrachromat than a woman whose sons all have normal vision?
- c. Suggest a scenario based on Fig. 7.33d that could explain how extremely rare males might be tetrachromats.

## chapter 8

# Gene Expression: The Flow of Information from DNA to RNA to Protein



*The ability of an aminoacyl-tRNA synthetase (red) to recognize a particular tRNA (blue) and couple it to its corresponding amino acid (not shown) is central to the molecular machinery that converts the language of nucleic acids into the language of proteins.*

## chapter outline

- 8.1 The Genetic Code
- 8.2 Transcription: From DNA to RNA
- 8.3 Translation: From mRNA to Protein
- 8.4 Differences in Gene Expression Between Prokaryotes and Eukaryotes
- 8.5 The Effects of Mutations on Gene Expression and Function

A DEDICATED EFFORT to determine the complete nucleotide sequence of the genome in a variety of organisms has been underway since 1990. This massive endeavor has been more successful than many scientists thought possible. By the time of this writing in 2016, the DNA sequence in the genomes of more than 8100 different species had already been deposited in databases, and sequencing projects for more than 35,000 additional species were in progress. With this sequence information in hand, geneticists can consult the genetic code—the cipher equating nucleotide sequence with amino acid sequence—to decide what parts of a genome are likely to be genes.

As a result, modern geneticists can discover the number and amino acid sequences of all the polypeptides that determine phenotype. Knowledge of DNA sequence thus opens up powerful new possibilities for understanding an organism’s growth and development at the molecular level.

In this chapter, we describe the cellular mechanisms that carry out **gene expression**, the means by which genetic information can be interpreted as phenotype. As intricate as some of the details may appear, the general scheme of gene expression is elegant and straightforward: *Within each cell, genetic information flows from DNA to RNA to protein.* This statement was set forward as the *Central Dogma* of molecular biology by Francis Crick in 1957. As Crick explained, “Once information has passed into protein, it cannot get out again.”

The Central Dogma maintains that genetic information flows in two distinct stages (**Fig. 8.1**). The conversion of the information in DNA to its equivalent in RNA is known as **transcription**. The product of transcription is a **transcript**: a molecule of **messenger RNA (mRNA)** in prokaryotes, a molecule of RNA that undergoes processing to become an mRNA in eukaryotes.

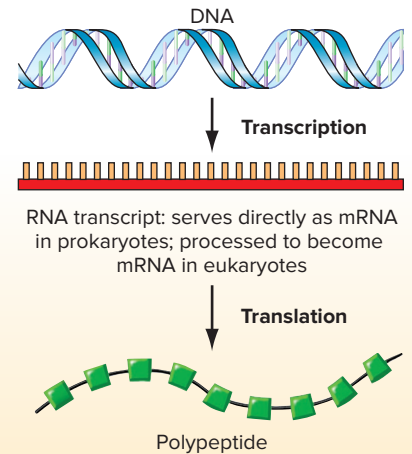
In the second stage of gene expression, the cellular machinery decodes the sequence of nucleotides in mRNA into a sequence of amino acids—a **polypeptide**—by

the process known as **translation**. It takes place on molecular workbenches called **ribosomes**, which are composed of proteins and **ribosomal RNAs (rRNAs)**. Translation depends on the dictionary known as the **genetic code**, which defines each amino acid in terms of specific sequences of three nucleotides. Translation also requires **transfer RNAs (tRNAs)**, small RNA adapter molecules that place specific amino acids at the correct position in a growing polypeptide chain.

The Central Dogma does not explain the behavior of all genes. As Crick himself realized, many genes are transcribed into RNAs that are never translated into proteins. You will see in this chapter that many nontranslated RNAs are critical to various steps of gene expression. The genes encoding rRNAs and tRNAs belong to this group.

Four general themes emerge from our discussion of gene expression. First, the pairing of complementary bases is key to the transfer of information from DNA to RNA, and from RNA to protein. Second, the polarities (directionality) of DNA, RNA, and polypeptides help guide the mechanisms of gene expression. Third, like DNA replication and recombination, gene expression requires an input of energy and the participation of specific proteins, RNAs, and macromolecular assemblies, such as ribosomes. Finally, mutations that change genetic information or obstruct the flow of its expression can have dramatic effects on phenotype.

**Figure 8.1 Gene expression: The flow of genetic information from DNA via RNA to protein.** In transcription, the enzyme RNA polymerase copies DNA to produce an RNA transcript. In translation, the cellular machinery uses instructions in mRNA to synthesize a polypeptide, following the rules of the genetic code.



## 8.1 The Genetic Code

### Learning Objectives

1. Explain the reasoning that established a sequence of three nucleotides (a triplet codon) as the basic unit of the code relating DNA to protein.
2. Summarize the evidence showing that the sequence of nucleotides in a gene is colinear with the sequence of amino acids in a protein.
3. Define *reading frame* and discuss its significance to the genetic code.
4. Describe experiments that determined which codons are associated with each amino acid and which are stop codons.
5. Explain how mutations were used to verify the genetic code.
6. Discuss evidence that the genetic code is almost universal, and cite some exceptions.

A code is a system of symbols that equates information in one language with information in another. A useful analogy for the genetic code is the Morse code, which uses dots and dashes to transmit messages over radio or telegraph wires. Various groupings of the dot-dash symbols represent the 26 letters of the English alphabet. And because many

more letters exist than the two symbols (dot or dash), groups of one, two, three, or four dots or dashes in various combinations represent individual letters. For example, the symbol for C is dash dot dash dot (– · – ·), the symbol for O is dash dash dash (– – –), D is dash dot dot (– · ·), and E is a single dot (·). Because anywhere from one to four symbols specify each letter, the Morse code requires a symbol for *pause* (in practice, a short interval of time) to signify where one letter ends and the next begins.

### Triplet Codons of Nucleotides Represent Individual Amino Acids

The language of nucleic acids is written in four nucleotides—A, G, C, and T in the DNA dialect; A, G, C, and U in the RNA dialect—while the language of proteins is written in amino acids. The first hurdle to be overcome in deciphering how sequences of nucleotides can determine the order of amino acids in a polypeptide is to determine how many amino acid “letters” exist.

Over lunch one day at a local pub, Watson and Crick produced the now accepted list of the 20 common amino acids that are encoded directly by DNA. They created the list by analyzing the known amino acid sequences of a variety of naturally occurring polypeptides. Amino acids present in only a small number of proteins or in only certain tissues or organisms did not qualify as standard building blocks; Crick and Watson correctly assumed

that most such amino acids arise when proteins undergo chemical modification after their synthesis. By contrast, amino acids present in most proteins made the list. The question then became: How can four nucleotides encode 20 amino acids?

Like the Morse code, the four nucleotides encode 20 amino acids through specific groupings of A, G, C, and T (in DNA) or A, G, C, and U (in RNA). Researchers initially arrived at the number of letters per grouping by deductive reasoning and later confirmed their guess by experiment. They reasoned that if only one nucleotide represented an amino acid, information would exist for only four amino acids: A would encode one amino acid, G a second amino acid, and so on. If two nucleotides represented each amino acid,  $4^2 = 16$  possible combinations of doublets would be possible.

Of course, if the code consisted of groups containing one *or* two nucleotides, it would have  $4 + 16 = 20$  groups and could account for all the amino acids, but nothing would be left over to signify the pause required to denote where one group ends and the next begins. Groups of three nucleotides in a row would provide  $4^3 = 64$  different triplet combinations, more than enough to code for all the amino acids. If the code consisted of doublets and triplets, a signal denoting a pause would once again be necessary. But a triplets-only code would require no symbol for *pause* if the mechanism for counting to three and distinguishing among successive triplets was very reliable.

Although this kind of reasoning generates a hypothesis, it does not prove it. As it turned out, however, the experiments described later in this chapter did indeed demonstrate that groups of three nucleotides represent all 20 amino acids. Each nucleotide triplet is called a **codon**. Each codon, designated by the bases defining its three nucleotides, specifies one amino acid. For example, GAA is a codon for glutamic acid (Glu), and GUU is a codon for valine (Val). Because the code comes into play only during the translation part of gene expression, that is, during the decoding of messenger RNA to polypeptide, geneticists usually present the code in the RNA dialect of A, G, C, and U, as depicted in **Fig. 8.2**. When speaking of genes, they can substitute T for U to show the same code in the DNA dialect.

If you knew the sequence of nucleotides in a gene or its transcript as well as the sequence of amino acids in the corresponding polypeptide, you could then deduce the genetic code without understanding how the underlying cellular machinery actually works. Although techniques for determining both nucleotide and amino acid sequence are available today, this was not true when researchers were trying to crack the genetic code in the 1950s and 1960s. At that time, they could establish a polypeptide's amino acid sequence, but not the nucleotide sequence of DNA or RNA. Because of their inability to read nucleotide sequence, scientists used an assortment of genetic and biochemical techniques

**Figure 8.2** The genetic code: 61 codons represent the 20 amino acids, while 3 codons signify stop. To read the code, find the first letter in the *left column*, the second letter along the *top*, and the third letter in the *right column*; this reading corresponds to the 5'-to-3' direction along the mRNA.

		Second letter				
		U	C	A	G	
U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U C A G	
	UUC } Phe	UCC } Ser	UAC } Tyr	UGC } Cys		
	UUA } Leu	UCA } Ser	UAA Stop	UGA Stop		
	UUG } Leu	UCG } Ser	UAG Stop	UGG Trp		
C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U C A G	
	CUC } Leu	CCC } Pro	CAC } His	CGC } Arg		
	CUA } Leu	CCA } Pro	CAA } Gln	CGA } Arg		
	CUG } Leu	CCG } Pro	CAG } Gln	CGG } Arg		
A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U C A G	
	AUC } Ile	ACC } Thr	AAC } Asn	AGC } Ser		
	AUA } Met	ACA } Thr	AAA } Lys	AGA } Arg		
	AUG Met	ACG } Thr	AAG } Lys	AGG } Arg		
G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U C A G	
	GUC } Val	GCC } Ala	GAC } Asp	GGC } Gly		
	GUA } Val	GCA } Ala	GAA } Glu	GGA } Gly		
	GUG } Val	GCG } Ala	GAG } Glu	GGG } Gly		

to fathom the code. They began by examining how different mutations in a single gene affected the amino acid sequence of the gene's polypeptide product. In this way, they were able to use the abnormal (specific mutations) to understand the normal (the general relationship between genes and polypeptides).

### A Gene's Nucleotide Sequence Is Colinear with the Amino Acid Sequence of the Encoded Polypeptide

As you know, DNA is a linear molecule with base pairs following one another down the intertwined chains. Proteins, by contrast, have complicated three-dimensional structures. Even so, if unfolded and stretched out from N terminus to C terminus, proteins have a one-dimensional, linear structure—a specific primary sequence of amino acids. If the information in a gene and its corresponding protein are colinear, the consecutive order of bases in the DNA from the beginning to the end of the gene would stipulate the consecutive order of amino acids from one end to the other of the outstretched protein.

In the 1960s, Charles Yanofsky was the first to compare maps of mutations within a gene to the particular amino acid substitutions that resulted. He began by generating a

large number of  $\text{Trp}^-$  auxotrophic mutants in *E. coli* that carried mutations in the *trpA* gene for a subunit of the enzyme tryptophan synthase. He next made a fine structure recombination map of these mutations analogous to Benzer's fine structure map for the *rII* region of bacteriophage T4, which was discussed in Chapter 7. Yanofsky then purified and determined the amino acid sequences of the mutant tryptophan synthase subunits.

As **Fig. 8.3a** illustrates, Yanofsky's data showed that the order of mutations mapped within the DNA of the gene by recombination was indeed colinear with the positions of the amino acid substitutions occurring in the resulting mutant proteins. By carefully analyzing his results, Yanofsky deduced two other key features of the relationship between nucleotides and amino acids.

### Evidence that a codon is composed of more than one nucleotide

Yanofsky observed that point mutations altering different nucleotide pairs may affect the same amino acid. In one example shown in Fig. 8.3a, mutation 23 changed the glycine (Gly) at position 211 of the wild-type polypeptide chain to arginine (Arg), while mutation 46 yielded glutamic

acid (Glu) at the same position. In another example, mutation 78 changed the glycine at position 234 to cysteine (Cys), while mutation 58 produced aspartic acid (Asp) at the same position. These are all **missense mutations** that change a codon for one amino acid into a codon that specifies a different amino acid.

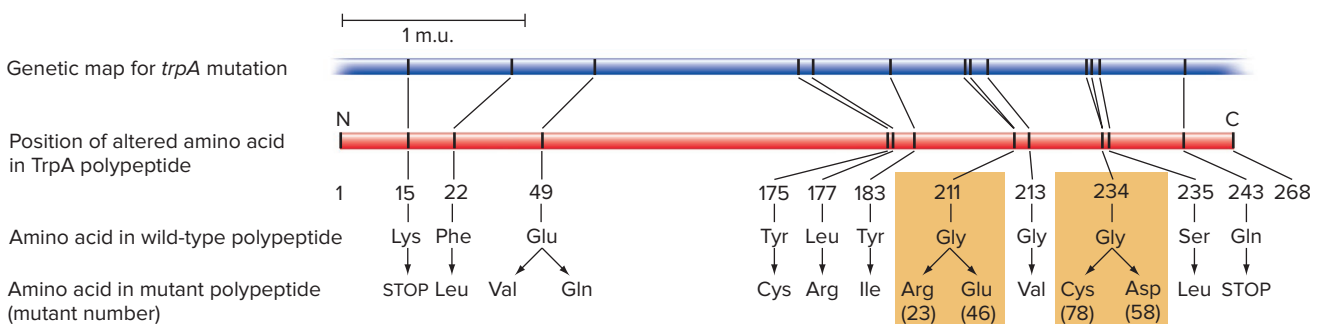
In both cases, Yanofsky found that recombination could occur between the two mutations that changed the identity of the same amino acid; such recombination would produce a wild-type tryptophan synthase gene (**Fig. 8.3b**). Because the smallest unit of recombination is the base pair, two mutations capable of recombination—in this case, in the same codon because they affect the same amino acid—must be in different (although nearby) nucleotides. Thus, a codon must contain more than one nucleotide.

### Evidence that each nucleotide is part of only one codon

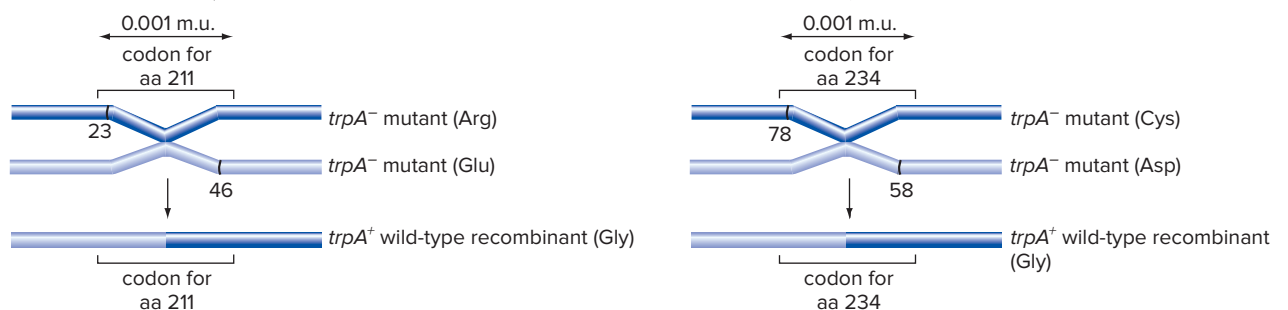
As Fig. 8.3a illustrates, each of the point mutations in the tryptophan synthase gene characterized by Yanofsky alters the identity of only a single amino acid. This is also true of the point mutations examined in many other genes, such as the human genes for rhodopsin and hemoglobin

**Figure 8.3** Mutations in a gene are colinear with the sequence of amino acids in the encoded polypeptide. (a) The relationship between the genetic map of *E. coli*'s *trpA* gene and the positions of amino acid substitutions in mutant tryptophan synthase proteins. (b) Codons must include two or more base pairs. When two mutant strains with different amino acids at the same position were crossed, recombination could produce a wild-type allele.

#### (a) Colinearity of genes and proteins



#### (b) Recombination within a codon



(see Chapter 7). Because point mutations that change only a single nucleotide pair affect only a single amino acid in a polypeptide, each nucleotide in a gene must influence the identity of only a single amino acid. In contrast, if a nucleotide were part of more than one codon, a mutation in that nucleotide would affect more than one amino acid.

### Nonoverlapping Triplet Codons Are Set in a Reading Frame

Although the most efficient code to specify 20 amino acids requires three nucleotides per codon, more complicated scenarios are possible. But in 1955, Francis Crick and Sydney Brenner obtained convincing evidence for the triplet nature of the genetic code in studies of mutations in the bacteriophage T4 *rIIB* gene originally characterized by Seymour Benzer (Chapter 7). They induced the mutations with *proflavin*, an intercalating mutagen that can insert itself between the paired bases stacked in the center of the DNA molecule (recall Fig. 7.14c). Crick and Brenner's original assumption was that proflavin would act like other mutagens, causing single-base substitutions. If this were true, it would be possible to generate revertants through treatment with other mutagens that might restore the wild-type DNA sequence.

Surprisingly, genes with proflavin-induced mutations did not revert to wild-type upon treatment with other mutagens known to cause nucleotide substitutions. Only further exposure to proflavin caused proflavin-induced mutations to revert to wild-type. Crick and Brenner had to explain this observation before they could proceed with their phage experiments. With keen insight, they correctly guessed that proflavin does not cause base substitutions; instead, it causes insertions or deletions of a single base pair. This hypothesis explained why base-substituting mutagens could not cause reversion of proflavin-induced mutations.

#### Evidence for a triplet code

Crick and Brenner began their experiments with a particular proflavin-induced *rIIB*<sup>-</sup> mutation they called FC0. They next treated this mutant strain with more proflavin to isolate an *rIIB*<sup>+</sup> revertant (Fig. 8.4a). By recombining this revertant with wild-type bacteriophage T4, Crick and Brenner were able to show that the revertant's chromosome actually contained two different *rIIB*<sup>-</sup> mutations (Fig. 8.4b). One was the original FC0 mutation; the other was the newly induced FC7. Either mutation by itself yields a mutant phenotype, but their simultaneous occurrence in the same gene yielded an *rIIB*<sup>+</sup> phenotype. Crick and Brenner reasoned that if the first mutation was the addition of a single base pair, represented by the symbol (+), then the counteracting mutation must be the deletion of a base pair, represented as (-). The restoration of gene

function by one mutation canceling another in the same gene is known as **intragenic suppression**.

Crick and Brenner supposed not only that each codon is a trio of nucleotides, but that each gene has a single starting point. This starting point establishes a **reading frame**: the sequential partitioning of nucleotides into groups of three to generate the correct order of amino acids in the resulting polypeptide chain (Fig. 8.4a). Changes that alter the grouping of nucleotides into codons are called **frameshift mutations**; they shift the reading frame for all codons beyond the point of insertion or deletion, almost always abolishing the function of the polypeptide product.

If codons are read in order from a fixed starting point, a deletion (-) can counterbalance an insertion (+) to restore the reading frame (Fig. 8.4a). Note that the gene would regain its wild-type activity only if the portion of the polypeptide encoded between the two mutations of opposite sign is not required for protein function, because in the double mutant, this region would have an improper amino acid sequence. Also, the incorrect amino acids must not prevent the protein from folding into a functional conformation.

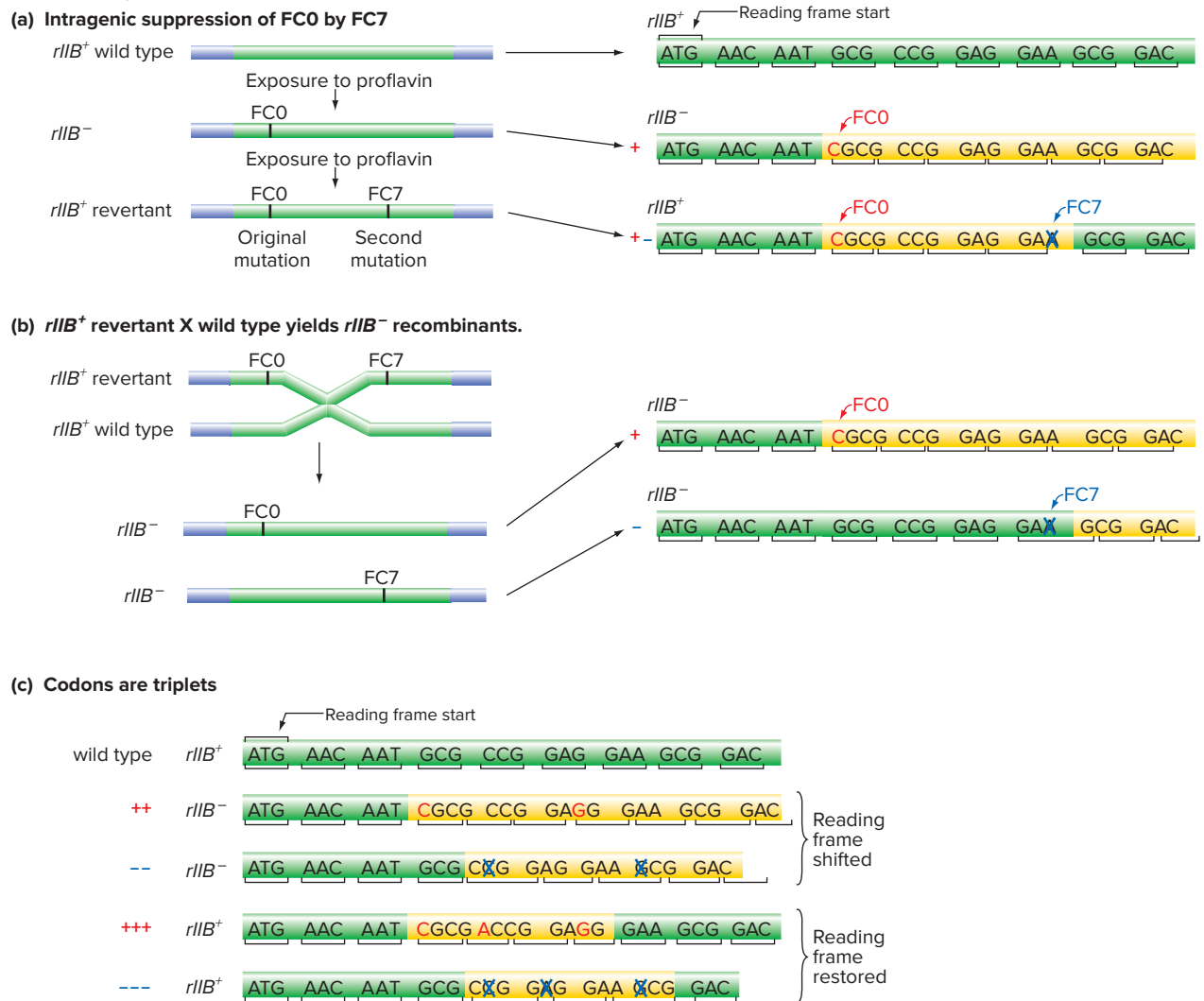
Crick and Brenner realized that they could use + and - mutations in *rIIB* to test the hypothesis that codons were indeed nucleotide triplets. If codons are composed of three nucleotides, then combining two different *rIIB*<sup>-</sup> mutations of the same sign (+ + or - -) in the same gene should never lead to intragenic suppression (an *rIIB*<sup>+</sup> phenotype). Combinations of three + or three - mutations, however, should sometimes result in an *rIIB*<sup>+</sup> revertant. These predictions were exactly verified by the results (Fig. 8.4c).

#### Evidence that most amino acids are specified by more than one codon

As Fig. 8.4c illustrates, intragenic suppression occurs only if, in the region between two frameshift mutations of opposite sign, a gene still dictates the appearance of amino acids—even if these amino acids are not the same as those appearing in the normal protein. If the frameshifted part of the gene instead encodes instructions to stop protein synthesis by introducing a triplet that does not correspond to any amino acid, then production of a functional polypeptide will not be possible. The reason is that polypeptide synthesis would stop before the compensating mutation could re-establish the correct reading frame.

The fact that intragenic suppression occurs as often as it does suggests that the code includes more than one codon for some amino acids. Recall that there are 20 common amino acids but  $4^3 = 64$  different combinations of three nucleotides. If each amino acid corresponded to only a single codon, there would be  $64 - 20 = 44$  possible triplets not encoding an amino acid. These noncoding triplets would act as *stop* signals and prevent further polypeptide synthesis. In this scenario, more than half of all frameshift

**Figure 8.4** Studies of frameshift mutations in the bacteriophage T4 *rII*B gene showed that codons consist of three nucleotides. **(a)** Treatment with proflavin produces an *rII*B<sup>-</sup> frameshift mutation at one site (FC0) by insertion of a single nucleotide; the reading frame of all codons downstream of the insertion is shifted (yellow). A second proflavin exposure results in a second mutation (FC7), deletion of a single nucleotide within the same gene, which suppresses FC0 by restoring the proper reading frame (green). **(b)** When the revertant is crossed with a wild-type strain, crossing-over separates the two *rII*B<sup>-</sup> frameshift mutations (FC0 and FC7) onto separate DNA molecules. The reversion to an *rII*B<sup>+</sup> phenotype was thus the result of intragenic suppression. **(c)** When recombined onto a single DNA molecule, two addition (++) or two deletion (--) mutations do not supply *rII*B<sup>+</sup> function, but three mutations of the same sign (+++ or ---) restore the reading frame.



mutations (44/64) would cause protein synthesis to stop at the first codon after the mutation, and the chances of extending the protein would diminish exponentially with each additional amino acid. As a result, intragenic suppression would rarely occur. However, we have seen that many frameshift mutations of one sign can be offset by mutations of the other sign. The distances between these mutations, estimated by recombination frequencies, are in some cases large enough to code for more than 50 amino acids, which would be possible only if most of the 64 possible triplet codons specified amino acids. Thus, the data of Crick and Brenner provide strong support for the idea that the genetic

code is **degenerate**: In most cases, two or more nucleotide triplets specify a single one of the 20 amino acids (see the genetic code in Fig. 8.2).

### Cracking the Code: Which Codons Represent Which Amino Acids?

Although the genetic experiments just described allowed remarkably prescient insights about the nature of the genetic code, they did not establish a correspondence between specific codons and specific amino acids. The discovery of

messenger RNA and the development of techniques for synthesizing simple messenger RNA molecules had to occur first so that researchers could manufacture simple proteins in the test tube.

**The discovery of messenger RNAs**

In the 1950s, researchers exposed eukaryotic cells to amino acids tagged with radioactivity and observed that protein synthesis incorporating the radioactive amino acids into polypeptides takes place in the cytoplasm, even though the genes for those polypeptides are sequestered in the cell nucleus. From this discovery, they deduced the existence of an intermediate molecule, made in the nucleus and capable of transporting DNA sequence information to the cytoplasm, where it can direct protein synthesis. RNA was a prime candidate for this intermediary information-carrying molecule.

Because of RNA’s potential for base pairing with a strand of DNA, one could imagine the cellular machinery copying a strand of DNA into a complementary strand of RNA in a manner analogous to the DNA-to-DNA copying of DNA replication. Subsequent studies in eukaryotes using radioactive uracil, a base found only in RNA, showed that although the molecules are synthesized in the nucleus, at least some of them migrate to the cytoplasm. Among those RNA molecules that migrate to the cytoplasm are the messenger RNAs, or mRNAs, depicted in Fig. 8.1. They arise in the nucleus from the transcription of DNA sequence information and then move (after processing) to the cytoplasm, where they determine the order of amino acids during protein synthesis.

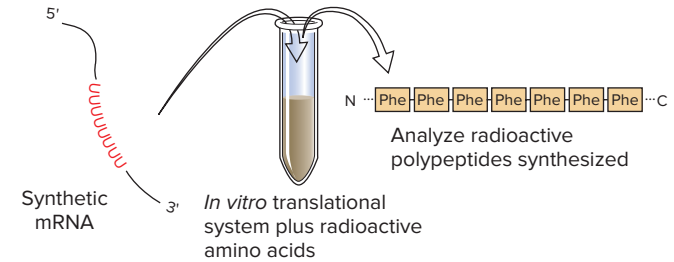
**Using synthetic mRNAs and *in vitro* translation**

Knowledge of mRNA served as the framework for two experimental breakthroughs that led to the deciphering of the genetic code. In the first, biochemists obtained cellular extracts that, with the addition of mRNA, synthesized polypeptides in a test tube. They called these extracts *in vitro translational systems*. The second breakthrough was the development of techniques enabling the synthesis of artificial mRNAs containing only a few codons of known composition. When added to *in vitro* translational systems, these simple, synthetic mRNAs directed the formation of simple polypeptides.

In 1961, Marshall Nirenberg and Heinrich Matthaei added a synthetic poly-U (5’ . . . UUUUUUUUUUU . . . 3’) mRNA to a cell-free translational system derived from *E. coli*. With the poly-U mRNA, phenylalanine (Phe) was the only amino acid incorporated into the resulting polypeptides (Fig. 8.5a). Because UUU is the only possible triplet in poly-U, UUU must be a codon for phenylalanine. In a similar fashion, Nirenberg and Matthaei showed that CCC encodes proline (Pro), AAA is a codon for lysine (Lys), and GGG encodes glycine (Gly) (Fig. 8.5b).

**Figure 8.5** How geneticists used synthetic mRNAs to limit the coding possibilities. (a) Poly-U mRNA generates a poly-phenylalanine polypeptide. (b) Polydi-, polytri-, and polytetra-nucleotides encode simple polypeptides. Some synthetic mRNAs, such as poly-GUAA, contain stop codons in all three reading frames and thus specify the construction only of short peptides.

**(a) Poly-U mRNA encodes polyphenylalanine.**



**(b) Analyzing the coding possibilities.**

Synthetic mRNA	Polypeptides synthesized
<b>Polypeptides with one amino acid</b>	
poly-U UUUU ...	Phe-Phe-Phe ...
poly-C CCCC ...	Pro-Pro-Pro ...
poly-A AAAA ...	Lys-Lys-Lys ...
poly-G GGGG ...	Gly-Gly-Gly ...
<b>Repeating dinucleotides</b>	
poly-UC UCUCUC ...	Ser-Leu-Ser-Leu ...
poly-AG AGAGAG ...	Arg-Glu-Arg-Glu ...
poly-UG UGUGUG ...	Cys-Val-Cys-Val ...
poly-AC ACACAC ...	Thr-His-Thr-His ...
<b>Repeating trinucleotides</b>	
poly-UUC UUCUUCUUC ...	Phe-Phe.... and Ser-Ser.... and Leu-Leu....
poly-AAG AAGAAGAAG ...	Lys-Lys.... and Arg-Arg.... and Glu-Glu....
poly-UUG UUGUUGUUG ...	Leu-Leu.... and Cys-Cys.... and Val-Val....
poly-UAC UACUACUAC ...	Tyr-Tyr.... and Thr-Thr.... and Leu-Leu....
<b>Repeating tetranucleotides</b>	
poly-UAUC UAUCUAUC ...	Tyr-Leu-Ser-Ile-Tyr-Leu-Ser-Ile...
poly-UUAC UUACUUAC ...	Leu-Leu-Thr-Tyr-Leu-Leu-Thr-Tyr...
poly-GUAA GUAAGUAA ...	none
poly-GAUA GAUAGAUA ...	none

The chemist Har Gobind Khorana later made mRNAs with repeating dinucleotides, such as poly-UC (5’ . . . UCUCUCUC . . . 3’), repeating trinucleotides, such as poly-UUC, and repeating tetranucleotides, such as poly-UAUC, and used them to direct the synthesis of slightly more complex polypeptides. As Fig. 8.5b shows, his results limited the coding possibilities, but some ambiguities remained. For example, poly-UC encodes the polypeptide N . . . Ser-Leu-Ser-Leu-Ser-Leu . . . C in which serine and leucine alternate with each other. Although the mRNA contains only two different codons (5’ UCU 3’ and 5’ CUC 3’), it is not obvious which corresponds to serine and which to leucine.

Nirenberg and Philip Leder resolved these ambiguities in 1965 with experiments in which they added short,

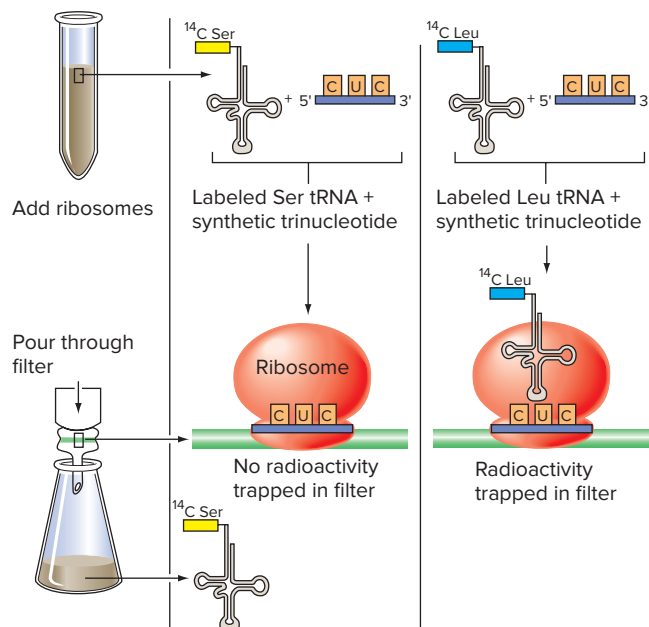


synthetic mRNAs only three nucleotides in length to an *in vitro* translational system containing tRNAs attached to amino acids, where only one of the 20 amino acids was radioactive. They then poured through a filter a mixture of a synthetic mRNA and the translational system containing a tRNA-attached, radioactively labeled amino acid (**Fig. 8.6**). tRNAs carrying an amino acid normally go right through a filter. If, however, a tRNA carrying an amino acid binds to a ribosome, it will stick in the filter, because this larger complex of ribosome, amino-acid-carrying tRNA, and small mRNA cannot pass through.

Nirenberg and Leder used this approach to see which small mRNA caused the entrapment of which radioactively labeled amino acid. For example, they knew from Khorana's earlier work that CUC encoded either serine or leucine. When they added the synthetic triplet CUC to an *in vitro* system where the radioactive amino acid was serine, this tRNA-attached amino acid passed through the filter, and the filter thus emitted no radiation (**Fig. 8.6**). But when they added the same triplet to a system where the radioactive amino acid was leucine, the filter lit up with radioactivity, indicating that the radioactively tagged leucine attached to a tRNA had bound to the ribosome-mRNA complex and gotten stuck in the filter (**Fig. 8.6**). CUC thus encodes leucine, not serine. Nirenberg and Leder used this technique to determine most of the codon-amino

### Figure 8.6 Cracking the genetic code with mini-mRNAs.

Nirenberg and Leder added trinucleotides of known sequence, in combination with a mixture of amino acid-charged tRNAs where only one amino acid was radioactive, to an *in vitro* extract containing ribosomes. If the trinucleotide specified this amino acid, the radioactive charged tRNA formed a complex with the ribosomes that could be trapped on a filter. The experiments shown here indicate that the codon CUC specifies leucine, not serine.



acid correspondences shown in the genetic code table (see **Fig. 8.2**).

### Polarities: 5' to 3' in mRNA corresponds to N to C in the polypeptide

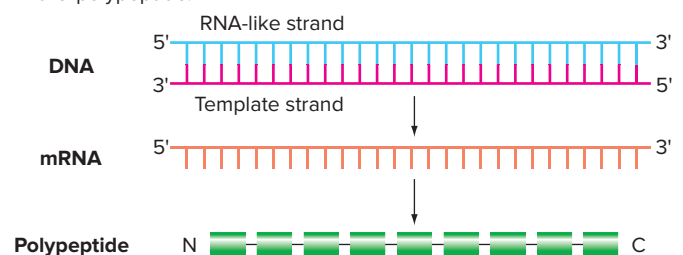
In studies using synthetic mRNAs, when investigators added the six-nucleotide-long 5' AAAUUU 3' to an *in vitro* translational system, the product N Lys-Phe C emerged, but no N Phe-Lys C appeared. Because AAA is the codon for lysine and UUU is the codon for phenylalanine, this result means that the codon closest to the 5' end of the mRNA encoded the amino acid closest to the N terminus of the corresponding polypeptide. Similarly, the codon nearest the 3' end of the mRNA encoded the amino acid nearest the C terminus of the resulting polypeptide.

To understand how the polarities of the macromolecules participating in gene expression relate to each other, remember that although the gene is a segment of a DNA double helix, only one of the two strands serves as a template for the mRNA. This strand is known as the **template strand**. The other strand is the **RNA-like strand**, because it has the same polarity and sequence (written in the DNA dialect) as the RNA. Note that some scientists use the terms *sense strand* or *coding strand* as synonyms for the RNA-like strand; in these alternative nomenclatures, the template strand would be the *antisense strand* or the *noncoding strand*. **Figure 8.7** diagrams the respective polarities of a gene's DNA, the mRNA transcript of that DNA, and the resulting polypeptide.

### Nonsense codons and polypeptide chain termination

Although most of the simple, repetitive RNAs synthesized by Khorana were very long and thus generated very long polypeptides, a few did not. These RNAs had signals that stopped construction of a polypeptide chain. As it turned out, three different triplets—UAA, UAG, and UGA—do not correspond to any of the amino acids. When these

**Figure 8.7 Correlation of polarities in DNA, mRNA, and polypeptide.** The template strand of DNA is complementary to both the RNA-like DNA strand and the mRNA. The 5'-to-3' direction in an mRNA corresponds to the N terminus-to-C terminus direction in the polypeptide.

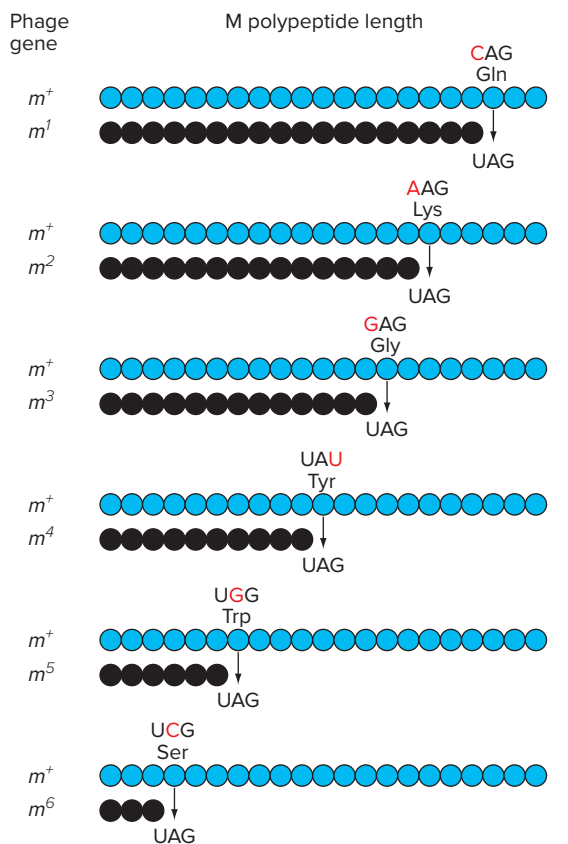


codons appear in frame, translation stops. As an example of how investigators established this fact, consider the case of poly-GUAA (review Fig. 8.5b). This mRNA will not generate a long polypeptide because in all possible reading frames, it contains the **stop codon** UAA.

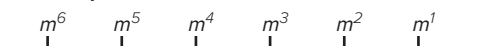
Sydney Brenner helped establish the identities of the stop codons in an alternative way, through ingenious experiments involving point mutations in a T4 phage gene named *m*, encoding a protein component of the phage head capsule. As shown in **Fig. 8.8a**, Brenner determined that certain mutant alleles ( $m^1$ – $m^6$ ) encoded *truncated polypeptides* that were shorter than the wild-type M protein. Brenner found that the final amino acid at the C terminus in each of the truncated proteins would have been followed in the normal, full-length protein by an amino acid specified by a codon

**Figure 8.8** Sydney Brenner's experiment showing that **UAG is a stop signal**. (a) The T4 phage  $m^+$  gene encodes a polypeptide M whose amino acids are shown with blue circles. Mutant alleles  $m^1$ – $m^6$  direct synthesis of truncated M proteins (black circles). In the wild-type M protein, the amino acid that would follow the final amino acid in each truncated protein is encoded by a triplet that differs from UAG by a single nucleotide. (b) The genetic map positions of the  $m^1$ – $m^6$  mutations are colinear with the sizes of the corresponding truncated M proteins.

**(a) Nonsense mutations**



**(b) Fine structure map**



that differed from the triplet UAG by a single nucleotide. These data suggested that each *m* mutant had a point mutation that changed a codon for an amino acid into the stop codon UAG. Such a mutation is called a **nonsense mutation** because it changes a codon that signifies an amino acid (a *sense codon*) into one that does not (a *nonsense codon*). (It was not a coincidence that all of the truncation mutants had nonsense mutations where a codon was changed to a particular stop codon—in this case UAG. Problem 56 at the end of this chapter explains why this was the case.)

Brenner later established that a fine structure map of mutations  $m^1$ – $m^6$  corresponds in a linear manner to the size of the truncated polypeptide chains (**Fig. 8.8b**). It makes sense that the M protein encoded by  $m^6$ , for example, is shorter than that encoded by  $m^5$  because the  $m^6$  nonsense mutation is closer to the beginning of the reading frame than  $m^5$ .

Brenner also isolated analogous sets of nonsense mutations that defined UAA and UGA as stop codons. For historical reasons, researchers often refer to UAG as the *amber* codon, UAA as the *ochre* codon, and UGA as the *opal* codon. The historical basis of this nomenclature is the last name of one of the early investigators—Bernstein—which means *amber* in German; ochre and opal derive from their similarity with amber as semiprecious materials.

## The Genetic Code: A Summary

The genetic code is a complete, unabridged dictionary equating the four-letter language of the nucleic acids with the 20-letter language of the proteins. The following list summarizes the code's main features:

1. **Triplet codons:** As written in Fig. 8.2, the code shows the 5'-to-3' sequence of the three nucleotides in each mRNA codon; that is, the first nucleotide depicted is at the 5' end of the codon.
2. The codons are *nonoverlapping*. In the mRNA sequence 5' GAAGUUGAA 3', for example, the first three nucleotides (GAA) form one codon; nucleotides 4 through 6 (GUU) form the second; and so on. Each nucleotide is part of only one codon.
3. The code includes three *stop*, or *nonsense*, *codons*: UAG, UAA, and UGA. These codons do not usually encode an amino acid and thus terminate translation.
4. The code is *degenerate*, meaning that more than one codon may specify the same amino acid. The code is nevertheless unambiguous because each codon specifies only one amino acid.
5. The cellular machinery scans mRNA from a fixed starting point that establishes a *reading frame*. As we will see later, the nucleotide triplet AUG, which specifies the amino acid methionine wherever it appears in the reading frame, also serves as the **initiation codon**, marking where in an mRNA the code for a particular polypeptide begins.

6. *Corresponding polarities of codons and amino acids:* Moving in the 5'-to-3' direction along an mRNA, each successive codon is sequentially decoded into an amino acid, starting at the N terminus and moving toward the C terminus of the resulting polypeptide.
7. Mutations may modify the message encoded in a sequence of nucleotides in three ways. *Frameshift mutations* are nucleotide insertions or deletions that alter the genetic instructions for polypeptide construction by changing the reading frame. *Missense mutations* change a codon for one amino acid to a codon for a different amino acid. *Nonsense mutations* change a codon for an amino acid to a stop codon.

## The Effects of Mutations on Polypeptides Helped Verify the Code

The experiments that first cracked the genetic code by assigning codons to amino acids were all *in vitro* studies using cell-free extracts and synthetic mRNAs. A logical question thus arose: Do living cells construct polypeptides according to the same rules? Early evidence that they do came from studies analyzing how mutations actually affect the amino acid composition of the polypeptides encoded by a gene. Most mutagens change a single nucleotide in a codon. As a result, most missense mutations that change the identity of a single amino acid should be single-nucleotide substitutions, and analyses of these substitutions should conform to the code. Yanofsky, for example, found two *trpA*<sup>-</sup> auxotrophic mutations in the *E. coli* tryptophan synthase gene that produced two different amino acids (arginine, or Arg, and glutamic acid, or Glu) at the same position—amino acid 211—in the polypeptide chain (**Fig. 8.9a**). According to the code, both of these mutations could have resulted from single-base changes in the GGA codon that normally inserts glycine (Gly) at position 211.

Even more informative were the *trpA*<sup>+</sup> revertants of these mutations subsequently isolated by Yanofsky. As Fig. 8.9a illustrates, single-base substitutions in the gene could also explain the amino acid changes in these revertants. Note that some of these substitutions restore Gly to position 211 of the polypeptide, while others place amino acids such as Ile, Thr, Ser, Ala, or Val at this site in the tryptophan synthase molecule. The substitution of these other amino acids for Gly at position 211 in the polypeptide chain is compatible with (that is, largely conserves) the enzyme's function.

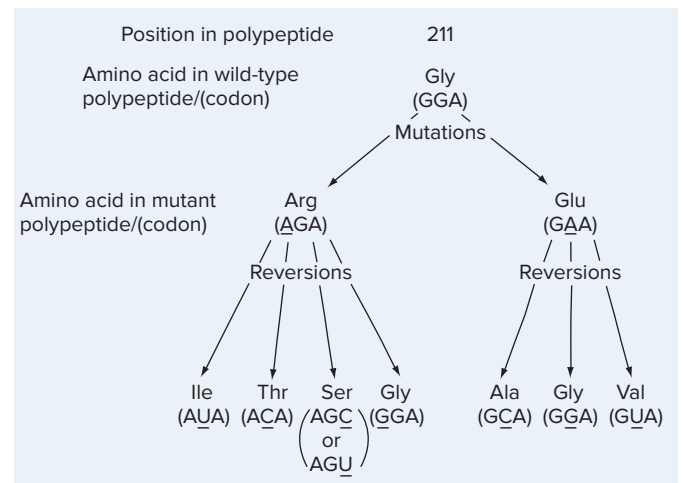
Yanofsky obtained better evidence yet that cells use the genetic code *in vivo* by analyzing proflavin-induced frameshift mutations of the tryptophan synthase gene (**Fig. 8.9b**). He first treated populations of *E. coli* with proflavin to produce *trpA*<sup>-</sup> mutants. Subsequent treatment of these mutants with more proflavin generated some

*trpA*<sup>+</sup> revertants among the progeny. The most likely explanation for the revertants was that their tryptophan synthase gene carried both a single-base-pair deletion and a single-base-pair insertion (- +). Upon determining the amino acid sequences of the tryptophan synthase enzymes made by the revertant strains, Yanofsky found that he could use the genetic code to predict the precise amino acid alterations that had occurred by assuming the revertants had a specific single-base-pair insertion and a specific single-base-pair deletion (**Fig. 8.9b**).

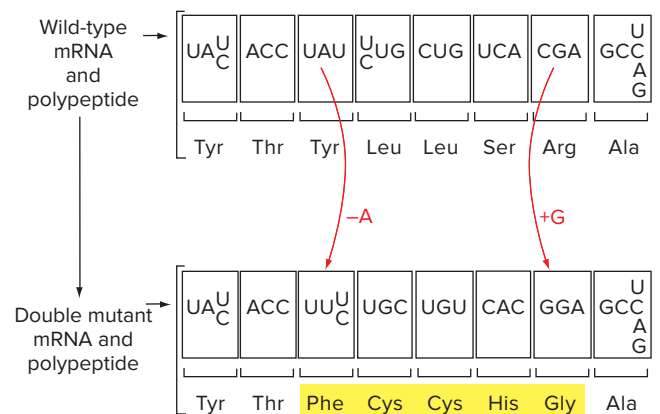
Yanofsky's results helped confirm not only amino acid codon assignments but other parameters of the code as well. His interpretations make sense only if codons do not overlap and are read from a fixed starting point, with no pauses or commas separating the adjacent triplets.

**Figure 8.9** Experimental verification of the genetic code. (a) Single-base substitutions can explain the amino acid substitutions caused by *trpA*<sup>-</sup> mutations and *trpA*<sup>+</sup> reversions. (b) The genetic code predicts the amino acid alterations (yellow) that would arise from single-base-pair deletions and suppressing insertions.

(a) Altered amino acids in *trpA*<sup>-</sup> mutations and *trpA*<sup>+</sup> revertants



(b) Amino acid alterations that accompany intragenic suppression



## The Genetic Code Is Almost, but Not Quite, Universal

We now know that virtually all cells alive today use the same basic genetic code. One early indication of this uniformity was that a translational system derived from one organism could use the mRNA from another organism to convert genetic information to the encoded protein. Rabbit hemoglobin mRNA, for example, when injected into frog eggs or added to cell-free extracts from wheat germ, directs the synthesis of rabbit hemoglobin proteins. More recently, comparisons of DNA and protein sequences have revealed a perfect correspondence according to the genetic code between codons and amino acids in almost all organisms examined.

### Conservation of the genetic code

The universality of the code is an indication that it evolved very early in the history of life. Once it emerged, the code remained constant over billions of years, in part because evolving organisms would have little tolerance for change. A single change in the genetic code could disrupt the production of hundreds or thousands of proteins in a cell—from the DNA polymerase essential for replication to the RNA polymerase required for gene expression to the tubulin proteins that compose the mitotic spindle—and such a change would therefore be lethal.

### Exceptional genetic codes

Researchers were thus quite amazed to observe a few exceptions to the universality of the code. In some species of the single-celled eukaryotic protozoans known as *ciliates*, the codons UAA and UAG, which are nonsense codons in most organisms, specify the amino acid glutamine; in other ciliates, UGA, the third stop codon in most organisms, specifies cysteine. Ciliates use the remaining nonsense codons as stop codons.

Other systematic changes in the genetic code exist in mitochondria, the semiautonomous, self-reproducing organelles within eukaryotic cells that are the sites of ATP formation. Each mitochondrion has its own chromosomes and its own apparatus for gene expression (which we describe in detail in Chapter 15). In the mitochondria of yeast, for example, CUA specifies threonine instead of leucine. Yet another exception to the code is seen in certain prokaryotes that sometimes use the triplet UAG to specify insertion of the rare amino acid pyrrolysine (see Fig. 7.28c and also Problem 57 at the end of this chapter.)

The experimental evidence presented so far helped define a nearly universal genetic code. But although cracking the code made it possible to understand the

broad outlines of information flow between gene and protein, these results did not explain exactly how the cellular machinery accomplishes gene expression. This question is our focus as we present in the next sections the details of transcription and translation.

### essential concepts

- The nearly universal *genetic code* consists of 64 *codons*, each one composed of three nucleotides. Sixty-one codons specify amino acids, while three—UAA, UAG, UGA—are *stop codons*. The code is *degenerate* in that more than one codon can specify one amino acid.
- The codon AUG specifies methionine; it also serves as the *initiation codon* establishing the *reading frame* that groups nucleotides into successive, nonoverlapping codon triplets.
- *Missense mutations* change a codon so that it specifies a different amino acid; *frameshift mutations* alter the reading frame for all codons following the mutation; and *nonsense mutations* change a codon for an amino acid into a stop codon.

## 8.2 Transcription: From DNA to RNA

### learning objectives

1. Describe the three stages of transcription: initiation, elongation, and termination.
2. Compare transcription initiation in prokaryotes and eukaryotes.
3. List three ways by which eukaryotes process mRNA after transcription.

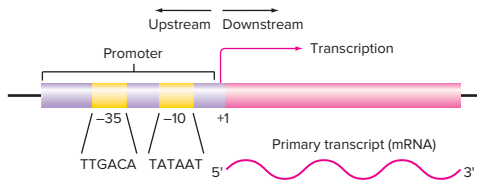
*Transcription* is the process by which the polymerization of ribonucleotides, guided by complementary base pairing, produces an RNA transcript of a gene. The template for the RNA transcript is one strand of that portion of the DNA double helix that constitutes the gene.

### RNA Polymerase Synthesizes a Single-Stranded RNA Copy of a Gene

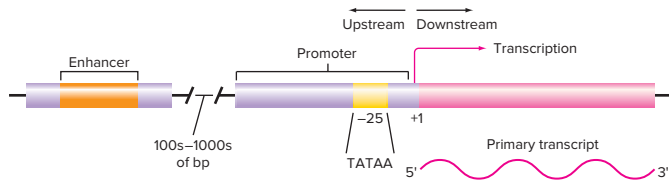
**Figure 8.10** depicts the basic components of transcription and illustrates key events in the process as it occurs in the bacterium *E. coli*. This figure divides transcription into

**Figure 8.11 Control regions of bacterial and eukaryotic genes.** Only the sequence of the RNA-like strand is shown; numbering starts at the first transcribed nucleotide (+1). **(a)** All promoters in *E. coli* share two different short stretches of nucleotides (yellow) essential for promoter recognition by RNA polymerase. The most common nucleotides in these short regions constitute the *consensus sequences* shown. **(b)** Eukaryotic genes transcribed by RNA pol II have a promoter, and also one or more distant DNA elements called *enhancers* (orange) that bind to protein factors aiding transcription.

**(a) Transcription initiation region in bacterial genes**



**(b) Transcription initiation region in eukaryotic genes transcribed by pol II**



successive phases of *initiation*, *elongation*, and *termination*. The following four points are of particular importance:

1. The enzyme **RNA polymerase** catalyzes transcription.
2. DNA sequences near the beginning of genes, called **promoters**, signal RNA polymerase to begin transcription. Most bacterial gene promoters have almost identical nucleotide sequences in each of two short regions (**Fig. 8.11a**). These are the sites at which RNA polymerase makes particularly strong contact with the promoters.
3. RNA polymerase adds nucleotides to the growing RNA polymer in the 5'-to-3' direction. The chemical mechanism of this nucleotide-adding reaction is similar to the formation of phosphodiester bonds between nucleotides during DNA replication (review Fig. 6.21), with one exception: Transcription uses ribonucleotide triphosphates (ATP, CTP, GTP, and UTP) instead of deoxyribonucleotide triphosphates. Hydrolysis of the high-energy bonds in each ribonucleotide triphosphate provides the energy needed for elongation.
4. Sequences in the RNA products, known as **terminators**, tell RNA polymerase where to stop transcription.

As you examine Fig. 8.10, bear in mind that a gene consists of two antiparallel strands of DNA, as mentioned earlier. One—the *RNA-like strand*—has the same polarity

and sequence (except for T instead of U) as the emerging RNA transcript. The second—the *template strand*—has the opposite polarity and a complementary sequence that enables it to serve as the template for making the RNA transcript. When geneticists refer to the *sequence of a gene*, they usually mean the sequence of the RNA-like strand.

## Transcription Initiation Varies Between Eukaryotes and Prokaryotes

Although the transcription of all genes in all organisms roughly follows the general scheme diagrammed in Fig. 8.10, prokaryotic and eukaryotic organisms vary in important details. In eukaryotes, promoters are more complicated than those in bacteria, and three different kinds of RNA polymerase exist that can transcribe different classes of genes. One of these is eukaryotic RNA polymerase II (pol II), which transcribes genes that encode proteins. **Figure 8.11b** illustrates the general structure of the DNA regions of eukaryotic genes that allow pol II to initiate transcription. A key difference with prokaryotes is that sequences called *enhancers* that can be thousands of base pairs away from the promoter are often also required for efficient transcription of eukaryotic genes.

Chapters 16 and 17 will describe how prokaryotic and eukaryotic cells can exploit these and other variations to control when, where, and at what level a given gene is expressed. Finally, the Genetics and Society Box *HIV and Reverse Transcription* describes how the AIDS virus uses an exceptional form of transcription, known as **reverse transcription**, to construct a double strand of DNA from an RNA template.

The result of transcription is a single strand of RNA known as a **primary transcript** (see Figs. 8.10 and 8.11). In prokaryotic organisms, the RNA produced by transcription is the actual messenger RNA that guides protein synthesis. In eukaryotic organisms, by contrast, most primary transcripts undergo **RNA processing** in the nucleus before they migrate to the cytoplasm to direct protein synthesis. As we see in the following section, this processing has played a fundamental role in the evolution of complex organisms.

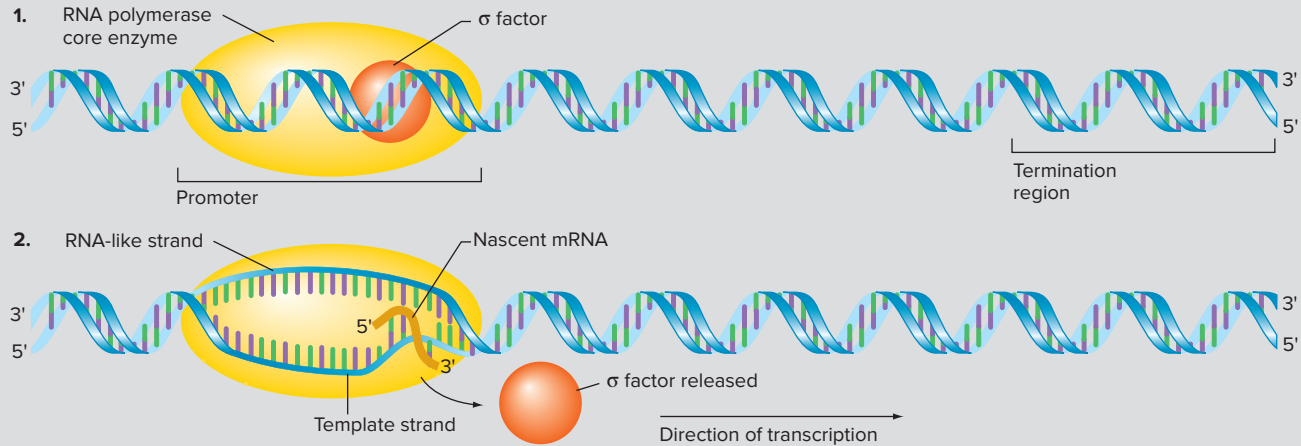
## In Eukaryotes, RNA Processing After Transcription Produces a Mature mRNA

Some RNA processing in eukaryotes modifies only the 5' or 3' ends of the primary transcript, leaving the information content of the rest of the mRNA untouched. Other processing deletes blocks of information from the middle of the primary transcript, so the content of the mature mRNA is

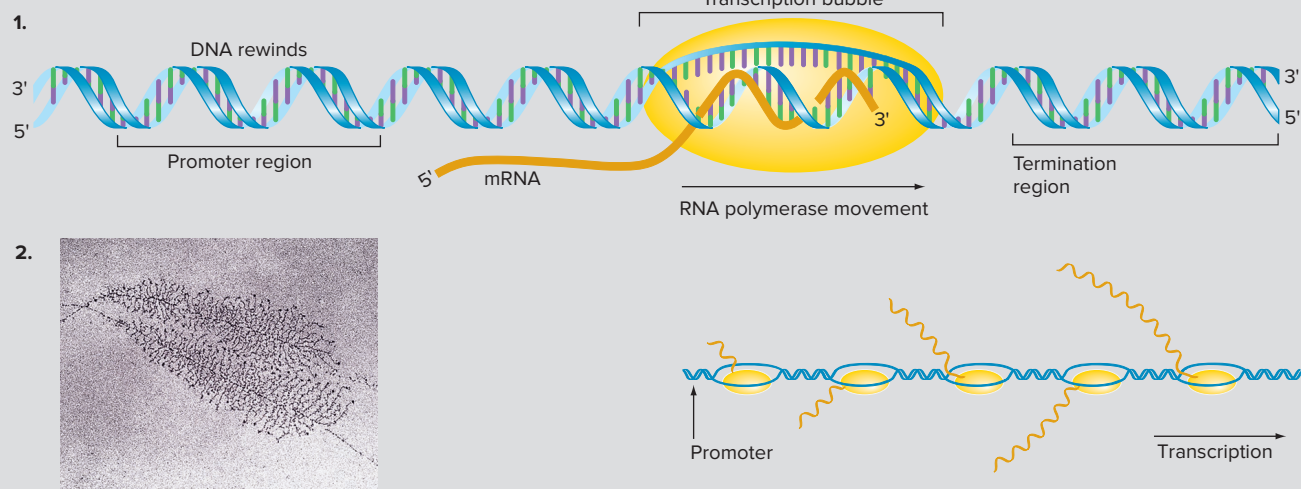
**FEATURE FIGURE 8.10**

**Transcription in Bacterial Cells**

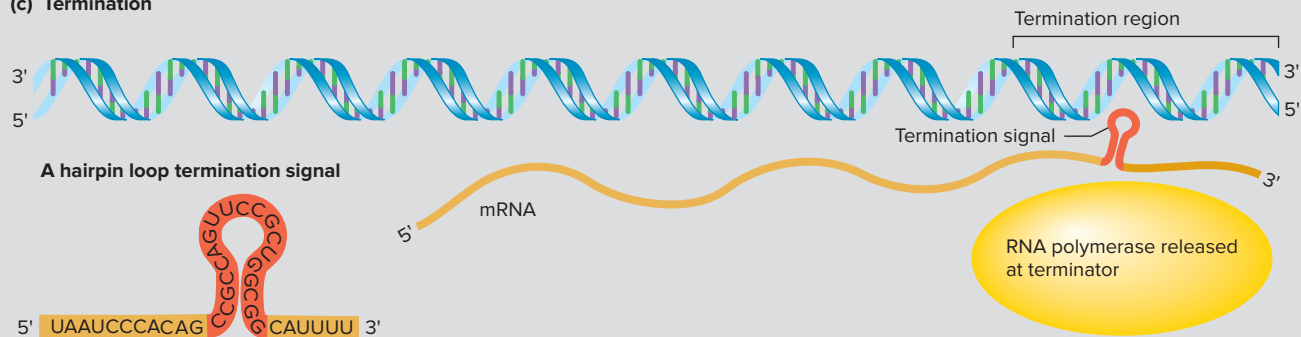
**(a) The initiation of transcription**



**(b) Elongation**



**(c) Termination**



**FEATURE FIGURE 8.10 (Continued)****(a) The Initiation of Transcription**

1. *RNA polymerase binds to double-stranded DNA at the beginning of the gene to be copied.* RNA polymerase recognizes and binds to **promoters**, specialized DNA sequences near the transcription start site. Although specific promoters vary substantially, all promoters in *E. coli* contain two characteristic short sequences of 6–10 base pairs (Fig. 8.11a). In bacteria, the complete RNA polymerase (the *holoenzyme*) consists of a *core enzyme*, plus a  $\sigma$  (*Sigma*) subunit involved only in initiation. The  $\sigma$  subunit reduces RNA polymerase's general affinity for DNA but simultaneously increases RNA polymerase's affinity for the promoter. As a result, the RNA polymerase holoenzyme can home in on a promoter and bind tightly to it, forming a *closed promoter complex*.

2. *After binding to the promoter, RNA polymerase unwinds part of the double helix, exposing unpaired bases on the template strand.* The complex formed between the RNA polymerase holoenzyme and an unwound promoter is called an *open promoter complex*. The enzyme identifies the template strand and chooses the two nucleotides to be copied. Guided by base pairing with these two nucleotides, RNA polymerase aligns the first two ribonucleotides at the 5' end of the new RNA. The DNA transcribed as the 5' end of the mRNA is the *5' end of the gene*. RNA polymerase then forms a phosphodiester bond between the first two ribonucleotides. Soon thereafter, the RNA polymerase releases the  $\sigma$  subunit, marking the end of initiation.

**(b) Elongation: Constructing an RNA Copy of the Gene**

1. *When the  $\sigma$  subunit is released, RNA polymerase loses its enhanced affinity for the promoter sequence and regains its strong generalized affinity for any DNA.* These changes enable the core enzyme to leave the promoter yet remain bound to the gene. The core enzyme moves along the chromosome, unwinding the DNA to expose the next single-stranded region of the template. The enzyme extends the RNA by adding the correct ribonucleotide to the 3' end of the growing chain. As the enzyme extends the mRNA in the 5'-to-3' direction, it moves in the antiparallel 3'-to-5' direction along the DNA template strand. RNA polymerase synthesizes RNA at an average speed of about 50 nucleotides per second.

The region of DNA unwound by RNA polymerase is the **transcription bubble**. Within the bubble, the nascent RNA chain remains base paired with the DNA template, forming a DNA–RNA hybrid. However, in those parts of the gene behind the bubble that have already been transcribed, the DNA double helix re-forms, displacing the RNA, which hangs out of the transcription complex as a single strand with a free 5' end.

2. *Once an RNA polymerase has moved off the promoter, other RNA polymerase molecules can move in to initiate transcription.* If the promoter is very strong, that is, if it can attract RNA polymerase rapidly, many enzyme molecules can transcribe it simultaneously. Here we show an electron micrograph and an artist's interpretation of simultaneous transcription by several RNA polymerases. The promoter for this gene lies very close to where the shortest RNA is emerging from the DNA.

Geneticists often use the direction traveled by RNA polymerase as a reference when discussing gene structure. If, for example, you started at the 5' end of a gene at point A and moved along the gene in the same direction as RNA polymerase to point B, you would be moving **downstream**. If, by contrast, you started at point B and moved in the opposite direction to point A, you would be traveling **upstream**.

**(c) Termination: The End of Transcription**

RNA sequences that signal the end of transcription are known as **terminators**. Two types of terminators exist: *intrinsic terminators*, which cause the RNA polymerase core enzyme to terminate transcription on its own, and *extrinsic terminators*, which require additional proteins—particularly a polypeptide known as *Rho*—to bring about termination. All terminators, whether intrinsic or extrinsic, are specific sequences in the mRNA that are transcribed from the gene. Terminators often form **hairpin loops** (also called **stem loops**) in which nucleotides within the mRNA pair with complementary nucleotides in the same molecule. Upon termination, RNA polymerase and a completed RNA chain are both released from the DNA.

related, but not identical, to the complete set of DNA nucleotide pairs in the original gene.

**Adding a 5' methylated cap and a 3' poly-A tail**

The nucleotide at the 5' end of a eukaryotic mRNA is a G in reverse orientation from the rest of the molecule; it is

connected through a triphosphate linkage to the first nucleotide in the primary transcript. This “backward G” is not transcribed from the DNA. Instead, a special *capping enzyme* adds it to the primary transcript after polymerization of the transcript's first few nucleotides. Enzymes known as *methyl transferases* then add methyl ( $-\text{CH}_3$ ) groups to the backward G and to one or more of the

## GENETICS AND SOCIETY



Crowd: © Image Source/Getty Images RF

## HIV and Reverse Transcription

The AIDS-causing human immunodeficiency virus (HIV) is the most intensively analyzed virus in history. From laboratory and clinical studies spanning more than three decades, researchers have learned that each viral particle is a rough-edged sphere consisting of an outer envelope enclosing a protein matrix, which, in turn, surrounds a cut-off cone-shaped core (Fig. A). Within the core lies an enzyme-studded genome: two identical single strands of RNA associated with many molecules of an unusual DNA polymerase known as **reverse transcriptase**.

During infection, the AIDS virus binds to and injects its cone-shaped core into cells of the human immune system (Fig. B). The virus next uses reverse transcriptase to copy its RNA genome into double-stranded DNA molecules in the cytoplasm of the host cell. The double helixes then travel to the nucleus where another enzyme, called *integrase*, inserts them into a host chromosome. Once integrated into a host-cell chromosome, the viral genome can do one of two things. It can commandeer the host cell's protein synthesis machinery to make hundreds of new viral particles that bud off from the parent cell, taking with them part of the cell membrane. This process sometimes results in the host cell's death, which weakens the person's immune system. Alternatively, the HIV genome can lie

latent inside the host chromosome, which then copies and transmits the viral genome to two new cells with each cell division.

The events of this life cycle make HIV a **retrovirus**: an RNA virus that after infecting a host cell copies its own single strands of RNA into double helixes of DNA, which a viral enzyme (integrase) then integrates into a host chromosome.

Reverse transcription, the foundation of the retroviral life cycle, is inconsistent with the one-way, DNA-to-RNA-to-protein flow of genetic information. Because it was so unexpected, the phenomenon of reverse transcription encountered great resistance in the scientific community when first reported by Howard Temin of the University of Wisconsin and David Baltimore, then of MIT. Now, however, it is an established fact. Reverse

Figure A Structure of the AIDS virus

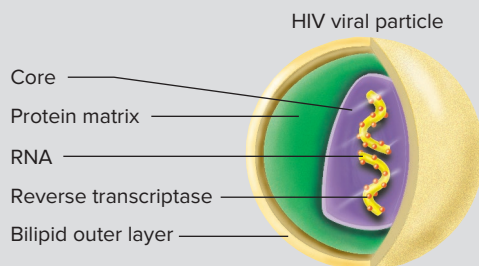
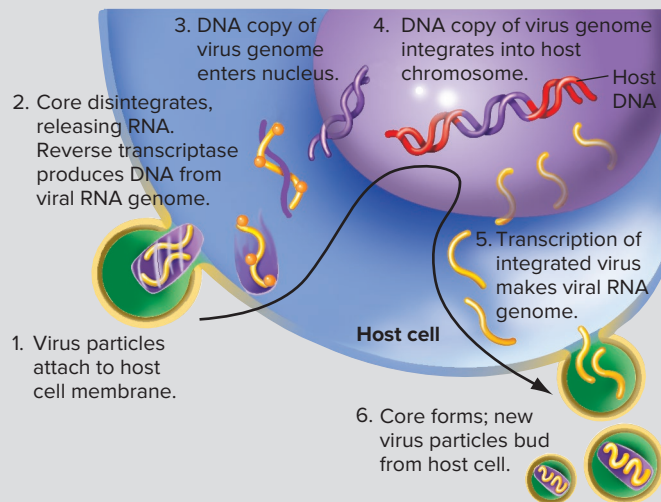


Figure B Life cycle of the AIDS virus



succeeding nucleotides in the RNA, forming a so-called **methylated cap** (Fig. 8.12).

Like the 5' methylated cap, the 3' end of most eukaryotic mRNAs is not encoded directly by the gene. In a large majority of eukaryotic mRNAs, the 3' end consists of 100–200 As, referred to as a **poly-A tail** (Fig. 8.13). Addition of the tail is a two-step process. First, a ribonuclease cleaves the primary transcript to form a new 3' end; cleavage depends on the sequence AAUAAA, which is found in poly-A-containing mRNAs 11–30 nucleotides upstream of the position where the tail is added. Next, the enzyme *poly-A polymerase* adds As onto the 3' end exposed by cleavage.

Unexpectedly, both the methylated cap and the poly-A tail are crucial for efficient translation of the mRNA into protein, even though neither helps specify an

amino acid. Recent data indicate that particular eukaryotic *translation initiation factors* bind to the 5' cap, while *poly-A binding protein* associates with the tail at the 3' end of the mRNA. The interaction of these proteins in many cases shapes the mRNA molecule into a circle. This circularization both enhances the initial steps of translation and stabilizes the mRNA in the cytoplasm by increasing the length of time it can serve as a messenger.

## Removing introns from the primary transcript by RNA splicing

Another kind of RNA processing became apparent in the late 1970s, after researchers had developed techniques that enabled them to analyze nucleotide sequences in both DNA



transcriptase is a remarkable DNA polymerase that can construct a DNA polymer from either an RNA or a DNA template.

In addition to its comprehensive copying abilities, reverse transcriptase has another feature not seen in most DNA polymerases: inaccuracy. As we saw in Chapter 7, normal DNA polymerases replicate DNA with an error rate of one mistake in every million nucleotides copied. Reverse transcriptase, however, introduces one mutation in every 5000 incorporated nucleotides.

HIV uses this capacity for mutation to gain a tactical advantage over the immune response of its host organism. Cells of the immune system seek to overcome an HIV invasion by multiplying in response to the proliferating viral particles (*virions*). The numbers are staggering. Each day of infection in every patient, from 100 million to a billion HIV particles are released from infected immune-system cells. As long as the immune system is strong enough to withstand the assault, it responds by producing as many as 2 billion new cells daily. Many of these new immune system cells produce antibodies targeted against proteins on the surface of the virus.

But just when an immune response wipes out those viral particles carrying the targeted protein, virions incorporating new forms of the protein resistant to the current immune response make their appearance. After many years of this complex chase, capture, and destruction by the immune system, the changeable virus outruns the host's immune response and gains the upper hand. Thus, the intrinsic infidelity of HIV's reverse transcriptase, by enhancing the virus's ability to compete in the evolutionary marketplace, increases its threat to human life and health.

This inherent mutability has undermined two potential therapeutic approaches toward the control of AIDS: drugs and vaccines. Some of the antiviral drugs approved in the United States for treatment of HIV infection—AZT (zidovudine), ddC (dideoxycytidine), and ddI (dideoxyinosine)—block viral replication by interfering with the action of reverse transcriptase. Each drug is similar to one of the four nucleotides, and when reverse

transcriptase incorporates one of the drug molecules rather than a genuine nucleotide into a growing DNA polymer, the enzyme cannot extend the chain any further. However, the drugs are toxic at high doses and thus can be administered only at low doses that do not destroy all virions. Because of this limitation and the virus's high rate of mutation, mutant reverse transcriptases soon appear that work even in the presence of the drugs.

Similarly, researchers are having trouble developing effective vaccines. Even if a vaccine could generate a massive immune response against one, two, or even several HIV proteins, such a vaccine might be effective for only a short while—until enough mutations build up to make the virus resistant.

For these reasons, the AIDS virus will most likely not succumb entirely to drugs or vaccines that target proteins active at various stages of its life cycle. However, combinations of these therapeutic tools have nonetheless proven remarkably effective at prolonging an AIDS patient's life. In 2013, AIDS patients who received combination therapy had on average two-thirds of a normal life span. Newer drugs added to the cocktail include protease inhibitors that prevent the activity of enzymes needed to produce viral coat proteins, drugs that prevent viral entry into human cells, and inhibitors of the viral integrase protein.

A self-preserving capacity for mutation, perpetuated by reverse transcriptase, is surely one of the main reasons for HIV's success. Ironically, it may also provide a basis for its subjugation. Researchers are studying what happens when the virus increases its mutational load. If reverse transcriptase's error rate determines the size and integrity of the viral population in a host organism, greatly accelerated mutagenesis might push the virus beyond the error threshold that allows it to function. In other words, too much mutation might destroy the virus's infectivity, virulence, or capacity to reproduce. If geneticists could figure out how to make this happen, they might be able to give the human immune system the advantage it needs to overcome the virus.

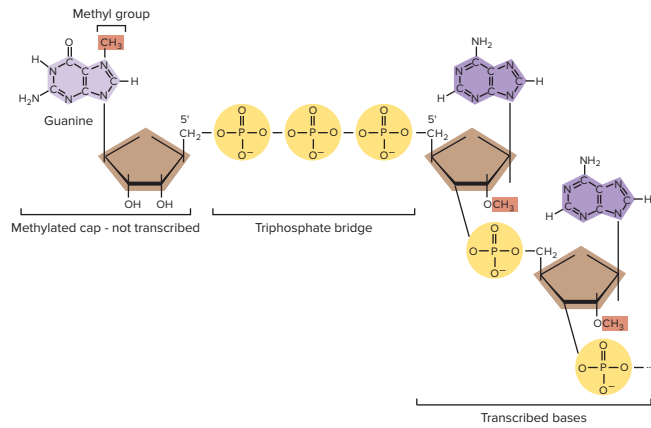
and RNA. Using these techniques, which we describe in Chapter 9, they began to compare eukaryotic genes with the mRNAs derived from them. Their expectation was that just as in prokaryotes, the DNA nucleotide sequence of a gene's RNA-like strand would be identical to the RNA nucleotide sequence of the messenger RNA (with the exception of U replacing T in the RNA). Surprisingly, the investigators found that the DNA nucleotide sequences of many eukaryotic genes are much longer than their corresponding mRNAs. This fact suggested that RNA transcripts, in addition to receiving a methylated cap and a poly-A tail, undergo extensive internal processing.

An extreme example of the length difference between primary transcript and mRNA is seen in the human gene *DMD*, which encodes the protein Dystrophin. Abnormalities

in the *DMD* gene underlie the genetic disorder Duchenne muscular dystrophy (DMD). The *DMD* gene is 2.5 million nucleotides—or 2500 kilobases (kb)—long, whereas the corresponding mRNA is roughly 14,000 nucleotides, or 14 kb, in length. Obviously the gene contains DNA sequences that are not present in the mature mRNA. Those regions of the gene that do end up in the mature mRNA are scattered throughout the 2500 kb of DNA.

**Exons and introns** Sequences found in both a gene's DNA and the mature messenger RNA are called **exons** (for *expressed regions*). The sequences found in the DNA of the gene but not in the mature mRNA are known as **introns** (for *intervening regions*). Introns interrupt, or separate, the exon sequences that actually end up in the mature mRNA.

**Figure 8.12** Structure of the methylated cap at the 5' end of eukaryotic mRNAs. Capping enzyme connects a backward G to the first nucleotide of the primary transcript through a triphosphate linkage. Methyl transferase enzymes then add methyl groups (orange) to this G and to one or two of the nucleotides first transcribed from the DNA template.



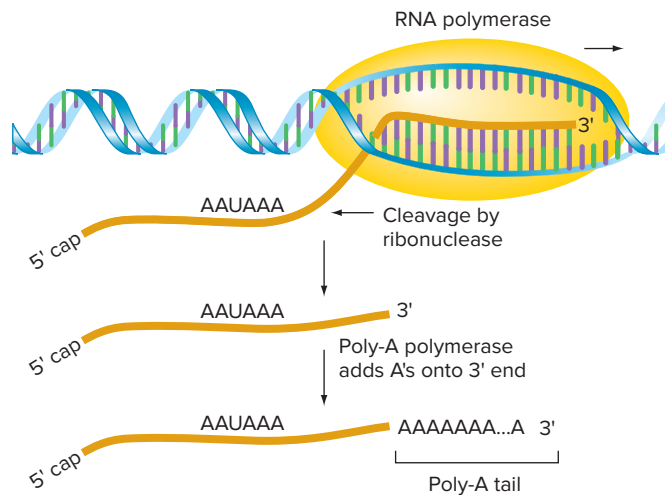
The gene for collagen (an abundant protein in connective tissue) shown in **Fig. 8.14** has two introns. By contrast, the *DMD* gene has more than 80 introns; the mean intron length is 35 kb, but one of its introns is an amazing 400 kb long. Other genes in humans generally have many fewer introns, while a few have none—and the introns range from 50 bp to over 100 kb. Exons, in contrast, vary in size from about 50 bp to a few kilobases; in the *DMD* gene, the mean exon length is 200 bp. The greater size variation seen in introns compared to exons reflects the fact that introns do not encode polypeptides and do not appear in mature mRNAs. As a result, fewer restrictions exist on the sizes and base sequences of introns.

Mature mRNAs must contain all the codons that are translated into amino acids, including the initiation and termination codons. In addition, mature mRNAs have sequences at their 5' and 3' ends that are not translated, but that nevertheless play important roles in regulating the efficiency of translation. These sequences, called the **5' and 3' untranslated regions (5' and 3' UTRs)**, are located just after the methylated cap and just before the poly-A tail, respectively (**Fig. 8.14a**). Excepting the cap and tail themselves, all of the sequences in a mature mRNA, including all the codons and both UTRs, must be transcribed from the gene's exons.

Introns can interrupt a gene at any location, even between the nucleotides making up a single codon. In such a case, the three nucleotides of the codon are present in two different (but successive) exons. You should also note that because introns can interrupt the 5' and/or 3' UTRs, the start codon is not always in the first exon, and neither is the stop codon always in the final exon.

How do cells make a mature mRNA from a gene whose coding sequences are interrupted by introns? The answer is that cells first make a primary transcript containing all of a

**Figure 8.13** How RNA processing adds a tail to the 3' end of eukaryotic mRNAs. A ribonuclease recognizes AAUAAA in a particular context of the primary transcript and cleaves the transcript 11–30 nucleotides downstream to create a new 3' end. The enzyme poly-A polymerase then adds 100–200 As onto this new 3' end.



gene's introns and exons, and then they remove the introns from the primary transcript by **RNA splicing**, the process that deletes introns and joins together successive exons to form a mature mRNA consisting only of exons (**Fig. 8.14a**). Because the first and last exons of the primary transcript become the 5' and 3' ends of the mRNA, while all intervening introns are spliced out, a gene must have one more exon than it does introns. To construct the mature mRNA, splicing must be remarkably precise. For example, if an intron lies within a codon, splicing must remove the intron and reconstitute the codon without disrupting the reading frame of the mRNA.

**The mechanism of RNA splicing** **Figure 8.15** illustrates how RNA splicing works. Three types of short sequences within the primary transcript—**splice donors**, **splice acceptors**, and **branch sites**—help ensure the specificity of splicing. These sites make it possible to sever the connections between an intron and the exons that precede and follow it, and then to join the formerly distant exons.

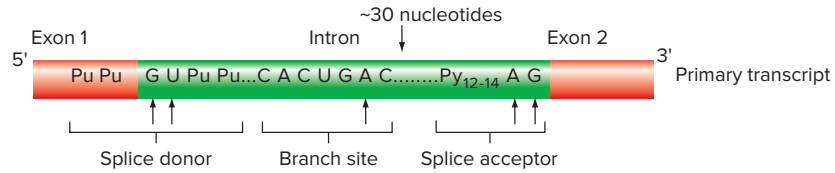
The mechanism of splicing involves two sequential cuts in the primary transcript. The first cut is at the splice-donor site, at the 5' end of the intron. After this first cut, the new 5' end of the intron attaches, via a novel 2'–5' phosphodiester bond, to an A at the branch site located within the intron, forming a so-called *lariat*. The second cut is at the splice-acceptor site, at the 3' end of the intron; this cut removes the intron. The discarded intron is degraded, and the precise splicing of adjacent exons completes the process of intron removal (**Fig. 8.15**).

**SnRNPs and the spliceosome** Splicing normally requires a complicated intranuclear machine called the **spliceosome**, which ensures that all of the splicing reactions take

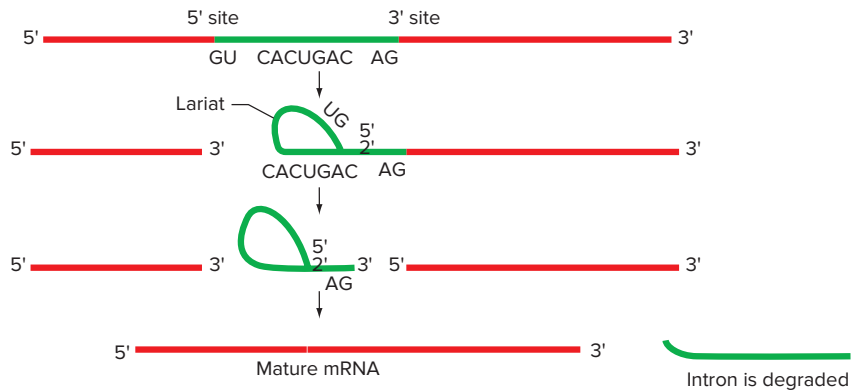


**Figure 8.15** How RNA processing splices out introns and joins adjacent exons. Exons are in red and introns in green. **(a)** Three short sequences within the primary transcript are needed for splicing. (1) The splice-donor site occurs where the 3' end of an exon abuts the 5' end of an intron. In most splice-donor sites, a GU dinucleotide (arrows) that begins the intron is flanked on either side by a few purines (Pu; that is, A or G). (2) The splice-acceptor site is at the 3' end of the intron where it joins with the next exon. The final nucleotides of the intron are always AG (arrows) usually preceded by 12–14 pyrimidines (Py; that is, C or U). (3) The branch site, which is located within the intron about 30 nucleotides upstream of the splice acceptor, must include an A (arrow) and is usually rich in pyrimidines. **(b)** Two sequential cuts, the first at the splice-donor site and the second at the splice-acceptor site, remove the intron, allowing precise splicing of adjacent exons.

**(a) Short sequences dictate where splicing occurs.**



**(b) Two sequential cuts remove the intron.**



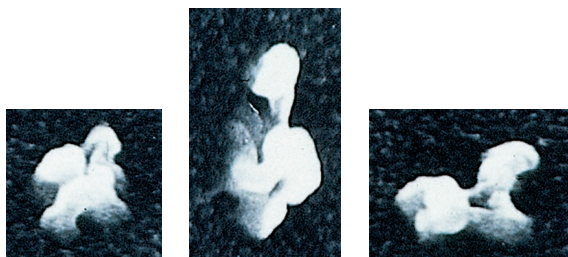
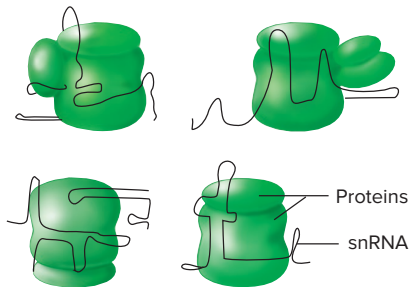
**Figure 8.16** Splicing is catalyzed by the spliceosome.

(Top) The spliceosome is assembled from four snRNP subunits, each of which contains one or two snRNAs and several proteins. (Bottom) Views of three spliceosomes in the electron microscope.

(bottom): © Dr. Thomas Maniatis, Thomas H. Lee Professor of Molecular and Cellular Biology, Harvard University

**Spliceosome components**

Five snRNAs (small nuclear RNAs) + ~50 proteins  
 ↓  
 Four snRNPs (small nuclear ribonucleic particles), which assemble into a spliceosome

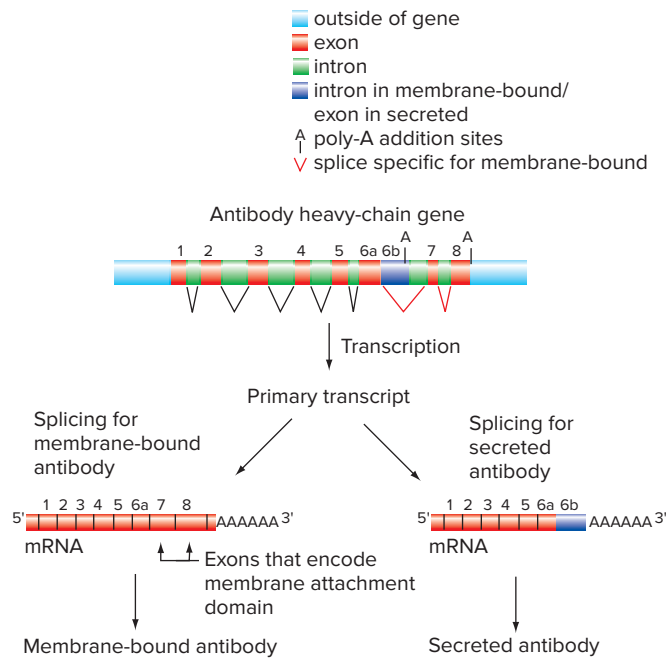


**Alternative splicing: Different mRNAs from the same primary transcript**

Sometimes RNA splicing joins together the splice donor and splice acceptor at the opposite ends of an intron, resulting in removal of the intron and fusion of two successive—and now adjacent—exons. Often, however, RNA splicing during development is regulated so that at certain times or in certain tissues, some splicing signals may be ignored. As an example, splicing may occur between the splice donor site of one intron and the splice acceptor site of a different intron downstream. Such **alternative splicing** produces different mRNA molecules that may encode related proteins with different—though partially overlapping—amino acid sequences and functions. In effect then, alternative splicing can tailor the nucleotide sequence of a primary transcript to produce more than one kind of polypeptide. Alternative splicing largely explains how the 27,000 genes in the human genome can encode the hundreds of thousands of different proteins estimated to exist in human cells.

In mammals, alternative splicing of the gene encoding the antibody heavy chain determines whether the antibody proteins become embedded in the membrane of the B lymphocyte that makes them or are instead secreted into the blood. The gene for antibody heavy chains has eight exons and seven introns; exon number 6 has a splice-donor site within it. To make the membrane-bound antibody, all exons except for the right-hand part of number 6 are joined to create an mRNA encoding a hydrophobic (water-hating, lipid-loving)

**Figure 8.17** Different mRNAs can be produced from the same primary transcript. Alternative splicing of the primary transcript for the antibody heavy chain produces mRNAs that encode different kinds of antibody proteins.



C terminus (Fig. 8.17). For the secreted antibody, only the first six exons (including the right part of 6) are spliced together to make an mRNA encoding a heavy chain with a hydrophilic (water-loving) C terminus. These two kinds of mRNAs formed by alternative splicing thus encode slightly different proteins that are directed to different parts of the body.

### essential concepts

- *Transcription* is the process by which *RNA polymerase* synthesizes a single-stranded *primary transcript* from a DNA template.
- *Transcription initiation* requires a DNA sequence called the *promoter* that signals *RNA polymerase* to begin copying. In eukaryotes, initiation requires an additional DNA sequence called an *enhancer*.
- During transcription *elongation*, *RNA polymerase* adds nucleotides to the growing RNA strand in the 5'-to-3' direction.
- A *terminator* in the RNA transcript tells *RNA polymerase* to cease transcription.
- In prokaryotes, the primary transcript is the *messenger RNA (mRNA)*.
- In eukaryotes, *RNA processing* after transcription produces a mature mRNA; the RNA transcript is modified by the addition of a 5' *cap* and a *poly-A tail*, along with the excision of *introns* when *exons* are joined by *splicing*.
- Exons can be spliced together in alternative ways; *alternative splicing* produces different mRNA sequences and therefore different polypeptides from the same primary transcript.

## 8.3 Translation: From mRNA to Protein

### learning objectives

1. Relate tRNA's structure to its function.
2. Describe the key steps of translation, indicating how each depends on the ribosome.
3. List three categories of posttranslational processing and provide examples of each.

*Translation* is the process by which the sequence of nucleotides in a messenger RNA directs the assembly of the correct sequence of amino acids in the corresponding polypeptide. Translation takes place on ribosomes that coordinate the movements of transfer RNAs carrying specific amino acids with the genetic instructions of an mRNA. As we examine the cell's translation machinery, we first describe the structure and function of tRNAs and ribosomes; and we then explain how these components interact during translation.

### Transfer RNAs Mediate the Translation of mRNA Codons to Amino Acids

No obvious chemical similarity or affinity exists between the nucleotide triplets of mRNA codons and the amino acids they specify. Rather, **transfer RNAs (tRNAs)** serve as adapter molecules that mediate the transfer of information from nucleic acid to protein.

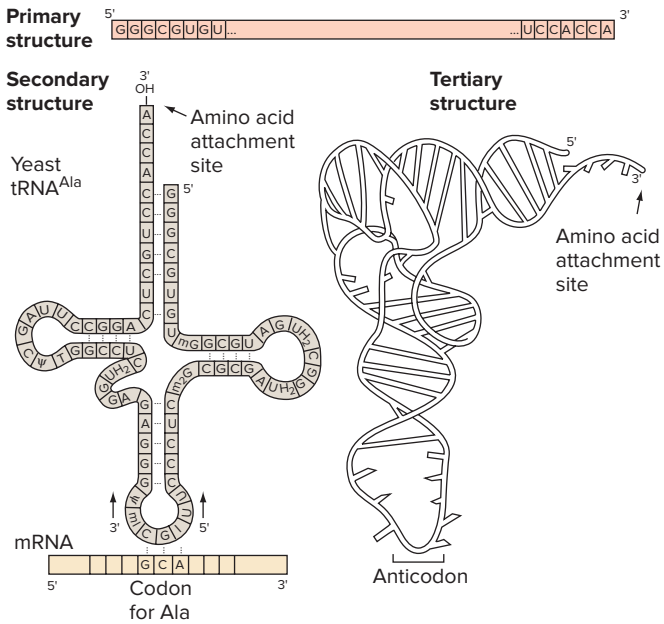
#### The structure of tRNA

Transfer RNAs are short, single-stranded RNA molecules 74–95 nucleotides in length. Several of the nucleotides in tRNAs contain chemically modified bases produced by enzymatic alterations of the principal A, G, C, and U nucleotides. Each tRNA carries one particular amino acid, and all cells must have at least one tRNA for each of the common 20 amino acids specified by the genetic code. The name of a tRNA reflects the amino acid it carries. For example, tRNA<sup>Gly</sup> carries the amino acid glycine.

As Fig. 8.18 shows, it is possible to consider the structure of a tRNA molecule on three levels.

1. The nucleotide sequence of a tRNA constitutes the primary structure.
2. Short complementary regions within a tRNA's single strand can form base pairs with each other to create a characteristic cloverleaf shape; this is the tRNA's secondary structure.
3. Folding in three-dimensional space creates a tertiary structure that looks like a compact letter L.

**Figure 8.18 tRNA structure.** The nucleotide sequence of a tRNA (the primary structure) folds to form characteristic secondary and tertiary structures. The anticodon and the amino acid attachment site are at opposite ends of the L-shaped tertiary structure. Several unusual bases of the tRNA, indicated as I,  $\psi$ , UH<sub>2</sub>, mI, m<sub>2</sub>G, and mG, are enzymatically modified variants of A, G, C, and U.



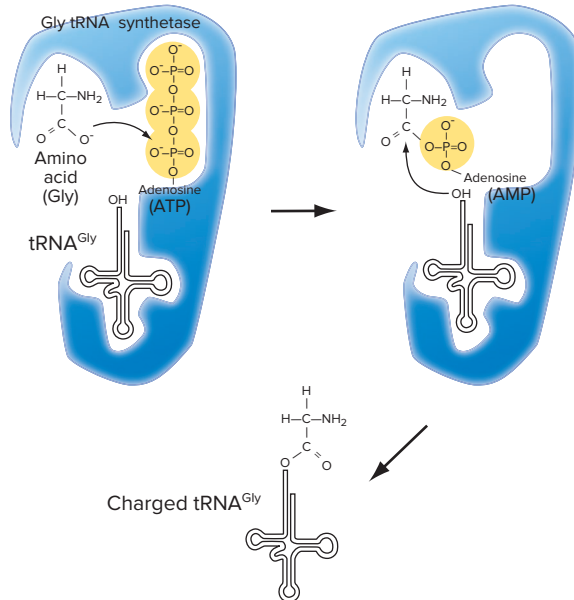
At one end of the L, the tRNA carries an **anticodon**: three nucleotides complementary to an mRNA codon specifying the amino acid carried by the tRNA (Fig. 8.18). The anticodon never forms base pairs with other regions of the tRNA; it is always available for base pairing with its complementary mRNA codon. As with other complementary base sequences, during pairing at the ribosome, the strands of anticodon and codon run antiparallel to each other. If, for example, the anticodon is 3' CCU 5', the complementary mRNA codon is 5' GGA 3', specifying the amino acid glycine. At the other end of the L, where the 5' and 3' ends of the tRNA strand are found (Fig. 8.18), the appropriate amino acid is attached to the tRNA's 3' end.

### Aminoacyl-tRNA synthetases: The molecular translators of the genetic code

Enzymes known as **aminoacyl-tRNA synthetases** connect the tRNA to the amino acid that corresponds to the anticodon. These enzymes are extraordinarily specific, recognizing unique features of a particular tRNA including the anticodon, while also recognizing the corresponding amino acid (see the opening figure of this chapter).

Aminoacyl-tRNA synthetases are, in fact, the only molecules that read the languages of both nucleic acid and protein. Normally, one aminoacyl-tRNA synthetase exists for each of the 20 common amino acids. Each synthetase functions with only one amino acid, but the enzyme may recognize several different tRNAs specific for that amino

**Figure 8.19 Aminoacyl-tRNA synthetases attach tRNAs to their corresponding amino acids.** The aminoacyl-tRNA synthetase has recognition sites for an amino acid, the corresponding tRNA, and ATP. The synthetase first activates the amino acid, forming an AMP-amino acid. The enzyme then transfers the amino acid's carboxyl group from AMP to the hydroxyl (–OH) group of the ribose at the 3' end of the tRNA, producing a charged tRNA.



acid. **Figure 8.19** shows the two-step process that establishes the covalent bond between an amino acid and the 3' end of its corresponding tRNA. A tRNA covalently coupled to its amino acid is called a **charged tRNA**. The bond between the amino acid and tRNA contains substantial energy that is later used to drive peptide bond formation.

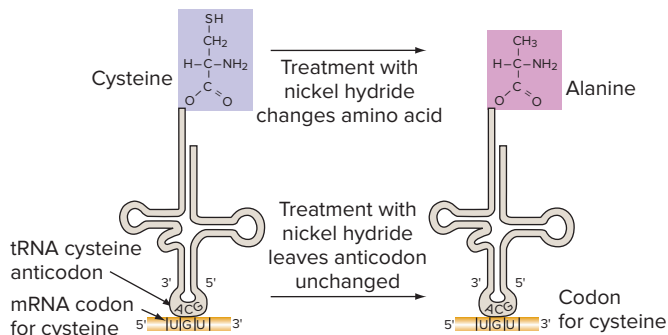
### The crucial role of base pairing between codon and anticodon

While attachment of the appropriate amino acid charges a tRNA, the amino acid itself does not play a significant role in determining where it becomes incorporated in a growing polypeptide chain. Instead, the specific interaction between a tRNA's anticodon and an mRNA's codon makes that decision. A simple experiment illustrates this point (**Fig. 8.20**). Researchers can subject a charged tRNA to chemical treatments that, without altering the structure of the tRNA, change the amino acid it carries. One treatment replaces the cysteine carried by tRNA<sup>Cys</sup> with alanine. When investigators then add the tRNA<sup>Cys</sup> charged with alanine to a cell-free translational system, the system incorporates alanine into the growing polypeptide wherever the mRNA contains a cysteine codon complementary to the anticodon of the tRNA<sup>Cys</sup>.

### Wobble: One tRNA, more than one codon

Although at least one kind of tRNA exists for each of the 20 common amino acids, cells do not necessarily carry tRNAs with anticodons complementary to all of the 61 possible

**Figure 8.20** Base pairing between an mRNA codon and a tRNA anticodon determines which amino acid is added to a growing polypeptide. A tRNA with an anticodon for cysteine, but carrying the amino acid alanine, adds alanine whenever the mRNA codon for cysteine appears.



codon triplets in the genetic code. *E. coli*, for example, makes 79 different tRNAs containing 42 different anticodons. Although several of the 79 tRNAs in this collection obviously have the same anticodon,  $61 - 42 = 19$  of 61 potential anticodons are not represented. Thus 19 mRNA codons will not find a complementary anticodon in the *E. coli* collection of tRNAs. How can an organism construct proper polypeptides if some of the codons in its mRNAs cannot locate tRNAs with complementary anticodons?

The answer is that some tRNAs can recognize more than one codon for the amino acid with which they are charged. That is, the anticodons of these tRNAs can interact with more than one codon for the same amino acid, in keeping with the degenerate nature of the genetic code. Francis Crick spelled out a few of the rules that govern the promiscuous base pairing between codons and anticodons.

Crick reasoned first that the 3' nucleotide in many codons adds nothing to the specificity of the codon. For example, 5' GGU 3', 5' GGC 3', 5' GGA 3', and 5' GGG 3' all encode glycine (review Fig. 8.2). It does not matter whether the 3' nucleotide in the codon is U, C, A, or G as long as the first two letters are GG. The same is true for other amino acids encoded by four different codons, such as valine, where the first two bases must be GU, but the third base can be U, C, A, or G.

For amino acids specified by two different codons, the first two bases of the codon are, once again, always the same, while the third base must be either one of the two purines (A or G) or one of the two pyrimidines (U or C). Thus, 5' CAA 3' and 5' CAG 3' are both codons for glutamine; 5' CAU 3' and 5' CAC 3' are both codons for histidine. If Pu stands for either purine and Py stands for either pyrimidine, then CAPu represents the codons for glutamine, while CAPy represents the codons for histidine.

In fact, the 5' nucleotide of a tRNA's anticodon can often pair with more than one kind of nucleotide in the 3' position of an mRNA's codon. (Recall that after base pairing, the bases in the anticodon run antiparallel to the bases

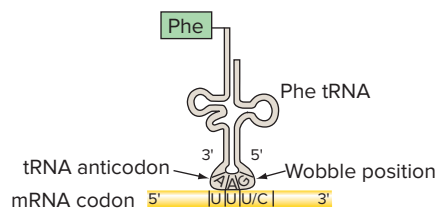
in the codon.) A single tRNA charged with a particular amino acid can thus recognize several or even all of the codons for that amino acid. This flexibility in base pairing between the 3' nucleotide in the codon and the 5' nucleotide in the anticodon is known as **wobble** (Fig. 8.21a). The combination of normal base pairing at the first two positions of a codon with wobble at the third position clarifies why multiple codons for a single amino acid usually start with the same two letters.

An important aspect of wobble is the chemical modification of certain bases at the 5' end of the anticodon (the *wobble position*) (Fig. 8.21b and c). An A in the wobble position of a tRNA is almost always modified to inosine

**Figure 8.21** Wobble: Some tRNAs recognize more than one codon for the same amino acid. (a) The G at the 5' end of the anticodon shown here can pair with either U or C at the 3' end of the codon. (b) The table shows the pairing possibilities for nucleotides at the 5' end of an anticodon (the wobble position).

xo<sup>5</sup>U only rarely pairs with C. k<sup>2</sup>C occurs only in certain bacteria. (c) Chemical structures of the modified bases in anticodons.

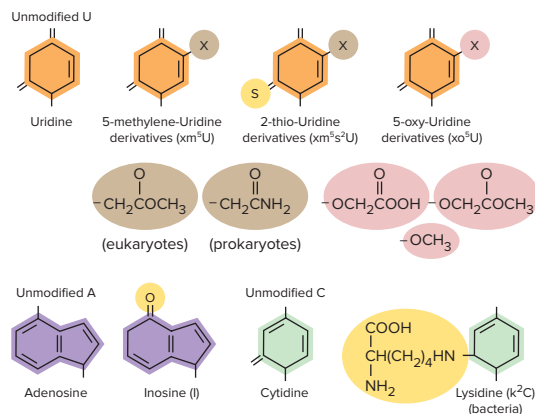
(a)



(b)

Wobble Rules		
5' end of anticodon	can pair with	3' end of codon
G		U or C
C		G
I		U, C, or A
xm <sup>5</sup> U		G
xm <sup>5</sup> s <sup>2</sup> U		A or G
xo <sup>5</sup> U		A, G, U, or (C)
k <sup>2</sup> C		A

(c) Modified bases in anticodon wobble position



(I), and a U in the wobble position is always modified in one of three possible ways. By contrast, G in the anticodon wobble position is always unmodified, while modification of C occurs only in the tRNAs of some bacterial species. Wobble bases are modified by specific enzymes that act on the tRNA after it has been synthesized by transcription.

The wobble rules in Fig. 8.21c delimit the anticodon sequences and the wobble base modifications consistent with the genetic code. For example, methionine (Met) is specified by a single codon (5' AUG 3'). As a result, Met-specific tRNAs must either have a C at the 5' end of their anticodons (5' CAU 3') or a U that is modified to xm<sup>5</sup>U, because these are the only nucleotides at that position that can base pair only with the G at the 3' end of the Met codon. By contrast, a single isoleucine-specific tRNA with the modified nucleotide inosine (I) at the 5' position of the anticodon can recognize all three codons (5' AUU 3', 5' AUC 3', and 5' AUA 3') for isoleucine.

### A special tRNA for selenocysteine

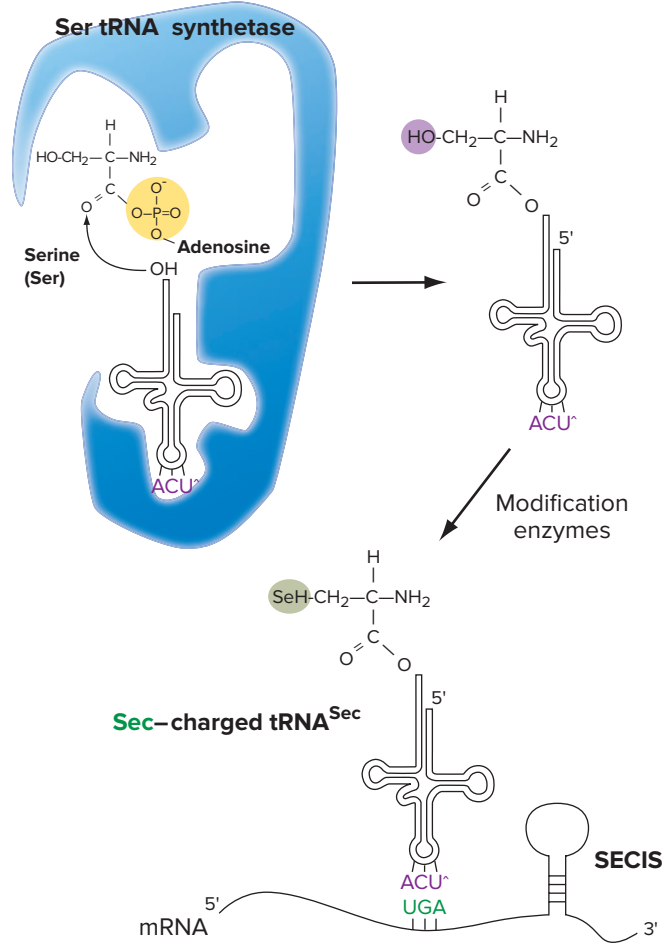
Most mRNAs direct the synthesis of proteins containing only the 20 common amino acids. Exceptional mRNAs in bacteria and eukaryotes direct the synthesis of *selenoproteins*, which contain the amino acid selenocysteine (Sec), sometimes referred to as amino acid 21. Selenoproteins are rare; in humans, only 25 are known to exist.

As shown in Fig. 8.22, a dedicated selenocysteine tRNA (tRNA<sup>Sec</sup>) is recognized by serine tRNA synthetase and charged with serine. Modification enzymes subsequently convert the Ser to Sec. The Sec-charged tRNA<sup>Sec</sup> interacts with 5' UGA 3' triplets found only in mRNAs that contain a special structure called the *Sec insertion sequence* (SECIS) element. The SECIS element is a region of the mRNA that forms a particular stem-loop (hairpin) structure through intramolecular complementary base pairing (Fig. 8.22). This stem loop prevents termination of polypeptide synthesis at the UGA triplet, which would otherwise act as a stop codon. The anticodon of the Sec-charged tRNA<sup>Sec</sup> binds to the UGA triplet in the mRNA, allowing the incorporation of Sec into the polypeptide product.

### Ribosomes Are the Sites of Polypeptide Synthesis

Ribosomes facilitate polypeptide synthesis in various ways. First, they recognize mRNA features that signal the start of translation. Second, they help ensure accurate interpretation of the genetic code by stabilizing the interactions between tRNAs and mRNAs; without a ribosome, codon-anticodon recognition, mediated by only three base pairs, would be extremely weak. Third, ribosomes supply the enzymatic activity that links the amino acids in a growing polypeptide chain. Fourth, by moving 5' to 3' along an mRNA molecule, they expose the mRNA codons in

**Figure 8.22** How rare proteins incorporate selenocysteine. (a) The serine carried by tRNA<sup>Sec</sup> with the anticodon 5' UCA 3' is modified to selenocysteine (Sec). The Sec-charged tRNA recognizes the triplet UGA only in rare mRNAs with a downstream SECIS element. The U in the wobble position of this tRNA is modified in an unusual manner (indicated as U<sup>⌘</sup>) and so it recognizes only A.



sequence, ensuring the linear addition of amino acids. Finally, ribosomes help end polypeptide synthesis by dissociating both from the mRNA directing polypeptide construction and from the polypeptide product itself.

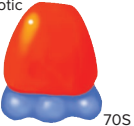
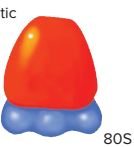
### The structure of ribosomes

In *E. coli*, ribosomes consist of three different ribosomal RNAs (rRNAs) and 52 different ribosomal proteins (Fig. 8.23a). These components associate to form two different ribosomal subunits called the 30S subunit and the 50S subunit. (S designates a coefficient of sedimentation related to the size and shape of the subunit; the 30S subunit is smaller than the 50S subunit). Before translation begins, the two subunits exist as separate entities in the cytoplasm. Soon after the start of translation, they come together to reconstitute a complete ribosome. Eukaryotic ribosomes have more components than their prokaryotic counterparts, but they still consist of two dissociable subunits.

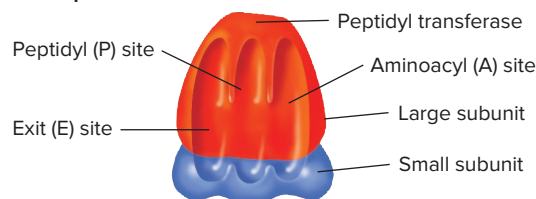


**Figure 8.23 The ribosome: Site of polypeptide synthesis.** (a) A ribosome has two subunits, each composed of rRNA and various proteins. (b) The small subunit initially binds to mRNA. The large subunit contributes the enzyme peptidyl transferase, which catalyzes the formation of peptide bonds. The two subunits together form the A, P, and E tRNA binding sites.

(a) A ribosome has two subunits composed of RNA and protein.

Complete Ribosomes	Subunits	Nucleotides	Proteins
Prokaryotic 	50S	23S rRNA 3000 nucleotides 5S rRNA 120 nucleotides	31
	30S	16S rRNA 1700 nucleotides	21
Eukaryotic 	60S	28S rRNA 5000 nucleotides 5.8S rRNA 160 nucleotides	~45
	40S	5S rRNA 120 nucleotides 18S rRNA 2000 nucleotides	~33

(b) Different parts of a ribosome have different functions.



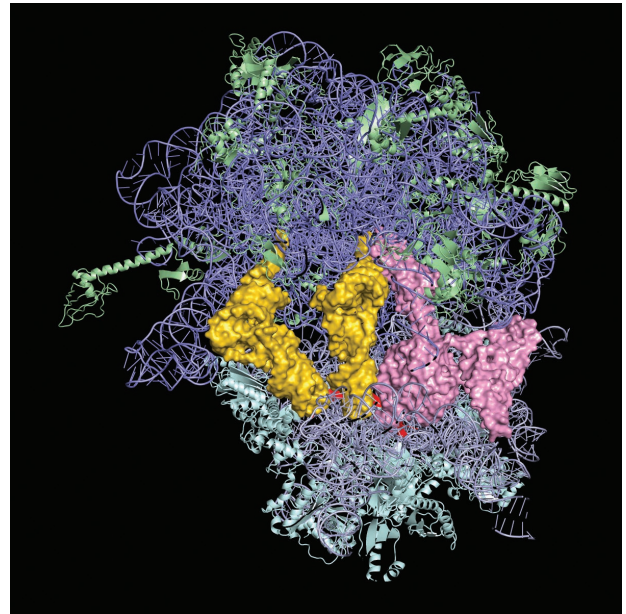
### Functional domains of ribosomes

The small 30S subunit is the part of the ribosome that initially binds to mRNA. The larger 50S subunit contributes an enzyme known as **peptidyl transferase**, which catalyzes formation of the peptide bonds joining adjacent amino acids (Fig. 8.23b). Both the small and the large subunits contribute to three distinct tRNA binding areas known as the **aminoacyl** (or **A**) **site**, the **peptidyl** (or **P**) **site**, and the **exit** (or **E**) **site**. Finally, other regions of the ribosome distributed over the two subunits serve as points of contact for some of the additional proteins that play roles in translation.

Using X-ray crystallography and elegant techniques of electron microscopy, researchers have gained a remarkably detailed view of the complicated structure of the ribosome. Figure 8.24 shows the interior of a ribosome nearing the completion of the translation of an mRNA; some parts of the ribosome that extend out of the page toward the reader were removed computationally so you can better see the tRNAs occupying the E site and P site.

With this illustration, you can see that the rRNAs occupy most of the space in the central part of the ribosome, while the various ribosomal proteins are studded around the exterior. Surprisingly, no proteins are found close to the region where the peptide bonds are formed during translation. This finding supports the conclusions of biochemical experiments that peptidyl transferase is actually a function

**Figure 8.24 High resolution view of a bacterial ribosome in action.** The large subunit is at the top; its 23S and 5S rRNA components are in *bright blue* and its various protein components in *bright green*. The small subunit is at the bottom, with the 16S rRNA in *gray* and its protein components in *aqua*. Two tRNA molecules are in *gold*, with the tRNA on the left at the E site and the tRNA to its right at the P site. The A site is occupied by a protein release factor shown in *pink*. A few of the nucleotides in the mRNA (*red*) can be seen near the bottoms of the tRNAs and release factor. This ribosome is acting during a stage of translation just prior to the ejection of the tRNA from the E site during the termination phase depicted in the left panel of Fig. 8.25c. © Yuxin Mao, Ph.D., Cornell University, Ithaca, NY



of the large subunit's rRNA rather than any protein component of the ribosome. In other words, the rRNA acts as a *ribozyme* that joins amino acids together.

During translation, the ribosome associates briefly with various proteins that aid steps in the process. For example, Fig. 8.24 shows that late in translation when the completed polypeptide is released from the ribosome, a protein called a *release factor* binds to the ribosome's A site. Remarkably, the release factor can associate with the A site because part of this protein folds in three-dimensional space in a way that mimics the structure of a tRNA.

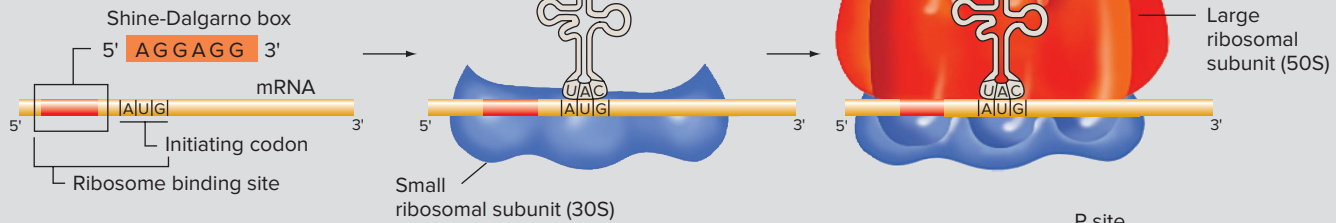
### Ribosomes and Charged tRNAs Collaborate to Translate mRNAs into Polypeptides

As was the case for transcription, translation consists of three phases: an **initiation** phase that sets the stage for polypeptide synthesis; **elongation**, during which amino acids are added to a growing polypeptide; and a **termination** phase that brings polypeptide synthesis to a halt and enables the ribosome to release a completed chain of amino acids. Figure 8.25 illustrates the details of the process, focusing on translation as it occurs in bacterial cells. As

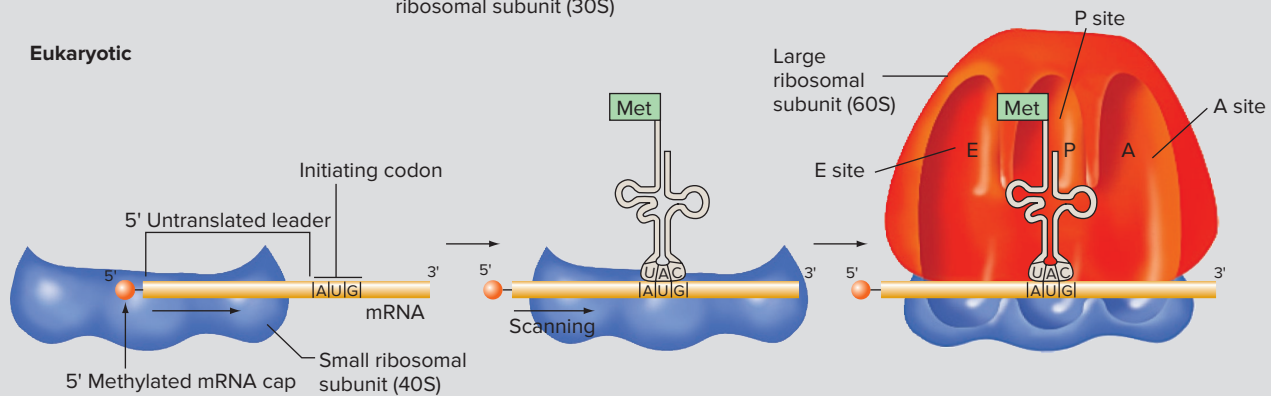
Translation of mRNAs on Ribosomes

Initiation phase

Prokaryotic



Eukaryotic



**(a) Initiation: Setting the stage for polypeptide synthesis** The first three nucleotides of an mRNA are not the first codon. Instead, a special signal indicates where along the mRNA translation should begin. In prokaryotes, this signal is the **ribosome binding site**, and it has two important elements. The first is a short sequence of six nucleotides—usually 5' . . . AGGAGG . . . 3'—named the **Shine-Dalgarno box**. The second element in an mRNA's ribosome binding site is the triplet 5' AUG 3', which serves as the initiation codon.

The 5' CAU 3' anticodon of a special initiator tRNA recognizes the AUG in the ribosome binding site. The initiator tRNA carries **N-formylmethionine (fMet)**, a modified methionine whose amino end is blocked by a formyl group. A different tRNA that is charged with an unmodified methionine recognizes AUG codons located within an mRNA's reading frame; this tRNA cannot start translation.

During initiation, the 3' end of the 16S rRNA in the 30S ribosomal subunit binds to the mRNA's Shine-Dalgarno box (*not shown*), the fMet tRNA binds to the mRNA's initiation codon, and a large 50S ribosomal subunit associates with the small subunit to round out the ribosome. At the end of initiation, the fMet tRNA sits in the P site of the completed ribosome. Proteins known as **initiation factors** (*not shown*) play a transient role in the initiation process.

In eukaryotes, the small ribosomal subunit binds first to the methylated cap at the 5' end of the mature mRNA. The small subunit then migrates to the initiation site—usually the first AUG it encounters as it scans the mRNA in the 5'-to-3' direction. The initiator tRNA in eukaryotes carries unmodified methionine (Met) instead of fMet.

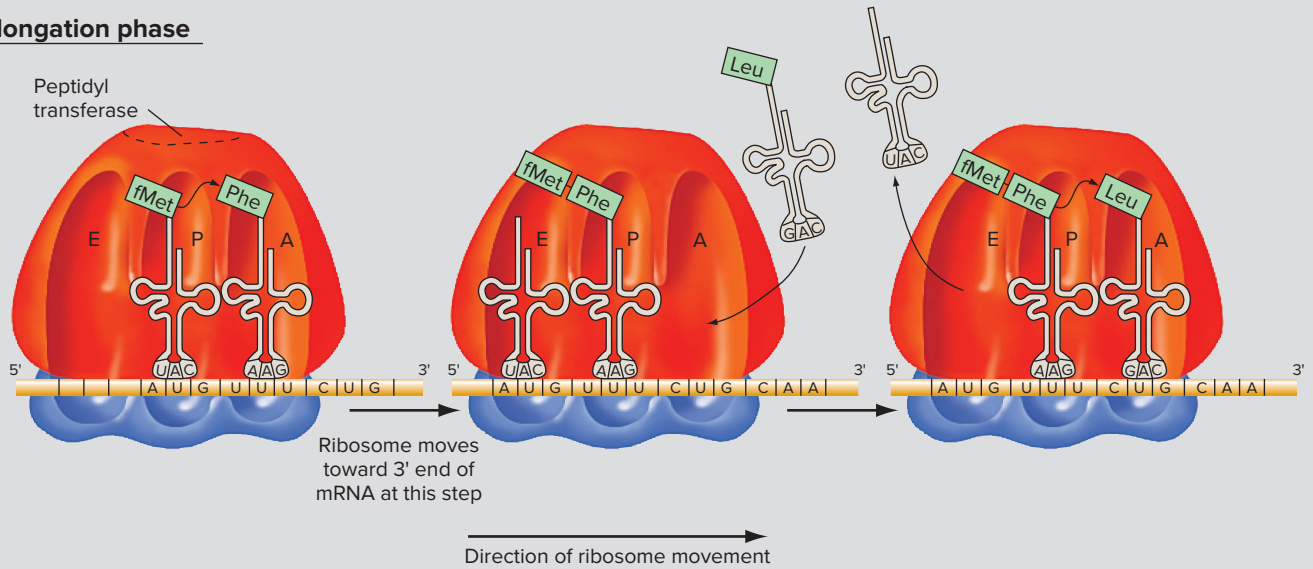
**(b) Elongation: The addition of amino acids to a growing polypeptide** Proteins known as **elongation factors** (*not shown*) usher the appropriate tRNA into the A site of the ribosome. The anticodon of this charged tRNA must recognize the next codon in the mRNA. The ribosome simultaneously holds the initiating tRNA at its P site and the second tRNA at its A site so that peptidyl transferase can catalyze formation of a peptide bond between the amino acids carried by the two tRNAs. As a result, the tRNA at the A site now carries two amino acids. The N terminus of this dipeptide is fMet or Met; the C terminus is the second amino acid, whose carboxyl group remains covalently linked to its tRNA.

After the first peptide bond forms, the ribosome moves with the help of elongation factors, exposing the next mRNA codon. As the ribosome moves, the initiating tRNA, which no longer carries an amino acid, is transferred to the E site, and the other tRNA, which carries the dipeptide, shifts from the A site to the P site.

The empty A site now receives another tRNA, whose identity is determined by the next codon in the mRNA. The uncharged initiating tRNA is bumped off the E site and leaves the ribosome. Peptidyl transferase then catalyzes the formation of a second peptide bond, generating a chain of three amino acids connected at its C terminus to the tRNA currently in the A site. With each subsequent round of ribosome movement and peptide bond formation, the peptide chain grows one amino acid longer. Note that each tRNA moves from the A site to the P site to the E site (excepting the initiating tRNA, which first enters the P site).

Because the ribosome adds amino acids to the C terminus of the growing chain, polypeptide synthesis proceeds from the N terminus to the C terminus. As a result, the initial fMet in prokaryotes (Met in eukaryotes), will be the N-terminal amino

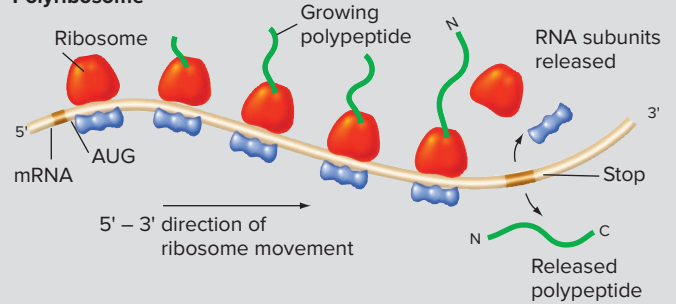
## Elongation phase



acid of all finished polypeptides prior to protein processing. Moreover, the ribosome must move along the mRNA in the 5'-to-3' direction so that the polypeptide can grow in the N-to-C direction.

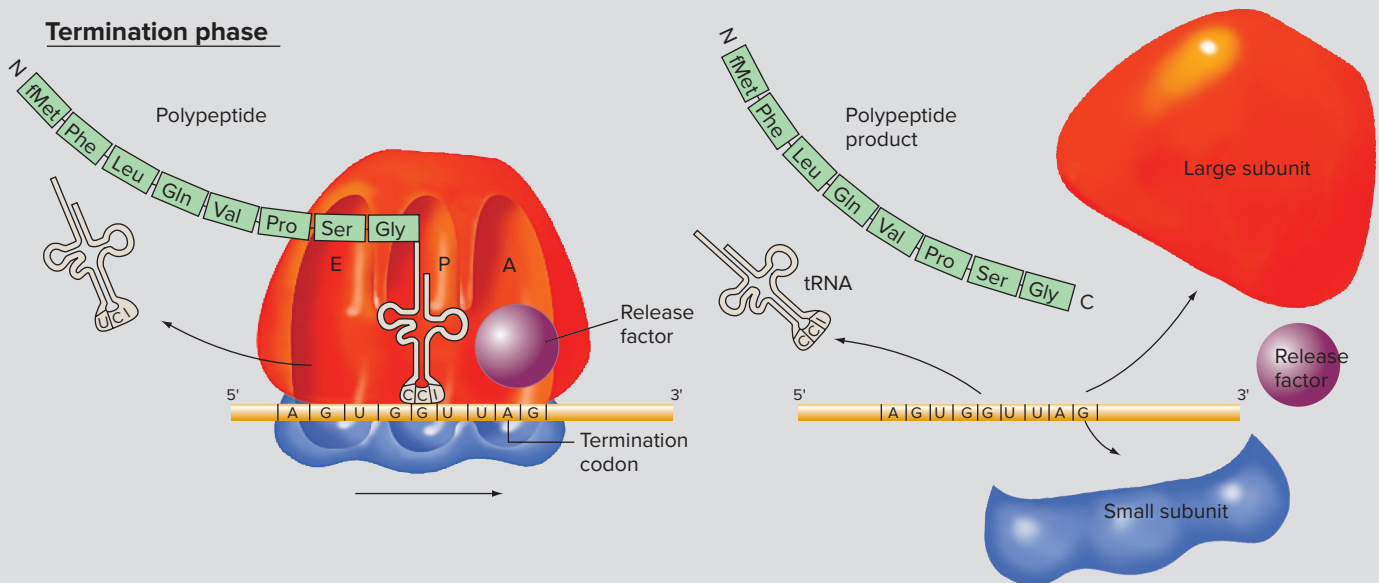
Once a ribosome has moved far enough away from the mRNA's ribosome binding site, that site becomes accessible to other ribosomes. In fact, several ribosomes can work on the same mRNA at one time. A complex of several ribosomes translating from the same mRNA is called a **polyribosome**. This complex allows the simultaneous synthesis of many copies of a polypeptide from a single mRNA.

## Polyribosome



- (c) **Termination: The ribosome releases the completed polypeptide** No normal tRNAs (except for tRNA<sup>Sec</sup>) carry anticodons complementary to any of the three nonsense (stop) codons UAG, UAA, and UGA. Thus, when a nonsense codon moves into the ribosome's A site, no tRNAs can bind to that codon. Instead, proteins called **release factors** recognize the stop codons and halt polypeptide synthesis. The tRNA specifying the C-terminal amino acid releases the completed polypeptide, the same tRNA as well as the mRNA separate from the ribosome, and the ribosome dissociates into its large and small subunits.

## Termination phase



you examine the figure, note the following points about the flow of information during translation:

- The first codon to be translated—the **initiation codon**—is an AUG set in a special context at the 5' end of the gene's reading frame (never precisely at the 5' end of the mRNA).
- Special initiating tRNAs carrying a modified form of methionine called *formylmethionine* (*fMet*) recognize the initiation codon.
- The ribosome moves along the mRNA in the 5'-to-3' direction, revealing successive codons in a stepwise fashion.
- At each step of translation, the polypeptide grows by the addition to its C terminus of the next amino acid in the chain.
- Translation terminates when the ribosome reaches a UAA, UAG, or UGA nonsense codon at the 3' end of the gene's reading frame.

These points explain the biochemical basis of colinearity, that is, the correspondence between the 5'-to-3' direction in the mRNA and the N-terminus-to-C-terminus direction in the resulting polypeptide.

During elongation, the translation machinery adds about 2–15 amino acids per second to the growing chain. The speed is higher in prokaryotes and lower in eukaryotes. At these rates, construction of an average-size 300-amino-acid polypeptide (from an average-length mRNA that is about 1000 nucleotides) could take as little as 20 seconds or as long as 2.5 minutes.

Several details have been left out of Fig. 8.25 so that you can concentrate on the flow of information during translation. In particular, this figure does not depict the

important roles played by protein translation factors, which help shepherd mRNAs and tRNAs to their proper locations on the ribosome. Some translation factors also carry GTP to the ribosome. Hydrolysis of the high-energy bonds in the GTP helps power certain molecular movements, including translocation of the ribosome along the mRNA.

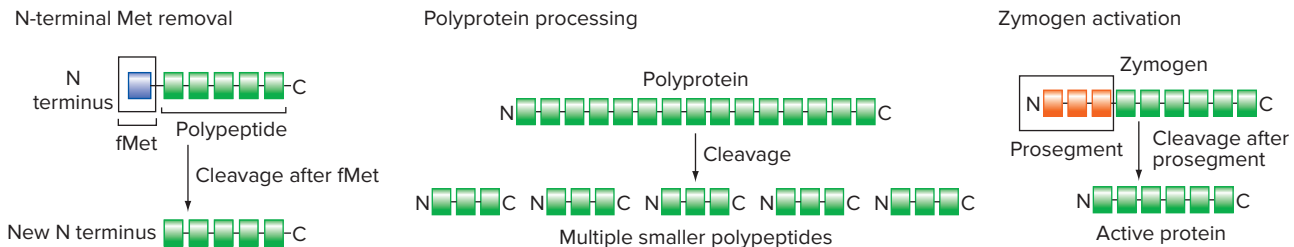
## Polypeptides Can Be Modified After Translation

Protein structure is not irrevocably fixed at the completion of translation. Several different processes may subsequently modify a polypeptide's structure. Cleavage may remove amino acids, such as the N-terminal *fMet*, from a polypeptide, or it may generate several smaller polypeptides from one larger product of translation (**Fig. 8.26a**). In the latter case, the larger polypeptide made before it is cleaved into smaller polypeptides is often called a **polyprotein**. Also, some proteins are synthesized in inactive forms called *zymogens* that are activated by enzymatic cleavage that removes an N-terminal *prosegment*.

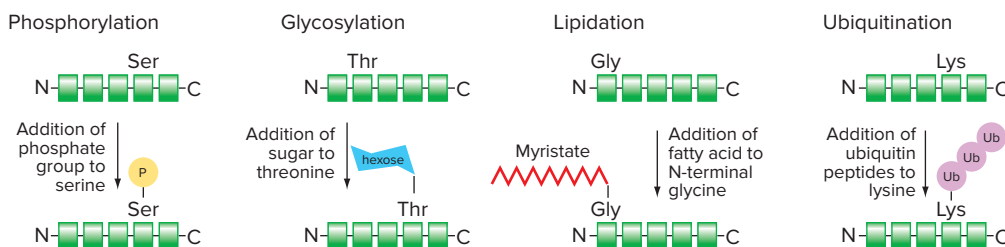
Enzymatic addition of chemical constituents, such as phosphate groups, carbohydrates, fatty acids, or even other small peptides to specific amino acids may also modify a polypeptide after translation (**Fig. 8.26b**). Such changes to polypeptides are known as **posttranslational modifications**. Posttranslational changes to a protein can be very important: For example, the biochemical function of many enzymes depends directly on the addition (or sometimes removal) of phosphate groups. Posttranslational modifications may alter the way a protein folds, its ability to interact with other proteins, its stability, its activity, or its location in the cell.

**Figure 8.26** Posttranslational processing can modify polypeptide structure. (a) Enzymatic cleavage processes many proteins into their mature forms. (b) Enzymes add various functional groups to specific amino acids.

### (a) Enzymatic cleavage may remove an amino acid, split a polyprotein, or activate a zymogen.



### (b) Addition of chemical constituents may alter protein structure, activity, or cellular location.



### essential concepts

- *Translation* is the process by which *ribosomes* synthesize proteins according to the instructions in mRNAs. Ribosomes have specific binding sites for *tRNAs* (the A, P, and E sites) and supply *peptidyl transferase*, the ribozyme that forms peptide bonds between amino acids.
- *Transfer RNAs (tRNAs)* are the adapters that link mRNA codons to amino acids at the ribosome. *Aminoacyl-tRNA synthetases* connect the correct amino acids to their corresponding tRNAs.
- Each tRNA has an *anticodon* complementary to the mRNA codon specifying the particular amino acid. Because of *wobble*, a tRNA may recognize more than one codon.
- Translation *initiation* begins when a charged tRNA<sup>Met</sup> (or tRNA<sup>Met</sup>) binds the *start codon*, AUG, at the ribosomal P site.
- During *elongation*, the amino acid connected to the tRNA at the P site forms a peptide bond with the amino acid connected to the tRNA at the A site. The ribosome then moves in the 5'-to-3' direction along the mRNA to the next codon.
- *Termination* occurs when the ribosome encounters an in-frame *stop codon* in the mRNA.
- *Posttranslational processing* enzymes may cleave a polypeptide or add chemical constituents to it.

## 8.4 Differences in Gene Expression Between Prokaryotes and Eukaryotes

### learning objectives

1. Explain how the nuclear membrane affects gene expression in eukaryotes.
2. Discuss the function of enhancer sequences in eukaryotic transcription.
3. Describe the differences in translation initiation between prokaryotes and eukaryotes.
4. List the steps in mRNA formation that occur in eukaryotes but not in prokaryotes.

The processes of transcription and translation in eukaryotes and prokaryotes are similar in many ways but also are affected by certain differences, including (1) the presence of a nuclear membrane in eukaryotes, (2) eukaryotic-specific complexities in the mechanisms by which RNA polymerase recognizes promoters to start transcription, (3) variations in the way in which translation is initiated, and (4) the need for additional transcript processing in eukaryotes.

### In Eukaryotes, the Nuclear Membrane Prevents the Coupling of Transcription and Translation

In *E. coli* and other prokaryotes, transcription takes place in an open intracellular space undivided by a nuclear membrane. Translation occurs in the same open space and is sometimes coupled directly with transcription (**Table 8.1**). This coupling is possible because transcription extends mRNAs in the same 5'-to-3' direction as the ribosome moves along the mRNA. As a result, ribosomes can begin to translate a partial mRNA that the RNA polymerase is still in the process of transcribing from the DNA.

The coupling of transcription and translation has significant consequences for the regulation of gene expression in prokaryotes. For example, in an important regulatory mechanism called *attenuation*, which we describe in Chapter 16, the rate of translation of some mRNAs directly determines the rate at which the corresponding genes are transcribed into these mRNAs.

Such coupling cannot occur in eukaryotes because the nuclear envelope physically separates the sites of transcription and RNA processing in the nucleus from the site of translation in the cytoplasm. As a result, translation in eukaryotes can affect the rate at which genes are transcribed only in more indirect ways.

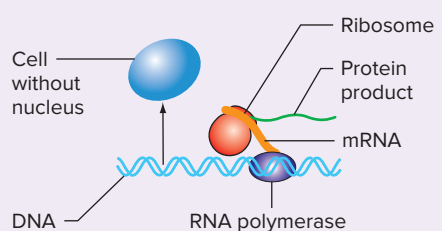
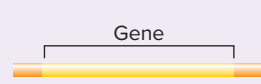
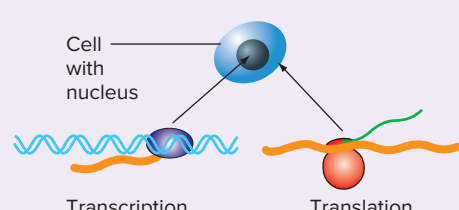
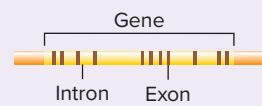
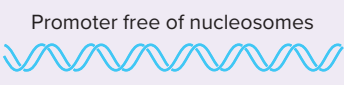

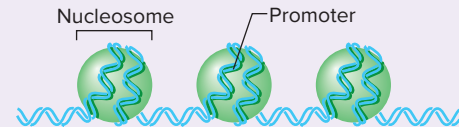

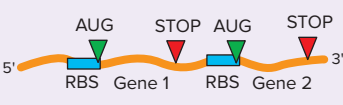
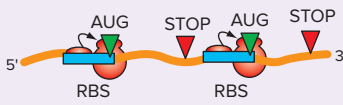

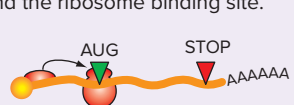
### Distant Enhancer Sequences and Interactions with Chromatin Influence Eukaryotic Promoters

In eukaryotes, the promoters recognized by RNA polymerase to initiate transcription are affected by two situations not seen in prokaryotes (Table 8.1). First, as previously seen in Fig. 8.11, the stability of RNA polymerase's interaction with the promoter is often affected by enhancer sequences located far from the promoter. In prokaryotes, the DNA sequences that regulate transcription are all found much closer to the promoter. Second, eukaryotic chromosomes are tightly wound around *histone proteins* in a DNA/protein complex called *chromatin*. To be recognized by RNA polymerase, the promoter of a eukaryotic gene must first be unwound from chromatin. Interestingly, clearing the histones from the promoter is an important function of enhancers. (Histones and chromatin and their roles in transcription will be discussed in Chapters 12 and 17.)

### Prokaryotes and Eukaryotes Initiate Translation Differently

In prokaryotes, translation begins at a ribosome binding site on the mRNA, which is defined by a short, characteristic sequence of nucleotides called a *Shine-Dalgarno box* adjacent to an initiating AUG codon (review Fig. 8.25a).

**TABLE 8.1** Differences Between Prokaryotes and Eukaryotes in the Details of Gene Expression

	Prokaryotes	Eukaryotes
<b>Overview</b>	<p>1. No nucleus. Transcription and translation take place in the same cellular compartments, and translation is often coupled to transcription.</p>  <p>2. Genes are not divided into exons and introns.</p> 	<p>1. Nucleus separated from the cytoplasm by a nuclear membrane. Transcription takes place in the nucleus, while translation occurs in the cytoplasm. Direct coupling of transcription and translation is not possible.</p>  <p>2. The DNA of a gene consists of exons separated by introns; the exons are defined by posttranscriptional splicing, which deletes the introns.</p> 
<b>Transcription</b>	<p>1. One RNA polymerase consisting of five subunits.</p> <p>2. DNA sequences needed for transcription initiation are located close to the promoter.</p> <p>3. Promoters are not wound up in chromatin.</p>  <p>4. Primary transcripts are the actual mRNAs; they have a triphosphate start at the 5' end and no tail at the 3' end.</p> 	<p>1. Several kinds of RNA polymerase, each containing 10 or more subunits; different polymerases transcribe different genes.</p> <p>2. Enhancer sequences far from the promoter are often needed for transcription initiation.</p> <p>3. Transcription initiation requires promoters to be cleared of chromatin to allow access to RNA polymerase.</p>  <p>4. Primary transcripts undergo processing to produce mature mRNAs that have a methylated cap at the 5' end and a poly-A tail at the 3' end.</p> 
<b>Translation</b>	<p>1. Unique initiator tRNA carries formylmethionine.</p> <p>2. mRNAs have multiple ribosome binding sites (RBSs) and can thus direct the synthesis of several different polypeptides.</p>  <p>3. Small ribosomal subunit immediately binds to the mRNA's ribosome binding site.</p> 	<p>1. Initiator tRNA carries methionine.</p> <p>2. mRNAs have only one start site and can thus direct the synthesis of only one kind of polypeptide.</p>  <p>3. Small ribosomal subunit binds first to the methylated cap at the 5' end of the mature mRNA and then scans the mRNA to find the ribosome binding site.</p> 

There is nothing to prevent an mRNA from having more than one ribosome binding site, and in fact, many prokaryotic messages are **polycistronic**: They contain the information of several genes (sometimes referred to as *cistrons*), each of which can be translated independently starting at its own ribosome binding site (Table 8.1).

In eukaryotes, by contrast, the small ribosomal subunit first binds to the methylated cap at the 5' end of the mature mRNA and then migrates through the 5' UTR to the initiation site. This site is almost always the first AUG codon encountered by the ribosomal subunit as it moves along, or *scans*, the mRNA in the 5'-to-3' direction (see Fig. 8.25a and Table 8.1). Because of this scanning mechanism, initiation in eukaryotes takes place at only a single site on the mRNA, and each mRNA is **monocistronic**—it contains the information for translating only a single kind of polypeptide.

Another difference between prokaryotic and eukaryotic translation is in the composition of the initiating tRNA. In prokaryotes, as already mentioned, this tRNA carries a modified form of methionine known as *N*-formylmethionine, while in eukaryotes, it carries an unmodified methionine (Table 8.1). Thus, immediately after translation, eukaryotic polypeptides all have Met (instead of fMet) at their N termini. Posttranslational cleavage in both prokaryotes and eukaryotes, however, often creates mature proteins that no longer have N-terminal fMet or Met (see Fig. 8.26a).

## Eukaryotic mRNAs Require More Processing than Prokaryotic mRNAs

Table 8.1 reviews other important differences in gene structure and expression between prokaryotes and eukaryotes. In particular, introns interrupt eukaryotic, but not prokaryotic, genes such that the splicing of a primary transcript is necessary for eukaryotic gene expression. Other types of RNA processing that occur in eukaryotes but not prokaryotes add a methylated cap and a poly-A tail, respectively, to the 5' and 3' ends of the mRNAs.

### essential concepts

- In prokaryotes, transcription and translation occur simultaneously. In eukaryotes, the nuclear membrane restricts transcription to the nucleus; mRNAs are translated only after transport into the cytoplasm.
- In eukaryotes, transcription initiation involves *enhancer sequences* located far from the promoter. In addition, the *chromatin* of eukaryotic chromosomes must be unwound to allow access by RNA polymerase.
- Prokaryotic mRNAs are *polycistronic* such that ribosomes can translate several different polypeptides from a single mRNA. Eukaryotes have *monocistronic* mRNAs that can be used to translate only a single protein.
- In prokaryotes, ribosomes bind to a sequence called the *Shine-Dalgarno box* adjacent to the AUG initiation codon.

In eukaryotes, the small ribosome subunit binds at the 5' cap and migrates until it encounters the initiation site.

- In prokaryotes, the primary transcripts are mRNAs immediately ready for translation. In eukaryotes, primary transcripts are processed prior to translation into mature mRNAs through the addition of 5' caps and poly-A tails, as well as the removal of introns.

## 8.5 The Effects of Mutations on Gene Expression and Function

### learning objectives

1. Compare silent mutations, missense mutations, nonsense mutations, and frameshift mutations in terms of how they alter a gene product.
2. Discuss mutations outside the coding sequence that could affect gene expression.
3. Explain why most loss-of-function alleles (hypomorphic or amorphic) are recessive to wild-type alleles, but some are incompletely dominant or dominant.
4. Contrast the actions of hypermorphic, neomorphic, and antimorphic gain-of-function alleles.
5. Give examples of mutations that can have global effects on gene expression.

We have seen that the information in DNA is the starting point of gene expression. The cell transcribes that information into mRNA and then translates the mRNA information into protein. Mutations that alter the nucleotide pairs of DNA can modify any of the steps or products of gene expression.

## Mutations in a Gene's Coding Sequence May Alter the Gene Product

Because of the nature of the genetic code, mutations in a gene's amino acid–encoding exons generate a range of repercussions (**Fig. 8.27a**).

### Silent mutations

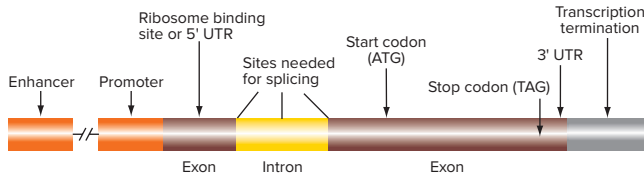
One consequence of the code's degeneracy is that some mutations, known as **silent mutations**, can change a codon into a mutant codon that specifies exactly the same amino acid. The majority of silent mutations change the third nucleotide of a codon, the position at which most codons for the same amino acid differ. For example, a change from GCA to GCC in a codon would still yield alanine in the protein product. Because silent mutations do not alter the amino acid composition of the encoded polypeptide, such mutations usually affect neither gene expression nor phenotype.

**Figure 8.27** How mutations in a gene can affect its expression. **(a)** Mutations in a gene's coding sequences. *Silent mutations* do not alter the protein's primary structure. *Missense mutations* replace one amino acid with another. *Nonsense mutations* shorten a polypeptide by replacing a codon with a stop signal. *Frameshift mutations* change the reading frame downstream of the addition or deletion. **(b)** Mutations outside the coding region can also disrupt gene expression.

**(a) Types of mutation in a gene's coding sequence**

Wild-type mRNA	5' ATG GGA GCA CCA GGA CAA GAU GGA 3'
Wild-type polypeptide	N Met Gly Ala Pro Gly Gln Asp Gly C
Silent mutation	ATG GGA GGC CCA GGA CAA GAU GGA Met Gly Ala Pro Gly Gln Asp Gly
Missense mutation	ATG GGA GCA CCA AGA CAA GAU GGA Met Gly Ala Pro Arg Gln Asp Gly
Nonsense mutation	ATG GGA GCA CCA GGA UAA GAU GGA Met Gly Ala Pro Gly Stop
Frameshift mutation	ATG GGA GGC ACC AGG ACA AGA UGG A Met Gly Ala Thr Arg Thr Arg Trp

**(b) Mutations outside the coding sequence**



### Missense mutations

Mutations that change a codon into a mutant codon that specifies a different amino acid are called **missense mutations**. If the substituted amino acid has chemical properties similar to the one it replaces, then this change may have little or no effect on protein function. Such mutations are **conservative substitutions**. For example, a mutation that alters a GAC codon for aspartic acid to a GAG codon for glutamic acid is a conservative substitution because both amino acids have acidic R groups.

By contrast, **nonconservative substitutions**, missense mutations that cause substitution of an amino acid with very different properties, are likely to have more noticeable consequences. A change of the same GAC codon for aspartic acid to GCC, a codon for alanine (an amino acid with an uncharged, nonpolar R group), is an example of a nonconservative substitution.

The effect on phenotype of any missense mutation is difficult to predict because it depends on how a particular amino acid substitution changes a protein's structure and function.

### Nonsense mutations

**Nonsense mutations** change an amino acid–specifying codon to a premature stop codon. Nonsense mutations therefore result in the production of *truncated proteins* lacking all amino acids between the amino acid encoded by the mutant codon and the C terminus of the normal polypeptide.

The mutant polypeptide will be unable to function if it requires the missing amino acids for its activity.

### Frameshift mutations

**Frameshift mutations** result from the insertion or deletion of nucleotides within the coding sequence. As discussed earlier, if the number of extra or missing nucleotides is not divisible by 3, the insertion or deletion will skew the reading frame downstream of the mutation. As a result, frameshift mutations usually result in the formation of truncated proteins (because of the appearance of premature stop codons) with incorrect amino acids at their C termini.

## Mutations Outside the Coding Sequence Can Also Alter Gene Expression

Mutations that produce a variant phenotype are not restricted to alterations in codons. Because gene expression depends on several signals other than the actual coding sequence, changes in any of these critical signals can disrupt the process (**Fig. 8.27b**).

We have seen that promoters and termination signals in the DNA of a gene instruct RNA polymerase to start and stop transcription. Changes in the sequence of a promoter that make it hard or impossible for RNA polymerase to associate with the promoter diminish or prevent transcription. Likewise, mutations in enhancers that disrupt them from being recognized by transcription factors also diminish the transcription of eukaryotic genes. Mutations in a termination signal can diminish the amount of mRNA produced and thus the amount of gene product.

In eukaryotes, most primary transcripts have splice-acceptor sites, splice-donor sites, and branch sites that allow splicing to join exons together with precision in the mature mRNA. Changes in any one of these sites can obstruct splicing. In some cases, the result will be the absence of mature mRNA and thus no polypeptide. In other cases, the splicing errors can yield aberrantly spliced mRNAs that encode altered forms of the protein.

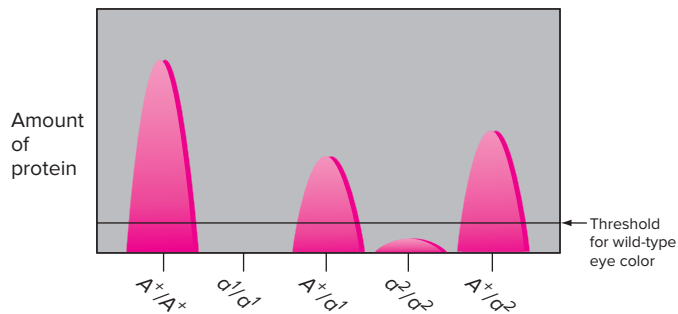
Mature mRNAs have ribosome binding sites and in-frame stop codons indicating where translation should start and stop. Mutations affecting a ribosome binding site would lower the affinity of the mRNA for the small ribosomal subunit; such mutations are likely to diminish the efficiency of translation and thus the amount of polypeptide product. Mutations in a stop codon would produce longer-than-normal proteins that might be unstable or nonfunctional.

## Most Mutations that Affect Gene Expression Reduce Gene Function

Mutations affect phenotype by changing either the amino acid sequence of a protein or the amount of the gene product produced. Any mutation inside or outside a coding region that reduces or abolishes protein activity in one of the many ways previously described is a **loss-of-function mutation**.



**Figure 8.28** Why most loss-of-function mutant alleles are recessive to wild-type alleles. Pink ellipses represent amounts of an enzyme in *Drosophila* called xanthine dehydrogenase. Flies need only 10% of the enzyme produced in wild-type strains ( $A^+/A^+$ ) to have normal eye color. Null allele  $a^1$  and hypomorphic allele  $a^2$  are recessive to wild-type because  $A^+/a^1$  or  $A^+/a^2$  heterozygotes have enough enzyme for normal eye color.



### Recessive loss-of-function alleles

Loss-of-function alleles that block the function of a protein completely are called **null mutations**, or **amorphic mutations**. For protein-encoding genes, the mutations either prevent synthesis of the polypeptide or promote synthesis of a protein incapable of carrying out any function.

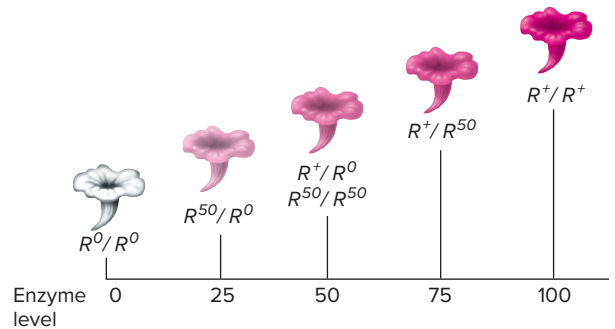
It is easy to understand why amorphic alleles are usually recessive to wild-type alleles. Consider an  $A^+/a^1$  heterozygote, in which the wild-type  $A^+$  allele generates functional protein, while the null  $a^1$  allele does not (Fig. 8.28). If the amount of protein produced by the single  $A^+$  allele (usually, though not always, half the amount produced in an  $A^+/A^+$  cell) is above the threshold amount sufficient to fulfill the normal biochemical requirements of the cell, the phenotype of the  $A^+/a^1$  heterozygote will be wild type. The vast majority of genes function in this way;  $A^+/A^+$  cells actually make more than twice as much of the protein needed for the normal phenotype. Mendel's alleles for green pea color or wrinkled pea shape were likely null alleles and recessive to wild-type alleles for this reason. (Recall Fig. 2.20.)

A **hypomorphic mutation** is a loss-of-function allele that produces either less of the wild-type protein or a mutant protein that functions less effectively than the wild-type protein ( $a^2$  in Fig. 8.28). Hypomorphic alleles are usually recessive to wild-type alleles for the same reason that amorphic alleles are usually recessive.

### Incompletely dominant loss-of-function alleles

Some combinations of alleles generate phenotypes that vary continuously with the amount of functional gene product, giving rise to incomplete dominance. For example, loss-of-function mutations in a single pigment-producing gene can generate a red-to-white spectrum of flower colors, with the white resulting from the absence of an enzyme in a biochemical pathway (Fig. 8.29). Consider three alleles of the gene encoding enzyme R:  $R^+$  specifies the wild-type amount of the enzyme;  $R^{50}$  generates half the normal amount of the same enzyme (or the full amount of an altered form that has half the

**Figure 8.29** When a phenotype varies continuously with levels of protein function, incomplete dominance results.



normal level of activity); and  $R^0$  is a null allele.  $R^+/R^0$  heterozygotes produce pink flowers whose color is halfway between red and white because one-half the  $R^+/R^+$  level of enzyme activity is not enough to generate a full red. Combining  $R^+$  or  $R^0$  with the  $R^{50}$  allele produces pigmentation intermediate between red and pink or between pink and white.

### Unusual dominant loss-of-function alleles

With phenotypes exquisitely sensitive to the amount of functional protein produced, even a relatively small change of two-fold or less can cause a switch between distinct phenotypes. Therefore, a heterozygote for a loss-of-function mutation that generates less than the normal amount of functional gene product may look different from the wild-type organism. Geneticists use the term **haploinsufficiency** to describe relatively rare situations in which one wild-type allele does not provide enough of a gene product to avoid a mutant phenotype.

The number of haploinsufficient genes in humans is estimated to be about 800. One example of a human haploinsufficient gene is *GLI3*, which encodes a transcription factor important for the specification of digits. Heterozygosity for loss-of-function mutations in *GLI3* causes one form of *polydactyly*—the presence of extra fingers and toes (Fig. 8.30).

### Unusual Gain-of-Function Alleles Are Almost Always Dominant

Because there are many ways to interfere with a gene's ability to make sufficient amounts of active protein, the large majority of mutations in most genes are loss-of-function alleles. However, rare mutations that either enhance a protein's function, confer a new activity on a protein, or express a protein at the wrong time or place act as **gain-of-function alleles**. Because a single such allele by itself usually produces a protein that can alter phenotype even in the presence of the normal protein, these unusual gain-of-function alleles are almost always dominant to wild-type alleles. Many dominant mutant alleles are lethal when homozygous.

### Hypermorphic alleles

A **hypermorphic mutation** is one that generates either more normal protein product than the wild-type allele, or a

**Figure 8.30 Haploinsufficiency: Some loss-of-function mutant alleles are dominant to wild-type alleles.** The human *GLI3* gene is haploinsufficient. *GLI3/GLI3<sup>+</sup>* heterozygotes have extra fingers and toes, a condition known as *polydactyly*. One particular mutant *GLI3* allele is a nonsense mutation that changes codon 643 from arginine (R) to stop. (Wild-type *GLI3* protein has 1580 amino acids.)

© Dinodia/agefotostock.com



					643			
	V	T	K	K	Q	<b>R</b>	G	D
Normal allele	GTC	ACC	AAG	AAG	CAG	CGA	GGG	GAC
( <i>GLI3<sup>+</sup></i> )						↓		
Mutant allele	GTC	ACC	AAG	AAG	CAG	TGA	GGG	GAC
( <i>GLI3</i> )	V	T	K	K	<b>Q</b>	<b>STOP</b>		
					642			

more efficient mutant protein. For example, a hypermorphic mutation in the human *FGFR3* gene results in achondroplasia, the most common form of dwarfism (**Fig. 8.31a**). The *FGFR3* gene encodes a signaling protein (fibroblast growth factor receptor 3) that inhibits bone growth. *FGFR3* protein is normally activated only when a small protein called FGF (fibroblast growth factor) binds to it (**Fig. 8.31b**). Most people with achondroplasia carry a mutant allele called *FGFR3<sup>G480R</sup>*, which encodes an *FGFR3* protein with arginine instead of the normal glycine at amino acid 480. This single amino acid change causes the mutant protein to become activated even in the absence of FGF. The mutant protein is thus a *constitutively active* receptor that is activated all the time (**Fig. 8.31c**). The hypermorphic allele (*FGFR3<sup>G480R</sup>*) is dominant to the wild-type allele because the mutant protein remains active and continues to inhibit bone growth even if the normal protein is present.

**Neomorphic alleles**

A rare class of dominant gain-of-function alleles arises from **neomorphic mutations** that generate a novel phenotype. Some neomorphic alleles produce mutant proteins with a new function, while others cause genes to produce the normal protein but at an inappropriate time or place (**ectopic expression**).

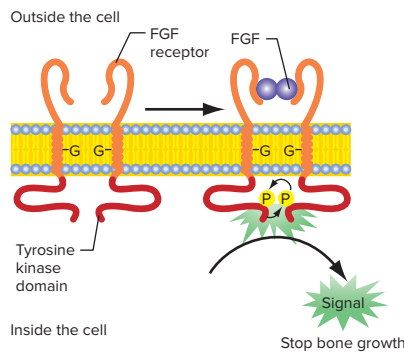
**Figure 8.31 Some hypermorphic alleles encode overactive proteins.** (a) Achondroplasia, a form of dwarfism, is caused by a dominant hypermorphic mutant allele of the *FGFR3* gene, *FGFR3<sup>G480R</sup>*. (b) The *FGFR3* gene encodes a dimeric transmembrane receptor protein that is normally activated only when it is bound to the small protein hormone FGF. The tyrosine kinase domain of one activated *FGFR3* subunit adds phosphate groups (P in yellow circles) to the other subunit and *vice versa*. These phosphorylations initiate a signal that ultimately stops bone growth. (c) Mutant *FGFR3<sup>G480R</sup>* protein is always activated, whether FGF is present or not, leading to improper bone development.

a: © Frazer Harrison/Getty Images

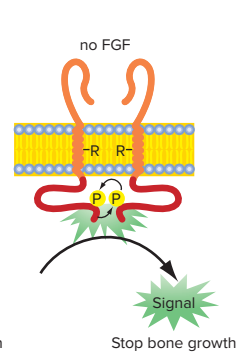
(a) Achondroplasia



(b) Normal *FGFR*



(c) Hypermorphic *FGFR<sup>G480R</sup>*



The dominant Huntington disease allele (*HD*) is an example of a neomorphic allele that makes a mutant protein. Recall from the Fast Forward Box in Chapter 7 (*Trinucleotide Repeat Diseases: Huntington Disease and Fragile X Syndrome*) that *HD<sup>+</sup>* is a polyQ-type trinucleotide repeat gene. Mutant *HD* proteins have an expanded run of glutamine (Q) amino acids, and for unknown reasons such mutant *HD* proteins cause neural degeneration. The *HD* disease allele is dominant to the normal allele because the presence of the normal *HD* protein (with fewer Qs) does not prevent the mutant *HD* protein from damaging nerve cells.

A striking example of a neomorphic allele that expresses a normal protein ectopically is the *Antp<sup>Ns</sup>* mutant allele of the *Drosophila* gene *Antennapedia*. Flies that are *Antp<sup>Ns</sup>/Antp<sup>+</sup>* heterozygotes have legs on their heads in place of antennae (Fig. 8.32a). The *Antp* gene encodes a protein that promotes leg development; accordingly, the wild-type allele *Antp<sup>+</sup>* is transcribed in tissues that will become the fly's legs. A mutation within the transcriptional control region of the gene instead causes the *Antp<sup>Ns</sup>* allele to express normal protein in tissues destined to become the antennae (Fig. 8.32b). *Antp<sup>Ns</sup>* is dominant because the *Antp<sup>+</sup>* allele does not prevent the ectopic expression of Antp protein in the cells normally destined to become antennae.

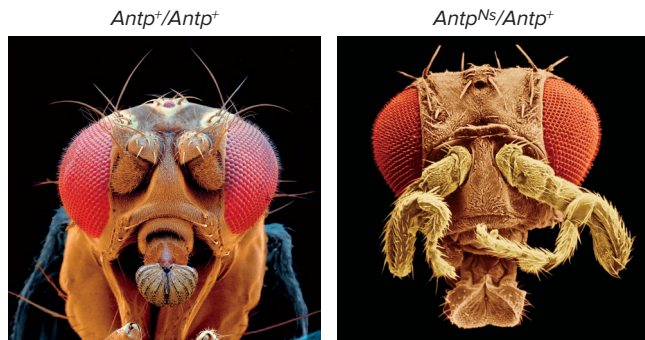
**Antimorphic alleles**

Some dominant mutant alleles of genes encode proteins that not only fail to provide the activity of the wild-type protein, but also prevent the normal protein from functioning. Such alleles are called **dominant-negative**, or **antimorphic**, alleles.

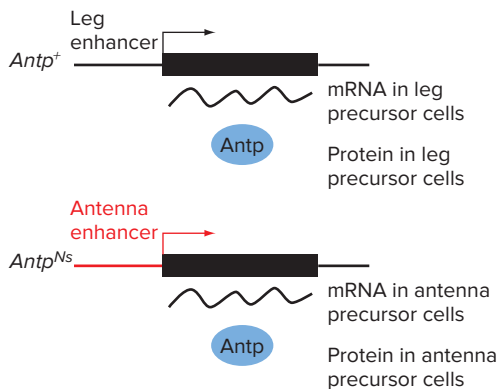
**Figure 8.32** Neomorphic alleles can express a normal protein ectopically. (a) A neomorphic dominant mutation (*Antp<sup>Ns</sup>*) in the fly *Antennapedia* gene produces flies with two legs growing out of the head (right); a normal fly head is shown at left. (b) *Antp<sup>Ns</sup>* has a mutant transcriptional control region that results in ectopic expression of the leg-determining Antp protein in cells normally destined to become antennae.

a.1: © Eye of Science/Science Source; a.2: © Juergen Berger/Science Source

(a) The dominant neomorphic allele *Antp<sup>Ns</sup>* causes antenna-to-leg transformation



(b) *Antp<sup>Ns</sup>* expresses Antp protein ectopically



Consider, for example, a gene encoding a polypeptide that associates with three other identical polypeptides in a four-subunit enzyme. All four subunits are products of the same gene. If a dominant mutant allele *D* directs the synthesis of a *poison subunit* whose presence in the multimer—even as one subunit out of four—abolishes enzyme function, then active tetramers composed solely of functional wild-type *d<sup>+</sup>* subunits are only 1/16 of all tetramers produced (Fig. 8.33a). As a result, total enzyme activity in *D/d<sup>+</sup>* heterozygotes is far less than that seen in wild-type *d<sup>+</sup>/d<sup>+</sup>* homozygotes. The *Kinky* allele of the *Axin* gene in mice, which results in a malformed (kinky) tail, is an example of a dominant negative mutation with such a mechanism of action (Fig. 8.33b).

**Figure 8.33** Why some dominant mutant alleles are antimorphic. (a) With proteins composed of four subunits encoded by a single gene, a dominant negative mutation may inactivate 15 out of every 16 multimers. (b) The *Kinky* allele of the mouse gene *Axin* (*Axin<sup>Kinky</sup>*) is a dominant negative mutation that causes a kink in the tail. The *Axin* protein is a subunit of a protein complex; the protein encoded by *Axin<sup>Kinky</sup>* prevents the complex from working.

b: © Tom Vasicek

(a) The antimorphic allele *D* encodes a mutant poison subunit

<i>d<sup>+</sup>/d<sup>+</sup></i>	<i>D/d<sup>+</sup></i>			
 Functional enzyme				
Normal subunit encoded by <i>d<sup>+</sup></i> allele				
Mutant subunit encoded by <i>D</i> allele				
	Functional enzyme			

(b) Phenotypic effect of the poison subunit encoded by the antimorphic allele *Axin<sup>Kinky</sup>*



**TABLE 8.2** Mutations Classified by Their Effects on Protein Function

	Loss-of-Function		Gain-of-Function		
<b>Mutation Type</b>	Amorphic (null)	Hypomorphic (leaky)	Hypermorphic	Neomorphic	Antimorphic (dominant negative)
<b>Occurrence</b>	Common	Common	Rare	Rare	Rare
<b>Possible Dominance Relations</b>	Usually recessive Can be incompletely dominant if phenotype varies continuously with gene product Can be dominant in cases of haploinsufficiency		Usually dominant or incompletely dominant	Usually dominant	Usually dominant or incompletely dominant

## The Effects of a Mutation Can Be Difficult to Predict

As noted previously, most mutations constitute loss-of-function alleles. The reason is that many changes in amino acid sequence are likely to disrupt a protein's function, and because most alterations in gene regulatory sites, such as promoters, will make those sites less efficient. Nonetheless, rare mutations at almost any location in a gene can result in a gain of function.

Even when you know how a mutation affects gene function, you cannot always predict whether the mutant allele will be dominant or recessive to a wild-type allele (Table 8.2). Although most loss-of-function mutations are recessive and almost all gain-of-function mutations are dominant, exceptions to these generalizations do exist. The reason is that dominance relations between the wild-type and mutant alleles of genes in diploid organisms depend on how drastically a mutation influences protein production or activity, and how thoroughly phenotype depends on the normal wild-type level of the protein.

## Mutations in Genes Encoding the Molecules that Implement Expression May Have Global Effects

Gene expression depends on an astonishing number and variety of proteins and RNAs, each encoded by a separate gene. The genes for all the proteins (RNA polymerases, ribosomal protein subunits, aminoacyl-tRNA synthetases, etc.) are transcribed and translated the same as any other gene. The genes for all the rRNAs, tRNAs, and snRNAs are **noncoding genes** that are transcribed but not translated. Mutations in almost any of these genes, whether protein-coding or noncoding, can have a dramatic effect on phenotype.

### Lethal mutations affecting the machinery of gene expression

Loss-of-function mutations in the genes encoding molecules that implement gene expression, such as ribosomal

proteins or rRNAs, are often lethal in homozygotes because such mutations adversely affect the synthesis of all proteins in a cell. Even a 50% reduction in the amount of some of the proteins or RNAs required for gene expression can have severe repercussions. In *Drosophila*, for example, null mutations in many of the genes encoding the various ribosomal proteins are lethal when homozygous. Due to haploinsufficiency, the same mutations in heterozygotes cause a dominant *Minute* phenotype, in which the slow growth of cells delays the fly's development.

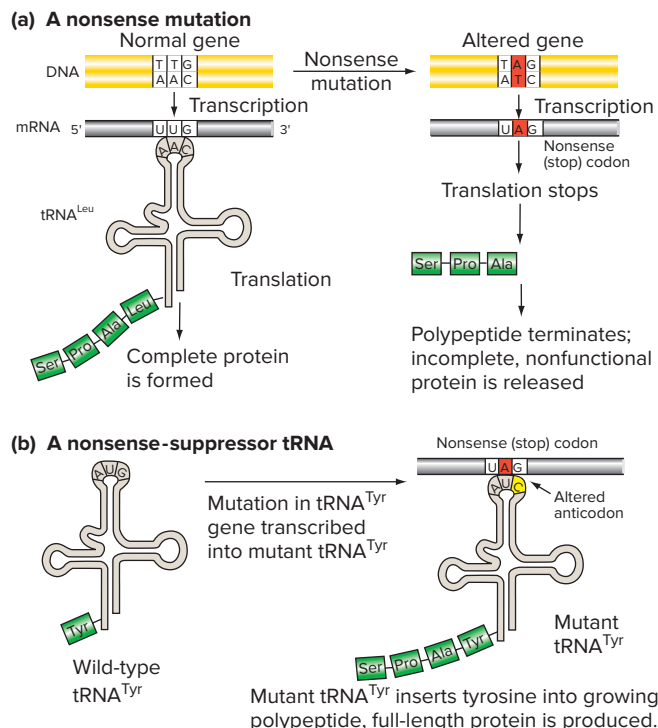
### Suppressor mutations in tRNA genes

If more than one gene encoded the same molecule with a role in gene expression, a mutation in one of these genes would not necessarily be lethal and might even be advantageous. Bacterial geneticists have found, for example, that mutations in certain tRNA genes can suppress the effects of nonsense mutations in other genes. The tRNA-gene mutations that have this effect give rise to **nonsense suppressor tRNAs**.

Consider, for instance, an otherwise wild-type *E. coli* population with an in-frame UAG nonsense mutation in the tryptophan synthase gene. All cells in this population make a truncated, nonfunctional form of the tryptophan synthase enzyme and are thus tryptophan auxotrophs ( $\text{Trp}^-$ ) unable to synthesize tryptophan (Fig. 8.34a). Subsequent exposure of these auxotrophs to mutagens, however, generates some  $\text{Trp}^+$  cells that carry two mutations: One is the original tryptophan synthase nonsense mutation, and the second is a mutation in the gene that encodes a tRNA for the amino acid tyrosine. Evidently, the mutation in the tRNA gene suppresses the effect of the nonsense mutation, restoring the function of the tryptophan synthase gene.

As Fig. 8.34b illustrates, the basis of this nonsense suppression is that the  $\text{tRNA}^{\text{Tyr}}$  mutation changes an anticodon that recognizes the codon for tyrosine to an anticodon complementary to the UAG stop codon. The mutant tRNA can therefore insert tyrosine into the polypeptide at the position of the in-frame UAG nonsense mutation, allowing the cell to make at least some full-length enzyme.

**Figure 8.34 Nonsense suppression.** (a) A nonsense mutation that generates a stop codon causes production of a truncated, nonfunctional polypeptide. (b) A second, nonsense-suppressing mutation in a tRNA gene causes the addition of an amino acid in response to the stop codon, allowing production of a full-length polypeptide.



Similarly, mutations in the anticodons of other tRNA genes can suppress UGA or UAA nonsense mutations.

Cells with a nonsense-suppressing mutation in a tRNA gene can survive only if two conditions coexist with the mutation. First, the cell must have other tRNAs that recognize the same codon as the suppressing tRNA recognized before mutation altered its anticodon. Without such tRNAs, the cell has no way to insert the proper amino acid in response to that codon (in our example, the codon for tyrosine). Second, the suppressing tRNA must respond only

inefficiently to the stop codons normally found at the ends of mRNA coding regions. If this were not the case, the suppressing tRNA would wreak havoc in the cell, producing a whole array of aberrant polypeptides that are longer than normal. One way cells guard against this possibility is to place two stop codons in a row at the ends of many genes. Because a suppressing tRNA's chance of inserting an amino acid at both of these codons is very low, only a small number of extended proteins arise.

**essential concepts**

- *Point mutations* in the coding sequences of a gene may modify the amino acid sequence of the polypeptide product.
- Mutations outside the coding sequences can modify gene expression by altering the amount, time, or place of protein production.
- *Loss-of-function mutations* reduce or abolish gene expression. Most loss-of-function alleles are recessive to wild-type alleles. When phenotype varies continuously with the amount of gene product, loss-of-function alleles are incompletely dominant. In *haploinsufficiency*, half the normal amount of gene product is not enough for a normal phenotype, so a loss-of-function mutant allele has dominant effects.
- Rare *gain-of-function mutations* cause increased protein production, synthesis of an altered protein, or production of the normal protein in the wrong context. Most gain-of-function alleles are dominant to wild-type alleles.
- Whether a mutation is recessive or dominant to wild type depends on how drastically the protein product is altered and how sensitive phenotype is to the abnormal gene function.
- Mutations in genes that encode molecules of the gene expression machinery are often lethal. Exceptions include mutations in tRNA genes that produce *nonsense suppressor tRNAs*.

**WHAT'S NEXT**

Our knowledge of gene expression enables us to redefine the concept of a gene. A gene is not simply the DNA that is transcribed into the mRNA codons specifying the amino acids of a particular polypeptide. Rather, *a gene is all the DNA sequences needed for expression of the gene as a polypeptide product*. A gene therefore includes the promoter sequences that govern where transcription begins and, at the opposite end, signals for the termination of transcription. A gene also must include sequences dictating where translation of the mRNA starts and

stops. In addition to all of these features, eukaryotic genes contain introns that are spliced out of the primary transcript to make the mature mRNA. Because of introns, most eukaryotic genes are much larger than prokaryotic genes.

Even with introns, a single gene carries only a very small percentage of the nucleotide pairs in the chromosomes that make up a genome. In humans, the average gene is 16,000 nucleotide pairs in length. But the haploid human genome has roughly 3 billion (3,000,000,000) nucleotide

pairs distributed among 23 chromosomes containing an average of 130 million nucleotide pairs apiece.

In Chapters 9, 10 and 11, we describe how researchers analyze the mass of genetic information in the chromosomes of a genome as they try to discover what parts of the DNA are genes and how those genes influence phenotype. They begin their analysis by breaking the DNA into pieces

of manageable size, making many copies of those pieces to obtain enough material for study, and characterizing the pieces down to the level of nucleotide sequence. The scientists then try to reconstruct the DNA sequence of an entire genome by determining the spatial relationship between the many pieces. Finally, they use this knowledge to examine the genomic variations that make individuals unique.

## SOLVED PROBLEMS

- I. A geneticist examined the amino acid sequence of a particular protein in a variety of *E. coli* mutants. The amino acid in position 40 in the normal enzyme is glycine. The following table shows the substitutions the geneticist found at amino acid position 40 in six mutant forms of the enzyme.

mutant 1	cysteine
mutant 2	valine
mutant 3	serine
mutant 4	aspartic acid
mutant 5	arginine
mutant 6	alanine

Determine the nature of the base substitution that must have occurred in the DNA in each case. Which of these mutants would be capable of recombination with mutant 1 to form a wild-type gene?

### Answer

To determine the base substitutions, use the genetic code table (see Fig. 8.2). The original amino acid was glycine, which can be encoded by GGU, GGC, GGA, or GGC. Mutant 1 results in a cysteine at position 40; Cys codons are either UGU or UGC. A change in the base pair in the DNA encoding the first position in the codon (a G–C to T–A transversion) must have occurred, and the original glycine codon must therefore have been either GGU or GGC. Valine (in mutant 2) is encoded by GUN (with N representing any one of the four bases), but assuming that the mutation is a single base change, the Val codon must be either GUU or GUC. The change must have been a G–C to T–A transversion in the DNA for the second position of the codon. To get from glycine to serine (mutant 3) with only one base change, the GGU or GGC would be changed to AGU or AGC, respectively. A transition occurred (G–C to A–T) at the first position. Aspartic acid (mutant 4) is encoded by GAU or GAC, so the DNA of mutant 4 is the result of a G–C to A–T transition at position 2. Arginine (mutant 5) is encoded by CGN, so the DNA of mutant 5 must have undergone a G–C to C–G transversion at position 1. Finally, alanine (mutant 6) is encoded by GCN, so the DNA of mutant 6 must have undergone a G–C to C–G transversion

at position 2. Mutants 2, 4, and 6 affect a base pair different from that affected by mutant 1, so they could recombine with mutant 1.

In summary, the sequence of nucleotides on the RNA-like strand of the wild-type and mutant genes at this position must be:

wild type	5' G G T/C 3'
mutant 1	5' T G T/C 3'
mutant 2	5' G T T/C 3'
mutant 3	5' A G T/C 3'
mutant 4	5' G A T/C 3'
mutant 5	5' C G T/C 3'
mutant 6	5' G C T/C 3'

- II. The double-stranded circular DNA molecule that forms the genome of the SV40 tumor virus can be denatured into single-stranded DNA molecules. Because the base composition of the two strands differs, the strands can be separated on the basis of their density into two strands designated W(atson) and C(rick).

When each of the purified preparations of the single strands was mixed with mRNA from cells infected with the virus, hybrids were formed between the RNA and DNA. Closer analysis of these hybridizations showed that RNAs that hybridized with the W preparation were different from RNAs that hybridized with the C preparation. What does this tell you about the transcription templates for the different classes of RNAs?

### Answer

An understanding of transcription and the polarity of DNA strands in the double helix are needed to answer this question. Some genes use one strand of the DNA as a template; others use the opposite strand as a template. Because of the different polarities of the DNA strands, one set of genes would be transcribed in a clockwise direction on the circular DNA (using, say, the W strand as the template), and the other set would be transcribed in a counterclockwise direction (with the C strand as template).

- III. Geneticists interested in human hemoglobins have found a very large number of mutant forms. Some of

these mutant proteins are of normal size (but have amino acid substitutions) while others are short, due to deletions or nonsense mutations. The first extra-long example was named Hb Constant Spring, in which the  $\beta$ -globin has all of its normal amino acids plus several extra amino acids attached after the normal C-terminal end of the protein.

- What is the most plausible explanation for its origin?
- Is it possible that Hb Constant Spring arose from failure to splice out an intron?
- Estimate how many extra amino acids might be added to the C terminal end of the mutant protein.

### Answer

An understanding of the principles of translation and RNA splicing are needed to answer this question.

- Because an extension on the C-terminal end of the protein is present, the mutation probably affected the termination (nonsense) codon rather than affecting splicing of the RNA. This alteration could have been a base change or a frameshift or

a deletion that altered or removed the termination codon. The information in the mRNA beyond the normal stop codon would be translated until another stop codon in the mRNA was reached.

- A splicing defect could explain Hb Constant Spring only if an intron existed at a location just before the stop codon and the mutation prevented removal of the intron from the mature mRNA. In this case, the nucleotides in the intron would be read in frame as triplets until a stop codon was reached. The necessity for the presence of an intron in a specific location makes this scenario much less likely than that in part (a).
- Regardless of whether the explanation for the Hb Constant Spring protein is a change to the termination codon or a change in splicing, you can estimate the number of amino acids added to the end of the protein by assuming they are encoded by a random DNA sequence. In the genetic code, 3 out of the 64 triplets (~1 in 21) are stop codons. So you would roughly estimate that on average you would add about 21 amino acids to the end of the mutant protein until the reading frame encountered a stop codon.



## PROBLEMS

### Vocabulary

- For each of the terms in the left column, choose the best matching phrase in the right column.

- |                                   |  |                          |   |
|-----------------------------------|--|--------------------------|---|
| a. codon                          | 1. removing base sequences corresponding to introns from the primary transcript  | i. RNA-like strand       | 9. AUG in a particular context  |
| b. colinearity                    | 2. UAA, UGA, or UAG  | j. intron                | 10. the linear sequence of amino acids in the polypeptide corresponds to the linear sequence of nucleotide pairs in the gene                  |
| c. reading frame                  | 3. the strand of DNA that has the same base sequence as the primary transcript   | k. RNA splicing          | 11. produces different mature mRNAs from the same primary transcript  |
| d. frameshift mutation            | 4. a transfer RNA molecule to which the appropriate amino acid has been attached   | l. transcription         | 12. addition or deletion of a number of base pairs other than three into the coding sequence  |
| e. degeneracy of the genetic code | 5. a group of three mRNA bases signifying one amino acid   | m. translation           | 13. a sequence of base pairs within a gene that is not represented by any bases in the mature mRNA  |
| f. nonsense codon                 | 6. most amino acids are not specified by a single codon  | n. alternative splicing  | 14. the strand of DNA having the base sequence complementary to that of the primary transcript  |
| g. initiation codon               | 7. using the information in the nucleotide sequence of a strand of DNA to specify the nucleotide sequence of a strand of RNA | o. charged tRNA          | 15. using the information encoded in the nucleotide sequence of an mRNA molecule to specify the amino acid sequence of a polypeptide molecule |
| h. template strand                | 8. the grouping of mRNA bases in threes to be read as codons   | p. reverse transcription | 16. copying RNA into DNA  |

## Section 8.1

2. Match the hypothesis from the left column to the observation from the right column that gave rise to it.
- |   |  |
|---|--|
| a. existence of an intermediate messenger between DNA and protein               | 1. two mutations affecting the same amino acid can recombine to give wild type   |
| b. the genetic code is nonoverlapping   | 2. one or two base deletions (or insertions) in a gene disrupt its function; three base deletions (or insertions) are often compatible with function |
| c. the codon is more than one nucleotide  | 3. artificial messages containing certain codons produce shorter proteins than messages not containing those codons                                  |
| d. the genetic code is based on triplets of bases                               | 4. protein synthesis occurs in the cytoplasm, while DNA resides in the nucleus   |
| e. stop codons exist and terminate translation                                  | 5. artificial messages with different base sequences gave rise to different proteins in an <i>in vitro</i> translation system                        |
| f. the amino acid sequence of a protein depends on the base sequence of an mRNA | 6. single base substitutions affect only one amino acid in the protein chain   |

3. How would the artificial mRNA

5'... GUGUGUGU ... 3'

be read according to each of the following models for the genetic code?

- two-base, not overlapping
  - two-base, overlapping
  - three-base, not overlapping
  - three-base, overlapping
  - four-base, not overlapping
4. An example of a portion of the T4 *rIIB* gene in which Crick and Brenner had recombined one + and one – mutation is shown here. (The RNA-like strand of the DNA is shown.)
- |           |                                   |
|-----------|-----------------------------------|
| wild type | 5' AAA AGT CCA TCA CTT AAT GCC 3' |
| mutant    | 5' AAA GTC CAT CAC TTA ATG GCC 3' |
- Where are the + and – mutations in the mutant DNA?
  - The double mutant produces wild-type plaques. What alterations in amino acids occurred in this double mutant?
  - How can you explain the fact that amino acids are different in the double mutant than in the wild-type sequence, yet the phage has a wild-type phenotype?

5. Consider Crick and Brenner's experiments in Fig. 8.4, which showed that the genetic code is based on nucleotide triplets.

- Crick and Brenner obtained FC7, an intragenic suppressor of FC0, that was a mutation in a second site

in the *rIIB* gene near the FC0 mutation. Describe a different kind of mutation in the *rIIB* gene these researchers might have recovered by treating the FC0 mutant with proflavin and looking for restored *rIIB*<sup>+</sup> function.

- How could Crick and Brenner tell the difference between the occurrence described in part (a) and an intragenic suppressor like FC7?
  - When FC7 was separated from FC0 by recombination, the result was two *rIIB*<sup>–</sup> mutant phages: One was FC7 and the other was FC0. How could they discriminate between the *rIIB*<sup>–</sup> recombinants that were FC7 and those that were FC0?
  - Explain how Crick and Brenner could obtain different deletion (–) or addition (+) mutations so as to make the various combinations such as ++, ––, ++++, and ––– shown in Fig. 8.4c.
6. The *Hbβ*<sup>S</sup> (sickle-cell) allele of the human β-globin gene changes the sixth amino acid in the β-globin chain from glutamic acid to valine. In *Hbβ*<sup>C</sup>, the sixth amino acid in β-globin is changed from glutamic acid to lysine. What would be the order of these two mutations within the map of the β-globin gene?
7. The following diagram describes the mRNA sequence of part of the *A* gene and the beginning of the *B* gene of phage φX174. In this phage, some genes are read in overlapping reading frames. For example, the code for the *A* gene is used for part of the *B* gene, but the reading frame is displaced by one base. Shown here is the single mRNA with the codons for proteins A and B indicated.

aa#	5	6	7	8	9	10	11	12	13	14	15	16
A	A	L	y	S	G	I	u	T	r	A	s	n
mRNA	G	C	U	A	A	G	A	A	U	G	G	A
B				M	e	t	G	l	u	G	l	n
aa#		1	2	3	4	5	6	7	8	9		

Given the following amino acid (aa) changes, indicate the base change that occurred in the mRNA and the consequences for the other protein sequence.

- Asn at position 10 in protein A is changed to Tyr.
  - Leu at position 12 in protein A is changed to Pro.
  - Gln at position 8 in protein B is changed to Leu.
  - The occurrence of overlapping reading frames is very rare in nature. When it does occur, the extent of the overlap is not very long. Why do you think this is the case?
8. The amino acid sequence of part of a protein has been determined:

N... Gly Ala Pro Arg Lys ... C

A mutation has been induced in the gene encoding this protein using the mutagen proflavin. The resulting



mutant protein can be purified and its amino acid sequence determined. The amino acid sequence of the mutant protein is exactly the same as the amino acid sequence of the wild-type protein from the N terminus of the protein to the glycine in the preceding sequence. Starting with this glycine, the sequence of amino acids is changed to the following:

N . . . Gly His Gln Gly Lys . . . C

Using the amino acid sequences, one can determine the sequence of 14 nucleotides from the wild-type gene encoding this protein. What is this sequence?

9. The results shown in Fig. 8.5 may have struck you as incongruous because many synthetic RNAs that lacked AUG start codons (such as poly-U) were nonetheless translated into polypeptides *in vitro*. The reason this experiment was possible is that Marshall Nirenberg found that a high concentration of  $Mg^{2+}$  ions in the test tube, much higher than that found in cells, allows ribosomes to initiate translation at any position on an RNA molecule. Predict the outcomes of *in vitro* translation with each of the following synthetic mRNAs at both high and low  $Mg^{2+}$  concentrations:
  - a. poly-UG (UGUGUG . . .)
  - b. poly-CAUG (CAUGCAUGCAUG . . .)
  - c. poly-GUAU (GUAUGUAUGUAU . . .)
10. Identify all the amino acid-specifying codons in the genetic code where a point mutation (a single base change) could generate a nonsense codon.
11. Before the technology existed to synthesize RNA molecules of defined sequence like those in Fig. 8.5, similar experiments were performed with synthetic mRNAs of undefined sequence. For example, RNAs consisting only of Us and Gs could be synthesized *in vitro*, but they would have random sequences. Suppose a pool of random-sequence RNAs was synthesized in a reaction mixture containing three times as much UTP as GTP, and that the resulting RNAs were translated *in vitro*.
  - a. How many different codons exist in the RNAs?
  - b. How many different amino acids would you find in the polypeptides synthesized?
  - c. Why are your answers to (a) and (b) not the same?
  - d. How often would you expect to find each of the codons in (a)?
  - e. In what proportions would you expect to find each of the amino acids in the polypeptides?
  - f. If you did this experiment—that is, synthesized random-sequence RNAs containing a 3:1 ratio of U:G, and quantified the amount of each amino acid in the polypeptides produced—prior to knowledge of the genetic code table, what would the results have told you?
12. A particular protein has the amino acid sequence
 

N . . . Ala-Pro-His-Trp-Arg-Lys-Gly-Val-Thr . . . C

 within its primary structure. A geneticist studying mutations affecting this protein discovered that several of the mutants produced shortened protein molecules that terminated within this region. In one of them, the His became the terminal amino acid.
  - a. What DNA single-base change(s) would cause the protein to terminate at the His residue?
  - b. What other potential sites do you see in the DNA sequence encoding this protein where mutation of a single base pair would cause premature termination of translation?
13. How many possible *open reading frames* (frames without stop codons) exist that extend through the following sequence?
 

5' . . . CTTACAGTTTATTGATACGGAGAAGG . . . 3'  
3' . . . GAATGTCAAATAACTATGCCTCTCC . . . 5'
14. a. In Fig. 8.3, the physical map (the number of base pairs) is not exactly equivalent to the genetic map (in map units). Explain this apparent discrepancy.  
 b. In Fig. 8.3, which region shows the highest frequency of recombination, and which the lowest?
15. Charles Yanofsky isolated many different *trpA*<sup>-</sup> mutants (Fig. 8.3).
  - a. Explain how he could identify Trp<sup>-</sup> auxotrophs of *E. coli* using replica plating (recall Fig. 7.6).
  - b. Assuming that the role of TrpA enzyme in the tryptophan biosynthesis pathway was known, explain how Yanofsky could have identified *trpA*<sup>-</sup> mutants among his Trp<sup>-</sup> auxotrophs. (*Hint*: Recall Beadle and Tatum's one gene, one enzyme experiments in Chapter 7.)
16. The sequence of a segment of mRNA, beginning with the initiation codon, is given here, along with the corresponding sequences from several mutant strains.
 

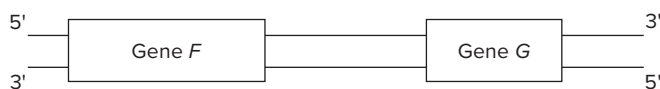
Normal	AUGACACAUCGAGGGGUGGUAACCCUAAG . . .
Mutant 1	AUGACACAUCGAGGGGUGGUAACCCUAAG . . .
Mutant 2	AUGACACAUCGAGGGGUGGUAACCCUAAG . . .
Mutant 3	AUGACGCAUCGAGGGGUGGUAACCCUAAG . . .
Mutant 4	AUGACACAUCGAGGGGUGGUAACCCUAAG . . .
Mutant 5	AUGACCAUUGAGGGGUGGUAACCCUAAG . . .
Mutant 6	AUGACAUUUACCCCCUCGAUGCCCUAAG . . .

  - a. Indicate the type of mutation present in each, and translate the mutated portion of the sequence into an amino acid sequence in each case.
  - b. Which of the mutations could be reverted by treatment with EMS (ethylmethane sulfonate; see Fig. 7.14)? With proflavin?

17. You identify a proflavin-generated allele of a gene that produces a 110-amino acid polypeptide rather than the usual 157-amino acid protein. After subjecting this mutant allele to extensive proflavin mutagenesis, you are able to find a number of intragenic suppressors located in the part of the gene between the sequences encoding the N terminus of the protein and the original mutation, but no suppressors located in the region between the original mutation and the sequences encoding the usual C terminus of the protein. Why do you think this is the case?
18. Using recombinant DNA techniques (which will be described in Chapter 9), it is possible to take the DNA of a gene from any source and place it on a chromosome in the nucleus of a yeast cell. When you take the DNA for a human gene and put it into a yeast cell chromosome, the altered yeast cell can make the human protein. But when you remove the DNA for a gene normally present on yeast mitochondrial chromosomes and put it on a yeast chromosome in the nucleus, the yeast cell cannot synthesize the correct protein, even though the gene comes from the same organism. Explain. What would you need to do to ensure that such a yeast cell could make the correct protein?

### Section 8.2

19. Describe the steps in transcription that require complementary base pairing.
20. Chapters 6 and 7 explained that mistakes made by DNA polymerase are corrected either by proofreading mechanisms during DNA replication or by DNA repair systems that operate after replication is complete. The overall rate of errors in DNA replication is about  $1 \times 10^{-10}$ , that is, one error in 10 million base pairs. RNA polymerase also has some proofreading capability, but the overall error rate for transcription is significantly higher ( $1 \times 10^{-4}$ , or one error in each 10,000 nucleotides). Why can organisms tolerate higher error rates for transcription than for DNA replication?
21. The coding sequence for gene *F* is read from left to right on the accompanying figure. The coding sequence for gene *G* is read from right to left. Which strand of DNA (*top* or *bottom*) serves as the template for transcription of each gene?



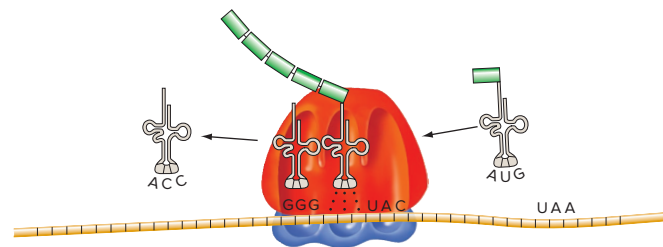
22. If you mixed the mRNA of a human gene with the genomic DNA for the same gene and allowed the RNA and DNA to form a hybrid molecule by base complementarity, what would you be likely to see in

the electron microscope? Your figure should include hybridization involving both DNA strands (template and RNA-like) as well as the mRNA.

23. In studying normal and mutant forms of a particular human enzyme, a geneticist came across a particularly interesting mutant form of the enzyme. The normal enzyme is 227 amino acids long, but the mutant form was 312 amino acids long. The extra 85 amino acids occurred as a block in the middle of the normal sequence. The inserted amino acids do not correspond in any way to the normal protein sequence. What are possible explanations for this phenomenon? How would you distinguish among them?
24. The *Drosophila* gene *Dscam1* encodes proteins on the surface of nerve cells (neurons) that govern neuronal connections. Each neuron has on its surface a single *Dscam1* protein of the tens of thousands that exist. The particular *Dscam1* protein a neuron expresses is thought to tag the cell uniquely to determine the paths of the axons and dendrites it will grow. Eukaryotic genes are monocistronic. How then can a single *Dscam1* gene encode tens of thousands of different proteins?

### Section 8.3

25. Describe the steps in translation that require complementary base pairing.
26. Locate as accurately as possible the listed items that are shown on the following figure. Some items are not shown. (a) 5' end of DNA template strand; (b) 3' end of mRNA; (c) ribosome; (d) promoter; (e) codon; (f) an amino acid; (g) DNA polymerase; (h) 5' UTR; (i) centromere; (j) intron; (k) anticodon; (l) N terminus; (m) 5' end of charged tRNA; (n) RNA polymerase; (o) 3' end of uncharged tRNA; (p) a nucleotide; (q) mRNA cap; (r) peptide bond; (s) P site; (t) aminoacyl-tRNA synthetase; (u) hydrogen bond; (v) exon; (w) 5' AUG 3'; (x) potential wobble interaction.



27. Concerning the figure for Problem 26:
- Which process is being represented?
  - What is the next building block to be added to the growing chain in the figure? To what end of the

- growing chain will this building block be added?  
How many building blocks will there be in the chain when it is completed?
- c. What other building blocks have a known identity?
- d. What details could you add to this figure that would be different in a eukaryotic cell versus a prokaryotic cell?
28. a. Can a tRNA exist that has the anticodon sequence 5' IAA? If so, which amino acid would it carry?
- b. Answer the same question for the anticodon sequence 5' xm<sup>5</sup>s<sup>2</sup>UAA.
29. For parts (a) and (b) of Problem 28, consider the DNA sequence of the gene encoding the tRNA. What is the sequence of the RNA-like strand of the tRNA gene corresponding to the tRNA's anticodon? What is the sequence of the template strand of the gene for these same three nucleotides? Be sure to indicate polarities.
30. Remembering that the wobble base of the tRNA is the 5' base of the anticodon:
- a. In human tRNAs, what are the sequences of all possible anticodons that were originally transcribed with A in the wobble position? (Assume this A is always modified to I.)
- b. In human tRNAs, what are the sequences of all possible anticodons that were originally transcribed with U in the wobble position? (*Note:* Any single type of tRNA with a U at the wobble position can be modified only in a single way.)
- c. How might the wobble Us in each of the anticodons in (b) be modified and still be consistent with the genetic code?
- d. What is the theoretical *minimal* number of different tRNA genes that must exist in the human genome? (Assume that xo<sup>5</sup>U pairs with A, G, or U only.)
31. The human genome contains about 500 genes for tRNAs.
- a. Do you think that each one of these tRNA genes has a different function?
- b. Can you explain why the human genome might have evolved so as to house so many tRNA genes?
32. The yeast gene encoding a protein found in the mitotic spindle was cloned by a laboratory studying mitosis. The gene encodes a protein of 477 amino acids.
- a. What is the minimum length in nucleotides of the protein-coding part of this yeast gene?
- b. A partial sequence of one DNA strand in an exon containing the middle of the coding region of the yeast gene is given here. What is the sequence of nucleotides of the mRNA in this region of the gene? Show the 5' and 3' directionality of your strand.
- 5' GTAAGTTAACTTTTCGACTAGTCCAGGGT 3'
- c. What is the sequence of amino acids in this part of the yeast mitotic spindle protein?
33. The sequence of a complete eukaryotic gene encoding the small protein Met Tyr Arg Gly Ala is shown here. All of the written sequences on the template strand are transcribed into RNA.
- 5' CCCCTATGCCCCCTGGGGGAGGATCAAAACACTTACCTGTACATGGC 3'  
3' GGGGATACGGGGGGACCCCTCCTAGTTTTGTGAATGGACATGTACCG 5'
- a. Which strand is the template strand? In which direction (right-to-left or left-to-right) does RNA polymerase move along the template as it transcribes this gene?
- b. What is the sequence of the nucleotides in the processed mRNA molecule for this gene? Indicate the 5' and 3' polarity of this mRNA.
- c. A single base mutation in the gene results in synthesis of the peptide Met Tyr Thr. What is the sequence of nucleotides making up the mRNA produced by this mutant gene?
34. Arrange the following list of eukaryotic gene elements in the order in which they would appear in the genome and in the direction traveled by RNA polymerase along the gene. Assume the gene's single intron interrupts the open reading frame. Note that some of these names are abbreviated and thus do not distinguish between elements in DNA versus RNA. For example, *splice-donor site* is an abbreviation for *DNA sequences transcribed into the splice-donor site* because splicing takes place on the gene's RNA transcript, not on the gene itself. Geneticists often use this kind of shorthand for simplicity, even though it is imprecise. (a) splice-donor site; (b) 3' UTR; (c) promoter; (d) stop codon; (e) nucleotide to which methylated cap is added; (f) initiation codon; (g) transcription terminator; (h) splice-acceptor site; (i) 5' UTR; (j) poly-A addition site; (k) splice branch site.
35. Concerning the list of eukaryotic gene elements in Problem 34:
- a. Which of the element names in the list are abbreviated? (That is, which of these elements actually occur in the gene's primary transcript or mRNA rather than in the gene itself?)
- b. Which of the elements in the list are found partly or completely in the first exon of this gene (or in the RNA transcribed from this exon)? In the intron? In the second exon?

36. The human gene for  $\beta 2$  lens crystallin has the components listed below. The numbers represent nucleotide pairs that make up the particular component. Assume for simplicity that no alternative splicing is involved.

5' UTR	174
1st exon	119
1st intron	532
2nd exon	337
2nd intron	1431
3rd exon	208
3rd intron	380
4th exon	444
4th intron	99
5th exon	546
3' UTR	715

Answer the following questions about the  $\beta 2$  lens crystallin gene, primary transcript, and gene product. Questions asking *where* should be answered with one of the 11 components from the list or with *None*. Assume poly-A tails contain 150 As.

- How large is the  $\beta 2$  lens crystallin gene in bp (base pairs)?
- How large is the primary transcript for  $\beta 2$  lens crystallin in bases?
- How large is the mature mRNA for  $\beta 2$  lens crystallin in bases?
- Where would you find the base pairs encoding the initiation codon?
- Where would you find the base pairs encoding the stop codon?
- Where would you find the base pairs encoding the 5' cap?
- Where would you find the base pairs that constitute the promoter?
- Which intron interrupts the 3' UTR?
- Where would you find the sequences encoding the C terminus?
- Where would you find the sequence encoding the poly-A tail?
- How large is the coding region of the gene in bp (base pairs)?
- How many amino acids are in the  $\beta 2$  lens crystallin protein?
- Which intron interrupts a codon?
- Which intron is located between codons?
- Where would you be likely to find the site specifying poly-A addition?

You find in lens-forming cells from several different people small amounts of a polypeptide that has the same N terminus as the normal  $\beta$  lens crystallin, but it has a different C terminus. The polypeptide is 114 amino acids long, of which 94 are shared with the

normal protein and 20 are unrelated *junk*. No mutation is involved in the production of this 114-amino acid protein.

- Outline a hypothesis for the process that would produce this protein. Your hypothesis should explain why 94 amino acids are the same as in the normal  $\beta 2$  lens crystallin.
- Explain why you would expect that a polypeptide such as the 114-mer described above would on average have 20 amino acids of junk.

### Section 8.4

- In prokaryotes, a search for genes in a DNA sequence involves scanning the DNA sequence for long open reading frames (that is, reading frames uninterrupted by stop codons). What problem can you see with this approach in eukaryotes?
- The genetic code table shown in Fig. 8.2 applies both to humans and to *E. coli*. Suppose that you have purified a piece of DNA from the human genome containing the entire gene encoding the hormone insulin. You now transform this piece of DNA into *E. coli*. Why can't *E. coli* cells containing the human insulin gene actually make insulin?
  - Pharmaceutical companies have actually been able to obtain *E. coli* cells that make human insulin; such insulin can be purified from the bacterial cells and used to treat diabetic patients. How were the pharmaceutical companies able to create such bacterial factories for making insulin?
- Very few if any eukaryotic genes contain tracts with more than 25 As or Ts in a row, yet almost all eukaryotic mRNAs have a tract with more than 100 As in a row. How is this possible?
  - Scientists know the nucleotide sequences that direct the termination of bacterial gene transcription, but they generally have little idea about the nature of the nucleotide sequences that direct transcription termination in eukaryotic cells. Explain the basis of this statement.
- Explain how differences in the initiation of translation dictate that eukaryotic mRNAs are monocistronic while prokaryotic mRNAs may be polycistronic.

### Section 8.5

- Do you think each of the following types of mutations would have very severe effects, mild effects, or no effect at all?
  - Nonsense mutations occurring in the sequences encoding amino acids near the N terminus of the protein
  - Nonsense mutations occurring in the sequences encoding amino acids near the C terminus of the protein
  - Frameshift mutations occurring in the sequences encoding amino acids near the N terminus of the protein

- d. Frameshift mutations occurring in the sequences encoding amino acids near the C terminus of the protein
- e. Silent mutations
- f. Conservative missense mutations
- g. Nonconservative missense mutations affecting the active site of the protein
- h. Nonconservative missense mutations not in the active site of the protein
42. Null mutations are valuable genetic resources because they allow a researcher to determine what happens to an organism in the complete absence of a particular protein. However, it is often not a trivial matter to determine whether a mutation represents the null state of the gene.
- Geneticists sometimes use the following test for the *nullness* of an allele in a diploid organism: If the abnormal phenotype seen in a homozygote for the allele is identical to that seen in a heterozygote (where one chromosome carries the allele in question and the homologous chromosome is known to be completely deleted for the gene) then the allele is null. What is the underlying rationale for this test? What limitations might there be in interpreting such a result?
  - Can you think of other methods to determine whether an allele represents the null state of a particular gene?
43. The following is a list of mutations that have been discovered in a gene that has more than 60 exons and encodes a very large protein of 2532 amino acids. Indicate whether or not each mutation could cause a detectable change in the size or the amount of mRNA and/or a detectable change in the size or the amount of the protein product. (Detectable changes in size or amount must be greater than 1% of normal values.) What kind of change would you predict?
- Lys576Val (changes amino acid 576 from lysine into valine)
  - Lys576Arg
  - AAG576AAA (changes codon 576 from AAG to AAA)
  - AAG576UAG
  - Met1Arg (at least two possible scenarios exist for this mutation)
  - promoter mutation
  - one base pair insertion into codon 1841
  - deletion of codon 779
  - IVS18DS, G–A, + 1 (this mutation changes the first nucleotide in the eighteenth intron of the gene, causing exon 18 to be spliced to exon 20, thus skipping exon 19)
  - deletion of the poly-A addition site
  - G-to-A substitution in the 5' UTR
  - insertion of 1000 base pairs into the sixth intron (this particular insertion does not alter splicing)
44. Considering further the mutations described in Problem 43:
- Which of the mutations could be null mutations?
  - Which of the mutations would be most likely to result in an allele that is recessive to a wild-type allele?
  - Which of the mutations could result in an allele dominant to a wild-type allele? What mechanism(s) could explain this dominance?
45. Adermatoglyphia (described previously in Problem 18 in Chapter 3) is an extremely rare condition where people are born without fingerprints; only four families on earth are known to have this condition. The condition is inherited in an autosomal dominant fashion and is due to point mutations in a gene on chromosome 4 called *SMARCAD1*.
- The following figure shows that different point mutations—all near the 5' end of the same intron of *SMARCAD1*—were found in each of the four families. All four mutations prevent the expression of a skin-specific transcript that uniquely contains exon 1, the first exon of this transcript; no other *SMARCAD1* mRNAs contain this exon. In the figure, the final three bases in the RNA-like strand of exon 1 are *shaded*, while the first five bases of intron 1 are *unshaded*.
- |          | Exon 1     | Intron 1 |
|----------|------------|----------|
| Normal   | CTG GTAAGT |          |
| Family 1 | CTG TAAAGT |          |
| Family 2 | CTG GCAAGT |          |
| Family 3 | CTG GTAACT |          |
| Family 4 | CTG ATAAGT |          |
- No ATG sequence normally exists in exon 1 upstream of the sequence shown. Which part of the skin-specific mRNA corresponds to exon 1?
  - What aspect of gene expression is likely to be affected most directly by these mutations?
  - Are these mutations more likely to cause loss of function or gain of function?
46. Homozygosity for extremely rare mutations in a human gene called *SCN9A* cause complete insensitivity to pain (*congenital pain insensitivity* or *CPA*) and a total lack of the sense of smell (*anosmia*). The *SCN9A* gene encodes a sodium channel protein required for transmission of electrical signals from particular nerves in the body to the brain. The failure to feel pain is a dangerous condition as people cannot sense injuries.
- The *SCN9A* gene has 26 exons and encodes a 1977-amino acid polypeptide. Consanguineous matings in three different families have resulted in individuals with CPA/anosmia. In Family 1, a G-to-A transition in exon 15 results in a truncated protein that is

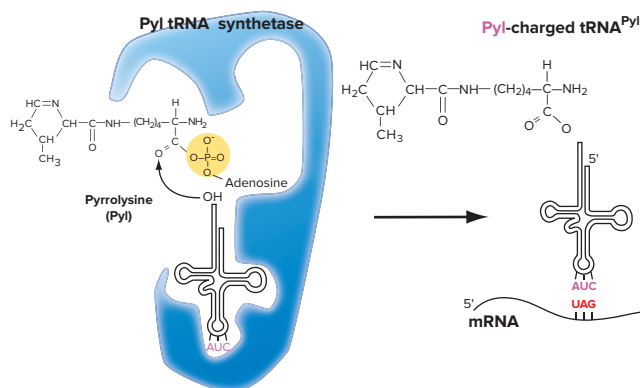
898 amino acids long; in Family 2, deletion of a single base results in a 766-amino acid polypeptide; and in Family 3, a C-to-G transversion in exon 10 yields a 458-amino acid protein.

- a. Hypothesize as to how each of the three *SCN9A* mutations affects gene structure: Why are truncated proteins made in each case?
  - b. How would you classify the mutant alleles? Do these cause loss of function or gain of function? Are they amorphs, hypomorphs, hypermorphs, neomorphs, or antimorphs?
  - c. Explain in molecular terms why CPA/anosmia is a recessive condition.
47. You learned in Problem 21 in Chapter 7 that the neurodegenerative disease ALS can be caused by expansion of a hexanucleotide repeat region (5'-GGGGCC-3') outside of the open reading frame (but within the first intron) of the gene called *C9ORF72*. While a normal *C9ORF72* allele has 2–23 copies of the hexanucleotide repeat unit, dominant disease-causing alleles have hundreds or even thousands of copies.
- Researchers observed that the first intron of the *C9ORF72* disease allele is transcribed not only from the normal template strand of DNA, but also from the nontemplate strand. Even more unusual, both types of repeat-region transcripts are translated in all six reading frames in an AUG-independent manner—a process called *repeat-associated non-ATG translation*, or *RAN translation*. These discoveries led to the hypothesis that the proteins made from the repeats might contribute to ALS.
- a. What polypeptides are made from the repeat-region transcripts?
  - b. According to the RAN translation hypothesis, why are disease-causing *C9ORF72* alleles dominant to normal alleles?
  - c. How would you classify the mutant alleles? Do they cause a loss of function or a gain of function? Are they amorphic, hypomorphic, hypermorphic, neomorphic, or antimorphic? (*Note*: More than one answer might be possible.)
48. When 1 million cells of a culture of haploid yeast carrying a *met<sup>-</sup>* auxotrophic mutation were plated on petri plates lacking methionine (Met), five colonies grew. You would expect cells in which the original *met<sup>-</sup>* mutation was reversed (by a base change back to the original sequence) would grow on the media lacking methionine, but some of these apparent reversions could be due to a mutation in a different gene that somehow suppresses the original *met<sup>-</sup>* mutations. How would you be able to determine if the mutations in your five colonies were due either to a precise reversion of the original *met<sup>-</sup>* mutation or to the generation of a suppressor mutation in a gene on another chromosome?

49. Why is a nonsense suppressor tRNA<sup>Tyr</sup>, even though it has a mutant anticodon that cannot recognize a tyrosine codon, charged with tyrosine by Tyr tRNA synthetase? (*Hint*: Refer to Fig. 8.19.)
50. A mutant *B. adonis* bacterium has a nonsense suppressor tRNA that inserts glutamine (Gln) to match UAG (but not other nonsense) codons. [This species does not modify wobble position C residues to k<sup>2</sup>C and does not have tRNA<sup>Pyl</sup> (see Problem 57).]
  - a. What is the anticodon of the suppressing tRNA? Indicate the 5' and 3' ends.
  - b. What is the sequence of the template strand of the wild-type *tRNA<sup>Gln</sup>* encoding gene that was altered to produce the suppressor, assuming that only a single-base-pair alteration was involved?
  - c. What is the *minimum* number of *tRNA<sup>Gln</sup>* genes that could be present in a wild-type *B. adonis* cell? Describe the corresponding anticodons.
51. You are studying mutations in a bacterial gene that codes for an enzyme whose amino acid sequence is known. In the wild-type protein, proline is the fifth amino acid from the amino terminal end. In one of your mutants with nonfunctional enzyme, you find a serine at position number 5. You subject this mutant to further mutagenesis and recover three different strains. Strain A has a proline at position number 5 and acts just like a wild-type strain. Strain B has tryptophan at position number 5 and also acts like wild type. Strain C has no detectable enzyme function at any temperature, and you can't recover any protein that resembles the enzyme. You mutagenize strain C and recover a strain (C-1) that has enzyme function. The second mutation in C-1 that is responsible for the recovery of enzyme function does not map at the enzyme locus.
  - a. What is the nucleotide sequence in both strands of the wild-type gene at this location?
  - b. Why does strain B have a wild-type phenotype? Why does the original mutant with serine at position 5 lack function?
  - c. What is the nature of the mutation in strain C?
  - d. What is the second mutation that arose in C-1?
52. Another class of suppressor mutations, not described in the chapter, are mutations that suppress missense mutations.
  - a. Why would bacterial strains carrying such missense suppressor mutations generally grow more slowly than strains carrying nonsense suppressor mutations?
  - b. What other kinds of mutations can you imagine in genes encoding components needed for gene expression that would suppress a missense mutation in a protein-coding gene?

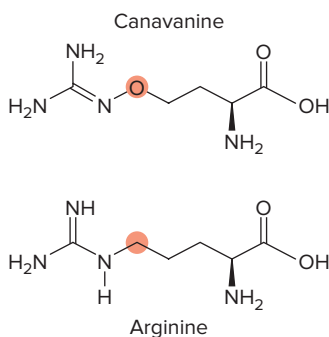
53. Yet another class of suppressor mutations not described in the chapter are mutations in tRNA genes that can suppress frameshift mutations. What would have to be true about a tRNA that could suppress a frameshift mutation involving the insertion of a single base pair?
54. At least one nonsense suppressing tRNA is known that can suppress more than one type of nonsense codon.
- What is the anticodon of such a suppressing tRNA?
  - What stop codons would it suppress?
  - Could this tRNA possibly also function as a missense suppressor?
  - What are the amino acids most likely to be carried by this suppressing tRNA?
55. An investigator was interested in studying UAG nonsense suppressor mutations in bacteria. In one species of bacteria, she was able to select two different mutants of this type, one in a  $tRNA^{Tyr}$  gene and the other in a  $tRNA^{Gln}$  gene, but in a second species, she was not able to obtain any such nonsense suppressor mutations, even after very extensive effort. What could explain the difference between the two species?
56. Brenner's  $m$  mutant phages ( $m^1$ – $m^6$ ) described in Fig. 8.8 were suppressed when grown in suppressor ( $su^-$ ) mutant bacteria; they produced full-length M proteins that functioned like wild-type M protein.
- What gene do you think was mutant in the  $su^-$  bacteria?
  - When the  $m^-$  phages were propagated in the  $su^-$  bacterial strain, not all of the proteins made by the mutant  $m$  alleles were identical to wild-type M protein. How did some of them differ?
57. In certain bacterial species, pyrrolysine (Pyl), sometimes called amino acid 22, is incorporated into polypeptides through an unusual use of the genetic code: Pyl is specified by UAG triplets in the middle of the open reading frame of certain rare genes. These bacteria have a pyrrolysine tRNA synthetase that attaches Pyl to a tRNA with the anticodon 5' CUA 3' (see

accompanying diagram). For Pyl to be incorporated into a protein, Pyl-charged tRNA<sup>Pyl</sup> must arrive at the ribosome before translation is terminated.



- Explain two ways in which the mechanism for Pyl specification differs from that of selenocysteine (Sec) incorporation.
  - How is the mechanism for Pyl specification similar to nonsense suppression? (See Fig. 8.34.)
58. Canavanine is an amino acid similar to arginine (see accompanying figure) that is normally synthesized by some plants. Usually, in plants or animals that don't make canavanine, arginine aminoacyl-tRNA synthetase cannot distinguish between canavanine and arginine, and tRNA<sup>Arg</sup> can be charged with canavanine. Incorporation of canavanine in proteins in place of arginine can cause misfolding and destroys protein structure and function.
- Can you think of a reason why a plant might have evolved the ability to make canavanine?
  - How do you suppose plants that make canavanine escape its toxicity?
  - A particular vining legume called *Dioclea megacarpa* makes canavanine and yet still has a single insect predator, a beetle, *Caryedes brasiliensis*. The beetle lays its eggs on the ripe fruit of the vine, and after hatching, the beetle larvae live in the fruit until they mature into adults. How do you suppose that the beetle evades canavanine toxicity?

(middle photo): © Mark W. Skinner, hosted by the USDA-NRCS PLANTS Database; (right photo): Source: Sarah McCaffrey/ Museum Victoria, <http://www.padil.gov.au/pests-and-diseases/pest/main/142145/41386> CC-BY

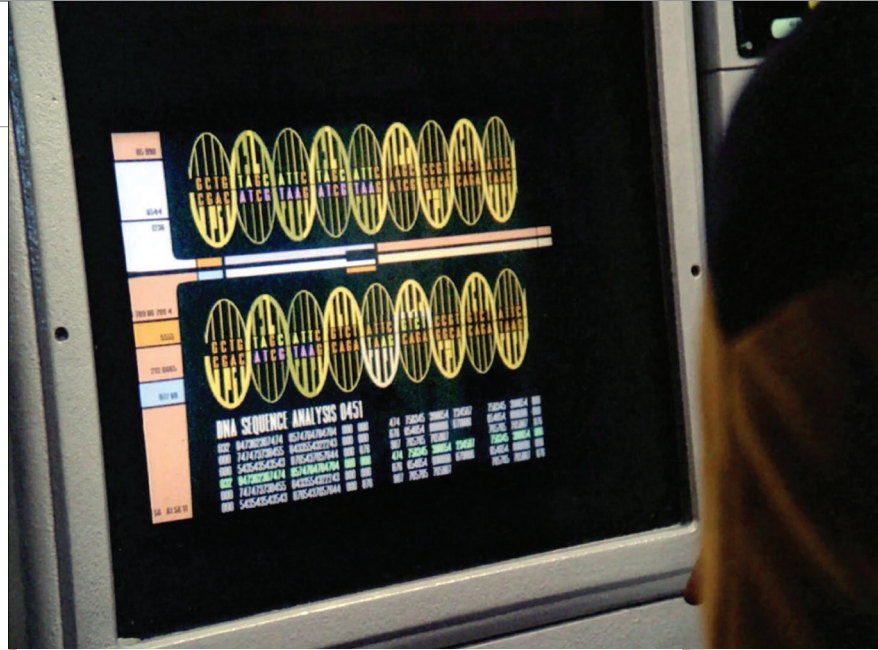


*Dioclea megacarpa*



*Caryedes brasiliensis*

## chapter 9

Digital Analysis  
of DNA

*In 1989, when an episode of Star Trek: The Next Generation featured this shot of DNA sequences on an imaginary computer screen, the ability to sequence the human genome appeared to be a distant dream in the realm of science fiction. Amazingly, the Human Genome Project, which began the following year, achieved this goal less than 15 years later.*

© CBS Photo Archive/Getty Images

SINCE THE MID-NINETEENTH CENTURY, three advances have transformed the field of genetics radically: Mendel's discovery of fundamental principles in the 1860s, Watson and Crick's elucidation of DNA structure in 1953, and the Human Genome Project from 1990 to the present. In this chapter and the next, we discuss the Human Genome Project and the field of **genomics** (the study of genomes) that it spawned.

The **Human Genome Project** was initiated to sequence and analyze the human genome in conjunction with the genomes of several model organisms. A **genome** is the total digital information contained within the DNA sequences of an organism's chromosomes. The haploid human genome contains a total of approximately 3 billion nucleotide pairs.

Prior to the inception of the Human Genome Project, the genome's enormous size caused many biologists to regard the objective of sequencing it as science fiction achievable only in the distant future. Nonetheless, some scientists could foresee the emergence of very fast and reliable automated (*high-throughput*) DNA sequencing methods as well as the computational tools necessary for capturing, storing, and analyzing the vast amounts of data involved. Persuaded by these arguments, agencies of the United States government agreed in 1990 to commit \$3 billion over a projected 15-year period toward completion of the human genome sequence. Several international organizations also joined the enterprise.

Remarkably, investigators were able to determine a rough sequence of the human genome by February 2001. In this *draft*, the sequence had some gaps and did not yet have an appropriate level of accuracy (an error rate of 1/10,000 or less). An accurate sequence covering 97% of the genome was completed shortly thereafter in 2003, two years ahead of schedule. The early finish was prodded by the 1998 promise of Celera, a private company, to complete a draft of the genome in just three years at much lower cost, employing a novel sequencing strategy. The internationally supported genome effort reacted by moving its timetable ahead by several years.

## chapter outline

- 9.1 Fragmenting DNA
- 9.2 Cloning DNA Fragments
- 9.3 Sequencing DNA
- 9.4 Sequencing Genomes



The techniques and approaches developed by the public and private Human Genome Projects also catalyzed efforts to sequence the genomes of many species other than humans. By 2016, whole-genome sequences had been completed for more than 8100 distinct species, revolutionizing study in many areas such as microbiology and plant biology. The availability of genome sequences for these other organisms in turn has important benefits for our understanding of the human genome through the identification of genes and other DNA elements that are conserved across evolutionary lines.

In this chapter, we describe the methods that scientists developed to determine the sequence of the human genome. The general ideas behind genome sequencing are in fact not very complicated. Genomic researchers first fragment the genome into much smaller pieces, and then isolate and amplify (that is, *clone*) individual pieces by making so-called *recombinant DNA molecules*. Next, the scientists determine the DNA sequence of individual purified, bite-sized fragments of the genome. Finally, computer programs analyze the sequence of millions of these snippets to reconstruct the sequence of the whole genome from which the pieces originated.

## 9.1 Fragmenting DNA

### learning objectives

1. Distinguish between digesting DNA with restriction enzymes and mechanical shearing of DNA.
2. Describe how certain restriction enzymes generate DNA fragments with sticky ends, while others generate blunt-ended fragments.
3. Calculate the average sizes and numbers of DNA fragments produced by digesting human genomic DNA with a given restriction enzyme.
4. Summarize the process by which gel electrophoresis separates DNA fragments.

Every intact diploid human cell, including the precursors of red blood cells, carries two nearly identical sets of 3 billion base pairs of information that, when unwound, extend 2 meters in length. This is much too much material and information to study as a whole. To reduce its complexity, researchers first cut the genome into bite-sized pieces that can be analyzed individually. One strategy to accomplish this goal is to use enzymes to cut the genome at specific DNA sequences; an alternative technique is to fragment the genome at random positions by shearing genomic DNA with mechanical forces. Both of these methods have their uses.

### Restriction Enzymes Cut the Genome at Specific Sites

Researchers use *restriction enzymes* to cut the DNA released from the nuclei of cells at specific locations. These

well-defined cuts generate fragments suitable for manipulation and characterization. A **restriction enzyme** recognizes a specific sequence of bases anywhere within the genome and then severs two phosphodiester bonds at that sequence, one in the sugar-phosphate backbone of each strand. The fragments generated by restriction enzymes are referred to as **restriction fragments**, and the act of cutting DNA is often called **digestion**.

Restriction enzymes originate in and can be purified from bacterial cells. As explained in the Tools of Genetics Box *Serendipity in Science: The Discovery of Restriction Enzymes*, these enzymes digest viral DNA to protect prokaryotic cells from viral infection. Bacteria shield their own genomes from digestion by these restriction enzymes through the selective addition of methyl groups ( $-\text{CH}_3$ ) to the restriction recognition sites in their genomic DNA. In the test tube, restriction enzymes from bacteria recognize target sequences of four to eight base pairs (bp) in DNA isolated from any other organism and cut the DNA at or near these sites. **Table 9.1** lists the names, recognition sequences, and microbial origins of just 10 of the close to 300 commonly used restriction enzymes.

For the majority of these enzymes, the recognition site consists of four to six base pairs and exhibits a kind of palindromic symmetry in which the base sequences of each of the two DNA strands are identical when read in the 5'-to-3' direction. Because of this fact, base pairs on either side of a central line of symmetry are mirror images of each other. Each enzyme always cuts at the same place relative to its specific recognition sequence, and most enzymes make their cuts in one of two ways: either straight through both DNA strands right at the line of symmetry to produce fragments with **blunt ends**, or displaced equally in opposite directions from the line of symmetry by one or more bases to generate fragments with single-stranded

**TABLE 9.1** Ten Commonly Used Restriction Enzymes

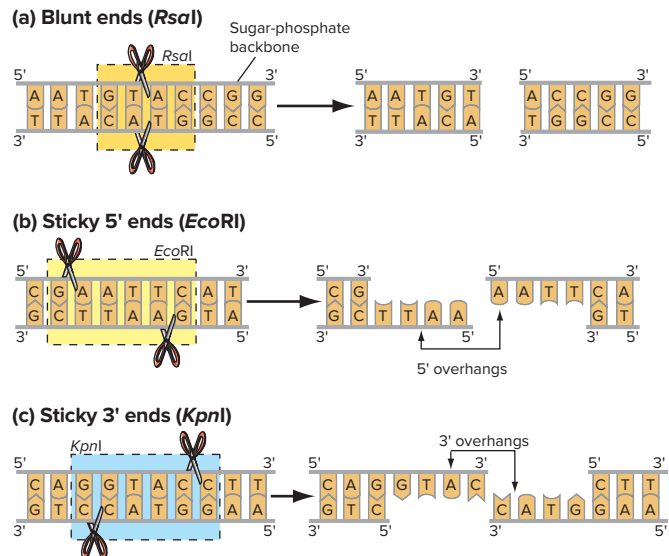
Enzyme	Sequence of Recognition Site	Microbial Origin
<i>TaqI</i>	5' TCGA 3' 3' AGCT 5'	<i>Thermus aquaticus</i> YTI
<i>RsaI</i>	5' GTAC 3' 3' CATG 5'	<i>Rhodopseudomonas sphaeroides</i>
<i>Sau3AI</i>	5' GATC 3' 3' CTAG 5'	<i>Staphylococcus aureus</i> 3A
<i>EcoRI</i>	5' GAATTC 3' 3' CTTAAG 5'	<i>Escherichia coli</i>
<i>BamHI</i>	5' GGATCC 3' 3' CCTAGG 5'	<i>Bacillus amyloliquefaciens</i> H
<i>HindIII</i>	5' AAGCTT 3' 3' TTAGCA 5'	<i>Haemophilus influenzae</i>
<i>KpnI</i>	5' GGATCC 3' 3' CCTAGG 5'	<i>Klebsiella pneumoniae</i> OK8
<i>Clal</i>	5' ATCGAT 3' 3' TAGCTA 5'	<i>Caryophanon latum</i>
<i>BssHIII</i>	5' GGGGGC 3' 3' CCCCAG 5'	<i>Bacillus stearothermophilus</i>
<i>NotI</i>	5' GCGGCCGC 3' 3' CGGCCGCG 5'	<i>Nocardia otitidiscaviarum</i>

ends (Fig. 9.1). Geneticists often refer to these protruding single strands as **sticky ends**. They are considered *sticky* because they are free to base pair with a complementary sequence from the DNA of any organism cut by the same restriction enzyme.

### Restriction Enzymes with Longer Recognition Sites Produce Larger DNA Fragments

Researchers often need to produce DNA fragments of a particular length—larger ones to study the organization of a chromosomal region, smaller ones to examine a whole gene, and ones that are smaller still for DNA sequence analysis (that is, for the determination of the precise order of bases in a DNA fragment). To make these different-sized fragments, scientists can cut DNA with different restriction enzymes that recognize different sequences.

**Figure 9.1** Restriction enzymes cut DNA molecules at specific locations to produce restriction fragments with either blunt or sticky ends. (a) The restriction enzyme *RsaI* produces blunt-ended restriction fragments. (b) *EcoRI* produces sticky ends with a 5' overhang. (c) *KpnI* produces sticky ends with a 3' overhang.

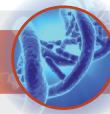


You can estimate the average length of the fragments that a particular restriction enzyme generates if you make two simplifying assumptions: first, that each of the four bases occurs in equal proportions such that a genome is composed of 25% A, 25% T, 25% G, and 25% C; and second, that the bases are distributed randomly in the DNA sequence. These assumptions enable us to estimate the average distance between recognition sites of any length by the general formula  $4^n$ , where  $n$  is the number of bases in the site (Fig. 9.2a).

According to the  $4^n$  formula, *RsaI*, which recognizes the four-base-sequence 5' GTAC 3', will cut on average once every  $4^4$ , or every 256 bp, creating fragments averaging 256 bp in length. By comparison, the enzyme *EcoRI*, which recognizes the six-base-sequence 5' GAATTC 3', will cut on average once every  $4^6$ , or 4096 bp; because 1000 base pairs = 1 kilobase pair, researchers often round off this large number to roughly 4.1 kilobase pairs, abbreviated 4.1 kb. Similarly, an enzyme such as *NotI*, which recognizes the eight bases 5' GCGGCCGC 3', will cut on average every  $4^8$  bp, or every 65.5 kb. Note, however, that because the actual distances between restriction sites for any enzyme vary considerably, very few of the fragments produced by the three enzymes mentioned here will be precisely 256 bp, 4.1 kb, or 65.5 kb in length.

Once you know the average lengths of the fragments produced with a particular restriction enzyme, you can also estimate the number of the fragments that could be produced by treating a genome with that enzyme. For example, we have seen that the four-base cutter *RsaI* cuts the

## TOOLS OF GENETICS



Blue DNA: © MedicalRF.com

### Serendipity in Science: The Discovery of Restriction Enzymes

Most of the tools and techniques for cloning and analyzing DNA fragments emerged from studies of bacteria and the viruses that infect them. Molecular biologists had observed, for example, that viruses able to grow abundantly on one strain of bacteria grew poorly on a closely related strain. While examining reasons for this discrepancy, these scientists discovered restriction enzymes.

To follow the story, one must know that researchers compare rates of viral proliferation in terms of *plating efficiency*: the fraction of viral particles that enter and replicate inside host bacterial cells, causing the cells to lyse and release viral progeny. These progeny go on to infect neighboring cells, which in turn lyse and release further virus particles. When a petri dish is coated with a continuous *lawn* of bacterial cells, an active viral infection forms as a visibly cleared spot, or *plaque*, where bacteria have been eliminated (see Fig. 7.24). The plating efficiency of lambda virus grown on the *E. coli* C strain is nearly 1.0 (Fig. A.1). This means that 100 original virus particles will cause close to 100 plaques on a lawn of *E. coli* C bacteria.

The plating efficiency of the same virus grown on *E. coli* K12 is only 1 in  $10^4$ , or 0.0001. The ability of a bacterial strain to prevent the replication of an infecting virus, in this case the growth of lambda on *E. coli* K12, is called **restriction**.

Restriction is rarely absolute. Although lambda virus grown on *E. coli* K12 produces almost no progeny (the viruses infect cells but can't replicate inside them), a few viral particles inside a few cells do manage to proliferate. If their progeny viruses are then tested on *E. coli* K12, the plating efficiency is nearly 1.0. The phenomenon in which growth on a restricting host modifies a virus so that succeeding generations grow more efficiently on that same host is called **modification**.

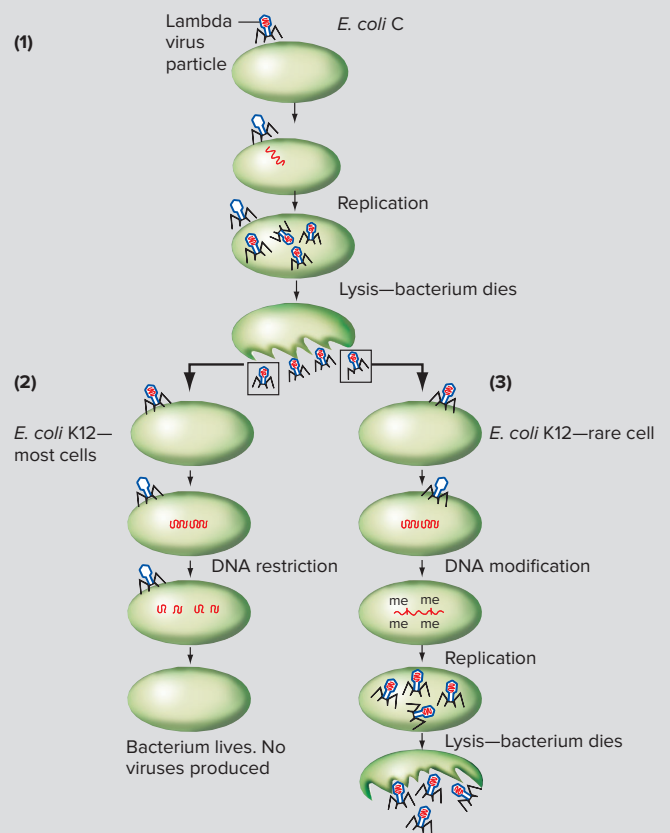
What mechanisms account for restriction and modification? Studies following viral DNA after bacterial infection found that during restriction, the viral DNA is broken into pieces and degraded (Fig. A.2). The enzyme responsible for the initial breakage was found to be an *endonuclease*, an enzyme that breaks phosphodiester bonds, usually making double-strand cuts at specific locations in the viral chromosome. Because this breakage restricts the biological activity of the viral DNA, researchers called the enzymes that accomplish it **restriction enzymes**.

Subsequent studies showed that the small percentage of viral DNA that escapes digestion and goes on to generate new viral particles has been modified by the addition of methyl groups during its replication in the host cell (Fig. A.3). Researchers named the enzymes that add methyl groups to specific DNA sequences **modification enzymes**.

Biologists have identified complementary restriction-modification systems in a wide variety of bacterial strains. Purification of the systems has yielded a mainstay of recombinant DNA technology: the battery of restriction enzymes used to cut DNA *in vitro* for cloning, mapping, and ligation (see Table 9.1).

This example of serendipity in science sheds light on the debate between administrators who distribute and oversee research funding and scientists who carry out the research.

**Figure A** Operation of the restriction enzyme/modification system in nature. (1) *E. coli* strain C does not have a restriction enzyme/modification system and is susceptible to infection by the lambda phage. (2) In contrast, *E. coli* strain K12 generally resists infection by viral particles produced by a previous infection of *E. coli* C. Cells of *E. coli* K12 make the *EcoRI* restriction enzyme, which cuts the lambda DNA molecule before its genes can be expressed. (3) In rare K12 cells, modification enzymes add methyl groups (*me*) to lambda DNA, protecting it from the restriction enzymes. Modified lambda DNA can now replicate, and as the DNA methylation marks are copied during DNA replication, progeny viruses that readily form plaques on K12 bacteria are generated.



Researchers did not set out to find restriction enzymes; they could not have known these enzymes would be one of their finds. Rather, they sought to understand the mechanisms by which viruses infect and proliferate in bacteria. Along the way, they discovered restriction enzymes and how they work. The politicians and administrators in charge of allocating funds often want to direct research spending to urgent health or agricultural problems, while scientists often call for a broad distribution of funds to all projects investigating interesting biological phenomena. The validity of both views suggests the need for a balanced approach to the funding of research activities.

**Figure 9.2** The number of base pairs in a recognition site determines the average size of the fragments produced. (a) *RsaI* recognizes and cuts at a 4 bp site, *EcoRI* cuts at a 6 bp site, and *NotI* cuts at an 8 bp site. (b) *RsaI*, *EcoRI*, and *NotI* restriction sites in a 200 kb region of human chromosome 11, followed by the names and locations of genes in this region. Numbered tick marks at the top are spaced 50 kb apart.

**(a) Calculating Average Restriction Fragment Size**

1. Probability that a four-base recognition site will be found at a given position in a genome =

$$1/4 \times 1/4 \times 1/4 \times 1/4 = 1/256$$

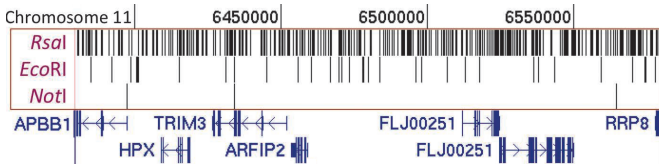
2. Probability that a six-base recognition site will be found =

$$1/4 \times 1/4 \times 1/4 \times 1/4 \times 1/4 \times 1/4 = 1/4096$$

3. Probability that an eight-base recognition site will be found =

$$(1/4)^8 = 1/65,536$$

**(b)**



genome on average every  $4^4$  (256) bp. If you exposed the haploid human genome with its 3 billion bp to *RsaI* for a sufficient time under appropriate conditions, you would ensure that all of the recognition sites in the genome that can be cleaved will be cleaved, and you would get

$$\frac{3,000,000,000 \text{ bp}}{\sim 256 \text{ bp}} = \sim 12,000,000 \text{ fragments that are } \sim 256 \text{ bp in average length}$$

By comparison, the six-base cutter *EcoRI* cuts the DNA on average every  $4^6$  (4096) bp, or every 4.1 kb. If you exposed the haploid human genome with its 3 billion bp, or 3 million kb, to *EcoRI* cleavage, you would get

$$\frac{3,000,000,000 \text{ bp}}{\sim 4100 \text{ bp}} = \sim 700,000 \text{ fragments that are } \sim 4.1 \text{ kb in average length.}$$

And if you exposed the same haploid human genome to the eight-base cutter *NotI*, which cuts on average every  $4^8$  (65,536) bp, or 65.5 kb, you would obtain

$$\frac{3,000,000,000 \text{ bp}}{\sim 65,500 \text{ bp}} = \sim 46,000 \text{ fragments that are } \sim 65.5 \text{ kb in average length.}$$

**Figure 9.2b** summarizes these relationships by depicting the results of cutting one small region of the human genome (containing only seven genes) with these three different restriction enzymes. Clearly, enzymes that recognize larger sites (like the 8 bp site for *NotI*) produce fewer fragments of larger average size than enzymes recognizing smaller sites (like the 4 bp sequence recognized by *RsaI*).

## Mechanical Shearing Forces Break DNA at Random Locations

As will be seen later in this chapter, some types of experiments require the random cutting of DNA so that in a given sample, different copies of the genome will be broken at different positions, rather than always at the same locations as with restriction enzymes. Random cutting of DNA can be achieved by subjecting the molecules to mechanical stress, such as passing the sample through very thin needles at high pressure, or by *sonication* (that is, the application of ultrasound energy). By pulling different parts of a DNA molecule in different directions, these mechanical forces can break phosphodiester bonds at random positions and thus fragment the DNA in the sample. Researchers can obtain fragments of various sizes by changing the amount of mechanical stress; for example, higher-energy ultrasound produces smaller fragments.

The ends of the DNA fragments produced by mechanical shearing are sometimes blunt, or they may have protruding single-stranded regions. If the latter, these single-stranded overhangs are not *sticky* in the same sense as the sticky ends produced by restriction enzymes because they are made up of random sequences and are thus not complementary with other overhangs. Molecular biologists have nonetheless developed elegant techniques that can convert any type of DNA end into any other type of end. All DNA fragments obtained by any procedure can thus ultimately be used in similar ways.

## Gel Electrophoresis Separates DNA Fragments According to Size

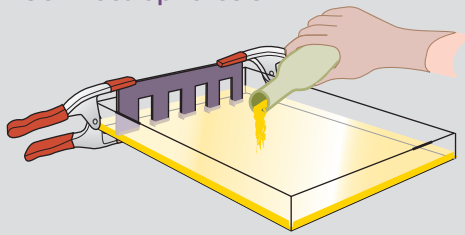
To analyze the DNA in a sample, biologists employ a technique called **electrophoresis**, the movement of charged molecules in an electric field. Biologists use electrophoresis to separate many different types of molecules, for example, DNA of one length from DNA of other lengths, DNA from protein, or one kind of protein from another. In this discussion, we focus on its application to the separation in a gel of DNA fragments of varying length (**Fig. 9.3**).

To carry out such a separation, you place a solution of DNA molecules into indentations called *wells* at one end of a porous gel-like matrix. You then place the gel in a buffered aqueous solution and set up an electric field between bare wires at either end connected to a power supply. The electric field causes all charged molecules in the wells to migrate in the direction of the electrode having an opposite charge. Because all of the phosphate groups in the backbone of DNA carry a net negative charge in a solution near neutral pH, DNA molecules are pulled through the gel toward the wire with a positive charge.

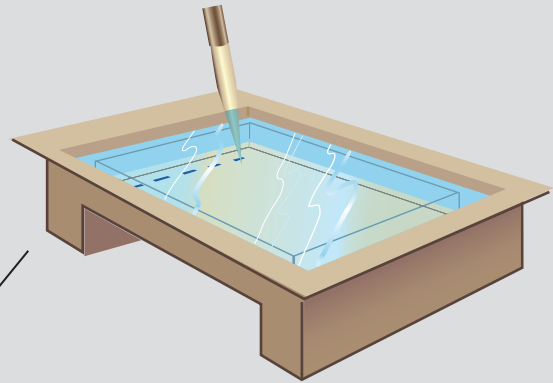
Several variables determine the rate at which DNA molecules (or any other molecules) move during electrophoresis.

## FEATURE FIGURE 9.3

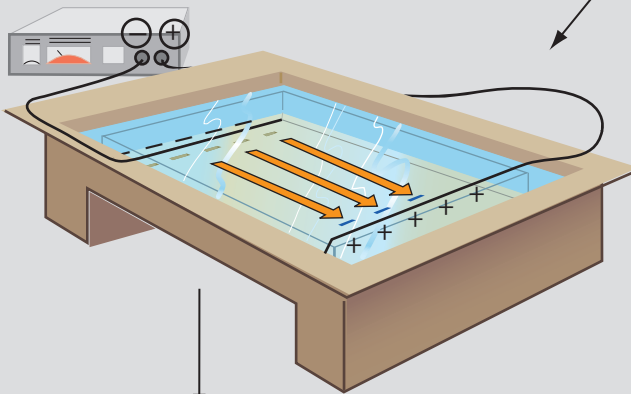
### Gel Electrophoresis



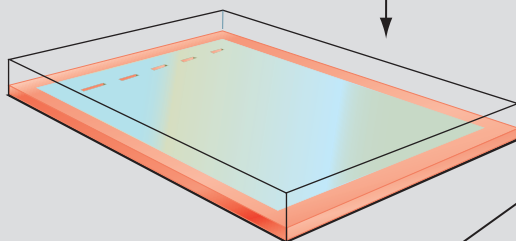
1. Pour heated molten agarose into a clear acrylic plate to which a comb has been attached with clamps. Allow the agarose to cool and harden.



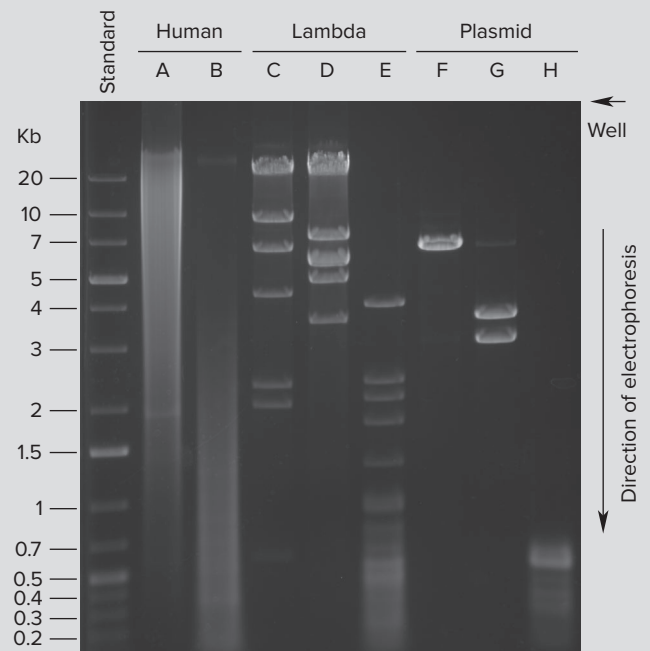
2. Remove the comb; shallow wells will be left in the gel. Remove the gel from the acrylic plate and transfer it to a tank containing a buffered solution. Use a micropipette to load a different DNA sample into each well of the gel. Each sample contains a blue dye to make it easier to see. One sample should contain DNA molecules of known length to serve as size markers.



3. The tank contains electrode wires placed along each end of the gel. Attach these electrodes to a power supply. When you switch on the current, the negatively charged DNA molecules in each sample migrate toward the "+" end of the box, along the paths (*lanes*) shown by the orange arrows. Smaller DNA molecules will move faster toward the "+" end than larger DNA molecules.



4. Remove the gel from the tank. Incubate in a solution containing ethidium bromide (which binds to DNA), then wash with water to remove excess dye from the gel.



5. Expose the gel to ultraviolet (UV) light. DNA molecules will fluoresce as orange bands because the ethidium bromide bound to the DNA absorbs UV photons and gives off photons in the visible red range. You can estimate the size of the DNA molecules in the unknown samples by comparing their migration in the gel with that of the size markers (*standards*) in the lane at the left.

(5): © Lee Silver, Princeton University

**Separating DNA molecules according to their size by agarose gel electrophoresis.** To prepare an agarose gel with wells for samples, you pour the gel as shown in Step 1. You then transfer the gel to a tank containing a buffered solution with ions that allow current to flow, and load DNA samples in the wells (Step 2). You then connect the gel tank to a power supply and allow electrophoresis to run from 1 to 20 hours (depending on the DNA size and the voltage; Step 3). After incubating the gel with the fluorescent dye ethidium bromide (Step 4), you then expose the gel to UV light (Step 5). DNA molecules will appear as orange bands because they bind to the fluorescent dye.

Step 5 shows actual results from gel electrophoresis; because black-and-white film was used, DNA appears white rather than orange. The standard lane at left has DNA fragments of known sizes. Human genomic DNA was cut with *EcoRI* in lane A and with *RsaI* in lane B. Smears containing hundreds of thousands of fragments are produced with an average size of about 4.1 kb for *EcoRI* and 256 bp for *RsaI*. In C, D, and E, the chromosome of bacteriophage  $\lambda$  was cut with *HindIII*, *EcoRI*, and *RsaI*, respectively. The sizes of the fragments in any one lane add up to 48.5 kb, the size of the viral genome. In F, G, and H plasmid DNA of total length 6.9 kb was cut with the same three enzymes. Note that the larger the genome analyzed, the more fragments are produced; moreover, the more bases in the restriction enzyme recognition site, the larger is the average size of the fragments produced.

These variables are the strength of the electric field applied across the gel, the composition of the gel, the charge per unit volume of the DNA molecule (known as *charge density*), and the physical size of the molecule. The only one of these variables that actually differs among any set of linear DNA fragments migrating in a particular gel is size. The reason is that all the DNA molecules are subjected to the same electric field and the same gel matrix, and they all have the same charge density (because the charge of all nucleotide pairs is nearly identical). As a result, only differences in size cause different linear DNA molecules to migrate at different speeds during electrophoresis.

The longer a molecule of linear DNA, the larger the volume it occupies as a random coil. The larger the volume a molecule occupies, the less likely it is to find a pore in the gel matrix big enough to squeeze through, and the more often it will bump into the matrix. And the more often the molecule bumps into the matrix, the lower its rate of migration (also referred to as its *mobility*). Thus, in any given period of electrophoresis, smaller DNAs will travel greater distances from the wells than larger DNAs.

When electrophoresis is completed, the gel is incubated with a fluorescent DNA-binding dye called *ethidium bromide*. After the unbound dye has been washed away, it is easy to visualize the DNA by placing the gel under an ultraviolet light, which causes the dye bound to the DNA fragments to glow in an orange color. You can determine the actual sizes of DNA molecules on gels by comparing their migration distances to those of known *marker fragments* subjected to electrophoresis in an adjacent lane of the gel.

Figure 9.3 (step 5) illustrates the types of results obtained by analyzing different DNA samples. If a genome is small, such as the 48.5 kb constituting the chromosome of the bacterial virus bacteriophage lambda ( $\lambda$ ), then *EcoRI* restriction enzyme digestion of this DNA will produce a small number of discrete bands that can be discriminated from each other easily by gel electrophoresis and whose sizes total 48.5 kb. In contrast, electrophoresis of the hundreds of thousands of different fragments created when a sample of human genomic DNA is treated with the same enzyme will generate a smear centered around the average fragment size (about 4.1 kb for *EcoRI*, as previously discussed). Random breakage of DNA by mechanical forces will also produce a distribution of fragments whose average size will reflect the intensity of the shearing forces applied to the sample (*not shown*).

DNA molecules range in size from small fragments of fewer than 10 bp to whole human chromosomes that have an average length of 130,000,000 bp. No one sizing procedure has the capacity to separate molecules throughout this enormous range. To detect DNA molecules in different size ranges, researchers use a variety of protocols based mainly on two kinds of gels: *polyacrylamide* (formed by covalent bonding between acrylamide monomers), which is good for distinguishing smaller DNA fragments (less than 1 kb);

and *agarose* (formed by the noncovalent association of agarose polymers), which is suitable for looking at larger fragments up to about 20 kb as in Fig. 9.3.

### essential concepts

- *Restriction enzymes* cut DNA molecules at specific sequences; *mechanical shearing* breaks DNA at random locations.
- The longer the sequence recognized by a restriction enzyme, the fewer but larger will be the fragments the enzyme produces when cutting genomic DNA.
- Certain restriction enzymes can produce fragments that all have the same *sticky ends*.
- *Gel electrophoresis* separates DNA fragments according to their sizes. The smaller the fragment, the farther it will migrate in the gel.

## 9.2 Cloning DNA Fragments

### learning objectives

1. Diagram the process by which restriction enzymes and DNA ligase are used to make recombinant DNA molecules.
2. Describe how scientists produce cellular clones of recombinant DNA molecules.
3. Contrast the use of plasmid vectors with that of BAC or YAC (bacterial or yeast artificial chromosome) vectors.
4. Explain why genomic DNA libraries require more colonies than are contained by a single genome equivalent.

The smear of hundreds of thousands of different DNA fragments seen in the gel of *EcoRI*-cut human DNA in Fig. 9.3 suggests that the genomes of animals, plants, and even microorganisms like *E. coli* are so complex that you can make sense of them only by looking at a small piece at a time. Ideally, you would like to purify just one of these fragments—a tiny bit of the genome—away from all the other fragments. You would then like to *amplify* this particular DNA fragment—that is, make many identical DNA copies of it. Amplification would allow you to obtain enough DNA to study, the most obvious type of analysis being to determine the sequence of nucleotides that constitute this particular fragment. If you could sequence separately each of the hundreds of thousands of fragments, you might ultimately be able to figure out the genome's entire DNA sequence.

The process that uses living cells both to isolate a single fragment of DNA from a complex mixture and to make many exact replicas of that fragment is called **molecular cloning**. This technique was central to the initial success of

the Human Genome Project. Sophisticated methods have been developed recently to circumvent the need for molecular cloning in determining the sequence of genomes, and some of these techniques will be described in later chapters. However, molecular cloning remains today an essential component of many important approaches to the analysis and manipulation of DNA.

Molecular cloning consists of two basic steps. In the first, DNA fragments are inserted into specialized chromosome-like carriers called **cloning vectors**, which ensure the transport, replication, and purification of individual DNA inserts. In the second step, the combined vector-insert molecules are transported into living cells, and the cells make many copies of these molecules. Because all the copies of a given fragment are identical, the group of replicated DNA molecules is known as a **DNA clone**. DNA clones may be purified for immediate study or stored within cells or viruses as collections of clones known as *libraries* for future analysis. We now describe each step of molecular cloning.

## Ligating Inserts to Vectors Produces Recombinant DNA Molecules

On their own, small fragments of human genomic DNA cannot reproduce themselves in a cell. To make replication possible, it is necessary to splice each fragment to a vector. Vectors must contain two kinds of specialized DNA sequences: one to provide a means of replication for the vector and the foreign DNA inserted into it, and the second to signal the vector's presence to an investigator by conferring a detectable property on the host cell. A vector must also have distinguishing physical traits, such as size or shape, by which it can be purified away from the host cell's genome. Several types of vectors are in use, and each one behaves as a minichromosome capable of accepting foreign DNA inserts and replicating independently of the host cell's genome. The cutting and ligating together of vector and inserted fragment—DNA from two different origins—creates a **recombinant DNA molecule**.

### Sticky ends and base pairing

Two characteristics of sticky ends provide a basis for the efficient production of a vector-insert recombinant: First, the single-strand overhangs are available for base pairing. Second, no matter what the origin of the DNA (bacterial or human, for example), two sticky ends produced with the same enzyme are always *compatible*, that is, complementary in sequence.

To make recombinant DNA molecules, you simply cut the vector with the same restriction enzyme used to generate the fragment of genomic DNA, and then you mix the digested vector and genomic DNAs together in the presence of the enzyme DNA ligase (**Fig. 9.4**). The complementary

sticky ends will form base pairs and the ligase will stabilize the molecule by forming phosphodiester bonds between adjacent nucleotides (one from the vector and one from the genomic DNA insert).

Laboratory tricks can increase the efficiency and general utility of molecular cloning. For example, certain procedures prevent two or more genomic fragments from joining with each other rather than with the vectors. Other methods minimize the chance that vector molecules can reseat themselves without including an insert of genomic DNA. Yet other manipulations can be performed to connect fragments of genomic DNA that do not have sticky ends to vectors. These techniques ensure that researchers can reliably produce the molecular clones they intend.

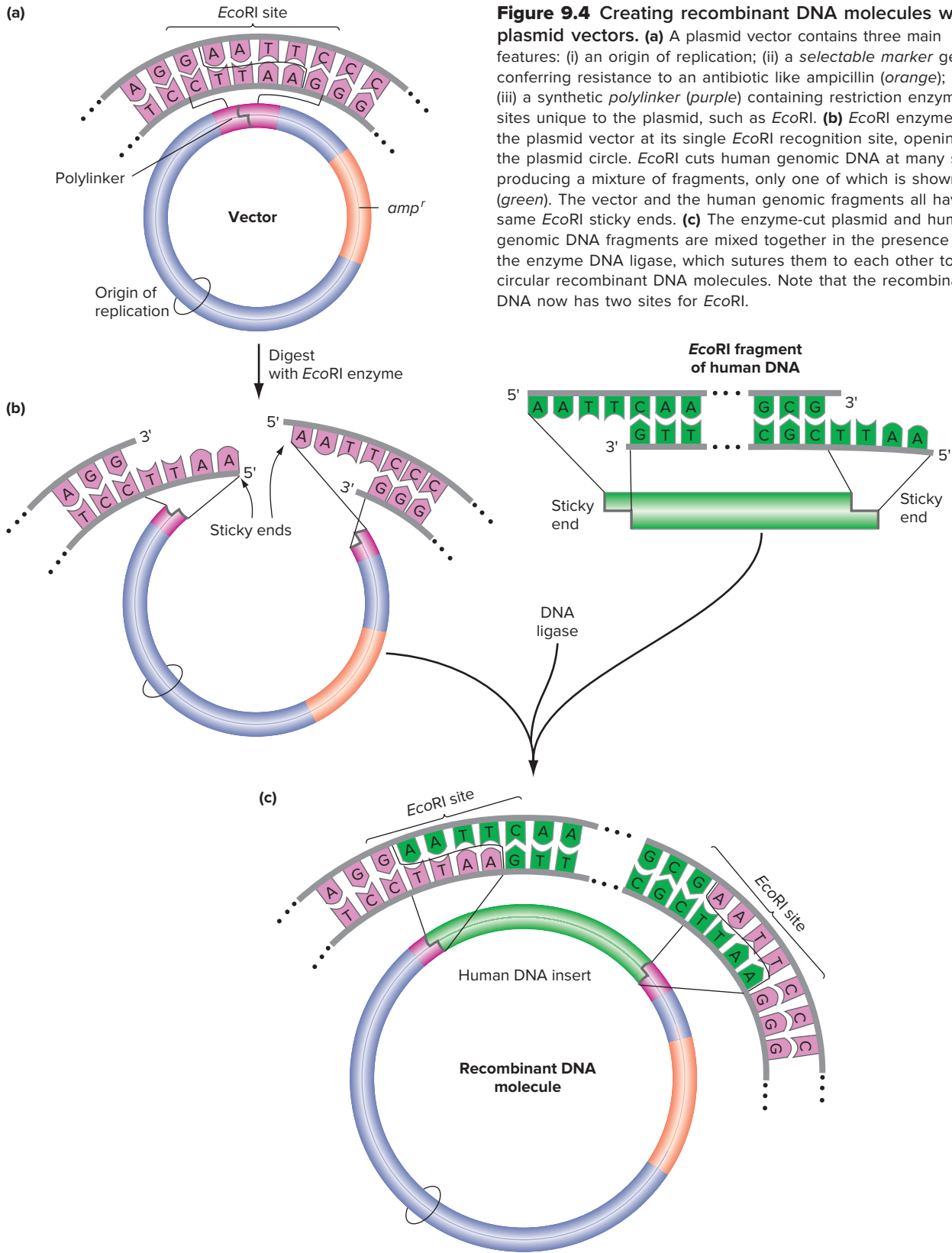
### Choice of vectors

Available vectors differ from one another in biological properties, carrying capacity, and the type of host they can infect. Different types of vectors have different experimental uses.

The simplest vectors are small circles of double-stranded DNA known as **plasmids** that can gain admission to and replicate within many kinds of bacterial cells, independently of the bacterial chromosomes (Fig. 9.4a). The most useful plasmids contain a **polylinker**, which is a short, synthetic DNA sequence that contains a number of different restriction enzyme restriction sites (Fig. 9.4a). Each of these sites is found once in the polylinker but nowhere else in the plasmid vector. The polylinker provides flexibility in the choice of enzymes that can be used to digest the DNA containing the fragment of interest. Exposure to any one of these restriction enzymes opens up the vector at the corresponding recognition site, allowing the insertion of a foreign DNA fragment cut with the same enzyme, without at the same time splitting the plasmid into many pieces (Fig. 9.4b and c). Plasmid vectors can carry only relatively small foreign DNA fragments less than about 20 kb long.

Each plasmid vector carries an origin of replication and a gene for resistance to a specific antibiotic (Fig. 9.4a). The origin of replication enables the plasmid to replicate independently inside a bacterium. The gene for antibiotic resistance confers on the host cell the ability to survive in a medium containing a specific antibiotic; the resistance gene thereby enables experimenters to select for propagation only those bacterial cells that contain a plasmid. Antibiotic resistance genes and other vector genes that make it possible to pick out cells harboring a particular DNA molecule are called **selectable markers**. Plasmids fulfill the final requirement for vectors—ease of purification—because they can be purified away from the genomic DNA of the bacterial host by several techniques that take advantage of size and other differences.

The largest-capacity vectors are *artificial chromosomes*: recombinant DNA molecules that combine replication and segregation elements in such a way that they



**Figure 9.4** Creating recombinant DNA molecules with plasmid vectors. (a) A plasmid vector contains three main features: (i) an origin of replication; (ii) a *selectable marker* gene conferring resistance to an antibiotic like ampicillin (*orange*); and (iii) a synthetic *polylinker* (*purple*) containing restriction enzyme sites unique to the plasmid, such as *EcoRI*. (b) *EcoRI* enzyme cuts the plasmid vector at its single *EcoRI* recognition site, opening up the plasmid circle. *EcoRI* cuts human genomic DNA at many sites, producing a mixture of fragments, only one of which is shown (*green*). The vector and the human genomic fragments all have the same *EcoRI* sticky ends. (c) The enzyme-cut plasmid and human genomic DNA fragments are mixed together in the presence of the enzyme DNA ligase, which sutures them to each other to form circular recombinant DNA molecules. Note that the recombinant DNA now has two sites for *EcoRI*.



behave like normal chromosomes when introduced into a host cell. A bacterial artificial chromosome (BAC) can accommodate a DNA insert of 300 kb. Yeast artificial chromosomes (YACs) can incorporate even larger DNA inserts up to 2000 kb (2 Mb). In addition to their use in molecular cloning, YACs can help investigators analyze functional elements of chromosomes such as centromeres; we will thus discuss YACs in greater detail in Chapter 12 on The Eukaryotic Chromosome.

## Host Cells Take Up and Amplify Recombinant DNA

Although each type of vector functions in a slightly different way and enters a specific kind of host, the general scheme of entering a host cell and taking advantage of the cellular environment to replicate itself is the same for all. **Figure 9.5** illustrates how scientists obtain *E. coli* cells that contain recombinant DNA molecules in which human DNA fragments were ligated into a plasmid vector. The procedure starts with vector and human genomic DNAs cut with the same restriction enzyme, which are then mixed together in the presence of DNA ligase to create hundreds of thousands of different recombinant DNAs, each with a different fragment of the human genome (Fig. 9.5a). Researchers must then introduce these molecules into *E. coli* such that each cell contains only a single type of recombinant DNA.

### Transformation of host cells

Transformation, as you saw in Chapter 6, is the process by which a cell or organism takes up a foreign DNA molecule, changing the genetic characteristics of that cell or organism. What we now describe is similar to what Avery and his colleagues did in the transformation experiments that determined DNA was the molecule of heredity (recall Fig. 6.4), but the method outlined here is more efficient.

Recombinant DNA molecules are first added to a suspension of specially prepared *E. coli* that are sensitive to the antibiotic ampicillin. Under conditions favoring entry, such as suspension of the bacterial cells in a cold  $\text{CaCl}_2$  solution or treatment of the solution with high-voltage electric shock (a technique known as *electroporation*), the plasmids will enter about 1 in 1000 cells (**Fig. 9.5b**). These protocols increase the permeability of the bacterial cell membrane, in essence punching temporary holes through which the DNA gains entry. The probability that any one plasmid will enter any one cell is so low (0.001) that the probability of simultaneous entry of two plasmids into a single cell is insignificant ( $0.001 \times 0.001 = 0.000001$ ).

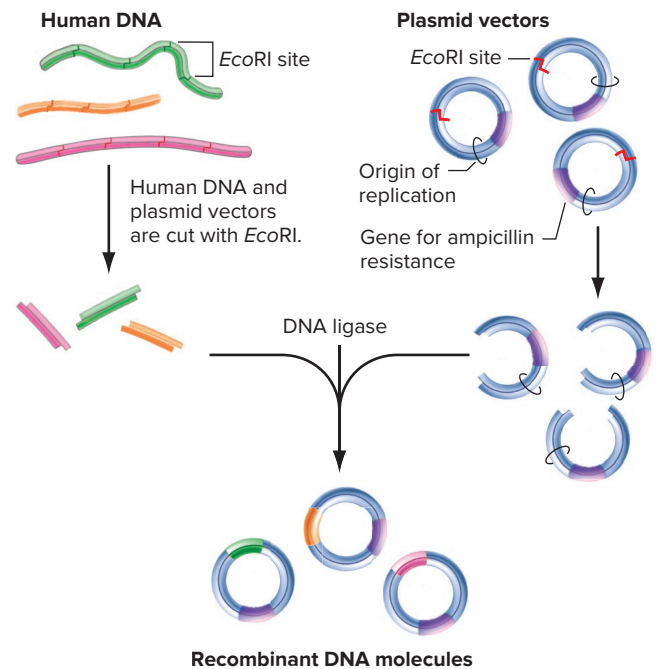
### Identification and isolation of transformed cells

To identify the 0.1% of cells housing a plasmid, the bacteria-plasmid mixture is decanted onto a plate containing agar,

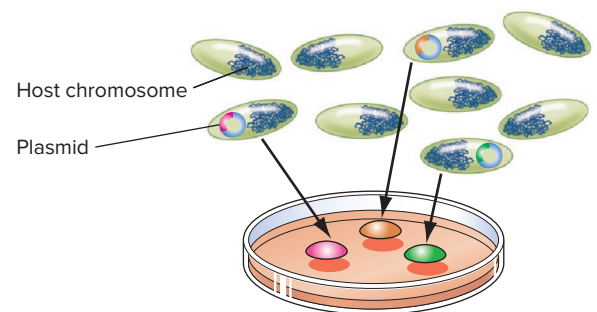
## Figure 9.5 Cloning recombinant DNA molecules.

**(a)** Recombinant DNA construction. Cutting genomic DNA with a restriction enzyme produces many fragments, each of which can form a different recombinant DNA molecule. **(b)** Obtaining clones of bacterial cells containing recombinant plasmids. Recombinant DNAs [from part (a)] are added to ampicillin-sensitive *E. coli* cells. Only cells transformed with recombinant plasmids (or more rarely with a religated vector lacking a foreign DNA insert) will grow on a petri plate containing ampicillin. Each colony on the plate contains millions of identical descendants from a single bacterial cell transformed with a single recombinant DNA molecule.

### (a) Constructing recombinant DNA molecules.



### (b) Transforming *E. coli* cells with recombinant DNAs



*E. coli* plated onto medium containing ampicillin. Only cells containing plasmids are able to grow.

nutrients, and ampicillin. Only cells transformed by a plasmid providing resistance to ampicillin will be able to grow and multiply in the presence of the antibiotic (Fig. 9.5b). The plasmid's origin of replication enables it to replicate in the bacterial cell independently of the bacterial chromosome; in fact, most plasmids replicate so well that a single bacterial

cell may end up with hundreds of identical copies of the same plasmid molecule.

Each viable plasmid-containing bacterial cell will multiply to produce a distinct spot on an agar plate, consisting of a colony of tens of millions of genetically identical cells. The colony as a whole is considered a **cellular clone**. Such clones can be identified when they have grown to about 1 mm in diameter. The millions of identical plasmid molecules contained within a colony together make up a **DNA clone** (Fig. 9.5b).

## Libraries Are Collections of Cloned Fragments

Moving step by step from the DNA of any organism to a single purified DNA fragment is a long and tedious process. Fortunately, scientists do not have to return to step 1 every time they need to purify a new genomic fragment from the same organism. Instead, they can build a **genomic library**: a long-lived collection of cellular clones that contains copies of every sequence in the whole genome inserted into a suitable vector (Fig. 9.6). Like traditional book libraries, genomic libraries store large amounts of information for retrieval upon request. They make it possible to start a new cloning project at an advanced stage, when the initial step of recombinant DNA construction has already been completed and the only difficult task left is to determine which of the many clones in the library contains the DNA sequence of interest. Once the correct cellular clone is identified, it can be amplified to yield a large amount of the desired genomic fragment.

If you digested the genome of a single cell with a restriction enzyme and ligated every fragment to a vector with 100% efficiency, and you then transformed all of these

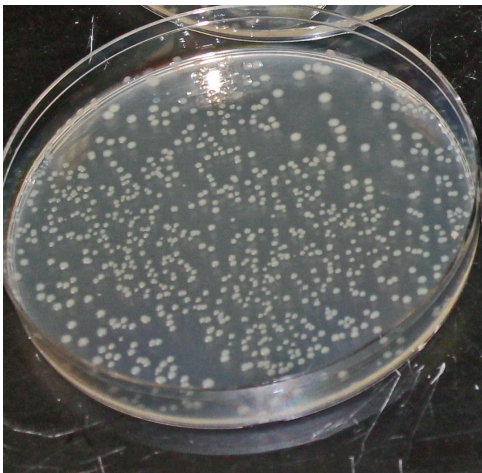
recombinant DNA molecules into host cells also with 100% efficiency, the resulting set of clones would represent the entire genome in a fragmented form. A hypothetical collection of cellular clones that includes one copy—and one copy only—of every sequence in the entire genome would be a single *complete genomic library*.

How many clones are present in this hypothetical library? If you started with the 3,000,000 kb of DNA from a haploid human sperm and reliably cut it into a series of 150 kb restriction fragments, you would generate  $3,000,000/150 = 20,000$  genomic fragments. If you placed each and every one of these fragments into BAC cloning vectors that were then transformed into *E. coli* host cells, you would create a perfect library of 20,000 clones that collectively carry every locus in the genome. The number of clones in this perfect library defines a **genomic equivalent**. To find the number of clones that constitute one genomic equivalent for any library, you simply divide the length of the genome (here, 3,000,000 kb) by the average size of the inserts carried by the library's vector (in this case, 150 kb).

In real life, it is impossible to obtain a perfect library. Each step of cloning is far from 100% efficient, and the DNA of a single cell does not supply sufficient raw material for the process. Researchers must thus harvest DNA from among the millions of cells in a particular tissue or organism. If you make a genomic library with this DNA by collecting only one genomic equivalent (20,000 clones for a human library in BAC vectors), then by chance some human DNA fragments will appear more than once, while others will not be present at all. Including four to five genomic equivalents produces an average of four to five clones for each region (locus) of the genome, and a 95% probability that any individual locus is present at least once.

**Figure 9.6** Part of a human genomic DNA library. Each colony on these plates contains a different recombinant plasmid with a different fragment of the human genome.

© McGraw-Hill Education. Lisa Burgess, photographer



### essential concepts

- To form *recombinant DNA molecules*, DNA ligase links together restriction-enzyme-cut vector and genomic DNA fragments with the same sticky ends.
- Vectors include an origin for DNA replication and a *selectable marker* such as a gene for antibiotic resistance. Plasmid vectors are used for small DNA inserts (<20 kb), while BAC or YAC vectors can carry much larger inserts.
- *Recombinant DNAs* are transformed into host cells. When a single transformed cell is grown into a *cellular clone*, each cell in the clone has the same recombinant DNA.
- *Genomic libraries* contain a species' entire genome as a collection of random DNA fragments; multiple *genome equivalents* are needed to ensure representation of all parts of the genome.

## 9.3 Sequencing DNA

### Learning Objectives

1. Explain the roles of DNA polymerase, the template, and the primer in a Sanger sequencing reaction.
2. Describe the role of dideoxynucleotides in generating DNA fragments for analysis.
3. Interpret the fluorescent peaks obtained during a DNA sequencing run as a sequence of nucleotides with the proper polarity.

Looking at Fig. 9.6, you see petri plates containing thousands of separate colonies constituting part of a human genomic library. Each clone contains a different recombinant DNA molecule, each with a plasmid vector attached to a different fragment of human genomic DNA. Note that the colonies are scattered randomly around the plate, so their arrangement on the plate has no correspondence to their relative order in the genome. How can you then tell which colony contains which fragment of human DNA?

With the technology available today, the simplest way to answer this question is to sequence the human DNA insert in each clone. The DNA sequencing technology in widest current use is based on an original method developed in the mid-1970s by Fred Sanger, who won one of his two Nobel Prizes for this work (the other was awarded for his determination of the amino acid sequence of the hormone insulin). Sanger's methodology can be automated easily, providing the power needed to sequence the 3 billion nucleotides in the human genome.

### Sanger Sequencing Depends on DNA Polymerase

Sanger based his technique on his knowledge of the way DNA is replicated in cells. You will recall from Chapter 6 that the enzyme DNA polymerase catalyzes DNA replication. As summarized in Fig. 9.7a, this enzyme's minimal requirements are: (1) a **template**; that is, a single strand of DNA to copy; (2) **deoxyribonucleotide triphosphates** (dATP, dCTP, dGTP, and dTTP) that are the basic building blocks for newly synthesized DNA; and (3) a **primer**; that is, a short single-stranded DNA molecule (an *oligonucleotide*) that is complementary to part of the template and that provides the free 3' end to which DNA polymerase can attach new nucleotides.

To sequence DNA using Sanger's methods, you would need a template, part of whose sequence is known, but the remainder of which is unknown (because that is what you are trying to determine). One strand of the DNA of a recombinant plasmid could serve as such a template: You know the DNA sequence of the vector (which will be the same in all

clones in the library) but not of the genomic DNA insert (which will vary in different clones) (Fig. 9.7a).

Next, you would need a short oligonucleotide primer complementary to the known sequence of the vector just adjacent to the unknown human DNA insert (Fig. 9.7a). Primers are made to order in *DNA synthesizers*, machines that can manufacture large quantities of any given DNA oligonucleotide up to 100 bases in length. The user simply types in the desired sequence of nucleotides into the computer controlling the DNA synthesizer, and the machine then strings those nucleotides together in the proper order using chemical reactions. You can design the primer because you already know the sequence of the vector, which was determined by alternative chemical techniques (not described here) that do not require prior knowledge.

Sanger sequencing allows the template and primer depicted in Fig. 9.7a to interact through the process of **hybridization**: the natural tendency of complementary single-stranded molecules of DNA or RNA to base pair and form double helices. To make the template, you could simply amplify and purify double-stranded recombinant DNA from one particular clone, and then melt the DNA into single strands by raising the temperature so as to disrupt the hydrogen bonds that would otherwise keep the strands together. Although both strands of a DNA fragment are present in a typical DNA sample, only one is used as a template for sequencing. You then mix in large amounts of the previously made primer. As the temperature of the mixture is gradually lowered, hydrogen bonds will form between complementary nucleotides of the primer and the template strand of the recombinant DNA. The primer you make in the DNA synthesizer must be long enough to ensure that it will form a stable double-stranded region (that is, *anneal*) only with the one complementary sequence in the template; usually primers are between 17 and 25 bases long. The interaction of primer and template creates the substrate for the action of DNA polymerase (Fig. 9.7a).

### Sanger Sequencing Generates Nested Sets of Single-Stranded DNA Fragments

To reveal the order of base pairs in an isolated DNA molecule, Sanger sequencing uses DNA polymerase to create a series of single-stranded fragments, where part of each fragment is complementary to the unknown portion of the DNA template under analysis (Fig. 9.7b). Each fragment differs in length by a single nucleotide from the preceding and succeeding fragments; the graduated set of fragments is known as a *nested array*. A critical feature of the fragments is that each one is distinguishable according to its terminal 3' base. Thus, each fragment has two defining attributes—relative length and one of four possible terminating nucleotides.

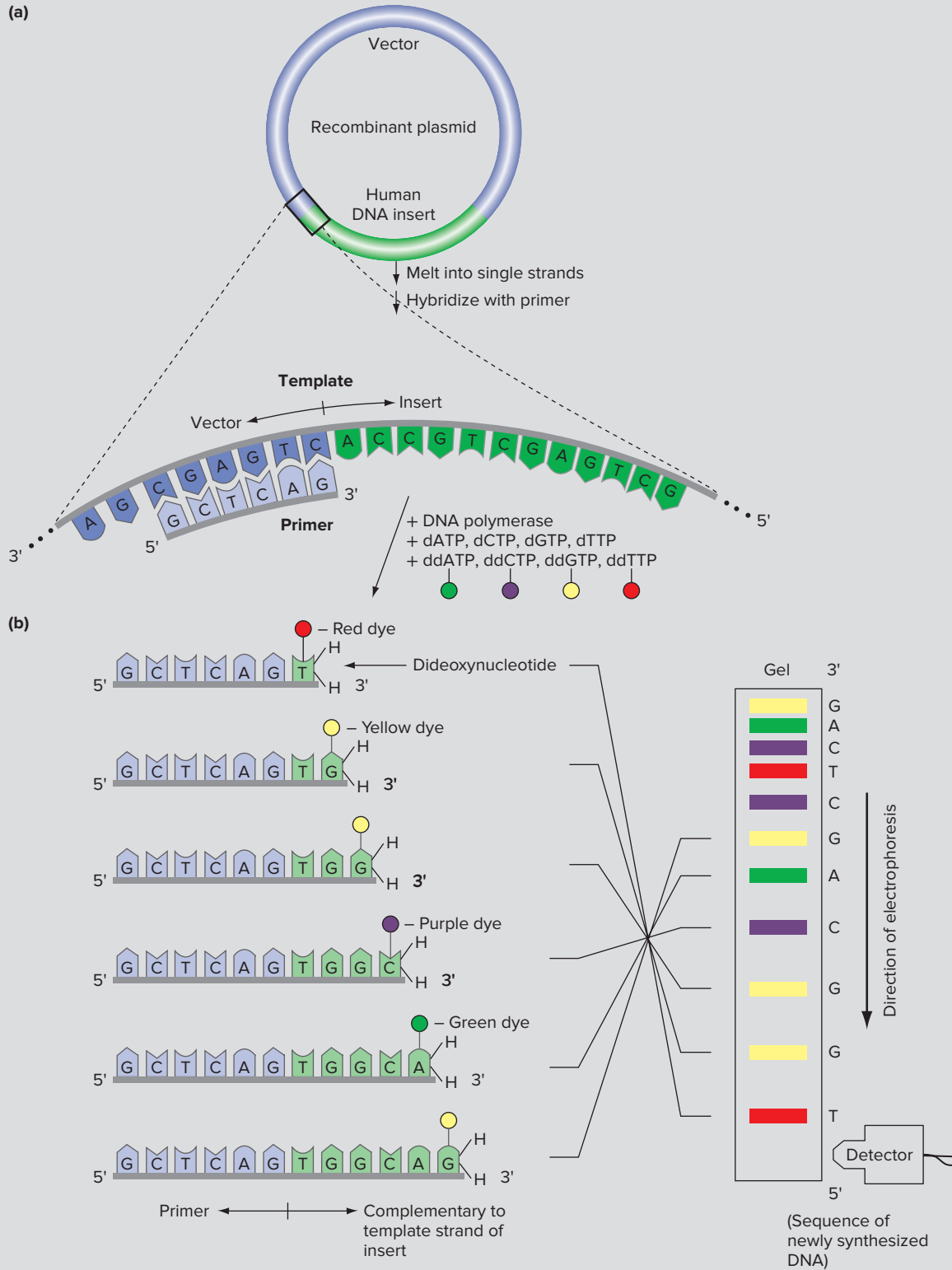
The sequencing procedure to create the nested array begins by adding DNA polymerase to the annealed

## FEATURE FIGURE 9.7

### Automated Sanger Sequencing

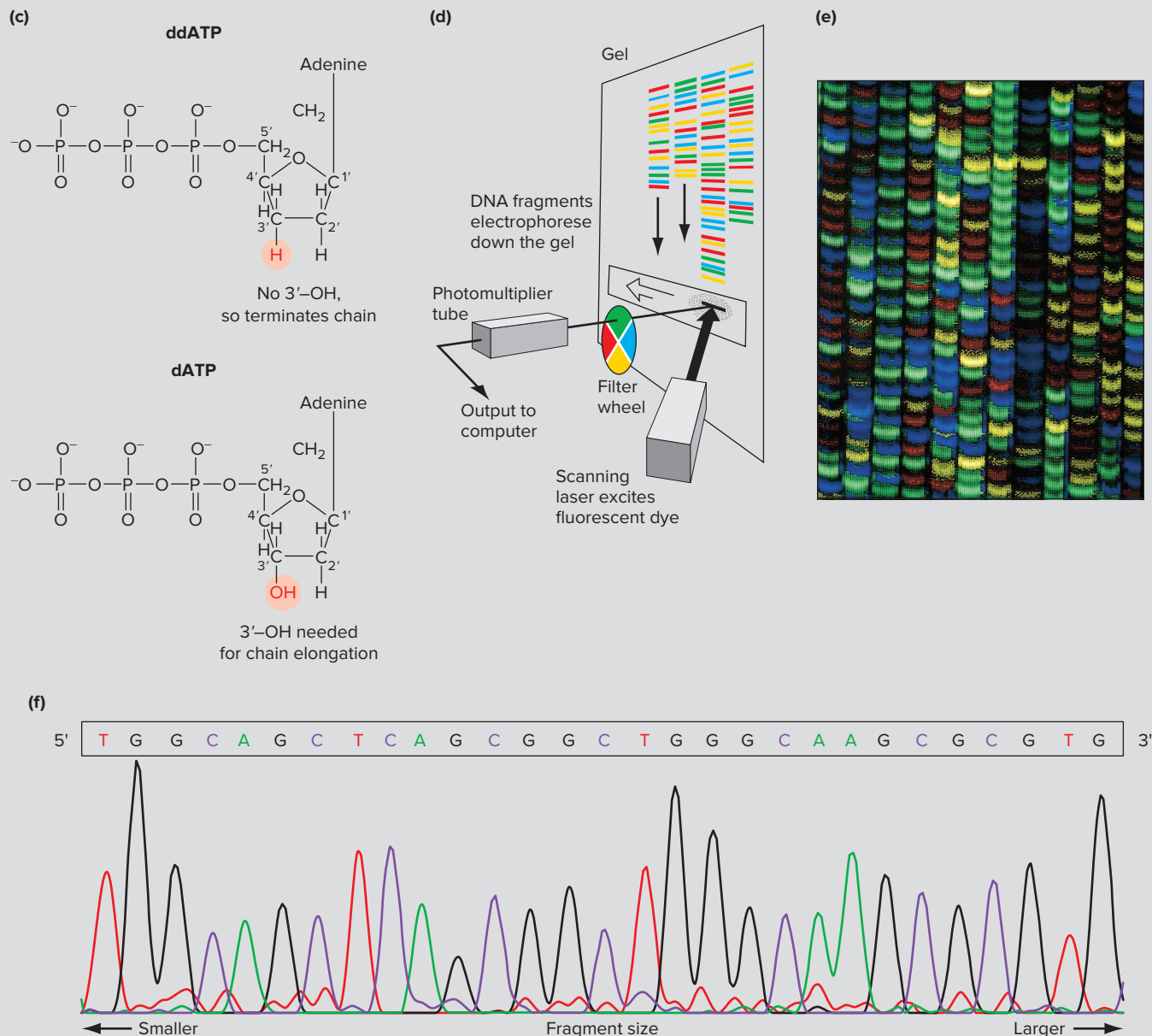
(a) Hybridization of template and primer. A cloned recombinant DNA is **denatured** (melted into single strands), using heat to break apart the hydrogen bonds joining the two strands. One of

these strands serves as the **template**. The recombinant DNA is mixed with an **oligonucleotide primer** (previously made in a DNA synthesizer) whose sequence is complementary to about



20 bases in the vector portion of the template strand. As the temperature is lowered, the template and primer **anneal (hybridize)** together. **(b)** Generating a nested set of polymerization products. The hybrid template-primer is now mixed with DNA polymerase, large amounts of the four **deoxynucleotide triphosphates (dNTPs)**, and smaller amounts of the four **dideoxynucleotide triphosphates (ddNTPs)**. Each ddNTP has a fluorescent tag of a different color. DNA polymerase synthesizes a new strand of DNA complementary to the template by adding nucleotides sequentially onto the 3' end of the primer. Synthesis terminates when a dideoxynucleotide is added to the chain. The reaction generates a nested set of products, each with the same 5' end but a different 3' end. The terminating dideoxynucleotide at the 3' end color-codes each product. After melting apart the newly synthesized DNA from the template, these products

are electrophoresed on a special gel that can separate DNAs that differ in size by a single nucleotide. As each fragment moves past a laser beam, the color of the terminal base is detected and recorded. **(c)** Dideoxynucleotide structure. Because ddNTPs lack an -OH group on the 3' carbon atom of deoxyribose, DNA polymerase cannot add any nucleotides onto a chain with a dideoxynucleotide at the 3' end. **(d)** Analyzing nested sets of products on gels. **(e)** Image of a sequencing gel. Each lane displays the sequence obtained with a separate sample. **(f)** A DNA sequence trace from one gel lane. The raw data are displayed as peaks of four different colors. Here, yellow is pseudocolored as black for easier visualization. The base-calling software produces a text sequence of the newly synthesized DNA strand. It is possible to read almost 1000 bases from a single reaction.



e: © Jean Claude Révy/ISM/Phototake; f: Courtesy of Joshua J. Filter, Cornell University, Ithaca, New York

template and primer, along with a carefully calibrated mixture of eight nucleotide triphosphates (Fig. 9.7b). Four of these are the normal deoxyribonucleotide triphosphates dATP, dCTP, dGTP, and dTTP. The other four are unusual and are added at lower concentrations: They are the **dideoxyribonucleotide triphosphates** (sometimes just called **dideoxynucleotides**) ddATP, ddCTP, ddGTP, and ddTTP (Fig. 9.7c). These dideoxynucleotides lack the 3' hydroxyl group crucial for the formation of the phosphodiester bonds that extend the chain during DNA polymerization (review Fig. 6.21). Moreover, each dideoxynucleotide is labeled with a different color fluorescent dye; for example, ddATP can carry a dye that fluoresces in green, ddCTP has a purple dye, etc.).

The sequencing reaction tube contains billions of originally identical hybrid DNA molecules in which the oligonucleotide primer has hybridized to the template DNA strand at the same location. On each molecule, the primer supplies a free 3' end for DNA chain extension by DNA polymerase. The polymerase adds nucleotides to the growing strand that are complementary to those of the sample's template strand. The addition of nucleotides continues until, by chance, a dideoxynucleotide is incorporated instead of a normal nucleotide. The absence of a 3' hydroxyl group in the dideoxynucleotide prevents the DNA polymerase from forming a phosphodiester bond with any other nucleotide, ending the polymerization for that new strand of DNA (Fig. 9.7b).

When the reaction is completed, the newly synthesized strands are released from the template strands by denaturing the DNA at high temperature. The result is a nested set of fragments that all have the same 5' end (the 5' end of the primer) but different 3' ends. The length and fluorescent color of each fragment making up the set is determined by the last nucleotide incorporated; that is, the single chain-terminating dideoxynucleotide the fragment contains (Fig. 9.7b).

## The Fluorescence of DNA Fragments Reveals the Nucleotide Sequence

Biologists analyze the mixture of DNA fragments created by the sequencing reaction through polyacrylamide gel electrophoresis, under conditions that allow the separation of DNA molecules differing in length by just a single nucleotide (Fig. 9.7b, d, and e). The gel is examined by a DNA sequencing machine that has a laser to activate the dideoxynucleotide fluorescent tags and a sensitive detector that can distinguish the resultant colored fluorescence. As each DNA fragment passes under the laser, it will glow in one of the four fluorescent colors dictated by the dye attached to the dideoxynucleotide at the 3' end of the chain. Each successive fluorescent signal represents a chain that is one nucleotide longer than the previous one.

The detector transmits information about the signals to a computer, which shows them as a series of different-colored peaks (Fig. 9.7f). The computers in DNA sequencing machines also have base-calling software that interprets the peaks as specific bases and that generates a digital file, called a **read**, of the sequence of As, Cs, Gs, and Ts comprising the newly synthesized DNA. Of course, this sequence is complementary to that of the template strand under analysis.

DNA sequencing machines available since the late 1990s can determine about 700–1000 bases from any single sample. These machines can also run hundreds of samples in parallel on separate gel lanes, each recorded with a separate fluorescence detector (Fig. 9.7d and e). Thus, a single machine running for a few hours can determine hundreds of thousands of bases of DNA sequence information.

### essential concepts

- In *Sanger DNA sequencing*, the DNA molecule to be sequenced serves as a template for DNA synthesis by DNA polymerase.
- Sanger DNA sequencing requires a short oligonucleotide *primer* that *hybridizes* to the template. DNA polymerase extends the primer by adding (to its 3' end) nucleotides that are complementary to the template.
- In automated DNA sequencing, chain synthesis terminates when DNA polymerase incorporates a *dideoxynucleotide* that has a fluorescent label.
- The DNA fragments made in the polymerization reaction are separated by size on a gel, and a detector reads the color of the fluorescent tag at the 3' end of each fragment to determine the nucleotide sequence.

## 9.4 Sequencing Genomes

### learning objectives

1. Explain why overlap between individual DNA sequences is required to reconstruct the sequence of a genome.
2. Describe the differences between the hierarchical and shotgun strategies for genome sequencing.

Genomes range from the 700,000 base pairs (700 kb) in the smallest known microbial genome, to more than 3 billion base pairs (3 gigabase pairs, or 3 Gb) distributed among the 23 chromosomes of humans, to even larger genomes. **Table 9.2** gives the genome sizes of representative microbes, plants, and animals. To put these numbers in perspective, the human genome is more than 700 times larger than that of *E. coli* and 45 times smaller than the

**TABLE 9.2** Genome Comparisons

Organism		Number of Chromosomes <sup>a</sup>	Number of Genes <sup>b</sup>	Genome Size (Mb)
Type	Species			
Bacterium	<i>Escherichia coli</i>	1	~4400	4.6 <sup>d</sup>
Yeast	<i>Saccharomyces cerevisiae</i>	16	~6000	12.5
Worm	<i>Caenorhabditis elegans</i>	6	~22,000	100.3
Fly	<i>Drosophila melanogaster</i>	4	~17,000	122.7
Mustard weed	<i>Arabidopsis thaliana</i>	5	~28,000	135
Mouse	<i>Mus musculus</i>	20	~27,000	2,700
Human	<i>Homo sapiens</i>	23	~27,000	3,300
Lungfish	<i>Protopterus aethiopicus</i>	14	??	133,000
Canopy plant	<i>Paris japonica</i>	5 <sup>c</sup>	??	152,400

<sup>a</sup>Haploid chromosome complement except where indicated.

<sup>b</sup>Includes non-protein-coding genes.

<sup>c</sup>This species is an octoploid; 5 is the *basic chromosome number* (see Chapter 13).

<sup>d</sup>*E. coli* genomes vary in size; 4.6 Mb is a representative length (see Chapter 14.)

genome of the plant *Paris japonica*. Thus, the information content of a genome is not necessarily proportional to the complexity of the organism.

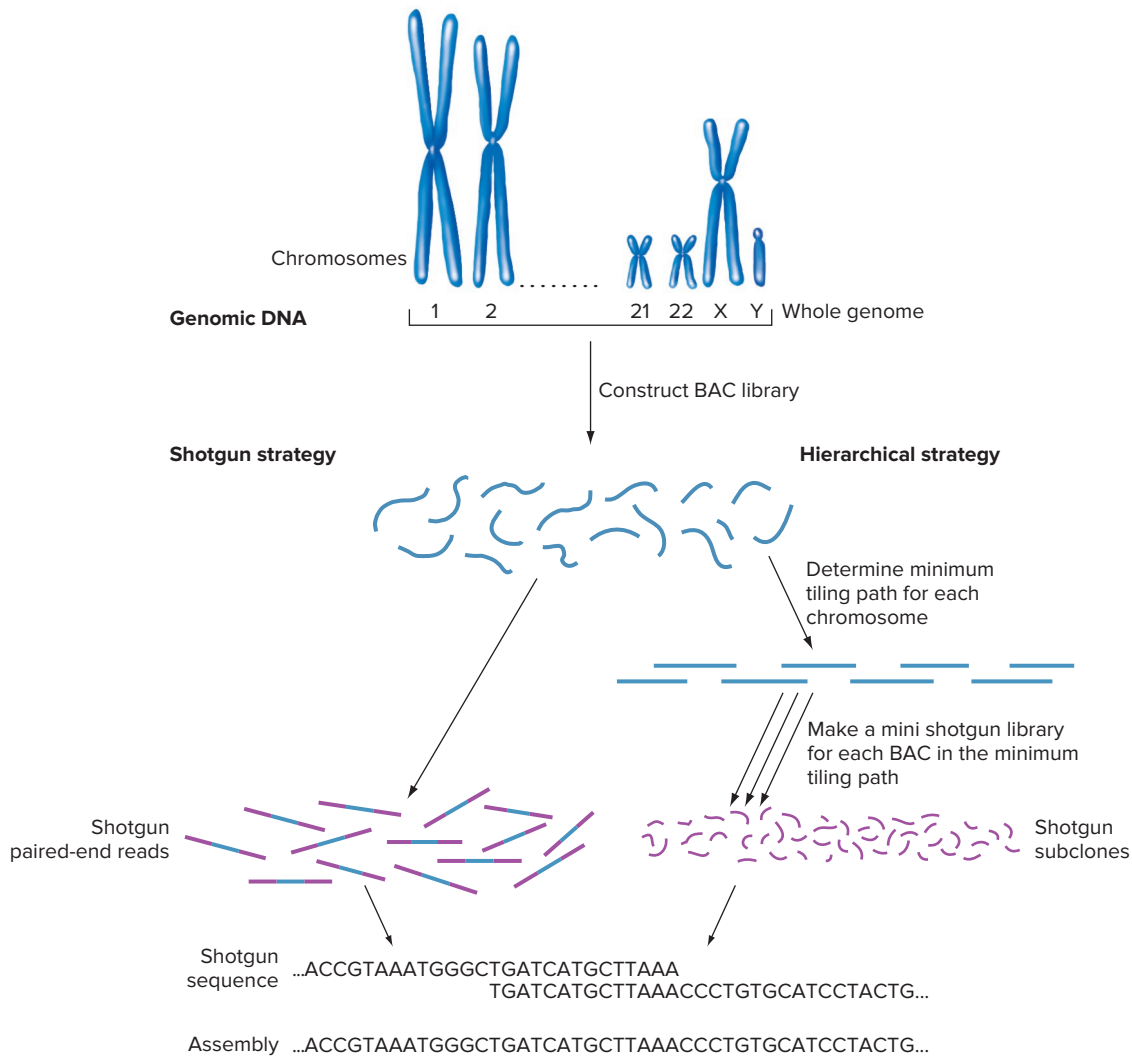
The large size of some genomes, including the human genome, presents major challenges for their ultimate characterization and analysis. If any single DNA sequencing run can yield at most 1000 bases of information, then you might think you would need to obtain at least 3 million such sequences to determine the human genome's entire sequence. In fact this is a gross underestimate, because as we discussed previously, you would really need to examine at least five times this number of clones from a genomic library to ensure just a 95% chance that each portion of the genome would be represented once. How can you do so many DNA sequencing runs? And how can you deal with the immense amount of data you would obtain so that you could somehow figure out how these millions of small, 1000-base snippets are ordered with respect to each other in the intact genome?

The basic concept behind the method now used to sequence complex genomes, called the **whole-genome shotgun strategy**, is easy to explain: Determine DNA sequences, each about 1000 bases long, from both ends (*paired ends*) of random human genomic DNA inserts of millions of individual BAC (bacterial artificial chromosome) clones from a genomic library, and then look for overlaps between the sequences so that they can be assembled to reconstruct the sequence of the entire genome (**Fig. 9.8**). (*Shotgun* refers to the fact that the clones are chosen randomly for sequencing.) Ideally, in the case of the human genome, the ultimate output would be 24 linear strings of nucleotide sequences, one for each chromosome (the autosomes and the X and the Y).

When the Human Genome Project began in 1990, whole-genome shotgun strategies were thought to be impossible. One problem was recognized very early: As will be discussed later, genomes contain many kinds of repetitive DNA sequences, each of which can be located in many positions scattered throughout the genome. Many repeated sequences are longer than a typical sequence read of 1000 bp. This fact makes it impossible to assemble the genome from random reads. The reason is that unique sequences on one side of a long repeat from one particular genomic location cannot be present in the same read as unique sequences from the other side of the repeat (**Fig. 9.9a**). Scientists eventually realized that *paired-end sequencing* of random clones would make the whole-genome shotgun strategy possible. We will explain this method, but first we describe an alternative strategy researchers used before this conceptual breakthrough.

To get around the assembly problem presented by the existence of long repeats, the first genome scientists tried a divide-and-conquer method called a **hierarchical strategy** (**Fig. 9.8**). They first separated the genome into large chunks by cloning 200–300 kb fragments in BAC vectors, and then they applied strategies (not discussed here) to determine the order of the inserts with respect to each other in the original genome. Note that the genomic DNA fragments were generated using a method (such as sonication) that cleaves different copies of the genome at different locations, resulting in overlapping fragments (**Fig. 9.8**). These methods allowed the researchers to determine the smallest set of BAC clones with the least amount of overlap that could cover the entire genome (the so-called *minimal tiling path*). The scientists then determined the DNA sequence of the entire insert in each BAC clone of the minimal tiling path so as to reconstruct the genome

**Figure 9.8 Strategies for genome sequencing.** In the whole-genome shotgun (Celera) approach, BAC libraries are made from fragmented genomic DNA. Both ends of millions of clones are directly sequenced; a computer looks for overlaps between these sequences and assembles the sequence of the genome. The more systematic but less efficient hierarchical approach employs intermediate steps in which the BAC clones of a genomic library (with inserts of 200–300 kb) are characterized so as to determine the *minimum tiling path*. Each BAC clone in the path is fragmented into smaller pieces to make a mini-shotgun library, and the DNAs in the mini-libraries are sequenced. Computers reassemble the sequences of each BAC clone, and then look for the overlaps between BAC clones to reassemble the complete genome sequence.



(Fig. 9.8). As most BAC clones would contain only a single copy of a particular repetitive element, assembling the genome one BAC clone at a time avoided the problem illustrated in Fig. 9.9a.

Although the hierarchical approach was ultimately successful, a private company called Celera astonished the scientific community by simultaneously undertaking and completing their own separate effort to sequence the human genome using the whole-genome shotgun strategy thought by many at the time to be hopeless. As mentioned earlier, the key to Celera's success was the idea of performing **paired-end sequencing**; that is, they obtained two sequence reads from each BAC clone, one from each end of the insert (Fig. 9.9b and c). Paired-end sequencing

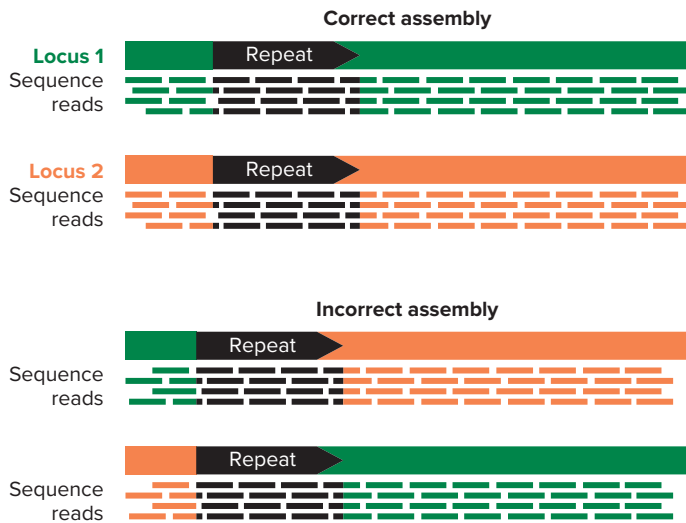
provided the Celera scientists with the information that the two sequences from each BAC clone must have originated from the same region of the genome and must have been 200–300 kb apart. As shown in Fig. 9.9b, this information allowed the scientists to properly align the unique sequences that flank repeat elements.

The whole-genome shotgun strategy has two important advantages over the hierarchical method. First, the shotgun approach does not require time-consuming mapping of the BAC clones to generate the tiling path. Second, the shotgun procedure can be highly automated. Celera invested in a huge facility containing hundreds of DNA sequencing machines fed by robots that first prepared DNA from the clones of genomic libraries, placed these

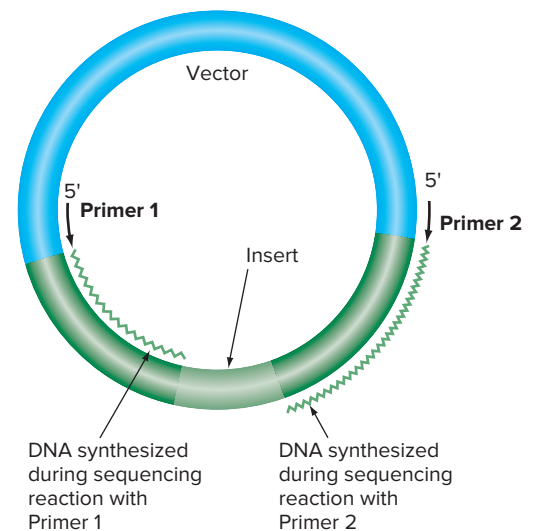


**Figure 9.9 Whole-genome shotgun sequencing.** (a) Repeated elements longer than a sequence read would prevent the assembly of shotgun sequences of the human genome because it is impossible to know which unique sequences (*green* or *orange*) flanking different copies of a repeat belong together. (b) The paired-end sequencing method. A recombinant BAC clone is melted into single strands and hybridized in one reaction with Primer 1 and in a second independent reaction with Primer 2. These primers correspond to the sequence of the vector just flanking the human DNA insert on either side. The two primers hybridize with opposite strands of the recombinant DNA and are oriented such that DNA polymerase will synthesize DNA corresponding to either end of the human DNA insert. [Note, as shown in part (c), that the *light green* portion of the insert is in fact much longer than the *dark green* portions sequenced.] (c) Paired-end sequencing allows correct assembly of genomes containing repeats because paired reads will include unique sequences on both sides of a repeat. In other words, the unique sequence reads (*green* or *orange* boxes at right) align, so the repeat-containing sequence reads (*green and black* or *orange and black* boxes at left) must align also.

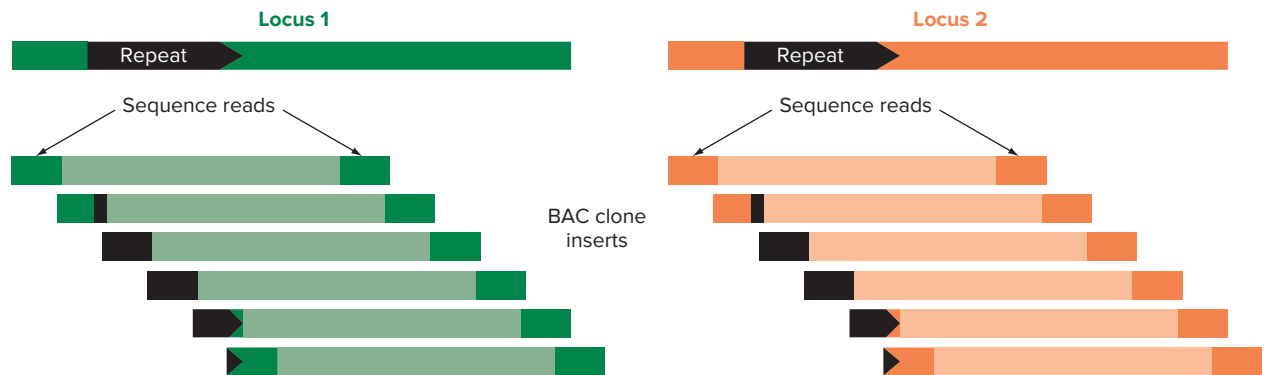
(a) Repeats prevent correct assembly of single shotgun sequence reads



(b) Paired-end sequence reads from a BAC done



(c) Paired-end sequences direct correct assembly of unique sequences flanking repeats



DNAs into sequencing reactions, and then loaded the reactions into the sequencing machines. This automation allowed Celera to obtain relatively cheaply the millions of DNA sequence reads required to provide about 10-fold genomic equivalent coverage. The DNA sequencing machines fed their data into a centralized supercomputer, whose complex software then assembled all these sequences into the chromosomal strings. The whole-genome shotgun approach had such large relative efficiencies that variants of this method have become the standard way to sequence genomes.

### essential concepts

- The *whole-genome shotgun sequencing* strategy involves sequencing inserts of millions of clones selected randomly from libraries constructed with mechanically sheared genomic DNA to ensure overlap between fragments.
- Sequencing both ends of DNA inserts (*paired-end sequencing*) provides information useful for genome assembly.

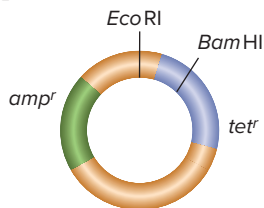
## WHAT'S NEXT

The Human Genome Project did not end with the determination of 3 billion base pairs of DNA sequence. An essential part of the project's task was to make sense of this vast amount of information. Where in all of these As, Cs, Gs, and Ts are the genes? What do the sequences of genes predict about the kinds of proteins and RNAs the genes encode and the possible

functions of these proteins and RNAs? What kinds of DNA sequences make up other important features of chromosomes such as centromeres and telomeres? In the next chapter, we explain how scientists identified functional elements of the genome and how their findings revealed, at the level of DNA sequence, the architecture of the human genome.

## SOLVED PROBLEMS

- I. The following map of the plasmid cloning vector pBR322 shows the locations of the ampicillin (*amp*) and tetracycline (*tet*) resistance genes as well as two unique restriction enzyme recognition sites, one for *EcoRI* and one for *Bam*HI. You digested this plasmid vector with both *EcoRI* and *Bam*HI enzymes and purified the large *EcoRI*–*Bam*HI vector fragment. You also digested human genomic DNA that you want to insert into the vector with both *EcoRI* and *Bam*HI. After mixing the plasmid vector and the human genomic fragments together and ligating, you transformed an ampicillin-sensitive strain of *E. coli* and selected for ampicillin-resistant colonies.



- If you test all of your selected ampicillin-resistant transformants for tetracycline resistance, what result do you expect, and why?
- Why is it important that the *EcoRI* site *not* be located in the gene conferring ampicillin resistance?
- Diagram the positions and orientations of two oligonucleotide primers that you could use to sequence the two ends of the human DNA insert found in any recombinant DNA molecule made by this method.
- Why would a library made in this fashion represent less than one genome equivalent?

## Answer

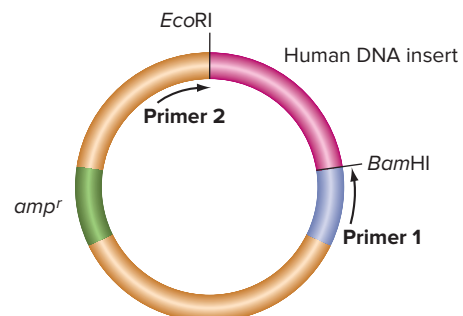
This problem requires an understanding of vectors and the process of combining DNAs using sticky ends generated by restriction enzymes.

- The plasmid must be circular to replicate in *E. coli*, and in this case, a circular molecule will be formed

DNA: © Design Pics/Bilderbuch RF

only if the insert fragment joins with the cut vector DNA. The cut vector will not be able to re-ligate without an inserted fragment because the *Bam*HI and *EcoRI* sticky ends are not complementary and cannot base pair. All ampicillin-resistant colonies therefore contain a *Bam*HI–*EcoRI* fragment of human DNA ligated to the *Bam*HI–*EcoRI* sites of the vector. Fragments cloned at the *Bam*HI–*EcoRI* sites interrupt and therefore inactivate the tetracycline resistance gene. **All ampicillin-resistant clones will be tetracycline sensitive.**

- If the gene for ampicillin resistance contained an *EcoRI* site while the tetracycline resistance gene had a *Bam*HI site, the cloning process would destroy the activity of both genes. **Thus, you would be unable to select bacterial cells transformed with recombinant DNA molecules.**
- The primers would have to flank the human genomic DNA inserts (*pink*) as shown in the following diagram.



- By chance, some regions of the human genome would have two *EcoRI* sites in a row without a *Bam*HI site in between. Other regions would have two *Bam*HI sites in a row without an *EcoRI* site in the middle. When such regions are cut with both enzymes, the resultant fragments would not have the two different kinds of sticky ends required for insertion into the vector cut with both enzymes. **Thus, many regions of the human genome could not be cloned into the vector by this procedure.**

II. In the course of sequencing a genome, a computer is trying to assemble the following six DNA sequences into *contigs* (that is, stretches of contiguous sequences that can be obtained from overlapping clones):

```

5' AGCAAATTACAGCAATATGAAGAGATC 3'
5' AAAATGCCCTAAAGGAAATGAGATTTT 3'
5' TGATCTCTTCATATTGCTGTAATTTGC 3'
5' TCCTTTTAAAAATCTCATTTCCTTTAG 3'
5' TACAGCAATATGAAGAGATCATAACAGT 3'
5' AAATGCCCTAAAGGAAATGAGATTTT 3'

```

- How many contigs are represented by this set of DNA sequences, and what is the sequence of each contig?
- Some of these sequences are complementary to each other in the region of overlap, while other sequences overlap but represent the same DNA strand. How is this possible?
- If these sequences are all derived at random from the human genome, why would you actually expect them not to overlap with each other?
- If you had enough sequence reads to cover all the base pairs in the human genome, how many contigs would there be?

### Answer

- a. Two contigs exist.

Sequences 1, 3, 5:

```

5' AGCAAATTACAGCAATATGAAGAGATCATAACAGT 3'
3' TCGTTTAAATGTCGTTATACTTCTCTAGTATGTCA 3'

```

Sequences 2, 4, 6:

```

5' TCCTTTTAAAAATCTCATTTCCTTTAGGGCATTTT 3'
3' AGGAAAATTTTATAGAGTAAAGGAAATCCCGTAAAA 5'

```

The orientation in which these fragments are written does not matter.

- You are sequencing different molecules of DNA in clones that are overlapping. The clones are ligated into a vector in random orientations, and it's the orientation that determines which DNA strand is used as a template. Therefore, some of the sequences read the same strand but start in different places, and others read the complementary strand.
- If you only had six short sequences from the entire 3 billion bp human genome, the chances would be vanishingly small that these few sequences all come at random from the same small region of the genome. Thus, six random sequences of the human genome should not overlap. Clearly, these sequences were not derived at random but were instead selected. (For example, these could be sequences all derived from a mini-library made from a particular human DNA insert in a BAC vector.)
- Each human chromosome is a contig. A human male genome sequence would have 24 contigs, and a female genome 23 contigs.

### Vocabulary

1. Match each of the terms in the left column to the best-fitting phrase from the right column.

- |                        |  |
|------------------------|--|
| a. oligonucleotide     | 1. gene in a vector that enables isolation of transformants  |
| b. vector              | 2. a collection of the DNA fragments of a given species, inserted into a vector                                    |
| c. sticky ends         | 3. synthetic DNA element in a cloning vector with unique restriction sites used for insertion of foreign DNA       |
| d. recombinant DNA     | 4. stable binding of single-stranded DNA molecules to each other   |
| e. ddNTPs              | 5. method for separating DNA molecules by size   |
| f. genomic library     | 6. oligonucleotide extended by DNA polymerase during replication   |
| g. genomic equivalent  | 7. contains genetic material from two different organisms  |
| h. gel electrophoresis | 8. the number of DNA fragments sufficient in aggregate length to contain the entire genome of a specified organism |
| i. selectable marker   | 9. short single-stranded sequences found at the ends of many restriction fragments                                 |
| j. hybridization       | 10. a short DNA fragment that can be synthesized by a machine  |
| k. primer              | 11. DNA chain-terminating subunits   |
| l. polylinker          | 12. a DNA molecule used for transporting, replicating, and purifying a DNA fragment                                |

When solving the problems in this chapter, unless instructed otherwise, make the simplifying assumptions that base pair sequences are random and that the number of A–T and G–C base pairs are equivalent.

### Section 9.1

- For each of the restriction enzymes listed below:
  - Approximately how many restriction fragments would result from digestion of the human genome ( $3 \times 10^9$  bases) with the enzyme?
  - Estimate the average size of the pieces of the human genome produced by digestion with the enzyme.
  - State whether the fragments of human DNA produced by digestion with the given restriction enzyme would have sticky ends with a 5' overhang, sticky ends with a 3' overhang, or blunt ends.
  - If the enzyme produces sticky ends, would all the overhangs on all the ends produced on all fragments of the human genome with that enzyme be identical, or not? (The recognition sequence on one strand for each enzyme is given in parentheses, with the 5' end written at the left. N means any of the four nucleotides; R is

any purine—that is, A or G; and Y is any pyrimidine—that is, C or T. ^ marks the site of cleavage.)

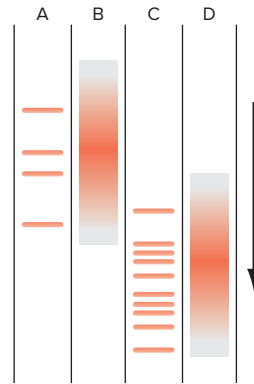
- Sau3A* (^GATC)
- BamHI* (G^GATCC)
- HpaII* (C^CGG)
- SphI* (GCATG^C)
- NaeI* (GCC^GGC)
- BanI* (G^GYRCC)
- BstYI* (R^GATCY)
- BsI* (CCNNNNN^NNGG)
- SbfI* (CCTGCA^GG)

- The calculations of the average restriction fragment size in Fig. 9.2 assume that DNA is composed equally of the four possible nucleotides. However, many genomes are somewhat enriched for certain nucleotides relative to others. As an example, the human genome is 29.6% A, 29.6% T, 20.4% C and 20.4% G. With this more accurate information in hand, re-estimate the average sizes of the pieces created by cleaving the human genome with enzymes (a–i) in the preceding Problem 9.2.
- The DNA molecule whose entire sequence follows is digested to completion with the enzyme *EcoRI* (5' G^AATTC 3'). How many molecules of DNA would result from this reaction? Write out the entire sequence(s) of the resultant DNA molecule(s), indicating all relevant 5'-to-3' polarities. What about this problem appears unusual (though by no means impossible) in relationship to DNA made of random nucleotide sequences?

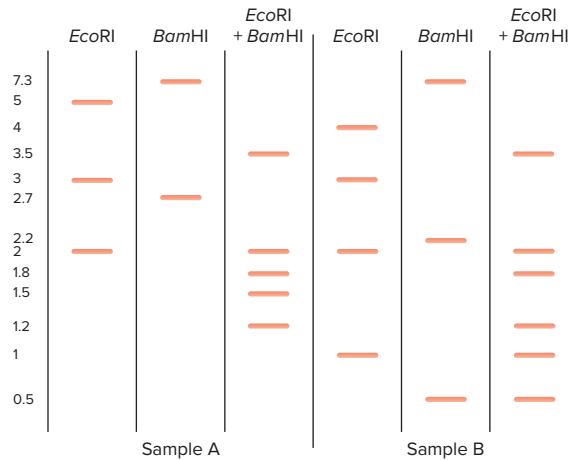
5' AGATGAATTTCGCTGAAGAACCAAGAATTCGATT 3'  
 3' TCTACTTAAGCGACTTCTTGGTTCTTAAGCTAA 5'

- Why do longer DNA molecules move more slowly than shorter ones during electrophoresis?
- Agarose gels with different average pore sizes are needed to separate DNA molecules of different size classes. For example, optimal separation of 1100 bp and 1200 bp fragments would require a gel with a larger average pore size than optimal separation of 8500 bp and 8600 bp fragments. How do you think that scientists prepare gels of different average pore sizes? (*Hint:* Agarose gels are made in a manner similar to gelatin desserts such as JELL-O.)
- The following picture shows the ethidium bromide–stained bands revealed by gel electrophoresis of two different DNA samples digested with two different restriction enzymes. One of the DNAs is human genomic DNA, the other is the small genome of a bacteriophage (bacterial virus) that infects *E. coli* cells. One of the restriction enzymes is *EcoRI* (5' G^AATTC 3'); the other is *HpaII* (5' C^CGG 3'). For each of the four lanes on the gel (A–D), identify which of the two DNA

samples was analyzed and which of the two restriction enzymes was used to digest that DNA. The arrow represents the direction of electrophoresis.



- The linear bacteriophage  $\lambda$  genomic DNA has at each end a single-strand extension of 20 bases. (These are sticky ends but are not, in this case, produced by restriction enzyme digestion.) These sticky ends can be ligated to form a circular piece of  $\lambda$  DNA. In a series of separate tubes, either the linear or circular forms of the DNA are digested to completion with *EcoRI*, *BamHI*, or a mixture of the two enzymes. The results are shown here.



- Which of the samples (A or B) represents the circular form of the DNA molecule? How do you know?
  - What is the total length of the linear form of the  $\lambda$  DNA molecule?
  - What is the total length of the circular form of the  $\lambda$  DNA molecule?
  - Draw diagrams of the circular and linear  $\lambda$  DNA molecules, showing the locations of the *EcoRI* and *BamHI* sites.
- Consider a *partial restriction digestion*, in which genomic DNA is exposed to a small, limiting amount of a restriction enzyme for a very short period of time.
    - Would the resultant fragments be longer or shorter or the same size as those produced by a complete digestion?

- b. If you prepared genomic DNA from a tissue sample containing millions of cells, would the fragments produced by partial digestion of DNA from all of these cells be the same or different?
10. The text stated that molecular biologists have developed elegant techniques that can convert any type of DNA end into any other type of DNA end. In this problem, consider genomic DNA that is broken by mechanical shearing into random pieces. Some of the ends of these pieces are blunt, some have 5'-overhangs, and others have 3'-overhangs.
- Must the two ends of any one genomic DNA fragment be of the same type?
  - Explain why the ends with 5' or 3' overhangs are not sticky.
  - Researchers can convert ends with overhangs into blunt ends using either DNA polymerase (plus the four dNTPs), or nuclease S1, which degrades single-stranded regions of DNA but not double-stranded regions. Which kinds of ends with overhangs (5' or 3') could be converted into blunt ends using DNA polymerase? With S1 nuclease?

## Section 9.2

- What is the purpose of molecular cloning?
  - What purpose do selectable markers serve in vectors?
  - What is the purpose of the origin of replication in a plasmid vector?
  - Why do cloning vectors have polylinkers?
- Which of the enzymes from the following list would you need to make a recombinant DNA molecule? What is the function of those enzyme(s) in the process?
  - DNA polymerase
  - RNA polymerase
  - A restriction enzyme
  - DNA ligase
  - An aminoacyl-tRNA synthetase
  - Peptidyl transferase
  - Reverse transcriptase
- Is it possible that two different restriction enzymes could cut the human genome into exactly the same number of fragments and with exactly the same distribution of fragment sizes, yet the ends produced by the two enzymes could not be joined together by DNA ligase? Explain.
- A plasmid vector pBS281 is cleaved by the enzyme *Bam*HI (5' G<sup>^</sup>GATCC 3'), which recognizes only one site in the DNA molecule. Human DNA is digested with the enzyme *Mbo*I (5' <sup>^</sup>GATC 3'), which recognizes many sites in human DNA. These two digested DNAs are now ligated together. Consider only those molecules in which the pBS281 DNA has been joined with a fragment of human DNA. Answer the following questions concerning the junction between the two different kinds of DNA.
  - What proportion of the junctions between pBS281 and all possible human DNA fragments can be cleaved with *Mbo*I?
  - What proportion of the junctions between pBS281 and all possible human DNA fragments can be cleaved with *Bam*HI?
  - What proportion of the junctions between pBS281 and all possible human DNA fragments can be cleaved with *Xor*II (5' C<sup>^</sup>GATCG 3')?
  - What proportion of the junctions between pBS281 and all possible human DNA fragments can be cleaved with *Bst*YI (5' R<sup>^</sup>GATCY 3')? (R and Y stand for purine and pyrimidine, respectively.)
  - What proportion of all possible junctions that can be cleaved with *Bam*HI will result from cases in which the cleavage site in human DNA was not a *Bam*HI site in the human chromosome?
- A recombinant DNA molecule is constructed using a plasmid vector called pMBG36 that is 4271 bp long. The pMBG36 plasmid contains a so-called *polylinker* that has a single site for each of several restriction enzymes, including *Bam*HI (5' G<sup>^</sup>GATCC 3') and *Eco*RI (5' G<sup>^</sup>AATTC 3'). The sequence of the polylinker region of pMBG36 is shown below; the dots indicate the large majority of the vector that is not shown. You now cut pMBG36 with *Eco*RI and insert into it a fragment of the DNA previously shown in Problem 4 that is also cut with *Eco*RI.
 

```
5'...CGGATCCCCTAAGATGAATTCCGCGCGCATCGGC...3'
3'...GCCTAGGGGATTCTACTTAAGGCGCGCTAGCCG...5'
```

  - Write out as much of the DNA sequence of the resultant recombinant DNA molecule as is possible. Two answers are possible; you need to show only one.
  - Why are there two possible answers to part (a)?
  - How many recognition sites for *Bam*HI will be found in the recombinant DNA molecule shown in your answer to part (a)?
  - If you cut this recombinant DNA molecule with *Bam*HI and run the digest on a gel, how many bands would you see and how large would they be?
  - How many recognition sites for *Eco*RI will be found in the recombinant DNA molecule shown in your answer to part (a)?
  - If you cut this recombinant DNA molecule with *Eco*RI and run the digest on a gel, how many bands would you see and how large would they be?

16. Suppose you are using a plasmid cloning vector that has no *EcoRI* sites (5' G<sup>A</sup>AATTC 3') in its polylinker because the particular drug resistance gene your vector contains has an *EcoRI* site within it.

- a. How could you use the following two oligonucleotides (and ligase enzyme) to ligate an insert that is an *EcoRI* fragment into the *Bam*HI site (5' G<sup>A</sup>GATCC 3') in the polylinker of your vector?



- b. How many *EcoRI* sites will the recombinant DNA contain? How many *Bam*HI sites?
- c. In part (a), you used the two oligonucleotides to make a so-called *adapter*. Adapters can also be used to ligate blunt-ended inserts into vectors cut with sticky-ended enzymes. Design an adapter that would allow you to ligate blunt-ended inserts into the *Bam*HI site of your vector's polylinker. (Note: Two blunt-ended DNA fragments can be ligated together, although the reaction is much less efficient than sticky-end ligation.)
17. As a molecular biologist and horticulturist specializing in snapdragons, you have decided that you need to make a genomic library to characterize the flower color genes of snapdragons.
- a. How many genomic equivalents would you like to have represented in your library to be 95% confident of having a clone containing each gene in your library?
- b. How do you determine the number of independent clones that should be screened to guarantee this number of genomic equivalents?
18. Suppose you are constructing a human genomic library in BAC vectors where the human DNA fragments are on average 100,000 bp.

- a. What is the minimum number of different recombinant BACs you need to construct in order to have a greater than zero chance of having a complete library—meaning one in which the entire genome is represented?

The simple statistical equation that follows allows you to determine the size that a genomic library needs to be (that is, the number of independent recombinant clones you need to make) for a given likelihood that the entire genome is represented in the library.

$$N = \frac{\ln(1 - P)}{\ln(1 - f)}$$

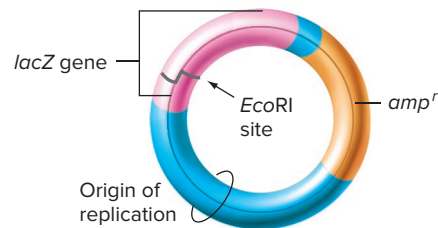
In the equation,  $N$  is the number of independent recombinant clones;  $P$  is the probability that any particular part of the genome is represented at least one time;  $f$  is the

fraction of the genome in a single recombinant clone. (Note:  $\ln$  is the natural log, sometimes written as  $\log_e$ .)

- b. Calculate  $f$  for the genomic library described in part (a).
- c. How many different recombinant BAC clones would you need to have a 99% chance that a specific 100,000 bp region of the genome is represented? How many clones for a 99.9% chance?
- d. How many *genomic equivalents* correspond to each of your answers in part (c)?
- e. Suppose that after you ligated the human DNA inserts with the BAC vectors and transformed *E. coli* with the mixture, you find that you have only 30,000 drug-resistant colonies transformed with recombinant plasmids. What is the chance that any specific 100,000 bp region of the genome is represented in a recombinant plasmid?
- f. If you want to construct a complete human genomic library that contains the smallest number of independent recombinant clones possible, what is the key variable that you should adjust?

One difficulty in molecular cloning using plasmid vectors is that the restriction enzyme-digested vector can be resealed by DNA ligase without an insert of genomic DNA. The next two problems investigate methods to deal with this issue.

19. The *lacZ* gene from *E. coli* encodes the enzyme  $\beta$ -galactosidase, which can catalyze the conversion of a colorless compound called X-gal into a blue product. Molecular biologists have taken advantage of this property by constructing plasmid vectors that contain the *lacZ* gene with an *EcoRI* site in its middle (see figure that follows). After cutting this vector with the *EcoRI* enzyme, scientists ligate it together with *EcoRI*-digested human genomic DNA, transform the resultant molecules into ampicillin-sensitive *E. coli* cells, and plate these cells on petri plates containing ampicillin and X-gal. Some of the colonies growing on this plate are white in color, while others are blue. Why?



20. Your undergraduate research advisor has assigned you a task: Insert an *EcoRI*-digested fragment of frog DNA into the vector shown in Problem 19. Your advisor suggests that after you digest your plasmid with *EcoRI*, you should treat the plasmid with the enzyme alkaline phosphatase. This enzyme







chapter **10**Genome  
Annotation

**DETERMINATION OF THE 3 BILLION BASE PAIR SEQUENCE** of the human genome, although an amazing achievement, was only the first step of the Human Genome Project. The nucleotide sequence by itself does not answer several key questions: Where are the genes? How many are there? What are their products? What lies in the genome other than genes? How are the genes and other genomic elements organized along the chromosomes? Without these answers, we cannot begin to understand how the human genotype (that is, the sequence) determines the complexity of the human phenotype.

In this chapter, we describe the **annotation** of the human genome; that is, the process of parsing out which sequences of DNA do which tasks. The process of annotation requires the compilation of data from diverse methods of investigation, including a variety of molecular experiments as well as complex computer algorithms to analyze the vast amount of data obtained. One lesson of this discussion is that the genomes of species other than humans provide important clues to the annotation of the human genome. We further describe some of the key findings from the Human Genome Project so that you can picture in a general fashion how these 3 billion nucleotides are organized.

A key feature of this chapter is an introduction to Internet-based resources, particularly a large database based at the National Center for Biotechnology Information (NCBI) at the U.S. National Institutes of Health, that you yourself can use to explore the human genome and other sequenced genomes. The chapter concludes with a comprehensive example illustrating how the sequence of the human genome has helped us to understand the nature of genetic diseases called *hemoglobinopathies* that disrupt our blood's ability to carry oxygen.

*Once the sequence of a genome has been determined, researchers need to determine where functional elements such as genes reside within these billions of base pairs.*

© fredex/Shutterstock RF

**chapter outline**

- 10.1 Finding the Genes in Genomes
- 10.2 Genome Architecture and Evolution
- 10.3 Bioinformatics: Information Technology and Genomes
- 10.4 A Comprehensive Example: The Hemoglobin Genes

## 10.1 Finding the Genes in Genomes

### learning objectives

1. Explain why a long open reading frame suggests the existence of a protein-coding exon.
2. Describe how scientists predict the location of genes by identifying sequences conserved in the genomes of widely divergent species.
3. Discuss the use of reverse transcriptase in the construction of a cDNA library.
4. Compare the information that can be obtained from genomic and cDNA libraries.

Genes are the key functional elements of genomes. In this section, we focus on methods to locate genes within genomic DNA sequences. You will see that information useful for the annotation of the genes within the human genome can be found in the sequence of the genome itself, the sequences of the genomes of species other than humans, and from the characterization of RNA molecules in human cells. These methods have successfully located and characterized more than 27,000 genes in the human genome, but in spite of all of these efforts, the task is still incomplete; some genes undoubtedly remain to be found.

### Open Reading Frames (ORFs) Help Locate Protein-Coding Genes

One way to look specifically for regions that might correspond to the exons of protein-coding genes is to scan genomic DNA sequences for long **open reading frames (ORFs)**; that is, stretches of nucleotides that have a reading frame of triplets uninterrupted by a stop codon. As you remember from Chapter 8's discussion of the genetic code, the four nucleotides can be arranged into  $4^3 = 64$  possible triplets, of which three (TAA, TAG, and TGA written as DNA) signify stop. Thus, as a very rough estimate, if you looked at any random sequence of DNA starting at any one nucleotide, you would on average run into a stop codon after about  $64/3 \approx 21$  triplets. If that nucleotide begins a reading

frame that continues without a stop for significantly more than 21 triplets, there is a good chance that the DNA in this region is not a random set of nucleotides, but instead actually encodes amino acids within a protein (**Fig. 10.1**).

This method is useful but far from foolproof. Genomes are so large that regions that do not correspond to genes might rarely contain a long ORF by chance. On the other hand, because many genes in higher eukaryotes are interrupted by introns, some protein-coding exons are so small that they would not be identified as ORFs unless other information was available.

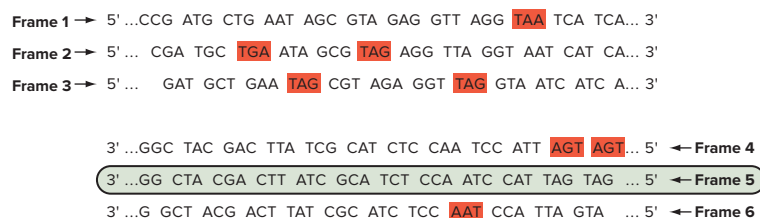
One type of additional information that could potentially aid computer programs in identifying genes is the fact that the splice acceptor and splice donor sites at intron/exon boundaries are composed of characteristic consensus sequences (review Fig. 8.15). Genome analysis programs can thus connect potential exons together and see if a long ORF suggestive of a gene would result.

### Whole-Genome Comparisons Distinguish Genomic Elements Conserved by Natural Selection

The whole-genome shotgun approach to the sequencing of genomes described in Chapter 9 has been so successful that scientists have already deciphered the genomes of thousands of different species. Researchers can exploit this tremendous amount of information to look for regions of DNA that are similar in diverse organisms. Such regions usually, though not always, correspond to genes.

The justification for comparing genomes goes all the way back to Charles Darwin. Nearly a century before the DNA double helix was discovered, he proposed the evolution of species from now-extinct ancestors by a process of *descent with modification*. We now know that the actual entity undergoing descent with modification is the DNA sequence that defines an organism's genome. The modifications are random mutations that occur in DNA. **Natural selection** is the process whereby mutations that confer an advantage to the individuals carrying them will spread throughout a population, while deleterious mutations will disappear. The challenge is to trace such molecular evolution at the DNA level.

**Figure 10.1** Open reading frames (ORFs). Any sequence of DNA can be read in any of six different reading frames (three from one strand, three from the other strand). Reading frames uninterrupted by stop codons (red) are ORFs. A long ORF suggests that the region may be part of a protein-coding exon. In this example, only one reading frame (Frame 5) is open.



### Finding conserved DNA sequences

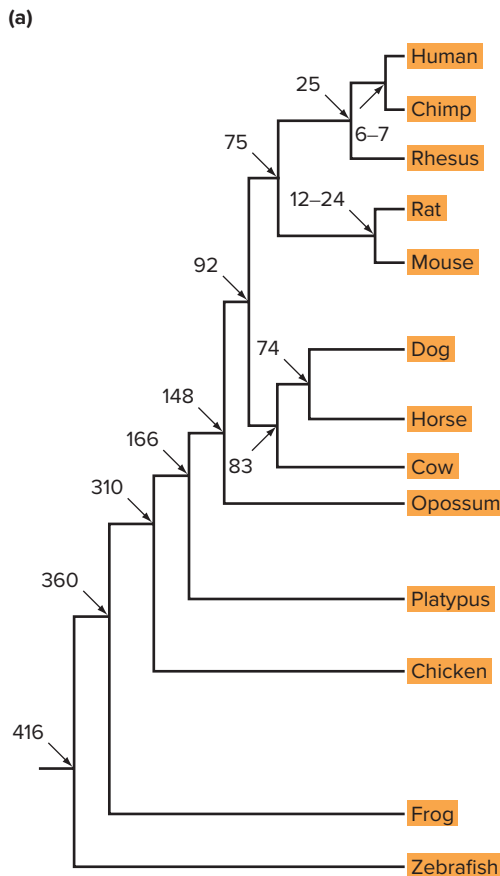
How can you tell whether DNA sequences from two sources are similar by chance or instead by common origin? As an example of a null hypothesis, consider a specific, but random, 50 bp sequence and calculate the probability that an independently derived DNA segment could be 100% identical, just by chance. The probability of the occurrence of any DNA sequence of length  $n$  is obtained simply by raising 0.25 (the chance occurrence of the same base at a particular position) to the 50th power (the number of independent chance events required):  $(0.25)^{50} = 8 \times 10^{-31}$ . This probability is very close to zero, which negates the null hypothesis and tells us that two perfectly matched 50 bp DNA sequences found in nature are almost certainly derived from the same ancestral sequence, rather than by chance.

A segment of DNA is said to be a **homolog** of a DNA segment in another species when the two show evidence of derivation from the same DNA sequence in a common ancestor. For perfectly matched sequences ~50 bp in length or longer, the evidence is clear. But evidence for homology of imperfectly matched DNA regions requires a more sophisticated statistical analysis, a task that is readily performed by specialized bioinformatics programs. When homologs of a DNA sequence are found in many different species, the sequence is said to be **conserved**.

### The landscape of DNA sequence conservation

A traditional *phylogenetic tree*, like the one shown in **Fig. 10.2a**, was made by comparing genomic DNA sequences. The tree depicts the relatedness of multiple species to each other, with branch points that represent a series of nested common ancestors. When the human genome is compared as a whole with other representative vertebrate species, the percentage of sequence conservation is very high for chimps and monkeys, but it decreases as the elapsed time to a common ancestor increases (**Fig. 10.2b**). At a distance of over 400 million years, the fish genome contains only 2% of the DNA sequences present in the human genome. In contrast, when comparisons are restricted to human protein-coding sequences, conservation levels remain high—at more than 82%—throughout vertebrate evolution.

Mutations that disrupt the function of functional DNA sequences such as protein-coding regions may lessen the evolutionary fitness of the organism. As a result, functionally important sequences evolve more slowly than nonfunctional sequences, which do not contribute to phenotype. Unconstrained divergence of nonfunctional sequences would eventually eliminate all evidence of common ancestry. Thus, whole-genome comparisons can distinguish functional and nonfunctional DNA sequences by the degree of sequence conservation.

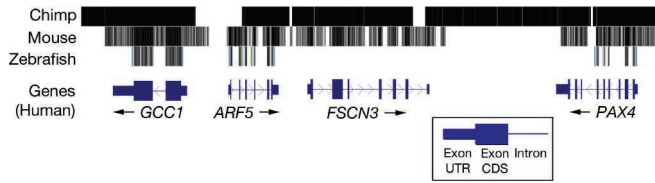


**Figure 10.2** Species relatedness and genome conservation between *H. sapiens* and other vertebrates.

(a) A phylogenetic tree showing branch points at which organisms diverged; the number at each branch point represents millions of years before the present. (b) Relatedness of the *H. sapiens* genome to that of other vertebrates. Column 1 shows the proportion of the complete human genome sequence found in the species being compared; column 2 indicates the proportions of human protein-coding sequences found in each vertebrate genome.

Scientific name	Common name	1	2
<i>Homo sapiens</i>	Human	100%	100%
<i>Pan troglodytes</i>	Chimp	93.9%	96.58%
<i>Macaca mulatta</i>	Rhesus	85.1%	96.31%
<i>Rattus norvegicus</i>	Rat	35.7%	94.47%
<i>Mus musculus</i>	Mouse	37.6%	95.36%
<i>Canis familiaris</i>	Dog	55.4%	95.18%
<i>Equus caballus</i>	Horse	58.8%	92.70%
<i>Bos taurus</i>	Cow	48.2%	94.78%
<i>Monodelphis domestica</i>	Opossum	11.1%	91.43%
<i>Ornithorhynchus anatinus</i>	Platypus	8.2%	86.43%
<i>Gallus gallus</i>	Chicken	3.8%	88.61%
<i>Xenopus tropicalis</i>	Frog	2.6%	87.44%
<i>Danio rerio</i>	Zebrafish	2.0%	82.38%

**Figure 10.3** Homology map for a 100 kb region of the human genome. Regions in *black* are homologous between the human genome and the genome of the indicated species. Most DNA sequences conserved between humans and zebrafish are found in protein-coding exons. Some sequences outside of the exons are also conserved evolutionarily, suggesting that they may play functional roles that are currently unknown. (UTR: untranslated region; CDS: protein-coding sequence).



With a computerized genome visualization tool, it becomes possible to explore DNA sequence conservation directly along the genome, as well as across evolutionary time. An example of cross-species homology analysis is shown in **Fig. 10.3** for a 100 kb region containing four genes. The bottom row of the figure displays the locations and exon/intron structures of the four genes in the human genome. Above this row are homology maps for three representative vertebrate species; highly conserved DNA sequences are indicated with dark lines or blocks.

As anticipated from the close relationship between human and chimpanzee species, nearly complete conservation of human sequences exists across the entire region in a chimp genome. In other mammals, represented here by the mouse, conservation is also apparent across the entire region, but the pattern is choppy, indicating small regions of conservation interspersed with small, nonconserved regions.

As we move farther across the phylogenetic landscape to fish, we can distinguish sequences subject to evolutionary constraints more clearly from those that are not. Note in particular that large parts of the coding regions of three of the four genes are highly conserved in all the species examined (**Fig. 10.3**). This conservation suggests that the protein products of the three genes are crucial to the survival of all vertebrates. However, a homolog of the fourth gene is not found in zebrafish, indicating that its function is dispensable to fish. Regions of homology between the human and mouse or zebrafish genomes are much less frequent in introns, in the noncoding parts of exons (corresponding to the 5' and 3' UTRs of the genes), and in the spaces between genes.

Sequence conservation over long evolutionary periods, such as the time since humans last shared a common ancestor with mice or fish, therefore usually predicts the location of genes. However, exceptions do exist: Conserved DNA sequences can be observed rarely at locations outside of the coding regions. The fact that these features are so well conserved suggests strongly that they have a function that is subject to evolutionary constraints—even if in most cases we do not yet know what these functions may be. Scientists are actively exploring the potential roles of these conserved noncoding sequences; for example, some might represent

enhancer elements (see **Fig. 8.11**) that help determine when and where nearby genes are transcribed into mRNA.

## The Most Direct Method to Find Genes Is to Locate Transcribed Regions

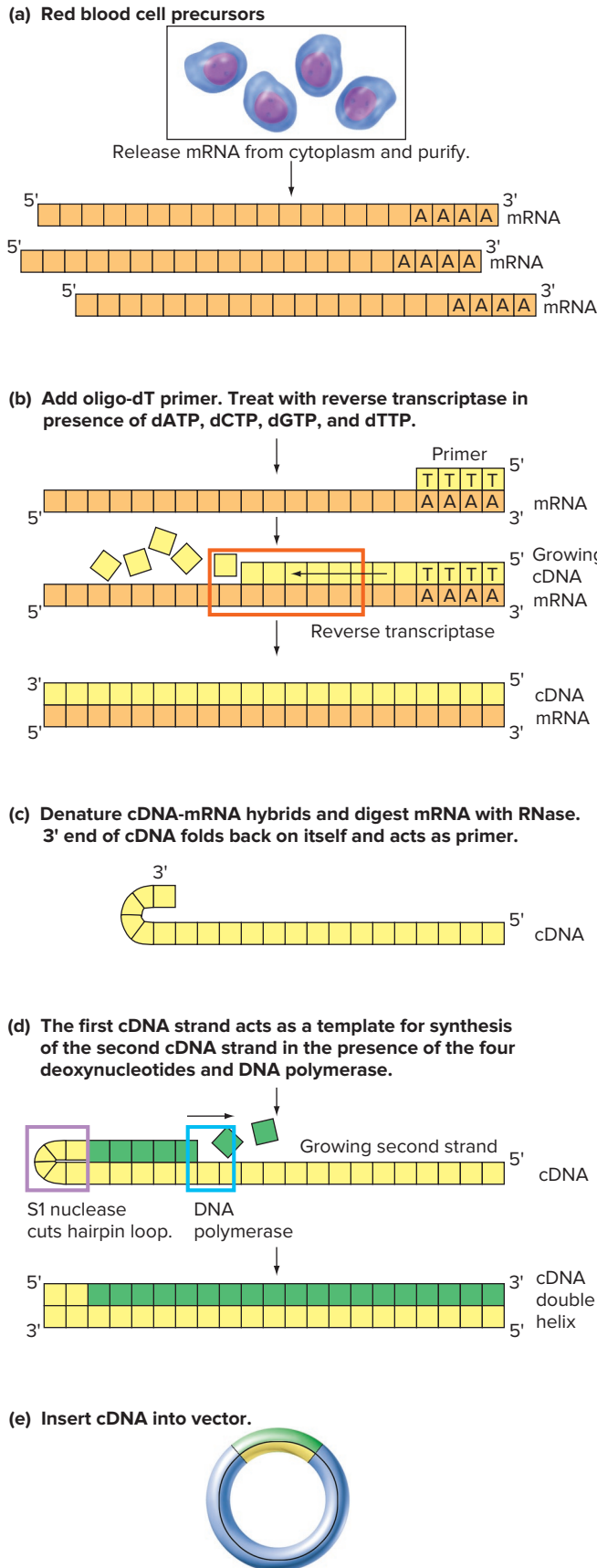
Many genes encode proteins while some others, such as the genes for rRNAs and tRNAs, do not. However, all genes are transcribed into RNAs, even if some RNAs are not translated. If you knew the sequence of the RNA produced from a gene, it would be easy to find that gene in genomic DNA simply by looking for the DNA sequence complementary to the RNA. This approach in fact works well for RNAs that can be purified in large amounts like rRNAs (which can be isolated from other RNAs because they form part of the ribosome).

In contrast, most mRNAs are so relatively rare in cells that they cannot be purified readily. Moreover, although technologies for determining the nucleotide sequence of RNAs do exist, they are less widely available and much more difficult to perform than the methods available for sequencing DNA. As a result, the easiest way to study mRNAs is to copy them into DNA, to clone the resultant DNA molecules, and then to sequence these clones by the same methods already described for genomic DNA.

### Making cDNA libraries

To produce DNA clones from mRNA sequences, researchers rely on a series of *in vitro* reactions that mimics part of the life cycle of viruses known as **retroviruses**. Retroviruses, which include among their ranks the HIV virus that causes AIDS, carry their genetic information in molecules of RNA. As part of their gene-transmission kit, retroviruses also contain the unusual enzyme known as **RNA-dependent DNA polymerase**, or simply **reverse transcriptase** (review the Genetics and Society Box in Chapter 8 entitled *HIV and Reverse Transcription*). After infecting a cell, a retrovirus uses reverse transcriptase to copy its single strand of RNA into a strand of **complementary DNA**, often abbreviated as **cDNA**. The reverse transcriptase, which can also function as a DNA-dependent DNA polymerase, then makes a second strand of DNA complementary to this first cDNA strand (and equivalent in sequence to the original RNA template). Finally, this double-stranded DNA copy of the retroviral RNA chromosome integrates into the host cell's genome. Although the designation *cDNA* originally meant a single strand of DNA complementary to an RNA molecule, it now refers to any DNA—single- or double-stranded—derived from an RNA template.

Let's see how you could use reverse transcriptase to make cDNA copies of all the mRNAs that are transcribed in a particular cell type such as red blood cell precursors. You would first isolate by simple chemical means the total population of RNA molecules in these cells (**Fig. 10.4a**).



**Figure 10.4** Converting RNA transcripts to cDNA. **(a)** Obtain mRNA from red blood cell precursors. **(b)** Create a hybrid cDNA-mRNA molecule using reverse transcriptase and oligo-dT primer. **(c)** Heat the mixture to separate mRNA and cDNA strands, and then eliminate the mRNA transcript. The 3' ends of the cDNA strands bind by chance to complementary nucleotides within the same strand, forming a *hairpin loop* that can prime DNA polymerization. **(d)** Create a second cDNA strand complementary to the first. The enzyme S1 nuclease is used to cleave the hairpin loop. **(e)** Insert the newly created double-stranded DNA molecule into a vector for cloning.

Next, because mRNAs constitute only a small fraction of all the RNAs in the cell (1–5% depending on the cell type), it would be desirable to separate the mRNAs from the much more abundant rRNAs and tRNAs. This goal is possible because mRNAs in eukaryotic cells have poly-A tails at their 3' ends. mRNAs will hybridize through their poly-A tails to the *oligo-dT* (single-stranded fragments of DNA containing about 20 Ts in a row). mRNA will thus bind to magnetic beads linked to oligo-dT, while other kinds of RNA will not. This interaction provides the basis for a separation technique (*not shown*) that will allow you to obtain a purified preparation of mRNA. The preparation will contain all of the mRNAs that are expressed in red blood cell precursors (Fig. 10.4a).

The addition of reverse transcriptase to this total mRNA—as well as ample amounts of the four deoxyribonucleotide triphosphates and primers to initiate synthesis—generates single-stranded cDNA bound to the mRNA template (**Fig. 10.4b**). The primers used in this reaction are also oligo-dT so as to initiate polymerization of the first cDNA strand from the 3' ends of all mRNAs. After synthesis is finished, you can **denature** (separate) the mRNA-cDNA hybrids into single strands by heating the hybrids to high temperature. The addition of an RNase enzyme that digests the original RNA strands leaves intact single strands of cDNA (**Fig. 10.4c**). Most of these fold back on themselves at their 3' ends to form transient hairpin loops that serve as primers for synthesis of the second DNA strand. Now the addition of DNA polymerase, in the presence of the requisite deoxyribonucleotide triphosphates, initiates the production of a second cDNA strand from the just-synthesized single-stranded cDNA template (**Fig. 10.4d**). The products are double-stranded cDNA molecules.

After using restriction enzymes and ligase to insert the double-stranded cDNA into a suitable vector (**Fig. 10.4e**) and then transforming the vector-insert recombinants into appropriate host cells, you would have a library of double-stranded cDNA fragments. The cDNA fragment in each individual clone will correspond to an mRNA molecule in the red blood cell precursors that served as your sample. It is important to note that this **cDNA library** includes only the exons from that part of the genome that these cells were actively transcribing for translation into protein. The clones in cDNA libraries do not contain introns because the mature mRNAs from which they were produced do not have introns. You

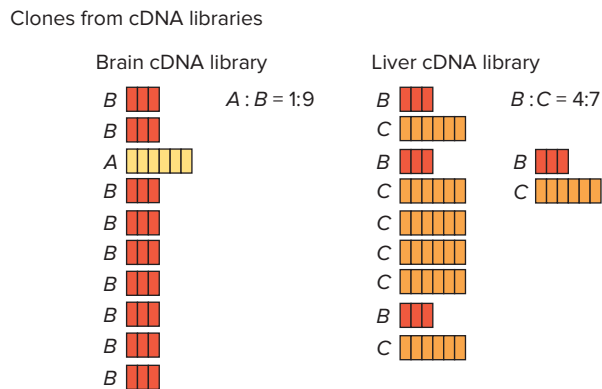
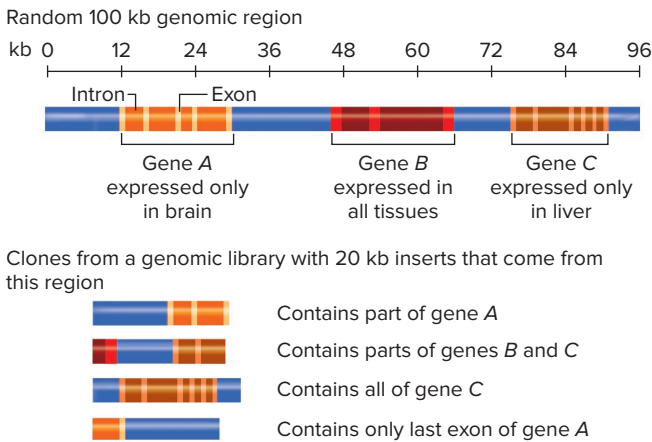
should also understand that the cDNA library made from red blood cell precursors would contain many clones corresponding to mRNAs that are highly expressed in this tissue, but only a few clones that reflect genes that are expressed rarely.

**Genomic versus cDNA libraries**

**Figure 10.5** compares genomic and cDNA libraries. The clones within genomic libraries represent all regions of DNA equally and show what the intact genome looks like in the region of each clone. The clones in cDNA libraries reveal which parts of the genome contain the information used in making proteins in specific tissues. The prevalence of the mRNAs for specific genes also gives some indication, though imperfect, of the relative amounts of the various proteins made in those cells.

As described previously, one of the main purposes of making cDNA libraries is to annotate genomes by finding

**Figure 10.5 A comparison of genomic and cDNA libraries.** Every tissue in a multicellular organism can generate the same genomic library, and the DNA fragments in that library collectively carry all the DNA of the genome. On average, the clones of a genomic library represent every locus an equal number of times. By contrast, different tissues in a multicellular organism generate different cDNA libraries. Clones of a cDNA library represent only the fraction of the genome that is transcribed in that tissue. The frequency with which particular fragments appear in a cDNA library is proportional to the level of the corresponding mRNA in that tissue.



transcribed regions (genes). The idea is very simple: You would determine the sequences of many cDNA clones, and then compare these cDNA sequences with that of the genome. Regions of identity between cDNA and genome represent the exons of genes, and the sequence of a complete cDNA (copied from a full-length mRNA) allows you to determine the exon/intron structure of the corresponding gene.

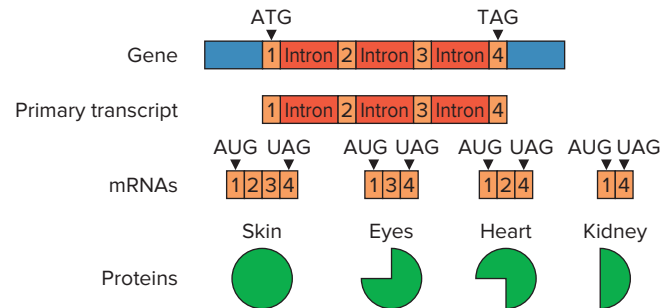
Although the basic idea of annotating genomes by comparing cDNA and genomic sequences is straightforward, putting it into practice on a large scale is not trivial. Because certain genes are expressed only rarely or only in certain tissue types, genomic scientists need to sequence millions of cDNA clones in multiple cDNA libraries made from mRNA derived from many diverse types of tissue. For this reason, it is likely that some infrequently expressed genes may not yet be recognized as genes in genome databases.

**cDNAs and alternative splicing**

Alternative splicing presents an additional challenge for genome annotation, one that is particularly important for predicting the amino acid sequences of the **proteome**—that is, all the proteins made in an organism. The problem is that a single primary transcript can be spliced in a variety of different ways, some of which can result in different proteins being made by a single gene (review Fig. 8.17).

The sequencing of many individual cDNA clones provides a solution to this issue caused by alternative splicing because each cDNA clone represents an individual mature mRNA. Analysis of these cDNAs is aided by the fact that alternative splicing of a primary transcript often occurs in a cell type-specific manner, allowing different kinds of cells to generate different (though related) proteins. This fact provides another reason why geneticists need to sequence cDNAs from libraries made using mRNAs from a variety of different tissues. The cDNA sequences will reveal which exons appear in the processed mRNAs in particular cell types, and thus will predict the amino acid sequences of the proteins present in those tissues (**Fig. 10.6**).

**Figure 10.6 Alternative splicing complicates human genome annotation.** Exons (orange) and introns (red) in the primary transcript can be alternatively spliced, often in cell type-specific ways; as a result, the same gene can express different proteins. Researchers analyze alternative splicing by sequencing multiple cDNA clones from libraries made from each of many different tissues.



**essential concepts**

- Long *open reading frames (ORFs)* in genomic DNA usually identify protein-coding exons.
- DNA sequences *conserved* between the genomes of widely divergent species often correspond to the exons of genes.
- *Reverse transcriptase* produces *complementary DNA (cDNA)* from mRNA transcripts; cDNA clones thus represent only the exons of genes.
- cDNA sequences reveal how a primary transcript is spliced in a given cell type, and thus predict the amino acid sequence(s) of a gene's protein product(s) in that cell type.

**10.2 Genome Architecture and Evolution****learning objectives**

1. Discuss the arrangement of genes in genomes, including the number of genes, transcription direction, and gene density.
2. Explain how gene duplication and divergence lead to the formation of gene families and pseudogenes.
3. List three ways in which genomes can change over evolutionary time.
4. Describe how mechanisms at the DNA, RNA, and protein levels can produce complexity from a small number of genes.

The complete sequences of the human and other genomes have provided striking new insights into the organization and evolution of genomes. Our detailed knowledge of genomic sequences has changed profoundly the practice of biology. We now briefly describe some of the main lessons and surprises from these genome projects, focusing on the following three questions: How are the genes arranged in genomes? How do genes and genomes change during evolution? Lastly, how can genomes with a relatively small number of genes produce the vast complexity of phenotypes that results in living organisms, including humans?

**The Arrangement of Genes in the Genome Is Not Uniform**

A major shock to emerge from the completion of the human genome sequence was the discovery of only about 27,000 genes. Of these genes, roughly 19,000 encode proteins, while the remainder are transcribed into RNAs

that are not translated, such as rRNAs and tRNAs that participate in translation, and the snRNAs that function in spliceosomes. These numbers are much lower than expected. Estimates made before the initiation of the Human Genome Project had suggested that 100,000 or more genes might exist. Many of these estimates anticipated that because humans have much greater biological complexity than simpler model organisms such as bacteria, yeasts, nematodes, and fruit flies, our genomes must have many more genes. Although the human genome indeed has more genes than these organisms, the difference in gene number is not nearly as great as had been thought (review Table 9.2). Mechanisms other than changes in gene number must therefore underly metazoan (multi-cellular animal) complexity.

The total length of genomes varies much more markedly over the course of evolution than the number of genes (compare especially the genomes of mammals such as mice and humans with model eukaryotes like worms and flies; Table 9.2). This generalization holds because the **exome**, the part of the genome corresponding to the exons of all known genes, constitutes only a small proportion of genomes, roughly 1.5–2.0% of the total in humans. The vast majority of DNA sequences instead are found in introns, in the spaces between genes (*intergenic regions*), in transposable elements that can move from one chromosomal position to another, and in structural features like centromeres and telomeres.

The tremendous variation seen in the size of genomes of different species is thus mostly due to expansions and contractions of noncoding DNA outside of the exome, rather than changes in gene number or gene sizes. For example, half or more of the human genome appears to be composed of transposable elements, often regarded as *selfish* or parasitic DNA that uses our genomes as a host for their own propagation. In a second example, the human genome also contains many simple repeating sequences (such as CGCGCGCG, etc.).

In this section, we focus on the small proportion of the genome corresponding to the genes, with an emphasis on those that encode proteins. Later chapters will describe in greater detail the nature of DNA sequences that constitute other chromosomal elements about which some information is available (centromeres, telomeres, and transposable elements). However, you should know that a large percentage of the human genome is “dark matter” whose presence we do not yet understand. Some of this DNA may have functions that are currently obscure, but much of it may in fact not have any function at all and may instead be the vestiges of chance events that occurred during evolution.

**Random orientation of transcription of most genes**

Students sometimes assume that all the genes on a chromosome are transcribed in the same direction, always using the same strand of double-stranded genomic DNA as the

template. This assumption is absolutely incorrect. As previously shown in Fig. 10.3, neighboring genes can be transcribed either in the same or in opposite directions with respect to each other, and either toward the centromere or toward the telomere with respect to the chromosome as a whole. Gene maps such as that in Fig. 10.3 typically indicate the 5'-to-3' direction of a gene's transcription (that is, the direction that RNA polymerase moves as it copies the gene into RNA) with an arrow.

Because neighboring genes can be transcribed in opposite directions, RNA polymerase uses the chromosome's *Watson* strand as the template for some genes, while for other genes, the template is the *Crick* strand. For most genes, the direction of transcription appears to be chosen at random, or at least no definitive patterns can yet be discerned. However, in a few exceptional genomic regions, such as those containing the hemoglobin genes that will be described later in this chapter, specific mechanisms of gene regulation require that neighboring genes have the same transcriptional orientation.

### Variable gene density

On average, the density of genes in the human genome is slightly less than 1 gene in every 100 kb of DNA (27,000 genes in a 3,000,000 kb genome). However, this rough value obscures the fact that the packing of genes can be very different in various parts of the genome. Some regions on some chromosomes are gene-rich, with little space between densely packed genes. The most gene-rich region of the human genome is a 700 kb stretch of chromosome 6 that contains 60 genes encoding histocompatibility proteins with diverse functions (Fig. 10.7).

Other regions, called **gene deserts**, contain few or no genes. The largest known desert in the human genome is 5.1 Mb on chromosome 5 without a single identified gene. Some deserts are gene-poor because they contain so-called *big genes* whose nuclear transcript spans 500 kb or more of chromosomal DNA. The largest of the big genes in humans is the gene for dystrophin, which spans 2.3 Mb, most of which is composed of introns. Interestingly, because big

genes can be expressed only through the production of enormous primary transcripts, their transcription cannot be completed in rapidly dividing cells. Many big genes are thus expressed only in neurons, which do not divide. It is possible that scientists have not yet detected some big genes, which might be transcribed into RNA only rarely.

A fundamental unanswered question is whether gene-rich and/or gene-poor regions have biological meaning. Is there a functional explanation for these variations in gene density, or do they instead reflect random fluctuations in evolutionary events that shape chromosomal architecture?

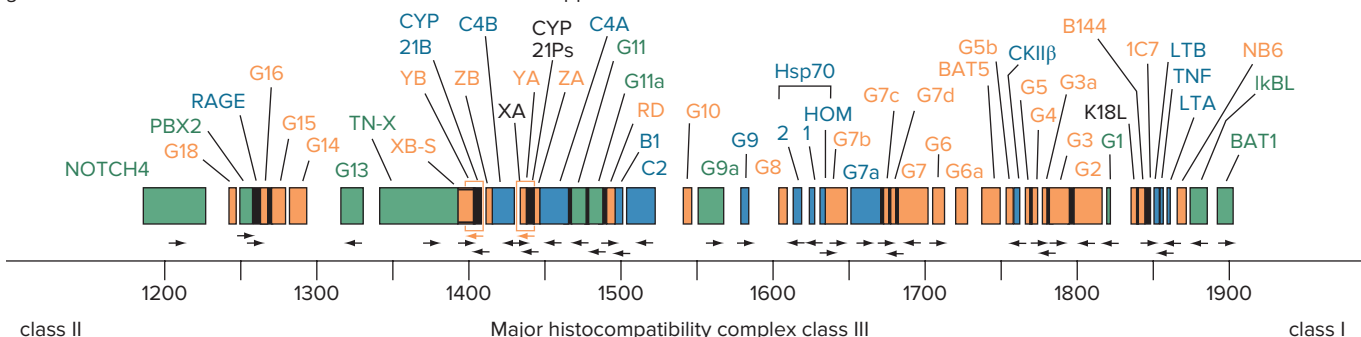
## Genomes Undergo Evolutionary Change

Genomes continually undergo many different kinds of DNA sequence changes that provide the raw material for natural selection. In earlier chapters of this book, you have encountered some of the events that can alter the nucleotide sequence of pre-existing genes: In particular, environmental mutagens and mistakes in DNA replication can result in nucleotide substitutions. The accumulation of such point mutations within a gene can certainly change the gene's function over time. However, as we now describe, analysis of the genomes of humans and other species indicates that evolution can also create new genes and reorganize genomes by reshuffling blocks of DNA larger than a single nucleotide. Many kinds of processes promote genome plasticity over the evolutionary timeframe.

### Alterations in domain architecture

Genome annotation has revealed that exons often encode discrete **protein domains**, each of which is a linear sequence of amino acids that folds up in three-dimensional space so as to act as a single functional unit. Genes with multiple exons often encode proteins with multiple domains analogous to the cars of a train. Each train is composed of many different cars, and each kind of car (engine, flat car, dining car, caboose) has a discrete function. Different trains may carry different combinations of cars and thus fulfill

**Figure 10.7** Class III region of the human major histocompatibility complex. This densely packed 700 kb-long region contains 60 genes (colored boxes). Arrows below the genes indicate the direction of each gene's transcription; just as in Fig. 10.3, some genes are transcribed in one direction and others in the opposite direction.



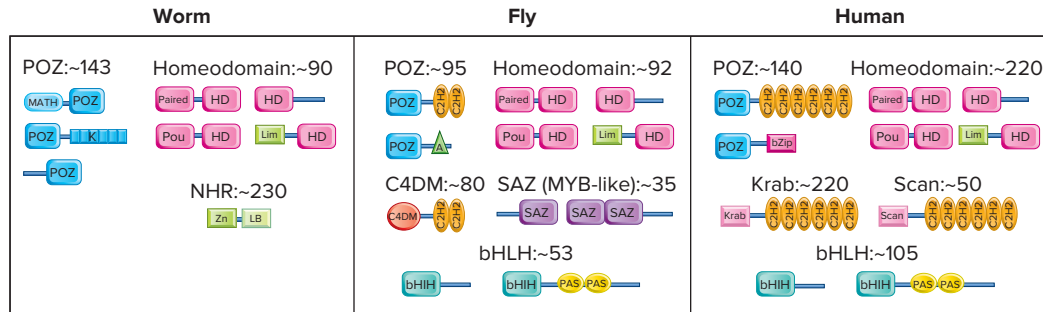
class II

Major histocompatibility complex class III

class I



**Figure 10.8 Domain architecture of transcription factors.** Protein domains are indicated as colored icons labeled POZ, HD (homeodomain), etc. Horizontal lines connect domains found in the same protein. The numbers and types of transcription factors vary considerably between different species due to protein domain reshuffling during evolution. As an example, worms make about 143 different transcription factors containing POZ domains, of which three are shown, while fruit flies make about 93 POZ-containing proteins, of which two are shown. Some of the domains shown govern DNA binding, while others facilitate protein–protein interactions.



**Figure 10.9 Homeodomain consensus sequence.** The **consensus sequence** of amino acids shows the most commonly found amino acid at a given position within all known homeodomains in all organisms. Subsequent rows show matches to the consensus (*purple*) of homeodomains in nine *Drosophila* proteins that dictate key aspects of the animal's development. (These genes and proteins will be discussed in Chapter 19.)

Consensus	RRRKR	TAYTRY	QLELE	KEFHF	NRYL	TRRRR	IELA	HSNL	TERQV	KIWF	QNR	RHK	KWK	KEN
Ubx	RRRGR	QTYTRY	QTL	ELEKE	FHTN	HYL	TRRRR	IE	MAHAL	SL	TERQ	I	K	IWFQ
Abd-A	RRRGR	QTYTR	FQTL	ELEKE	FHN	HYL	TRRRR	IE	I	AHAL	SL	TERQ	I	K
Abd-B	VRKKR	KPYSK	FQTL	ELEKE	FL	FNAV	SKQ	KRW	L	M	R	NAQ	S	L
lab	NNSGR	TNFTN	KQL	TELEK	EFHN	RNYL	TRRRR	IE	I	A	N	L	Q	L
pb	PRRLR	TAYTN	TQL	ELEKE	FHTN	KYL	CR	PRR	IE	I	A	A	S	L
Dfd	PKRQR	TAYTL	LHQ	I	ELEKE	FHN	RNYL	TRRRR	IE	I	A	H	L	V
Scr	TKRQR	TSYTRY	QTL	ELEKE	FHN	RNYL	TRRRR	IE	I	A	H	A	L	S
Antp	RKRGR	QTYTRY	QTL	ELEKE	FHN	RNYL	TRRRR	IE	I	A	H	A	L	S

different purposes. Similarly, many genes are composed of multiple exons that encode discrete protein domains. The shuffling, addition, or deletion of exons during evolution can create new genes whose protein products have novel **domain architectures** (different numbers and kinds of domains in different orders) and thus can assume new roles in cells and organisms.

**Figure 10.8** shows an example of the domains associated with various *transcription factors*, proteins that bind to regions of DNA such as enhancers that control the transcription of nearby genes. Exon shuffling over evolution has produced different transcription factors with differing domains that enable these proteins to recognize particular DNA sequences and also to interact uniquely with cofactors such as other proteins.

Biologists may guess at the function of a new protein (or the gene that encodes it) by analogy, if they find by computerized analysis that it contains a domain known to play a specific role in other proteins. As an example, many proteins that include a *homeodomain* (a particular DNA-binding motif) are transcription factors important for the development of multicellular organisms. Computer algorithms determine that a particular gene encodes a

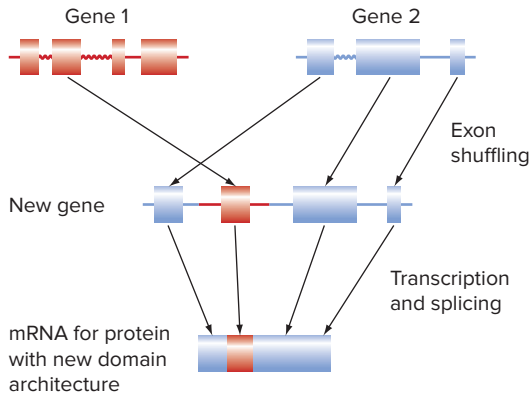
homeodomain-containing protein by comparing its putative amino acid sequence to those of known homeodomains and searching for similarity (**Fig. 10.9**).

The mechanism of RNA splicing facilitates this kind of exon rearrangement in eukaryotic genomes (and thus the creation of new genes) because the reshuffling does not have to be precise. Suppose, as shown in **Fig. 10.10**, that the exon of one gene plus its flanking introns is moved to a new location in the intron of a different gene. This exon can now be spliced together with the second gene's exons to make a single mRNA molecule, regardless of where within the introns these pieces of DNA were brought together.

### Gene families

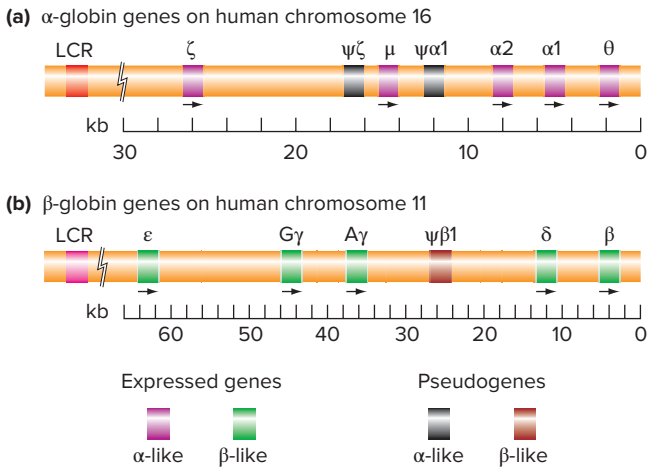
**Gene families** are groups of genes closely related in sequence and function; such gene families are abundant throughout genomes. Examples of gene families include the genes that encode the hemoglobins that allow us to transport oxygen in our blood (**Fig. 10.11**), the immunoglobins (antibodies) that help us ward off infections, and the olfactory receptors critical for our sense of smell.

**Figure 10.10 Exon shuffling.** Suppose two genes are broken in introns and joined together as shown. Transcription of the newly reshuffled gene will produce a primary transcript that can be spliced into a mature mRNA encoding a novel protein, regardless of where in the introns the breakages occurred. If the different exons encode different protein domains, the domain architecture of proteins can change over the course of evolution, as seen for the transcription factors in Fig. 10.8.



**Figure 10.11 The genes for human hemoglobin polypeptides are located in two genomic clusters.**

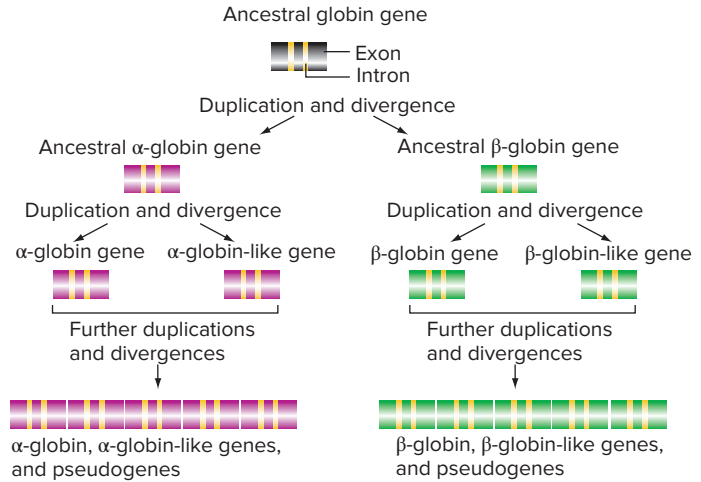
(a) Schematic representation of the  $\alpha$ -globin locus. The five functional genes are indicated with purple boxes, the two pseudogenes with black boxes. All of these genes are transcribed in the same direction (left-to-right on the map). The red box is the locus control region (LCR) described later in this chapter. (b) Schematic representation of the  $\beta$ -globin locus; this cluster has five functional genes (green) and one pseudogene (brown).



With the use of bioinformatics, researchers can see that each gene family evolved by a process of **duplication and divergence** from an ancestral gene. The two DNA sequence products of a duplication event, which start out identical, eventually diverge as they accumulate different mutations (Fig. 10.12). Additional rounds of duplication and divergence can further increase the number of related genes. For example, the human genome has ten functional members of the hemoglobin gene family, while the olfactory

**Figure 10.12 Evolution of the globin gene family.**

Duplication of an ancestral gene, followed by divergence of the duplication products, established the  $\alpha$ - and  $\beta$ -globin lineages. Further rounds of duplication and divergence within the separate lineages generated the genes and pseudogenes of the current-day globin gene family.

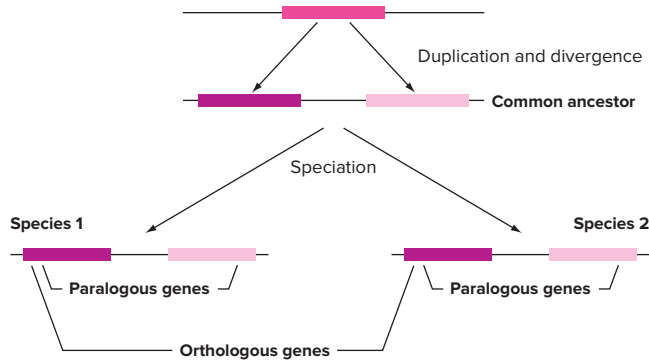


receptor family includes about 1000 genes. The duplication and divergence process is crucial for the creation of new raw material for evolution. Once a gene has duplicated, divergence allows either or both of the copies to assume new specialized but related functions, as long as one or both of the copies still fulfills the role of the original gene.

The genes in such families may be clustered together on one chromosome or dispersed on several chromosomes. In the case of the hemoglobin family, the  $\alpha$ -globin gene cluster (also called the  $\alpha$ -globin locus) on chromosome 16 contains five functional genes, while the  $\beta$ -globin cluster ( $\beta$ -globin locus) on chromosome 11 also has five genes (Fig. 10.11). The sequences of all the  $\alpha$ -like genes are more similar to each other than they are to the  $\beta$ -like sequences, and *vice versa*. The  $\beta$ -like genes are exactly the same length, and the five  $\beta$ -like genes have two introns at exactly the same position; in fact, the  $\alpha$ -like genes also have two introns at the same positions.

These comparisons suggest that all of the globin genes can be traced back to a single ancestral DNA sequence (Fig. 10.12). Hundreds of millions of years ago, this ancestral globin gene duplicated, and one copy moved to another chromosome. With time, one of the two copies gave rise to the  $\alpha$ -lineage, the other to the  $\beta$ -lineage. Each lineage then underwent further duplications to generate the present array of  $\alpha$ -like and  $\beta$ -like genes in humans. By comparing genomes of different organisms, it is possible to estimate when these various duplication events occurred. For example, the  $\beta$ -globin clusters of humans and chimpanzees have the same genes in the same order, but some other primates have one

**Figure 10.13 Gene family nomenclature.** *Orthologous genes* are separated by a speciation event. *Paralogous genes* are separated by a duplication event. *Homologous genes* are related to each other by descent from a common ancestral DNA sequence regardless of the mechanism of separation; all the genes shown in this figure are thus homologous.



fewer  $\beta$ -like gene. Thus, the last gene duplication event in the  $\beta$ -globin cluster must have occurred in a common primate ancestor of humans and chimps.

The existence of gene families requires the definition of new terms to describe the relationship of the genes that compose them (Fig. 10.13). **Orthologous genes** are genes in two different species that arose from the same gene in the species' common ancestor; usually but not always, orthologous genes retain the same function. The genes for the  $\epsilon$ -globin in humans and chimpanzees are orthologs because an  $\epsilon$  gene already existed in their last common ancestor. By contrast, **paralogous genes** arise by duplication; this term is usually used to denote the different members of a gene family. Thus, the genes for  $\delta$ -globin and  $\epsilon$ -globin in the human  $\beta$ -globin locus (Fig. 10.11b) are close paralogs, and both are more distant paralogs of genes in the  $\alpha$ -globin cluster. Finally, **homology** is a blanket term for all evolutionarily related sequences; all hemoglobin genes in all species are thus homologous, and all these genes share weaker homologies with myoglobin genes encoding more distantly related oxygen-carrying proteins in muscle tissues rather than red blood cells.

The duplications that gave rise to multiple functional hemoglobin genes also produced genes that eventually lost the ability to function. Molecular geneticists made this last deduction from data showing two additional  $\alpha$ -like sequences within the  $\alpha$  locus and one  $\beta$ -like sequence within the  $\beta$  locus that no longer have the capacity for proper expression (Fig. 10.11). The reading frames are interrupted by frameshifts, missense mutations, and nonsense codons, while regions needed to control the expression of the genes have lost key DNA signals. Sequences that look like, but do not function as, genes are known as **pseudogenes**; they occur in many gene families throughout all higher eukaryote genomes. Interestingly, the same pseudogene with almost all the same gene-inactivating mutations is found in the

$\beta$  cluster of both the human and chimpanzee genomes, indicating that the duplication giving rise to the pseudogene, as well as many of the mutations that disrupt its function, must have existed in a common primate ancestor of both species.

Because they serve no function, pseudogenes are subject to mutation without selection and thus accumulate mutations at a far faster pace than coding or regulatory sequences of a functional gene. Eventually, nearly all pseudogene sequences mutate past a boundary beyond which it is no longer possible to identify the functional genes from which they have been derived. Continuous mutation can thus turn a once functional sequence into an essentially random sequence of DNA.

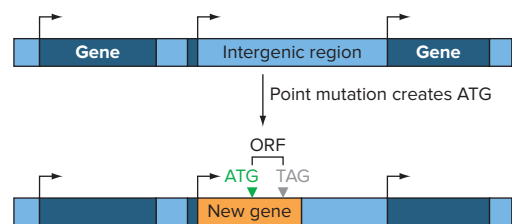
### De novo genes

Most annotated genes in any sequenced genome belong to gene families and are also homologs of genes that exist in many distantly related species. However, many genes discovered by genome sequencing appear either to lack homologs in any other species or to have homologs only in closely related species. For example, a few hundred genes in the human genome are human-specific. Genes without homologs are called **de novo genes**. The term *de novo* means *from new* in Latin.

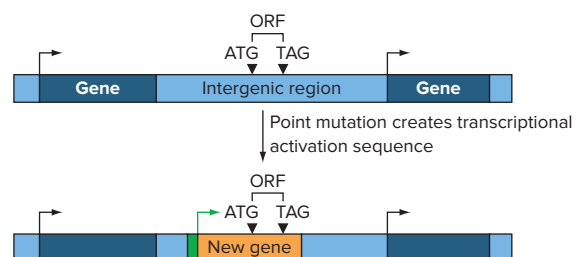
De novo genes are young genes that evolved recently from ancestral intergenic sequences. Evidence exists for two different mechanisms of de novo gene evolution through mutation: Either transcribed intergenic regions gained an ATG and thus a short ORF (Fig. 10.14a), or small

**Figure 10.14 Origins of de novo genes.** Genes without homologs can arise either when (a) transcribed intergenic DNA mutates to generate an ATG and thus a small ORF, or when (b) a small ORF in intergenic DNA acquires transcriptional activation sequences.

#### (a) Transcribed intergenic DNA acquires ORF



#### (b) ORF in intergenic DNA acquires transcriptional activation sequences



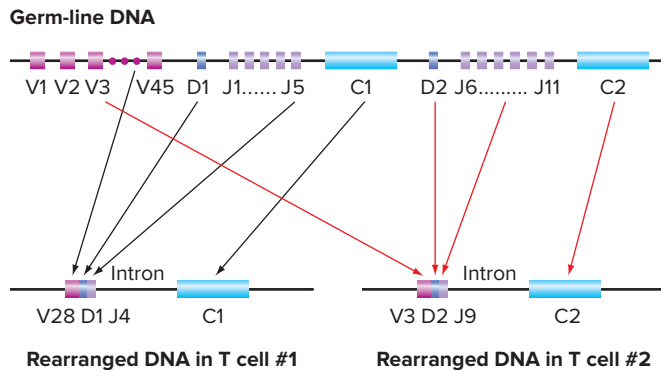


**Combinatorial strategies at the DNA level**

T-cell receptor genes, among the best-studied examples of DNA-level combinatorial amplification, are encoded by a multiplicity of gene segments that become rearranged in one type of somatic cells—T cells—but not in the germ line or any other type of cell (Fig. 10.16). The human T-cell receptor family has 45 functional variable (V) gene segments, two functional diversity (D) gene segments, 11 functional joining (J) gene segments, and two almost identical constant (C) segments. In an individual T cell, any D element may first join to any J element by deletion of the intervening DNA. This joined D-J element may, in turn, join to any V element—once again by deletion of the intervening DNA—to generate a complete V-D-J exon. This combinatorial process can generate 990 different V-D-J exons ( $45 \times 2 \times 11 = 990$ ), although in a given T cell, only one such functional rearrangement occurs. Thus, from 58 gene elements ( $45 + 2 + 11$ ) within a single gene, a combinatorial joining mechanism can generate 990 different kinds of T-cell receptor proteins.

T-cell receptors are capable of interacting with foreign molecular structures, which are termed *antigens*. T cells are driven by contact with antigens to divide and expand their numbers 1000-fold or more. This antigen-triggered expansion by mitosis to a clone of genetically identical cells is a key part of every immune response. The particular combinatorial gene arrangements in a few of the original population of T cells by chance produce T-cell receptors fitted more precisely to a specific antigen. Binding with the antigen then triggers the clonal expansion of the cells that carry the tightly fitting receptors, amplifying the useful combinatorial information. The specificity and strength of the immune response thus increases over time because of

**Figure 10.16 Gene for the human  $\beta$  T-cell receptor chain.** In the germ line and in most somatic cells, the gene is composed of about 45 V elements, 2 D segments, about 11 J elements, and 2 nearly identical C (constant) regions. During T cell development, any D may join with any J. Subsequently, any V may join with any D-J. Finally, the rearranged V-D-J exon is spliced to a C exon. As a result of these sequential rearrangements of a single gene, different T cells can express one of almost 1000 different kinds of  $\beta$  receptor chains.



the proliferation of T cells that have particular V-D-J rearrangements encoding the best receptors for the antigen to which the individual was exposed.

**Combinatorial strategies at the RNA level**

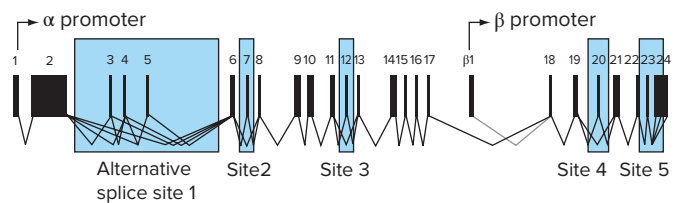
The splicing together of RNA exons in different orders—*alternative splicing*—is another way in which combinatorial strategies can increase information and generate diversity. Further diversity results from the initiation of transcription at distinct promoter regions, which creates transcripts with different numbers of exons.

The three neurexin genes, which encode proteins that help bind neurons together at synapses, illustrate both of these combinatorial RNA strategies (Fig. 10.17). Each neurexin gene contains two promoter regions (producing  $\alpha$ - and  $\beta$ -class mRNAs) and five sites at which alternative splicing can occur. Together, these three genes can probably generate more than 2000 alternatively spliced forms of mRNA. Key questions include how many of the splice variants encode functionally distinct proteins (rather than proteins with the same function), and whether different variants represent different addresses for telling neurons where to go during embryonic development. By looking at the sequences of many cDNA clones, scientists have detected some splice variants that are specific for particular subsets of nerve tissue, suggesting the importance of this combinatorial strategy for nervous system organization.

**Posttranslational modification of proteins**

Human proteins may be modified by more than 400 different chemical reactions, each capable of altering the proteins' functions. Some examples of these posttranslational modifications were shown in Fig. 8.26, and they include reactions such as protein cleavage and protein phosphorylation. Thus, the typical human cell might have perhaps 50,000 different types of mRNAs (the primary transcripts of many genes are alternately spliced in a single cell type) but perhaps 1 million different proteins. Human cells can make more types of protein modifications than can cells of their simpler model-organism counterparts.

**Figure 10.17 The organization of the human neurexin genes.** The human genome has three genes encoding neurexin. Each gene has two promoters ( $\alpha$  and  $\beta$ ) to initiate mRNA synthesis and five sites at which alternative RNA splicing can occur. The *blue rectangles* indicate exons affected by alternative splicing. Numbers at the top of the figure designate individual exons.



## Genome Sequence Studies Affirm Evolution from a Common Ancestor

Comparisons of complete genomic sequences from nearly 10,000 species to date have resoundingly supported the ideas that began with Darwin and Mendel: All living organisms have similar genetic components for accomplishing basic cellular processes. This conclusion strongly supports the idea that we and other living organisms are all descendants of a single, fortuitous life-producing biochemistry. The similarity of basic genetic components also affirms that the analysis of appropriate biological systems in model organisms can provide fundamental insights into how the corresponding systems function in humans.

### essential concepts

- Even the most complex genomes have surprisingly few genes (about 27,000 in the human genome).
- Gene density varies considerably within a genome, reflecting differences in intron size and in the spacing between genes.
- In most regions of the genome, the orientations of individual genes and thus the direction of gene transcription appears to be chosen at random.
- New genes can arise during evolution through: (i) *exon shuffling*, which can alter the domain structure of proteins; (ii) *duplication and divergence* that generates gene families; and (iii) *de novo mutations* in intergenic DNA sequences.
- *Combinatorial strategies* at the DNA level and RNA level, as well as *posttranslational modifications* of proteins, allow the production of highly diversified gene products even from a single gene.
- Genome comparisons affirm that all present life descended from a single common ancestor.

## 10.3 Bioinformatics: Information Technology and Genomes

### learning objectives

1. Explain the relevance of a species RefSeq to bioinformatic studies.
2. Describe the uses of BLAST searches in comparative genomics.

At the time of this writing in 2016, the genomes of more than 8000 species including our own have already been characterized by DNA sequence analysis, and this number is increasing continually. The amount of sequence data

available to researchers is staggering. Scientists must therefore rely on computers to store and help interpret this vast supply of information. The digital language used by computers for information storage and processing is ideally suited to handle the digital A, C, G, T code that exists naturally in genomes. These four values can be represented in two digits of binary code (00, 01, 10, and 11).

Keeping pace with the 1980s revolution in biological data generation fostered by the advent of automated DNA sequencing, a parallel revolution was occurring in information technology. The Internet came into existence along with personal computers that were linked together to establish rapid transmission of electronic data from one lab to another. It was a straightforward task to channel the output of DNA sequencer machines directly into electronic storage media, from which sequences were available for analysis and transmission to other scientists.

The **GenBank** database, established by the National Institutes of Health in 1982, still serves as the most widely used online repository of sequence data. The information is generated in molecular biology laboratories around the world, which deposit their sequences into GenBank electronically. From its establishment, the GenBank database has doubled in size roughly every 18 months, so that by 2016 it contained more than 300 billion annotated nucleotides of sequence information. One of the great powers of GenBank is that anyone in the world with an Internet connection can access this incredible storehouse of information easily.

## Bioinformatics Provides Tools for Visualizing and Analyzing Genomes

**Bioinformatics** is the science of using computational methods—specialized software—to decipher the biological meaning of information contained within organismal systems. This section provides some examples of bioinformatics tools that can be accessed through any web browser to examine and interpret publicly available genome data.

### The species RefSeq

Comparisons of experimental data involving DNA sequences generated by different laboratories depend on the use of a universally agreed-upon standard for analysis. This role is played by a *species reference sequence*, abbreviated as **RefSeq**. A RefSeq is a single, complete, annotated version of the species genome. RefSeqs are maintained by the National Center for Biotechnology Information (NCBI: <http://www.ncbi.nlm.nih.gov>), which was established in 1988 to oversee GenBank and other public databases of biological information and to develop bioinformatic applications for analyzing, systematizing,

and disseminating the data. A RefSeq need not be derived from a single individual, and it need not contain the most common genetic variants found in species members. Rather, it is simply an arbitrary, but well-characterized, example against which all newly obtained sequences from that species can be compared.

### Visualizing genes and genomes

Several web-based programs have been developed that allow a user to examine visual representations of genome data. One such program is the UCSC Genome Browser (<https://genome.ucsc.edu/>) that visualizes RefSeq genes and their associated annotations, showing features such as exon/intron structure and the location of protein-coding regions. Fig. 10.3 showed an example of the Genome Browser output, focusing on a 100 kb region of the human genome containing four genes. The transcription units are indicated at the bottom of the figure with large blue arrows that depict the extent of the gene, the direction of transcription, and each gene's exon/intron structure (exons represented as wider than the introns). Researchers can adjust their view of the browser to show many additional genomic features of interest, such as alternative splice variants, the location of repetitive DNA sequences, similarities with the genomes of other organisms, and the location of possible transcriptional regulatory elements.

### BLAST Searches Automate the Finding of Homologous Sequences

Suppose that you have identified a gene, for example from the fruit fly *Drosophila*, that is of interest to you. You would like to know whether the human genome contains a homolog of this fly gene. One tool you could use is an NCBI program called BLAST (Basic Local Alignment Search Tool), which allows you to find nucleotide or amino acid sequences related to any given nucleotide or amino acid sequence. **Figure 10.18** displays a typical output of a BLAST search, in this case looking for human proteins that share similarity with a *Drosophila* protein of interest. The

**Figure 10.18** Output from a BLAST search. The program was asked to find a human protein related to a protein in *Drosophila*. The Query shows part of the sequence of the fly protein (from amino acids 688–720); the Subject (Sbjct) indicates the corresponding amino acids in the human protein found by the search. Some of these amino acids are identical in the fly and human proteins. Positions marked with a plus (+) are conservative substitutions in which the substituted amino acids have similar chemical properties. At some positions the amino acids are very different, suggesting that the identities of these particular amino acids are not crucial to protein function.

Query	688	GPLTASYKSDEIKHLIRALFQDTDWRAKAITQI	720
		GPL A++ S E+K LIRALFQ+T+ RA A+ +I	
Sbjct	583	GPLAAAFSSSEVKALIRALFQNTERRAAALAKI	615

Query is the sequence you already know; here, the amino acid sequence of the *Drosophila* protein written in the one-letter code. The Subject is the homologous sequence found by the BLAST program; in this case, the related human protein. The row between the Query and the Subject indicates the conserved amino acids, with a + symbol denoting conservative amino acid replacements (missense substitutions in which an amino acid is replaced by a different amino acid with similar chemical properties).

To appreciate the power of bioinformatics programs such as the Genome Browser and BLAST search tool, you really need to access and use them yourself. Problems 23 and 24 at the end of this chapter involve some simple exercises that will place a few of these vast genomic databases at your disposal.

#### essential concepts

- *Bioinformatics* applications that are freely accessible online provide gateways for the exploration of genomic data.
- *Genome browsers* show the arrangement and structure of genes within RefSeq genomes.
- A *BLAST search* allows rapid, automated matching of particular DNA or amino acid sequences across multiple species for analysis of evolutionary relationships.

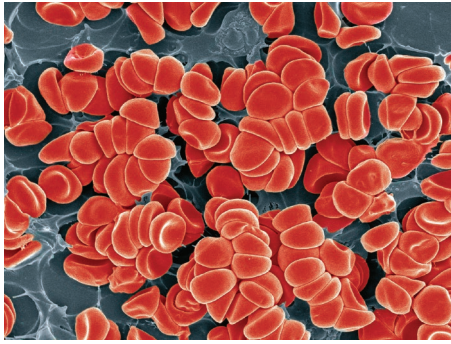
## 10.4 A Comprehensive Example: The Hemoglobin Genes

#### learning objectives

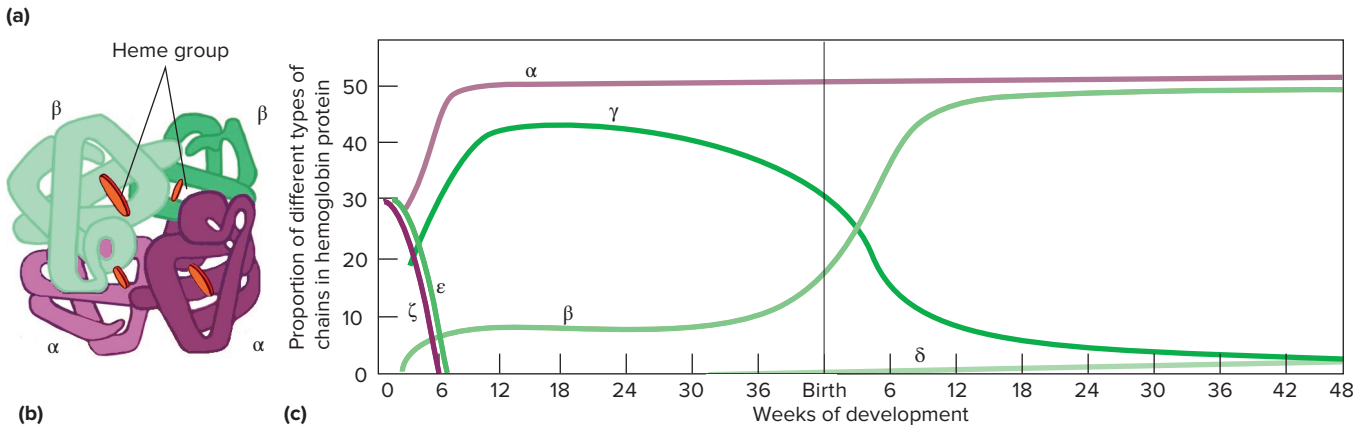
1. Discuss why it is advantageous for humans to produce different hemoglobins at different stages of development.
2. Explain how the clustering of hemoglobin genes impacts the cellular strategy to regulate their expression.
3. Predict the phenotypic severity of particular mutations in the  $\alpha$  and  $\beta$  clusters.

The vivid red color of our blood arises from its life-sustaining ability to carry oxygen. This ability, in turn, derives from billions of red blood cells, each one packed with close to 280 million molecules of the protein pigment known as *hemoglobin* (**Fig. 10.19a**).

A normal adult hemoglobin molecule consists of four polypeptide chains—two alpha ( $\alpha$ ) and two beta ( $\beta$ ) globins—each surrounding an iron-containing small molecular structure



**Figure 10.19 Hemoglobin is composed of four polypeptide chains that change during development.** (a) Scanning electron micrograph of adult human red blood cells loaded with hemoglobin. (b) Adult hemoglobin consists of two  $\alpha$  and two  $\beta$  polypeptide chains, each associated with an oxygen-carrying heme group. (c) Hemoglobin switches during human development from an embryonic form containing two  $\alpha$ -like  $\zeta$  chains and two  $\beta$ -like  $\epsilon$  chains to a fetal form containing two  $\alpha$  chains and two  $\beta$ -like  $\gamma$  chains, and finally to the adult form containing two  $\alpha$  and two  $\beta$  chains. In a small percentage of adult hemoglobin molecules, a  $\beta$ -like  $\delta$  chain replaces the usual  $\beta$  chain. The  $\alpha$ -like chains are in magenta, and  $\beta$ -like chains are in green. Levels of protein expression from the  $\mu$  and  $\theta$  genes shown in Fig. 10.11 are very low. a: © Science Photo Library RF/Getty Images



known as a *heme group* (Fig. 10.19b). The iron atom within the heme sustains a reversible interaction with oxygen, binding it firmly enough to hold it on the trip from lungs to body tissue but loosely enough to release it where needed. The intricately folded  $\alpha$  and  $\beta$  chains protect the iron-containing hemes from substances in the cell's interior. Each hemoglobin molecule can carry up to four oxygen atoms, one per heme, and these oxygenated hemes impart a scarlet hue to the pigment molecules and thus to the blood cells that carry them.

### Different Hemoglobins Are Expressed at Different Developmental Stages

The genetically determined molecular composition of hemoglobin changes several times during human development, enabling the molecule to adapt its oxygen-transport function to the varying environments of the embryo, fetus, newborn, and adult (Fig. 10.19c). In the first five weeks after conception, the red blood cells carry *embryonic hemoglobin*, which consists of two  $\alpha$ -like zeta ( $\zeta$ ) chains and two  $\beta$ -like epsilon ( $\epsilon$ ) chains. Thereafter, throughout the rest of gestation, the cells contain *fetal hemoglobin*, composed of two bona fide  $\alpha$  chains and two  $\beta$ -like gamma ( $\gamma$ ) chains. Then, shortly before birth, production of *adult hemoglobin*, composed of two  $\alpha$  and two  $\beta$  chains, begins to climb. By the time an infant reaches

three months of age, almost all of his or her hemoglobin is of the adult type.

Evolution of the various forms of hemoglobin maximized the delivery of oxygen to an individual's cells at different stages of development. The early embryo, which is not yet associated with a fully functional placenta, has the least access to oxygen in the maternal circulation. Both embryonic and fetal hemoglobin evolved to bind oxygen more tightly than adult hemoglobin does; they thus facilitate the transfer of maternal oxygen to the embryo or fetus. All the hemoglobins release their oxygen to cells, which have an even lower level of oxygen than any source of the gas. After birth, when oxygen is abundantly available in the lungs, adult hemoglobin, with its more relaxed kinetics of oxygen binding, allows for the most efficient delivery of the vital gas to other organs.

We have already seen that the hemoglobin genes occur in two clusters: the approximately 28 kb-long  $\alpha$ -globin locus on chromosome 16 and the approximately 45 kb-long  $\beta$ -globin locus on chromosome 11 (review Fig. 10.11). As explained previously, the five functional genes plus two pseudogenes in the  $\alpha$ -globin locus, and the five functional genes plus one pseudogene in the  $\beta$ -globin locus, all can be traced back to a single ancestral DNA sequence through multiple rounds of duplication and divergence.

Here, we show how the DNA sequence of these loci obtained from the Human Genome Project has provided



fundamental insights into the mechanisms that change globin expression during normal development from embryonic to fetal to adult forms. Furthermore, the DNA sequence of these clusters reveals how various mutations give rise to a range of globin-related disorders. Hemoglobin disorders are the most common genetic diseases in the world and include *sickle-cell anemia*, which arises from an altered  $\beta$  chain, and *thalassemia*, which results from decreases in the amount of either  $\alpha$  or  $\beta$  chain production.

### The Order of the Hemoglobin Genes in the $\alpha$ and $\beta$ Clusters Reflects the Timing of Their Expression

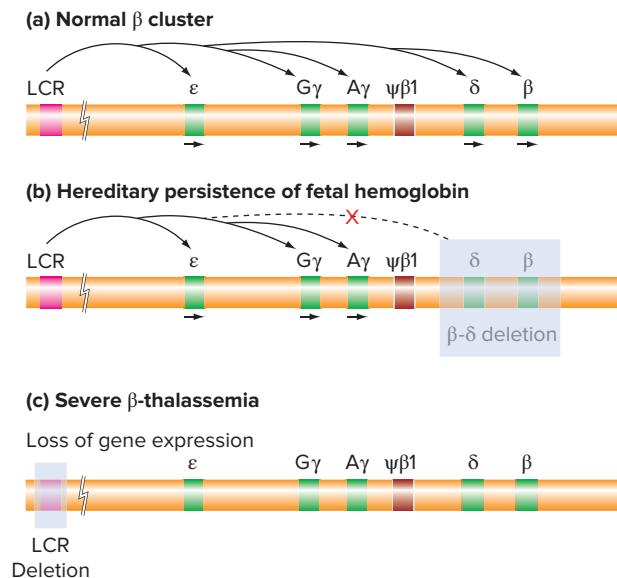
For the  $\alpha$ -like chains, the temporal order of protein expression is  $\zeta$ -globin during the first five weeks of embryonic life, followed by  $\alpha$ -globin (encoded by both the  $\alpha 1$  and  $\alpha 2$  genes) during fetal and adult life. For the  $\beta$ -like chains, the order of protein production is  $\epsilon$ -globin during the first five weeks of embryonic life; then  $\gamma$ -globin (encoded by the  $A\gamma$  and  $G\gamma$  genes) during fetal life; and finally, within a few months of birth, mostly  $\beta$  but also some  $\delta$  chains (see Fig. 10.19c).

If you compare Figs. 10.11 and 10.19c, you will note that within each cluster, the order of globin genes on the chromosomes parallels the order of their expression during development. Furthermore, all the genes in the  $\alpha$ -globin locus are oriented in the same direction relative to chromosome 16; that is, they all use the same strand of DNA as the template for transcription. The genes in the  $\beta$ -globin locus are also all oriented in the same direction, here relative to chromosome 11. The organization of these globin-gene-containing regions contrasts with most regions of the genome, where adjacent genes appear to be oriented randomly. These facts taken together suggest that whatever mechanism turns the globin genes on and off during different stages of development takes advantage of their relative positions and orientations.

We now understand what that mechanism is: Each of the two globin loci contains a **locus control region** (or **LCR**) at one end that controls sequential gene expression from that locus (Fig. 10.11). The LCR at each locus is a collection of regulatory elements called *enhancers* that are discussed in detail in Chapter 17. Through their interactions with proteins called *transcription factors*, the enhancers activate transcription of each gene in the right cells at the appropriate time.

One interesting consequence of the  $\beta$ -globin locus's organization and its control through the LCR is seen in a rare medical condition with a surprising prognosis. In some adults, the red blood cell precursors express neither the  $\beta$  nor the  $\delta$  genes. Although this should be a lethal situation, these adults remain healthy. Sequence analysis of the  $\beta$ -globin locus from affected adults shows that they

**Figure 10.20** Effects of deletions in the  $\beta$ -globin gene cluster. **(a)** Normal situation. The locus control region (LCR) sequentially turns on the transcription of the  $\epsilon$  gene in embryos; the two  $\gamma$  genes during fetal development, and the  $\beta$  and  $\delta$  genes in adults. **(b)** Loci with a deletion of the  $\beta$  and  $\delta$  genes cannot switch gene expression at birth, so the  $\gamma$  fetal polypeptides are still produced in adults. This hereditary persistence of fetal hemoglobin is benign. **(c)** Deletion of the LCR prevents the expression of all genes in the cluster, causing severe  $\beta$ -thalassemia.



have certain deletions extending across the  $\beta$  and  $\delta$  genes (compare Fig. 10.20a and b). Because of these specific deletions, the LCR can't switch, as it normally would near the time of birth, from  $\gamma$ -globin production to  $\beta$ - and  $\delta$ -globin production. People with this rare condition, called *hereditary persistence of fetal hemoglobin*, continue to produce large enough amounts of fetal  $\gamma$ -globin throughout adulthood to maintain near-normal health.

### Globin-Related Diseases Result from a Variety of Mutations

By comparing DNA sequences from affected individuals with those from healthy individuals, researchers have learned that two general classes of disorders arise from alterations in the hemoglobin genes.

In one class, mutations change the amino acid sequence and thus the three-dimensional structure of the  $\alpha$ - or  $\beta$ -globin chain. These structural changes result in an altered protein whose malfunction causes the destruction of red blood cells. Diseases of this type are known as *hemolytic anemias*. An example is sickle-cell anemia, caused by an A-to-T substitution in the sixth codon of the  $\beta$ -globin chain. This simple change in DNA sequence alters the sixth amino acid in the chain from glutamic acid to valine. Red blood cells carrying these altered molecules often have abnormal

shapes that cause them to block blood vessels or be degraded (review Fig. 7.29).

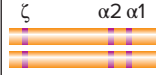

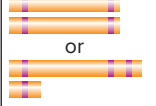


The second major class of hemoglobin-related genetic diseases arises from DNA mutations that reduce or eliminate the production of one of the two globin polypeptides. The disease state resulting from such mutations is known as *thalassemia*, from the Greek words *thalassa* meaning *sea* and *emia* meaning *blood*; the name arose from the observation that a relatively high rate of this blood disease occurs among people who live near the Mediterranean Sea. Several different types of mutation can cause thalassemia, including those that delete an entire globin gene or locus; those that alter the sequence in regions that are outside of a globin gene but necessary for its regulation; or those that alter the sequence within the gene such that no protein can be produced, such as nonsense or frameshift mutations. The consequence of these changes in DNA sequence is the total absence or a deficient amount of one or the other of the normal hemoglobin chains.

Because each  $\alpha$ -globin locus contains two  $\alpha$ -globin genes ( $\alpha 1$  and  $\alpha 2$ ), normal people have four copies of this gene. Individuals carrying deletions within the  $\alpha$ -globin locus may be missing anywhere from one to four of these copies (Fig. 10.21). A person lacking only one would be a heterozygote for the deletion of one of two  $\alpha$  genes; a person missing all four would be a homozygote for deletions of both  $\alpha$  genes. Given that  $\alpha 1$  and  $\alpha 2$  are expressed more or less equally beginning a few weeks after conception, the range of mutational possibilities explains the range of phenotypes seen in  $\alpha$ -thalassemia. Individuals missing only one of four possible copies of the  $\alpha$  genes are normal; those lacking two of the four have a mild anemia, and those without all four die before birth.

The fact that the  $\alpha 1$  and  $\alpha 2$  genes are expressed early in fetal life explains why the  $\alpha$ -thalassemias are detrimental *in utero*. By contrast,  $\beta$ -thalassemia major, the disease occurring in people who are homozygotes for most deletions of the single  $\beta$ -globin gene, also usually results in death, but not until soon after birth. These individuals survive that long because the  $\beta$ -like protein  $\delta$ -globin is expressed in the fetus (review Fig. 10.19c).

In some thalassemias, disease symptoms arise from mutations that alter the LCR found at one end of the  $\alpha$  cluster or the LCR in the  $\beta$  cluster. Deletions of either LCR can produce severe thalassemias because none of the genes in

**Figure 10.21** Thalassemias associated with deletions of genes in the  $\alpha$ -globin cluster. The fewer copies of the  $\alpha$  genes that remain, the more severe the clinical symptoms of thalassemia.

Clinical condition	Genotype	Number of functional $\alpha$ genes	$\alpha$ -chain production
Normal	 $\alpha\alpha/\alpha\alpha$	4	100%
Silent carrier	 $\alpha\alpha/\alpha-$	3	75%
Heterozygous $\alpha$ -thalassemia—mild anemia	 $\alpha-/alpha-$ or $\alpha\alpha/---$	2	50%
HbH ( $\beta_4$ ) disease—moderately severe anemia	 $\alpha-/--$	1	25%
Homozygous $\alpha$ -thalassemia—lethal	 $---/--$	0	0%

the affected cluster are correctly transcribed (see Fig. 10.20c for an example). The fact that all the globin genes in such patients are intact while DNA off to the side of one of the clusters is missing was one of the first clues to the existence of LCRs.

### essential concepts

- Embryonic and fetal forms of hemoglobin bind oxygen more tightly than does the adult form, helping to ensure that the growing embryo/fetus receives sufficient oxygen from the mother's blood.
- The sequential expression of globin genes over the course of development is regulated by *locus control regions (LCRs)* in the  $\alpha$  and  $\beta$  clusters.
- *Thalassemias* are blood diseases caused by mutations that eliminate or reduce the production of globin polypeptides from one of the clusters but not the other. These mutations can include deletions of specific genes or the LCR in either cluster.

### WHAT'S NEXT

Determination of the nucleotide sequences of the human genome and the genomes of many other species constitutes an incredible milestone in our understanding of biology. We now know the fundamental blueprints for the lives of

cells and organisms, and we have some idea of how differences in DNA lead to the emergence of different species.

However, in many ways the term “the human genome” lacks precision. People are not identical clones; instead,

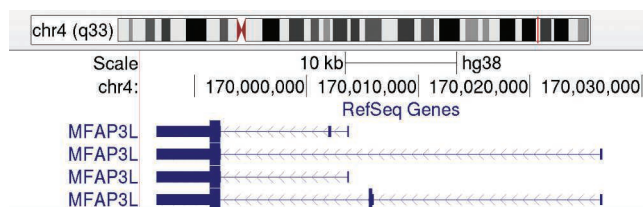
each of us has our own human genome that is closely related to that of all other humans but is also distinct and unique. It is the differences between the genomes of individuals that cause each of us to possess our own distinct and unique phenotype.

The sequence and even the annotation of one human genome is only the beginning. The human RefSeq provides a reference mark toward identifying and analyzing the

differences between the genomes of many individuals so we can understand the genetic basis of phenotypic variation; for example, finding the nucleotide differences responsible for far-ranging and varied effects on human health. In Chapter 11, we describe how geneticists can now look at the genomes of many individuals to track genetic variation and to identify those differences in DNA sequence that underlie important traits.

## SOLVED PROBLEMS

- I. The following figure shows a screen shot from the UCSC Genome Browser, focusing on a region of the human genome encoding a gene called *MFAP3L*. (Note: *hg38* refers to version 38 of the human genome RefSeq.) If you do not remember how the browser represents the genome, refer to the key at the bottom of Fig. 10.3.



Source: University of California Genome Project, <https://genome.ucsc.edu>

- Describe in approximate terms the genomic location of *MFAP3L*.
  - Is the gene transcribed in the direction from the centromere-to-telomere or from the telomere-to-centromere?
  - How many alternative splice forms of *MFAP3L* mRNA are indicated by the data?
  - How many different promoters for *MFAP3L* are suggested by the data?
  - How many different proteins does the *MFAP3L* gene appear to encode? Which alternatively spliced forms of the mRNA encode which proteins? Do the different forms vary at their N termini, their C termini, or somewhere in the middle? Estimate how many amino acids each of these proteins contains.
- The arrows within the introns of the gene show that the direction of transcription is from the telomere of 4q toward the centromere of chromosome 4.
  - The data indicate four alternatively spliced forms of the mRNA. In the following parts, we list these as A to D from top to bottom.
  - The data suggest two promoters. One is roughly at position 170,037,000 and allows the transcription of a primary RNA alternatively spliced to produce mRNAs B and D. The other is roughly at position 170,013,000 and leads to the transcription of a primary RNA alternatively spliced to generate mRNAs A and C.
  - The data indicate that the *MFAP3* gene can encode two different but closely related proteins. mRNAs A, B, and C all encode the same protein; mRNA D a slightly larger protein that includes at its N terminus additional amino acids not found in the other protein. Otherwise these two proteins appear to be the same. The ORF that encodes the A B C protein form is about 880 bp long (a rough estimate); this corresponds to about  $(880/3 = 293)$  amino acids. The D protein is about 50 amino acids longer.

### Answer

- The gene is located on the long (q) arm of human chromosome 4; this position is denoted by the thin red vertical line on the chromosome representation (an *idiogram*) at the top of the figure. This location (in a band called 4q33) is roughly 170 million bp from the telomere of the small arm of chromosome 4 (from where the numbering begins); the total length of this chromosome is about 190 million bp.
- If this couple has many conceptions, what percentage of these conceptions is expected to result in hydrops fetalis?
- Two other parents, who come from North Africa, also both have  $\alpha$ -thalassemia trait, but a genetics counselor told them that none of their conceptions would result in hydrops fetalis. Explain how the genetic counselor's advice could be correct.

**Answer**

- a. Both of these parents from Southeast Asia must have been heterozygotes with one normal copy of the *HbA* locus and the other copy deleted for both *Hba1* and *Hba2*. One fourth of their conceptions would lead to hydrops fetalis (homozygosity for the deletion), one half to  $\alpha$ -thalassemia ( $Hba1^+ Hba2^+ / - -$ ), one fourth to a normal situation ( $Hba1^+ Hba2^+ / Hba1^+ Hba2^+$ ).

- b. These parents from North Africa both were homozygotes for a deletion of one of the two *Hba* genes. Thus, all of their children would have  $\alpha$ -thalassemia (two *Hba* genes instead of the normal four), and none of their conceptions would lead to hydrops fetalis. Note that the genetic counselor could have given this advice only if the counselor knew from analysis of the parents' genomes what kind of defects were present in their *HbA* loci.


**PROBLEMS**
**Vocabulary**

1. Match each of the terms in the left column to the best-fitting phrase from the right column.

- |                             |   |
|-----------------------------|---|
| a. exome                    | 1. a discrete part of a protein that provides a unit of function  |
| b. de novo gene             | 2. a nonfunctional member of a gene family  |
| c. gene desert              | 3. the joining together of exons in a gene in different combinations  |
| d. pseudogene               | 4. most frequent residues, either nucleotide or amino acid, found at each position in a sequence alignment  |
| e. syntenic block           | 5. set of genes related by processes of duplication and divergence  |
| f. orthologs                | 6. chromosomal region with the same genes in the same order in two different species                        |
| g. natural selection        | 7. genes with sequence similarities in two different species that arose from a common ancestral gene        |
| h. consensus sequence       | 8. genes that arose by duplication within a species   |
| i. gene family              | 9. genomic DNA sequences containing exons   |
| j. paralogs                 | 10. gene-poor region of the genome  |
| k. alternative RNA splicing | 11. recently evolved from intergenic DNA sequences  |
| l. protein domain           | 12. progressive elimination of individuals whose fitness is low and survival of individuals of high fitness |

**Section 10.1**

2. List three independent techniques you could use to identify DNA sequences encoding human genes within a cloned genomic region.
3. Figure 10.2a has numbers indicating the approximate number of millions of years ago that species on separate branches of the tree last shared a common ancestor.
- a. About how many millions of years ago did humans last share a common ancestor with chimpanzees, mice, dogs, chickens, and frogs?

These estimates for evolutionary events were obtained in part by comparing the genomic sequences of various current-day species. The basic supposition behind these estimates is that of a *molecular clock*: Differences in particular types of genomic sequences accumulate at a relatively linear rate during evolutionary time. Consider the three following kinds of nucleotide changes: (i) missense mutations in coding regions that alter amino acid identity; (ii) silent (synonymous) mutations that change a codon for a particular amino acid into a different codon for the same amino acid; and (iii) mutations in introns. Which of the three types of mutations would . . .

- b. . . represent the slowest-ticking clock? (That is, which type of mutation would accumulate the least rapidly in genomes? *Hint*: See Fig. 10.3.)
- c. . . you most likely use to estimate the divergence times of species that last shared a common ancestor more than 400 million years ago?
- d. . . be most likely to vary in the rate at which they would accumulate in different genes?
4. Which of the enzymes from the following list would you need to make a cDNA library? What is the function of those enzyme(s) in the process?
- DNA polymerase
  - RNA polymerase
  - A restriction enzyme
  - DNA ligase
  - An aminoacyl-tRNA synthetase
  - Peptidyl transferase
  - Reverse transcriptase
5. One of the following sequences was obtained from a cloned piece of a genome that includes parts of two exons of a gene. The other sequence was obtained from the corresponding part of a cDNA clone representing the mRNA for this gene. (*Note*: For

simplicity, the intron is unrealistically short; some but not all sequence features needed for splicing are present.)

Sequence 1:  
5' TAGGTGAAAGAGTAGCCTAGAAATCAGTTA 3'

Sequence 2:  
5' TAACTGATTCTTTCACCTA 3'

- a. Which sequence is the genomic fragment and which is the cDNA fragment?
  - b. Write the RNA-like strand of the genomic sequence and indicate the 5' and 3' ends. Draw vertical lines between the bases that are the exon/intron boundaries. (Refer to Fig. 8.15 for splice junction sequences.)
  - c. What sequence features needed for splicing are missing from this problem?
  - d. Assuming both exons are made only of protein-coding nucleotide sequences, what can you determine about the amino acid sequence of the protein product of the gene? (Indicate the N-to-C orientation.)
6. a. What sequence information about a gene is lacking in a cDNA library?
  - b. Can clones in a cDNA library contain 5' UTR sequences? 3' UTR sequences?
  - c. Would you be likely to find on average longer ORFs in cloned sequences from a genomic library or from a cDNA library? Explain.
7. Why do geneticists studying eukaryotic organisms often construct cDNA libraries, whereas geneticists studying bacteria almost never do? Why might bacterial geneticists have difficulties constructing cDNA libraries even if they wanted to?
  8. Consider three different kinds of human libraries: a genomic library, a brain cDNA library, and a liver cDNA library.
    - a. Suppose that all three of these libraries are sufficiently large so as to represent all of the different human nucleotide sequences that the library could possibly include. Which of these libraries would then correspond to the largest fraction of the total human genome?
    - b. Would you expect any of these libraries not to overlap the others at all in terms of the sequences it contains? Explain.
    - c. How do these three libraries differ in terms of the starting material for constructing the clones in the library?
    - d. Why would you need to sequence many clones from many cDNA libraries to annotate a genome?
  9. The human genome has been sequenced, but we still don't have an accurate count of the number of genes. Why not?
  10. This problem investigates issues encountered in sequencing the inserts in cDNA libraries.
    - a. If you sequenced many clones individually, wouldn't you spend many of your resources inefficiently sequencing cDNAs for the same type of mRNA molecule over and over again? Explain. Does this apparently inefficient process provide any useful information beyond the sequences of individual mRNAs?
    - b. Suppose that you identified a clone with a cDNA insert that was 4 kb long. You could determine the entire sequence of the clone by shearing the DNA into small random fragments, cloning these fragments into a vector to make a mini-shotgun library, and then sequencing hundreds of these clones to allow the computer to assemble the full sequence of the 4 kb-long insert. However, this procedure would be inefficient.
 

An alternative that requires many fewer sequencing reactions is called *primer walking*. This technique involves the synthesis of additional oligonucleotide primers corresponding to cDNA sequences you have just obtained. Diagram how you would sequence the entire 4 kb-long cloned cDNA using primer walking, indicating the vector and insert, all primers that you would use, and all the sequences you would obtain. Assume that each sequence read is 1 kb.
  11. For the sake of simplicity, Fig. 10.4 omitted one step of cDNA library construction. The figure implied that the last step of the process is the ligation of blunt-ended cDNAs into plasmid cloning vectors. Although such ligation reactions can occur, in reality they are highly inefficient. Instead, scientists convert blunt-ended cDNA molecules into sticky-ended molecules using *adapters*, and then they ligate the cDNAs into vectors with compatible sticky ends.
 

Adapters are short, partly double-stranded DNA molecules made by hybridization of two single-stranded oligonucleotides made in a DNA synthesizer. Suppose that the following two oligonucleotides were synthesized and then mixed together at high concentration and at a temperature that promotes hybridization of complementary DNA sequences:

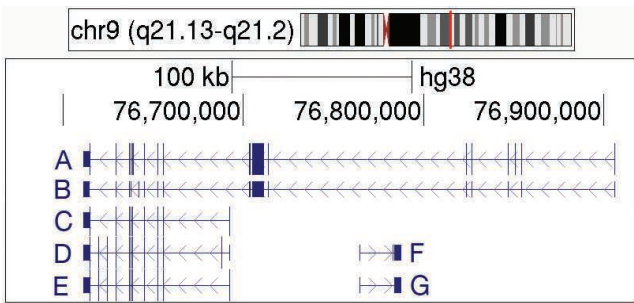
5' CCCCCG 3'  
5' AATTCGGGG 3'

    - a. Draw the hybridized DNA molecules. These are the adapters.
    - b. Suppose you added the adapters and ligase enzyme to blunt-ended cDNAs at a very high molar ratio of adapters to cDNAs, so that each cDNA molecule is ligated to one adapter at each of its ends. Draw a picture of a resulting cDNA molecule.

- c. The particular adapters discussed in this problem allow the cDNAs to be ligated efficiently into a vector treated with a commonly used restriction enzyme listed in Table 9.1. Name this restriction enzyme.

### Section 10.2

12. Give two different reasons for the much higher ratio of total DNA to protein-encoding DNA in the human genome as compared to bacterial genomes.
13. Using a cDNA library, you isolated two different cDNA clones that have sequences indicating that they both correspond to mRNAs transcribed from the same nerve growth factor gene. The beginning and ending sequences of the clones are the same, but the middle sequence is different. How can you explain the different cDNAs?
14. The figure that follows shows part of a modified screen shot of part of the human genome as displayed on the UCSC Genome Browser. A through G map the sequences within individual cDNA clones to the genome sequence. Refer to the key in Fig. 10.3 if you need help in interpreting this diagram. Pay close attention to the vertical widths of the icons indicating exons.



Source: University of California Genome Project, <https://genome.ucsc.edu>

- a. How many annotated genes do you think are present in this region of the human genome?
- b. For all annotated genes in this region, indicate whether they are transcribed in the direction from centromere to telomere or from telomere to centromere.
- c. How many promoters are suggested by the data? Approximately where are these promoters located?
- d. How many different proteins are encoded by the DNA sequences in this region?
- e. What is unusual about this region of the human genome?
15. In Problem 14, cDNAs F and G could not be found in cDNA libraries (from any tissue) prepared using the method shown in Fig. 10.4. The reason is that the corresponding transcripts do not have poly-A tails.

- a. Why is the lack of poly-A tails not surprising in light of your answer to part (d) of Problem 14?
- b. Why does the lack of poly-A tails present a difficulty for the method diagrammed in Fig. 10.4?
- c. Outline how you might adjust the protocol in Fig. 10.4 so as to find the cDNAs F and G annotated in the Genome Browser.
16. Fig. 10.10 presents a model for exon shuffling in which chromosomal fragments broken into introns can be restitched together to make novel genes that did not exist before. However, only a certain fraction of events like those shown on the figure could actually produce genes encoding proteins with functional domains from both original polypeptides. What is this fraction?
17. An interesting phenomenon found in vertebrate DNA is the existence of *pseudogenes*, nonfunctional copies of a gene found elsewhere in the genome. Some pseudogenes appear to have originated as double-stranded DNA copies of mature mRNA inserted into the chromosome; these copies later underwent mutations to make them into pseudogenes.
- a. What sequence information might provide clues that the original source of some of these pseudogenes is cDNA copied in cells from mRNA and then inserted into the genome?
- b. Would this mechanism of generating pseudogenes be more likely to have operated if the pseudogene was part of a gene family clustered in one region of the genome, or if it was instead part of a gene family whose members are scattered around the genome? Explain.
18. a. If you found a *zinc-finger domain* (which facilitates DNA binding) in a newly identified gene, what kinds of hypotheses could you make about the gene's function?
- b. Suppose that this newly identified gene shares a high percentage of similarity throughout its length with a previously characterized gene in the same organism. What does this fact suggest about the origin of the two genes? Would you categorize these genes as being: (i) homologous, (ii) paralogous, or (iii) orthologous? (More than one answer may apply.)
19. You sequence the genomes of four different organisms and compare their sequences over a short region as shown below.

```

5' AGGTATATAATTTGCG 3'
5' CAATATAAACCTAC 3'
5' GCGTATAAAGAGCTA 3'
5' TTATATATAAGAAGT 3'

```

- a. Determine the *consensus sequence* common to the four regions above.

- b. Why would you want to define the consensus sequence? How would you decide whether the four sequences were worth comparing to define a consensus?
- c. How could you use this general strategy for defining a consensus sequence to determine which amino acids of a protein are most crucial for its function?
20. In the human immune system, so-called *B cells* can make more than a billion different types of antibody molecules that protect us from infection. However, our genomes have only three genes that encode the polypeptides found in antibodies. What experiments could you perform to determine what kind of combinatorial events occur at the DNA level (V-D-J joining) and RNA level (alternative splicing) for any of these genes?
21. Chimpanzees have a set of hemoglobin genes very similar to the set in humans that was shown in Fig. 10.11. For example, the genomes of both species have  $\alpha 1$ ,  $\alpha 2$ ,  $\beta$ ,  $\text{G}\gamma$ ,  $\text{A}\gamma$ ,  $\delta$ ,  $\epsilon$ , and  $\zeta$  genes.
- Of the human and chimpanzee hemoglobin genes, which would be considered homologous? Which paralogous? Which orthologous?
  - When comparing genomes, geneticists would usually like to know which genes are the most likely to perform similar if not identical functions in different species. This determination can be somewhat complicated in the case of gene families. Would paralogous genes or orthologous genes be more likely to be functionally equivalent? Explain.
  - Which gene would have the greatest degree of nucleotide similarity to the human  $\beta$  gene: the chimpanzee  $\beta$  gene, or the human  $\gamma$  gene? Explain.
  - Rationalize the pattern of hemoglobin genes in the two species with the existence of duplication and divergence events among the hemoglobin genes depicted in Fig. 10.12.
22. Complete genome sequences indicate that the human genome has roughly 27,000 genes, while the worm (nematode) genome has about 22,000 genes. Explain how the human genome with only about 20% more genes can encode a creature enormously more complex than the worm.
- to the proper window.) At the top of this window are control buttons that allow you to move your view to the left or right, zoom in (even to the level of the nucleotide sequence), zoom out, or (on the second row) jump to a different chromosomal position. Below these buttons is a diagram, called an *idiogram*, of the chromosome you are viewing, with a region in red indicating the particular region of the chromosome you are looking at. (You can also click on the idiogram to move around.)
- How many exons are in the *CFTR* gene?
  - Is the *CFTR* gene located on the short arm (the *p* arm) or the long arm (the *q* arm) of human chromosome 7?
  - In which direction is the *CFTR* gene transcribed: toward the centromere, or away from the centromere?
- Now zoom out the view by 10 $\times$ .
- What are the names of the genes that flank *CFTR* on either side? Are these genes transcribed from the same strand of chromosome 7 as *CFTR*, or from the other strand?
- Now zoom out 100 $\times$  until the entirety of chromosome 7 is visible.
- What is the approximate size of chromosome 7 in Mb?
  - What is the approximate location of the centromere on human chromosome 7?
  - What is the significance of the RefSeq genes' appearing to pile up when you are viewing the whole chromosome?
24. On your computer's browser, view the page accessed by the URL: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
- The heading "Basic BLAST" lists various programs that allow you to search for DNA or protein sequences related to any DNA or protein sequence query. For this problem, choose "nucleotide blast." On the window that comes up, make sure that the database "Human genomic + transcript" is selected (so that you will search through the RefSeq for humans and not any other species). Now, in the large box under "Enter Query Sequence" you should type in the nucleotide sequence from Fig. 9.7f. Then hit the blue "BLAST" button at the bottom and wait for the response from the NCBI computer, which may take up to a few minutes.
- The response will come in several sections; you should migrate to the Descriptions. The "E value" column in the list is a statistical measure of the probability that a given sequence is related to your query by chance in the RefSeq database you are searching. The lower the E value (the closer to zero), the more

### Section 10.3

23. On your computer's browser, view the page accessed by the URL: <http://genome.ucsc.edu/cgi-bin/hgGateway>
- In the Search Term box at the top, type *CFTR* (for the *CFTR* gene responsible for cystic fibrosis), then hit "GO." You will be directed to a window showing the organization of the *CFTR* gene on human chromosome 7. (If a list appears instead of a picture, click on the first link at the top of the list, and you will be directed

significant is the match. For this exercise, look only at the first entry in the list, with the lowest E value.

a. What is the human gene that best matches the query sequence?

Now migrate down on the same page to the list of Alignments. Look only at the first entry, which corresponds to the first entry under the Descriptions.

b. Is the match exact or is it imperfect?

Go back to the nucleotide blast page. Using the same query sequence (from Fig. 9.7f), now choose “mouse genomic + transcript” as the database to search.

c. Is this gene conserved in the mouse? If so, how well conserved?

25. Use the UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>) to identify the two major genes illustrated in the figure accompanying Problem 10.14.

#### Section 10.4

26. Certain individuals with mild forms of  $\beta$ -thalassemia produce, in addition to normal adult hemoglobin with two  $\alpha$  chains and two  $\beta$  chains, lower levels of an unusual, so-called *Lepore hemoglobin* with two  $\alpha$  chains and two chains in each of which the N-terminal half comes from a normal  $\delta$  chain and the C-terminal half comes from a normal  $\beta$  chain. Certain other individuals who are asymptomatic produce a different, unusual *anti-Lepore hemoglobin* that contains two  $\alpha$  chains and two chains in which the N-terminal half comes from a normal  $\beta$  chain and the C-terminal half comes from a normal  $\delta$  chain.

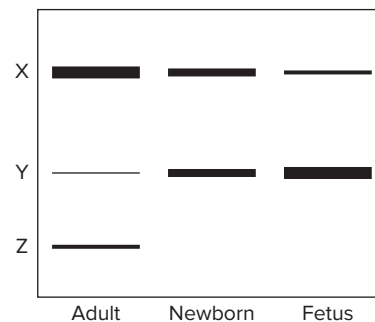
a. Describe an event that could give rise to both Lepore and anti-Lepore hemoglobins.

b. Are the mildly thalassemic individuals with Lepore hemoglobin homozygotes or heterozygotes for the unusual allele?

c. Why might these mildly thalassemic people produce less Lepore hemoglobin than normal adult hemoglobin?

27. The  $\alpha 1$  and  $\alpha 2$  genes in humans are identical in their coding regions. With this fact and your answer to Problem 26 in mind, describe a mechanism that might frequently produce a deletion of one of these two genes.

28. The following figure shows an electrophoretic analysis of the hemoglobin proteins present in normal adults, newborns, and fetuses. Each band represents a complete hemoglobin protein with all of its subunits. The intensity of the band indicates the relative amount of that protein in the sample.

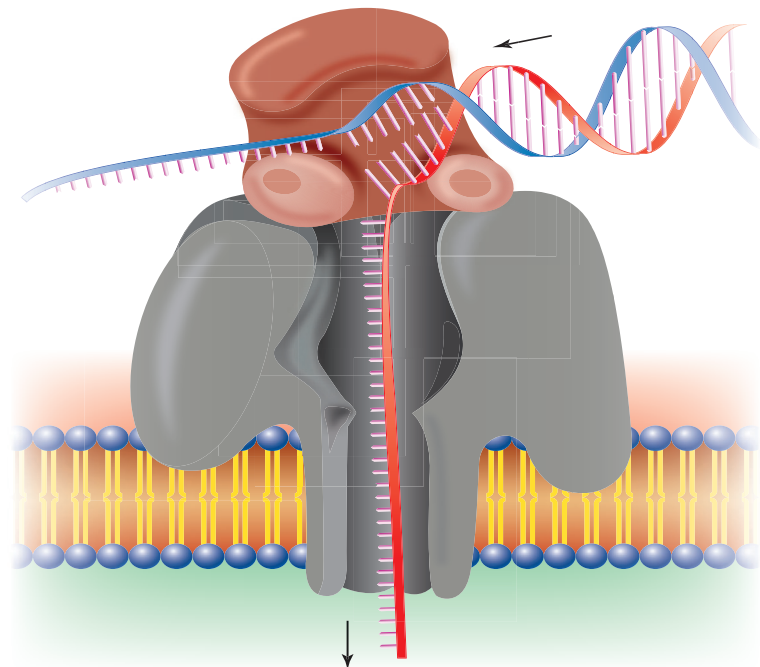


- a. What is the subunit structure of the hemoglobin molecules marked X, Y, and Z?
- b. Name one abnormal condition that should *increase* the percentage of hemoglobin in a newborn that is HbZ.
- c. Name one abnormal condition that should *increase* the percentage of hemoglobin in an adult that is HbY.



chapter **11**

## Analyzing Genomic Variation



The era of whole-genome sequencing is being made possible by remarkable innovations. One novel technique for massively parallel DNA sequencing uses an enzyme (brown) to thread a DNA strand through a narrow channel called a micropore (gray). The DNA in the channel limits (in a nucleotide-specific fashion) the flow of ions through the micropore. Recordings of the current flowing through the micropore as a function of time can be interpreted as the sequence of nucleotides.

**MILLIONS OF PEOPLE** worldwide suffer from a group of conditions, including ulcerative colitis and Crohn's disease, featuring chronic inflammations of the digestive tract (**Fig. 11.1**). The patients' immune systems overreact to bacteria in the gut and begin to attack cells in the intestinal mucosa (gut lining). The symptoms are usually not life threatening, although flare-ups of intense abdominal pain, vomiting, diarrhea, nausea, and fatigue can completely disrupt a person's life. Variations in more than 100 genes have been found to predispose individuals to inflammatory bowel diseases (IBDs).

At the age of 2, Nic Volker acquired an IBD of unprecedented severity. His digestive tract developed lesions that extended through his body cavity all the way to the outside of his skin. As a result, fecal matter leached into his system, causing dangerous sepsis (systemic bacterial infections). Unfortunately, Nic's condition did not respond to the usual treatments for IBD, such as immunosuppressants or anti-inflammatory steroids. By the age of 4, Nic had already undergone more than 140 surgeries to resect parts of his digestive tract and heal the wounds in his skin; he weighed only 17 pounds (**Fig. 11.2**). Nic's long-term prognosis was obviously dire.

Nic's parents and doctors enlisted a team of human geneticists to attempt a novel approach for his desperate case: determining the sequence of all the protein-coding nucleotides in Nic's genome. Remarkably, in 2009 the research team found the mutation that caused Nic's disease. The mutation was located in a gene known to be associated with another inherited condition called X-linked lymphoproliferative disease (XLPD). The symptoms of XLPD were so different from Nic's that no one had ever guessed at the connection. Blood marrow transplantation was known to be effective for XLPD, so Nic's doctors decided to try this method, even though IBDs had never before been treated in this way. Within a few months of the transplant, Nic's health underwent an astonishing rebound, allowing him to live the normal life of a six-year-old (**Fig. 11.3**).

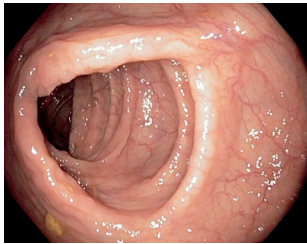
**chapter outline**

- 11.1 Variation Among Genomes
- 11.2 Genotyping a Known Disease-Causing Mutation
- 11.3 Sampling DNA Variation in a Genome
- 11.4 Positional Cloning
- 11.5 The Era of Whole-Genome Sequencing

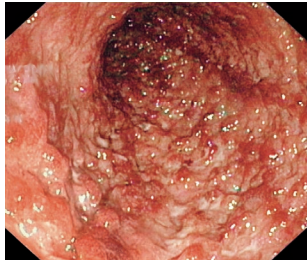
**Figure 11.1 Normal and Crohn's syndrome bowels.**

Colonoscopy (endoscopic examination of the colon using a fiber optic camera) of a normal person (*left*) and of a patient suffering from Crohn's disease (*right*).

© Gastrolab/Science Source



(a) Normal colon lining



(b) Crohn's disease

This case history illustrates one of the far-reaching medical consequences of our rapidly increasing power to detect genotype directly at the DNA level. Until very recently, scientists were more limited in their ability to look at human genomes. In the 1990s, researchers could examine an individual's genotype only a single gene at a time, and this was worthwhile only in the few cases where the disease-causing mutation was already identified. Even this limited amount of personal genetic information was often valuable in helping couples to make informed reproductive decisions.

By the turn of the twenty-first century, advances in genotyping allowed scientists to look at a much larger sample of the many nucleotide changes that differentiate one person's genome from another. For example, new methods including *DNA microarrays* allowed simultaneous examination of dozens or even millions of nucleotide variations at different positions, or **loci**, within individual genomes. As you will see, the ability to follow large numbers of nucleotide variations has many uses, even if these variations themselves have nothing to do with disease.

By 2013, scientists had the ability to genotype not just one or two or thousands of loci in a person's genome, but nearly all of the 6 billion nucleotides in a person's diploid genome. Costs are being driven down rapidly by new innovations in DNA sequencing technologies that will soon become a routine part of medical care. In this new and uncharted era of whole-genome DNA sequencing, the exponential increase in our knowledge of genome variations will provide details about people's genetic histories and destinies never before available.

**Figure 11.2 Nic Volker at age 4.**

© Gary Porter/Milwaukee Journal Sentinel/MCT/Newscom

**Figure 11.3 Nic Volker at age 6.**

© Andy Manis/Bloomberg via Getty Images



## 11.1 Variation Among Genomes

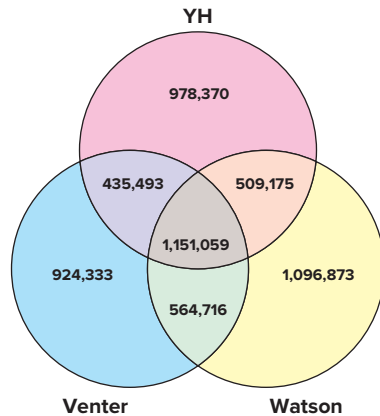
### learning objectives

1. Cite the approximate number of DNA polymorphisms that differentiate any two haploid human genomes.
2. Explain why most of these DNA polymorphisms are not responsible for the phenotypic differences between people.
3. Differentiate among different classes of DNA variants in terms of their structures, mechanisms of formation, and frequency in genomes.

There is no such thing as a wild-type human genome; instead, a staggering amount of variation exists between the genomes of any two people. With the advent of new technologies for whole-genome sequencing that will be described later in this chapter, the degree to which individual human genomes differ from each other is becoming apparent. Only a small minority of these DNA sequence variations is responsible for the phenotypic differences that characterize individuals. But even if certain DNA sequence differences have no effect on phenotypes, they are still highly useful as markers to track genes and chromosomes.

**Figure 11.4** Comparison of three personal genomes.

Single nucleotide substitutions in the genomes of J. Craig Venter, James D. Watson, and an anonymous Chinese man (YH), all relative to the human RefSeq. A substitution is counted once whether the individual is homozygous or heterozygous for that variant. Numbers of substitutions unique to each man's genome are in nonoverlapping portions of each circle. Variants not in the human RefSeq but shared by two of the three individuals are shown in the double overlap regions. The central three-way overlap indicates variants shared by all three men.



### Extensive DNA Variation Distinguishes Individuals Within a Species

The genomes of James Watson, co-discoverer of the DNA double helix; J. Craig Venter, a pioneer of DNA sequencing; and an anonymous Chinese man reveal in total more than 5.6 million single nucleotide differences from the standard human genome (the GenBank RefSeq; see Chapter 10) (Fig. 11.4). Each man's diploid genome contains about 1 million unique **DNA polymorphisms** (that is, sequence differences) not shared by either of the other men, while the remaining approximately 2.6 million polymorphisms are shared in the genomes of two or in some cases all three of these individuals.

Not only does no single wild-type human genome sequence exist, there is even no such thing as a wild-type human genome length. Deletions, insertions, and duplications of DNA result in genome lengths that differ by as

much as 1% in healthy people. For example, the genomes of Watson and Venter vary by small additions or subtractions of genetic material—insertions or deletions—at over 100,000 genomic sites.

### Most DNA Polymorphisms Do Not Influence Phenotype

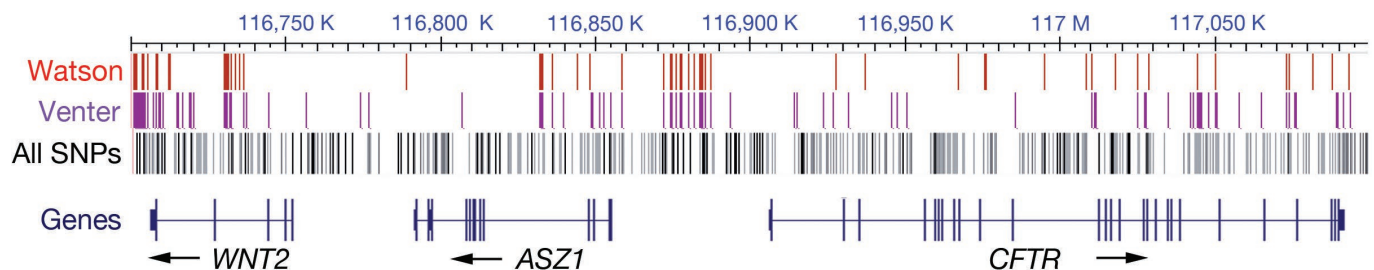
Some of the millions of DNA polymorphisms between the genomes of Watson and Venter must be responsible for the phenotypic differences that distinguish them as individuals. But in reality only a small fraction of these DNA sequence changes actually impacts phenotype. Only about 5000 of the millions of differences between these two people alter the amino acid sequences of proteins. This fact makes sense because:

- (1) less than 2% of the human genome consists of codons within genes;
- (2) even when they occur, many mutations of codons are silent (that is, they don't change the amino acid); and
- (3) if a particular mutation is not silent and has deleterious effects, natural selection could often lead to its disappearance from the human population.

In addition to the approximately 5000 amino-acid-altering mutations, a few thousand other polymorphisms between these two genomes likely affect gene expression, for example the frequency of transcription or the efficiency of primary transcript splicing to produce mRNA. But even after accounting for these, we are left with the conclusion that the vast majority of sequence differences between genomes are **anonymous DNA polymorphisms** affecting neither the nature nor the amounts of any protein in the body. (You will see later that **nonanonymous DNA polymorphisms** do affect gene expression, and thus can affect phenotype.)

**Figure 11.5** shows the actual distribution of polymorphisms that distinguish Watson and Venter from the human RefSeq within a 400 kb genomic region. This part of the genome includes the cystic fibrosis transmembrane receptor gene (*CFTR*), mutations in which cause cystic fibrosis, and two other genes. You can see that almost all of the

**Figure 11.5** SNP distribution in a 400 kb region. This part of chromosome 7 (from base pairs 116,700,001 to 117,100,000) contains *CFTR* and two other genes. Vertical marks indicate locations at which a genome is either heterozygous or homozygous for a single nucleotide polymorphism (SNP) different from the human RefSeq. Two rows show SNPs that were read from the personal genomes of Watson and Venter. The third track compiles all SNPs from all human genomes analyzed that were deposited in the central SNP database as of 2009.



polymorphisms in both men’s genomes, and indeed in all genomes, are located either between genes or in introns, consistent with the idea that most DNA changes do not alter phenotype.

The existence of such a vast number of anonymous DNA polymorphisms that distinguish different human genomes presents both challenges and opportunities for geneticists. The challenges are clear: How can we sort out the millions of polymorphisms in anyone’s DNA to find the few that are relevant to traits such as genetic diseases? The opportunities lie in the fact that even if a polymorphism is anonymous, with no effect on phenotype, it still serves as a signpost in the genome, a **DNA marker**. You will see that researchers can use anonymous polymorphic loci to help locate the nearby mutations that are actually responsible for inherited diseases.

### Genetic Variants Occur in Several Types

Geneticists usually place polymorphic DNA loci into one of the four categories shown in **Table 11.1** based on the number and kinds of nucleotide pairs involved. Although the borders between these classes are fuzzy and overlap to some extent, the categories help researchers describe what a particular genetic variant looks like. A useful generalization is that the smaller the number of nucleotide pairs involved in a given class of polymorphism, the more frequently variants of that class are found in the genome (Table 11.1).

### Single nucleotide polymorphisms (SNPs)

Far and away the most common type of genetic variant is the class of **single nucleotide polymorphisms**, or **SNPs** (pronounced *snips*). SNPs are particular base positions in the genome where alternative letters of the DNA alphabet distinguish some people from others. SNPs account for the vast majority of the total variation that exists between human genomes, occurring on average once every 1000 bases in any pairwise comparison (Table 11.1). Chapter 7 discussed

TABLE 11.1 Categories of Genetic Variants		
	Size	Frequency (1 per...)
<b>SNP</b> Single nucleotide polymorphism	1 bp	1 kb
<b>DIP or InDel</b> Insertion/deletion	1–100 bp	10 kb
<b>SSR</b> Simple sequence repeat	1–10 bp repeat unit	30 kb
<b>CNV</b> Copy number variant	10 bp–1 Mb	3 Mb

The right column shows how frequently on average you would find a polymorphism of the indicated class when comparing any two haploid human genomes.

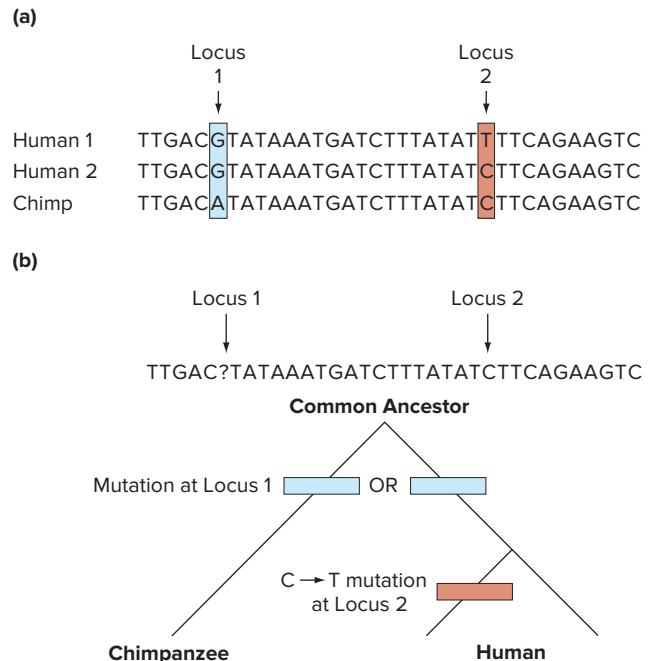
several of the many mechanisms that can give rise to SNPs, including rare mistakes in DNA replication, or exposure of the genome to mutagenic chemicals or radiation in the environment.

Despite the existence of so many types of events that can create SNPs, the per-base spontaneous mutation rate is still less than one in 30 million (by some estimates as low as one in 100 million) per generation. This number is so low that most SNPs are biallelic in human populations, with only two of the four possible nucleotide pairs represented. The low mutation rate of SNPs allows researchers to trace back each individual SNP to a genomic change that occurred once in a single ancestral genome. The low mutation frequency also means that those people who did not inherit this changed nucleotide (called the **derived allele**) have a more ancient **ancestral allele** that was probably present long before the human species took form.

If a SNP exists, geneticists can take advantage of the close relationship between the human and chimpanzee genomes to determine which allele is ancient and which is the derived allele resulting from a relatively recent mutational event. **Figure 11.6** compares a small region of

**Figure 11.6** Inferring the evolutionary history of SNPs.

(a) A comparison of two human genomic sequences to the chimp RefSeq. Loci 1 and 2 are invariant in many sequenced chimp genomes. (b) Cladogram (diagram of evolutionary lineages). Locus 1 (light blue) differs between chimps and humans, but all humans have the same allele (G). The mutation causing the locus 1 difference must have occurred since the species diverged, either in the lineage leading to chimps or in that leading to humans; the allele at this position in the most recent common ancestor of the two species cannot be determined. At locus 2 (red), the C allele shared between chimps and some humans must be *ancestral*, while the T allele in other humans must be *derived* (that is, caused by a recent mutation specifically in the lineage of some humans).



DNA in two different haploid human genomes and the chimp RefSeq genome. Two single-base changes have occurred in this small genomic region since the divergence of the two species. One is shared by all human genomes and is thus not polymorphic in humans. The second base change was from a C in the common chimpanzee-human ancestor to a T (the derived allele) in a chromosome of the ancestor of some people but not others. This means that if you and a friend share a derived allele at an anonymous SNP locus, you both got that allele from the same ancestor who must have lived since the human and chimp lineages diverged from each other. The fact that every random pair of human beings on the planet shares many unlinked, derived SNP alleles indicates common ancestry for all people.

To date, the analysis of thousands of human genomes has led to the identification of more than 50 million SNPs that are catalogued in a SNP database (dbSNP) at the National Center for Biotechnology Information (NCBI). About 15 million of these SNPs are commonly found in human populations. The mutations giving rise to the derived allele of common SNPs must have occurred far enough back in human history to have been disseminated to a significant proportion of current-day people. The SNP database already includes a large fraction of all common SNPs in human populations. However, you should realize that mutational events in the very recent past also create rare SNPs that would be found in only one or a few people of the billions on the earth. Very few of these rare SNPs are yet accounted for in dbSNP because so few human genomes have been analyzed relative to the entire human population.

This brief discussion about the origin of SNPs suggests that genome sequencing provides powerful tools for understanding human ancestry. Chapter 21 presents some of the surprising findings about human history revealed by the analysis of DNA sequences from present-day humans and the fossilized remains of our primate ancestors.

### Deletion-insertion polymorphisms (DIPs)

Short insertions or deletions of genetic material represent the second most common form of genetic variation in the human genome. These variants are referred to as **deletion-insertion polymorphisms (DIPs)** or **InDels**. While SNP loci occur about once per kilobase in the comparison of any two haploid genomes, DIPs are considerably rarer, occurring about once in every 10 kb of DNA (Table 11.1). DIPs range in length from one base pair to hundreds of base pairs, but their relative frequency declines steeply in relation to their length. As a result, DIPs involving only one or two nucleotides are the most common.

Several biochemical processes appear to contribute to the formation of DIPs. These include problems in DNA replication or recombination, and mistakes that occur when cells try to repair damage such as broken DNA strands.

You should remember that in protein-coding regions, DIP variants act as frameshift mutations unless the number of nucleotide pairs inserted or deleted is 3 or a multiple of 3.

### Simple sequence repeats (SSRs)

The genomes of humans and higher eukaryotes are loaded with loci defined by **simple sequence repeats (SSRs)**, sometimes also termed **microsatellites**. SSR loci consist of sequences of one to a few bases that are repeated in tandem less than 10 to more than 100 times. Different alleles of an SSR locus have different numbers of repeating units. The most common repeating units are one-, two-, or three-base sequences. SSRs with larger repeating units are less frequent, and we employ here a relatively arbitrary cutoff in which the largest repeating unit of an SSR has 10 bases (Table 11.1; those with larger repeating units will be classified as *CNVs* below). Examples of SSRs are AAAAAAAAAAAAAAAAAA (a one-base repeat) or CACACACACACACACACACA (a two-base repeating unit). SSRs of all types together account for about 3% of the total DNA in the human genome; an SSR locus can be found on average once in every 30 kb of human DNA.

As with all other polymorphisms, most SSRs occur outside the coding regions of genes and have no effect on phenotype. In contrast, SSR variations within genes can have profound phenotypic consequences. For example, we have already discussed in Chapter 7 (review the Fast Forward Box entitled *Trinucleotide Repeat Diseases: Huntington Disease and Fragile X Syndrome*) that long tracts of trinucleotide repeats are the molecular cause of several severe neurological conditions, including fragile X syndrome and Huntington disease.

SSRs arise spontaneously from rare, random events that initially produce a short repeated sequence with four to five repeat units. Once a short SSR mutates into existence, however, it can expand into a longer sequence by a form of faulty DNA replication called *slipped mispairing* or *stuttering*. Figure 7.12 showed in detail how this stuttering mechanism can change the number of repeat units at the SSR locus responsible for Huntington disease.

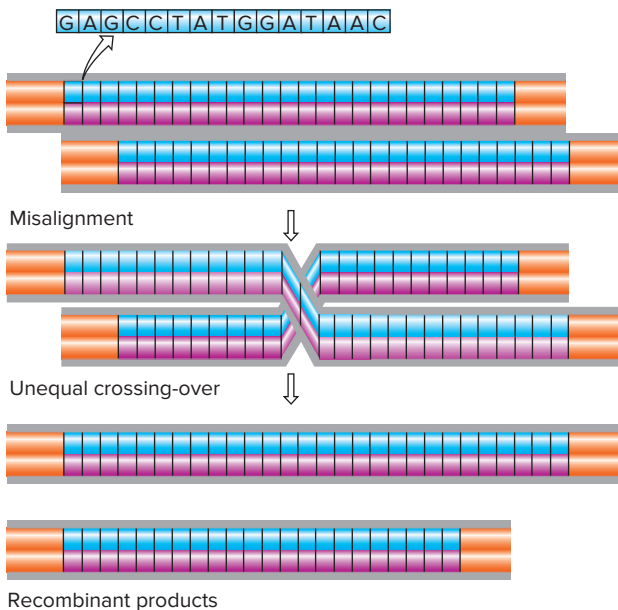
Because of events such as slipped mispairing, new alleles arise at SSR loci at an average rate of  $10^{-3}$  per locus per gamete (that is, one in every thousand gametes). This frequency is much greater than the single nucleotide mutation rate of  $10^{-9}$  and results in a large amount of SSR variation among unrelated individuals within a population. Unlike SNPs—which are biallelic and do not change after the mutational event that gave rise to them—SSRs are therefore highly polymorphic in the number of repeats they carry, often with 10 or more alleles distinguishable at a single SSR locus. The rate of SSR mutation is nonetheless low enough that changes usually do not occur within a few generations of even a large family. SSRs can thus serve as relatively stable, highly polymorphic DNA markers in linkage studies of many organisms, including humans.

### Copy number variants (CNVs)

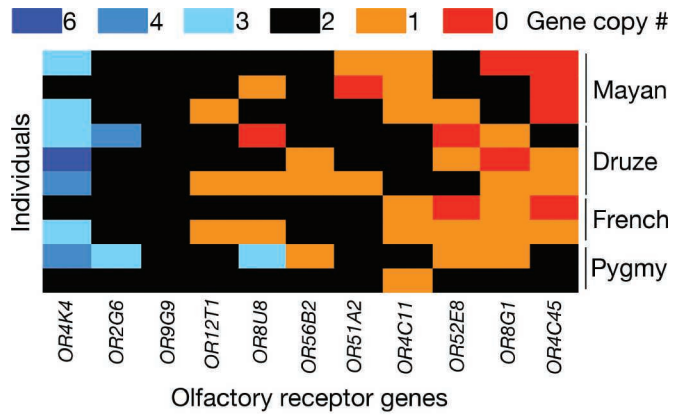
Individual human genomes also display DNA length polymorphisms involving more than just the few nucleotides characterizing SSRs and DIPs. Researchers were surprised to find that the genomes of many people showing no signs of any genetic disease carry a variable number of copies of large blocks of genetic material up to 1 Mb in length. This category of genetic variants is referred to as **copy number variants (CNVs)**. CNVs turn out to be quite common both in their distribution across the genome and in their frequency of occurrence within human populations (Table 11.1). Over 10,000 CNV loci have been identified in all human genomes examined to date, and pairwise comparison of any two genomes typically identifies differences at more than 1000 of these loci.

One of the most important mechanisms that can produce new alleles of a CNV locus is **unequal crossing-over** (Fig. 11.7). During meiosis I, tandem arrays of the repeating units on homologous chromosomes can pair out of register. If recombination takes place between mispaired repeating units, gametes are produced that have more or fewer copies of the repeating unit than the originals. Although mechanisms such as unequal crossing-over make CNV loci highly polymorphic, CNVs are still relatively stable when observed in families over a few generations: More than 99% of all CNV alleles in the current human population are thus derived from inheritance rather than new mutation.

**Figure 11.7** CNVs are highly polymorphic because of their potential for unequal crossing-over. CNVs are composed of tandem repeating units of identical or near identical sequences more than 10 bp long. (Blue and purple boxes are complementary strands of the repeating unit.) Misalignment and unequal crossing-over produce recombinant products—new alleles that have more or fewer repeating units than either parental allele.



**Figure 11.8** CNVs of olfactory receptor genes. Each row is a different person; each column a different olfactory receptor gene. The colors indicate the number of copies of a particular gene in a particular individual. Different humans vary substantially in copy numbers of olfactory receptor genes, accounting for much of the variation in people's ability to smell certain scents.



The olfactory receptor (*OR*) gene family, which encodes proteins that allow animals to smell a diverse array of odors, offers a fascinating example of variation in gene copy number. A typical mouse genome carries 1400 *OR* genes distributed at numerous chromosomal sites. But a keen sense of smell is no longer as important for human survival. As a result, *OR* genes can be lost without consequence, and people typically carry fewer than a thousand *OR* genes. However, individual humans vary widely around this mean. **Figure 11.8** shows the variation in copy number among 10 people at 11 representative *OR* loci. One gene, *OR4K4*, varies in copy number from two to six in different genomes, while five of the 11 genes are completely missing from some individuals. Some people can have hundreds more or hundreds fewer *OR* genes than others do, resulting in large differences in people's abilities to distinguish odors.

#### essential concepts

- When two or more alleles exist at a DNA locus, the locus is polymorphic, and the variations themselves are *DNA polymorphisms (DNA markers)*. Most polymorphisms are *anonymous*; they have no effect on phenotype.
- *Single nucleotide polymorphisms, or SNPs*, are the most frequently found DNA polymorphisms. The low rate of SNP formation allows investigators to estimate when particular SNP-causing mutations occurred during evolution.
- Addition and subtraction of DNA sequences also cause genetic variations, including *DIPs*, *SSRs*, and *CNVs*. These variations can result from stuttering of DNA polymerase during replication of repeated sequences or from unequal crossing-over during meiosis.

## 11.2 Genotyping a Known Disease-Causing Mutation

### learning objectives

1. Outline the steps by which the polymerase chain reaction (PCR) amplifies a specific region of a genome.
2. Describe how the sequencing or sizing of PCR products can elucidate genotypes.
3. Explain how PCR can be used to genotype fetuses *in utero* or embryos prior to implantation.

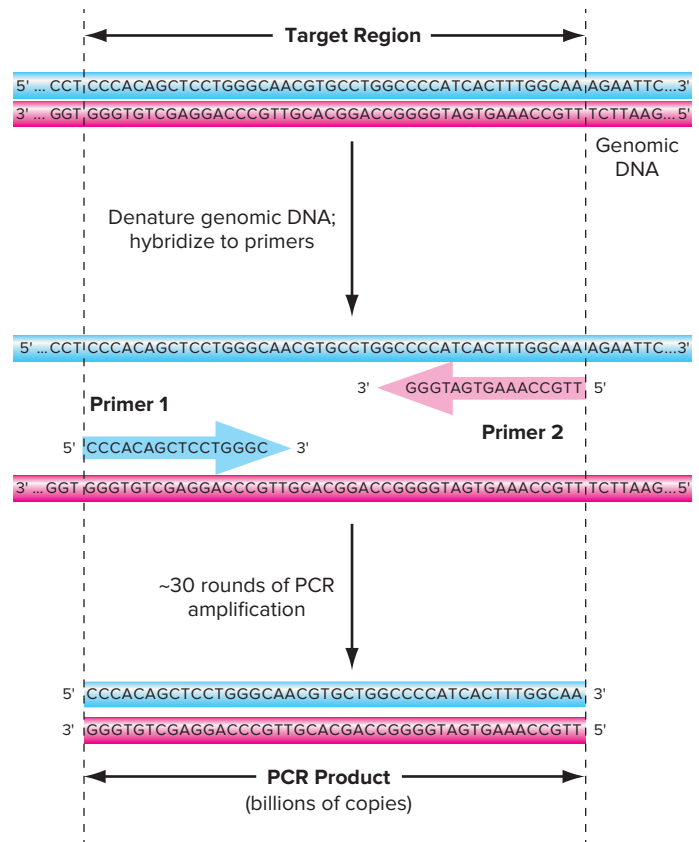
The ability to genotype individuals for genetic diseases provides information that can impact lives profoundly. If a person is diagnosed as having a genetic disease for which treatments are available, knowledge of the DNA genotype could save his or her life. Even if the condition cannot be treated, the genotyping of prospective parents, a fetus carried by a mother, or embryos created by *in vitro* fertilization allows families to make informed reproductive decisions.

The ability to determine whether a person is homozygous or heterozygous for a disease-causing allele of course presumes that scientists already know the precise change of nucleotides responsible for the disease. For certain diseases such as sickle-cell anemia, the identity of the disease-causing mutation is clear because we know how a particular protein (such as hemoglobin) is altered in the disease. But in most cases, the phenotype of the disease does not provide clear information about the disease gene. Later sections of this chapter will describe more general strategies to find mutations associated with various diseases. Once the mutation is identified, individuals can be genotyped by the methods we now discuss.

### The Polymerase Chain Reaction (PCR) Amplifies Defined Regions of a Genome

Determining whether a person is homozygous or heterozygous for a disease-causing allele, or homozygous for a normal allele of the same gene, implies that you can isolate that gene from the person's genome and analyze the alleles by looking at the purified DNA. But genes are rare targets in complex genomes: The gene for the  $\beta$  chain of hemoglobin, for example, spans only about 1400 of the 3,000,000,000 nucleotide pairs in the haploid human genome. In 1985, Kary Mullis invented one of the most powerful techniques in molecular biology, called the **polymerase chain reaction (PCR)**, to deal with this problem of looking for the needle of a gene in the haystack of the genome. PCR is remarkably fast and efficient. Starting

**Figure 11.9** Overview of the polymerase chain reaction (PCR). The reaction amplifies a target region of genomic DNA defined by the 5' ends of the two primers, making a PCR product. To help you understand the basis of PCR, the Watson and Crick strands of the same region of DNA are shown in different colors.



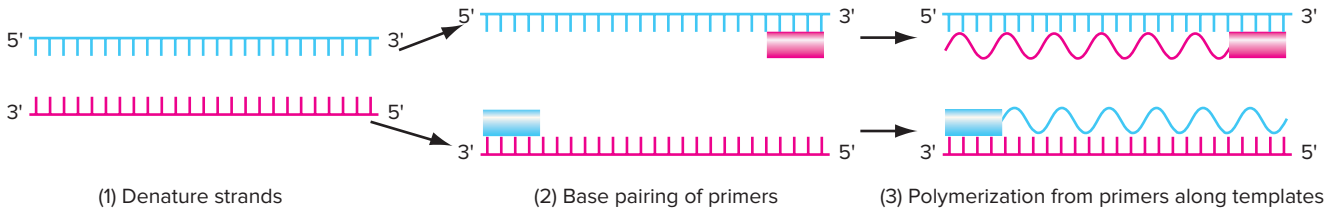
with minute amounts of DNA, such as that found in a single sperm or hair follicle, researchers can make a billion or more copies of a short, defined portion of that genome in just a few hours.

As shown in **Fig. 11.9**, PCR amplifies a *target region* of DNA. Two 16- to 30-base-long oligonucleotides, the *PCR primers*, define the ends of the target region. The investigator synthesizes these primers based on prior knowledge of the genome. One oligonucleotide is complementary to one strand of DNA at one end of the region; the other oligonucleotide is complementary to the other strand at the other end of the region. If these primers are drawn as 5'→3' arrows to indicate their polarity, the arrows will point toward each other through the target region (Fig. 11.9). For DNA genotyping, the object of the PCR is to amplify sequences within the target region that may have different allelic forms.

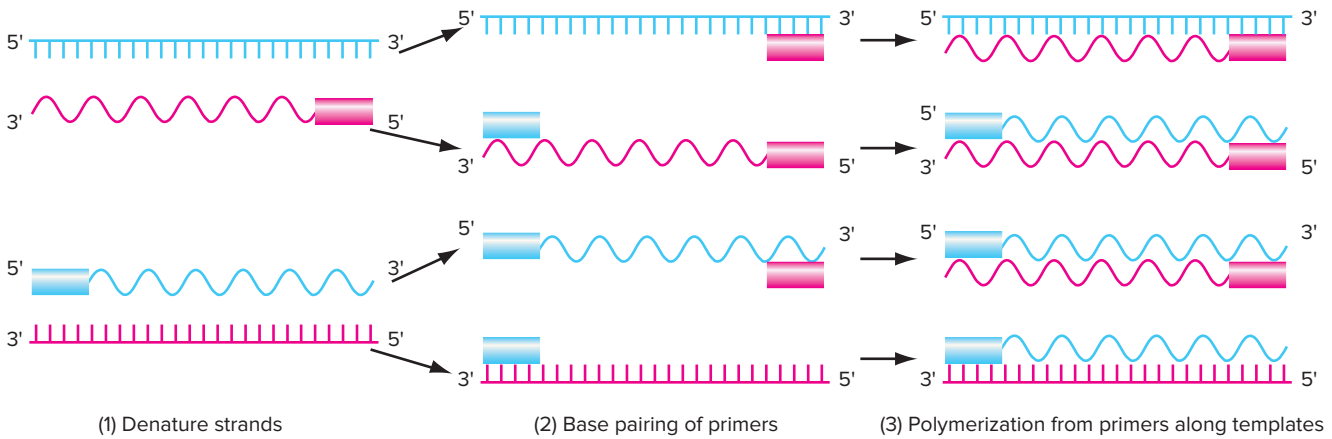
The process of amplification is initiated by the hybridization of these synthetic oligonucleotides to one or more denatured template DNA molecules (melted into single strands) within the sample of genomic DNA to be analyzed (**Fig. 11.10**). The oligonucleotides act as primers that allow

**Figure 11.10 PCR amplification of a target sequence.** In the first PCR cycle of this example, a single molecule of double-stranded genomic DNA is heated to denature it into single strands. The temperature is then lowered to allow these single strands to hybridize with the two PCR primers. A heat-stable DNA polymerase now polymerizes new DNA onto the 3' ends of each primer. In the second PCR cycle, the DNA in the reaction tube is denatured and then hybridized to the same primers. Both the original DNA strands and those made in the first cycle serve as templates in the second round. Each cycle of PCR thus doubles the amount of DNA in the target region. After the conclusion of the third round of PCR, the 5' ends of the majority of template strands are defined by the 5' ends of the primers, fixing the length of the accumulating PCR product.

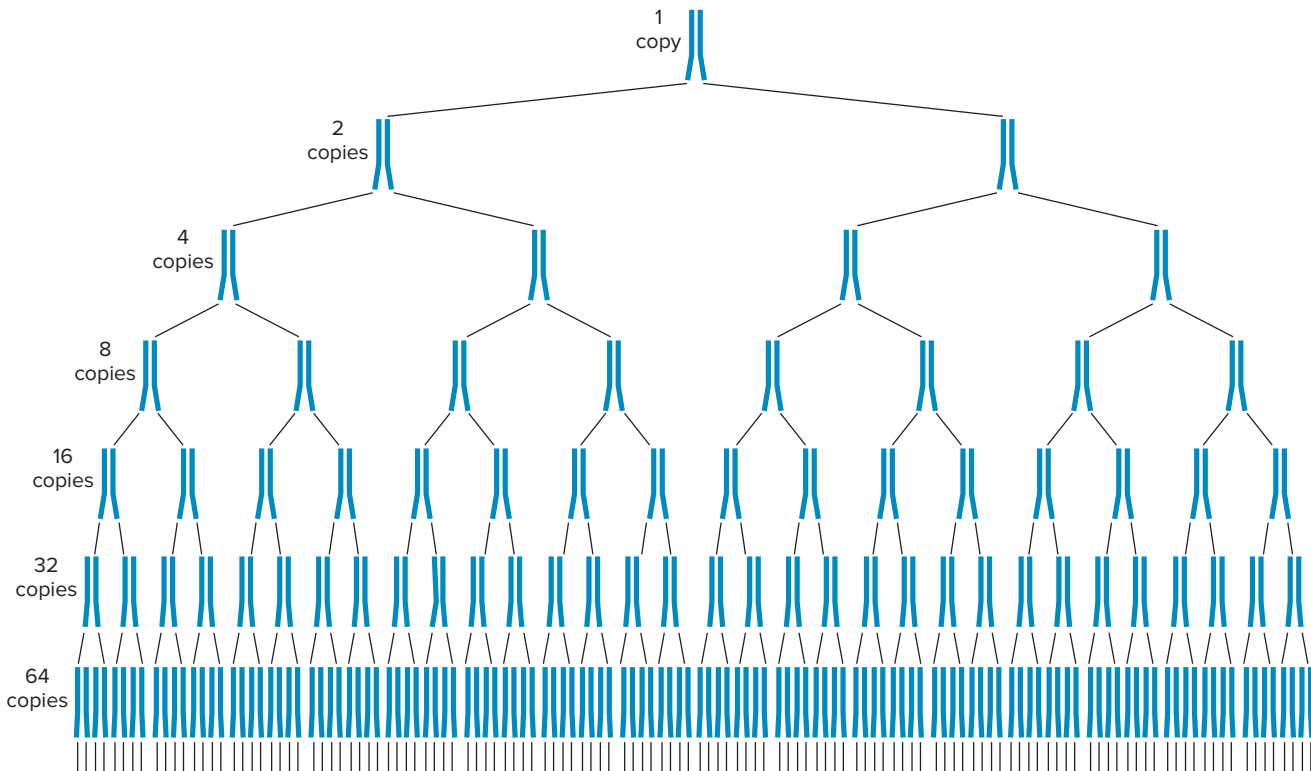
**PCR cycle 1:**



**PCR cycle 2:**



**Exponential amplification**





DNA polymerase to generate new strands of DNA complementary to both strands of genomic DNA between the primers; remember that DNA polymerase adds nucleotides sequentially onto the 3' end of a primer.

After sufficient time has elapsed to allow copying of the target region, the reaction is heated in order to melt apart the original template strands of DNA from the newly synthesized strands. The reaction tube is allowed to cool, so that the starting DNA and the copies synthesized in the previous step become templates for further replication, using the oligonucleotides remaining in the tube as primers. Performing this same sequence of steps—denaturation into single strands, hybridization of primers, and polymerization by DNA polymerase—as an iterative loop results in an exponential increase in the number of copies of the target region with each step (Fig. 11.10). Repeating the cycle just 22 times would generate more than a million double-stranded copies of the target region; after 32 repetitions, the reaction tube would have over a billion copies of this part of the genome.

The iterative steps of the protocol can be automated in a PCR machine that heats up and cools down the sample according to a preprogrammed schedule. The reaction tubes placed into the machine contain enough nucleotide triphosphates and oligonucleotide primers to support these multiple rounds of DNA replication. Moreover, the tubes contain a special DNA polymerase from a bacterium that grows in hot springs. This DNA polymerase remains active after being subjected to the high temperatures used to melt apart the DNA strands at each round of the PCR protocol.

It is crucial for you to remember from Fig. 11.9 that the two priming oligonucleotides dictate the nature of the final PCR product. The ultimate PCR product is a double-stranded fragment of DNA that extends from the position of one primer's 5' end to the position of the other primer's 5' end. The primers must be complementary to opposite strands and have 5'-to-3' polarities that point toward each other through the region of interest. In practice, PCR is inefficient if the primers are far apart, so the protocol generally cannot amplify DNA regions greater than 25 kb long.

## PCR Products Are Genotyped by Sequencing or Sizing

For Mendelian genetic diseases caused by changes involving only one or a few nucleotides in a single gene, all of the information that distinguishes a normal allele from a mutant allele resides within a discrete region of the genome that can be encompassed by a PCR product. The differences between alleles can be recognized either by direct sequencing of PCR products, or in cases in which mutations add or subtract nucleotide pairs from the genome, by simply looking at the sizes of the PCR products. More complex polymorphisms such as copy number variants

(CNVs) affect many nucleotide pairs that extend over regions much larger than can be amplified into a PCR product, so they must be analyzed by other methods that will be described later in the book.

## Sequencing PCR products

As you may remember, the mutation  $Hb\beta^S$  that causes sickle-cell anemia changes the identity of a single amino acid in the  $\beta$  chain of adult hemoglobin from glutamic acid to valine. The mutant allele is a single nucleotide substitution that changes an A to a T in the mRNA-like strand of the  $\beta$ -globin gene; the mutation is thus a single nucleotide polymorphism (SNP) (Fig. 11.11a). By genotyping the alleles of this SNP, we can identify people who will suffer from sickle-cell anemia or are carriers for this trait.

The process begins by PCR amplifying the locus from the person's genomic DNA using a pair of primers complementary to sequences on either side of the actual disease-causing mutation (Fig. 11.11a). Once the PCR product is made, its DNA sequence can be determined by the automated Sanger method shown previously in Fig. 9.7. Either one of the two PCR primers can serve as the primer for the sequencing reactions.

As Fig. 11.11b shows, the nucleotide substitution responsible for the disease shows up clearly in comparing the sequence obtained from the PCR products in  $Hb\beta^S Hb\beta^S$  sickle-cell patients and  $Hb\beta^A Hb\beta^A$  normal homozygotes. But importantly, both alleles are visible simultaneously in the sequence trace made from the PCR product generated using  $Hb\beta^A Hb\beta^S$  heterozygous genomic DNA as the template. Genomic DNA prepared from somatic cells of the heterozygote contains both allelic variants. Because the primers hybridize equally well with the two homologous chromosomes (given that the sickle-cell mutation does not alter the genomic sequences complementary to the primers), about half the DNA molecules in the final PCR product will contain the mutant sequence and the other half the wild-type sequence. Heterozygosity for the disease-causing SNP is thus seen as a double peak showing both A and T in the DNA sequence trace.

The technique of sequencing PCR products amplified from genomic DNA is a straightforward way to determine one's genotype for any SNP. The same method can also be used to genotype other kinds of polymorphisms involving small numbers of nucleotides, such as small deletions/insertions (DIPs) or expansions/contractions of the numbers of repeats in simple sequence repeats (SSRs).

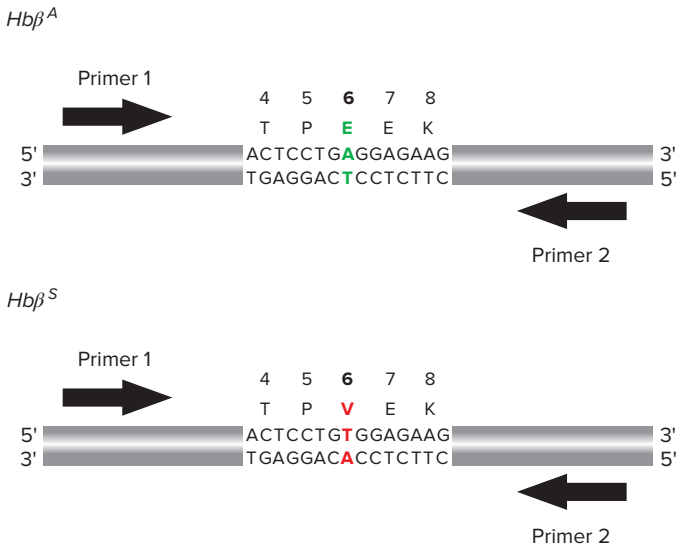
## Size variation in PCR products

In some cases it is possible to genotype a polymorphism in a PCR product without actually sequencing it. Gel electrophoresis can easily distinguish small variations in the actual size of a locus caused by DIPs or SSRs, as illustrated in Fig. 11.12. Again, you begin by using a pair of primers

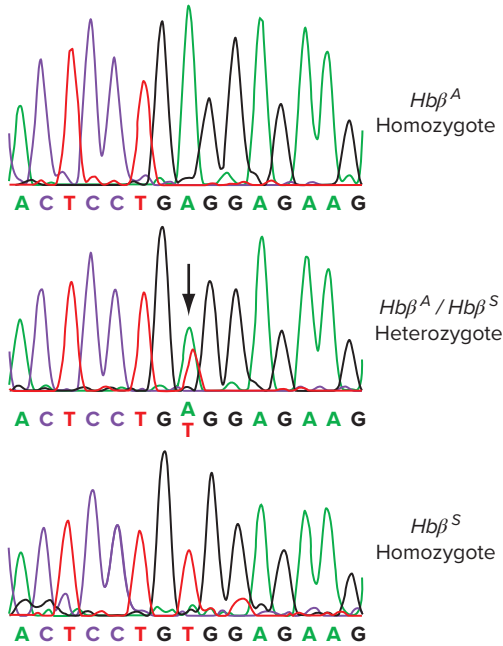
**Figure 11.11** Detection of the sickle-cell mutation by sequencing of PCR products. (a) The mutation responsible for sickle-cell anemia is a SNP that changes a single amino acid in the hemoglobin  $\beta$  polypeptide from glutamic acid (E) to valine (V). This polymorphism is PCR amplified by primers complementary to flanking sequences that do not vary between alleles. (b) Sequencing of PCR products made from genomic DNA templates. Note that the sequence of the PCR product from a heterozygous carrier shows both the normal ( $Hb\beta^A$ ) and mutant ( $Hb\beta^S$ ) nucleotides at the position of the substitution (black arrow).

Courtesy of Joshua J. Filter, Cornell University, Ithaca, New York

**(a) PCR amplifying alleles of the  $Hb\beta$  gene**



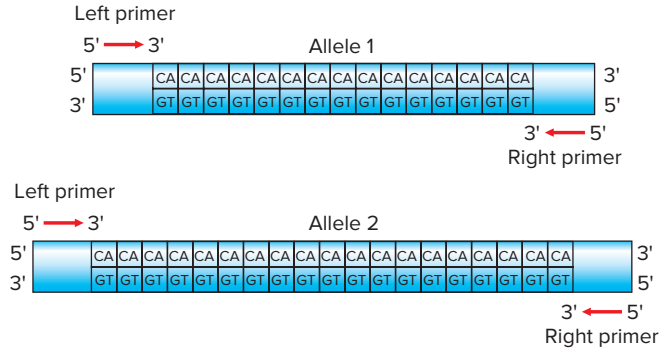
**(b) Genotyping for sickle-cell anemia**



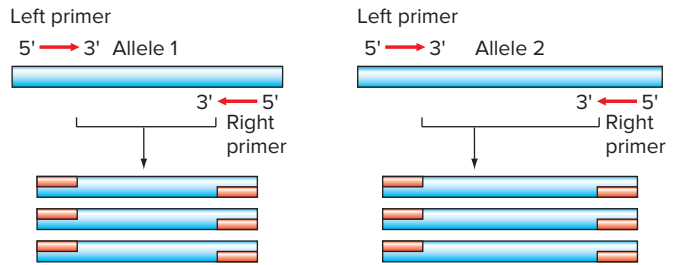
**Figure 11.12** Detection of simple sequence repeat (SSR) polymorphisms by electrophoresis of PCR products.

(a) SSR alleles differ in length. Left and right primers correspond to unique sequences that flank the SSR locus. (b) Genomic DNA is amplified by PCR with primers specific for the particular SSR locus. (c) Gel electrophoresis and ethidium bromide staining distinguish the alleles from each other. (d) SSRs are highly polymorphic with many different alleles present in a population, but each person has only two alleles of any given SSR locus.

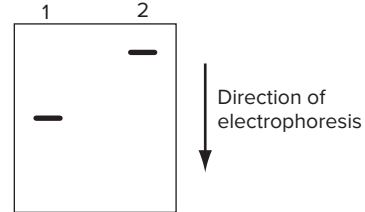
**(a) Synthesize primers corresponding to sequences flanking repeat locus.**



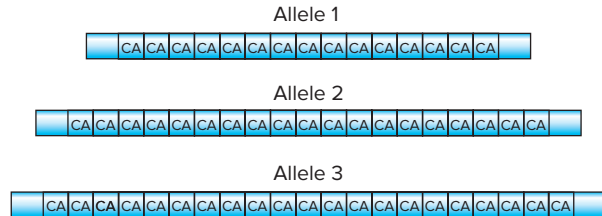
**(b) Amplify alleles by PCR.**



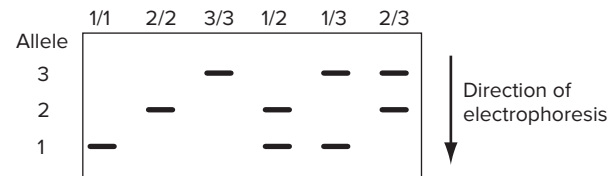
**(c) Analyze PCR products by gel electrophoresis.**



**(d) Example of population with three alleles**



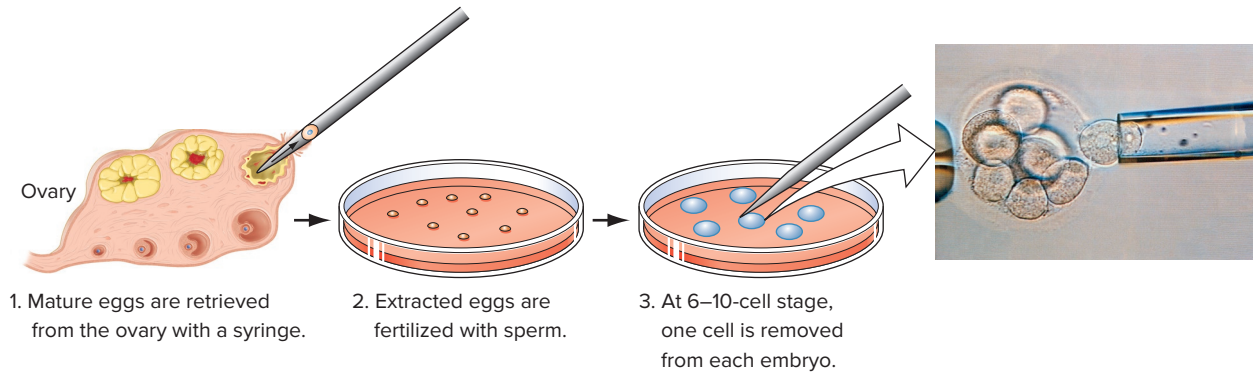
Six diploid genotypes are present in this population.





**Figure 11.14 Preimplantation embryo diagnosis.** Plucking one cell from an eight-cell embryo for the direct detection of genotype. The genomic DNA from the one cell is extracted and subjected to genotyping after PCR amplification. The remainder of the embryo survives and can be implanted into the mother's uterus.

© Benoit Rajau/Science Source



**embryo diagnosis** begins when a woman is injected with follicle stimulating hormone (FSH) to stimulate the maturation of about 10 eggs in her ovaries. An obstetrician then removes these eggs from the ovaries and fertilizes them *in vitro* with her partner's sperm. The fertilized eggs are then incubated for several days, allowing several cycles of mitotic division so as to produce early embryos containing 6–10 cells.

Specially trained technicians next use micropipettes to remove a single cell from each of these early embryos (Fig. 11.14). These early embryonic cells are not yet determined to become particular cell types or organs; indeed, embryos that split naturally at this stage can develop into healthy identical twins. Thus, the removal of a single cell does not harm the embryos nor prevent them from developing normally.

The technicians then prepare genomic DNA from the single cell obtained from each embryo and PCR amplify the specific region containing the site of the disease-causing mutation. They then analyze the PCR products by sequencing or sizing. In consultation with the physicians, the parents can select healthy embryos with genotypes that would not result in the disease (homozygous for the normal allele, or heterozygous for a recessive disease-causing allele). Usually two or three such embryos are placed into the mother's womb to improve the chances that at least one will implant properly into the uterus.

That embryos can be genotyped within a few days after fertilization by looking at the DNA from a single cell is conceptually and technically astonishing, but preimplantation embryo diagnosis has been used successfully in tens of thousands of pregnancies worldwide. The procedure is complex and expensive, costing thousands of dollars, but the information it provides can be invaluable to couples whose children would otherwise be at risk for serious genetic diseases.

### essential concepts

- PCR amplifies specific regions of DNA defined by two oligonucleotide *primers*. Repeated cycles of synthesis increase exponentially the number of copies of the target DNA region.
- PCR product sequencing constitutes a simple method for genotyping many polymorphisms. Small insertions or deletions of DNA (DIPs and SSRs) can also be genotyped by examining PCR product sizes on gels.
- Fetuses can be genotyped *in utero* by obtaining fetal cells via *amniocentesis*. In *preimplantation diagnosis*, a single cell from an early embryo produced by *in vitro* fertilization is genotyped.

## 11.3 Sampling DNA Variation in a Genome

### learning objectives

1. Explain why a relatively small number of SSR loci are sufficient to provide a DNA fingerprint of an individual.
2. Describe how a DNA microarray is constructed and how to genotype millions of loci on this microarray.

The vast majority of polymorphisms present in any genome do not cause disease or otherwise affect phenotypes. Several good reasons nonetheless exist for determining people's genotypes at these *anonymous loci*. Genotyping such loci allows people to be identified by their DNAs, which is highly useful for forensic purposes and for studying human

evolution and history. Of additional practical significance, anonymous loci serve as molecular markers for specific regions of the genome. Even if these DNA markers do not themselves cause disease, scientists can follow their inheritance to locate hard-to-find genes responsible for genetic diseases and other phenotypes.

### Forensic DNA Fingerprinting Examines Multiple SSR Loci

SSR loci are highly polymorphic: Many alleles that differ in the number of repeating units exist in the population, although any one person carries only two of these alleles for any given locus. The polymorphism of SSR loci makes them a powerful resource in identifying a person from his or her DNA.

The power comes from the possibility of examining multiple polymorphic SSR loci simultaneously. Suppose the likelihood that any two random people share exactly the same combination of two alleles of a particular SSR locus is 10% (0.1), and that the same is true for a second, independently segregating SSR locus. Using the product rule for independent events, the probability that two randomly chosen people will have the same alleles at the two SSR loci is  $(0.1) \times (0.1) = 0.01$  or 1 in 100. Now consider 13 such SSR loci. The chance that two people will have the same combinations of alleles at all 13 positions in the

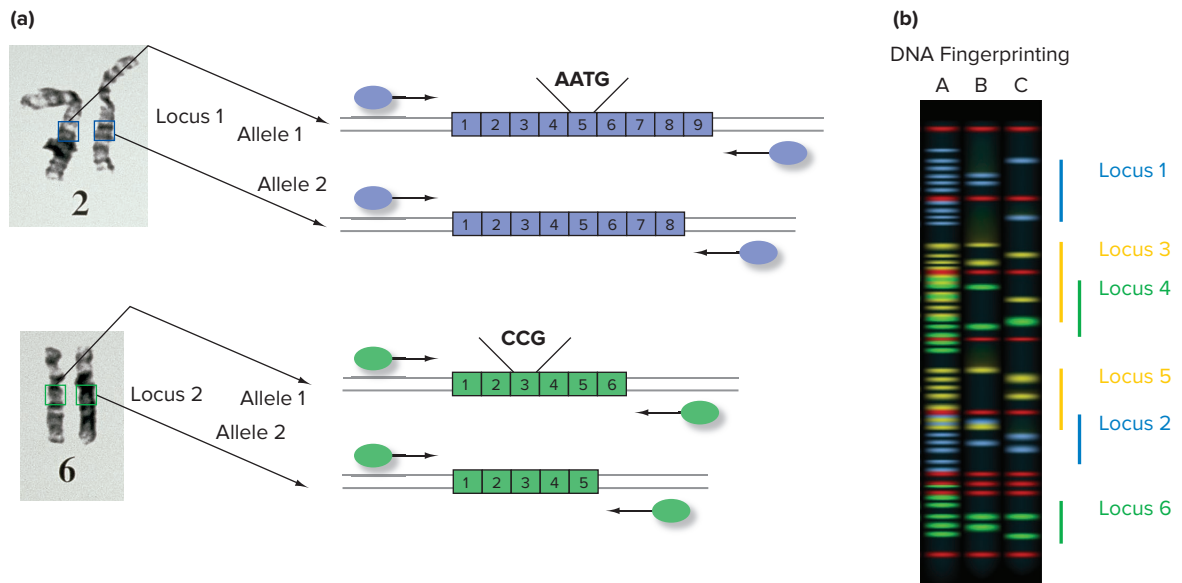
genome is  $(0.1)^{13}$ , or one chance in 10 trillion. (This calculation is simplified but gives you a rough idea of these probabilities; Chapter 21 on population genetics will show you how to make such calculations precisely from actual data.) Given that the earth currently has only about 7 billion human inhabitants, you can see that a genotype for 13 unlinked, polymorphic SSR loci would serve as a **DNA fingerprint** unique to any one person, excepting identical twins.

A simple extension of a method we have already discussed to genotype a single SSR locus allows the simultaneous genotyping of multiple SSRs. Figure 11.12 showed that the PCR products amplified from an SSR locus will have different sizes reflecting the number of repeating units in each allele. To examine 13 SSRs at the same time, you would label the 13 pairs of PCR primers with dyes that fluoresce in different colors, and then combine all the primer pairs in the same PCR reaction tube. After gel electrophoresis, you could then identify the allelic variants for each SSR locus based on the fluorescent colors and sizes of the PCR products (**Fig. 11.15**).

In the United States, the Federal Bureau of Investigation (FBI) maintains a database called CODIS (Combined DNA Index System) that allows forensic laboratories throughout the country to share and compare DNA profiles. All these laboratories use the same 13 primer pairs to amplify the 13 SSR loci. The laboratories carefully catalog the sizes of the PCR products and submit the results to the

**Figure 11.15 DNA fingerprinting.** (a) Basis of DNA fingerprinting by *PCR multiplexing* (simultaneous analysis of PCR products from multiple loci). PCR primer pairs (ovals) amplify separate SSR loci, usually from nonhomologous chromosomes. The primer pairs are labeled with different fluorescent molecules (*blue* and *green* in this example for two loci). (b) Gel electrophoresis of multiplexed PCR products. This example shows the analysis in three people (A–C) of six loci, two of each color (*blue*, *green*, and *yellow*). The alleles for the two different loci of the same color (i.e., locus 1 and locus 2) differ sufficiently in their sizes that it is clear which allele belongs with which locus. *Red* bands in each lane are size standards.

a: © Scott Camazine & Sue Trainor/Science Source b: © Alila Medical Images/Shutterstock RF



CODIS database. All 50 states mandate the collection of DNA fingerprint data for CODIS from felons convicted of certain crimes, such as sexual offenders; the database also includes profiles from missing persons.

As of 2016, CODIS has assisted more than 300,000 criminal investigations. Typically, forensic investigators use the database to match the DNA profile of evidence left at crime scenes with that of a felon. But DNA can also establish innocence: Suspects can be excluded if no match with crime scene evidence exists. In fact, a public policy organization called the Innocence Project has used DNA fingerprint evidence to help exonerate more than 300 people convicted of capital crimes, including several who were awaiting execution.

The power of DNA fingerprinting technology raises many concerns about privacy and possible discrimination in the collection of data. Just to give one interesting example, consider that siblings will share 50% of all SSR alleles; the same is true of parents and children. As a result, it is possible to identify the perpetrator of a crime not by a match to his or her own DNA, but instead by a partial match to the DNA of a close relative. This kind of familial DNA search was critical for the apprehension in 2010 of the major suspect for the “Grim Sleeper” serial killer in Los Angeles. The suspect’s son had recently been convicted on a felony weapons charge, so his DNA was analyzed. The son’s DNA profile partially matched DNA fingerprints from semen and saliva found on the Grim Sleeper’s victims. Policemen followed the father, and one detective posing as a waiter obtained a partially eaten slice of pizza with the father’s DNA. Stuningly, this DNA was a perfect match with the crime scene evidence. (The trial has just begun as of this writing in early 2016.) Should criminal investigators be allowed to conduct such familial searches given that family members of felons who have not committed any crime are in effect under lifelong genetic surveillance?

DNA fingerprints have many important uses beyond forensics for capital crimes. They now provide the most

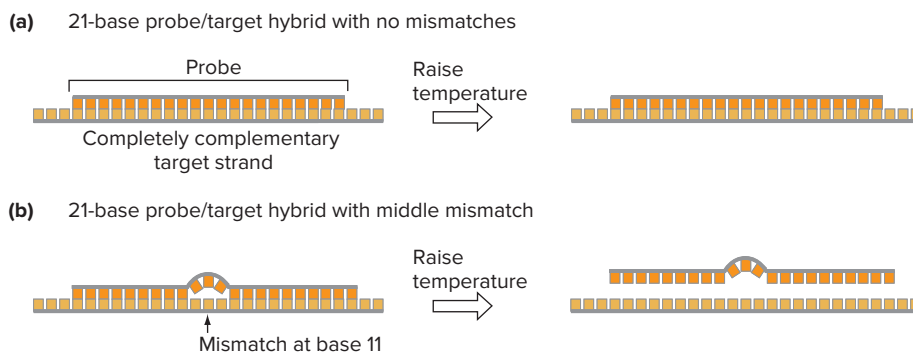
conclusive evidence in paternity suits, and DNA fingerprints can be used to identify human remains, as was the case for the victims of the World Trade Center disaster on September 11, 2001. The benefits of this technology are not restricted to DNA fingerprints of humans. Wildlife biologists study populations of endangered species by fingerprinting individual animals to increase the chance for success of captive breeding programs or to identify illegally poached animals. Owners of valuable domesticated animals such as show dogs, thoroughbred horses, or cattle can in some cases establish lineage through DNA fingerprints. In one fascinating if bizarre case from Argentina, scientists were enlisted to help apprehend a butcher who moonlighted as a cattle rustler. Meat hanging in the butcher’s shop had the same DNA profile as a tissue sample that a rancher had taken from one of his cows before it was stolen.

## DNA Microarrays Genotype Millions of SNPs

**Nucleic acid hybridization**, the ability of complementary single strands of DNA or RNA to come together to form double-stranded molecules, is the basis for many techniques in molecular biology. We have already discussed the importance of hybridizing oligonucleotide primers to DNA templates in Sanger DNA sequencing and in PCR. Both DNA sequencing and PCR assume that a perfect complementary match exists between all the nucleotides in the primers and templates. But what will happen if a mismatch exists between two single strands of nucleic acid?

Consider a 21-base oligonucleotide that hybridizes to a target strand that differs at a single base in the middle of the sequence (**Fig. 11.16**). The resulting double-stranded hybrid is significantly less stable than a similar hybrid in which all the nucleotides match. The reason is that the

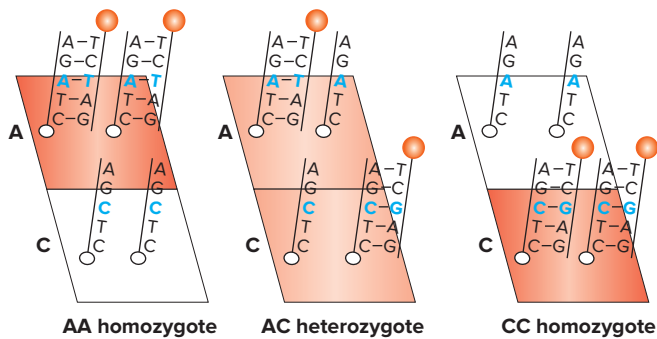
**Figure 11.16 Short hybridization probes can distinguish single-base mismatches.** Researchers allow hybridization between a short 21-base *probe* and two different target sequences. **(a)** A perfect match between probe and target extends across all 21 bases. When the temperature rises, this hybrid has enough hydrogen bonds to remain intact. **(b)** With a single-base mismatch in the middle of the probe, the effective length of the probe-target hybrid is only 10 bases. When the temperature rises, this hybrid falls apart.



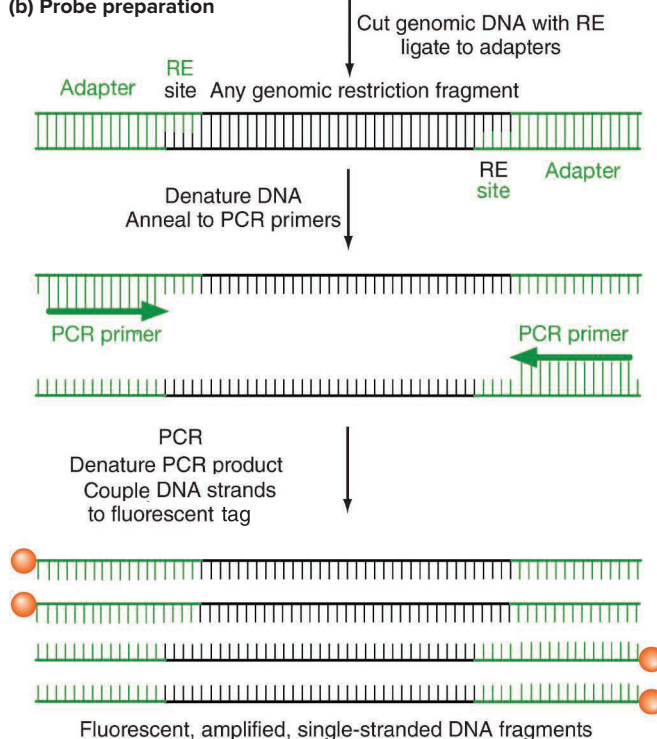
**Figure 11.17 DNA microarrays.** (a) Three identical DNA microarrays containing ASOs for two alleles of one SNP. (The ASOs are shown as 5 nucleotides long, but in practice they would be 20–40 nucleotides long.) A probe (with a red fluorescent tag) will hybridize only to perfectly complementary ASOs. The fluorescence intensity reflects the number of sequences in the genomic DNA that are complementary to the ASO. (b) Method to amplify genomic DNA for microarray analysis. Genomic DNA is cut with a restriction enzyme (RE), and the ends produced are ligated to a double-stranded oligonucleotide adapter. PCR then amplifies all genomic fragments using a single primer that hybridizes to part of the adapter. The resultant DNA fragments are denatured and fluorescently tagged (red). (c) A small region of a DNA microarray after hybridization with a genomic DNA probe.

c: Source: National Cancer Institute

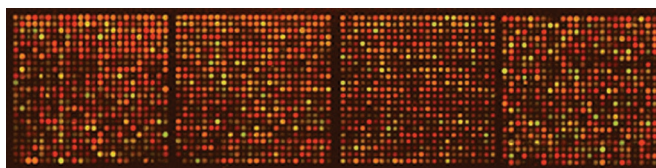
(a) Microarray schematic



(b) Probe preparation



(c) Signal from part of a large microarray



strength of the hydrogen-bond forces holding together the double helix depends on the longest stretch that does not contain any mismatches. When the two strands do not match exactly, there may not be enough weak hydrogen bonds in a row to hold them together. Thus, for small regions of up to about 40 bp, researchers can devise hybridization conditions (such as a particular temperature) under which the perfect hybrids will remain intact, while the less stable imperfect hybrids will not (Fig. 11.16).

Researchers can exploit the different stabilities of hybrid molecules for the genotyping of SNP loci. The idea, illustrated in Fig. 11.17a, is to attach to a solid support (such as a chip of silicon) short 20- to 40-base-long oligonucleotides that will hybridize under the right conditions to only one of the two alleles at a SNP locus. These oligonucleotides are logically called **allele-specific oligonucleotides**, or **ASOs**. The investigator now takes DNA from the genome to be analyzed and turns it into a **probe** by fragmenting the DNA, denaturing the fragments into single strands by heating, and attaching a fluorescent dye to these small pieces of single-stranded genomic DNA. The dye-carrying genomic DNA is now placed on the silicon chip (sometimes called a **DNA microarray**) and the temperature adjusted so that the genomic DNA probe will hybridize only to ASOs that match the probe perfectly. To visualize the hybridization signal, light is shined on the chip, and a detector records the amount of fluorescence emitted by each area containing a specific ASO. As you can see in Fig. 11.17a, the pattern of fluorescence allows straightforward determination of the genotype of the original genomic DNA for any SNP.

One additional feature of interest in Fig. 11.17a is that the intensity of the fluorescent signal over a particular ASO on the silicon chip is proportional to the number of copies of that allele in the genome. Fluorescence intensity can thus provide a way to monitor copy number variations (CNVs). We will discuss this use of ASO chips later in the book.

A single fluorescently-tagged genomic DNA molecule does not generate enough light to allow its detection on the microarray. Investigators must therefore amplify the genomic DNA so that many copies of each part of the genome can all be attached to fluorescent tags. **Figure 11.17b** illustrates one clever method for achieving this amplification. Researchers first digest the genome with a restriction enzyme that creates fragments with a sticky end, and then they ligate an *oligonucleotide adapter* to these restriction fragments. Part of the adapter can anneal to the overhang, while the rest of the oligonucleotide is complementary to a PCR primer. Fragments connected to adapters at the ends of both strands can now serve as templates for PCR amplification. In this way all parts of the genome could be amplified using a single adapter and a single PCR primer.

Rapid advances in DNA microarray manufacturing technology have led to the fabrication of chips capable of

detecting SNP alleles at more than 4 million loci (Fig. 11.17c). At the time of this writing (2016), the cost of analyzing a sample of genomic DNA is only a few hundred dollars, which works out to a per-SNP genotyping cost that is a small fraction of a penny. The SNP loci analyzed on commercially available microarrays include all single-nucleotide variants known to be associated with genetic diseases, but most of the loci on the chip are common SNPs likely to be without phenotypic effect. The widespread occurrence of these particular anonymous SNPs makes them invaluable for locating the mutations that do cause diseases, as will be explained in the next section.

### essential concepts

- In *DNA fingerprinting*, genotyping of multiple polymorphic loci such as SSRs provides enough information to identify individuals from their DNA.
- A *DNA microarray* contains *allele-specific oligonucleotides* (ASOs) for millions of SNP loci. Under the proper conditions, a *probe* made of fluorescently-labeled genomic DNA fragments binds only to complementary ASOs, allowing these loci to be genotyped.

## 11.4 Positional Cloning

### learning objectives

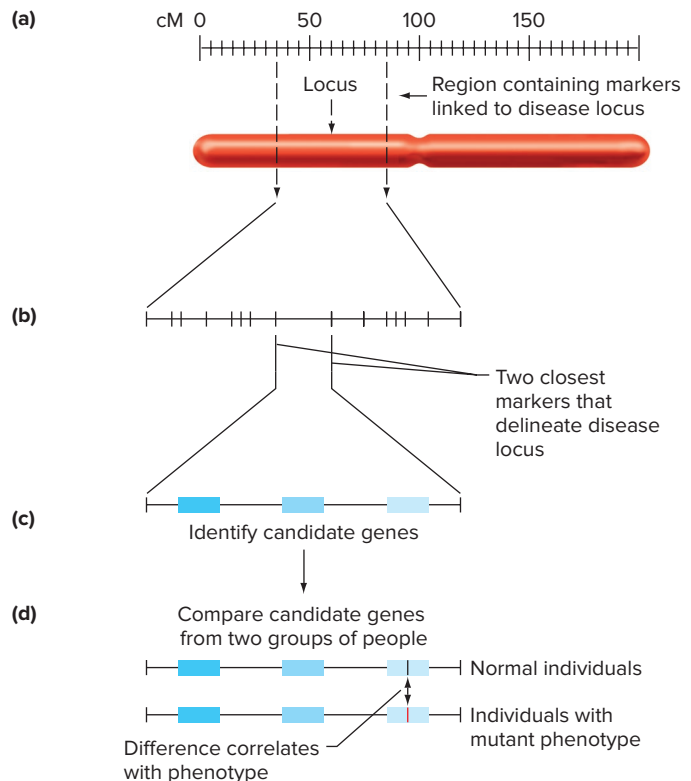
1. Describe the process of positional cloning and how it allows mapping of disease-causing mutations.
2. Examine the limitations of pedigree analysis in providing the information needed for positional cloning.
3. Explain how a Lod score is obtained and what information it provides.
4. Discuss the consequences of allelic heterogeneity, compound heterozygosity, and locus heterogeneity.

Of the thousands of known human **disease genes** (genes whose mutant alleles cause a disease phenotype), scientists can identify only a small number based on the specifics of the abnormal condition. For example, sickle-cell anemia and thalassemias are diseases affecting red blood cells. About 97% of the dry weight of a red blood cell consists of hemoglobin, so researchers directed their attention to the genes encoding the polypeptides making up this oxygen-carrying protein as likely causes of these diseases. More often, it is difficult to make an educated guess about which protein is changed by a disease-causing mutation, so a different approach is needed.

## Linkage Analysis with DNA Markers Gives Disease Genes an Approximate Chromosomal Address

A generally useful strategy to identify the defects causing hereditary diseases is called **positional cloning** (Fig. 11.18). The object is to obtain information about the unknown location of the disease gene by finding polymorphic loci to which the mutation is genetically linked. Because we know from the human genome sequence the exact position of each locus, discovering anonymous DNA polymorphisms closely linked to the disease gene allows researchers to focus their search for the mutation on a small region of a single chromosome. From the candidate genes within this region, the gene responsible for the disease can be found by looking for mutations that appear consistently in patients.

**Figure 11.18** Positional cloning: From phenotype to chromosomal location to guilty gene. (a) The disease gene is located less than 50 map units (~50 Mb) away from any markers linked to it. (b) Researchers narrow the region of interest by looking for the most closely linked markers to the left and right of the mutation. (c) Candidates for the disease gene (different shades of blue) must lie within the region of interest. (d) Comparing the structure and expression of each candidate gene in many diseased and nondiseased individuals pinpoints the causative mutation and thus the disease gene.







the T allele of the SNP. II-2, the unaffected mate of II-1, does not have the disease (so she is  $NF^+/NF^+$ ), and the DNA analysis indicates that she is homozygous for the G allele of SNP1. The children in generation III of the pedigree are thus in effect the progeny of a testcross. Examining each child both for the presence or absence of the disease and the SNP genotype will reveal whether the child obtained a parental-type (nonrecombinant) or a recombinant-type sperm from the doubly heterozygous father II-1.

As Fig. 11.20a shows, seven out of eight progeny in generation III resulted from parental-type sperm: either the combination of  $NF$  and G seen in the grandmother I-1, or the combination of  $NF^+$  and T in the grandfather I-2. Only one of the children in generation III was the product of a recombinant sperm (III-8, an affected son with  $NF$  and the T allele of the SNP). The data thus strongly suggest that the neurofibromatosis gene and this particular SNP are genetically linked, separated by a map distance  $(1/8) \times 100 = 12.5$  cM.

A family size of eight children is large by today's standards, but this number of progeny is still not sufficiently large to achieve great statistical significance. Nonetheless, the most straightforward interpretation of the data is that the neurofibromatosis gene is located on chromosome 17, within a region that extends roughly 12.5 Mb to either side of the SNP1 locus. Fig. 11.20b shows graphically how you can easily interpret the cross between II-1 and II-2 based on these provisional conclusions.

### Multipoint analysis

Microarray data include not just two-point information about linkage relationships between a disease gene and individual DNA loci, but also multipoint information about the behavior of millions of DNA loci with respect to each other. You saw in Chapter 5 the power of three-point crosses in fruit flies, but microarrays are like three-point crosses "on steroids." Researchers can now pinpoint particular crossovers occurring during the production of a single gamete to positions between two polymorphisms on the same chromosome. As you will see in Section 11.5, this information in turn allows researchers to pinpoint disease genes to short regions defined by mapped recombination events.

### Positional Cloning Has Several Limitations

For traits expressed in plants or small animals, researchers can easily set up crosses that generate hundreds of progeny so as to allow accurate genetic mapping. But scientists do not direct human breeding, so not every mating between two people provides interpretable information about the relative positions of any two given loci. In addition, human

family sizes are small, making it difficult to obtain sufficient data for precise mapping.

### The phase problem

You should note that the calculation just presented for the map distance between the neurofibromatosis gene and SNP1 does not include one of the crosses shown in the pedigree: the mating of I-1 and I-2 (see Fig. 11.20). The reason is that we don't know the configuration of the alleles (or *phase*) in the affected grandmother I-1. It is possible that her mutant  $NF$  allele was on the same chromosome 17 as the G allele of the SNP, but it is also possible that the SNP allele on this  $NF$ -bearing chromosome was instead T. Because we cannot say with certainty whether II-1 (her affected son) results from a parental-type or recombinant-type egg, we did not consider this gamete in calculating the map distance.

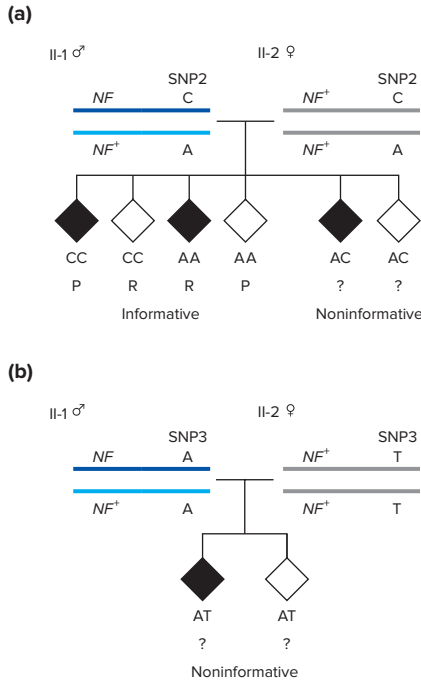
The phase problem can be resolved in either of two circumstances. First, if you know the genotypes of two loci in both of a person's parents, sometimes you can determine which alleles came from one parent and which from the other. This is precisely how we knew that the phase in the double heterozygous father II-1 was  $NF$  G/ $NF^+$  T (Fig. 11.20). So to determine the phase in I-1 by this method, we would need to have genotyping information about *her* affected parent. Second, if the two loci are sufficiently close together, you can infer the probable phase because the linked alleles should segregate with each other more often than not. In Fig. 11.20, where the parental classes considerably outnumber the recombinants in generation III, you could have inferred the likely arrangement of alleles in the doubly heterozygous II-1 even if you had no information about his parents.

### Informative and noninformative crosses

Even if you know the phase in a doubly heterozygous parent, a mating may not provide any useful information about whether the two loci are linked. One example is seen in **Fig. 11.21a**, which again examines the mating between male II-1 with neurofibromatosis and his unaffected partner II-2. However, we are now looking at a different SNP locus (SNP2), where both parents are heterozygotes for the alleles A and C. If a child were also a heterozygote for A and C, you would not be able to tell which parent contributed the A and which the C, so you could not determine if the child was the result of a parental or recombinant gamete. But if the child were a homozygote for either SNP2 allele, say of genotype AA, then you know both the egg and sperm must have carried the A allele.

You should remember from Chapter 5 that the basic requirement for genetic mapping is that at least one parent must be a double heterozygote. **Figure 11.21b** emphasizes this crucial point by showing that if neither parent is a double heterozygote, the cross cannot be informative.

**Figure 11.21** Some noninformative matings. (a) Even if you know that the *NF* gene and SNP2 are linked and you know the phase (arrangement of alleles) in the parents, the mating may or may not be informative. If the child is a CC or AA homozygote, the cross would be informative. If the child is an AC heterozygote, you don't know which allele came from which parent, so the cross would be noninformative. (b) If neither parent is a double heterozygote, it is impossible to perform linkage analysis. (SNP3 is a third SNP locus.)



Even if a mating is noninformative for the linkage of a disease gene with a particular SNP locus, multipoint analysis on microarrays usually provides a way for scientists to overcome this constraint. This is because the microarray will likely contain other nearby SNPs that will be informative.

**Obtaining sufficient pedigree data**

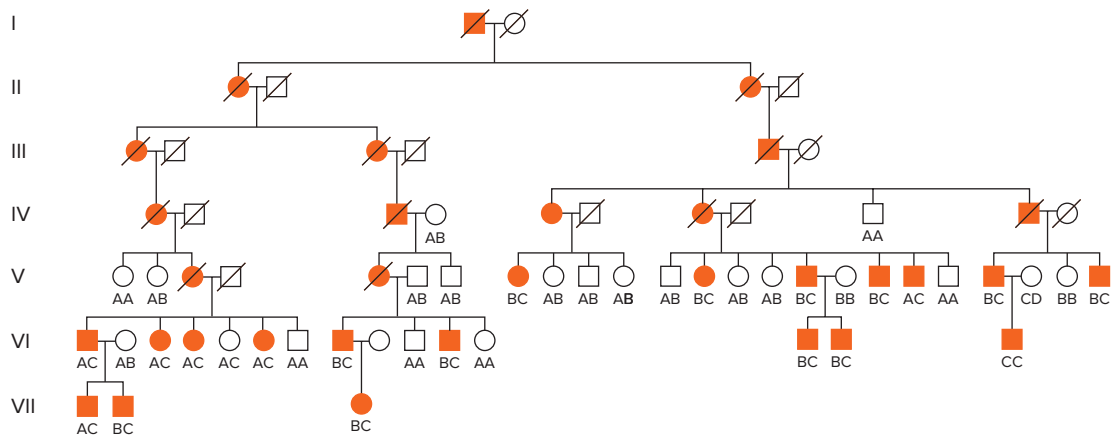
With millions of polymorphic loci on a DNA chip, it should be possible in theory to map disease genes very accurately, even if certain crosses are uninformative for linkage of the disease gene to certain DNA markers. However, the resolution of positional cloning is always limited in practice by the number of people human geneticists can track in families in which the disease is segregating. If scientists have mapped a disease gene to within 1 cM of a DNA polymorphism, this value means they have examined the phenotypes (affected or unaffected) of at least 100 members of such families, and they have also genotyped on microarrays the DNA of all these people. (Remember that 1 cM means 1 recombinant gamete out of 100 total gametes.)

For this reason, positional cloning achieved its first successes for diseases that can be found in extended families with a large number of children. In 1984, the Huntington disease (*HD*) locus became the first human disease gene to be mapped successfully by positional cloning precisely because such a family was available. **Figure 11.22** shows the seven-generation, 65-member family pedigree used to demonstrate tight linkage between a DNA marker called *G8* and the *HD* locus.

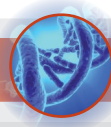
**The Lod Score Provides a Statistical Approach to Studying Linkage**

Positional cloning is rarely as straightforward as it was for *HD*. Most human families have only a few children, and it is difficult to obtain DNA and phenotype information about multiple generations in a pedigree. For these reasons, human geneticists have developed a statistical tool called a **Lod score** (log of the odds). The purpose of the Lod score is reminiscent of that of the  $\chi^2$  statistic used in Chapter 5: to determine whether the data are sufficient to conclude with confidence whether a disease gene and a

**Figure 11.22** A marker closely linked to the Huntington disease locus. Detection of linkage between the DNA marker *G8* and a locus responsible for Huntington disease (*HD*) was the first step in the cloning of the *HD* gene. The pedigree shows an extended Venezuelan family affected by *HD*. Alleles at the *G8* marker locus are indicated (A, B, C, and D), while affected individuals are indicated in orange. Cotransmission of marker alleles with the mutant and wild-type alleles at the *HD* locus is obvious.



## TOOLS OF GENETICS



Blue DNA: © MedicalRF.com

### The Lod Score Statistic

The Lod score is a mathematical answer to the question: How much more likely is it that the allele transmission pattern seen in a pedigree will occur if the loci are linked at a given recombination frequency (RF) less than 50%, than if they are not linked? The Lod score, as its name implies (log of the odds) is the logarithm of the ratio between these two probabilities:

$$\text{Lod} = \log \left[ \frac{P(\text{obtaining observed results if loci are linked at a given RF})}{P(\text{obtaining observed results if loci are unlinked})} \right]$$

Here, we illustrate the Lod score calculation for the pedigree in Fig. 11.20a. The pedigree suggests that the *NF* gene is linked to a particular SNP on chromosome 17. The calculation will allow us to determine our degree of confidence in this preliminary conclusion.

- 1. Tabulate which progeny are parental and which are recombinant.** In Fig. 11.20a, you can see that the first seven children in generation III have the parental (P) configuration of alleles and that only child III-8 has the recombinant (R) configuration. We'll abbreviate these data as *PPPPPPR*.
- 2. Calculate the Lod score denominator.** If two loci are unlinked, it is equally likely that any one child will be P or R (that is, the RF = 50%). The probability of P is thus 1/2, and the probability of R is also 1/2. The probability of obtaining children in the particular birth order *PPPPPPR* if the *NF* gene and the SNP locus are unlinked is:

$$P(\text{RF}_{50\%}) = \left(\frac{1}{2}\right)^8 = \frac{1}{256}$$

You can see that a generalized formula for this part of the calculation is simply:

$$\left(\frac{1}{2}\right)^n, \text{ where } n \text{ is the total number of tabulated individuals.}$$

- 3. Calculate the Lod score numerator.** Loci could be linked if the RF is any value less than 50%, but the calculation requires us to assume an RF value. The pedigree in Fig. 11.20a indicates an RF of 1/8 = 12.5%, so we will use this as our best current estimate. With RF = 1/8, the expected frequency of P progeny is 7/8, and R progeny is

1/8. The probability of seven parentals and one recombinant in the particular birth order *PPPPPPR* is:

$$P(\text{RF}_{12.5\%}) = \left(\frac{7}{8}\right)^7 \left(\frac{1}{8}\right)^1 \approx \frac{1}{20}$$

A generalized formula for calculating the Lod score numerator is:

$$(1 - \text{RF}_{\text{obs}})^{\#P} \times (\text{RF}_{\text{obs}})^{\#R}$$

where  $\text{RF}_{\text{obs}}$  is the RF indicated by the data, #P is the number of parentals, and #R is the number of recombinants.

- 4. Calculate the likelihood ratio.** This is simply the ratio of the values you found in steps 2 and 3. For this example,

$$P(\text{RF}_{12.5\%})/P(\text{RF}_{50\%}) = \left(\frac{1}{20}\right) / \left(\frac{1}{256}\right) = 12.8$$

This likelihood ratio means that it is 12.8 times more likely that the *NF* gene and the SNP are linked with RF = 12.5% than that they are not linked (RF = 50%).

- 5. Calculate the Lod score.** The Lod score is simply the base 10 logarithm of the likelihood ratio. For the example in Fig. 11.20a:

$$\text{Lod score} = \log(12.8) = 1.1$$

- 6. Interpret the Lod score.** The convention among human geneticists is that a Lod score  $\geq 3$  (that is, a likelihood ratio  $\geq 1000$ ) is required to be confident of linkage. The Lod score of 1.1 indicates that the data in Fig. 11.20a are insufficient to conclude that *NF* and SNP1 are linked.

#### Important points about Lod scores:

- The Lod score determined by assuming the precise RF implied by the data will always be the maximum Lod score obtainable for the data set.
- For a single pedigree, Lod scores can be calculated for any RF value less than 50%. A Lod score  $\geq 3$  indicates that the data obtained are compatible statistically with the particular distance (less than 50 m.u.) being tested.
- Likelihood ratios are converted into Lod scores because Lod scores calculated for the same RF value in different pedigrees may simply be added to see whether a Lod of 3 can be obtained.

DNA marker are genetically linked. The Lod score is used in human genetics instead of the  $\chi^2$  statistic because the Lod score better handles a small number of data points while allowing the data obtained from many different pedigrees to be combined.

The Lod score statistic is calculated from the ratio of two probabilities: the probability of obtaining a particular set of results in a pedigree if two loci are linked (assuming a particular RF value), and the chance of observing the same results if the loci are unlinked. The Lod score statistic is the base 10 logarithm (log) of this likelihood ratio. The

convention adopted by human geneticists is that a Lod score greater than or equal to 3 indicates two loci are linked. A Lod score of 3 means that it is 1000 times more likely that the two loci are linked than that they are not (because  $3 = \log 1000$ ). The beauty of the Lod score statistic is that because it is a log function, the Lod scores from different pedigrees may simply be added, so researchers will know when they have enough data to conclude that a disease allele is linked to a specific marker.

The Tools of Genetics Box entitled *The Lod Score Statistic* illustrates how to calculate a Lod score, using as

an example the pedigree in Fig. 11.20a. As you can see, the maximum Lod score for these data (which is obtained by assuming RF = 12.5%) is 1.1, indicating that this one pedigree is insufficient evidence for linkage of the SNP1 locus on chromosome 17 and the *NF* gene. However, if two additional pedigrees were available, each also with a Lod score of 1.1 (calculated for RF = 12.5%), the Lod score of the three pedigrees together would be 3.3, constituting strong evidence that the *NF* gene is on chromosome 17 and linked to SNP1.

## Genetic Diseases Can Display Allelic or Locus Heterogeneity

Suppose that by positional cloning you have been successful in narrowing down the location of a disease gene to a 1 Mb long region between two polymorphic markers. In the human genome, the average gene density is about 1 gene per 100 kb of DNA. More than 10 genes might therefore lie in the 1 Mb region. How could you discriminate among these candidates to find the right one?

In some cases, it might be possible to find clues by looking for changes in patients in the amounts or sizes of the mRNA transcripts or the protein products of genes (see Problem 38 at the end of this chapter). But by far the most generally useful strategy is to use PCR to amplify DNA from all the candidate genes in all available patients, and then sequence all of these PCR products. If you found that the patients all had identifiable mutations in one of these candidate genes, particularly mutations that might affect the amino acid sequence of the gene's protein product, the evidence would be powerful for identifying that candidate as the actual disease gene (review Fig. 11.18d).

Some genetic conditions are always caused by the same single mutation in a single gene; we've already seen that all patients with sickle-cell anemia are homozygous for exactly the same base pair substitution in the gene encoding the  $\beta$  chain of hemoglobin. DNA sequencing would thus reveal the same mutation in the genomic DNA of all patients and carriers of sickle-cell anemia.

### Allelic heterogeneity: Multiple mutant alleles in one gene

This simple scenario is not, however, always the case. Many other genetic diseases display **allelic heterogeneity**, meaning that they can be caused by a variety of different mutations in the same gene. An important example is cystic fibrosis, a recessive autosomal genetic condition inherited by 1 child in every 2500 born from two parents of European descent. Children with the disease have a variety of symptoms arising from abnormally

viscous secretions in the lungs, pancreas, sweat glands, and several other tissues. Most cystic fibrosis patients die before the age of 30.

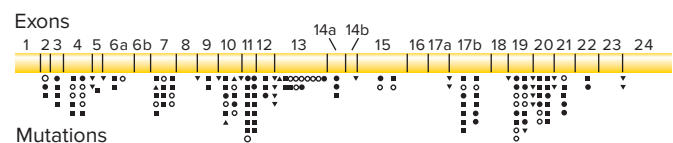
Positional cloning strategies allowed investigators to narrow their search for the causative gene to a 400 kb region between two DNA markers on chromosome 7 that contained only three candidate genes (previously shown in Fig. 11.5). One of these genes was *CFTR*, encoding the cystic fibrosis transmembrane receptor that allows chloride ions to pass through cell membranes; see Fig. 2.25). Significantly, both *CFTR* copies in all cystic fibrosis patients were found to contain mutations that would alter the amino acid sequence of the protein or would prevent normal amounts of the protein from being synthesized. Thus, as the name implies, *CFTR* is clearly the gene responsible for cystic fibrosis.

One mutation called  $\Delta F508$  (which removes the amino acid phenylalanine—F—from position 508 of the protein) accounts for about two-thirds of all mutant *CFTR* alleles worldwide. The remaining alleles consist of more than 1500 different rare mutations (Fig. 11.23). Many patients are thus so-called **compound heterozygotes** (sometimes known as **trans-heterozygotes**), in which one copy of chromosome 7 has one mutation in *CFTR* and the other copy of chromosome 7 has a different *CFTR* mutation. The disease results because neither chromosome 7 can encode a normal transmembrane receptor; in effect, the two different recessive *CFTR* mutations fail to complement each other.

The concept of allelic heterogeneity is central to understanding a drug that has very recently been developed to treat cystic fibrosis effectively, but only in a minority of patients. In 2012, the United States Food and Drug Administration approved the drug *ivacaftor* for patients who have one specific *CFTR* mutation called G551D (changing glycine at position 551 to aspartic acid). The mutant protein encoded by this allele assembles properly into the cell membrane, but the G551D protein is inefficient in

**Figure 11.23 Allelic heterogeneity in the *CFTR* gene.**

Every cystic fibrosis patient has a mutated *CFTR* gene on both copies of chromosome 7. The diagram indicates the location of mutations at different positions in *CFTR* relative to the 24 exons of the gene. *Compound heterozygotes* are patients who have one copy of *CFTR* with one of these mutations, while the other copy of *CFTR* has a different mutation.



- In-frame deletion
- Missense mutation
- Nonsense mutation
- Frameshift mutation
- ▾ Splicing mutation

transporting chloride ions across the membrane. Ivacaftor interacts specifically at the cell surface with the G551D mutant CFTR protein, enhancing its ability to transport chloride. This treatment has been remarkably effective in preventing the symptoms of cystic fibrosis from developing in young children, but unfortunately G551D accounts for only about 4% of all mutant *CFTR* alleles in the human population.

More recently (in 2015), researchers developed a treatment for the much more prevalent  $\Delta F508$  mutation. This allele encodes a protein that cannot fold properly and thus is not inserted into cell membranes. The new drug, called *lumacaftor*, ameliorates the folding problem, resulting in an increase in the number of CFTR molecules in cell membranes. The mutant proteins are still partially defective in chloride ion transport. Remarkably, a combination pill containing lumacaftor and ivacaftor prevents the development of cystic fibrosis in many patients homozygous for  $\Delta F508$ .

### Locus heterogeneity: Mutations in different genes cause the same disease

In this chapter, we deal exclusively with Mendelian genetic diseases caused by mutations in a single gene, but you already know that many other conditions display **locus heterogeneity**: They are caused by mutations in one of two or more different genes. A previously discussed example of a heterogeneous condition is deafness (review Fig. 3.23). In confronting a new genetic disease, researchers must always be aware of the possibility of locus heterogeneity.

In *complex traits* (also called *quantitative traits*) such as high blood pressure, many different genes can influence the phenotype even in a single person. Chapter 22 outlines some methods geneticists can use to study the genetic basis of these complex traits.

#### essential concepts

- *Positional cloning* identifies DNA polymorphisms that are linked to disease genes.
- *Lod scores* allow statistical assessment of linkage when data are limited, as in human pedigrees.
- After a disease gene is mapped approximately, researchers sequence candidate genes in the region to identify one that is altered consistently in affected individuals.
- In *allelic heterogeneity*, a variety of mutations in a single gene cause disease. *Compound heterozygotes* with two different recessive loss-of-function mutations in the same gene may display the mutant phenotype.
- In *locus heterogeneity*, mutations in one of two or more different genes can cause the same disease.

## 11.5 The Era of Whole-Genome Sequencing

### learning objectives

1. Describe a high-throughput, automated method by which millions of DNA templates may be sequenced simultaneously.
2. Summarize a sequence of investigative steps that can narrow the candidates for a disease-causing variant.
3. Explain how databases that catalog sequence variation in many people can facilitate the diagnosis of genetic diseases.

DNA microarrays with millions of SNPs sample only a small proportion of the variation between human genomes and can suggest only a disease gene's general chromosomal location. As we just saw, disease gene identification eventually requires DNA sequencing to correlate the disease phenotype with actual mutations. Suppose now that we could cheaply and accurately sequence all of the nucleotides—not just those of candidate genes—in an affected person's genome. The whole-genome sequence must somewhere include the causative mutation. Thus, unlike positional cloning, where the first goal is to find a molecular marker linked to the disease gene, the goal in the whole-genome approach is to find directly a DNA alteration that *is* the disease allele.

Startling developments are making the idea of routine and affordable whole-genome sequencing into a reality. Chapters 9 and 10 explained that the Human Genome Project, completed in 2003, sequenced the complete human genome at a cost of 3 billion dollars. Researchers have since invented imaginative new methods that have rapidly driven down the cost of DNA sequencing. In 2016, it is possible to sequence a person's whole genome (at a high coverage that will still miss some small regions) for about \$2000, and the cost will undoubtedly fall under \$1000 within a few years.

Whole-genome sequencing is still costly enough that researchers often economize by sequencing just that portion of the genome corresponding to the protein-coding exons. This is often informative because many, though far from all, disease-causing mutations alter the amino acid sequence of a protein. In **whole-exome sequencing**, investigators first enrich (by hybridization to cDNA sequences) for genomic DNA fragments that correspond to the exons of all genes, and then sequence these fragments. The **exome**, that is, the collection of all exons of all genes, constitutes less than 2% of whole-genome DNA, so whole-exome sequencing requires many fewer sequencing reads than whole-genome sequencing. When DNA sequencing becomes even cheaper, enriching for the exome will likely no longer have a significant cost advantage over whole-genome sequencing.

## New Techniques Sequence Millions of Individual DNA Molecules in Parallel

The major technical advances that are making exome and genome sequencing fast and cheap enough for use in identifying disease genes permit millions of individual DNA molecules to be sequenced simultaneously. Many creative methods have been invented to perform so-called *high-throughput* or *massively parallel* sequencing.

Several of these high-throughput methods for sequencing human genomes are straightforward extensions of the Sanger sequencing by synthesis approach you already learned in Chapter 9, but three things are new. First, individual DNA molecules being synthesized by DNA polymerase are anchored in one place. Second, these methods control base addition temporally so that each base can be identified before the next one is added. Third, in some systems the sensitivity of detection is so high that a single molecule of DNA can be monitored without the need for cloning or PCR amplification steps. As shown in **Fig. 11.24**, the

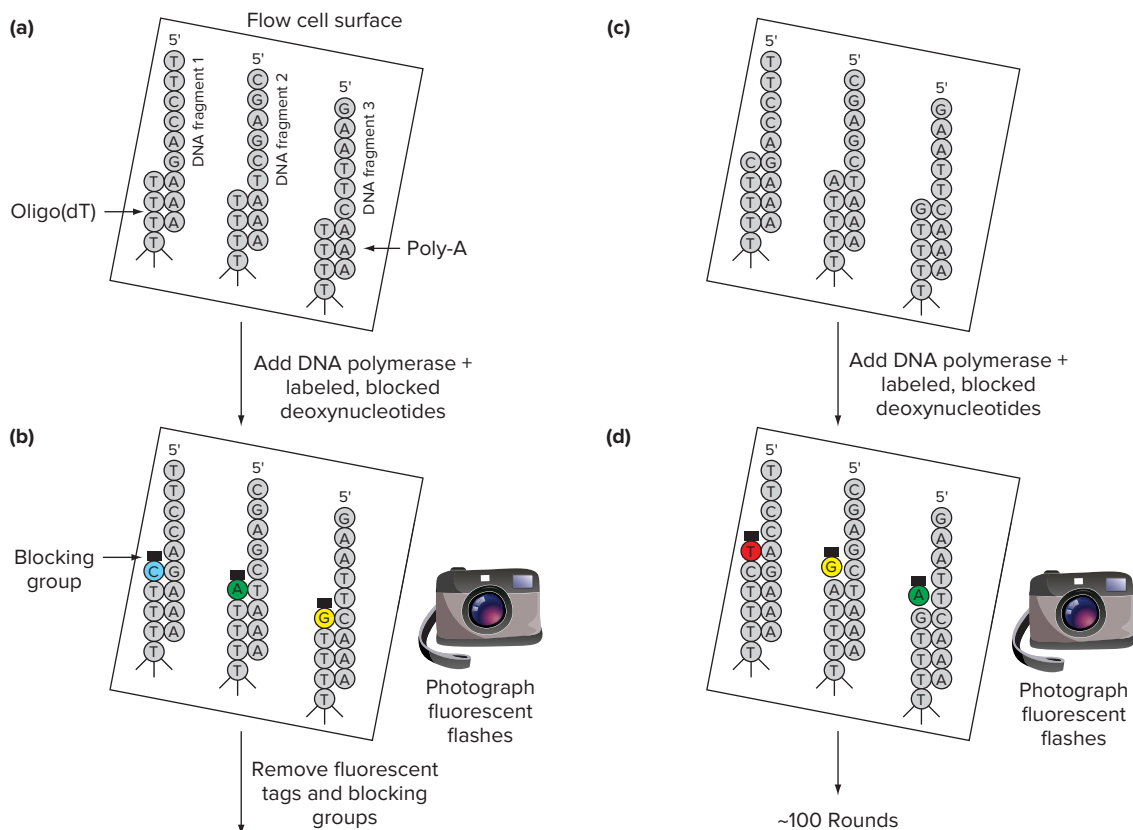
combination of these three innovations allows sequencing machines to record the successive addition of nucleotides to each of millions of growing DNA molecules in real time.

Figure 11.24 outlines only one of many ingenious technologies for inexpensive high-throughput sequencing currently under development. Other prototype systems are based on very different ideas; for example, the figure at the beginning of this chapter illustrates a novel method in which individual DNA molecules are threaded through small channels called *nanopores*. It is not clear which methods will become standard in the future as costs are steadily driven lower. But you should have no doubt that the era of whole-genome sequencing has already arrived.

## Disease-Causing Mutations Are Hidden in a Sea of Variation

A patient's whole-exome or whole-genome sequence should include the sequence difference(s) responsible for

**Figure 11.24** One method for high-throughput, single molecule DNA sequencing. **(a)** Millions of single-stranded genomic DNA fragments, to which poly-A has been enzymatically added at the 3' end, are hybridized to oligo-dT molecules attached to the surface of a special microarray called a *flowcell*. **(b)** Using the genomic fragment as template and the oligo-dT as primer, DNA polymerase synthesizes new DNA containing nucleotides with colored, base-specific fluorescent tags. These nucleotides are also blocked at their 3' ends so that only one nucleotide can be added at a time. **(c)** After a high-resolution camera photographs the fluorescence, chemicals applied to the flowcell remove the tag and the blocking group from the just-added nucleotide. **(d)** Each subsequent cycle begins by infusing the flowcell with a new dose of tagged nucleotides and polymerase, and then step (c) is iterated. The sequencing machine takes about 100 pictures that record a sequence of colored flashes at each of the millions of spots where a single DNA molecule is being synthesized. A computer rearranges the data into millions of short sequence reads of about 100 nucleotides and then assembles the genome sequence.



the genetic disease, but possession of this sequence information does not guarantee that geneticists will be able to identify the responsible mutation(s). One problem is technical: No genome sequence is 100% accurate or 100% complete. All sequencing methods have a low but real error rate in identifying nucleotides, and random sampling of DNA fragments will leave some regions of the genome unsequenced. These issues can be minimized by coverage of 10 or more genome equivalents, but they cannot be eliminated completely.

An even more fundamental problem is that the amount of variation among human genomes is huge. We saw at the beginning of this chapter that any person's genome differs at more than 3 million locations from the standard RefSeq human genome. How can we tell which of these millions of DNA polymorphisms causes a patient's disease? Our ability to deal with whole-genome sequences is still so limited that in many cases, the responsible mutation has yet to be identified. It should be lurking in the sequence, but it frustratingly remains hidden in front of our noses.

Despite these issues, investigators have been able to marshal the results of several types of data analysis, sometimes supported by inspired guesswork, to find an increasing number of disease genes. We focus in this section on the types of clues geneticists use to identify disease-causing mutations within whole-genome/exome sequences. However, it is crucial to keep in mind that these methods are not, at least not yet, always successful.

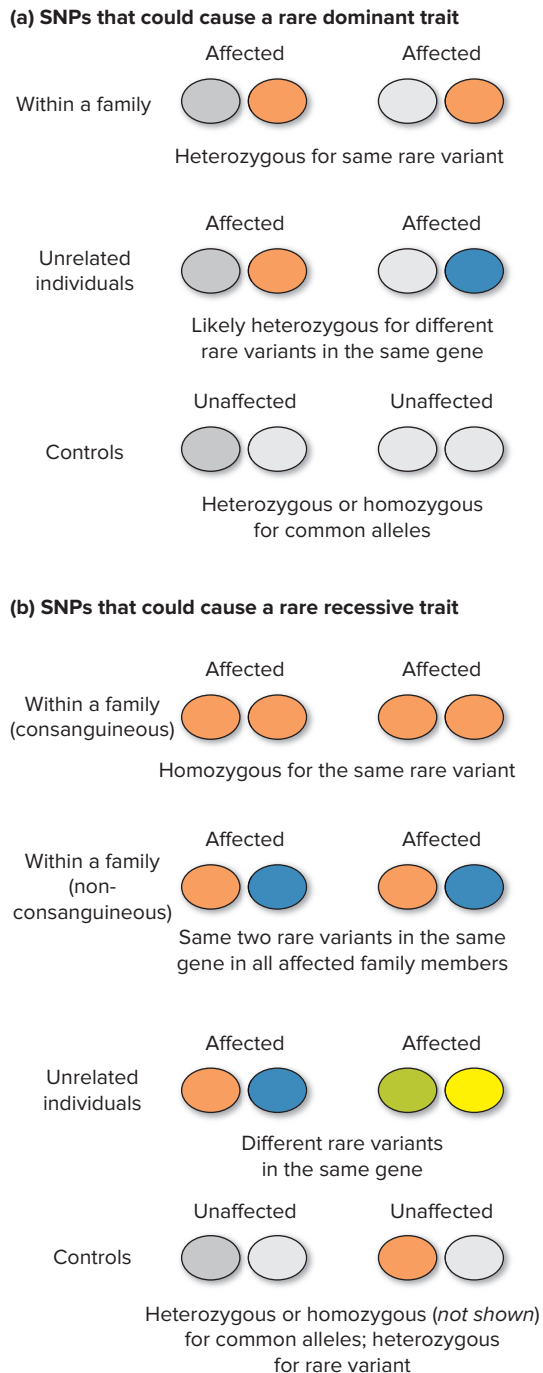
### Clues from disease transmission patterns

The underlying logic of whole-genome or whole-exome sequencing requires that the DNA variants that are disease alleles will be rare in the population. This basic assumption allows scientists to make predictions about which of the variations in a patient's genome could be responsible for the disease. These predictions depend on what pedigrees say about the disease's inheritance: Is the disease allele recessive or dominant? Is it sex-linked or autosomal? Is the penetrance complete or incomplete? Each of these inheritance modes is consistent only with particular molecular genotypes at a candidate locus.

In the case of a rare dominant condition, it is highly likely that the patient would be heterozygous for the causative allele. Related patients should have the same rare mutant allele, whereas unrelated patients might have different mutations in the same gene (Fig. 11.25a). If the condition is recessive, geneticists would first focus their attention on rare mutations that are homozygous in the patient's genome, particularly if the parents are related even distantly. If the condition is recessive and the parents are unrelated, the patient could instead be a compound heterozygote, with two different mutant alleles of the same gene (Fig. 11.25b). To check this latter scenario, geneticists would look in the patient's DNA for a gene affected

**Figure 11.25** SNP patterns consistent with inherited traits.

Each oval represents a copy of a gene, so each person corresponds to two ovals. Common variants are in different shades of gray; while orange, blue, yellow, or green symbolize different rare variants in the same gene. **(a)** Variants that could cause a dominant trait. Within a family, affected individuals will be heterozygotes for the same rare variant. Unrelated affected people may be heterozygotes for different rare variants in the same gene. **(b)** Variants that could cause a recessive trait. In consanguineous families, affected individuals will be homozygotes for a single variant they inherit by descent from a recent common ancestor. The affected children of unrelated people most likely are compound heterozygotes who inherit different rare mutations in the same gene, one from each parent. In both (a) and (b), unaffected controls may or may not be related.





by two different mutations. Finally, if an inheritance pattern shows sex linkage, the search for candidate genes would be limited to the X chromosome; if autosomal, the X chromosome would be excluded.

DNA sequence information from the patient's relatives is particularly useful in winnowing down the list of candidate polymorphisms. As an example, SNP genotyping of relatives using microarrays, as discussed earlier in this chapter, could narrow the search to a region between two known SNPs. Positional cloning and whole-genome sequencing are thus not mutually exclusive approaches to disease gene identification; instead, they can provide complementary information. Better yet, though more expensive, would be comparisons of the patient's whole-genome or exome sequence with those of parents and/or siblings.

A recent case study illustrates the power of DNA sequence information from related individuals (Fig. 11.26). A brother and sister had Miller syndrome, a rare condition affecting development of the face and limbs, but neither parent was affected (Fig. 11.26a). These facts suggest (but don't prove) that Miller syndrome is a recessive autosomal condition, with the two children inheriting mutant alleles from both of their heterozygous, carrier parents. To find the Miller syndrome gene, researchers sequenced the entire genomes of both parents and both children; in fact, this was the first time in history that the genomes of all the members of a nuclear family were sequenced completely.

**Figure 11.26** The first family with completely sequenced genomes. (a) Pedigree for Miller syndrome. (b) Map showing allele inheritance along chromosomes 16 and 17 in the affected children. In *identical* regions, the affected brother and sister share the same maternally and paternally derived alleles. In *nonidentical* regions, the siblings share no alleles. In *haploidentical maternal* regions, the siblings have the same allele from the mother but different alleles from the father. In *haploidentical paternal* regions, the brother and sister share a common allele from the father but have different alleles from the mother. If Miller syndrome is recessive, the responsible gene should lie in an identical region. This prediction was upheld when mutations in the *DHOD* gene on chromosome 16 were found to cause the disease.

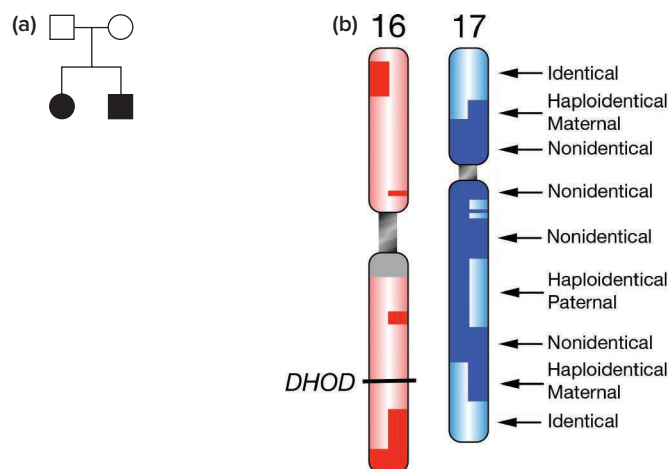


Figure 11.26b presents a graphical summary of part of this gigantic data set, showing the landscape of recombination that occurred on just two chromosomes during meiosis in the mother and father to produce the gametes resulting in the affected children. The hypothesis that Miller syndrome is recessive predicts that the responsible gene would lie in a region where the affected son and daughter share the same allele from the mother and also the same allele from the father (regions labeled *identical* in the figure). Geneticists studying the disease could thus focus their attention only on the approximately 25% of the genome where this was the case (Fig. 11.26b). We describe the outcome of this investigation later on in this chapter.

### Clues from a variant's predicted effect on gene function

Researchers first try to look for disease-causing mutations in the protein coding regions of the exome because these parts of the genome are the easiest to look at: Coding regions constitute only a small fraction of the total genome, and alterations in the coding region are the most straightforward to interpret. In particular, investigators would search for rare polymorphisms that change the identity of an amino acid (that is, SNPs causing missense mutations) or alter the reading frame (SNPs, DIPs, or SSRs causing nonsense or frameshift mutations). Most nonsense/frameshift mutations and a subset of missense mutations will be *nonanonymous DNA polymorphisms* that affect phenotype through changes in protein function. In contrast, silent mutations that change a codon into a different codon for the same amino acid will not affect phenotype; these anonymous SNPs can therefore be discarded as candidates.

The assumption that a particular genetic disease results from a mutation in a protein-coding exon is nonetheless very uncertain. Some genetic diseases are caused not by alterations in amino acid sequence, but rather by the amount of a protein that the organism produces. Mutations that reside in regions of the genome outside of the exome could, for example, lower or prevent the transcription of a gene or the splicing of its primary transcript. Either case would lower the amount of the gene's protein product or even prevent its synthesis. Such mutations would never be found if researchers focused their attention only on the exome. And unfortunately, we still understand so little about the DNA sequences that regulate transcription or splicing that many such mutations will be overlooked even when the patient's whole-genome sequence is available.

### Clues from previously determined genome sequences

Rare diseases are unlikely to be caused by variants common in the human population. As a result, variants that have been documented in databases as common are poor

**TABLE 11.2** Finding the Disease-Causing Mutation in Nic Volker's Genome

Analysis Step	Candidate Variants Remaining
Sequence of Nic's exome	16,124
Filter for missense mutations changing an amino acid	7,157
Filter for novel variants not previously reported in databases	878
Filter for variants that are X-linked or display a recessive pattern in Nic's exome	136
Filter for variants that change evolutionarily conserved amino acids	35
Filter for variants in genes that are not frequently mutated in the general population	5
Filter for variants in genes known to be mutated in other genetic diseases of some relevance	1 ( <i>XIAP</i> )

candidates for disease-causative mutations. In contrast, it is possible that different patients suffering from similar, or at least vaguely related symptoms, may share either the same rare mutation or they may have different rare mutations in the same gene.

Even genomes of other organisms can point genetic sleuths in the proper direction. For example, consider a gene that has been closely conserved during evolution so that some of the amino acids in the protein product are identical in diverse organisms, such as humans and fruit flies. Conservation suggests that these particular amino acids play crucial roles in the protein's function. If you now find a rare variant in a patient that changed such a highly conserved amino acid to a different one, this mutation would be a strong candidate because of the likelihood it would affect phenotype.

### Pinpointing a Disease Gene Requires a Combination of Approaches

The beginning of this chapter introduced the case of Nic Volker, one of the first patients to be treated successfully for a genetic disease based on identification of the causative gene through sequencing of his personal exome. When Nic's DNA was characterized in 2009, sequencing was still relatively expensive, so investigators sequenced only Nic's exome, rather than his whole genome or the exomes/genomes of his parents or siblings. **Table 11.2** examines how geneticists analyzed Nic's exome, using many sources of information to narrow down the candidate variants to the single responsible mutation.

Nic's exome contained about 16,000 variants relative to the human standard RefSeq (Table 11.2). These variants were mostly SNPs, but some DIPs and SSRs were also found. Researchers started to winnow this list by excluding likely anonymous variants in the exons (that is, silent mutations in codons, or mutations in the sequences encoding the 5' and 3' UTRs of mRNAs). They next focused their attention on novel variants that had not been recorded

in databases previously, in this way ignoring common variants known to exist in the genomes of normal individuals. The following step was to filter the list for mutations consistent with X-linked or recessive inheritance, as these were the most likely scenarios for a disease that affected Nic but not his parents or other relatives.

At this point, the list of good candidates for the cause of Nic's condition had been narrowed to 136 variants (Table 11.2). The researchers now took an evolutionary approach and asked whether any of these changes might have altered the identity of an amino acid that was tightly conserved in many diverse species. The next-to-last step of the analysis was to examine the remaining candidates for those that were in genes known from databases to be never or only infrequently inactivated (for example, by nonsense or frameshift mutations) in the general population. The idea behind this step was to ignore genes that are defective in many normal individuals and are thus unlikely to influence disease phenotypes.

Five candidate variants remained in the investigators' target list. The researchers realized that one of these candidate mutations was in a gene called *XIAP*. Other mutations in *XIAP* were known to cause a serious condition called *X-linked lymphoproliferative disease (XLPD)*, in which the blood contains too many lymphocytes (white blood cells of the immune system), crowding out the oxygen-carrying red blood cells and damaging the liver. These symptoms were very different from Nic's, but XLPD had some vague relevance to Nic's case because the immune system is clearly involved in gastrointestinal inflammation.

The researchers became particularly excited when they noticed that the variation in Nic's *XIAP* gene was a missense mutation changing in the protein product the identity of a single amino acid that is completely conserved among humans, frogs, flies, and many other species (**Fig. 11.27**). XLPD-causing mutations in the same gene are instead nonsense and frameshift mutations that would prevent synthesis of full-length protein, potentially explaining the difference in Nic's symptoms and in those of XLPD patients.

**Figure 11.27 A mutation in a conserved amino acid in the XIAP protein.** Amino acids 195–211 of the XIAP protein in the one-letter code. Compared to the human RefSeq XIAP protein (second row), Nic Volker's XIAP (first row) has an amino acid substitution at position 203, from cysteine (C) to tyrosine (Y). In all other species examined, cysteine is found at this position, suggesting that the mutation in Nic's genome might alter XIAP function.

	195	200	205	210													
Nic's XIAP	G	D	Q	V	Q	C	F	C	Y	G	G	K	L	K	N	W	E
Human	G	D	Q	V	Q	C	F	C	C	G	G	K	L	K	N	W	E
Chimpanzee	G	D	Q	V	Q	C	F	C	C	G	G	K	L	K	N	W	E
Mouse	D	D	Q	V	Q	C	F	C	C	G	G	K	L	K	N	W	E
Dog	D	D	Q	V	Q	C	F	C	C	G	G	K	L	K	N	W	E
Cow	D	D	Q	V	Q	C	F	C	C	G	G	K	L	K	N	W	E
Chicken	D	D	Q	V	Q	A	F	C	C	G	G	K	L	K	N	W	E
Zebrafish	D	D	N	V	Q	C	F	C	C	G	G	L	S	G	W	E	
Frog	R	D	H	V	K	C	F	H	C	D	G	L	R	N	W	E	
<i>Drosophila</i>	L	D	H	V	K	C	V	W	C	N	G	V	I	A	K	W	E

Although this extensive bioinformatics analysis did not prove that the *XIAP* variation caused Nic's condition, it was an excellent candidate that fulfilled all the criteria examined in Table 11.2. What is more, if this identification was correct, Nic's disease might be *actionable*, meaning that something could be done to alleviate his condition. Maybe he could be treated by a method known to help XLPD patients: namely, a transplant into his bone marrow of umbilical cord blood from a newborn infant. This treatment would in theory provide Nic with a self-renewing source of stem cells that could continuously produce normal lymphocytes. Within a year of the transplantation procedure, Nic's condition improved remarkably (see Fig. 11.3).

## The Study of Human Genetics Is an Ongoing Venture

Nic Volker is one of an increasing number of patients whose suffering has been ameliorated by information gleaned from whole-exome/genome sequencing. But the outcome of such studies is not always so favorable. For example, using bioinformatics filters similar to those employed in Nic's case, investigators were able to identify the mutations responsible for Miller syndrome that affected the brother and sister previously shown in Fig. 11.26. These siblings inherited one mutation in a gene called *DHOD* from their mother, and a different mutation in *DHOD* from their father; that is, they were compound heterozygotes. The Miller syndrome gene identification was unfortunately not actionable in terms of suggesting any kind of treatment. However, knowledge that *DHOD* is the Miller syndrome gene may nonetheless in the future help these patients to guide their reproductive decisions. If their

partners do not carry a mutation in *DHOD*, none of their children would be affected by this condition.

In many other cases, the whole-genome sequence has not yet even allowed researchers to identify the responsible mutation. Perhaps the mutation lies in a poorly covered part of the sequence; perhaps the mutation lies outside of the protein-coding exome in sequences whose function has not yet been determined; perhaps the researchers made an incorrect (even though reasonable) assumption in one step of their bioinformatics analysis.

One lesson from Nic's story is the importance of databases and shared information. Critical steps in the identification of his disease-causing mutation depended on knowledge of variants from many other people's genomes, including unaffected controls and individuals affected by other genetic conditions sharing little or no phenotypic similarity with his own symptoms. The practice of human genetics is thus a giant bootstrapping operation: The more genomes that are sequenced, the more information is available to aid the analysis of all new genomes. The progress of human genetics therefore requires that databases be kept up to date and that their vast information be made accessible to all investigators using common methods of archiving. Allowing this degree of access, while preserving the confidentiality of the individuals whose genomes are cataloged, is a significant challenge for the future.

One of the most important databases for studies of human genetics is called **Online Mendelian Inheritance in Man (OMIM; www.omim.org)**. OMIM is a catalog of human genes and the traits they control. The database lists the known variants in human genes that are associated with particular diseases or other traits and provides links to published research articles about these variants. Updated daily with new research findings, OMIM is an invaluable resource for researchers like those who figured out the genetic cause of Nic Volker's symptoms. This online database is so useful and easy to use that we encourage you to explore it on your own.

### essential concepts

- High-throughput technologies allow parallel sequencing of millions of individual DNA molecules. These new methods are rapidly driving down the cost of whole-exome and whole-genome sequencing.
- Finding disease-causing mutations among the many DNA variations that distinguish individual genomes involves sequential filtering of information. These steps may involve deduction of likely transmission patterns, analysis of relatives' genomes, knowledge of similar genetic diseases, and predictions regarding a variant's effect on protein function.

## WHAT'S NEXT

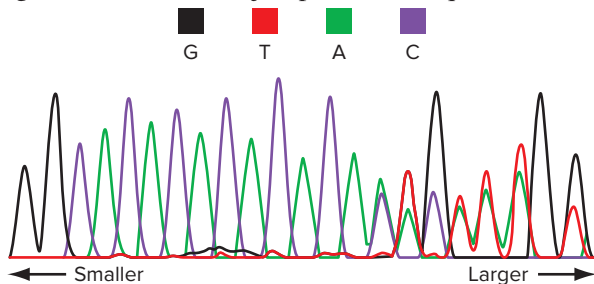
In this chapter and in Chapters 9 and 10, we have focused on the nucleotide content of genomes, particularly the 6 billion nucleotides organized into 46 chromosomes in each normal human diploid cell. In the next several chapters, we examine features of the chromosomes that allow these DNA sequences to function properly and to be transmitted from one generation to the next.

We begin by considering how in spite of the enormous complexity of DNA sequences, the DNA actually constitutes only about one-third of the total mass of a chromosome. The remainder of the chromosome is made of thousands of

different types of proteins that help package and manage the information carried by DNA. These proteins have many roles. Certain proteins help compact the chromosomes to fit in the nucleus. Some proteins ensure that the chromosomal DNA is properly duplicated during each cell cycle, while others govern the distribution of chromosomes to daughter cells. Yet other proteins are responsible for regulating the availability of genes to the transcriptional machinery so that the genes can be expressed into proteins. In Chapter 12, we examine how proteins interact with DNA to generate the functional complexity of a chromosome.

## SOLVED PROBLEMS

- I. Genomic DNA from a woman's blood cells is PCR amplified by a single pair of primers representing a unique locus in the genome. The PCR products are then sequenced by the Sanger method, using one of the PCR primers as a sequencing primer. The following figure shows a trace of just part of the sequence read.



- What kind of polymorphism is most likely represented?
- With your answer to part (a) in mind, determine the woman's genotype at this locus. Indicate all nucleotides that can be read from both alleles and their 5'-to-3' orientation.
- What kind of molecular event was likely to have generated this polymorphism?
- How would you know exactly where in the genome this locus is found?
- What is another way in which you could analyze the PCR products to genotype this locus?
- Suppose you wanted to genotype this locus based on single-molecule DNA sequencing of whole genomes as shown in Fig. 9.24. Would a single read suffice for genotyping the locus by this alternative method?

## Answer

To solve this problem, you need to understand that PCR will simultaneously amplify both copies of a locus (one on the

maternally derived chromosome and one on the paternally derived chromosome), as long as the primer can hybridize to both homologs as is usually the case. The DNA sequence trace has two nucleotides at several positions. This fact indicates that the woman must be a heterozygote and that the PCR is amplifying both alleles of the locus.

- Notice that both alleles contain multiple repeats of the dinucleotide CA. The most likely explanation for the polymorphism is therefore that **the locus contains an SSR polymorphism whose alleles have different numbers of CA repeats**. One allele has six repeats; the second allele must have more CA units.
- Writing out the first 14 nucleotides of both alleles is straightforward. If the assumption in part (a) is correct, then one allele should have more than six CA repeats. The trace shows evidence for two additional CA repeats in one allele at positions 15–18, for a total of eight CA repeats.

You can then determine the nucleotides beyond the repeats in the shorter allele by subtracting CACA from positions 15–18. The remaining peaks at these positions correspond to ATGT. Note that ATGT can also be found in the longer allele, but now at nucleotides 19–22, just past the two additional CACA repeats. You can determine the last four nucleotides in the shorter allele by subtracting ATGT from positions 19–22, revealing TAGG. **The sequences of the two alleles of this SSR locus (indicating only one strand of DNA each) are thus:**

Allele 1: 5'...GGCACACACACACAATGTTAGG...3'  
 Allele 2: 5'...GGCACACACACACACAATGT...3'

- The mechanism thought to be responsible for most SSR polymorphisms is **stuttering of DNA polymerase during DNA replication**.

- d. You actually knew the location of this locus even before starting the experiment. This is because you design the PCR primers from knowledge of the entire human genome sequence.
- e. The polymorphism involves a difference in the number of repeat units, and therefore the two alleles would produce PCR products that differ in length. You could genotype this locus by gel electrophoresis of the PCR products, as shown in Fig. 11.12.
- f. Direct Sanger sequencing of a PCR product from genomic DNA produces a trace that includes both alleles. This is not true of single-molecule DNA sequencing techniques. You would require enough sequence runs from individual genomic DNA molecules to ensure that you could see both alleles if the person was a heterozygote.

II. It is difficult to obtain accurate recombination frequencies in humans because family sizes are small. An interesting way to circumvent this problem is to genotype individual sperm cells so as to obtain large data sets for linkage studies. The table that follows shows the genotype of four SNP loci from 20 single sperm that one man provided for this research. The genotypes were determined by microarray analysis of four SNP loci amplified from these samples by PCR. In the table, A, C, G, and T are the alleles of the SNPs (that is, the nucleotides on one strand) and a dash (—) means that no DNA corresponding to the locus was amplified from the sample.

SNP:	Locus 1	Locus 2	Locus 3	Locus 4
<b>Sperm number</b>				
1	G	C	—	T
2	G	A	G	C
3	G	C	G	C
4	G	C	G	T
5	G	A	—	C
6	G	C	—	T
7	G	A	G	C
8	G	A	—	C
9	G	A	G	T
10	G	C	—	T
11	G	C	—	T
12	G	C	G	T
13	G	A	—	C
14	G	A	—	C
15	G	C	G	T
16	G	A	G	C
17	G	C	—	T
18	G	A	—	T
19	G	A	G	C
20	G	C	G	T

- a. Which SNP loci could be X-linked?
- b. Which SNP loci could be on the Y chromosome?
- c. Which SNP loci must be autosomal?
- d. For any autosomal SNP loci, what is the sperm donor's genotype in somatic tissue?
- e. Do any SNP loci appear to be linked to each other?
- f. What is the distance between any two linked SNP loci?

### Answer

SNP analysis by PCR is so sensitive that the single alleles present within individual sperm cells can be assayed, providing researchers with considerable information. You should also remember that a man's somatic cells have two copies of each autosome, one X chromosome, and one Y chromosome. Individual sperm will have one copy of each autosome and either an X or a Y chromosome.

- a. For any X-linked SNP locus, half of the sperm cells will carry the same SNP allele, but the other half of the sperm would not have an X chromosome and would thus not yield a PCR product. Locus 3 shows this type of pattern.
- b. Similarly, a locus on the Y chromosome would be found in only half the sperm, and all these sperm would have the same allele. Again, Locus 3 is a candidate for a Y-linked SNP. The data do not allow you to discriminate between an X or a Y chromosome location for Locus 3.
- c. For an autosomal SNP locus, all the sperm will have one copy of the locus. Loci 1, 2, and 4 are thus autosomal.
- d. If a man is homozygous for a single allele of an autosomal SNP locus, all the sperm he produces will have this one allele. If he is a heterozygote for two different alleles, approximately one-half of the sperm will have one allele and the other half of the samples will carry the other allele for the locus. The man's genotypes for the autosomal genes are SNP Locus 1: GG (homozygous); SNP Locus 2: CA (heterozygous); and SNP Locus 4: CT (heterozygous).
- e. Alleles at linked loci will segregate together (end up in the same sperm) more than 50% of the time. This is true of the C allele of SNP Locus 2 and the T allele of Locus 4. The reciprocal alleles (A for Locus 2 and C for Locus 4) are also transmitted together more often than not. SNP Loci 2 and 4 are linked.
- f. Sperm 3, 9, and 18 show evidence of recombination between alleles at Loci 2 and 4. Three out of 20, or 15%, of the sperm are recombinant. The distance between Loci 2 and 4 is therefore 15 cM.


**PROBLEMS**
**Vocabulary**

1. Choose the phrase from the right column that best fits the term in the left column.
 

a. DNA polymorphism	1. DNA element composed of short tandemly repeated sequences
b. phase	2. two different nucleotides appear at the same position in genomic DNA from different individuals
c. informative cross	3. arrangement of alleles of two linked genes in a diploid
d. ASO	4. location on a chromosome
e. SNP	5. a DNA sequence that occurs in two or more variant forms
f. DNA fingerprinting	6. a short oligonucleotide that will hybridize to only one allele at a chosen SNP locus
g. SSR	7. detection of genotype at a number of unlinked highly polymorphic loci
h. locus	8. allows identification of a gamete as recombinant or nonrecombinant
i. compound heterozygote	9. all exons in a genome
j. exome	10. individual with two different mutations in the same gene

**Section 11.1**

2. Would you characterize the pattern of inheritance of anonymous DNA polymorphisms as recessive, dominant, incompletely dominant, or codominant?
3. Would you be more likely to find single nucleotide polymorphisms (SNPs) in the protein-coding or in the noncoding DNA of the human genome?
4. A recent estimate of the rate of base substitutions at SNP loci is about  $1 \times 10^{-8}$  per nucleotide pair per gamete.
  - a. Based on this estimate, about how many *de novo* mutations (that is, mutations not found in the genomes of your parents) are present in your own genome?
  - b. Where and when did these *de novo* mutations in your genome most likely occur?
  - c. It has been calculated that each sperm made in a 25-year-old man is the result on average of about 300 rounds of cell division, starting with the first mitotic division of the male zygote. In contrast, each mature oocyte found in a 5-month-old female human fetus is the result of about 25 rounds of division, starting with the first mitotic division of the female zygote. What bearing do these calculations have on the estimate of the rate of base substitutions in humans, and on your answer to part (b)?
5. If you examine Fig. 11.5 closely, you will note that in some regions, such as between nucleotides 116,870 K and 116,890 K, James Watson and Craig Venter share the same SNPs, and these regions are surrounded by others in which these two men do not share any SNPs. What does this fact say about the relationship between these two men, and how do you think this pattern of shared and unshared SNPs arose?
6. Approximately 50 million SNPs have thus far been recorded after the characterization of thousands of human genomes.
  - a. About how many base pairs in the human genome are identical in these thousands of people?
  - b. Do you think that many other SNPs exist among the human population? If so, why haven't they been found?
  - c. Almost all of the SNP polymorphisms found to date are biallelic; that is, among all the genomes in the population studied to date, only two possible alleles can be found (for example, A and C). Provide a rough estimate for the number of triallelic SNP loci that could be found in the same group of humans (that is, the number of loci with three different alleles—for example, A, C, and T). At about how many loci would all four possible nucleotides be found among the human genomes studied to date?
7. Mutations at simple sequence repeat (SSR) loci occur at a frequency of  $1 \times 10^{-3}$  per locus per gamete, which is much higher than the rate of base substitutions at SNP loci (whose frequency is about  $1 \times 10^{-8}$  per nucleotide pair per gamete).
  - a. What is the nature of SSR polymorphisms?
  - b. By what mechanism are these SSR polymorphisms likely generated?
  - c. Copy number variants (CNVs) also mutate at a relatively high frequency. Do these mutations occur by the same or a different mechanism than that generating SSRs?
  - d. The SSR mutation rate is much higher than the mutation rate for new SNPs. Why then have geneticists recorded more than 50 million SNP loci but only about 100,000 SSR loci in human genomes?
8. Humans and gorillas last shared a common ancestor about 10 million years ago. Humans and chimps last shared a common ancestor about 6 million years ago. The table that follows shows the corresponding genomic region from two gorilla gametes, three chimpanzee gametes, and three human gametes.

- Draw a cladogram similar to that in Fig. 11.6b to show the evolutionary relationships among these three species.
- The data reveal six polymorphisms among these eight genomes, at positions 2 (A or G), 3 (A or T), 4 (G or T), 7 (C or T), 8 (C or T), and 9 (G or T). On your cladogram, indicate approximately when the mutations that produced each of these polymorphisms occurred. For each allele, state whether it is ancestral, derived, or that you can't tell.
- Infer the sequences in (i) the last common ancestor of humans and chimpanzees, and (ii) the last common ancestor of all three species. Use a question mark (?) to represent any uncertainty.

Genome (haploid)	Gorilla	Chimpanzee	Human
1	CATGTCCTGA	CGAGTCCTGA	CAAGTCCTGA
2	CATGTCCTTA	CAAGTCCTGA	CAATTCCTGA
3		CAAGTCCCGA	CAATTCCTGA

- In 2015, an international team of scientists assembled the complete genome sequences of two different woolly mammoths. Both specimens were discovered buried in the permafrost of Siberia, the coldest inhabited place on earth. Through radiocarbon dating, it was determined that one of the mammoths, found on Wrangel Island off the Siberian coast, died about 4000 years ago; the other mammoth, found in the town of Oimyakon, died about 45,000 years ago.

Analysis revealed that the genome sequences of these two animals differed significantly in the distribution of base pairs at which they are either homozygous or heterozygous. The Wrangel Island woolly mammoth had an extreme excess of *runs of homozygosity (ROHs)*, regions in which the animal was homozygous for all of the base pairs. About 23.4% of the Wrangel Island animal's genome was composed of ROHs that were greater than 500 kb in length; some of these ROHs were in excess of 5 Mb long. In contrast, only 0.83% of the Oimyakon animal's genome consisted of ROHs longer than 500 kb.

- Explain how polymorphisms are detected when sequencing a genome. How would researchers know, for any particular base pair, whether a genome is homozygous or heterozygous?
- What does the extreme excess of ROHs in the Wrangel Island mammoth genome suggest about that animal's parents?
- The Wrangel Island woolly mammoth is thought to have belonged to the last population on earth before the species went extinct about 4000 years ago. The answer to part (b) suggests one possible reason for the woolly mammoth's extinction. Explain.

## Section 11.2

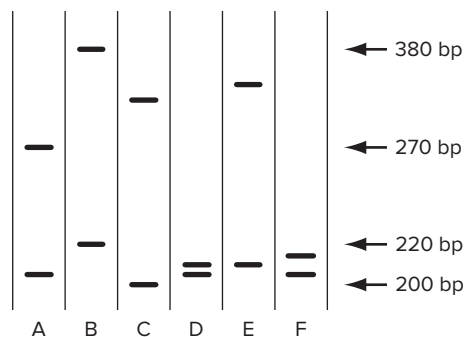
- Using PCR, you want to amplify an approximately 1 kb exon of the human autosomal gene encoding the enzyme phenylalanine hydroxylase from the genomic DNA of a patient suffering from the autosomal recessive condition phenylketonuria (PKU).
  - Why might you wish to perform this PCR amplification in the first place, given that the sequence of the human genome has already been determined?
  - Calculate the number of template molecules that are present if you set up a PCR reaction using 1 nanogram ( $1 \times 10^{-9}$  grams) of chromosomal DNA from blood cells as the template. Assume that each haploid genome contains only a single gene for phenylalanine hydroxylase and that the molecular weight of a base pair is 660 grams per mole. The haploid human genome contains  $3 \times 10^9$  base pairs.
  - Calculate the number of PCR product molecules you will obtain if you perform 25 PCR cycles and the yield from each cycle is exactly twice that of the previous cycle. What would be the mass of these PCR products taken together?
- Which of the following set(s) of primers a–d could you use to amplify the following target DNA sequence, which is part of the last protein-coding exon of the *CFTR* gene?

```
5' GGCTAAGATCTGAATTTTCCGAG ... TTGGGCAATAATGTAGCGCCTT 3'
3' CCGATTCTAGACTTAAAAGGCTC ... AACCCGTTATTACATCGCGGAA 5'
```

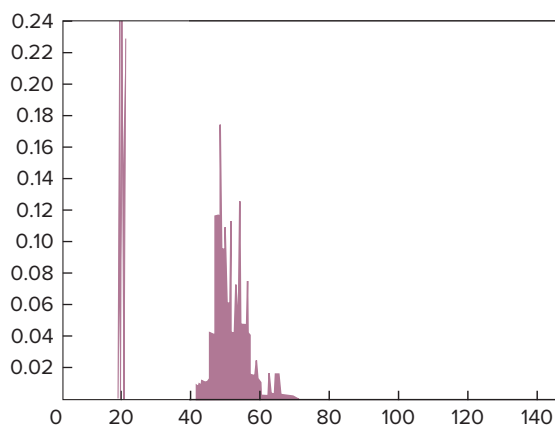
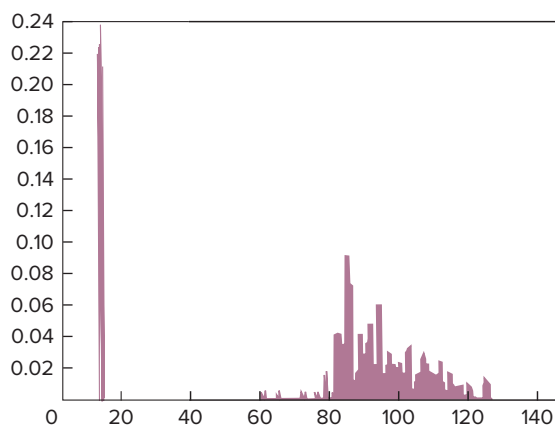
- 5' GGAAAATTCAGATCTTAG 3';  
5' TGGGCAATAATGTAGCGC 3'
  - 5' GCTAAGATCTGAATTTTC 3';  
3' ACCCGTTATTACATCGCG 5'
  - 3' GATTCTAGACTTAAAGGC 5';  
3' ACCCGTTATTACATCGCG 5'
  - 5' GCTAAGATCTGAATTTTC 3';  
5' TGGGCAATAATGTAGCGC 3'
- The previous problem raises several interesting questions about the design of PCR primers.
    - How can you be sure that the two 18-nucleotide-long primers you chose as your answer to Problem 11 will amplify only an exon of the *CFTR* gene, but no other region, from a sample of human genomic DNA?
    - Primers used in PCR are generally at least 16 nucleotides long. Why do you think the lower limit would be approximately 16?
    - Suppose one of the primers in your answer to Problem 11 had a mismatch with a single base in the genomic DNA of a particular individual. Would you be more likely to obtain a PCR product from this genomic DNA if the mismatch was at the 5' end or at the 3' end of the primer? Why?

13. You want to make a recombinant DNA in which a PCR product amplified from the human genome is inserted into a plasmid vector. The polylinker of this vector includes recognition sites for the enzymes *EcoRI* (5' G<sup>^</sup>AATTC 3') and *BamHI* (5' G<sup>^</sup>GATCC 3'). (The <sup>^</sup> symbolizes the cut site in the DNA.) PCR primers that could amplify the fragment of human DNA are: 5' GCTACTTCGCGTATTCCA 3' and 5' CCCAAGTCCTAGCCGATA 3'.
- Describe in detail how these primers would need to be modified to create a fragment of the human genome flanked by *EcoRI* sticky ends so that this fragment could be cloned easily into the plasmid vector. You will need to consider the fact that most restriction enzymes, including *EcoRI*, cannot cut DNA if the restriction site is directly at the end of the DNA molecule; the restriction enzyme recognition site must be at least six base pairs distant from the end.
  - Describe a potential feature of the PCR-amplified region of the human genome that could prevent you from using the strategy you described in part (a).
  - Now describe how the primers must be modified to create a human DNA fragment with an *EcoRI*-compatible single-stranded overhang at one end and a *BamHI*-compatible overhang at the other end. (Two possibilities exist; you need to describe only one. Assume that a *BamHI* site also must be at least six base pairs from the end of the DNA.) Why might you want to make such a fragment?
14. You sequence a PCR product amplified from a person's genome, and you see a double peak such as that seen in Fig. 11.11b. Most of the time, this result indicates that the person is a heterozygote for a SNP at that position. But it is also possible that the result is due to a mistake in DNA replication during the PCR amplification, with DNA polymerase misincorporating the wrong nucleotide.
- If you saw an artifactual double peak in the sequence trace, did the mistake happen in the first few rounds of PCR amplification or in the last few rounds?
  - Whether or not you see a double peak, is it more likely that a mistake would happen in the first few rounds of PCR amplification or in the last few rounds?
  - Given that mistakes can happen during PCR amplification, what could you do to be sure of a person's genotype? Why would this degree of certainty be difficult to achieve if you were doing preimplantation genotyping of embryos?
  - PCR relies on heat-stable DNA polymerases from thermophilic bacteria that grow in hot springs. The DNA polymerase originally used for PCR, from the bacterium *Thermus aquaticus*, lacks the 3'-to-5' exonuclease found in other DNA polymerases such as that from *E. coli* (review Fig. 7.9). Why do scientists now most often use DNA polymerase from a different thermophilic bacterium (*Pyrococcus furiosus*) that does contain this exonuclease function?
15. Problem 8 shows three different sequences of the same autosome in human populations. These sequences are each from a single chromatid. You know this to be true because the PCR amplifications were from individual haploid gametes. If you wanted to obtain the same information by PCR amplification of genomic DNA from somatic cells, the problem would be somewhat more complicated because the starting cells are diploid. Each PCR product to be sequenced would thus actually be amplified from two homologous chromosomes. You could still verify the existence of the three different haploid sequences shown in Problem 8 by analyzing the somatic genomic DNA from as few as two people (if they happened to be the right people).
- Indicate the diploid genotypes of two people from whom you could identify these three different haploid sequences. Account for all 10 nucleotides in the sequences. Three possible correct answers exist; you need to show only one.
  - If you PCR amplified DNA from somatic genomic DNA from a person with one particular genotype, you would not be able to conclude that their genome contains any of these three sequences. What is the genotype of this person? Explain why you could not reach this conclusion.
16. The trinucleotide repeat region of the Huntington disease locus (*HD*) in six individuals is amplified by PCR and analyzed by gel electrophoresis as shown in the following figure; the numbers to the right indicate the sizes of the PCR products in bp. Each person whose DNA was analyzed has one affected parent.
- Which individuals are most likely to be affected by Huntington disease, and in which of these people is the onset of the disease likely to be earliest?
  - Which individuals are least likely to be affected by the disease?
  - Consider the two PCR primers used to amplify the trinucleotide repeat region. If the 5' end of one of these primers is located 70 nucleotides upstream of the first CAG repeat, what is the maximum distance downstream of the last CAG repeat at which the 5' end of the other primer could be found? [Assume that the diagram shows the largest *HD*<sup>+</sup> allele possible (that is, 35 CAG repeats).]





17. Sperm samples were taken from two men just beginning to show the effects of Huntington disease. Individual sperm from these samples were analyzed by PCR for the length of the trinucleotide repeat region in the *HD* gene. In the graphs that follow, the horizontal axes represent the number of CAG repeats in each sperm, and the vertical axes represent the fraction of total sperm of a particular size. The first graph shows the results for a man whose mutant *HD* allele (as measured in somatic cells) contained 62 CAG repeats; the man whose sperm were analyzed in the second graph had a mutant *HD* allele with 48 repeats.



- a. What is the approximate CAG repeat number in the  $HD^+$  alleles from both patients?

- b. Assuming that these results indicate a trend, what can you conclude about the processes that give rise to mutant *HD* alleles? In what kinds of cells do these processes take place?
- c. How do these results explain why approximately 5–10% of Huntington disease patients have no family history of this condition?
- d. Predict the results if you performed this same PCR analysis on single skin cells from each of these patients instead of single sperm.

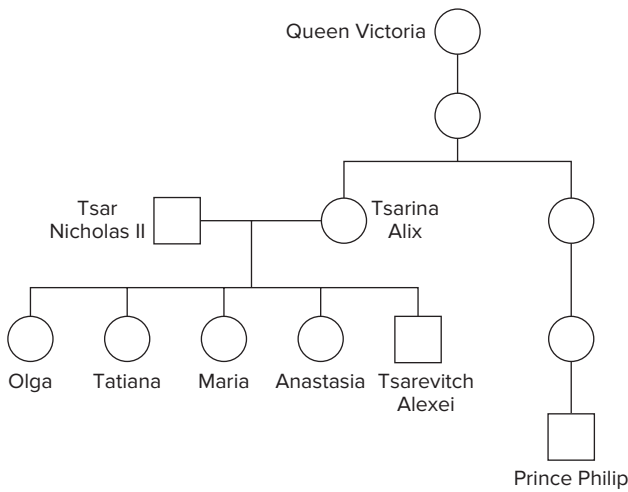
### Section 11.3

18. In 1993, the courts for the first time accepted plant DNA as evidence in a murder trial. The accused defendant owned a pickup truck in which the police found a few seed pods from a Palo Verde, the state tree of Arizona. The murdered woman was found abandoned in the Arizona desert. How could a prosecutor use DNA from the seed pods to build a strong case against the defendant?
19. a. It is possible to perform DNA fingerprinting with SNPs instead of SSRs as DNA markers, but in general you would need to examine more SNP markers than the 13 SSRs used in the CODIS database to be sure of a match. Explain why.
- b. DNA fingerprinting has been used to verify pedigrees of valuable animals such as show dogs, racing greyhounds, and thoroughbred horses. However, the technology is much harder to apply in these cases than it is in forensic applications for humans. In particular, many more DNA markers must be examined in domesticated animals to establish the identity or close familial relationship of two DNA samples. Why would you need to look at more polymorphic loci in these animals than you would in humans?
20. On July 17, 1918, Tsar Nicholas II; his wife the Tsarina Alix; their daughters Olga, Tatiana, Maria, and Anastasia; their son, the Tsarevitch (Crown Prince) Alexei; and four loyal retainers were murdered by Bolshevik revolutionaries. The bodies were not recovered for many years, fueling legends that Grand Duchess Anastasia had escaped, and allowing a woman named Anna Anderson to claim that she was Anastasia. In 1991 and in 2007, two mass graves with a total of nine skeletal remains were unearthed at Ekaterinburg in Russia's Ural Mountains. The table that follows presents partial DNA fingerprint analysis (using only five SSR loci and the sex chromosome marker Amel) of these skeletons. Entries separated by commas indicate alleles (number of repeating units).
- a. What is the most likely identification for each skeleton? (*Note:* You cannot differentiate among any of the daughters based on this information alone.)

Locus	A	B	C	D	E	F	G	H
Amel	XX	XX	XY	XX	XX	XX	XY	XX
D3S1358	17,18	17,18	14,17	16,17	16,18	13,15	14,16	17,18
D5S818	12,12	—	12,12	12,12	12,12	11,12	12,12	12,12
D13S317	11,11	11,11	—	11,11	11,11	10,14	11,12	11,11
D16S539	11,11	11,11	11,14	11,14	—	10,12	11,14	9,11
D8S1179	13,16	15,16	13,15	13,16	16,16	13,14	15,16	15,16

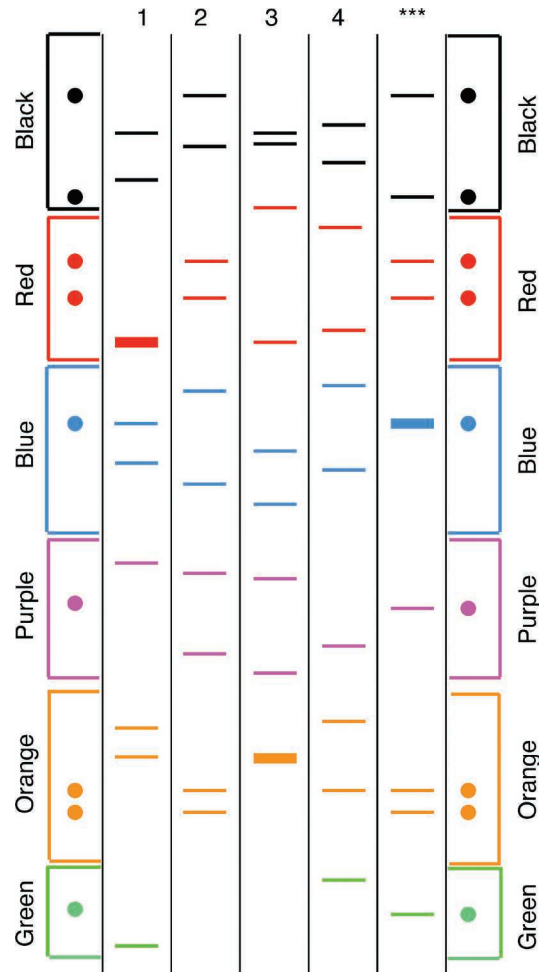
- b. Three PCR reactions failed to yield PCR products. If the reactions had worked properly, what alleles would you expect to see in each case?
- c. Are any of the daughters identical twins?
- d. What kind of evidence could you obtain from the skeletons to differentiate among the daughters?
- e. How do these DNA fingerprints repudiate the claims of Anna Anderson?

The DNA fingerprint data in the table are certainly consistent with the idea that some of these skeletons were members of the Tsar’s family, but they do not prove the hypothesis. To investigate further, forensic scientists obtained blood samples from Prince Philip (the consort of Queen Elizabeth II of Great Britain) and compared his DNA fingerprint to those obtained from the skeletal remains in Russia. A family genealogy is provided here. The results validated that the Tsar’s family was indeed interred in these graves.



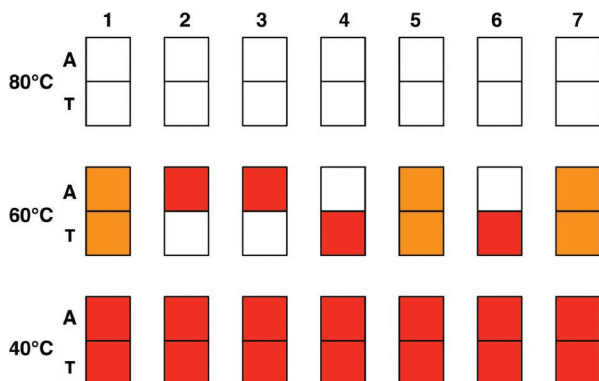
- f. For autosomal DNA markers, what percentage of alleles in the Tsarina’s skeleton should match with alleles in Prince Philip’s genome?
- g. For autosomal DNA markers, what percentage of alleles in the Tsarevitch’s skeleton should match with alleles in Prince Philip’s genome?
- h. A question for genealogy aficionados: What is Prince Philip’s relationship with the Tsarina?

- 21. The figure that follows shows DNA fingerprint analysis of the genomic DNA from semen associated with a rape (\*\*\*) and from mouth swabs (somatic cells) of individuals 1–4. This analysis involves the PCR amplification of six SSR loci, each from a different (nonhomologous) chromosome. All PCR primers used are 20 nucleotides long; the primers for each locus have fluorescent tags in a locus-specific color. In the gel, some bands are thicker because relatively more of the corresponding PCR product was obtained. The figure has dots aligned on both sides



that you can use to find the critical bands, using the edge of a piece of paper as a guide.

- Sperm are haploid, but the semen sample shows two different-sized PCR products for certain loci. How is this possible?
  - Is any locus on the X chromosome? If so, identify it.
  - Is any locus on the Y chromosome? If so, which one?
  - Explain why these results demonstrate that none of the four individuals is the rapist. What pattern would you expect by analyzing mouth swab DNA from the rapist?
  - Do these results nonetheless provide any information that could help catch the rapist? If so, be as specific as possible.
  - The two orange bands amplified by PCR from the semen are 200 and 212 bp long. How many tandem repeats of the SSR repeat unit are found in the two alleles of this locus in the rapist's genomic DNA? (Assume that the PCR products are the shortest possible and that the repeat unit for this locus is TCCG.)
22. Microarrays were used to determine the genotypes of seven embryos (made by *in vitro* fertilization) with regard to sickle-cell anemia. Each pair of squares in the accompanying figure represents two ASOs, one specific for the  $Hb\beta^A$  allele (A) and the other for the  $HB\beta^S$  allele (T), attached to a chip of silicon and hybridized with fluorescently labeled PCR product from a single cell from one of the embryos. The hybridizations were performed at three different temperatures (80°C, 60°C, and 40°C) as shown.



- Why do you think the PCR step is needed for this microarray analysis?
- Make a sketch of the location in genomic DNA of the PCR primers relative to the sickle-cell mutation. Indicate the 5'-to-3' polarities of all DNA molecules involved.

- Why is no hybridization seen at 80°C?
- Why do you see strong hybridization of all genomic DNA probes to both ASOs at 40°C?
- What are the genotypes of the seven embryos? Which of these embryos would you choose to implant into the mother's uterus to avoid the possibility that the child would have sickle-cell anemia?

23. A partial sequence of the wild-type  $Hb\beta^A$  allele is shown here (the top strand is the RNA-like coding strand, and the location of the disease-causing mutation is underlined):

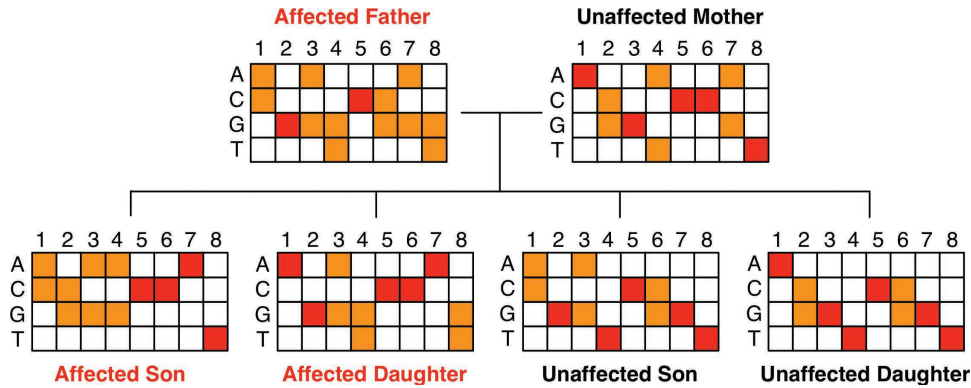
```
5' ATGGTGCACCTGACTCCTGAGGAGAAGTCGCCG 3'
3' TACCACGTGGACTGAGGACTCCTCTTCAGCGGC 5'
```

and the sickle-cell allele  $HB\beta^S$  sequence is:

```
5' ATGGTGCACCTGACTCCTGTGGAGAAGTCGCCG 3'
3' TACCACGTGGACTGAGGACACCTCTTCAGCGGC 5'
```

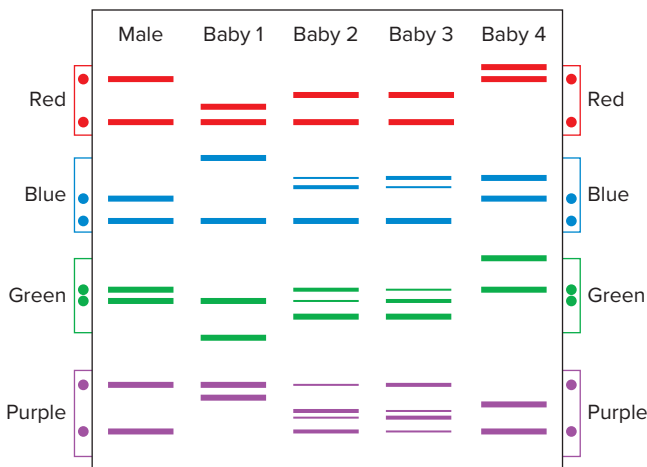
Design two 21-nucleotide-long ASOs that could be attached to a silicon chip for the microarray analysis performed in Problem 22. Two possibilities exist for each ASO; you only need to show one possibility for each.

- In Fig. 11.17b, PCR is performed to amplify genomic DNA for genotyping on microarrays. This PCR reaction requires only a single primer, but normally, PCR requires two primers. Why does only a single primer suffice in this case?
  - Again in Fig. 11.17b, genomic DNA was cut with a restriction enzyme before its PCR amplification. What kind of restriction enzyme would be most effective for this purpose: Would it create sticky ends or blunt ends? Would it recognize a site made of 4 bp, 6 bp, or 8 bp?
25. The following figure shows a partial microarray analysis for members of a nuclear family. The eight SNP loci examined are evenly spaced at about 10 Mb intervals on chromosome 4, and they are shown on the microarray in their actual order on this chromosome. For the time being, focus your attention only on the two parents and ignore whether they are affected or unaffected.
- Write out the complete genotype for all the DNA markers in both parents.
  - The microarray data indicate that one SNP locus has three alleles in this family. Which one?
  - How would you know that these loci are in fact on chromosome 4 and are about 10 Mb apart?
  - About what percentage of the total length of chromosome 4 is present in the region between DNA markers 1 and 8? (Chromosome 4 is 191 Mb long; it is the fourth largest in the human genome.)



26. Scientists were surprised to discover recently that marmosets, a kind of monkey, are often *chimeric*—they have cells that originated from two different zygotes. Marmoset chimeras occur in fraternal twins. While in their mother's womb, these twins can share cells through the blood supply because their placentas are fused. Any organ or tissue, including the germ line, can be chimeric.

To determine if a particular male fathered any of four baby marmosets from three different mothers, DNA fingerprinting of four SSR loci was performed using genomic DNA from hairs. In the accompanying diagram, the thickness of the bands correlates with the amount of DNA present. You can use the dots on both sides of the diagram, along with a ruler or paper edge, to help compare the positions of the bands.



- What is the maximum number of different alleles of any one SSR locus that the hair cells of any one marmoset could have?
- Which of the four babies could have been fathered by the male whose DNA fingerprint is shown?
- Identify the pair of chimeric fraternal twins.

- Does any evidence exist for chimerism in the father of the twins? in the mother?
- Explain how the biological mother of a baby marmoset can be its genetic uncle. (*Note:* This is not necessarily the situation for any of the baby marmosets in the figure.)

### Section 11.4

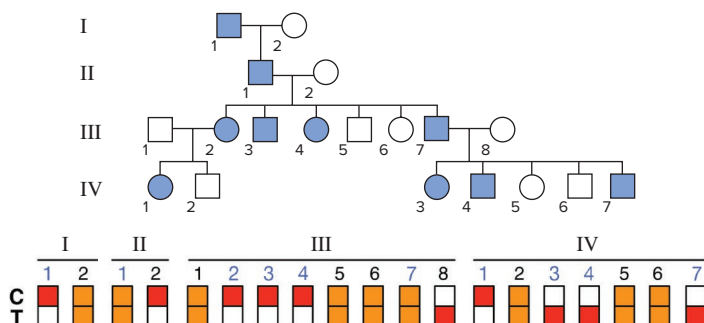
27. The microarray shown in Problem 25 analyzes genomic DNA from a nuclear family in which the father, one son, and one daughter have rare, late onset polycystic kidney disease; while the mother, the second son, and the second daughter are unaffected. As stated in Problem 25, the eight SNP loci examined are evenly spaced at about 10 Mb intervals on chromosome 4, and they are shown on the microarray in their actual order on this chromosome.

- Is the allele responsible for the disease dominant or recessive with respect to wild type? Is the disease gene autosomal or X-linked?
- For each of the four siblings, indicate the genotype of the sperm from which they were created. For each of these sperm, write the alleles for each of the eight loci on chromosome 4 in order.
- Identify the two SNP loci that are uninformative in this family (that is, you cannot determine whether or not either of these loci is linked to the disease gene).
- Assuming for the sake of simplicity that the four children shown would be completely representative even if the parents had 100 children, the data in the figure indicate that one locus is unlinked to the disease gene. Which one?
- The microarray results indicate that during meiosis in the father, two different crossovers occurred in the region including the disease gene and the SNP loci that are genetically linked to it. Draw a map of chromosome 4 showing the locations of

the disease gene, the linked SNP loci, and the two crossovers. Your map should indicate any uncertainties in these positions.

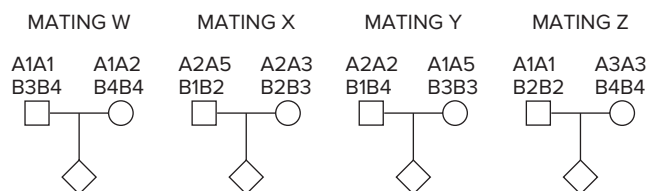
f. Diagram the location and arrangement (phase) of all alleles of all genes that are on the two chromosomes in the father's diploid genome.

28. The figure that follows shows the pedigree of a family in which a completely penetrant, autosomal dominant disease is transmitted through three generations, together with microarray analysis of each individual for a biallelic SNP locus (the alleles are C and T).



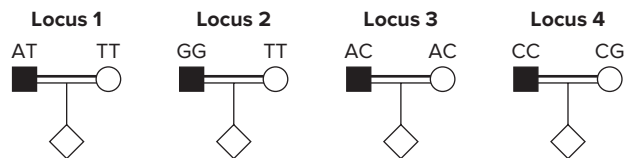
- Do the data suggest the existence of genetic linkage between the SNP locus and the disease locus? If so, what is the estimated genetic distance between the two loci?
- Calculate the maximum Lod score for linkage between the SNP and the disease locus for this pedigree. What does this value of the Lod score signify?

29. One of the difficulties faced by human geneticists is that matings are not performed with a scientific goal in mind, so pedigrees may not always provide desired information. As an example, consider the following matings (W, X, Y, and Z):



- Which of these matings are informative and which noninformative for testing the linkage between anonymous loci A and B? (A1 and A2 are different alleles of locus A, B1 and B2 are different alleles of locus B, etc.) Explain your answer for each mating.
- Is locus A more likely to be a SNP or an SSR? What about locus B? Explain.

30. Now consider a mating between consanguineous people involving a recessive genetic disease. The figures that follow show the genotypes of these two people at four SNP loci (1–4).



- For which of these loci is it possible to obtain information about the linkage of the SNP to the disease gene? Explain your answer for each locus and describe any special conditions that may apply.
- For any of the loci for which the mating is potentially informative, how would you tell whether the child is the product of a recombinant or nonrecombinant gamete? (That is, how could you solve the phase problem?) Be as specific as possible.

31. The pedigree shown in Fig. 11.22 was crucial to the identification of the Huntington disease gene *HD*, which is located on chromosome 4.

- The data show that the DNA marker G8 is clearly linked to *HD*. For the large majority of the people in the pedigree with Huntington disease, which allele of G8 (A, B, C, or D) did they inherit, along with the dominant disease-causing allele of *HD*, on the copy of chromosome 4 from their affected parent?
- How many people in the pedigree can you categorize absolutely as the product of parental or recombinant gametes from their affected parent, without making any assumptions at all (including the assumption of linkage)?
- If you now make the assumption that G8 and *HD* are linked, how many of the people in this pedigree must be the product of a recombinant gamete from their affected parent?

- Based solely on the data from this pedigree, what would be the best estimate for the map distance between G8 and *HD*?
- Considering your answers to parts (b) through (d), calculate the maximum Lod score. The pedigree contains 47 people resulting from informative matings. (Note that  $0^0 = 1$ .) What does this Lod score signify?

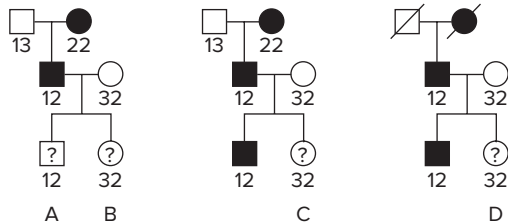
32. You have identified a SNP marker that in one large family shows no recombination with the locus causing a rare hereditary autosomal dominant disease. Furthermore, you discover that all afflicted individuals in the family have a G base at this SNP on their

mutant chromosomes, while all wild-type chromosomes have a T base at this SNP. You would like to think that you have discovered the disease locus and the causative mutation but realize you need to consider other possibilities.

- What is another possible interpretation of the results?
- How would you go about obtaining additional genetic information that could support or eliminate your hypothesis that the base-pair difference is responsible for the disease?

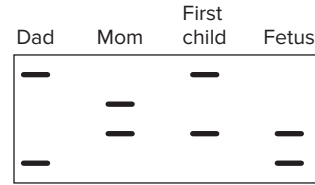
Problems 33 and 34 show that you can make predictions about a child's genotype by genotyping linked markers even if you don't directly examine the disease-causing mutation. This method can be valuable for diseases showing high allelic heterogeneity if the linkage is extremely tight.

33. The pedigrees indicated here were obtained with three unrelated families whose members express the same disease caused by a completely penetrant dominant allele. The disease allele is linked at a distance of 10 cM from an SSR marker locus with three alleles numbered 1, 2, and 3. The SSR alleles present within each living person's genotype are indicated below the pedigree symbol. The phenotypes of the newly born labeled individuals—A, B, C, and D—are unknown.



- What is the probability of disease expression in each of these newborn babies?
  - Why would a human geneticist be unlikely to use this SSR marker for diagnosis of the genetic disease?
- 34 Approximately 3% of the population carries a mutant allele at the *CFTR* gene responsible for the recessive disease cystic fibrosis. A genetic counselor is examining a family in which both parents are known to be carriers for a *CFTR* mutation. Their first child was born with the disease, and the parents have come to the counselor to assess whether the new fetus inside the mother's womb is also diseased, is a carrier, or is homozygous wild type at the *CF* locus. DNA samples from each family member and the fetus are tested by PCR and gel electrophoresis for an SSR marker

within one of the *CFTR* gene's introns. The following results are obtained:



- What is the probability that the child who will develop from this fetus will exhibit the disease?
  - When this child grows up, what is the probability that any one of her own children will be afflicted with the disease?
  - The cystic fibrosis gene displays extensive allelic heterogeneity: More than 1500 different mutations of the *CFTR* gene have been shown to be associated with cystic fibrosis worldwide. With this fact in mind, why might human geneticists choose to test the fetus in the indirect manner described in this problem rather than focusing directly on the mutations that actually caused the disease in the first child?
35. The drug *ivacaftor* has recently been developed to treat cystic fibrosis in children with the rare G551D mutant allele of *CFTR*.
- Do you think that *ivacaftor* would be effective only in patients homozygous for the G551D mutation, or might it work as well in compound heterozygotes in which one copy of chromosome 7 had G551D and the other copy a different allele of *CFTR*, such as the more prevalent allele  $\Delta F508$ ? (The protein encoded by G551D folds up properly and inserts into the cell membrane, but is inefficient in chloride ion transport. *Ivacaftor* increases the efficiency of G551D's ion transport. The  $\Delta F508$  protein does not fold up properly and therefore does not get inserted into the cell membrane.)
  - Why do you think *ivacaftor* would be more effective in children than in older cystic fibrosis patients?
  - The scientists who developed *ivacaftor* had a model for cystic fibrosis: a line of cells that grow in culture and that are homozygous for G551D. These cells accumulate mucus at their surfaces that prevent cilia (tiny hairs on the outside of cells) from beating. Explain how the scientists could use this disease model to screen for drugs that would be effective against G551D-associated cystic fibrosis.
  - Ivacaftor* is used in combination with an even newer drug called *lumacaftor* to treat individuals homozygous for the most common CF allele,  $\Delta F508$ . *Lumacaftor* helps the  $\Delta F508$  mutant

CFTR protein fold properly so that it can insert in the cell membrane. Why do you think that neither lumacaftor nor ivacaftor alone are effective in treating  $\Delta F508$  homozygotes, while the combination of both drugs is effective?

### Section 11.5

36. In the high-throughput DNA sequencing protocol shown in Fig. 11.24:
- What is the purpose of adding poly-A to fragments of single-stranded genomic DNA, and why is the poly-A added to the 3' end of these fragments?
  - Why at the end of every synthesis cycle do you need to remove the fluorescent tag on the incorporated nucleotide?
  - Why do the incorporated nucleotides have a blocking group, and why does this blocking group need to be removed each cycle?
37. A researcher sequences the whole exome of a patient suffering from Usher syndrome, a rare autosomal recessive condition that is nonetheless the leading cause for simultaneous deafness and blindness. The exome sequence does not show homozygosity for any polymorphisms different from the human RefSeq.
- How could the researcher examine the data already gathered to try to find the disease gene, assuming the sequence is accurate?
  - If the attempt described in part (a) was unsuccessful, the researcher might contemplate sequencing the patient's whole genome. What are the potential pitfalls of this strategy?
38. As explained in the text, the cause of many genetic diseases cannot yet be discerned by analyzing whole-exome/genome sequences. But in some of these seemingly intractable cases, important clues can be obtained by looking at mRNAs or proteins, rather than at the DNA.
- As you will see in more detail in later chapters, it is possible to use single-molecule methods to sequence cDNA copies of millions of mRNA molecules from any particular tissue cheaply. How could you sometimes use such information to find a disease gene? When would this information be noninformative?
  - A technique called *Western blotting* allows you to examine any protein for which you have an antibody; it is possible to see differences in size or amount of that protein. How could you sometimes use such information to find a disease gene? When would this information be noninformative?
39. Figure 11.26 portrayed the analysis of Miller syndrome through the sequencing of four complete genomes:
- those of a brother and sister both affected by the disease, and of both their parents.
- Researchers made the assumption that Miller syndrome is a recessive trait. Could Miller syndrome instead be due to a dominant mutation? If so, what scenarios would make this possible?
  - Why is it highly unlikely that Miller syndrome in this family is due to *de novo* mutations that occurred in the germ line of the mother, or of the father, or of both parents? Describe a scenario based on your understanding of cell divisions in human ovaries or testes (see Figs. 4.18 and 4.19) that make the *de novo* mutation hypothesis at least theoretically possible even if very unlikely.
  - On Fig. 11.26b, indicate the location on chromosome 16 closest to the *DHOD* gene at which recombination took place during meiosis in one of the parents of the Miller syndrome patients. In which parent did this recombination occur?
  - Do the number of crossovers you see in Fig. 11.26b fit previous estimates that in the human genome, 1 centiMorgan corresponds to about 1 Mb? Chromosome 16 is about 90 Mb long; chromosome 17 is about 81 Mb long.
  - How could researchers use all the sequence data from this family to estimate the per-nucleotide rate of mutation in humans?
40. A research paper published in the summer of 2012 presented a method to obtain the whole-genome sequence of a fetus without any invasive procedure such as amniocentesis that could on rare occasions cause miscarriage. This new technique is based on the fact that some fetal cells leak into the mother's bloodstream and then break down, releasing their DNA. Assume that exactly 10% of the DNA fragments in the mother's blood serum come from the fetus, while the remaining 90% of the DNA fragments in the serum come from the mother's genome.
- The investigators collected cell-free DNA from a pregnant woman's bloodstream and subjected it to an advanced high-throughput sequencing method. The table at the end of this problem looks at seven unlinked loci; the number of *reads* of particular alleles (identified by Greek letters) are shown. You should assume for the sake of simplicity that all numerical differences are statistically significant (even though actual data are never this clean).
- Determine whether each locus is autosomal, X-linked, or Y-linked.
  - Describe the diploid genomes of the mother and fetus by using Greek letters for the alleles, or a dash (–) if no Greek letter is appropriate.
  - Is the fetus male or female?

- d. At an eighth locus, 1500 reads of a single type of sequence were found. Provide a possible explanation for this result, being as specific as possible.

Locus	Sequences	Number of reads
1	$\alpha$ : TCTTTGGTAAACGCAAG	1000
2	$\alpha$ : GTACCGGAGGCAGCCTC	500
	$\beta$ : GTACCGCGGCAGCCTC	500
3	$\alpha$ : AGCCATTGCGGATCCGA	950
	$\beta$ : AGCTATTGCGGATCCGA	50
4	$\alpha$ : GGGGCCTTATGATAAGG	50
5	$\alpha$ : CAGTTCCTGGAGTTGTA	550
	$\beta$ : CAGTTCATGGAGTTGTA	450
6	$\alpha$ : GCAGCCCGTGCTGTTAA	500
	$\beta$ : GCAGCCCGTGCTGTCAA	450
7	$\alpha$ : CACTCAGTCCTACGGAC	500
	$\beta$ : CACTCGTCCTACGGAC	450
	$\gamma$ : CACTCAGTCCTAAGGAC	50

41. Table 11.2 and Fig. 11.27 together portray the search for the mutation causing Nic Volker's severe inflammatory bowel disease. Neither of Nic's parents had the condition, so geneticists narrowed their investigation by focusing on rare variants that showed a recessive pattern and those on the X chromosome.
- For candidate variants on an autosome, would the researchers have looked only for variants for which Nic is homozygous? Explain.
  - Apart from the recessive and X-linked hypotheses, do any other possible explanations exist for Nic's condition?
  - The causative mutation was pinpointed by analyzing only Nic's exome, because at the time of these investigations, whole-genome or whole-exome sequencing was too expensive to perform on his parents. How could you determine inexpensively whether or not this mutation occurred *de novo* in the germ line of one of his parents (that is, during the formation of the particular egg or sperm that produced Nic)? Your answer should not involve whole-genome or whole-exome sequencing.
42. The human RefSeq of the entire first exon of a gene involved in Brugada syndrome (a cardiac disorder characterized by an abnormal electrocardiogram and an increased risk of sudden heart failure) is:

5' CAACGCTTAGGATGTGCGGAGCCT 3'

The genomic DNA of four people (1–4), three of whom have the disorder, was subjected to single-molecule sequencing. The following sequences represent all those obtained from each person. Nucleotides different from the RefSeq are underlined.

Individual 1:  
5' CAACGCTTAGGATGTGCGGAGCCT 3'  
and  
5' CAACGCTTAGGATGTGCGGAGCCT 3'

Individual 2:  
5' CAACGCTTAGGATGTGAGGAGCCT 3'

Individual 3:  
5' CAACGCTTAGGATGTGCGGAGCCT 3'  
and  
5' CAACGCTTAGGATGGCGGAGCCT 3'

Individual 4:  
5' CAACGCTTAGGATGTGCGGAGCCT 3'  
and  
5' CAACGCTTAGGATGTGIGGAGCCT 3'

- The first exon of the RefSeq copy of this gene includes the start codon. Write as much of the amino acid sequence of the encoded protein as possible, indicating the N-to-C polarity.
  - Are any of these individuals homozygotes? If so, which person and what allele?
  - Is the inheritance of Brugada syndrome among these individuals dominant or recessive?
  - Is Brugada syndrome associated with allelic heterogeneity?
  - Are any of these individuals compound heterozygotes?
  - Do the data show any evidence for locus heterogeneity?
  - Which person has normal heart function?
  - For each variant from the RefSeq, describe: (i) what the mutation does to the coding sequence; and (ii) whether the variation is a loss-of-function allele, a gain-of-function allele, or a wild-type allele.
  - For each variant, indicate which of the following terms apply: null, hypomorphic, hypermorphic, nonsense, frameshift, missense, silent, SNP, DIP, SSR, anonymous.
  - Is the function of this gene haploinsufficient? Explain.
43. Mutations in the *HPRT1* gene in humans result in at least two clinical syndromes. Consult OMIM ([www.omim.org](http://www.omim.org)) by querying HPRT1; you will only need to look briefly at the top three hits (files #300322, 300323, and 308000).
- What is the full name of the HPRT1 enzyme?
  - On which chromosome is the *HPRT1* gene located?



- c. Mutations in *HPRT1* are associated with two different syndromes. What are these syndromes? For each, answer the following questions: (i) What are the symptoms associated with the syndrome? (ii) Is the mutant allele that causes the syndrome dominant, recessive, codominant, or incompletely dominant with respect to the normal allele, or do special conditions apply? (iii) Is the syndrome associated with a loss-of-function or a gain-of-function disease allele? (iv) Does the syndrome display allelic heterogeneity? (v) Does the syndrome display locus heterogeneity? (*Note:* You do not need to understand everything in the OMIM entries to answer these questions.)
44. We think of people as having two genomes (one from their mother and one from their father) that are identical in every somatic cell and in germ cell precursors. However, with every DNA replication and thus every cell division, a genome has an opportunity to mutate. When a mutation occurs in a single cell, all of that cell's mitotic descendants will thus have a mutation that the other cells in the body do not have. Humans are in this sense *mosaic*—we are a patchwork of cells with somewhat different genotypes. Explain how human mosaicism could complicate the process of pinpointing a recessive or dominant disease-causing mutation by positional cloning or by genome sequencing.

chapter **12**

## The Eukaryotic Chromosome



*CC, the cloned cat (left), and her nuclear donor Rainbow (right) share the same chromosomes yet have distinctive appearances.*

(left): © Texas A&M University/AP Photo  
(right): © Alpha/ZUMAPRESS/Newscom

**chapter outline**

- 12.1 Chromosomal DNA and Proteins
- 12.2 Chromosome Structure and Compaction
- 12.3 Chromosomal Packaging and Gene Expression
- 12.4 Replication of Eukaryotic Chromosomes
- 12.5 Chromosome Segregation
- 12.6 Artificial Chromosomes

**IN DECEMBER 2001**, veterinarians delivered by caesarean section the world's first cloned pet, the kitten CC (a play on the words *Copy Cat* and *Carbon Copy*; accompanying figure). CC was the result of a nuclear transfer experiment in which the nucleus from a female calico cat named Rainbow (*right*) was injected into an egg (from an unnamed donor cat) whose own nucleus had been removed previously. The reconstituted egg was then implanted into the uterus of a surrogate mother. CC turned out to be a perfectly normal cat who has survived well past the age of 12 and has mothered three of her own normal kittens.

All of the cells you see in both CC and Rainbow have the same nuclear DNA because they are the mitotic descendants of a single cell: the zygote that became Rainbow. Yet even though the nuclear DNAs of all these cells are identical, all of the cells clearly are not. Some cells have differentiated into eyes, others into whiskers, yet others to skin, and so on. And despite the same DNA, CC and Rainbow are dissimilar in many phenotypes, ranging from the color patterns of their fur to the disposition of their characters.

The packaging of DNA into chromosomes underlies much of the biology you see on display in these photographs. Such packaging allows the billions of base pairs making up the cat genome to fit into a cell's nucleus, and the packaging is a prerequisite for the faithful copying and distribution of the genome through innumerable mitotic cell divisions. Some of the proteins on the chromosomes tell genes when to turn on and off, and are thus responsible for cell differentiation. Other macromolecules (proteins and RNAs) associated with chromosomes are responsible for X chromosome inactivation, which accounts for many differences seen in the coat color patterns of Rainbow and CC.

In this chapter, we examine the structure and function of the eukaryotic chromosome. One general theme emerges from our discussion. Chromosomes have a versatile, dynamic structure that supports their many functions in the packaging, replication, segregation, and expression of the information in a single long molecule of DNA.

## 12.1 Chromosomal DNA and Proteins

### Learning Objectives

1. Diagram the DNA components of a chromosome, including the polarity of strands.
2. Contrast histone and nonhistone proteins in terms of structure and function.

When viewed under the light microscope, chromosomes change shape, character, and position as they pass through the cell cycle. During interphase, they look like tangled masses of spaghetti. By metaphase of mitosis, they appear as paired bars (the two sister chromatids) in the middle of the spindle apparatus. In this section, we describe what chromosomes are made of. Then, in succeeding sections of this chapter, we explain how these chromosomal components interact to produce the observed metamorphoses of structure.

### Each Chromosome Is Composed of a Single Long Molecule of DNA

Researchers learned from physical analyses that each chromosome (or after replication, each chromatid) within a cell nucleus contains one long linear molecule of double-stranded DNA. In one early study, they placed chromosomal DNA between two cylinders, stretched the DNA by rotating one of the cylinders, and measured the DNA's rate of recoil. Shorter molecules recoil faster than longer ones. When the investigators applied this technique to the DNA in a *Drosophila* cell, the length of the longest DNA molecule measured corresponded to the amount of DNA in the largest chromosome. This chromosome must therefore contain a single linear molecule of DNA.

You have examined previously (in Chapter 11) general aspects of the way genes are organized along the DNA molecule making up a typical eukaryotic chromosome. Chromosomes also display another organizational feature of importance: Substantial stretches of noncoding repetitive DNA such as **simple sequence repeats (SSRs)** and **transposable elements** are concentrated in specific chromosomal regions, particularly at **centromeres** and **telomeres**. The repetitive sequences at these locations are crucial for several aspects of chromosome biology to be discussed in later sections of this chapter.

### Chromosomes Contain Histone Proteins and Nonhistone Proteins

By itself, DNA does not have the ability to fold up small enough to fit in the cell nucleus. For sufficient compaction,

it depends on interactions with two categories of proteins: *histones* and *nonhistone chromosomal proteins*. **Chromatin** is the generic term for any complex of DNA and protein found in a cell's nucleus. A chromosome is a piece of chromatin that contains (prior to S phase) a single DNA molecule and behaves as a unit during cell division.

Although chromatin is roughly one-third DNA, one-third histones, and one-third nonhistone proteins by weight, it also contains significant amounts of RNA. The roles of chromatin-associated RNAs are not well understood in general, but later in this chapter we will describe the function of a specific RNA molecule called *Xist* in controlling gene expression.

### Histone proteins

Discovered in 1884, **histones** are relatively small proteins with a preponderance of the basic, positively charged amino acids lysine and arginine. The histones' strong positive charge enables them to bind to and neutralize the negatively charged DNA throughout the chromatin. Histones make up half of all chromatin protein by weight and are classified into five types of molecules: H1, H2A, H2B, H3, and H4. The last four types—H2A, H2B, H3, and H4—form the core of the most rudimentary DNA packaging unit—the **nucleosome**—and are therefore referred to as **core histones**. (We will examine the role of these histones in nucleosome structure momentarily.)

All five types of histones appear throughout the chromatin of nearly all diploid eukaryotic cells, and they are very similar in all eukaryotes. In the H4 proteins of pea plants and calves, for example, all but two of 102 amino acids in the polypeptide sequence are identical. That histones have changed so little throughout evolution underscores the importance of their contribution to chromatin structure.

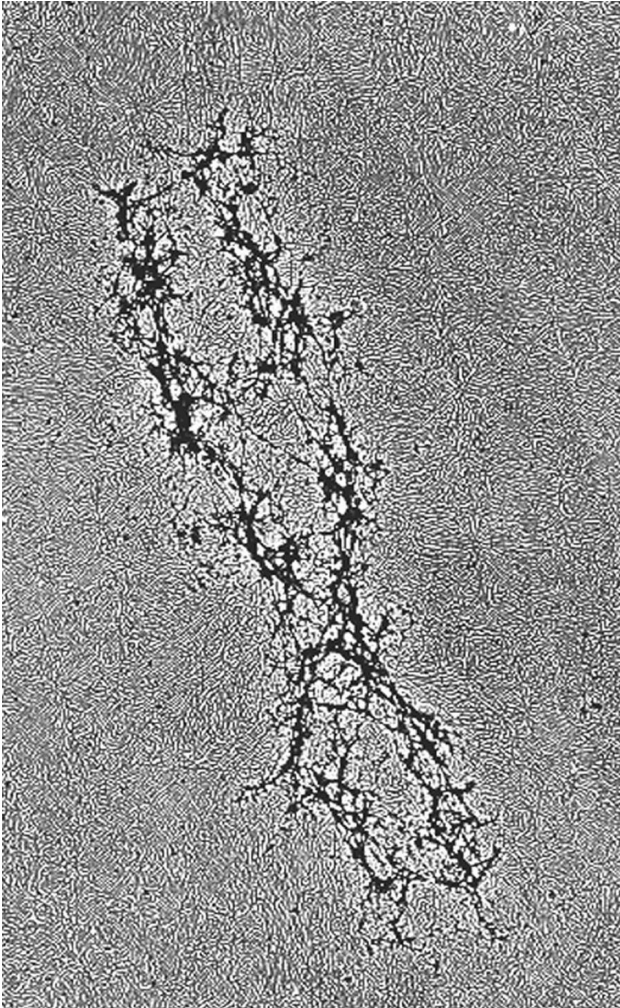
### Nonhistone proteins

The remaining half of the mass of protein in eukaryotic cell chromatin is not composed of histones. Rather, it consists of thousands of different kinds of **nonhistone chromosomal proteins**. The chromatin of a diploid genome contains from 200 to 2,000,000 molecules of each kind of nonhistone protein.

Not surprisingly, this large variety of proteins fulfills many different functions. Some nonhistone proteins play a purely structural role, helping to package DNA into more complex structures. The proteins that form the structural backbone, or *scaffold*, of the chromosome fall in this category (**Fig. 12.1**). Other nonhistone proteins, such as DNA polymerase, are active in replication. Still others are crucial for chromosome segregation: For example, the motor proteins of kinetochores help move chromosomes along

**Figure 12.1 The chromosome scaffold.** When this human chromosome was gently treated with detergents to remove the histones and some nonhistone proteins, a dark *scaffold* composed of the remaining nonhistone proteins became visible in the shape of the two sister chromatids. Loops of DNA freed by the detergent treatment surround the scaffold.

© Dr. Don Fawcett/J.R. Paulson, U.K. Laemmli/Science Source

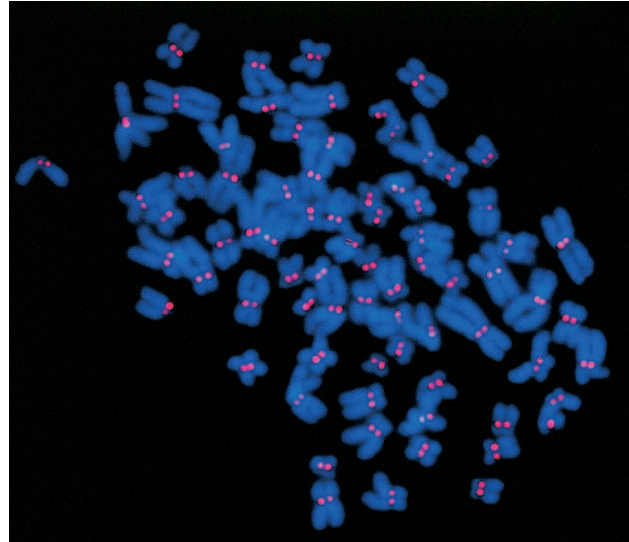


the spindle apparatus and thus expedite the transport of chromosomes from parent to daughter cells during mitosis and meiosis (**Fig. 12.2**).

By far the largest class of nonhistone proteins constitute those which foster or regulate transcription during gene expression. Mammals carry more than 5000 different proteins of this kind. By interacting with DNA, these proteins influence when, where, and how frequently genes are transcribed.

**Figure 12.2 Kinetochore proteins.** Some nonhistone proteins are required for chromosome movements along the spindle during cell division. Here, chromosomes are stained in *blue* and the nonhistone protein CENP-E in *red*. CENP-E molecules at the kinetochores help move sister chromatids toward the spindle poles during anaphase.

© Daniel A. Starr/University of Colorado



### essential concepts

- Each eukaryotic chromosome contains (prior to its replication) a single molecule of linear, double-stranded DNA, without breaks or changes of polarity.
- The *histone proteins* help package DNA into *chromatin*; the *nonhistone proteins* function in chromosome structure, replication, and segregation, as well as in the control of gene expression.

## 12.2 Chromosome Structure and Compaction

### learning objectives

1. Diagram the structure of a nucleosome.
2. Describe nucleosome supercoiling and its relationship to the radial loop–scaffold model of chromatin packaging.
3. Summarize the process of detecting G bands in a chromosome and how these bands are used in locating genes.
4. Describe FISH analysis and its application in finding specific DNA sequences in a chromosome.

**TABLE 12.1** Levels of Chromosome Compaction

Mechanism	Status	What It Accomplishes
Nucleosome	Confirmed by crystal structure	Condenses naked DNA 7-fold to a 100 Å fiber
Supercoiling	Hypothetical model (although the 300 Å fiber predicted by the model has been seen in the electron microscope)	Causes additional 6-fold compaction, achieving a 40- to 50-fold condensation relative to naked DNA
Radial loop–scaffold	Hypothetical model (preliminary experimental support exists for this model)	Through progressive compaction of 300 Å fiber, condenses DNA to rodlike mitotic chromosomes that are 10,000 times more compact than naked DNA

Stretched out in a thin, straight thread, the DNA of a single human cell would be 6 feet (about 2 m) in length. This is, of course, much longer than the cell nucleus in which the genome must be contained; the diameter of the average human cell nucleus is only about 6  $\mu\text{m}$  ( $6 \times 10^{-6}$  m). Several levels of compaction enable the DNA to fit inside the nucleus (**Table 12.1**). First, the winding of DNA around histones forms small *nucleosomes*. Next, tight coiling gathers nucleosomes together into higher-order structures. Additional levels of compaction, which researchers do not yet understand, produce the metaphase chromosomes observable in the microscope.

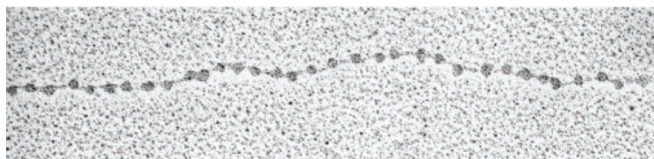
### The Nucleosome Is the Fundamental Unit of Chromosome Packaging

The electron micrograph of chromatin in **Fig. 12.3** shows long, nub-studded fibers bursting from the nucleus of a human blood cell. The nucleosomes resemble beads on a string, with the beads having a diameter of about 100 Å and the string a diameter of about 20 Å ( $1 \text{ Å} = 10^{-10} \text{ m} = 0.1 \text{ nm}$ ). The 20 Å string is DNA. **Figure 12.4a** illustrates how DNA wraps around histone cores to form the chromatin fiber's observed beads-on-a-string structure.

Each bead is a nucleosome containing roughly 160 bp of DNA wrapped around a core composed of eight histones—two each of H2A, H2B, H3, and H4, arranged as shown in **Fig. 12.4a**. The 160 bp of DNA wrap twice around this core histone octamer. An additional 40 or so base pairs form **linker DNA**, which connects one nucleosome with the next.

**Figure 12.3 Nucleosomes.** In the electron microscope, nucleosomes look like beads along a string of DNA.

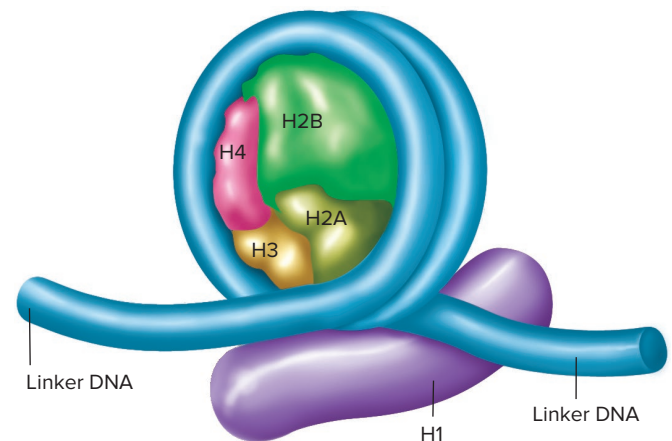
© Dr. Barbara A. Hamkalo



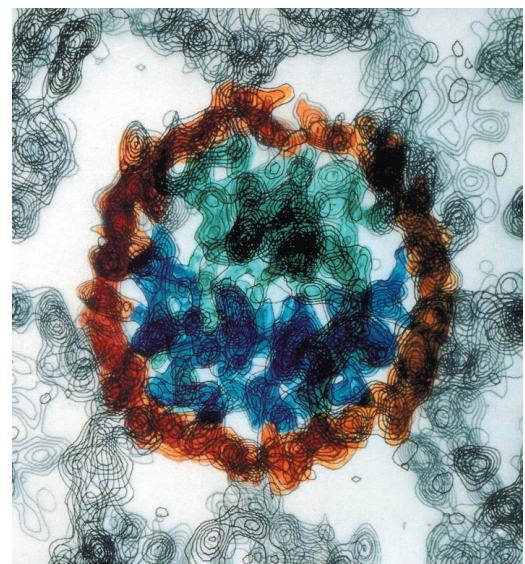
**Figure 12.4 Nucleosome structure.** (a) The DNA in each nucleosome wraps twice around a core made of two copies each of the histones H2A, H2B, H3, and H4. Histone H1 associates with the DNA as it enters and leaves the nucleosome. (b) Nucleosome structure as revealed by X-ray crystallography. The DNA (orange) bends sharply at several places as it wraps around the core histone octamer (blue and turquoise).

b: © Dr. Gerard J. Bunick, Oak Ridge National Laboratory

(a) Schematic diagram of a nucleosome



(b) Nucleosome structure at high resolution



Histone H1 lies outside the core, associating with DNA entering and leaving the nucleosome (Fig. 12.4a). When investigators use specific chemical reagents to remove H1 from the chromatin, some DNA unwinds from each nucleosome, but the nucleosomes do not fall apart; about 140 bp remain wrapped around each core.

Scientists can crystallize the nucleosome cores and subject the crystals to X-ray diffraction analysis. The pictures have led to the model of nucleosome structure just described, and they also indicate that the DNA does not coil smoothly around the histone core (Fig. 12.4b). Instead, the DNA bends sharply at some positions and barely at all at others. Because the sharp bending may occur only with some DNA sequences and not others, base sequence helps dictate preferred nucleosome positions along the chromosome.

Duplication of the basic nucleosomal structure occurs in conjunction with DNA replication. Synthesis of the four basic histone proteins increases during S phase of the cell cycle so as to incorporate histones onto the newly replicated DNA. Additional proteins mediate the assembly of nucleosomes. Special regulatory mechanisms tightly coordinate DNA and histone synthesis so that both occur at the appropriate time.

The spacing and structure of nucleosomes correlate with genetic function. The nucleosomes of each chromosome are not evenly spaced, but they do have a particular arrangement along the chromatin. This arrangement varies among different cell types, and it can change even in a single cell when conditions are altered. The spacing of nucleosomes along the chromosome matters because DNA in the regions between nucleosomes is more readily available than the DNA within nucleosomes for interactions with the proteins that initiate expression, replication, and further compaction.

Packaging into nucleosomes condenses naked DNA about sevenfold. With this condensation, the 2 m of DNA in a diploid human genome shortens to approximately 0.25 m (a little less than a foot) in length. This is still much too long to fit in the nucleus of even the largest cell, and additional compaction is required.

## Higher-Order Packaging Condenses Chromosomes Further

Many of the details of chromosomal condensation beyond the nucleosome remain unknown, but researchers have proposed several models to explain the different levels of compaction (see Table 12.1).

### Nucleosome supercoiling

One model of additional compaction beyond nucleosomal winding proposes that the 100 Å nucleosomal chromatin supercoils into a 300 Å superhelix, achieving a further sixfold chromatin condensation. Support

for this model comes in part from electron microscope images of 300 Å fibers that contain about six nucleosomes per turn (Fig. 12.5a). Whereas the 100 Å fiber is one nucleosome in width, the 300 Å fiber appears to be three beads wide. Histone H1 likely plays a special role in formation of the superhelix, because removal of some H1 causes the 300 Å to unwind to 100 Å, while adding back H1 reverses this process.

Although electron microscopists can actually see the 300 Å fiber, they do not know its exact structure. Higher levels of compaction are even less well understood.

### The radial loop–scaffold model

This model proposes that several nonhistone proteins bind to chromatin every 60–100 kb and tether the super-helical, nucleosome-studded 300 Å fiber into structural loops (Fig. 12.5b). These proteins may gather the loops into daisylike rosettes and then compress the rosette centers into a compact bundle (Fig. 12.5c). This proposal, known as the **radial loop–scaffold model**, offers a simple explanation of chromosome packaging. Progressive levels of chromosome compaction involving nucleosome formation, nucleosome supercoiling into the 300 Å fiber, looping of the fibers, gathering of loops into rosettes, and rosette compression have the potential to give rise to the highly condensed, rod-like shapes we see as mitotic chromosomes.

Several pieces of biochemical and micrographic evidence support the radial loop–scaffold model. For example, metaphase chromosomes from which experimenters have extracted all the histones still maintain their familiar X-like shapes (see Fig. 12.1). The proteinaceous scaffold that remains includes proteins such as the helicase enzyme topoisomerase II, which maintains DNA winding (review Fig. 6.23), and protein complexes called **condensins** that help condense interphase chromosomes into metaphase chromosomes. Moreover, electron micrographs of mitotic chromosomes extracted in this way show loops of chromatin at the periphery of the chromosomes (Fig. 12.6), as predicted by the model.

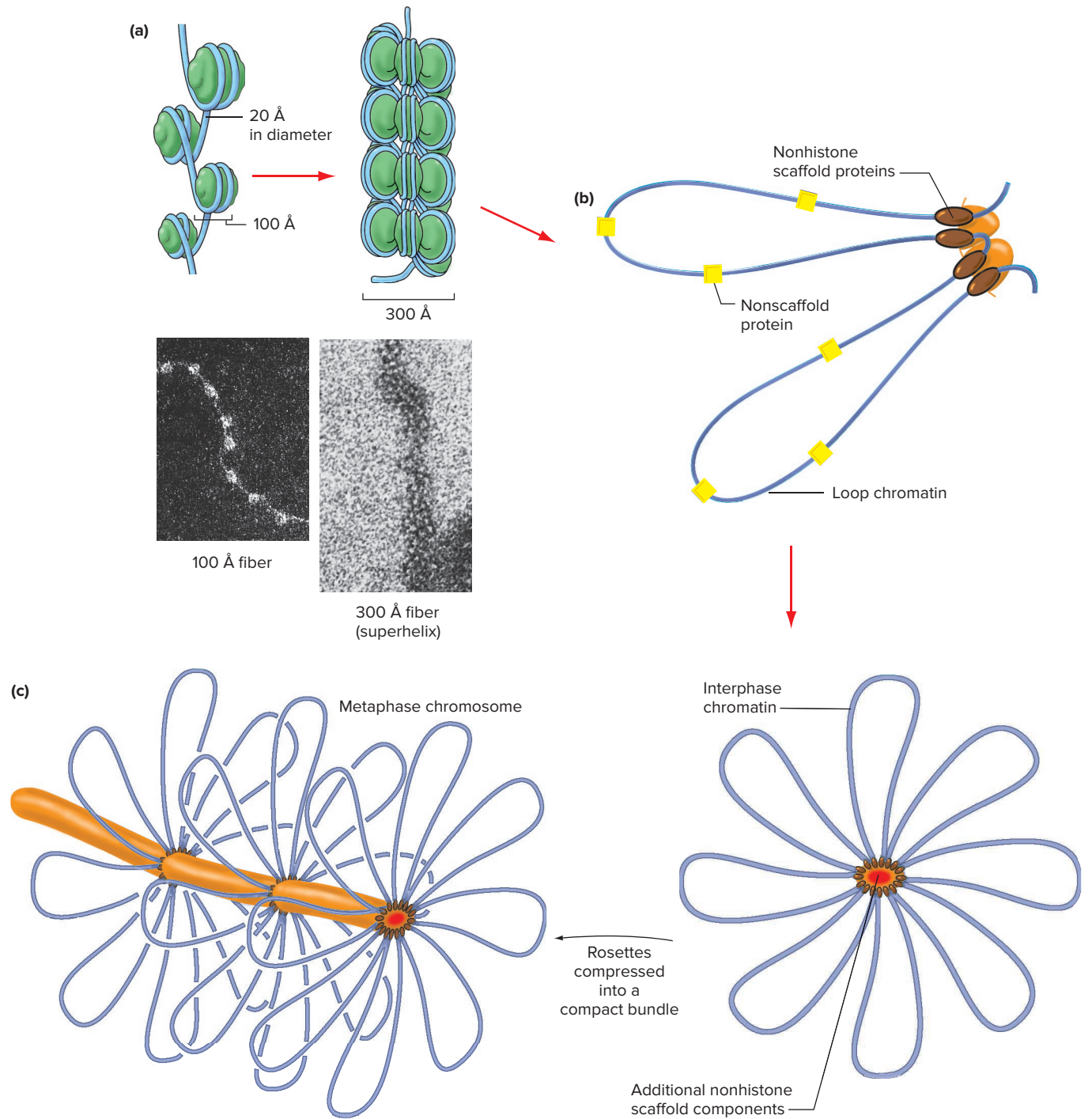
Despite bits and pieces of experimental evidence, studies that directly confirm or reject the radial loop–scaffold model have not yet been completed. Thus, the loops and scaffold concept of higher-order chromatin packaging remains a hypothesis. The hypothetical status of this higher-order compaction model contrasts sharply with nucleosomes, which are entities that investigators have isolated, crystallized, and analyzed in detail.

## Giemsa Staining Reveals Reproducible Chromosome Banding Patterns

We have just seen that different levels of packaging compact the DNA in human metaphase chromosomes

**Figure 12.5 Models of higher-order chromosome packaging.** (a) Artist's conception (top) and electron micrographs (bottom) contrasting the 100 Å fiber (left) with the 300 Å fiber (right). (b) The radial loop–scaffold model for higher levels of compaction. According to this model, the 300 Å fiber is first drawn into loops of 60–100 kb of DNA (blue) tethered by nonhistone scaffold proteins (brown and orange). (c) Additional nonhistone proteins might gather several loops together into daisylike rosettes and then compress the rosettes into bundles, forming metaphase chromosomes.

a1: © Dr. Barbara Hamkalo/University of California-Irvine, Department of Biochemistry;  
 a2: © Dr. Don Fawcett/H. Ris and A. Olins/Science Source



**Figure 12.6** Experimental support for the radial loop–scaffold model. A close-up of part of the image in Fig. 12.1, this electron micrograph shows long DNA loops emanating from the protein scaffold at the *bottom* of the picture.

© Dr. Don Fawcett/J.R. Paulson, U.K. Laemmli/Science Source



10,000-fold (see Table 12.1). With this amount of compaction, the centromere region and telomeres of each chromosome become visible. We have also seen (in Chapter 4) that various staining techniques reveal a characteristic banding pattern, size, and shape for each metaphase chromosome, establishing a karyotype. In *G-banding*, for instance, chromosomes are first gently heated and then exposed to Giemsa stain; this DNA dye preferentially darkens certain regions to produce alternating dark and light **G bands**. Each G band is a very large segment of DNA from 1 to 10 Mb in length, containing many loops. With high-resolution G-banding techniques, a standard diploid human karyotype of 46 chromosomes is seen to contain hundreds of dark and light G bands (**Fig. 12.7**).

The biochemical basis of banding is not yet understood. Most molecular geneticists think the bands produced by Giemsa staining probably reflect an uneven packaging of loops determined in some way by the spacing and density of short, repetitive DNA sequences along the chromosomes. Regardless of the underlying cause, every time a chromosome replicates, its banding pattern is faithfully reproduced. The fact that banding patterns are so highly reproducible from one generation to the next indicates they

**Figure 12.7** Human chromosomes examined by high-resolution G-banding.

© Scott Camazine & Sue Trainor/Science Source

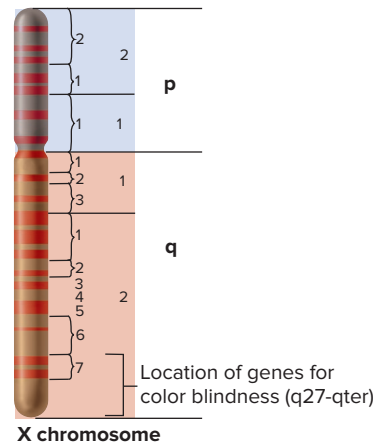


are an intrinsic property of each chromosome, initially determined by the DNA sequence itself.

The reproducibility of this pattern means that geneticists can designate the chromosomal location of a gene by describing its position in relation to the bands on the p (short; after the French *petit*) or q (long; for *queue*, the French word for *tail*) arm of a particular chromosome. For this purpose, the p and q arms are subdivided into regions, and within each region, the dark and light bands are numbered consecutively. Diagrams of the banding patterns, such as the one shown in **Fig. 12.8**, are called **idiograms**. As an example, the X-linked genes for color blindness reside at q27-qter, which indicates they are located on the X chromosome's long (q) arm somewhere between the beginning of the seventh band in the second region and the end of the telomere (terminus, or ter; Fig. 12.8).

**Figure 12.8** Idiogram of the human X chromosome.

Genes for color blindness in humans are located in a small region near the tip of the long arm (q) of the X chromosome.





## Fluorescent *In Situ* Hybridization (FISH) Helps Geneticists Characterize Genomes

Scientists are often faced with a problem of resolution when understanding the relationship of specific DNA sequences to the genome as a whole. Karyotypes show the G-banded chromosomes constituting the whole genome, but obviously at a resolution much lower than that of individual nucleotide pairs. In contrast, it is hard to step back from the mass of information derived from whole-genome sequencing to obtain a global view of genome organization. A technique called **fluorescent *in situ* hybridization (FISH)** provides a convenient bridge between the low resolution of a karyotype and the ultra-high resolution of a complete genomic sequence. In essence, FISH allows investigators to find the locations of specific DNA sequences with respect to the chromosomes in a karyotype.

As the name implies, the fundamental basis of FISH is inherent in nucleotide complementarity. Researchers first obtain cells in mitotic metaphase and then drop the cells onto a glass microscope slide. The cells are then subjected to treatments that successively burst the cells and nuclei open, spread the chromosomes apart, fix the chromosomes on the slide, and gently denature the chromosomal DNA. This latter denaturation step is performed in a way that preserves the overall chromosomal structure, even though the double helixes separate into single strands at numerous points. In a separate reaction, the researchers label a purified DNA sequence with a fluorescent tag, making a *DNA probe*. The probe DNA is denatured into single strands by heating, and the probe is then applied to the chromosomes on the plate. The probe will hybridize only with chromosomal regions that are complementary in nucleotide sequence, and the researchers can identify these regions by looking in a fluorescence microscope.

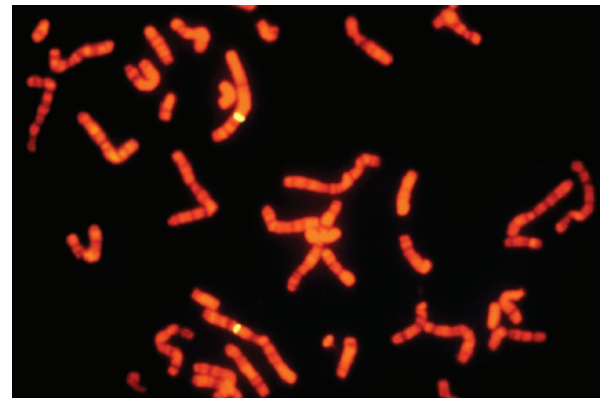
In one use of FISH, the probe is a short, defined sequence such as a cDNA clone. The results are fluorescent spots showing the location of the corresponding gene in the genome (**Fig. 12.9a**). Historically, this method was of considerable importance in verifying that the original draft of the Human Genome Project was properly assembled.

In a second variation of FISH called **spectral karyotyping (SKY)**, the probes are made from multiple DNAs that originated from positions scattered along the length of individual chromosomes. The probe for chromosome 1 is labeled with a mix of fluorescent tags (*dyes* or *fluors*) that together glow in one color, the probe for chromosome 2 is tagged with a different mix of fluors that yields a different color, and so forth, so that each of the 24 human chromosomes in a SKY karyotype can be recognized easily by its color (**Fig. 12.9b**). Chapter 13 will demonstrate that both of these FISH techniques remain important for characterizing chromosomal rearrangements such as deletions, duplications, inversions, and translocations that may cause genetic diseases.

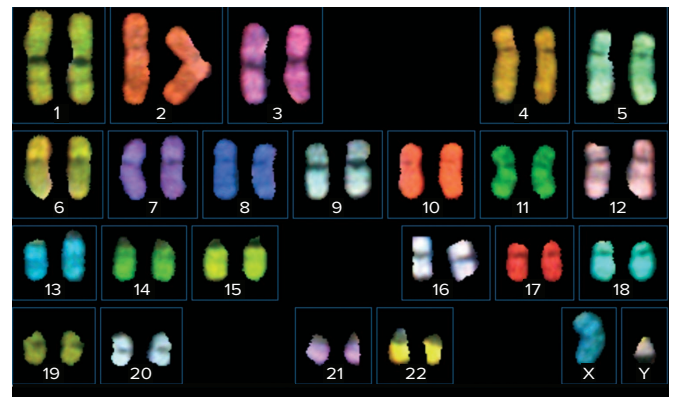
**Figure 12.9** Fluorescent *in situ* hybridization. (a) FISH analysis of a human cell's chromosomes. The *yellow* spots show where a probe made from a single gene hybridizes to the two sister chromatids on each of two homologous chromosomes. Because the two sister chromatids are extremely close together in this preparation, only one *yellow* spot appears on each homolog. (b) Spectral karyotyping (SKY). Probes made from DNA along the length of each chromosome are labeled with fluorescent tags of different colors.

a: © Patrick Landmann/Science Source; b: Courtesy Dr. Thomas Ried, Head, Cancer Genomic Section, Genetics Branch/CCR/NCI/NIH. Image created by Dr. Heses Padilla-Nash

(a) FISH locates a human gene



(b) SKY identifies human chromosomes



### essential concepts

- In a *nucleosome*, DNA wraps twice around a core composed of histones H2A, H2B, H3, and H4. Histone H1 controls entry and exit of DNA.
- Models of higher-order compaction suggest that the nucleosomal fiber is supercoiled into a shorter but wider fiber. Nonhistone proteins anchor this fiber to the chromosome *scaffold*, forming loops that are then gathered together to condense DNA even more.
- Giemsa stain produces *G bands* in metaphase chromosomes. The pattern of G bands is highly specific and reproducible, allowing identification of chromosomes and gene locations.

- The *FISH* technique allows researchers to localize specific DNA sequences with respect to chromosomal bands or to visualize entire chromosomes to facilitate *karyotype* interpretation.

## 12.3 Chromosomal Packaging and Gene Expression

### learning objectives

1. Describe how chromatin remodeling complexes allow gene expression to occur.
2. Differentiate between gene expression in heterochromatic and euchromatic regions.
3. Describe how scientists use position effect variegation to study the mechanisms underlying the formation of heterochromatin.
4. Outline how histone methylation and acetylation affect chromatin structure and gene expression.
5. Summarize the role of the *Xist* gene in X inactivation in mammalian cells.

The compaction of DNA into chromatin presents a problem for proteins that must recognize DNA sequences to carry out functions such as transcription, replication, and segregation. How do these proteins access particular nucleotide sequences that appear to be buried within complex chromatin structures? The answer is that chromatin structure is dynamic and can change to allow access of specific proteins when they need to act. These changes produce variations in chromatin structure necessary for different chromosomal functions.

In this section we focus on the relationship between chromatin structure and gene transcription. We first describe alterations in chromatin that allow RNA polymerase to recognize the promoters of genes and initiate transcription. We then discuss a type of chromatin structure called *heterochromatin* that is associated with chromosomal regions that are not transcribed. The formation of heterochromatin is the molecular basis for several important genetic phenomena, including X chromosome inactivation in mammals.

### Transcription Requires Changes in Chromatin Structure and Nucleosome Position

An important generalization concerning gene expression in eukaryotes is that the less frequently a segment of DNA is transcribed, the more it is compacted. For example, cells express their genes mainly during interphase of the cell cycle

when the chromosomes have decondensed, or decompacted. Little gene transcription occurs on highly compacted metaphase chromosomes. But even the relatively decompacted euchromatic interphase chromatin requires further unwinding to expose the DNA inside nucleosomes for transcription. Gene promoters are hidden from RNA polymerase and transcription factors when the promoter DNA is wrapped around the histone core of a nucleosome (**Fig. 12.10a**).

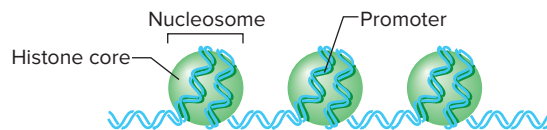
Studies of chromatin structure show that the promoters of most inactive genes are indeed wrapped in nucleosomes. In these studies, the positions of nucleosomes are investigated at the molecular level by treating chromatin with the enzyme DNase I, an enzyme that cleaves phosphodiester bonds in DNA. Sequences within nucleosomes are protected from DNase digestion, while chromosomal regions from which nucleosomes have been eliminated are recognizable by their hypersensitivity to DNase cleavage.

When a previously inactive gene prepares for transcription during a later step of cellular differentiation, the promoter region is observed to change from a DNase-resistant site to a **DNase hypersensitive site**. The reason is that transcription regulatory proteins (*transcription factors*) bind DNA at nearby *enhancers* and recruit proteins that reorganize the chromatin in the vicinity. In particular, these newly recruited proteins remove the promoter-blocking nucleosomes or reposition them in relation to the gene (**Fig. 12.10b**). One type of chromatin modulator consists of multisubunit **remodeling complexes** that use the energy of

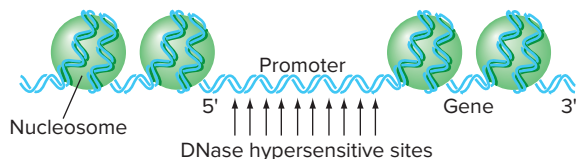
**Figure 12.10** Nucleosome packaging and gene expression.

(a) Gene promoters wrapped around nucleosome cores are not accessible to RNA polymerase and transcription factors. (b) Chromatin remodeling complexes can expose promoters by placing them in nucleosome-free regions that are hypersensitive to DNase. (c) DNA in heterochromatin is so tightly packaged that it is transcriptionally inactive.

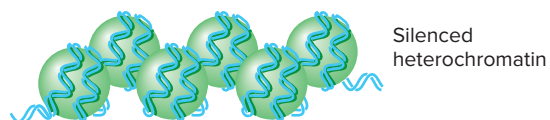
(a) Promoters are hidden when wrapped in nucleosomes.



(b) Chromatin remodeling complexes can expose gene promoters.



(c) Nucleosomes in heterochromatin are tightly packed.



ATP hydrolysis to alter nucleosome positioning. Other chromatin modulators chemically modify the tails of the histones in the nucleosome core (as will be explained later). Chromatin changes accomplished by either mechanism expose the previously hidden promoter, allowing its recognition by RNA polymerase and thus facilitating the gene's transcriptional activation (Fig. 12.10b).

When differentiated cells divide, the transcriptional regulatory proteins that bind DNA and establish the chromatin structure are distributed into both daughter cells. After DNA replication, these proteins rebind DNA and reestablish the chromatin configuration that was present in the parent cell. As a result, differentiated cells have specific patterns of chromatin configuration and gene expression that persist after the cells divide by mitosis.

### Most Genes in Heterochromatin Regions Are Silenced

One type of chromatin organization is widespread in genomes and is correlated with the strong suppression of gene expression. Chromatin organized in this way is visible even in the light microscope because it involves long stretches of DNA: In cells stained with certain DNA-binding chemicals, some chromosomal regions appear much darker than others. Geneticists call these darker regions **heterochromatin**; they refer to the contrasting lighter regions as **euchromatin**. The distinction between euchromatin and heterochromatin also appears in electron microscopy, where the heterochromatin appears much more condensed than the euchromatin. This observation reflects the much tighter packing of nucleosomes in heterochromatic regions (Fig. 12.10c).

Microscopists first identified dark-staining heterochromatin in the decondensed chromatin of interphase cells, where it tends to localize at the periphery of the nucleus. Even highly compacted metaphase chromosomes show the differential staining of heterochromatin versus euchromatin (Fig. 12.11). (This staining is distinct from, and should not

be confused with, the G-banding in karyotyping described earlier.) Most of the heterochromatin in highly condensed chromosomes is found in regions flanking the centromere, but in some animals, heterochromatin forms in other regions of the chromosomes. In *Drosophila*, the entire Y chromosome, and in humans, most of the Y chromosome, is heterochromatic. Chromosomal regions that remain condensed in heterochromatin at most times in all cells are known as **constitutive heterochromatin**.

Autoradiography (a method that detects radioactivity with photographic films) reveals that cells actively expressing genes incorporate radioactive RNA precursors into RNA almost exclusively in regions of euchromatin. This observation indicates that euchromatin contains most of the sites of transcription and thus almost all of the genes. By contrast, heterochromatin appears to be transcriptionally inactive for the most part; it is so tightly packaged that the enzymes required for transcription of the few genes it contains cannot access the correct DNA sequences (Fig. 12.10c).

A high proportion of the DNA located in regions of constitutive heterochromatin consists of long stretches of simple repetitive sequences like SSRs. Heterochromatic regions are also repositories for many *transposable elements*—segments of DNA that move around the genome. SSRs and transposable elements probably accumulate in constitutive heterochromatin because they are transcriptionally silenced there. Repetitive DNAs and transposable elements together constitute more than half of most genomes; their sequestration in transcriptionally inactive heterochromatin provides organisms with a way to minimize the effects of such *junk DNA* on normal cellular physiology.

We now discuss two specialized phenomena—position-effect variegation in *Drosophila* and Barr bodies in mammalian females—that illustrate clearly the correlation between loss of gene activity and heterochromatin formation. These phenomena also helped scientists investigate the biochemical differences between heterochromatin and euchromatin.

### Heterochromatin Can Spread Along a Chromosome and Silence Nearby Euchromatic Genes

The *white*<sup>+</sup> (*w*<sup>+</sup>) gene in *Drosophila* is normally located near the telomere of the X chromosome, in a region of relatively decondensed euchromatin. When a chromosomal rearrangement such as an inversion of a segment of DNA places the gene next to highly compacted heterochromatin near the centromere, the *w*<sup>+</sup> gene's expression may cease (Fig. 12.12). Such rearrangements silence *w*<sup>+</sup> gene expression in some cells and not others, producing **position-effect variegation (PEV)**.

In flies carrying the wild-type *w*<sup>+</sup> allele, cells in the eye with an active *w*<sup>+</sup> gene are red, while cells with an inactive

**Figure 12.11** Constitutive heterochromatin. Human metaphase chromosomes were stained using a special technique that darkens the constitutive heterochromatin, most of which is in regions surrounding the centromeres.

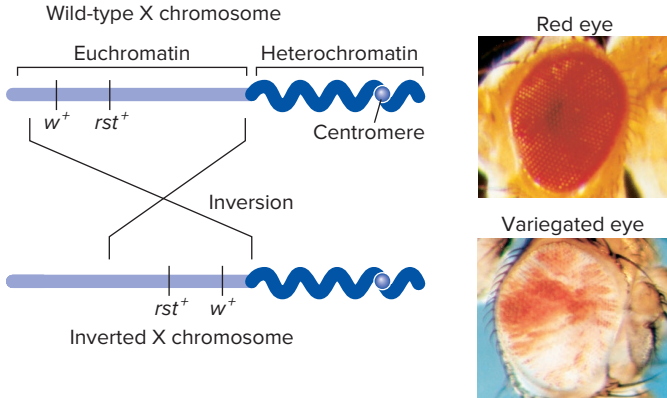
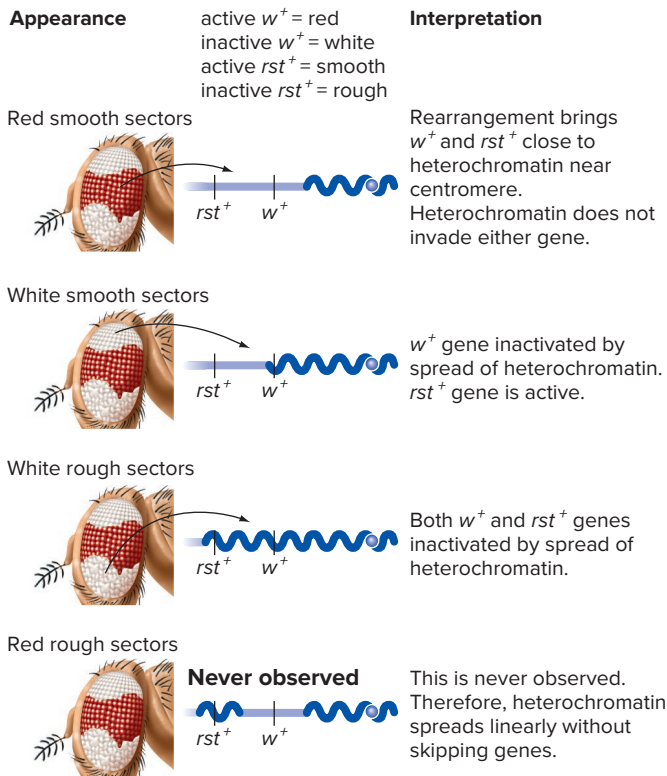
© Doug Chapman, University of Washington Medical Center Cytogenetics Laboratory



**Figure 12.12** Position-effect variegation in *Drosophila*.

(a) When the  $w^+$  gene is brought near an area of heterochromatin by chromosomal inversion, the fly's eyes become variegated, with some red cells and some white cells. (b) A model for position-effect variegation postulates that heterochromatin can spread linearly from its normal location surrounding the centromere to nearby genes, causing their inactivation.

a(both): © Dr. Clinton Bishop, Department of Biology, West Virginia University

**(a) Heterochromatin can turn off adjacent genes.****(b) Heterochromatin spreads linearly.**

$w^+$  gene are white. Apparently, when normally euchromatic genes like  $w^+$  come into the vicinity of heterochromatin, the heterochromatin can spread into the euchromatic regions, shutting off gene expression in those cells where the heterochromatic invasion takes place. In such a situation, the DNA of the gene has not been altered, but the relocation has

altered the gene's packaging in some cells. The phenomenon of position-effect variegation thus reflects the existence of **facultative heterochromatin**: regions of chromosomes (or even whole chromosomes) that are heterochromatic in some cells and euchromatic in other cells of the same organism.

Position-effect variegation of red and white eye color in *Drosophila* produces eyes that are a mosaic of red and white patches (Fig. 12.12a). The numbers, positions, and sizes of the patches vary from eye to eye. Such variation suggests that the decision determining whether heterochromatin spreads to the  $w^+$  gene in a particular cell is the result of a random process. Because patches composed of many adjacent cells have the same color, the decision must be made early in the development of the eye. Once made, the decision determining whether the *white* gene will be on or off is perpetuated by all of the cell's mitotic descendants. These descendants occupy a particular region of the eye, forming patches of red or white cells, respectively.

**Scope of heterochromatin effects**

One interesting property revealed by position-effect variegation is that heterochromatin can spread over more than 1000 kb of previously euchromatic chromatin. For example, some rearrangements that bring the  $w^+$  gene near heterochromatin also place the *roughest* gene in the same vicinity, although a little farther away from the centromeric heterochromatin (Fig. 12.12a). The wild-type *roughest* ( $rst^+$ ) gene normally produces a smooth eye surface. In flies carrying the rearrangements, some white-colored patches have smooth surfaces while others have rough surfaces (Fig. 12.12b). In the latter patches, the heterochromatin inactivated both the  $w^+$  and the  $rst^+$  gene. Red-colored, rough-surfaced patches never form, which means that the heterochromatin does not skip over genes as it spreads linearly along the chromosome.

If heterochromatin can spread, how is the boundary between heterochromatin and euchromatin normally formed? Research has identified DNA segments called **barrier elements** which block the spread of heterochromatin. The exact mechanism by which barrier elements work is unclear, but current models suggest these DNA elements recruit enzymes that modify histone proteins, as will be explained later in the chapter.

**Using PEV to identify heterochromatin components**

Scientists have taken advantage of the phenomenon of position-effect variegation in *Drosophila* to explore the molecules involved in heterochromatin formation. By looking for changes in the amount of variegation, researchers have obtained mutations that alter its efficiency. Enhancement of variegation (eyes more white) reflects gene inactivation in a larger number of cells; suppression of variegation (eyes more red) reflects gene inactivation in fewer cells.

The researchers later isolated several of the genes whose mutant alleles modified variegation, and they created antibodies against the mutant protein products of these genes. In this way, they discovered that some of the genes influencing heterochromatin formation encode proteins that localize selectively to the heterochromatin. As described in the next section, the identification of these proteins has provided important clues about the biochemical control of chromatin structure.

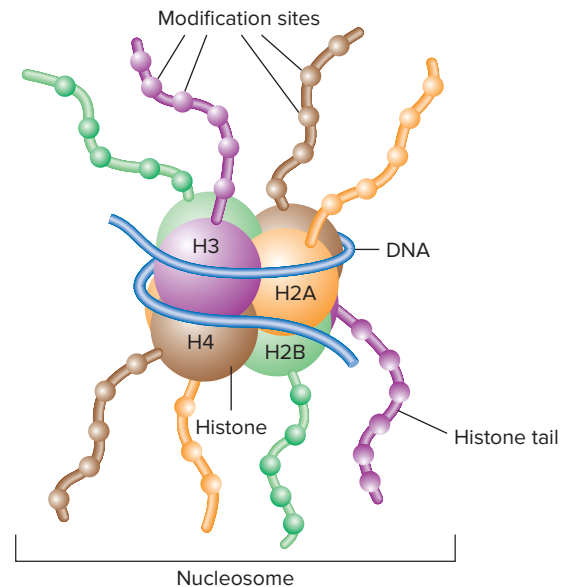
## Heterochromatin and Euchromatin Have Different Histone Modifications

Several interdependent mechanisms govern the distinction between active (or potentially active) euchromatin and silenced heterochromatin. We focus here on one of the most important of these mechanisms, involving covalent modifications of the histones in the nucleosomes. Discussion of the interactions of chromatin modifier proteins with the transcription factors that regulate gene expression will be deferred until Chapter 17.

### Histone tail modifications

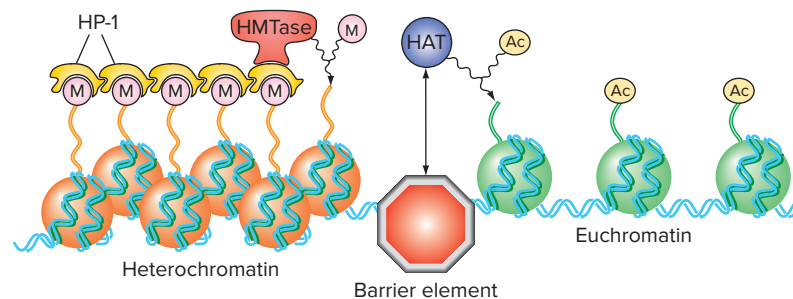
The N-terminal regions of the four core histones—H2A, H2B, H3, and H4—form tails that extend outward from the nucleosome (Fig. 12.13). Enzymes can add several different kinds of chemical groups (among others, methyl groups, acetyl groups, phosphate groups, and ubiquitin; review Fig. 8.26) to various amino acids along these tails, while other enzymes can remove groups that were added previously. Such modifications of these **histone tails** can influence the packing of nucleosomes, and the modified tails can also serve as platforms to which chromatin modifier proteins can bind. The histone tails of a nucleosome core potentially could be modified in more than 100 ways (Fig. 12.13).

**Figure 12.13 Histone tail modifications.** The N-terminal tails of the core histone proteins extend outward from the nucleosome. Various amino acids in these tails are targets for modifications such as methylation and acetylation that can alter chromatin structure.



The best understood of these histone tail modifications are additions of acetyl groups to specific lysines (acetylation) and additions of methyl groups to specific lysines and arginines (methylation) (Fig. 12.14). Lysine acetylation, accomplished by a family of enzymes called *histone acetyltransferases (HATs)*, opens chromatin by preventing the close packing of nucleosomes. Histone acetylation thus favors the expression of genes in euchromatic regions, as their promoters are now accessible to RNA polymerase and its associated proteins. Interestingly, the acetylated lysines on histone tails serve as binding sites for HAT enzymes, thus facilitating the spreading of histone acetylation to neighboring nucleosomes. *Histone deacetylase enzymes (HDACs)* that remove the acetyl groups reverse the process, resulting in closed chromatin and repressed transcription.

**Figure 12.14 Histone tail modifications alter chromatin structure.** In the heterochromatin (orange), K9-methylated (M) nucleosomes are bound by HP1 protein, which attracts HMTase enzymes that methylate adjacent nucleosomes and thus spread the inactive closed state. In euchromatic regions (green), HAT enzymes acetylate (Ac) core histones, resulting in open chromatin with nucleosomes sufficiently far apart so as to be accessible for transcription. *Barrier elements* that block the spread of heterochromatin are likely DNA sequences that attract HAT enzymes as shown, and/or demethylase enzymes that reverse K9 methylation (not shown).



Methylation of histone tails is more complex, and may either close or open chromatin depending on the particular amino acid methylated. The enzymes that methylate histone tail amino acids are *histone methyltransferases* (*HMTases*), and enzymes that reverse histone methylation are called *histone demethylases*.

One of the genes whose loss-of-function mutant alleles act as PEV suppressors in *Drosophila* codes for an HMTase enzyme that adds methyl groups to a specific lysine (K9) in histone H3. This specific methylation marks the chromatin for assembly into heterochromatin by providing binding sites for heterochromatin-specific proteins. The methylation of histone H3 K9 is a common feature of chromosomal regions that are transcriptionally silenced (Fig. 12.14).

### Heterochromatin-specific proteins

A different PEV suppressor gene in *Drosophila* encodes HP1, a key heterochromatin protein that binds to histone H3 tails containing methylated K9. The HP1 protein promotes chromatin compaction into heterochromatin in two ways (Fig. 12.14). First, it self-associates, and this helps to bring adjacent nucleosomes closer together. Second, HP1 binds other proteins, notably including the HMTase enzyme that adds methyl groups to the same K9 amino acid of histone H3. The recruited HMTase can methylate K9 on the histone H3s of adjacent nucleosomes. This autocatalytic effect provides at least a partial explanation for the linear spreading of heterochromatin observed in PEV.

As shown in Fig. 12.14, the choice between heterochromatic and euchromatic states involves competition between the enzymes that add and remove methyl and acetyl groups to various amino acids in the histone tails. Once the chromatin marks typical of euchromatin or heterochromatin predominate in a cluster of nucleosomes, they can spread to nearby nucleosomes until they reach barrier elements. The barriers that stop the spread of heterochromatin are probably DNA sequences that recruit HATs and histone demethylases, tilting the local competition toward the formation of euchromatin.

### Heterochromatin Formation Inactivates an X Chromosome in Cells of Female Mammals

You will remember from Chapter 4 that mammals compensate for the difference in dosage of X-linked genes between males and females by randomly inactivating one of the two X chromosomes in each of the female's somatic cells. In some cells, the X inherited from the mother is inactivated; in others, it is the X inherited from the father (review Fig. 4.25). The inactive X chromosomes, or **Barr bodies**, are examples of facultative heterochromatin: An entire X chromosome becomes nearly completely heterochromatic in

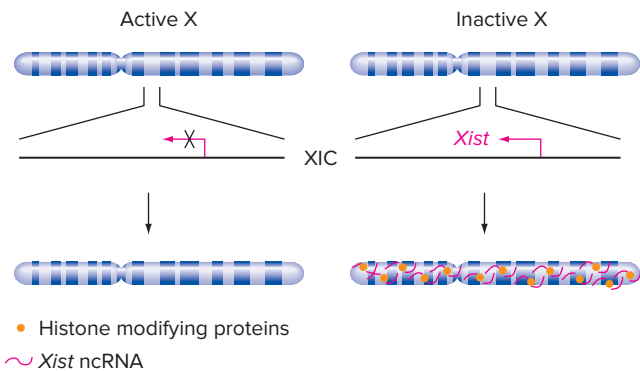
some cells, while other copies of this same X chromosome remain euchromatic in other cells. Most genes on an X chromosome are available for transcription only in cells where the chromosome is euchromatic. In contrast, only a few genes (mainly those in the pseudoautosomal regions) are available for transcription on an X chromosome that has become a heterochromatic Barr body.

Human X chromosomes contain a 450 kb region of DNA called the **X inactivation center (XIC)** that mediates dosage compensation (Fig. 12.15a). The role of the XIC

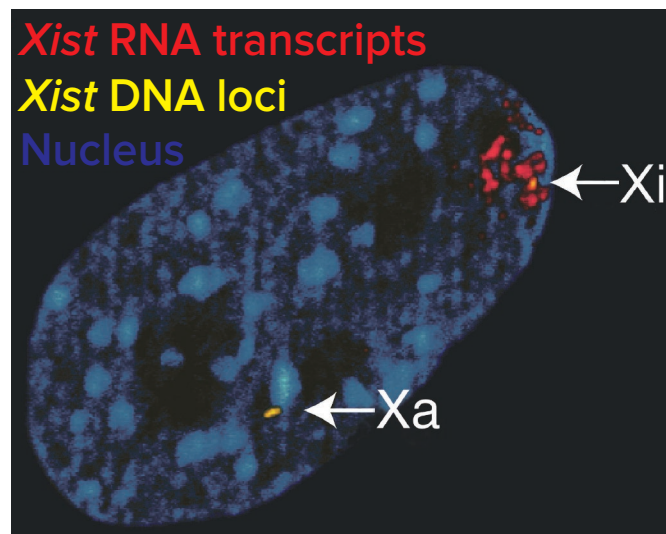
**Figure 12.15 X chromosome inactivation.** (a) One of the genes in the X inactivation center (XIC) is *Xist*, whose product is a nontranslated ncRNA. *Xist* is transcribed stably only from the inactive X chromosome. The *Xist* ncRNA binds to many sites on this chromosome; the *Xist* ncRNA then attracts histone-modifying proteins that silence the DNA. (b) Cells from an XX female mouse showing binding of *Xist* ncRNA (red) to the inactivated X chromosome ( $X_i$ ; the Barr body) but not the active X chromosome ( $X_a$ ). DNA is stained blue; the *Xist* genes on both X chromosomes are yellow.

b: From: B. Reinius et al. (2010), "Female-biased expression of long non-coding RNAs in domains that escape X-inactivation in mouse," *BMC Genomics*, 11:614, Fig. 5. © Reinius et al. Licensee BioMed Central Ltd. 2010. <http://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-614>

#### (a) *Xist* is transcribed from the inactive X chromosome



#### (b) *Xist* ncRNA binds the Barr body



was established by experiments in which a copy of the XIC was transferred to an autosome, and remarkably, that autosome then became a Barr body. The most important known gene in the XIC is called *Xist* (**X inactive specific transcript**). The *Xist* gene product is an unusually long (~17 kb) **noncoding RNA (ncRNA)** that, unlike most transcripts, never leaves the nucleus and is never translated into a protein.

*Xist* is transcribed stably only from the future inactive X chromosome, and it is the *Xist* ncRNA that triggers inactivation of the X chromosome from which it is transcribed. We know that *Xist* is a crucial gene that mediates inactivation because if *Xist* is deleted from one X chromosome, the other X always becomes the Barr body. Conversely, if a copy of *Xist* is added to an autosome, that autosome now becomes a heterochromatic Barr body. Researchers are currently exploring the possibility of adding *Xist* to one of the three copies of chromosome 21 in children with Down syndrome. In theory, this strategy could work as a kind of gene therapy to ameliorate the symptoms of Down syndrome, by turning off gene expression from the additional chromosome 21.

Studies using fluorescent molecular probes show that *Xist* ncRNA coats the X chromosome that produces it (**Fig. 12.15b**). The *Xist* ncRNA then recruits histone-modifying enzymes to this X chromosome. The enzymes methylate and deacetylate the histone tails in the nucleosomes of this chromosome, helping to condense it into a Barr body. Inactive X chromosomes also display other characteristics of silenced chromatin, such as the methylation of certain nucleotides in the DNA, as will be described in Chapter 17.

Although Fig. 12.15 shows some of the molecular details of X chromosome inactivation that have been worked out to date, several mysteries remain. Chapter 4 explains that in humans, about two weeks after fertilization, each of the 500–1000 cells in an XX female embryo decides independently to condense one of the two X chromosomes to a Barr body. It is unclear at present how individual cells choose which of the X chromosomes to inactivate. In addition, once an X chromosome is chosen to transcribe *Xist*, why does the *Xist* ncRNA bind only to that X chromosome?

Yet another area of active research involves the fact that X inactivation is inherited in somatic cells: Once the decision is made, it is clonally perpetuated so that all of the millions of cells descended by mitosis from a particular embryonic cell condense the same X chromosome to a Barr body. Scientists do not yet understand how the cell remembers which X chromosome to recondense into a Barr body following cell division.

### essential concepts

- *Remodeling complexes* use the energy from ATP hydrolysis to change the position of nucleosomes, exposing promoters and allowing gene transcription.

- In regions of *heterochromatin*, promoters are tightly wrapped in nucleosomes, preventing transcription and thus silencing those genes.
- *Position-effect variegation* occurs when a chromosomal rearrangement places a gene next to the heterochromatic region near the centromere. If the heterochromatin spreads into the gene, then the gene will be silenced.
- *Methylation* of a specific lysine (K9) in histone H3 is a feature of silenced regions. In contrast, *acetylation* of lysines in histone tails opens chromatin and allows gene expression.
- The *Xist* gene is responsible for X inactivation. The *Xist noncoding RNA (ncRNA)* coats the X chromosome that produces it and recruits histone-modifying enzymes to inactivate the chromosome, which becomes a *Barr body*.

## 12.4 Replication of Eukaryotic Chromosomes

### learning objectives

1. Explain how replication of the DNA in long eukaryotic chromosomes can occur in a short period of time.
2. Summarize the process by which nucleosomes are re-formed during replication.
3. Discuss the structure of telomeres and their roles in maintaining chromosome integrity.
4. Describe the action of telomerase and identify the cell types in which it continues to be synthesized.

Just as with transcription, replication of chromosomes requires reading of their DNA. Historically, scientists have been able to observe in the light microscope some aspects of the processes by which chromosomes replicate and become segregated into daughter cells. The mechanics at the molecular level, however, have been discovered only recently. In this section we review what is currently known about the replication of eukaryotic chromosomes.

In Chapter 6 we discussed in molecular detail how DNA polymerase cooperates with other factors to replicate DNA in bacterial cells. Although many aspects of DNA replication are similar in eukaryotes, these cells face additional challenges. First, eukaryotic cells have much more DNA than prokaryotic cells, all of which needs to be copied within the short span of a cell cycle. Second, the DNA replication machinery in eukaryotic cells must be able to operate even though the DNA is wrapped around nucleosomes. Finally, eukaryotic chromosomes are linear rather than circular, and as you will see, the ends of linear chromosomes are difficult to copy. We discuss in this

section how eukaryotic cells deal with these special issues in chromosome replication.

## Chromosomal DNA Replication Begins at Specific Origins

When DNA is copied, the enzyme DNA polymerase assembles a new string of nucleotides according to a DNA template, linking about 50 nucleotides per second in a typical human cell. At this rate and with only one origin of replication, it would take the polymerase about 800 hours, a little more than a month, to copy the 130 million base pairs in an average human chromosome. But the length of the cell cycle in actively dividing human tissues is much shorter, some 24 hours, and S phase (the period of DNA replication) occupies only about a third of this time. Eukaryotic chromosomes meet these time constraints by firing multiple origins of replication that can function simultaneously.

Most mammalian cells carry about 10,000 origins positioned strategically among the chromosomes. As you saw in Chapter 6, each origin of replication binds proteins that unwind the two strands of the double helix, separating them to produce two mirror-image replication forks. Replication then proceeds in two directions (bidirectionally), going one way at one fork and the opposite way at the other. As replication opens up a chromosome's DNA, a *replication bubble* becomes visible in the electron microscope, and with many origins, many bubbles appear (**Fig. 12.16**). These bubbles increase in size until adjacent bubbles run into each other, eventually allowing the entire chromosome to be copied.

The DNA running both ways from one origin of replication to the endpoints, where it merges with DNA from

adjoining replication forks, is called a **replication unit**, or **replicon**. As yet unidentified controls tie the number of active origins to the length of S phase. In *Drosophila*, for example, early embryonic cells replicate their DNA in less than 10 minutes. To complete S phase in this short a time, their chromosomes use many more origins of replication than are active later in development when S phase is much longer. Thus, not all origins of replication are necessarily active during all the mitotic divisions that create an organism. The 10,000 origins of replication scattered throughout the chromatin of each mammalian cell nucleus are separated from each other by 30–300 kb of DNA, which suggests that at least one origin of replication exists per loop of chromatin.

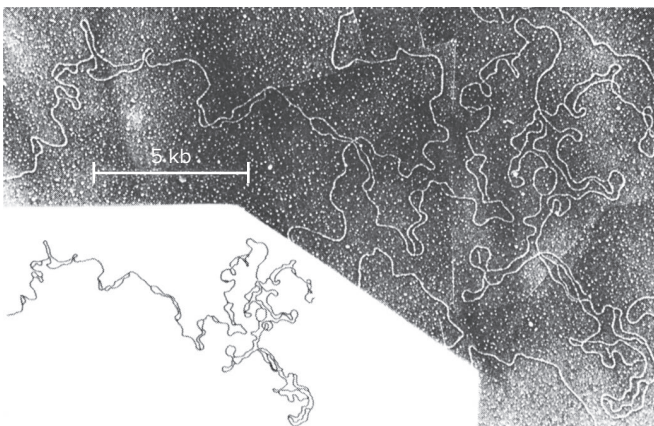
Origins of replication in yeast (known as *autonomously replicating sequences*, or *ARSs*) can be isolated by their ability to permit replication of plasmids in yeast cells. ARSs are capable of binding to the enzymes that initiate replication. Almost all ARSs contain a variant of the 11 bp AT-rich consensus sequence 5' T/ATTTAYRTTTT/A 3', where Y is either pyrimidine and R is either purine. Several other nearby sequence motifs also contribute to the function of ARSs as origins of replication. By digesting interphase chromatin with DNase I, an enzyme that fragments the chromatin only at points where the DNA is not protected by its association with a nucleosome, investigators have determined that origins of replication lie within accessible regions of DNA devoid of nucleosomes.

## New Nucleosomes Must Be Formed During DNA Replication

DNA replication is only one step in chromosome duplication. The complex process also includes the synthesis and incorporation of histone and nonhistone proteins to regenerate nucleosomes and chromatin structure. While some aspects of the following model are controversial, researchers have obtained evidence that the process works something like this:

- The synthesis and transport of histones must be tightly coordinated with DNA synthesis because the nascent DNA becomes incorporated into nucleosomes within minutes of its formation.
- As DNA replication takes place, nucleosomes assemble rapidly on newly formed daughter DNA molecules. The new nucleosomes are a mixture of old (recycled) and newly formed histones, distributed randomly on the two daughter DNA molecules (**Fig. 12.17**).
- Some histone modifications may be retained at the replication fork to some extent through self-propagation because some modified histone tail amino acids attract modification enzymes (**Fig. 12.14**). However, even in these cases retention is inefficient. The reason is that only recycled histones can have modifications and about half the histones on newly replicated DNA are new. Moreover, histone modifications are labile, and so not all recycled histones retain their original marks.

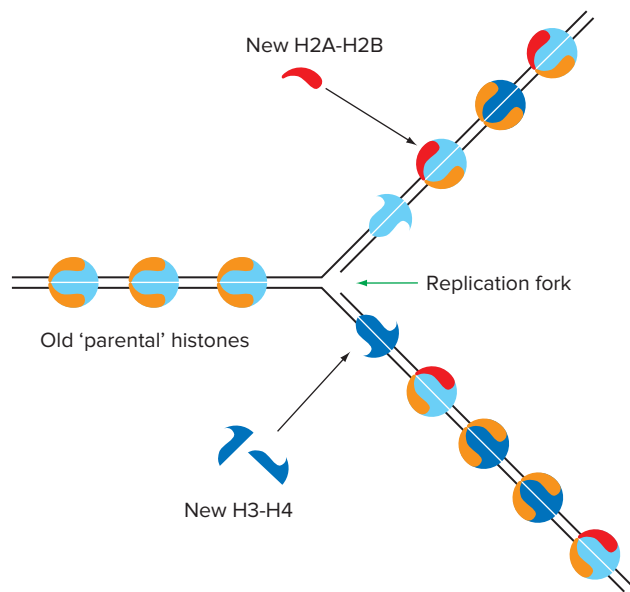
**Figure 12.16 Eukaryotic chromosomes have multiple origins of replication.** Electron micrograph and diagrammatic interpretation, showing replicating DNA from a *Drosophila* embryo. Many origins of replication are active, creating multiple replicons.  
© H. Kreigstein and D.S. Hogness, "Mechanism of DNA Replication in *Drosophila* Chromosomes: Structure of Replication Forks and Evidence of Bidirectionality", *PNAS*, 71(1974): 135-139





**Figure 12.17 Nucleosome creation after DNA replication.**

The replication fork disassembles the nucleosomes it encounters in the parental chromosome. New nucleosomes start assembling immediately after the replication fork passes. First, a tetramer containing two molecules each of histones H3 and H4 associates with DNA to form a half nucleosome, followed by two dimers each containing one molecule of H2A and one of H2B. The new nucleosomes can contain various random combinations of H3/H4 tetramers and H2A/H2B dimers that were either newly synthesized or that were previously found in parental nucleosomes.



If no robust mechanism exists for copying histone marks during DNA replication, why is chromatin structure often recapitulated after differentiated cells undergo mitosis? The answer appears to be that just after replication, chromatin formed from new nucleosomes is open. A brief window of opportunity thus exists for transcription regulatory proteins in the nucleus to bind to the daughter DNA molecules. These proteins then recruit histone-modifying enzymes, re-creating the chromatin structure of the parental chromosome even though the histone modifications were briefly lost when the DNA was replicated.

### Telomeres Protect the Ends of Linear Chromosomes and Allow Their Replication

The linear chromosomes of eukaryotic cells terminate at both ends in protective caps called **telomeres** (Fig. 12.18). Composed of special DNA sequences associated with specific proteins, these caps contain no genes but are crucial in preserving the structural integrity of each chromosome.

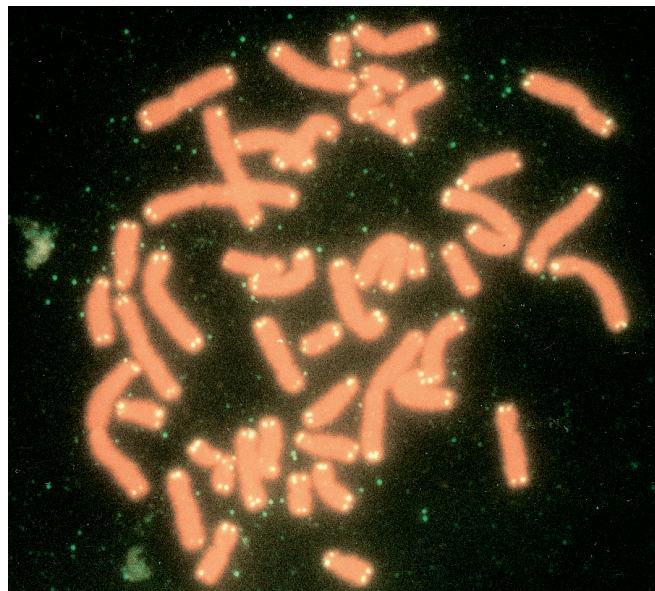
#### Telomeres and the replication of chromosome ends

The replication of the ends of linear chromosomes poses a difficult problem for cells. As you saw in Chapter 6, DNA

### Figure 12.18 Telomeres protect the ends of eukaryotic chromosomes.

Human telomeres light up in yellow upon FISH with probes that recognize the repeating base sequence TTAGGG.

© Dr. Robert Moyzis/University of California-Irvine, Department of Biochemistry

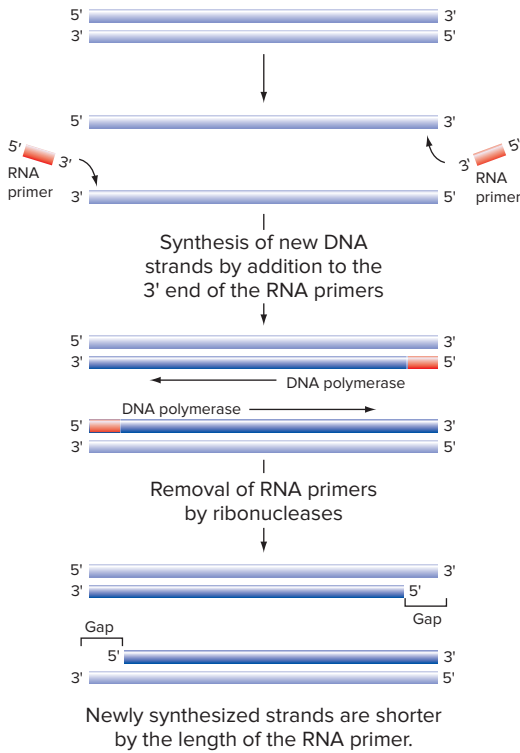


polymerase, a key component of the replication machinery, functions only in the 5'-to-3' direction: It can add nucleotides only to the 3' end of an existing chain. With this constraint, the enzyme on its own cannot possibly replicate some of the nucleotides at the 5' ends of the two DNA strands (one of which is in the telomere at one end of the chromosome, the other of which is in the telomere at the other end). As a result, DNA polymerase can reconstruct the 3' end of each newly made DNA strand in a chromosome, but not the 5' end (Fig. 12.19).

We saw in Chapter 5 that because DNA polymerase cannot start synthesis at the 5' end of the new chain, it relies on RNA primers to do the job. These RNA primers are eventually removed. Thus, at the 5' ends of linear chromosomes in eukaryotic cells, an RNA primer's length of nucleotides will be missing at the 5' end of every new chromosomal strand every time the DNA is copied (Fig. 12.19). As a result, the chromosomes in successive generations of cells would become shorter and shorter, losing crucial genes as their DNA diminished.

Telomeres and an enzyme called **telomerase** provide a countermeasure to this limitation of DNA polymerase. Telomeres consist of particular repetitive DNA sequences that do not encode proteins. Human telomeres are composed of the base sequence 5' TTAGGG 3' (read on one strand) repeated 250 to 1500 times. The number of repeats varies with the cell type; sperm have the longest telomeres. The same TTAGGG sequence occurs in the telomeres of all mammals as well as in birds, reptiles, amphibians, bony

**Figure 12.19 Replication at chromosome ends.** Even if an RNA primer (red) at the 5' end can begin synthesis of a new strand, a gap will remain when ribonucleases eventually remove the primer. DNA polymerase cannot fill this gap without a primer, so newly synthesized strands (dark blue) will be shorter than parental strands (light blue).



fish, and many plant species. Some much more distantly related organisms also have repeats in their telomeres but with slightly different sequences. For example, the telomeric repeat in the chromosomes of the ciliate *Tetrahymena* is TTGGGG. The close similarities of these repeated sequences across phyla suggests that they perform a vital function that emerged in the earliest stages of the evolutionary line leading to eukaryotic organisms.

Telomerase is an unusual enzyme consisting of protein in association with RNA. Because of this mix, it is called a *ribonucleoprotein*. The RNA portion of the enzyme contains 3' AAUCCC 5' repeats that are complementary to the 5' TTAGGG 3' repeats in telomeres, and they serve as a template for adding new TTAGGG repeats to the end of the telomere (Fig. 12.20). You can easily see how the addition of new repeats could counterbalance the loss of DNA that must occur when linear DNA molecules are copied.

**Telomerase activity and cell proliferation**

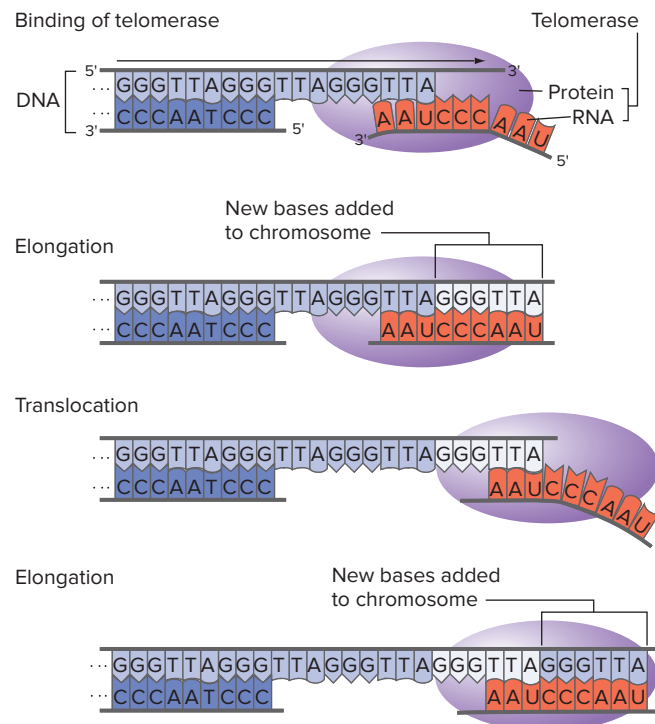
Some types of cells take advantage of the ability of telomerase to maintain the size of their chromosome ends upon cell division, but other cell types make very little if any telomerase, and in fact their chromosomes shorten during

every cell generation. Most differentiated somatic cells in humans are of this latter type. Even though these cells have the telomerase gene in their genomes, they do not express this enzyme, so the telomeres shorten slightly with each cell division. After 30 to 50 cell generations, the chromosomes begin to lose essential genes from their ends; the cells start showing signs of senescence and then die. The lack of telomerase thus ensures that differentiated somatic cells have a finite life span.

In contrast, germ-line cells do express telomerase and thus maintain their chromosomal ends through repeated rounds of DNA replication without the loss of genes. (If this were not the case, our species would have died out long ago.) Apparently, germ-line cells have some kind of feedback mechanism that maintains the optimal number of repeats at the telomeres so that the chromosome ends neither shorten nor lengthen appreciably during each generation of humans.

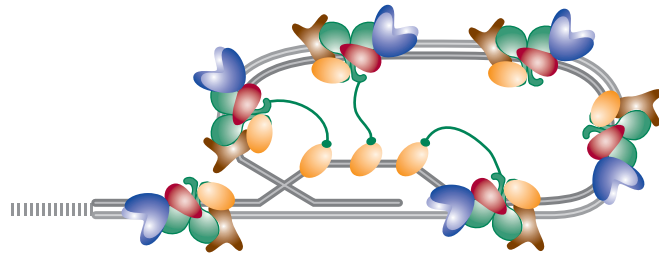
Two kinds of somatic cells in our bodies can also make telomerase, and thus have the potential to reproduce for many generations, if not forever. One class of such cells are the *stem cells* that allow tissue renewal, such as the continual production of blood cells. The second class consists of tumor cells: somatic cells gone awry that can

**Figure 12.20 How telomerase extends telomeres.** The 3' AAUCCC 5' repeats of telomerase RNA (red) are complementary to the 5' TTAGGG 3' repeats of telomeres. Telomerase RNA thus serves as a template for adding TTAGGG repeats to telomere ends. After a new repeat is added, telomerase moves (translocates) to the newly synthesized end, allowing additional rounds of telomere elongation.



**Figure 12.21** The shelterin complex protects telomeres.

The proteins of the shelterin complex (colored shapes) bind to telomeres, folding the DNA ends (gray) so they can neither be attacked by nucleases nor subjected to nonhomologous end-joining.



divide indefinitely, becoming seemingly immortal. Because high telomerase activity is a characteristic of many tumor cells, pharmaceutical companies are developing cancer treatment drugs that inhibit this important enzyme.

### Telomeres and chromosome integrity

Chromosomes that are broken (for example, by exposure to X-rays) and detached from their telomeres pose many dangers for cells. One problem is that cells contain nuclease enzymes that can progressively degrade DNA inward from the broken ends. A second problem results from the enzyme systems responsible for nonhomologous end-joining (NHEJ) that were described in Fig. 7.18. If two different chromosomes are broken, these NHEJ enzymes will join the chromosomes together. The end-to-end fusion of broken chromosomes produces entities with two centromeres. During anaphase of mitosis, if the two centromeres are pulled in opposite directions, the DNA between them will rupture, resulting in broken chromosomes that segregate poorly and eventually disappear from the daughter cells. These examples demonstrate that the telomeres on normal, unbroken chromosomes have important functions in protecting the chromosomes and maintaining the correct genetic complement in cells.

The protective function of telomeres is due to proteins, different from telomerase, that also bind to the TTAGGG repeats at the very ends of a chromosome. These proteins form a complex called **shelterin** that folds up the telomeres into a structure that shields single-stranded TTAGGG sequences from nucleases and NHEJ enzymes (Fig. 12.21).

### essential concepts

- In eukaryotic cells, DNA copying initiates nearly simultaneously at thousands of *origins of replication*.
- During replication, new nucleosomes assemble on the daughter DNA molecules. Regulatory proteins bind to the daughter DNAs and recruit histone-modifying enzymes to re-create the parental chromatin structure.

- Because DNA polymerase cannot copy the 5' ends of linear DNA molecules, chromosomes will shorten every time they are replicated, leading to gene loss and cell death. To counteract this shortening, *telomeres* contain repeating sequences that can be extended by the enzyme *telomerase*.
- Most somatic cells do not express telomerase and thus have limited life spans. Cells that continue to express telomerase include germ-line cells, stem cells, and tumor cells.

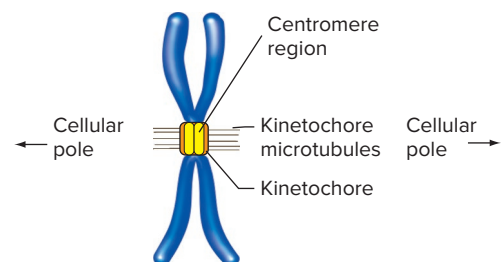
## 12.5 Chromosome Segregation

### learning objectives

1. Contrast the DNA sequences dictating centromere formation in yeast and in higher eukaryotes.
2. Describe spindle attachment during metaphase and the role tension on kinetochores plays in ensuring proper chromosome segregation.
3. Compare the behavior of cohesin complexes during mitosis, meiosis I, and meiosis II.

When cells divide during mitosis or meiosis II, the two chromatids of each replicated chromosome must separate from one another at anaphase and segregate such that each daughter cell receives one and only one chromatid from each chromosome. In contrast, during meiosis I, homologous chromosomes must pair and segregate such that each daughter cell receives one and only one chromosome from each homologous pair. The *centromeres* of eukaryotic chromosomes ensure this precise distribution during different kinds of cell division by serving as segregation centers. The centromeric region is where sister chromatids are most tightly bound together; in addition, structures at the centromeres called *kinetochores* are the locations at which chromosomes bind to spindle fibers (Fig. 12.22).

**Figure 12.22** Centromeres. Cohesin proteins at centromeres (yellow) hold sister chromatids together. Centromeres also contain information for the construction of kinetochores (orange), the structures that allow the chromosomes to bind to spindle fibers.



## Special DNA Sequences Allow the Formation of Centromeres

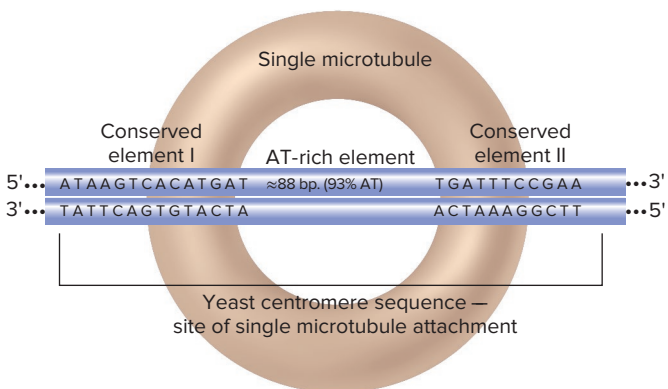
In the yeast *S. cerevisiae*, centromeres consist of two highly conserved nucleotide sequences, each only 10–15 bp long, separated by approximately 90 bp of AT-rich DNA (Fig. 12.23). Evidently, a short stretch of roughly 120 nucleotides is sufficient to specify a centromere in this organism. The centromere sequences of different yeast chromosomes are so closely related that the centromere of one chromosome can substitute for that of another. This fact indicates that all centromeres play the same role in chromosome segregation, and they do not help distinguish one chromosome from another.

The centromeres of higher eukaryotic organisms are much larger and more complex than those of yeast. These more complicated centromeres are contained within blocks of certain repetitive, noncoding sequences known as **satellite DNAs**. Many different kinds of satellite DNA exist, each consisting of short sequences 5–300 bp long, repeated in tandem thousands or millions of times to form large arrays. The predominant human satellite at centromeres,  $\alpha$ -satellite, is a noncoding sequence 171 bp in length; it is present in a block of tandem repeats extending over a megabase of DNA in the centromeric region of each chromosome. Various human centromeres also contain repetitive sequences unrelated to  $\alpha$ -satellite; these sequences impart heterochromatic characteristics to centromeric regions, as was seen in Fig. 12.11.

## Kinetochores Govern Attachment of Chromosomes to the Spindle

One of the important ways in which centromeres contribute to proper chromosome segregation is through the

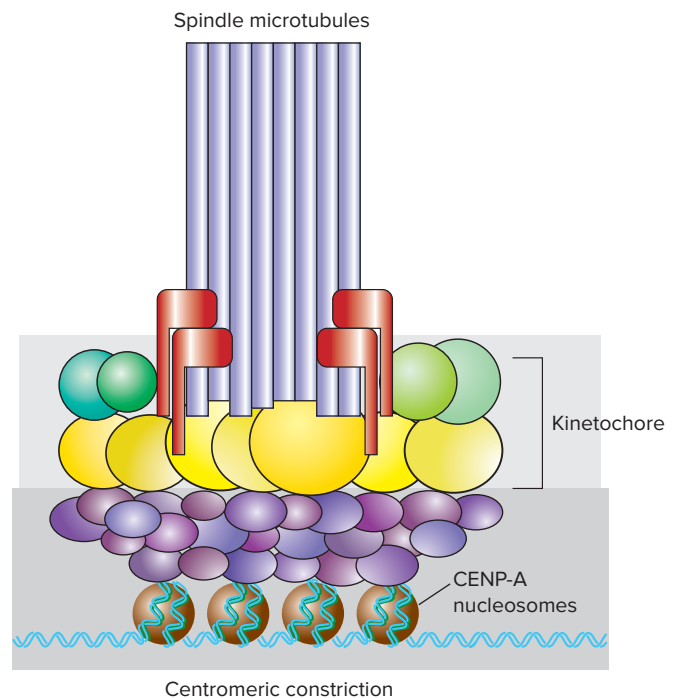
**Figure 12.23** Yeast centromeric DNA sequences. Each yeast centromere has two short, conserved DNA elements and binds through kinetochore proteins (not shown) to a single spindle microtubule (*tan*), whose size is shown for comparison.



elaboration of **kinetochores**: specialized structures composed of DNA and proteins that are the sites at which chromosomes attach to the spindle fibers. Each of the simple kinetochores in yeast cells connect only with a single spindle fiber (Fig. 12.23), but the kinetochores in higher eukaryotes attach to many spindle microtubules (Fig. 12.24). Researchers think that these complex kinetochores are likely to consist of repeating structural subunits, with each subunit responsible for attachment to one fiber.

The kinetochore-forming DNA, such as the  $\alpha$ -satellite in humans, has a different chromatin structure and different higher-order packaging than other chromosomal regions (Fig. 12.24). In this specialized chromatin, the normal histone H3 protein has been replaced by a histone variant called *CENP-A* in the nucleosome core. The *CENP-A* protein is very similar to histone H3 in its C-terminal region, but different from H3 in its N-terminal portion. Nucleosomes with this histone variant act as scaffolds to allow the assembly of many other proteins into kinetochores.

**Figure 12.24** Kinetochores. In higher eukaryotes, centromeric DNA consists of repeated sequences organized into nucleosomes containing *CENP-A*, a variant of histone H3. Kinetochores containing dozens of proteins organize around these nucleosomes. Some of these proteins govern kinetochore assembly (*purple*), some bind microtubules (*yellow*); some are motors that move the chromosomes along the spindle (*red*); and some (*green*) act in a checkpoint that makes sure sister chromatids (mitosis, meiosis II) or homologous chromosomes (meiosis I) do not separate before all the chromosomes are properly attached to spindle fibers.



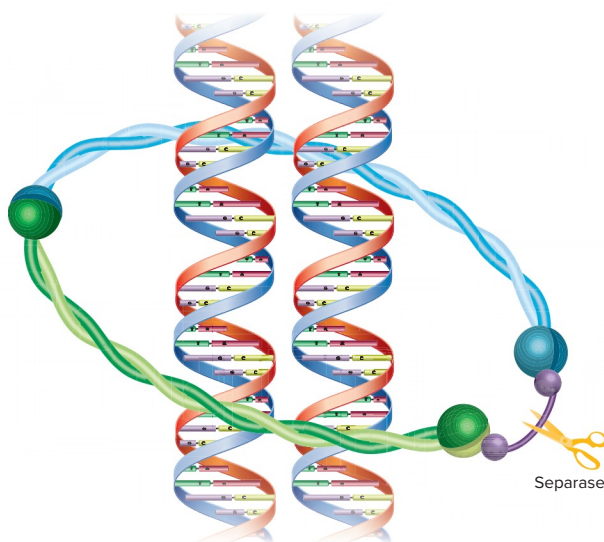
During mitosis, a kinetochore develops late in prophase on each sister chromatid, at the part of the centromere that faces one or the other cellular pole. By prometaphase, the kinetochores of the two sister chromatids attach to spindle fibers emanating from centrosomes at opposite poles of the cell. Some of the kinetochore proteins are motor proteins that exert pulling forces on chromatids toward the spindle pole to which they are attached.

At metaphase, tension arises when sister chromatids are pulled in opposite directions, yet are still held together at the centromere. Certain other kinetochore proteins monitor this tension to establish a cell cycle checkpoint: Only after all the kinetochores in the cell are under tension (meaning only after all chromosomes are properly attached to the spindle) do these proteins generate a molecular signal that allows the sister chromatids to disconnect from each other. The chromatids are then able to move to their respective poles during anaphase.

### Cohesin Complexes Hold Sister Chromatids Together

A highly conserved, multisubunit protein complex called **cohesin** acts as the glue that holds sister chromatids together during mitosis and meiosis until segregation takes place. After the chromosomes replicate in S phase of the cell cycle, cohesin proteins associate with and hold sister chromatids together along the arms and in the centromere region. Cohesin encircles the two double helices of the sister chromatids to keep them together. **Figure 12.25** shows

**Figure 12.25 A molecular model for cohesin.** The cohesin complex has protein subunits (*green, aqua, and purple*) that together surround the two sister chromatids (the two DNA molecules with strands in *blue and red*). Sister chromatids can separate when *separase (gold)* cleaves the *purple* cohesin subunit, freeing the two DNA molecules.



one current (though controversial) model for the topological relationship between cohesin and the two molecules of DNA. Cohesin rings are scattered along the length of the chromosome, but they are found in particularly high concentrations in the vicinity of centromeric heterochromatin.

During metaphase of mitosis, the cohesin complexes resist the forces pulling the chromatids poleward at their kinetochores, generating tension across the chromosomes, as just mentioned. At anaphase a proteolytic enzyme called *separase* cleaves the cohesin complexes, allowing the sister chromatids to separate and move to opposite spindle poles (**Fig. 12.25** and **Fig. 12.26a**). In support of this idea, dividing cells expressing a cohesin that cannot be cleaved by *separase* display many chromosome segregation errors.

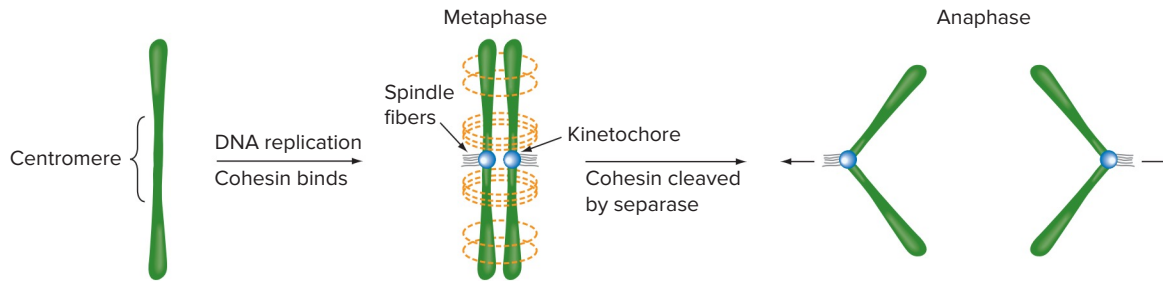
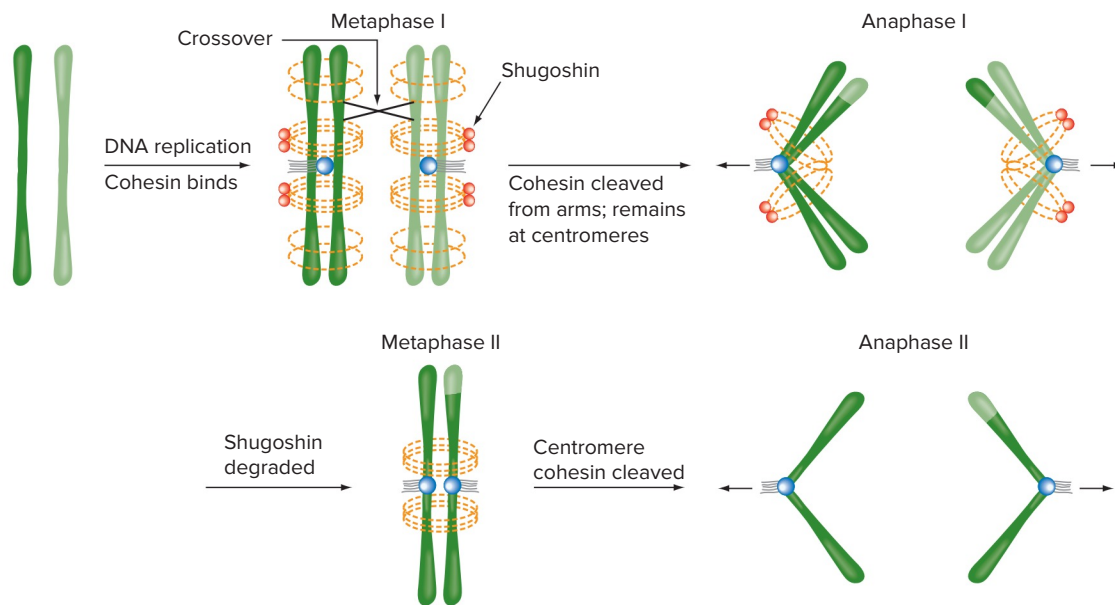
In meiosis, a special problem arises: Sister chromatids must stay together during the entire first meiotic division but then separate during the second meiotic division. Cells solve this problem by making cohesin complexes with a meiosis-specific subunit (replacing the purple subunit in **Fig. 12.25**) that can interact with a protein called *shugoshin* (meaning *guardian spirit* in Japanese; **Fig. 12.26b**). *Shugoshin* protects the cohesin at the centromere from being cleaved by *separase*. Upon entering meiosis II, *shugoshin* is removed, and *separase* can now cleave the centromeric cohesin at anaphase II, allowing sister chromatids to segregate to opposite poles.

Interestingly, *shugoshin* does not protect cohesin along the arms of the sister chromatids; the molecular mechanism that discriminates between cohesin on the arms from that at the centromere is not yet known. Thus, at anaphase of meiosis I, the cohesin along the arms of the sister chromatids is cleaved, while that at the centromere is not. This fact is very important because, as you will discover in Problem 30 at the end of this chapter, cohesin along the arms is the glue keeping homologous chromosomes together while the spindle tries to pull them apart during metaphase of the first meiotic division. Cleavage of the cohesin on the arms is therefore essential to allow the homologous chromosomes to segregate to opposite spindle poles during anaphase of meiosis I.

#### essential concepts

- Short DNA sequences define yeast *centromeres*. In higher eukaryotes, centromeres are much more complex and contain repetitive DNA sequences.
- *Kinetochores* at the centromeres are sites where spindle fibers attach to the chromosomes. A cell cycle checkpoint ensures that chromatids do not separate until all kinetochores are properly connected to the spindle.
- *Cohesin* complexes hold chromatids together until the enzyme *separase* cleaves the cohesin at anaphase. In meiosis I, *shugoshin* protects cohesin at the centromere and thus keeps sister chromatids together; in meiosis II, *shugoshin* is removed, so *separase* can cleave cohesin at anaphase of this division.

**Figure 12.26 Cohesin action in mitosis and meiosis.** (a) During mitosis, cohesin holds sister chromatids together through metaphase. Cleavage of cohesin by separase releases sister chromatids so they can segregate at anaphase. (b) In anaphase of meiosis I, cohesin is cleaved from the chromatid arms but is protected at the centromeres by shugoshin. Cohesin remains at the centromere to hold sister chromatids together until anaphase II.

**(a) Mitosis****(b) Meiosis**

## 12.6 Artificial Chromosomes

### learning objectives

1. List the elements that must be included in an artificial chromosome.
2. Discuss the reasons why scientists create artificial chromosomes.

In the 1980s, molecular geneticists constructed the first artificial eukaryotic chromosome, a **yeast artificial chromosome (YAC)**, by combining in a single DNA molecule yeast (*Saccharomyces cerevisiae*) versions of the three key chromosomal elements described in this chapter—centromeres, telomeres, and origins of replication

(Fig. 12.27). In 2014, a team of scientists at Johns Hopkins University reported the construction of a completely synthetic chromosome of this same organism. In this section, we discuss the distinction between YACs and synthetic chromosomes and discuss the uses of these two kinds of DNA molecules.

### Yeast Artificial Chromosome (YACs) Help Characterize DNA Elements Required for Chromosome Replication and Transmission

In Chapter 9, we described the strategy of whole-genome shotgun sequencing through paired-end sequencing of cloned genome fragments. When using this method, the genome is cut into large pieces up to 2 Mb long that are

**Figure 12.27 Yeast artificial chromosomes (YACs).** To function effectively as artificial chromosomes, YAC vectors containing an ARS (yeast origin of replication), a centromere sequence, and telomeres at the ends need to be ligated (here through *EcoRI* sites) with inserts of foreign DNA more than 100 kb long (not drawn to scale). Yeast cells transformed with YACs can be identified if the vector has a selectable marker such as the *TRP<sup>+</sup>* gene.



subsequently isolated, amplified, and characterized. Two types of vectors that allow such long DNA fragments to be cloned are BACs (bacterial artificial chromosomes) and YACs (yeast artificial chromosomes). We describe YACs in some detail here not only because they are useful for the cloning of large fragments of genomic DNA, but also because manipulation of these vector molecules has produced many insights into chromosome function.

Figure 12.27 illustrates how centromeres, telomeres, and origins of replication are combined to make a YAC. Once scientists join these elements using recombinant DNA technology, they transform the YAC into yeast cells, where it can be maintained as an autonomous chromosome. Researchers track the presence of the YAC by including a selectable marker, for example the *TRP<sup>+</sup>* gene that allows an auxotrophic host cell to grow in the absence of the amino acid typtophan in the growth medium.

Manipulation of the YAC construction process was historically significant because it allowed geneticists to isolate and analyze the chromosomal regions corresponding to ARSs (yeast origins of replication) and centromeres. Plasmids containing only origins of replication but no centromere or telomere replicate but do not segregate properly. Plasmids with origins of replication and a centromere but no telomeres replicate and segregate fairly well if they are circular; if they are linear, they degrade and eventually become lost from the cell.

Interestingly, small DNA molecules carrying all three elements replicate and segregate as linear chromosomes independent of the normal yeast chromosomes, but they do not segregate accurately because they do not have enough DNA. YACs carrying inserts of 11,000 bp of random DNA mis-segregate in 50% of cell divisions. YACs containing 55,000 bp show segregation errors in 1.5% of cell divisions. With artificial chromosomes containing more than 100,000 bp inserts, the rate of segregation error falls to 0.3%. This error rate is still 200 times greater than that seen with natural yeast chromosomes of normal size, indicating that some subtle aspects of chromosome structure and function remain to be discovered.

Researchers have used the lessons from these YAC studies to develop human artificial chromosomes. They

hope to use such chromosomes eventually to treat genetic diseases, the idea being they can serve as vectors for the transformation of human cells defective for a particular gene with a wild-type copy of that gene.

### Synthetic Chromosomes May Help Define Minimal Genomes

Chromosomes undoubtedly contain many DNA sequences that are not required for the survival of the organism. These include transposable elements, repetitive DNAs, some of the DNA between genes, and many intronic sequences. In addition, only about 1000 of the approximately 6000 yeast genes have been found to be essential individually for yeast cell growth under laboratory conditions. In other words, deletion of any one of the remaining 5000 genes allows yeast to survive. But perhaps many of these 5000 nonessential genes are actually essential under specific growth conditions found in nature, or maybe a given gene becomes essential when a different gene is removed simultaneously. What then is the minimal set of DNA sequences that yeast cells actually need for their survival?

To answer this question, molecular biologists aim to create yeast cells with an entirely synthetic set of 16 chromosomes. In this approach, a **synthetic chromosome** differs from an artificial chromosome like a YAC in two ways. First, all the DNA in a synthetic chromosome is entirely man-made in DNA synthesizer machines; in contrast, a YAC is constructed by splicing together DNA elements that were originally in yeast chromosomes. Second, a synthetic chromosome would include (before any subsequent manipulation) all genes present on the corresponding yeast chromosome. In contrast, YAC vectors usually have a single protein-coding yeast gene—a selectable marker like *TRP<sup>+</sup>* in Fig. 12.27.

Remarkably, in 2014, scientists reported the construction of Syn III, a synthetic version of yeast chromosome III, and the first synthetic eukaryotic chromosome. Even though Syn III lacks about 50,000 bp (mostly transposable elements and other intergenic sequences) of the normal (~317,000 bp) third chromosome, yeast cells containing Syn III instead of a normal third chromosome display

normal appearance and growth. As you will see in Problem 37, Syn III also contains special short sequences that will in the future allow the researchers to make deletions of various regions of the chromosome to see which genes or gene combinations are essential or nonessential.

Anticipating that viable yeast with 16 minimized synthetic chromosomes can be obtained, these scientists are now busy creating similar synthetic versions of the other 15 yeast chromosomes. If the experiment is successful, not only will these researchers have created an entire genome by chemical synthesis, but the yeast they create will serve as the basis for future investigations aimed at determining the minimal yeast genome that can support life.

### essential concepts

- Stable transmission of linear *yeast artificial chromosomes* (YACs) requires the inclusion of an origin of replication (ARS), a centromere, and telomeres; for unknown reasons, stably transmitted chromosomes must contain DNA in excess of 100 kb.
- *Artificial chromosomes* have been used as cloning vectors and in studies to identify the functional regions of natural chromosomes; they might be used in the future to treat genetic diseases in humans.
- A *synthetic chromosome* lacking most intergenic sequences functions normally. Such synthetic chromosomes may help to define a minimal yeast genome.

### WHAT'S NEXT

Eukaryotic chromosomes manage the genetic information in DNA through a modular chromatin design whose flexibility allows shifts between different levels of organization and compaction. These reversible changes in chromatin structure reliably sustain a variety of chromosome functions, including packaging in nuclei, proper copying and segregation during cell division, and coordinating gene expression.

Although the faithful function, replication, and transmission of chromosomes underlie the perpetuation of life within each species, chromosomal changes do occur. We

have already described two mechanisms of change: mutation of individual nucleotides (Chapter 7) and homologous recombination, which exchanges bases between homologs (Chapters 4, 5, and 6). In Chapter 13, we examine broader chromosomal rearrangements that produce different numbers of chromosomes, reshuffle genes between nonhomologous chromosomes, and reorganize the genes of a single chromosome. These large-scale modifications provide some of the important variations that fuel evolution.

### SOLVED PROBLEMS

- I. Scientists can construct YACs that range in size from 15 kb to 1 Mb. Based on DNA length, what level of chromosome compaction would you predict for a YAC of 50 kb compared with a YAC of 500 kb?

#### Answer

To answer this question, you need to apply information about the amount of DNA needed to achieve different levels of chromosome condensation.

The 500 kb YAC would probably be more condensed than the 50 kb YAC based on its larger size. The DNA of both YACs would be wound around histones to form the nucleosome structure (160 bp around the core histones plus 40 bp in linker region). That DNA would be further compacted into 300 Å fibers that contain six nucleosomes per turn. The 500 kb YAC would be compacted at a higher level of order, presumably in radial loops that occur every 60–100 kb in the chromosome, but the 50 kb YAC is not large enough to be packaged in this way.

- II. A protein called CBF1 was identified in yeast as a centromere-binding protein. The CBF1 protein is essential for proper chromosome segregation during cell division in yeast. You have identified a gene from the human genome that encodes a protein with some similarity in amino acid sequence with the yeast CBF1 protein.

- How could you establish whether the protein encoded by the human gene is associated with human centromere regions? (Assume that you can make an antibody that can bind specifically to this protein.) Why would your test not be a FISH experiment?
- Describe two methods to test if this human protein might actually participate in centromere function (as opposed to just being present at centromeres). For the first method, assume that you can easily produce mutations in any given gene in a human tissue culture cell or even in a whole mouse.



For the second method, you should use two different recombinant YACs. YAC-1 contains the yeast *CBF1*<sup>+</sup> gene and the yeast *URA3*<sup>+</sup> gene that allows *URA3*<sup>-</sup> yeast to grow in the absence of uracil. YAC-2 contains the wild type human *CBF1*-related gene and a *TRP*<sup>+</sup> gene that allows *TRP*<sup>-</sup> yeast cells to grow in the absence of the amino acid tryptophan. You will also want to use 5-FOA, a chemical closely related to the *URA3* enzyme's normal substrate. Yeast that make the *URA3* protein cannot grow in the presence of 5-FOA because the enzyme will convert 5-FOA to a lethal toxin.

### Answer

This question requires an understanding of the structure and function of centromeres in ensuring proper chromosome segregation during cell division.

- a. You would require a molecular probe that could specifically bind to the human protein related to yeast *CBF1*. An antibody that reacts with this human protein and that is labeled with a fluorescent tag would act as such a probe. You would purify large amounts of the human *CBF1*-related protein (Chapter 16 discusses how this can be easily accomplished using recombinant DNA technology), and then inject this protein into a rabbit or other experimental animal. The rabbit would develop antibodies against the protein (analogous to the way humans make antibodies against viral proteins in vaccines). You could obtain these antibodies from the rabbit's blood and label them with a fluorescent tag. Finally, you would place the tagged antibodies on a microscope slide on which human chromosomes are displayed. If the human *CBF1*-related protein is associated

with human centromeres, you would see a fluorescent pattern similar to that seen in Fig. 12.2, which is the result of exactly this kind of experiment using antibodies against a different centromeric protein. This is not a FISH experiment because the probe is not a nucleic acid that hybridizes specifically to DNA sequences on a chromosome, but is instead an antibody that binds to a protein possibly associated with centromeric DNA sequences.

- b. Two general lines of experimentation would allow you to ask if this human protein is involved in centromere function.

First, you could try to eliminate or reduce the function of the gene encoding this protein in human cells or cells of a related mammal like mice, and then determine if, when they divide, these cells mis-segregate their chromosomes at a high frequency. Chapter 18 discusses various ways to disrupt mammalian genes or interfere with their expression.

Second, you could determine if expression of the human protein can replace the function of the yeast *CBF1* gene. You would transform triple mutant yeast cells (*CBF1*<sup>-</sup> *URA3*<sup>-</sup> *TRP*<sup>-</sup>) with YAC-1 (yeast *CBF1*<sup>+</sup> and *URA3*<sup>+</sup>). The transformed yeast cells should be viable as long as they are grown in the presence of tryptophan. Next, you want to see if you can replace YAC-1 with YAC-2 (human *CBF1*<sup>+</sup> and *TRP*<sup>+</sup>). You will select cells that are transformed with YAC-2 and also have lost YAC-1 by growing the yeast in the absence of tryptophan (selects for the *TRP*<sup>+</sup> gene in YAC-2) and in the presence of 5-FOA (selects against the *URA3*<sup>+</sup> gene in YAC-1). Yeast cells containing YAC-2 instead of YAC-1 will be obtained only if human *CBF1* protein can substitute for the yeast *CBF1*<sup>+</sup> gene.



## PROBLEMS

### Vocabulary

1. For each of the terms in the left column, choose the best matching phrase in the right column.

- |                |   |
|----------------|---|
| a. telomere    | 1. protein complex that keeps sister chromatids together until anaphase |
| b. G bands     | 2. origin of replication in yeast                                       |
| c. kinetochore | 3. repetitive DNA found near the centromere in higher eukaryotes        |
| d. nucleosome  | 4. specialized structure at the end of a linear chromosome              |
| e. ARS         | 5. complexes of DNA, protein, and RNA in the eukaryotic nucleus         |

- |                  |  |
|------------------|--|
| f. satellite DNA | 6. small basic proteins that bind to DNA and form the core of the nucleosome       |
| g. chromatin     | 7. complex of DNA and proteins where spindle fibers attach to a chromosome         |
| h. cohesin       | 8. beadlike structure consisting of DNA wound around histone proteins              |
| i. histones      | 9. protein complex that protects telomeres from degradation and end-to-end fusions |
| j. shelterin     | 10. regions of a chromosome that are distinguished by staining differences         |