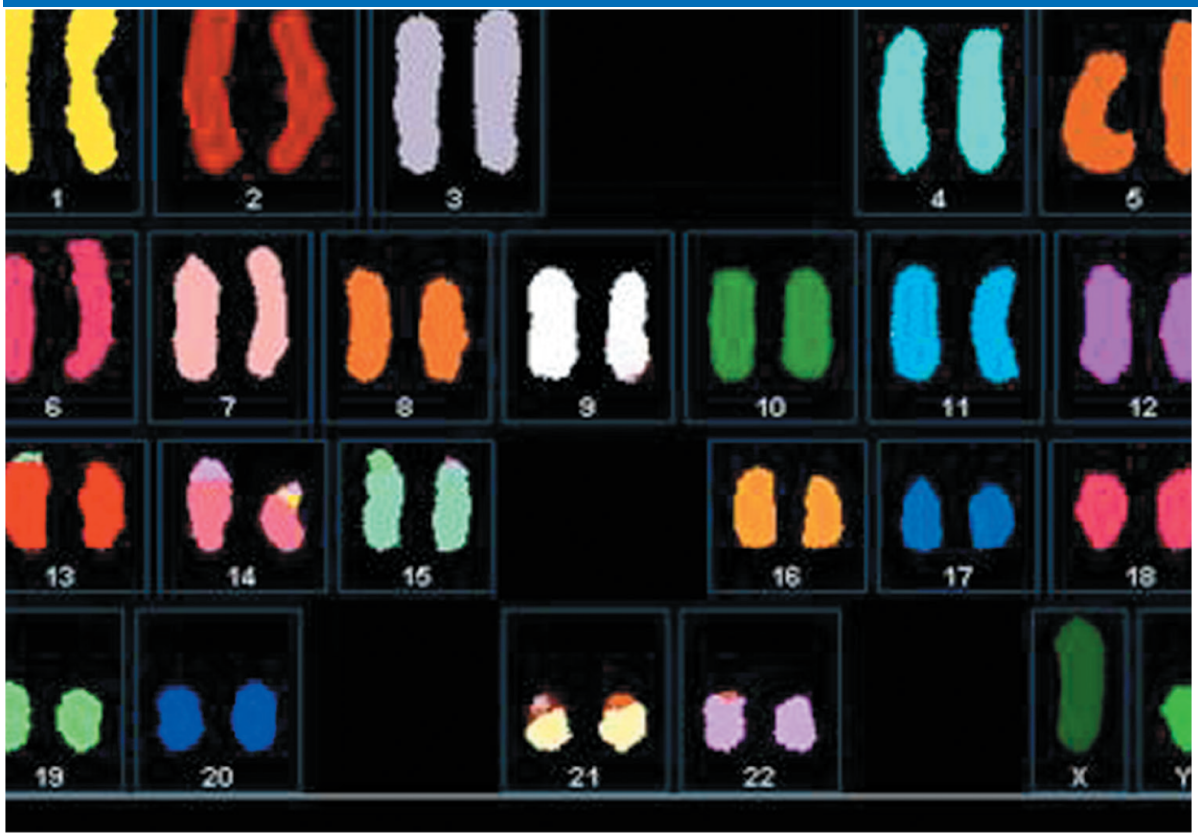
CHAPTER 14

Genomes and Genomics



Nallasivam Palanisamy, MSc., MPhil., PhD., Associate Professor of Pathology, Michigan Center for Translational Pathology, University of Michigan.

The human nuclear genome viewed as a set of labeled DNA. The DNA of each chromosome has been labeled with a dye that emits fluorescence at one specific wavelength (producing a specific color).

CHAPTER OUTLINE AND LEARNING OBJECTIVES

LO 14.4 Compare genomic methods used to identify mutations that have been associated with human disease thus far.

14.7 FUNCTIONAL GENOMICS AND REVERSE GENETICS

LO 14.5 Outline reverse genetic approaches to analyze the function of genes and genetic elements identified by genome sequencing and comparative genomics.

CHAPTER OBJECTIVE

In this chapter, we will see that the ability to sequence whole genomes has revolutionized the field of genetics. Our broad objective is to learn how a combination of experimental and computational methods are used to sequence genomes and to identify functional elements within those genomes.

In the summer of 2009, Dr. Alan Mayer, a pediatrician at Children's Hospital of Wisconsin in Milwaukee, wrote to a colleague about the heartbreaking and baffling case of a four-year-old patient of his (Figure 14-1). For two years, little Nicholas Volker had endured over 100 trips to the operating room as doctors tried to manage a mysterious disease that was destroying his intestines, leaving him vulnerable to dangerous infections, severely underweight, and often unable to eat.

Nicholas Volker

Gary Porter/Tribune News Service/WAUWATOSA/ WI/USA/ Newscom.

FIGURE 14-1 DNA sequencing of all the exons of Nicholas Volker's genome revealed a single mutation responsible for his debilitating, but previously unidentified, disease.

Neither Mayer nor any other doctors had ever seen a disease like Nicholas's; they were unable to diagnose it, or to stem its ravages by any medical, surgical, or nutritional treatment. It was difficult to treat a disease that no one could identify. So, Dr. Mayer asked his colleague, Dr. Howard Jacob at the Medical College of Wisconsin, "if there is some way we can get his genome sequenced. There is a good chance Nicholas has a genetic defect, and it is likely to be a new disease. Furthermore, a diagnosis soon could save his life and truly showcase personalized genomic medicine."[1]

Dr. Jacob knew that it would be a longshot. Finding a single mutation responsible for a disease would require sifting through thousands of variations in Nicholas's DNA. One key decision was to narrow the search to just the exon sequences in Nicholas's DNA. The rationale was that if the causal mutation was a protein-coding change, then it could be identified by sequencing all of the exons, or Nicholas's *exome*, which comprise a little over 1 percent of the entire human genome. Still, it would be an expensive search—the sequencing would cost about $75,000 with the technology available at the time. Nevertheless, the money was raised from donors, and Jacob and a team of collaborators undertook the task.

As Jacob expected, they found more than 16,000 possible candidate variations in Nicholas's DNA. They narrowed this long list by focusing on those mutations that had not been previously identified in humans, and that caused amino acid replacements that were not found in other species. Eventually, they identified a single base substitution in a gene called the *X*-linked *i*nhibitor of *ap*optosis (*XIAP*) that changed one amino acid at position 203 of the protein—an amino acid that was invariant among mammals, fish, and even the fruit-fly counterparts of the *XIAP* gene.

Fortunately, the identification of Nicholas's *XIAP* mutation suggested a therapeutic approach. The *XIAP* gene was previously known to have a role in the inflammatory response, and mutations in the gene were associated with a very rare but potentially fatal immune disorder (although not Nicholas's intestinal symptoms). Based on that knowledge, Nicholas's doctors boosted his immune system with an infusion of umbilical-cord blood from a well-matched donor. Over the next several months, Nicholas's health improved to the point where he was able to eat steak and other foods. And over the next two years, Nicholas did not require any further intestinal surgeries.

The diagnosis and treatment of Nicholas Volker illustrate the dramatic advances in the technology and impact of **genomics**—the study of genomes in their entirety. The long-awaited promise that

genomics would shape clinical medicine is now very much a reality. The technological and biological progress from what started as a trickle of data in the 1990s has been astounding. In 1995, the 1.8-Mb (1.8-megabase) genome of the bacterium *Haemophilus influenzae* was the first genome of a free-living organism to be sequenced. In 1996 came the 12-Mb genome of *Saccharomyces cerevisiae;* in 1998, the 100-Mb genome of *Caenorhabditis elegans;* in 2000, the 180-Mb genome of *Drosophila melanogaster;* in 2001, the first draft of the 3000-Mb human genome; and, in 2005, the first draft of our closest living relative, the chimpanzee. These species are just a small sample. By the end of 2017, over 130,000 bacterial genomes, and nearly 5500 eukaryotic genomes (including protists, fungi, plants, and animals) had been sequenced. At the beginning of 2018, the Earth BioGenome Project announced its bold intention to sequence all of the approximately 1.5 million known species of eukaryotes in the next 10 years.

It is no hyperbole to say that genomics has revolutionized how genetic analysis is performed and has opened avenues of inquiry that were not conceivable just a few years ago. Most of the genetic analyses that we have so far considered employ a **forward genetics** approach to analyzing genetic and biological processes. That is, the analysis begins by first screening for mutants that affect some observable phenotype, and the characterization of these mutants eventually leads to the identification of the gene and the function of DNA, RNA, and protein sequences. In contrast, having the entire DNA sequences of an organism's genome allows geneticists to work in both directions—forward from phenotype to gene, and in reverse from gene to phenotype (**Figure 14-2**). Without exception, genome sequences reveal many genes that were not detected from classical mutational analysis. Using so-called **reverse genetics**, geneticists can now systematically study the roles of such formerly unidentified genes. Moreover, a lack of prior classical genetic study is no longer an impediment to the genetic investigation of organisms. The frontiers of experimental analysis are growing far beyond the bounds of the very modest number of long-explored model organisms (for more, see the *Beyond Model Organisms* section of *A Brief Guide to Model Organisms*, at the back of this book).

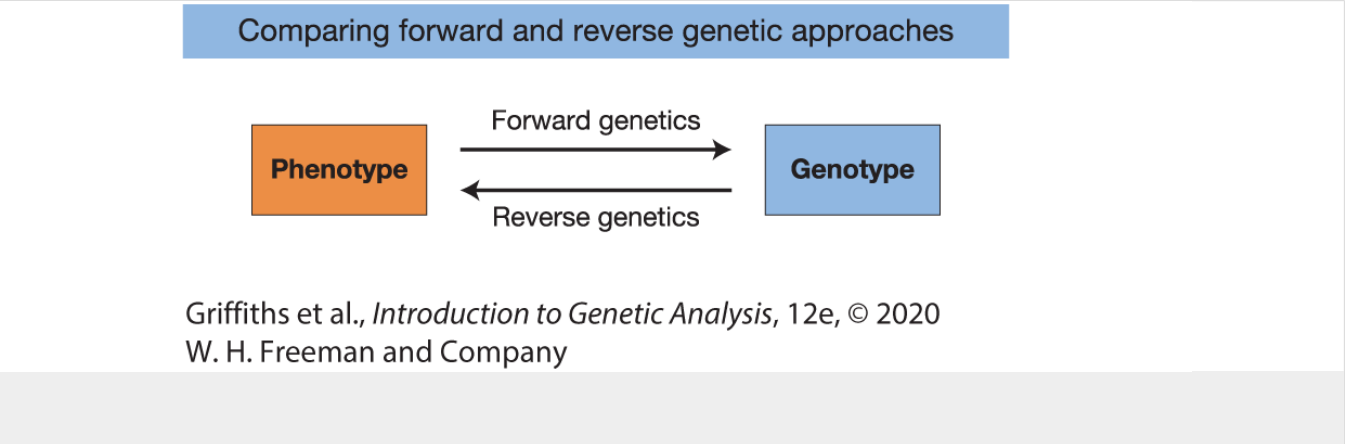


FIGURE 14-2 Forward genetics is phenotype driven, and asks *what genes underlie a particular phenotype*, while reverse genetics is genotype driven and asks *what phenotypes are associated with a particular gene*.

Analyses of whole genomes now contribute to every corner of biological research. In human genetics, genomics is providing new ways to locate genes that contribute to many genetic diseases, like Nicholas's, which had previously eluded investigators. The day is soon approaching when a person's genome sequence is a standard part of his or her medical record. The availability of genome sequences for long-studied model organisms and their relatives has dramatically accelerated gene identification, the analysis of gene function, and the characterization of noncoding elements of the genome. New technologies for the global, genome-wide analysis of the physiological role of all gene products are driving the development of the new field called *systems biology*. From an evolutionary perspective, genomics provides a detailed view of how genomes and organisms have diverged and adapted over geological time.

The DNA sequence of the genome is the starting point for a whole new set of analyses aimed at understanding the structure, function, and evolution of the genome and its components. In this chapter, we will focus on three major aspects of genomic analysis:

- *Bioinformatics*, the analysis of the information content of entire genomes. This information includes the numbers and types of genes and gene products as well as the location, number, and types of binding sites on DNA and RNA that allow functional products to be produced at the correct time and place.
- *Comparative genomics*, which considers the genomes of closely and distantly related species for evolutionary insight.
- *Functional genomics*, the use of an expanding variety of methods, including reverse genetics, to understand gene and protein function in biological processes.

14.1 THE GENOMICS REVOLUTION

After the development of recombinant DNA technology in the 1970s, research laboratories typically undertook the cloning and sequencing of one gene at a time (see Chapter 10), and then only after having had first found out something interesting about that gene from a classic mutational analysis. The steps in proceeding from the classical genetic map of a locus to isolating the DNA encoding a gene (*cloning*) to determining its sequence were often numerous and time consuming. In the 1980s, some scientists realized that a large team of researchers making a concerted effort could clone and sequence the *entire* genome of a selected organism. Such **genome projects** would then make the clones and the sequence publicly available resources. One appeal of having these resources available is that, when researchers become interested in a gene of a species whose genome has been sequenced, they need only find out where that gene is located on the map of the genome to be able to zero in on its sequence and potentially its function. By this means, a gene could be characterized much more rapidly than by cloning and sequencing it from scratch, a project that at the time could take several years to carry out. This quicker approach is now a reality for all model organisms.

Similarly, the Human Genome Project aimed to revolutionize the field of human genetics. The availability of human genome sequences, and the ability to sequence the genomes of patients and their relatives, has greatly aided the identification of disease-causing genes. Furthermore, the ability to determine gene sequences in normal and diseased tissues (for example, cancers) has been a great catalyst to the understanding of disease processes, and pointed the way to new therapies.

From a broader perspective, the genome projects had the appeal that they could provide some glimmer of the principles on which genomes are built. The human genome contains 3 billion base pairs of DNA. Having the entire sequence raised questions such as: How many genes does it contain? How are they distributed, and why? What fraction of the genome is coding sequence? What fraction is regulatory sequence? How is our genome similar to or different from other animals? Although we might convince ourselves that we understand a single gene of interest, the major challenge of genomics today is genomic literacy: How do we read the storehouse of information enciphered in the sequence of complete genomes?

The basic techniques needed for sequencing entire genomes were already available in the 1980s. But the scale that was needed to sequence a complex genome was, as an engineering project, far

beyond the capacity of the research community then. Genomics in the late 1980s and the 1990s evolved out of large research centers that could integrate these elemental technologies into an industrial-level production line. These centers developed robotics and automation to carry out the many thousands of cloning steps and millions of sequencing reactions necessary to assemble the sequence of a complex organism. Just as important, advances in information technology aided the analysis of the resulting data.

The first successes in genome sequencing set off waves of innovation that led to faster and much less expensive sequencing technologies. Now, individual machines can produce as much sequence in a day as centers used to accomplish in months. New technologies can now obtain more than 1×10^{12} bases of sequence in a working day on a single instrument. This represents an approximately *1 million-fold* increase in throughput over earlier instruments used to obtain the first human genome sequences (**Figure 14-3**).

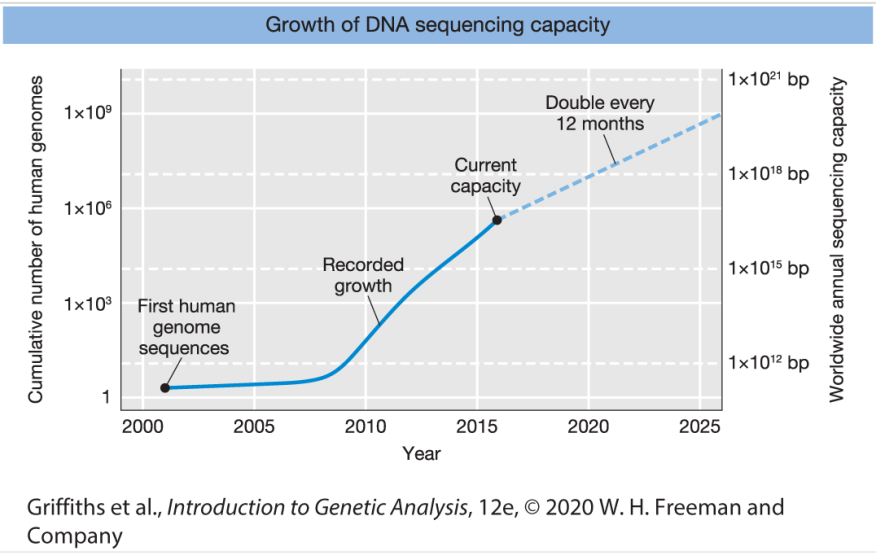


FIGURE 14-3 This plot shows the increase in DNA sequencing capacity from the publication of the first human genome sequences in 2001 until the end of 2015 (solid line), along with the projected growth if capacity doubles every year (dashed line). The total number of human genomes sequenced is indicated on the y-axis on the left, and the total worldwide annual sequencing capacity is indicated on the y-axis on the right. [*Data from Z. D. Stephens et al., "Big Data: Astronomical or Genomical?"* PLoS Biology, *13(7), 2015, e1002195, https://doi.org/10.1371/journal.pbio.1002195.*]

A Genomics, aided by the explosive growth in information technology, has encouraged researchers to develop ways of experimenting on the genome as a whole rather than simply one gene at a time. Genomics has also demonstrated the value of collecting large-scale data sets in advance so that they can be used later to address specific research problems. In the last sections of this

chapter, we will explore some ways that genomics now drives basic and applied genetics research. In subsequent chapters, we will see how genomics is catalyzing advances in understanding the dynamics of mutation, recombination, and evolution.

KEY CONCEPT Characterizing whole genomes is fundamental to understanding the entire body of genetic information underlying the physiology, development, and evolution of living organisms, and to the discovery of new genes such as those having roles in human genetic disease.

14.2 OBTAINING THE SEQUENCE OF A GENOME

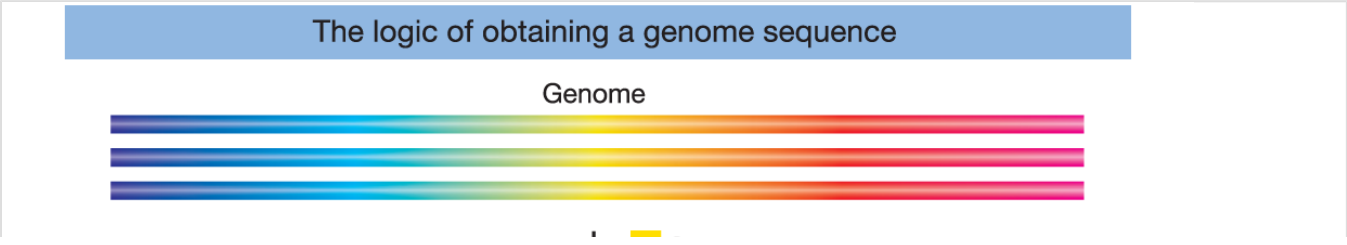
LO 14.1 Describe the combinations of strategies typically necessary for obtaining and assembling the complete DNA sequences of organisms.

When people encounter new territory, one of their first activities is to create a map. This practice has been true for explorers, geographers, oceanographers, and astronomers, and it is equally true for geneticists. Geneticists use many kinds of maps to explore the terrain of a genome. Examples are linkage maps based on inheritance patterns of gene alleles and cytogenetic maps based on the location of microscopically visible features such as rearrangement break points (see Chapters 4 and 17).

The highest-resolution map is the complete DNA sequence of the genome—that is, the complete sequence of nucleotides A, T, C, and G of each double helix in the genome. Because obtaining the complete sequence of a genome is such a massive undertaking of a sort not seen before in biology, new strategies must be used, all based on automation.

Turning sequence reads into an assembled sequence

You've probably seen a magic act in which the magician cuts up a newspaper page into a great many pieces, mixes it in his hat, says a few magic words, and *voila!* an intact newspaper page reappears. Basically, that's how genomic sequences are obtained. The approach is to (1) break the DNA molecules of a genome up into thousands to millions of more or less random, overlapping small segments; (2) read the sequence of each small segment; (3) computationally find the overlap among the small segments where their sequences are identical; and (4) continue overlapping ever larger pieces until all the small segments are linked (**Figure 14-4**). At that point, the sequence of a genome is assembled.



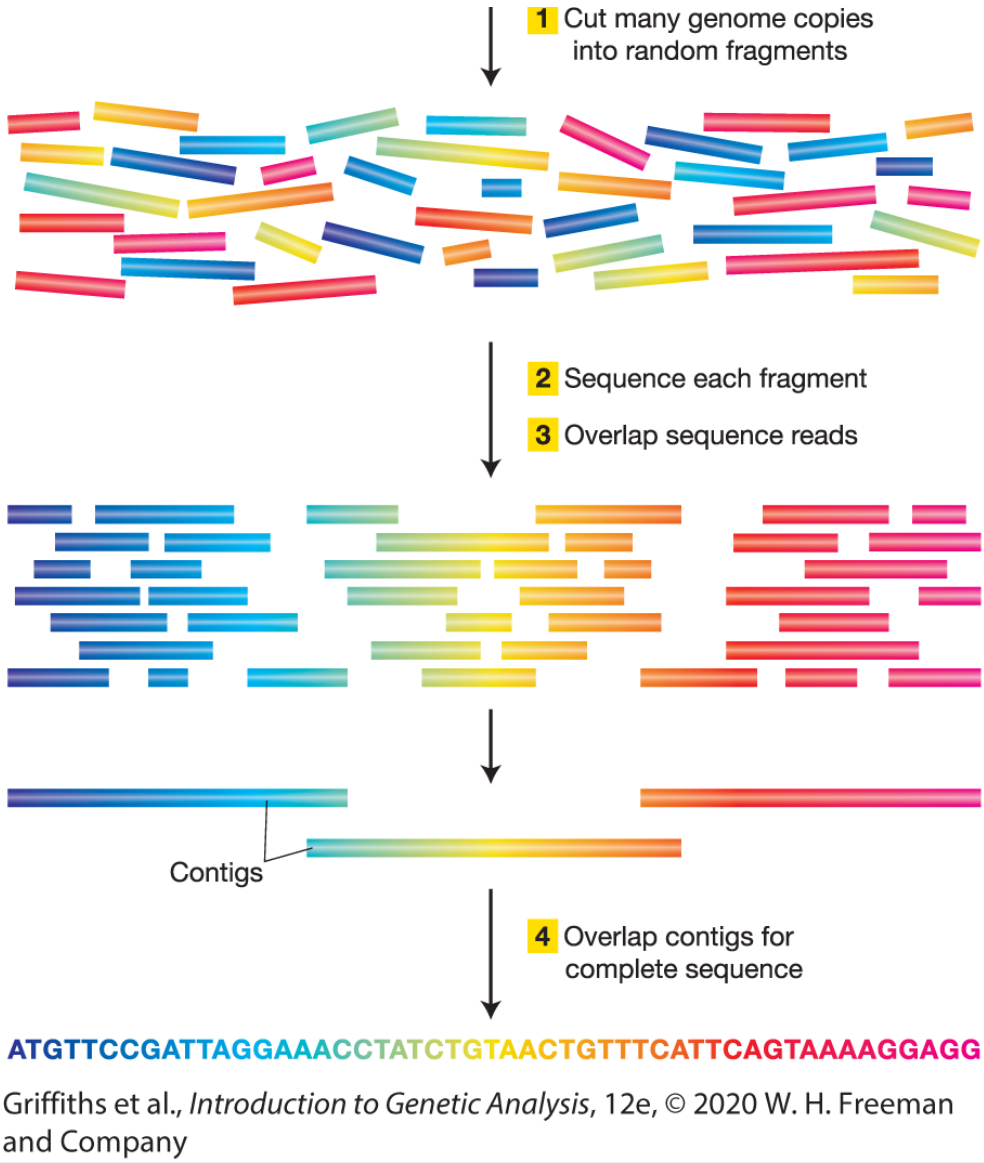


FIGURE 14-4 To obtain a genome sequence, multiple copies of the genome are cut into small pieces that are sequenced. The resulting sequence reads are overlapped by matching identical sequences in different fragments until a consensus sequence of each DNA double helix in the genome is produced.

Why does this process require automation? To understand why, let's consider the human genome, which contains about 3×10^9 bp of DNA, or 3 billion base pairs (3 gigabase pairs = 3 Gbp). Suppose we could purify the DNA intact from each of the 24 human chromosomes (the 22 autosomes, plus the X and the Y sex chromosomes), separately put each of these 24 DNA samples into a sequencing machine, and read their sequences directly from one telomere to the other. Obtaining a complete sequence would be utterly straightforward, like reading a book with 24 chapters—albeit a very, very long book with 3 billion characters (about the length of 3000 novels). Unfortunately, such a sequencing machine does not yet exist.

Rather, automated sequencing is the current state of the art in DNA sequencing technology. Initially based on the pioneering dideoxy chain-termination sequencing method developed by Fred

Sanger (discussed in Chapter 10; see Figure 10-18), automated sequencing now employs a variety of chemistries and optical-detection methods. The methods now available vary in the length of DNA sequence obtained, the bases determined per second, and raw accuracy. For large-scale sequencing projects that seek to analyze large individual genomes or the genomes of many different individuals or species, choosing a method requires balancing speed, cost, and accuracy.

Individual sequencing reactions (called *sequencing reads*) provide letter strings that, depending on the sequencing technique employed, range on average from about 100 to 15,000 bases long. Such lengths are tiny compared with the DNA of a single chromosome. For example, an individual read of 300 bases is only 0.0001 percent of the longest human chromosome (about 3×10^8 bp of DNA) and only about 0.00001 percent of the entire human genome. Thus, one major challenge facing a genome project is **sequence assembly**—that is, building up all of the individual reads into a **consensus sequence**, a sequence for which there is consensus (or agreement) that it is an authentic representation of the sequence for each of the DNA molecules in that genome.

Let's look at these numbers in a somewhat different way to understand the scale of the problem. As with any experimental observation, automated sequencing machines do not always give perfectly accurate sequence reads. Indeed, newer, higher-throughput sequencing technologies generate a *greater* frequency of errors than older methods; the error rate may range from less than 1 percent to as much as 15 percent, depending upon the technology. Thus, to ensure accuracy, genome projects conventionally obtain many independent sequence reads of each base pair in a genome. Many-fold coverage ensures that chance errors in the reads do not give a false reconstruction of the consensus sequence.

Given a sequence read of about 100 bases of DNA and a human genome of 3 billion base pairs, 300 million independent reads are required to give 10-fold average coverage of each base pair. However, not all sequences are represented equally, and so the number of reads required is even larger. Typically, 30-fold average coverage is desired when sequencing a genome. The amount of information to be tracked is enormous. Thus, genome sequencing has required many advances in automation and information technology.

What are the goals of sequencing a genome? First, we strive to produce a consensus sequence that is a true and accurate representation of the genome, starting with one individual organism or standard strain from which the DNA was obtained. This sequence will then serve as a reference sequence for the species. We now know that there are many differences in DNA sequence between different individuals within a species and even between the maternally and paternally

contributed genomes within a single diploid individual. Thus, no single genome sequence truly represents the genome of the entire species. Nonetheless, the genome sequence serves as a standard or reference with which other sequences can be compared, and it can be analyzed to determine the information encoded within the DNA, such as the inventory of encoded RNAs and polypeptides.

Like written manuscripts, genome sequences can range from *draft* quality (the general outline is there, but there are typographical errors, grammatical errors, gaps, sections that need rearranging, and so forth), to *finished* quality (a very low rate of typographical errors, some missing sections but everything that is currently possible has been done to fill in these sections), to truly *complete* (no typographical errors, every base pair absolutely correct from telomere to telomere). Although complete assemblies have been obtained for organisms with small genomes, such as bacteria and yeast, this is currently not possible for large and complex eukaryotic genomes, including human. In the following sections, we will examine the strategy and some methods for producing draft and finished genome-sequence assemblies. We will also encounter some of the features of genomes that challenge genome-sequencing projects.

Whole-genome sequencing

The current general strategy for obtaining and assembling the sequence of a genome is called **whole-genome shotgun (WGS) sequencing**. This approach is based on determining the sequence of many segments of genomic DNA that have been generated by breaking the long chromosomes of DNA into many short segments. Two approaches to WGS sequencing are responsible for most genome sequences obtained to date. The fundamental differences between them are in how the short segments of DNA are obtained and prepared for sequencing and the sequencing chemistry employed. The first method, used to sequence the first human genome, relied on the cloning of DNA in microbial cells and employed the dideoxy sequencing technique. We will refer to this approach as “traditional WGS sequencing.” Methods in the second group are generally cell-free methods that employ new techniques for sequencing and are designed for very high throughput (referring to the number of reads per machine per unit time). We will refer to this group of methods as “next-generation WGS sequencing.”

Traditional WGS sequencing

The traditional WGS approach begins with the construction of genomic libraries, which are collections of these short segments of DNA, representing the entire genome. The short DNA segments in such a library have been inserted into one of a number of types of *accessory* chromosomes (nonessential elements such as plasmids, modified bacterial viruses, or artificial chromosomes) and propagated in microbes, usually bacteria or yeast. These accessory chromosomes carrying DNA inserts are called vectors (see Chapter 10).

To generate a genomic library, a researcher first uses restriction enzymes, which cleave DNA at specific sequences, to cut up purified genomic DNA. Some enzymes cut the DNA at many places, whereas others cut it at fewer places; so the researcher can control whether the DNA is cut, on average, into longer or shorter pieces. The resulting fragments have short single strands of DNA at both ends. Each fragment is then joined to the DNA molecule of the accessory chromosome, which also has been cut with a restriction enzyme and which has ends that are complementary to those of the genomic fragments. In order for the entire genome to be represented, multiple copies of the genomic DNA are cut into fragments. By this means, thousands to millions of different fragment-vector recombinant molecules are generated.

As discussed in Chapter 10, the resulting pool of recombinant DNA molecules is then propagated, typically by introducing the molecules into bacterial cells. Each cell takes up one recombinant molecule. Then each recombinant molecule is replicated in the normal growth and division of its host so that many identical copies of the inserted fragment are produced for use in analyzing the fragment's DNA sequence. Because each recombinant molecule is amplified from an individual cell, each cell is a distinct *clone*. The resulting library of clones is called a *shotgun library* because sequence reads are obtained from clones randomly selected from the whole-genome library without any information on where these clones map in the genome.

Next, the genome fragments in clones from the shotgun library are partially sequenced. The sequencing reaction must start from a primer of known sequence. Because the sequence of a cloned insert is not known (and is the goal of the exercise), primers are based on the sequence of adjacent vector DNA. These primers are used to guide the sequencing reaction into the insert. Hence, short regions at one or both ends of the genomic inserts can be sequenced (**Figure 14-5**). After sequencing, the output is a large collection of random short sequences, some of them

overlapping. These sequence reads are assembled into a consensus sequence covering the whole genome by matching homologous sequences shared by reads from overlapping clones. The sequences of overlapping reads are assembled into units called **sequence contigs**, which are sequences that are contiguous, or touching.

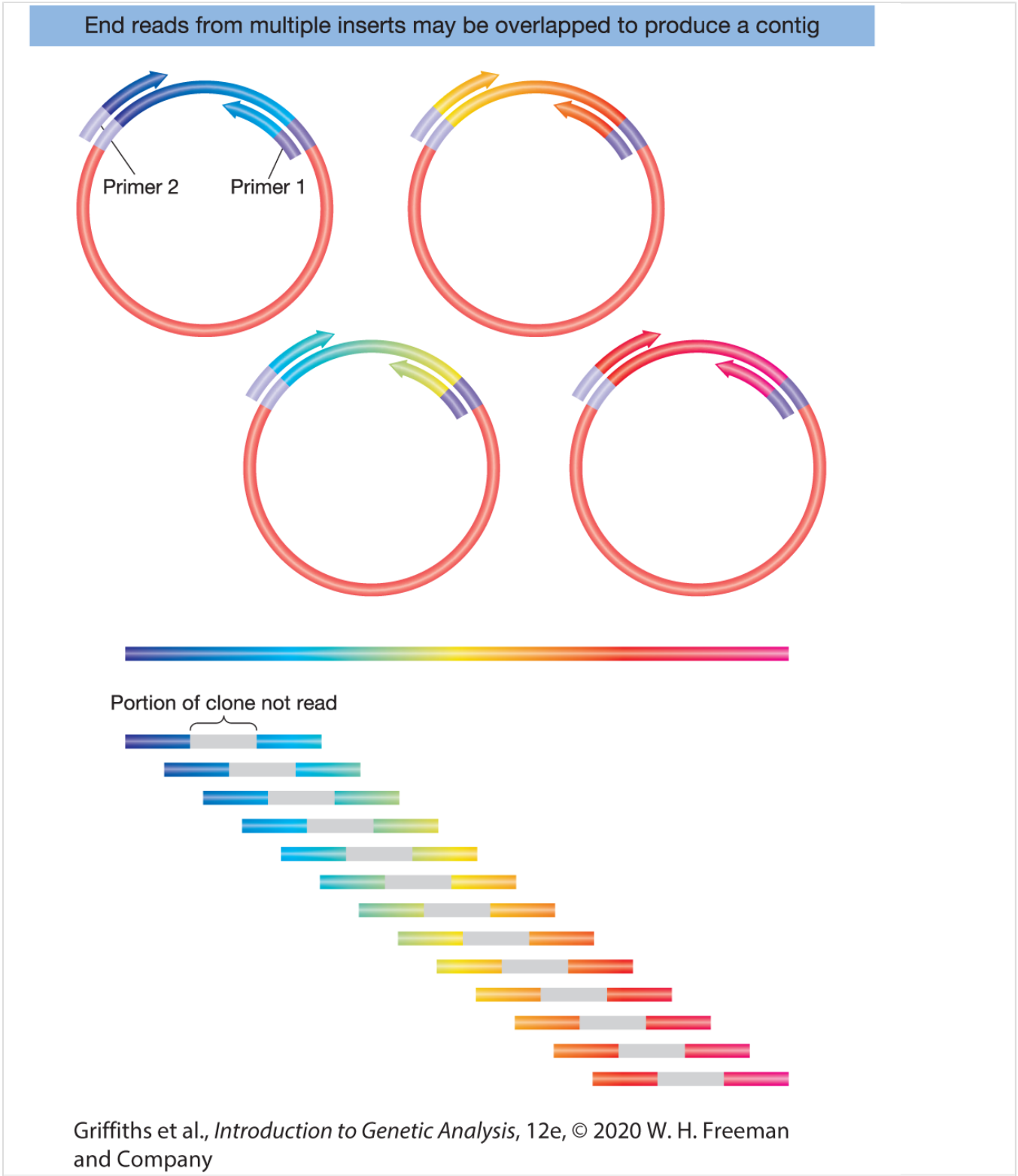


FIGURE 14-5 Sequencing reads are taken only of the ends of cloned inserts. The use of two different sequence-priming sites, one at each end of the vector, makes possible the sequencing of as many as 600 base pairs at each end of the genomic insert. If both ends of the same clone are sequenced, the two resulting sequence reads are called *paired-end reads*. When

paired-end reads from many different clones are obtained, they can be assembled into a sequence contig even though the sequence from the middle of each single clone is missing (gray bars).

KEY CONCEPT Whole genomes can be assembled from sequencing many short segments of DNA.

Next-generation WGS sequencing

The goal of next-generation WGS is the same as that of traditional WGS—to obtain a large number of overlapping sequence reads that can be assembled into contigs. However, the methodologies used differ in several substantial ways from traditional WGS. A few different systems have been developed that, while they differ in their sequencing chemistry and machine design, each employ three strategies that have dramatically increased throughput:

1. DNA molecules are prepared for sequencing in *cell-free* reactions, without cloning in microbial hosts.
2. Millions of individual DNA fragments are isolated and sequenced in parallel during each machine run.
3. Advanced fluid-handling technologies, cameras, and software make it possible to detect the products of sequencing reactions in extremely small reaction volumes.

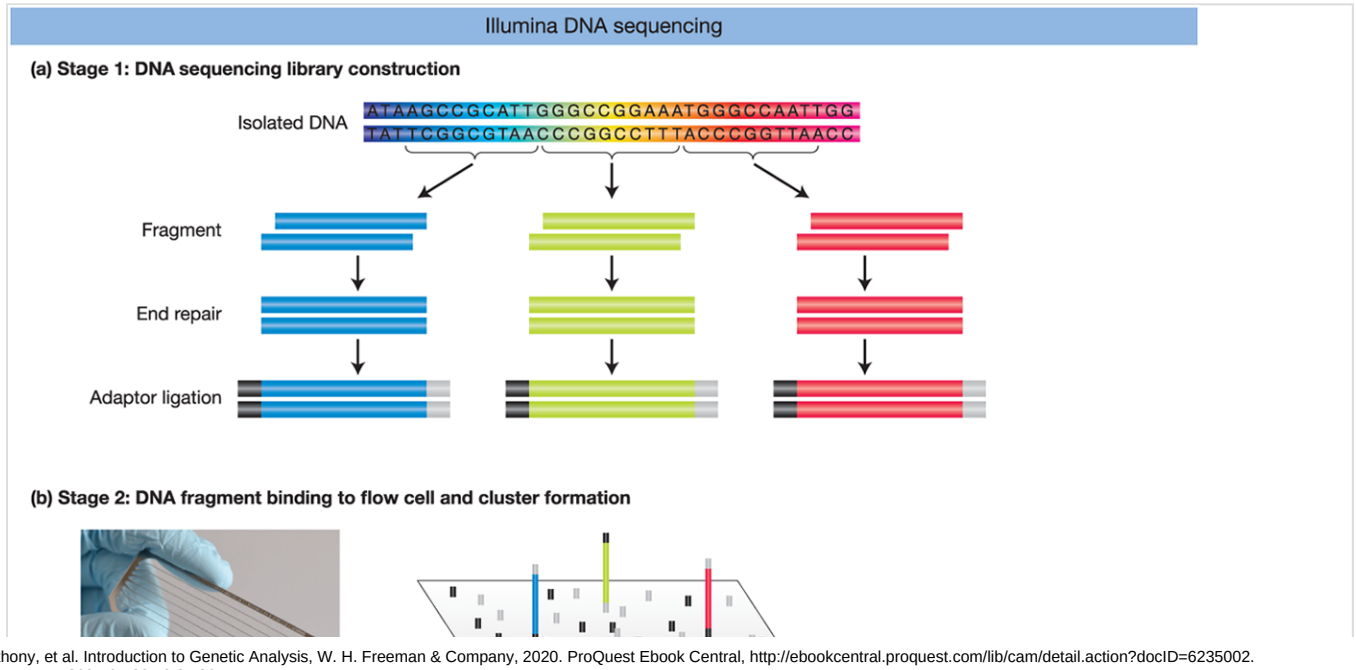
Since the field of genomic technology is evolving rapidly, we will not describe every next-generation system. Here, we will examine *Illumina sequencing*, which is currently the most widely used approach that employs all of these features. The Illumina approach illustrates the gains that have been made in throughput and what such gains enable geneticists to do. The approach can be considered to have three stages:

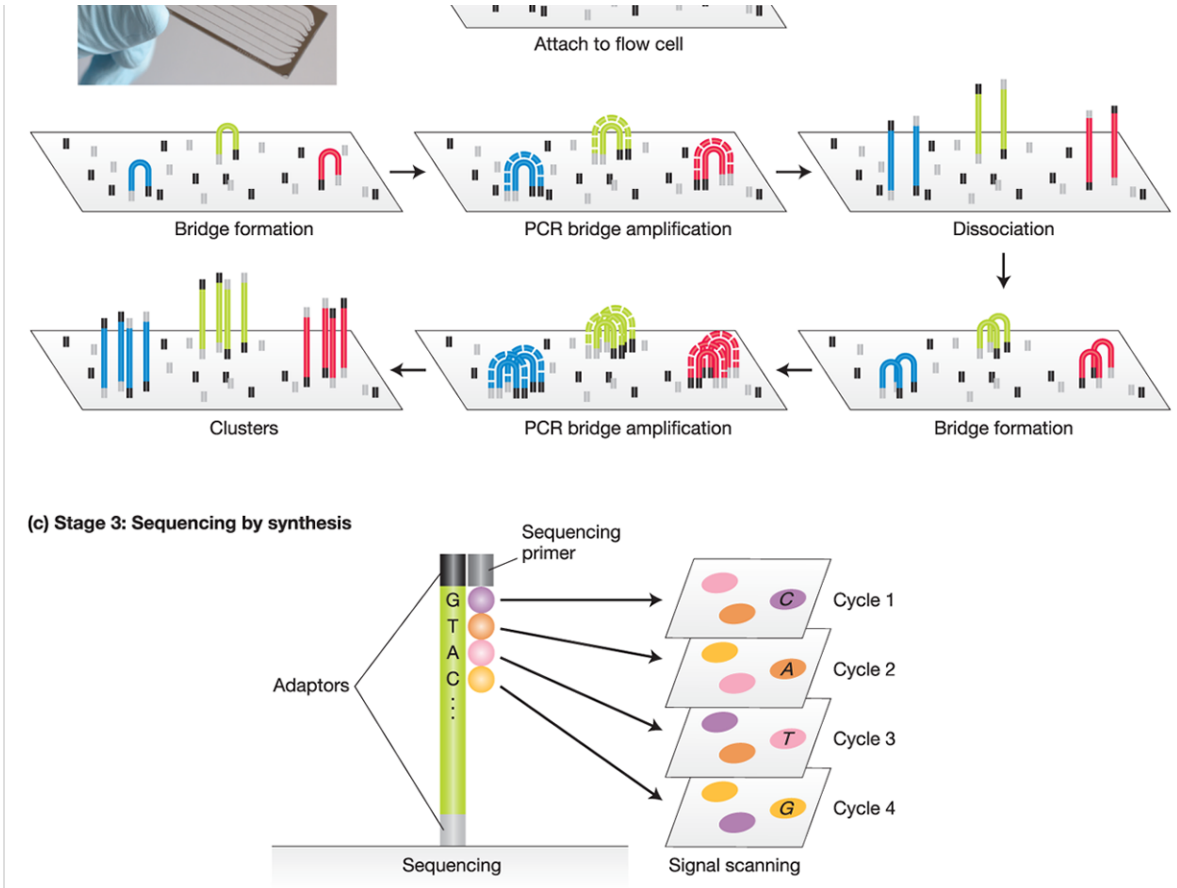
Stage 1. A **DNA sequencing library** of DNA molecules is constructed. After genomic DNA is isolated from an organism of interest, it is fragmented into smaller pieces of a uniform size. Then, short sequences called *adaptors* are added to both ends of the DNA fragments. There are two adaptor sequences; one sequence is added to one end of the DNA fragment, and the other sequence is added to the other end of the DNA fragment (**Figure 14-6a**).

Stage 2. The DNA fragments are bound to a sequencing *flow cell*. This is a glass slide with small channels that are coated with oligonucleotides containing sequences complementary to both adaptor sequences (Figure 14-6b, inset). A single DNA molecule will bind to a unique location in

the flow cell due to hybridization between the adaptor sequence at one end of the DNA molecule and the oligonucleotide on the flow cell. Then, the adaptor on the other end of the DNA molecule will bind to its complementary oligonucleotide, which is called *bridge formation*. Once immobilized, each DNA molecule is amplified across this bridge by the polymerase chain reaction (PCR; see Chapter 10). After one round of *PCR bridge amplification*, there will be two DNA molecules with complementary sequence on the same location on the flow cell. One end of each of the two DNA molecules will be dissociated from the flow cell. This *dissociation* allows for another round of bridge formation and PCR bridge amplification to take place. Repeating this process many times will generate *clusters*. Each cluster contains thousands of copies of the same DNA fragment in a tiny spot (Figure 14-6b). Each channel of the flow cell contains millions to billions of these clusters.

Stage 3. The sequencing of each cluster is performed using a novel "sequencing-by-synthesis" approach (Figure 14-6c). DNA polymerase and a primer are added to the flow cell to prime the synthesis of a complementary DNA strand. Each of the four deoxyribonucleotide triphosphates, dATP, dGTP, dTTP, and dCTP, is labeled with a different fluorescent dye that emits a signal at different wavelengths (and therefore appears as a different color). In each sequencing cycle, a single nucleotide will be added that is complementary to the next base in the template strand in a given cluster. When the nucleotide is incorporated, the reaction emits a unique wavelength depending upon which base was added. After each sequencing cycle, an image of the flow cell is taken. Each cluster will have added only one of the four bases and will therefore appear as a spot of a single color in the image. The reaction is repeated for at least 100 and up to 300 cycles, and the signals from each cluster over all of the cycles are integrated to generate the sequence reads from each cluster.





Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company
Bainscou, Wikimedia Commons, Creative Commons Attribution 3.0 Unported license, https://commons.wikimedia.org/wiki/File:Next_generation_sequencing_slide.jpg#filehistory.

FIGURE 14-6 Illumina DNA sequencing consists of three stages: (a) DNA sequencing library construction; (b) DNA fragment binding to flow cell and cluster formation; and (c) sequencing by synthesis. See text for details.

ANIMATED ART SaplingPlus

Next-generation sequencing

The pace of development of next-generation sequencing technologies has been astonishing and is continuing at a dizzying rate. Recently, so-called “third-generation” sequencing technologies have been developed to enable sequencing of single molecules of DNA. Third-generation methods like those developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies provide a number of advantages over second-generation sequencing methods such as Illumina. These include the ability to generate very long sequence reads, which greatly enables the assembly of whole genomes, as detailed in the next section. However, these newer sequencing methods currently have a lower throughput and a higher error rate. Thus, the method chosen by researchers depends a great deal on the application, and these choices will continue to evolve in the coming years.

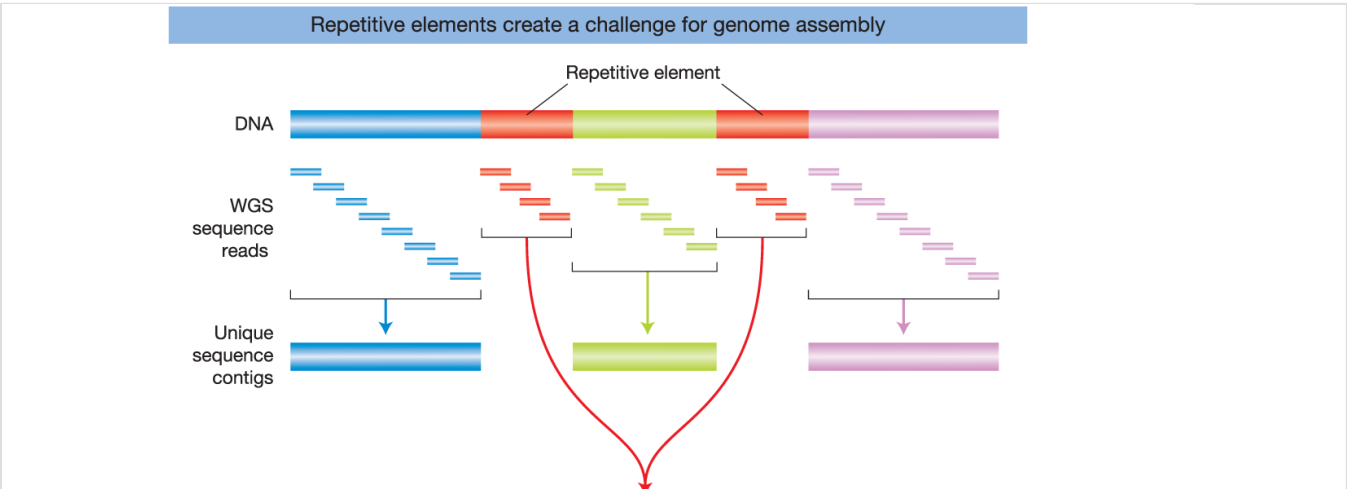
KEY CONCEPT Next-generation WGS sequencing methods have already enabled enormous gains in sequencing output and are continuing to evolve at a rapid rate.

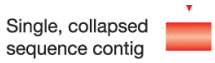
Whole-genome-sequence assembly

Whichever method of obtaining raw sequence is used, the challenge remains to assemble the contigs into the entire genome sequence. The difficulty of that process depends strongly on the size and complexity of the genome.

For instance, the genomes of bacterial species are relatively easy to assemble. Bacterial DNA is essentially *single-copy* DNA, with no repeating sequences. Therefore, any given DNA sequence read from a bacterial genome will come from one unique place in that genome. Owing to these properties, contigs within bacterial genomes can often be assembled into larger contigs representing most or all of the genome sequence in a relatively straightforward manner. In addition, a typical bacterial genome is only a few megabase pairs of DNA in size.

For eukaryotes, genome assembly often presents some difficulties. A big stumbling block is the existence of numerous classes of repeated sequences, some arranged in tandem and others dispersed (see Chapter 16). Why are they a problem for genome sequencing? In short, because a sequencing read of repetitive DNA fits into many places in the draft of the genome. Not infrequently, a tandem repetitive sequence is in total longer than the length of a maximum sequence read. In that case, there is no way to bridge the gap between adjacent unique sequences. Dispersed repetitive elements can cause reads from different chromosomes or different parts of the same chromosome to be mistakenly assembled together in a single, collapsed sequence contig (**Figure 14-7**).





Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

FIGURE 14-7 WGS reads from sequences that are found in only one location in the genome can be assembled into many unique sequence contigs. By contrast, WGS reads from repetitive elements found in many locations in the genome will be collapsed into a single sequence contig.

KEY CONCEPT The landscape of eukaryotic chromosomes includes a variety of repetitive DNA segments. These segments are difficult to assemble as sequence reads.

WGS sequencing is particularly good at producing draft-quality sequences of complex genomes with many repetitive sequences. As an example, we will consider the genome of the fruit fly *D. melanogaster*, which was initially sequenced by the traditional WGS sequencing method. The project began with the sequencing of libraries of genomic clones of different sizes (2 kb, 10 kb, 150 kb). Sequence reads were obtained from *both* ends of genomic-clone inserts and aligned by a logic identical to that used for bacterial WGS sequencing. Through this logic, sequence overlaps were identified and clones were placed in order, producing sequence contigs—consensus sequences for these single-copy stretches of the genome. However, unlike the situation in bacteria, the contigs eventually ran into a repetitive DNA segment that prevented unambiguous assembly of the contigs into a whole genome. The sequence contigs had an average size of about 150 kb. The challenge, then, was how to glue the thousands of such sequence contigs together in their correct order and orientation.

The solution to this problem was to make use of the pairs of sequence reads from opposite ends of the genomic inserts in the same clone—these reads are called **paired-end reads**. The idea was to find paired-end reads that spanned the gaps between two sequence contigs (**Figure 14-8**). In other words, if one end of an insert was part of one contig and the other end was part of a second contig, then this insert must span the gap between two contigs, and the two contigs were clearly near each other. Indeed, because the size of each clone was known (that is, it came from a library containing genomic inserts of uniform size, either the 2-kb, 100-kb, or 150-kb library), the distance between the end reads was known. Further, aligning the sequences of the two contigs by using paired-end reads automatically determines the relative orientation of the two contigs. In this manner, single-copy contigs could be joined together, albeit with gaps where the repetitive elements reside. These gapped collections of joined-together sequence contigs are called **scaffolds** (sometimes also referred to as **supercontigs**). Because most *Drosophila* repeats are large (3−8 kb) and widely spaced (one repeat approximately every 150 kb), this technique was extremely effective at

producing a correctly assembled draft sequence of the single-copy DNA. A summary of the logic of this approach is shown in **Figure 14-9**.

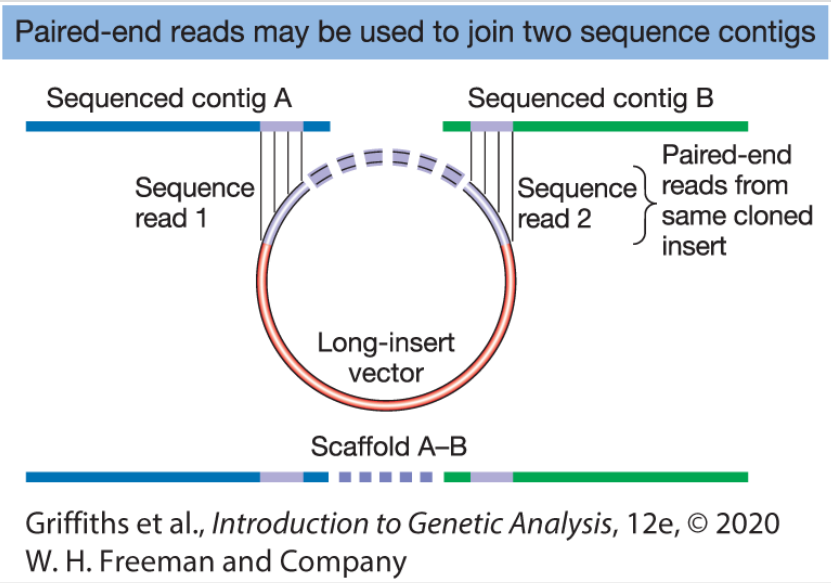


FIGURE 14-8 Paired-end reads can be used to join two sequence contigs into a single ordered and oriented scaffold.

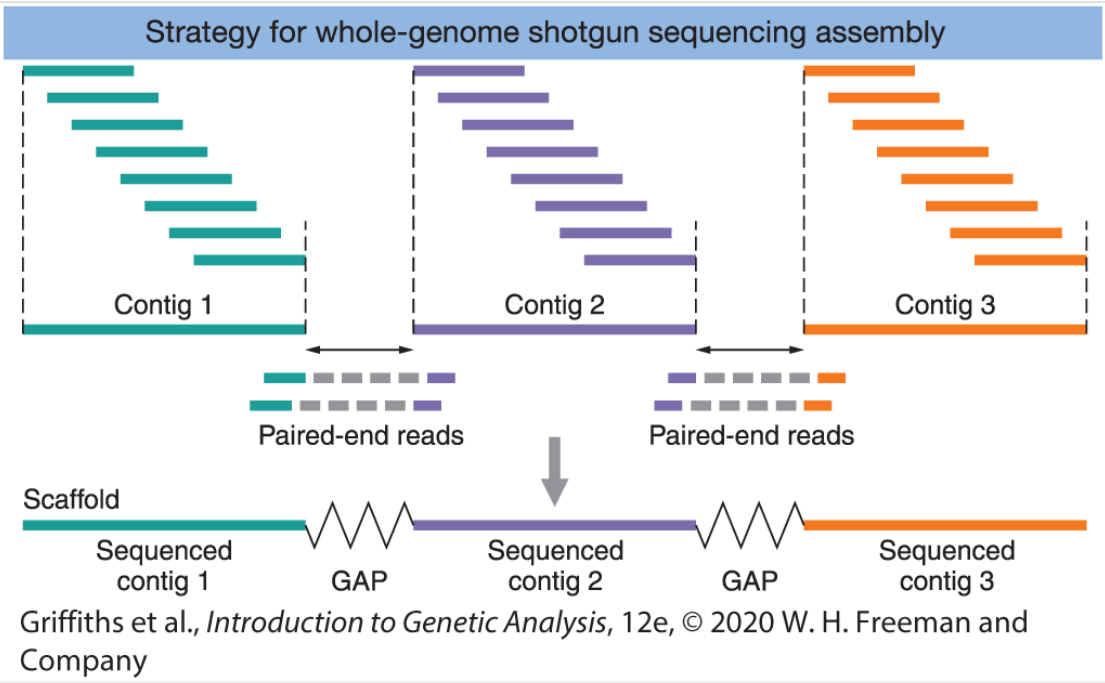


FIGURE 14-9 In whole-genome shotgun sequencing, first, the unique sequence overlaps between sequence reads are used to build contigs. Paired-end reads are then used to span gaps and to order and orient the contigs into larger units, called *scaffolds*.

Next-generation WGS sequencing does not circumvent the problem of repetitive sequences and gaps. Since this approach is intended to circumvent the construction of libraries, which would otherwise facilitate the bridging of gaps between contigs via paired-end reads, next-generation

WGS researchers had to devise a way to bridge these gaps without building genomic libraries in vectors. One solution was to build a library of circularized genomic DNA fragments of desired sizes. The circularization allows for short segments of previously distant sequences located at the ends of each fragment to be juxtaposed. Shearing of these circular molecules and amplification and sequencing of fragments containing the junction produces paired-end reads equivalent to those obtained from sequencing of traditional genomic-library inserts (Figure 14-10).

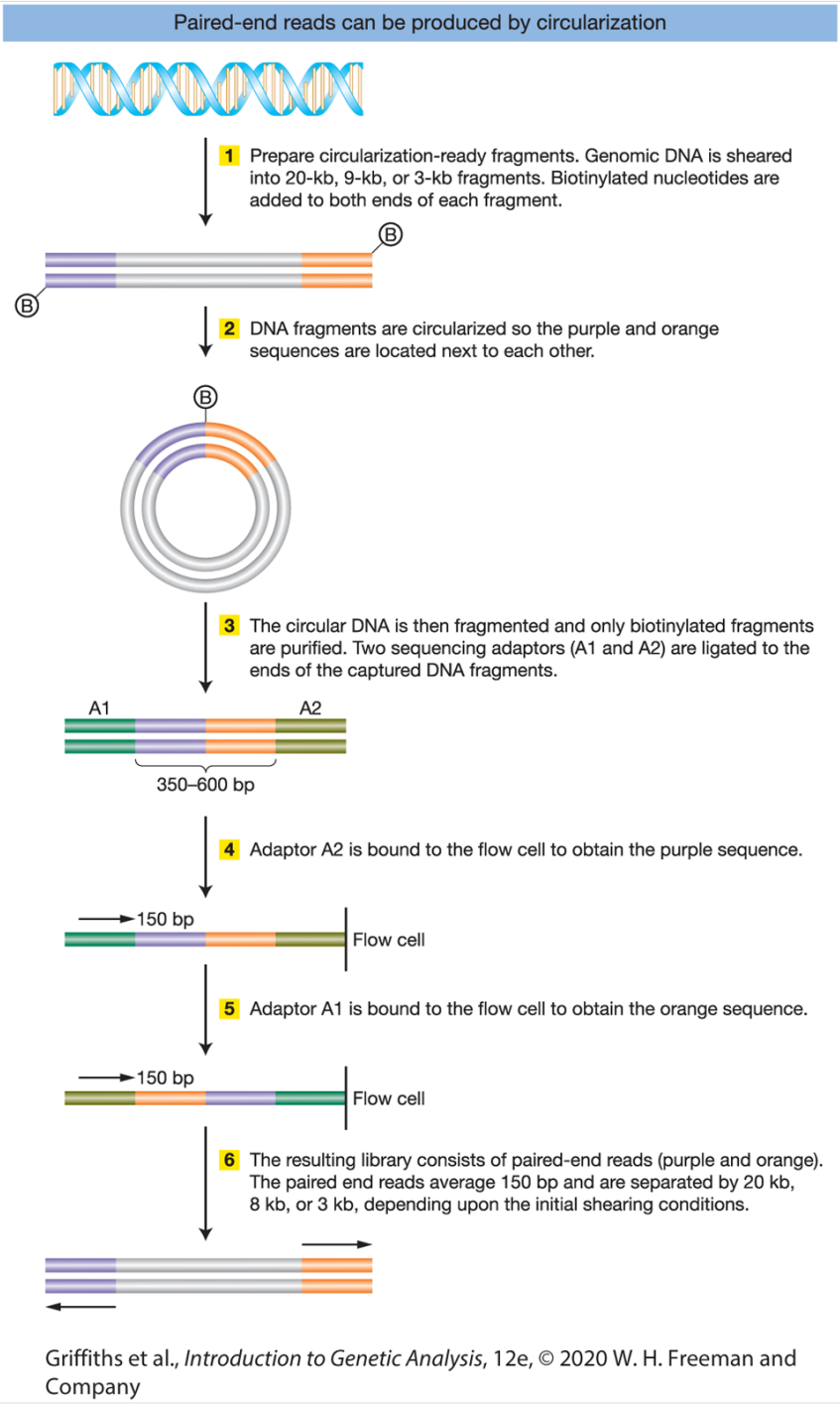


FIGURE 14-10 Paired-end reads for high-throughput sequencing can be produced without genomic-library construction.

KEY CONCEPT Paired-end reads are crucial for assembling genomes from both traditional and next-generation WGS sequencing data.

In both traditional and next-generation WGS sequencing, some gaps usually remain. Specific procedures targeted to individual gaps must be used to fill the missing data in the sequence assemblies. If the gaps are short, missing fragments can be generated by using the known sequences at the ends of the assemblies as primers to amplify and analyze the genomic sequence in between. If the gaps are longer, attempts can be made to isolate the missing sequences as parts of larger inserts that have been cloned into a vector, and then to sequence the inserts. In the future, the longer sequencing reads generated by third-generation sequencing methods will also contribute to filling the gaps in sequence assemblies, particularly in regions of the genome that contain many repeat sequences.

Whether a genome is sequenced to “draft” or “finished” standards is a cost–benefit judgment. Currently, it is relatively straightforward to create a draft but very hard to complete a finished sequence.

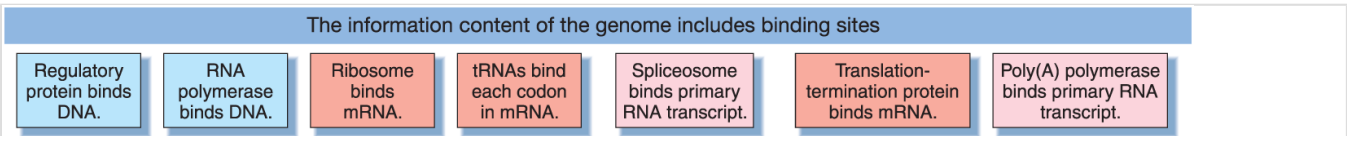
14.3 BIOINFORMATICS: MEANING FROM GENOMIC SEQUENCE

LO 14.2 Explain the role of various functional elements within genomes, and differentiate between computational and experimental methods used to identify these elements.

The genomic sequence is a highly encrypted code containing the raw information for building and operation of organisms. The study of the information content of genomes is called **bioinformatics**. We are far from being able to read this information from beginning to end in the way that we would read a book. Even though we know which triplets encode which amino acids in the protein-coding segments, much of the information contained in a genome is not decipherable from mere inspection.

The nature of the information content of DNA

DNA contains information, but in what way is it encoded? Conventionally, the information is thought of as the sum of all the gene products, both proteins and RNAs. However, the information content of the genome is more complex than that. The genome also contains binding sites for different proteins and RNAs. Many proteins bind to sites located in the DNA itself, whereas other proteins and RNAs bind to sites located in mRNA (**Figure 14-11**). The sequence and relative positions of those sites permit genes to be transcribed, spliced, and translated properly, at the appropriate time in the appropriate tissue. For example, regulatory protein-binding sites determine when, where, and at what level a gene will be expressed. At the RNA level in eukaryotes, the locations of binding sites for the RNAs and proteins of spliceosomes will determine the 5′ and 3′ splice sites where introns are removed. Regardless of whether a binding site actually functions as such in DNA or RNA, the site must be encoded in the DNA. The information in the genome can be thought of as the sum of all the sequences that encode proteins and RNAs, plus the binding sites that govern the time and place of their actions. As a genome draft continues to be improved, the principal objective is the identification of all of the functional elements of the genome. This process is referred to as **annotation**.



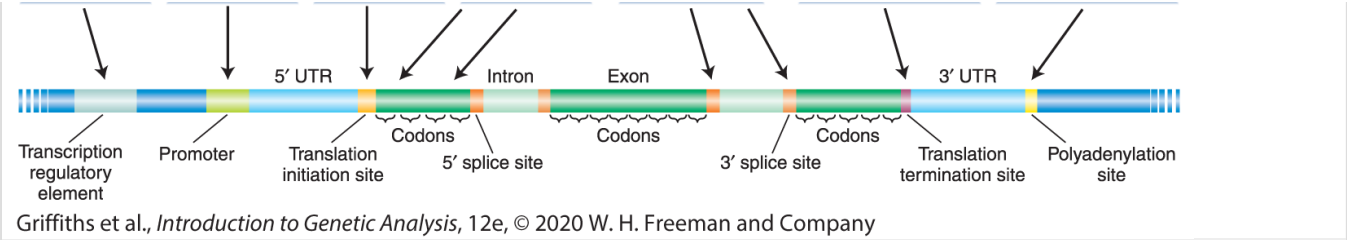


FIGURE 14-11 A gene within DNA may be viewed as a series of binding sites for proteins and RNAs.

KEY CONCEPT The functional elements of the genome include the sequences that encode proteins and RNAs, as well as the binding sites for the proteins and RNAs that regulate gene expression.

Deducing the protein-encoding genes from genomic sequence

Because the proteins present in a cell largely determine its morphology and physiological properties, one of the first orders of business in genome analysis and annotation is to try to determine an inventory of all of the polypeptides encoded by an organism's genome. This inventory is referred to as the organism's *proteome*. It can be considered a "parts list" for the cell. To determine the list of polypeptides, the sequence of each mRNA encoded by the genome must be deduced. Because of intron splicing, this task is particularly challenging in multicellular eukaryotes, where introns are the norm. In humans, for example, an average gene has about 10 exons. Furthermore, many genes encode alternative exons; that is, some exons are included in some versions of a processed mRNA but are not included in others (see Chapter 8). The alternatively processed mRNAs can encode polypeptides having much, but not all, of their amino acid sequences in common. Even though we have a great many examples of completely sequenced genes and mRNAs, we cannot yet identify 5′ and 3′ splice sites merely from DNA sequence with a high degree of accuracy. Therefore, we cannot be certain which sequences are introns. Predictions of alternatively used exons are even more error prone. For such reasons, deducing the total polypeptide parts list in higher eukaryotes is a large problem. Some approaches follow.

ORF detection

The main approach to producing a polypeptide list is to use the computational analysis of the genome sequence to predict mRNA and polypeptide sequences, an important part of

bioinformatics. The procedure is to look for sequences that have the characteristics of genes. These sequences would be gene-size and composed of sense codons after possible introns had been removed. The appropriate 5′- and 3′-end sequences would be present, such as start and stop codons. Sequences with these characteristics typical of genes are called **open reading frames (ORFs)**. To find candidate ORFs, computer programs scan the DNA sequence on both strands in each reading frame. Because there are three possible reading frames on each strand, there are six possible reading frames in all.

Direct evidence from cDNA sequences

Another means of identifying ORFs and exons is through the analysis of mRNA expression. This analysis can be done in two ways. Both methods involve the synthesis of libraries of DNA molecules that are complementary to mRNA sequences, called cDNA (see Chapter 10). The longest established method entails the cloning and amplification of these cDNA molecules in a vector. However, the next-generation sequencing technologies described in the previous section also allow for the direct sequencing of short cDNA molecules without the cloning step, called **RNA sequencing** or "**RNA-seq**" for short (this technique will be described in more detail later in Section 14.7). Whichever method is utilized, complementary DNA sequences are extremely valuable in two ways. First, they are direct evidence that a given segment of the genome is expressed and may thus encode a gene. Second, because the cDNA is complementary to the mature mRNA, the introns of the primary transcript have been removed, which greatly facilitates the identification of the exons and introns of a gene (**Figure 14-12**).

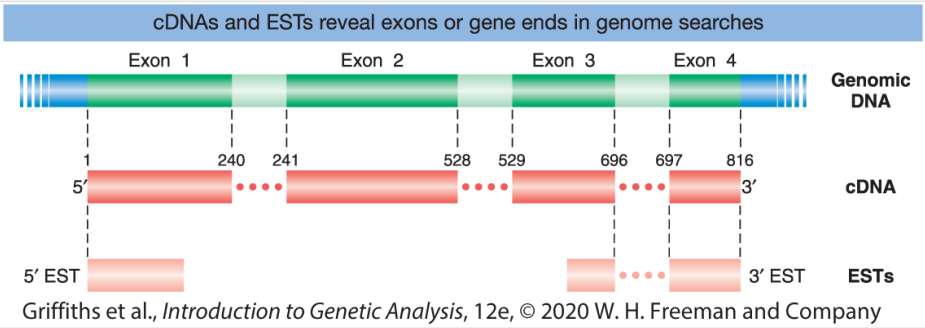


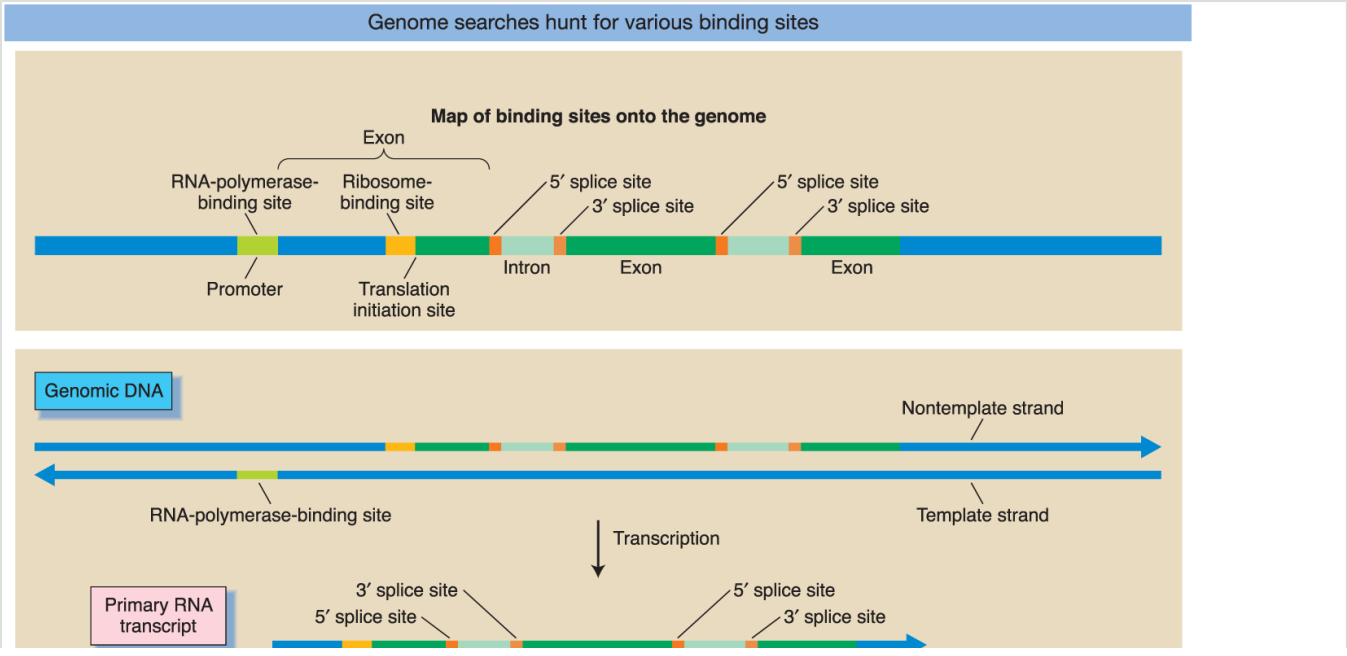
FIGURE 14-12 Alignment of fully sequenced complementary DNAs (cDNAs) and expressed sequence tags (ESTs) with genomic DNA. The dashed lines indicate regions of alignment; for the cDNA, these regions are the exons of the gene. The dots between segments of cDNA or ESTs indicate regions in the genomic DNA that do not align with cDNA or EST sequences; these regions are the locations of the introns. The numbers above the cDNA line indicate the base coordinates of the cDNA sequence, where base 1 is the 5′-most-most base and base 816 is the 3′-most-most base of the cDNA. For the ESTs, only a short sequence read is obtained from each end (5′ and 3′) of the corresponding cDNA. These sequence reads establish the boundaries of the transcription unit, but they are not informative about the internal structure of the transcript unless the EST sequences cross an intron (as is true for the 3′ EST depicted here).

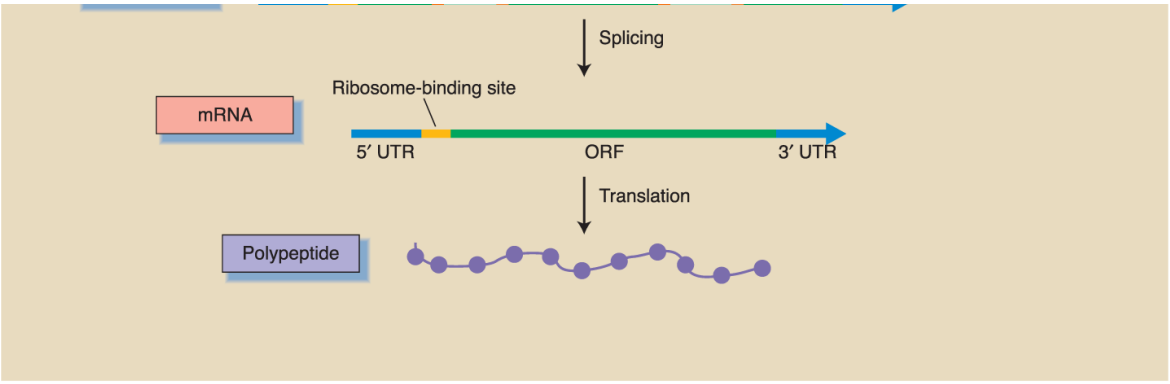
The alignment of cDNAs with their corresponding genomic sequence clearly delineates the exons, and hence introns are revealed as the regions falling between the exons. In the assembled cDNA sequence, the ORF should be continuous from initiation codon through stop codon. Thus, cDNA sequences can greatly assist in identifying the correct reading frame, including the initiation and stop codons. Full-length cDNA evidence is taken as the gold-standard proof that one has identified the sequence of a transcription unit, including its exons and its location in the genome.

In addition to full-length cDNA sequences, there are large data sets of cDNAs for which only the 5′ or the 3′ ends or both have been sequenced. These short cDNA sequence reads are called **expressed sequence tags (ESTs)**. Expressed sequence tags can be aligned with genomic DNA and thereby used to determine the 5′ and 3′ ends of transcripts—in other words, to determine the boundaries of the transcript as shown in Figure 14-12.

Predictions of binding sites

As already discussed, a gene consists of a segment of DNA that encodes a transcript as well as the regulatory signals that determine when, where, and how much of that transcript is made. In turn, that transcript has the signals necessary to determine its splicing into mRNA and the translation of that mRNA into a polypeptide (**Figure 14-13**). There are now statistical “gene-finding” computer programs that search for the predicted sequences of the various binding sites used for promoters, for transcription start sites, for 3′ and 5′ splice sites, and for translation initiation codons within genomic DNA. These predictions are based on consensus motifs for such known sequences, but they are not perfect.





Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

FIGURE 14-13 Eukaryotic information transfer from gene to polypeptide chain. Note the DNA and RNA "binding sites" that are bound by protein complexes to initiate the events of transcription, splicing, and translation.

Using polypeptide and DNA similarity

Because organisms have common ancestors, they also have many genes with similar sequences in common. Hence, a gene will likely have relatives among the genes isolated and sequenced in other organisms, especially in the closely related ones. Candidate genes predicted by the preceding techniques can often be verified by comparing them with all the other gene sequences that have ever been found. A candidate sequence is submitted as a "query sequence" to public databases containing a record of all known gene sequences. This procedure is called a BLAST search (BLAST stands for Basic Local Alignment Search Tool). The sequence can be submitted as a nucleotide sequence (a BLASTn search) or as a translated amino acid sequence (BLASTp). The computer scans the database and returns a list of full or partial "hits," starting with the closest matches. If the candidate sequence closely resembles that of a gene previously identified from another organism, then this resemblance provides a strong indication that the candidate gene is a real gene. Less-close matches are still useful. For example, an amino acid identity of only 35 percent, but at identical positions, is a strong indicator that two proteins have a common three-dimensional structure.

BLAST searches are used in many other ways, but always the goal is to find out more about some identified sequence of interest.

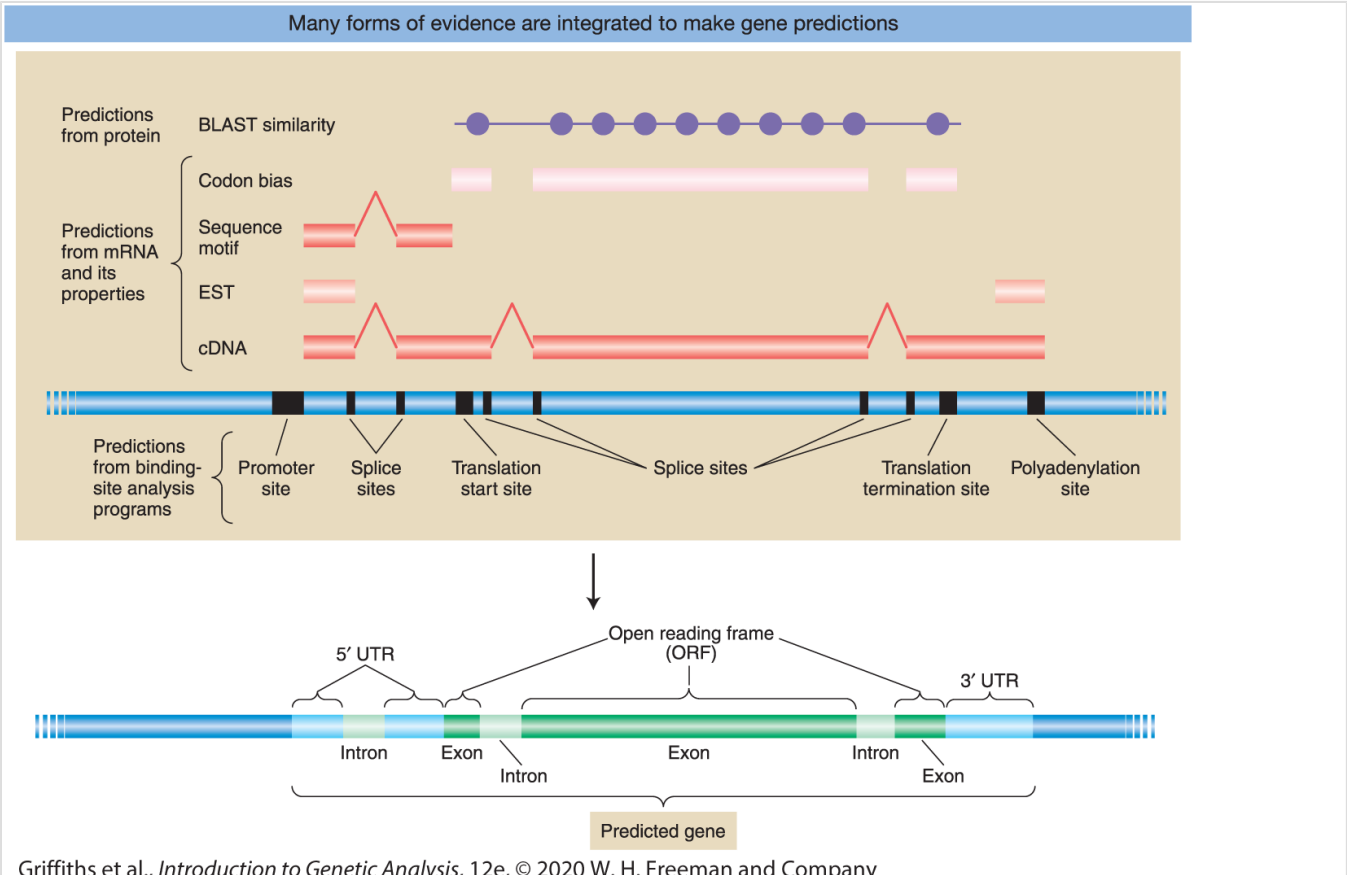
Predictions based on codon bias

Recall from Chapter 9 that the triplet code for amino acids is degenerate; that is, most amino acids are encoded by two or more codons (see Figure 9-8). The multiple codons for a single amino acid

are termed *synonymous codons*. In a given species, not all synonymous codons for an amino acid are used with equal frequency. Rather, certain codons are present much more frequently in mRNAs (and hence in the DNA that encodes them). For example, in *D. melanogaster*, of the two codons for cysteine, UGC is used 73 percent of the time, whereas UGU is used 27 percent. This usage is a diagnostic for *Drosophila* because, in other organisms, this "codon bias" pattern is quite different. Codon biases are thought to be due to the relative abundance of the tRNAs complementary to these various codons in a given species. If the codon usage of a predicted ORF matches that species' known pattern of codon usage, then this match is supporting evidence that the proposed ORF is genuine.

Putting it all together

A summary of how different sources of information are combined to create the best-possible mRNA and gene predictions is depicted in **Figure 14-14**. These different kinds of evidence are complementary and can cross-validate one another. For example, the structure of a gene may be inferred from evidence of protein similarity within a region of genomic DNA bounded by 5′ and 3′ ESTs. Useful predictions are possible even without a cDNA sequence or evidence of protein similarities. A binding-site-prediction program can propose a hypothetical ORF, and proper codon bias would be supporting evidence.



Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

FIGURE 14-14 The different forms of gene-product evidence—cDNAs, ESTs, BLAST-similarity hits, codon bias, and motif hits—are integrated to make gene predictions. Where multiple classes of evidence are found to be associated with a particular genomic DNA sequence, there is greater confidence in the likelihood that a gene prediction is accurate.

KEY CONCEPT Predictions of mRNA and polypeptide structure from genomic DNA sequence depend on the integration of information from cDNA and EST sequence, binding-site predictions, polypeptide similarities, and codon bias.

Let’s consider some of the insights from our first view of the overall genome structures and global parts lists of a few species whose genomes have been sequenced. We will start with ourselves. What can we learn by looking at the human genome by itself? Then we will see what we can learn by comparing our genome with others.

14.4 THE STRUCTURE OF THE HUMAN GENOME

LO 14.2 Explain the role of various functional elements within genomes, and differentiate between computational and experimental methods used to identify these elements.
LO 14.5 Outline reverse genetic approaches to analyze the function of genes and genetic elements identified by genome sequencing and comparative genomics.

In describing the overall structure of the human genome, we must first confront its repeat structure. A considerable fraction of the human genome, about 45 percent, is repetitive. Much of this repetitive DNA is composed of copies of transposable elements (discussed in Chapter 16). Indeed, even within the remaining single-copy DNA, a fraction has sequences suggesting that they might be descended from ancient transposable elements that are now immobile and have accumulated random mutations, causing them to diverge in sequence from the ancestral transposable elements. Thus, much of the human genome appears to be composed of genetic "hitchhikers."

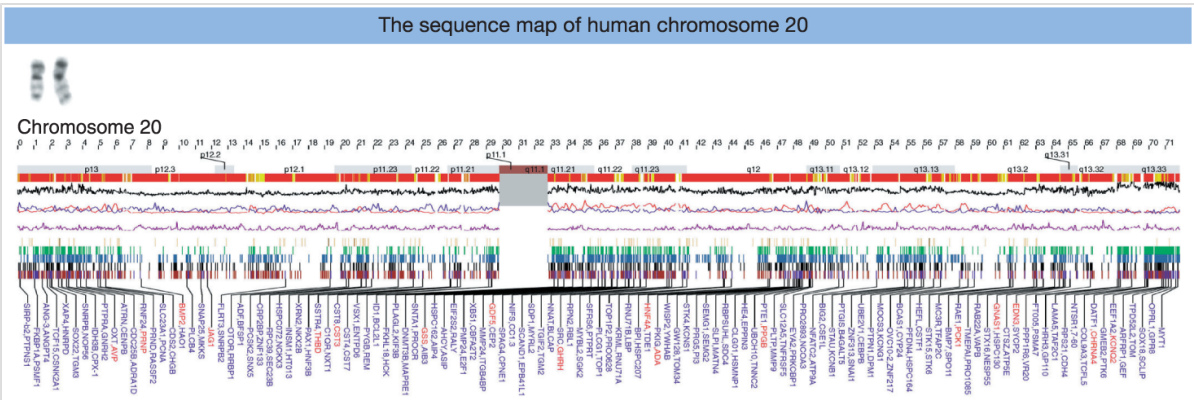
Only a small part of the human genome encodes polypeptides; that is, somewhat less than 3 percent of it encodes exons of mRNAs. Exons are typically small (about 150 bases), whereas introns are large, many extending more than 1000 bases and some extending more than 100,000 bases. Transcripts are composed of an average of 10 exons, although many have substantially more. Finally, introns may be spliced out of the same gene in locations that vary. This variation in the location of splice sites generates considerable added diversity in mRNA and polypeptide sequence. On the basis of current cDNA and EST data, at least 60 percent of human protein-coding genes are likely to have two or more splice variants. On average, there are several splice variants per gene. Hence, the number of distinct proteins encoded by the human genome is several-fold greater than the number of recognized genes.

KEY CONCEPT Only a small proportion of the human genome consists of protein-coding genes.

The number of genes in the human genome has not been easy to pin down. In the initial draft of the human genome, there were an estimated 30,000 to 40,000 protein-coding genes. However, the complex architecture of these genes and the genome can make annotation difficult. Some

sequences scored as genes may actually be exons of larger genes. In addition, there are approximately 15,000 **pseudogenes**, which are ORFs or partial ORFs that may at first appear to be genes but are either nonfunctional or inactive due to the manner of their origin or to mutations. So-called **processed pseudogenes** are DNA sequences that have been reverse-transcribed from RNA and randomly inserted into the genome. Seventy percent or so of human pseudogenes appear to be of this type. Most of the other pseudogenes in the human genome appear to have arisen from gene duplication events in which one of the duplicates has acquired one or more ORF-disrupting mutations in the course of evolution. As the challenges in annotation have been overcome, the estimated number of genes in the human genome has dropped steadily. A recent estimate is that there are about 20,000 protein-coding genes.

The annotation of the human genome progressed as the sequences of each chromosome were finished one by one. These sequences then became the searching ground in the hunt for candidate genes for human diseases. An example of gene predictions for a chromosome from the human genome is shown in **Figure 14-15**. Such predictions are being revised continually as new data become available. The current state of the predictions can be viewed at many Web sites, most notably at the public DNA databases in the United States and Europe (see Appendix B). These predictions are the current best inferences of the protein-coding genes present in the sequenced species and, as such, are works in progress.



Reprinted with permission from Macmillan Publishers Ltd: from Jim Kent, Ewan Birney, Darryl Leja, and Francis Collins, After the International Human Genome Sequencing Consortium, "Initial Sequencing and Analysis of the Human Genome," Nature, *2001, February 15; 409, 860–921, Figure 9, pdf 1. Permission conveyed through Copyright Clearance Center, Inc.*

FIGURE 14-15 Numerous genes have been identified on human chromosome 20. The recombinational and cytogenetic map coordinates are shown in the top lines of the figure. Various graphics depicting gene density and different DNA properties are shown in the middle sections. The identifiers of the predicted genes are shown at the bottom of the panel.

Noncoding functional elements in the genome

The discussion thus far has focused exclusively on the protein-coding regions of the genome. This emphasis is due more to analytical ease than to biological importance. Because of the simplicity and universality of the genetic code, and the ability to synthesize cDNA from mRNA, the detection of ORFs and exons is much easier than the detection of functional noncoding sequences. As stated earlier, only 3 percent of the human genome encodes exons of mRNAs, and fewer than half of these exon sequences, a little over 1 percent of the total genome DNA, encode protein sequences. So, nearly 99 percent of our genome does not encode proteins. How do we identify other functional parts of the genome?

Introns and 5′ and 3′ untranslated sequences are readily annotated by analysis of gene transcripts, while gene promoters are usually identified by their proximity to transcription units and signature DNA sequences. However, other regulatory sequences such as enhancers are not identifiable by mere inspection of DNA sequences, and other sequences that encode various kinds of RNA transcripts (microRNAs, small interfering RNAs, piwi-interacting RNAs, long noncoding RNAs; see Chapter 8) require detection and annotation of their transcripts. While many such noncoding elements have been identified in the course of the study of human molecular genetics, the potentially vast number of such elements warrants a more systematic approach. The Encyclopedia of DNA Elements (ENCODE) project was thus launched with the ambitious goal of identifying all functional elements within the human genome.

This large-scale collaborative endeavor has employed a diverse array of techniques to detect sequences potentially involved in the control of gene transcription, as well as all transcribed regions. Because such sequences are expected to be active in only individual or subsets of cell types, researchers studied 147 human cell types. By searching for regions that were associated with the binding of transcription factors, the ENCODE project estimated that there are approximately 500,000 potential enhancers and promoters associated with known genes. The project also detected transcripts emanating from nearly 80 percent of the human genome.

This is a much larger fraction of the genome than was expected. After all, as stated earlier, only a little over 1 percent of the genome is protein-coding sequence. However, the production of a transcript does not necessarily mean that the transcript contributes to human biology. It is possible that some proportion of these transcripts represent “noise” in the cell—transcripts that have no

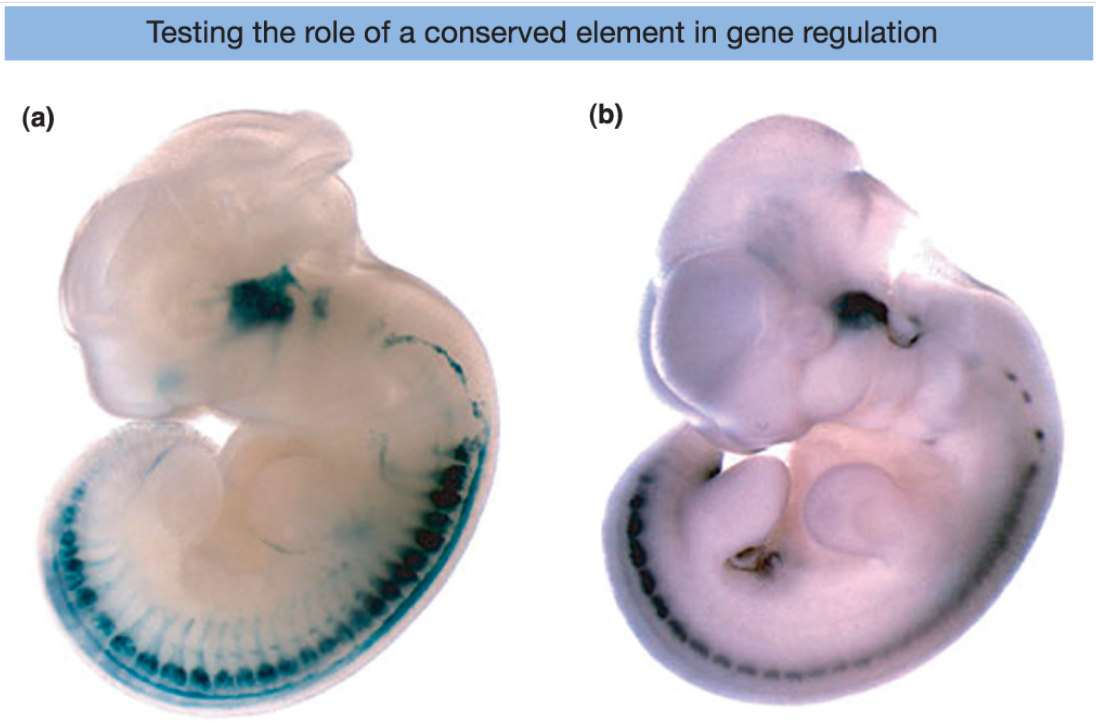
biological function, but also do no harm. It is not sound to ascribe function to a sequence without some form of additional data, so what kinds of additional data can be used to resolve questions of function?

Evolutionary conservation of sequences has proven to be a good indicator of biological function. Sequences will not be preserved over evolutionary time unless mutations that alter them are weeded out by natural selection. One way to locate potentially functional noncoding elements then is to look for conserved sequences, which have not changed much over millions of years of evolution.

For example, one can search for very highly conserved sequences of modest length among a few species or for less perfectly conserved sequences of greater length among a larger number of species. Comparisons of the human, rat, and mouse genomes have led to the identification of so-called *ultraconserved elements*, which are sequences that are perfectly conserved among the three species. Searches of these genomes have found more than 5000 sequences of more than 100 bp and 481 sequences of more than 200 bp that are absolutely conserved. Nearly all of these elements were highly conserved in the chicken genome, and about two-thirds were also conserved in a fish genome. Although many of these elements are found in gene-poor regions, they are most richly concentrated near regulatory genes important for development. The majority of highly conserved noncoding elements may largely take part in regulating the expression of the genetic toolkit for the development of mammals and other vertebrates (see Chapter 13).

How can we verify that such conserved elements play a role in gene regulation? These elements can be tested in the same manner as the transcriptional cis-acting regulatory elements examined in earlier chapters, with the use of reporter genes (see Figure 13-18). A researcher places candidate regulatory regions adjacent to a promoter and reporter gene and introduces the reporter gene into a host species. One such example is shown in **Figure 14-16**. An element that is highly conserved among mammalian, chicken, and a frog species lies 488 kb from the 3′ end of the human *ISL1* gene, which encodes a protein required for motor-neuron differentiation. This element was placed upstream of a promoter and the *β*-galactosidase (*lacZ*) reporter gene, and the construct was injected into the pronuclei of fertilized mouse oocytes (see Figure 10-25). The reporter protein is expressed along the spinal cord and in the head, as one would expect for the location of future motor neurons (see Figure 14-16). Most significantly, the expression pattern corresponds to part of the expression pattern of the native mouse *ISL1* gene (presumably other noncoding elements control the other features of *ISL1* expression). The expression pattern strongly suggests that the conserved element is a regulatory region for the *ISL1* gene in each species. The success of this

approach suggests that many additional human noncoding regulatory elements will likely be identified on the basis of sequence conservation and the activity of those elements in reporter assays.



Reprinted with permission from Macmillan Publishers Ltd: from G. Bejerono et al. "A distal enhancer and an ultraconserved exon are derived from a novel retroposon" Nature, *2006, April 16; 441: 87–90. Figure 3. Permission conveyed through Copyright Clearance Center, Inc.*

FIGURE 14-16 A transcriptional cis-acting regulatory element is identified in an ultraconserved element of the human genome. An ultraconserved element lying near the human *ISL1* gene was coupled to a reporter gene and injected into fertilized mouse oocytes. The regions where the gene is expressed are stained dark blue or black. (a) The reporter gene is expressed in the head and spinal cord of a transgenic mouse, as seen here on day 11.5 of gestation. This expression pattern corresponds to (b) the native pattern of expression of the mouse *ISL1* gene on day 11.5 of gestation. This experiment demonstrates how functional noncoding elements can be identified by comparative genomics and tested in a model organism.

KEY CONCEPT Noncoding regulatory elements can be identified through a combination of computational approaches and reporter gene assays.

14.5 THE COMPARATIVE GENOMICS OF HUMANS WITH OTHER SPECIES

LO 14.3 Infer the evolutionary direction of genomic changes among species based on their phylogenetic relationships.
LO 14.5 Outline reverse genetic approaches to analyze the function of genes and genetic elements identified by genome sequencing and comparative genomics.

Fundamentally, much of the science of genomics entails a comparative approach. For instance, most of what we know about the function of human proteins is based on the function of those proteins as analyzed in model species. And many of the questions that may be addressed through genomics are comparative. For example, we often want to know, as in the case of Nicholas Volker, how an individual with a trait or disease differs genetically from those without it.

Comparative genomics also has the potential to reveal how species diverge. Species evolve and traits change through changes in DNA sequence. The genome thus contains a record of the evolutionary history of a species. Comparisons among species' genomes can reveal events unique to particular lineages that may contribute to differences in physiology, behavior, or anatomy. Such events could include, for example, the gain and loss of individual genes or groups of genes. Here, we will explore the key principles underlying comparative genomics and look at a few examples of how comparisons reveal what is similar and different among humans and other species. In the next section we will examine how differences are identified among individual humans.

Phylogenetic inference

The first step in comparing species' genomes is to decide which species to compare. In order for comparisons to be informative, it is crucial to understand the evolutionary relationships among the species to be compared. The evolutionary history of a group is called an evolutionary tree, or a **phylogeny**. Phylogenies are useful because they allow us to infer how species' genomes have changed over time.

The second step in comparing genomes is the identification of the most closely related genes, called **homologous genes** (**Figure 14-17**). These genes can be recognized by similarities in their

DNA sequences and in the amino acid sequences of the proteins they encode. It is important to distinguish here two classes of homologous genes. Some homologs are genes at the same genetic locus in different species. These genes would have been inherited from a common ancestor and are referred to as **orthologs**. However, many homologous genes belong to families that have expanded (and contracted) in number in the course of evolution. These homologous genes are at different genetic loci in the same organism. They arose when genes within a genome were duplicated. Genes that are related by gene-duplication events in a genome are called **paralogs**. The history of gene families can be quite revealing about the evolutionary history of a group.

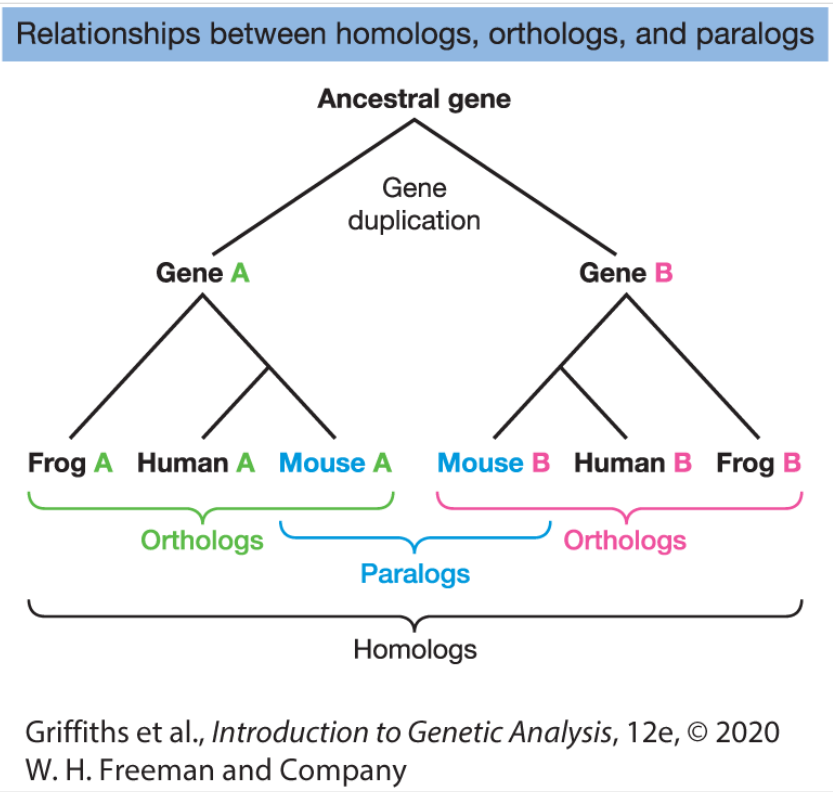


FIGURE 14-17 A gene in the common ancestor of a particular group of species (here frogs, mice, and humans) is duplicated, creating the A and the B genes, which are known as homologs. The A gene present in the frog genome is an ortholog of the A gene present in the mouse or human genome. Similarly, the B gene present in the frog genome is an ortholog of the B gene present in the mouse or human genome. The A gene present in the mouse genome is a paralog of the B gene in the mouse genome.

For example, suppose we would like to know how the mammalian genome has evolved over the history of the group. We would like to know whether mammals as a group might have acquired some unique genes, whether mammals with different lifestyles might possess different sets of genes, and what the fate was of genes that existed in mammalian ancestors.

Fortunately, we now have a large and rapidly expanding set of mammal genome sequences to compare that includes representatives of the three main branches of mammals—monotremes (for

example, platypus), marsupials (for example, wallaby, opossum), and eutherian mammals (for example, human, chimpanzee, dog, mouse). The relationships between these groups, some members within these groups, and other amniote vertebrates (amniotes are mostly land-dwelling vertebrates that have a terrestrially adapted egg) are shown in **Figure 14-18**.

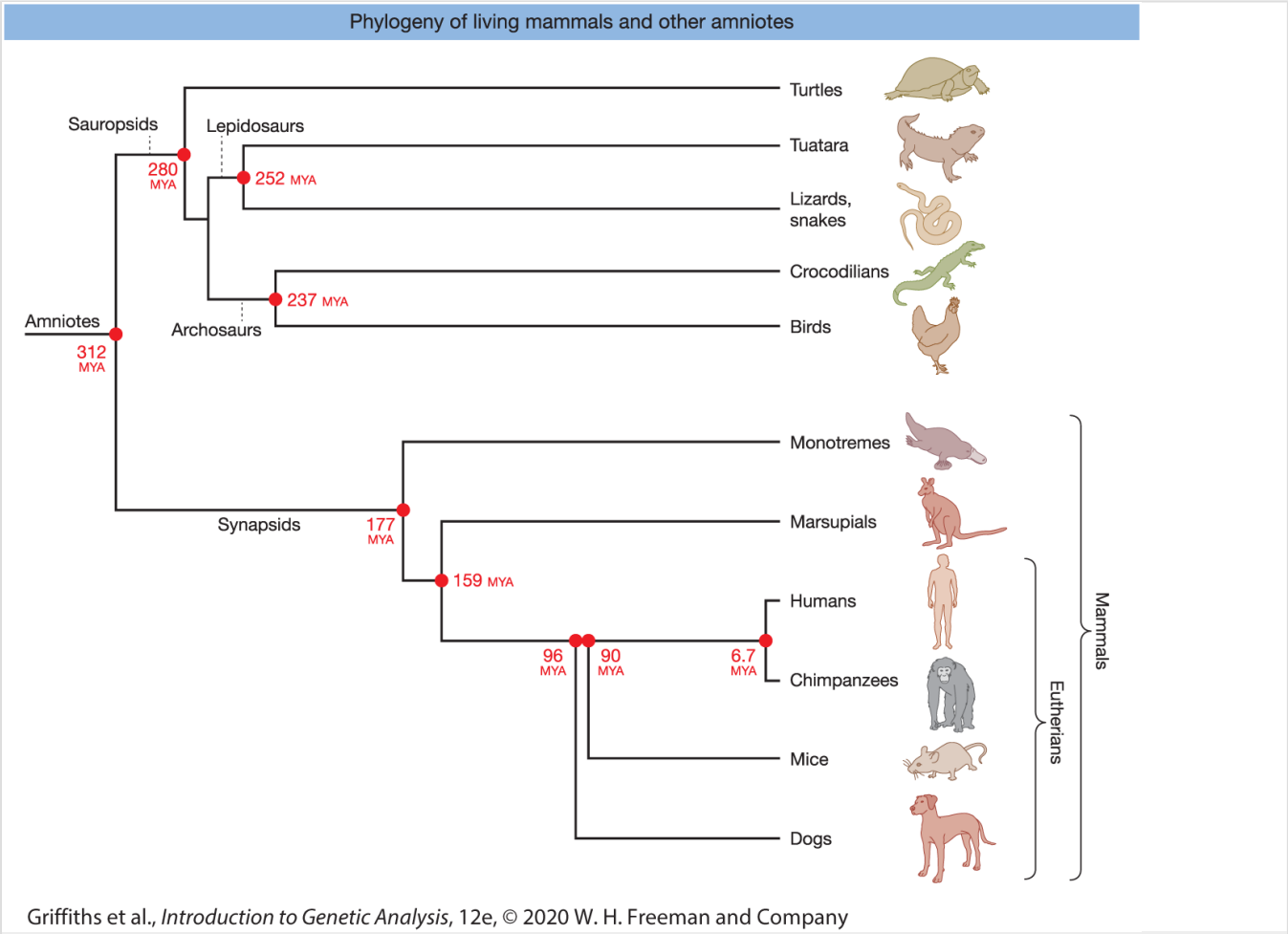


FIGURE 14-18 The phylogenetic tree depicts the evolutionary relationships among the three major groups of mammals (monotremes, marsupials, and eutherians) and other amniotes, including birds and various reptiles. By mapping the presence or absence of genes in particular groups onto known phylogenies, one can infer the direction of evolutionary change (gain or loss) in particular lineages.

INTERACTIVE RESOURCE SaplingPlus

Understanding evolutionary trees

To illustrate the importance of understanding phylogenies and how to utilize them, we consider the platypus genome. Monotremes differ from other mammals in that they lay eggs. Inspection of the platypus genome revealed that it contains one egg-yolk gene called vitellogenin. Analyses of marsupial and eutherian genomes revealed no such functional yolk genes. The presence of

vitellogenin in the platypus and its absence from other mammals could be explained in one of two ways: (1) vitellogenin is a novel invention of the platypus, or (2) vitellogenin existed in a common ancestor of monotremes, marsupials, and eutherians but was subsequently lost from marsupials and eutherians. The direction of evolutionary change is opposite in these two alternatives.

A simple pair-wise comparison between the platypus and another mammal does not distinguish between these alternatives. To do that, first we have to infer whether vitellogenin was likely to be present in the last common ancestor of the platypus, marsupials, and eutherians. We make this **phylogenetic inference** by examining whether vitellogenin is found in taxa outside of this entire group of mammals, what is referred to as an evolutionary **outgroup**. Indeed, three homologous vitellogenin genes exist in the chicken. Next, we consider the relationship of the chicken to mammals. Chickens belong to another major branch of the amniotes. Looking at the evolutionary tree in Figure 14-18, we can explain the presence of vitellogenins in chickens and the platypus as the result of two independent acquisitions (in the platypus lineage and the chicken lineage, respectively) or as the result of just one acquisition in a common ancestor of the platypus and chicken (which, based on the tree, would be a common ancestor of all amniotes) followed by the loss of vitellogenin genes in marsupials and eutherians.

How do we decide between these alternatives? When studying infrequent events such as the invention of a gene, evolutionary biologists prefer to rely on the principle of **parsimony**, that is, to favor the simplest explanation involving the smallest number of evolutionary changes. Therefore, the preferred explanation for the pattern of vitellogenin evolution in mammals is that this egg-yolk protein and corresponding gene were present in some egg-laying amniote ancestor and were retained in the egg-laying platypus and lost from non-egg-laying mammals.

As it turns out, there is one additional and very compelling piece of evidence that supports this inference. While inspection of eutherian genomes does not reveal any intact, functional vitellogenin genes, traces of vitellogenin gene sequences are detectable in the human and dog genomes at positions that are in the same position as (syntenic to) the vitellogenin genes of the platypus and chicken (**Figure 14-19**). These sequences are molecular relics of our egg-laying ancestors. As our mammalian ancestors shifted away from yolky eggs, natural selection was relaxed on the vitellogenin gene sequences such that they have been nearly eroded away by mutations over tens of millions of years. Our genome contains numerous relics of genes that once functioned in our ancestors, and as we will see again in this section, the identities of those pseudogenes reflect how human biology has diverged from that of our ancestors.

The human genome carries relics of our egg-laying ancestors

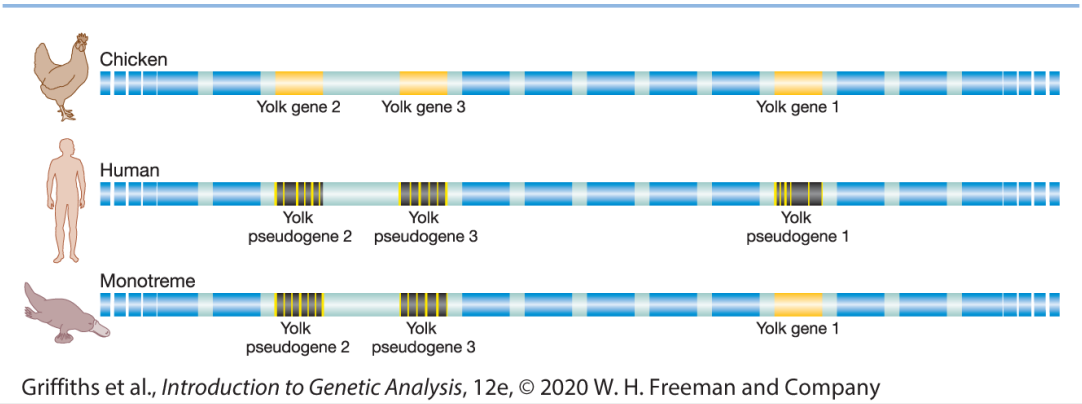


FIGURE 14-19 Strings of genes along chicken chromosome 8 and human chromosome 1 and in the platypus are in the same relative order (boxes). Whereas the chicken genome has three genes that encode egg-yolk proteins, the egg-laying platypus has one functional gene and two pseudogenes, and humans have fragmented, very short remnants of the yolk genes.

Of course, evolution is also about the acquisition of new traits. For example, milk production is a shared trait among all mammals. A family of genes encoding the casein milk proteins are unique to mammals and tightly clustered together in their genomes, including that of the platypus. Just this brief glance at a few mammalian genomes informs us that, indeed, some mammals have genes that others do not, some genes are shared by all mammals, and the presence or absence of certain genes correlates with mammals' lifestyle. The latter is a pervasive finding in comparative genomics.

KEY CONCEPT Determining which genomic elements have been gained or lost during evolution requires knowledge of the phylogeny of the species being compared. The presence or absence of genes often correlates with organism lifestyles.

Let's look at a few more examples that illuminate the evolutionary history of our genome and how we are different from, and similar to, other mammals.

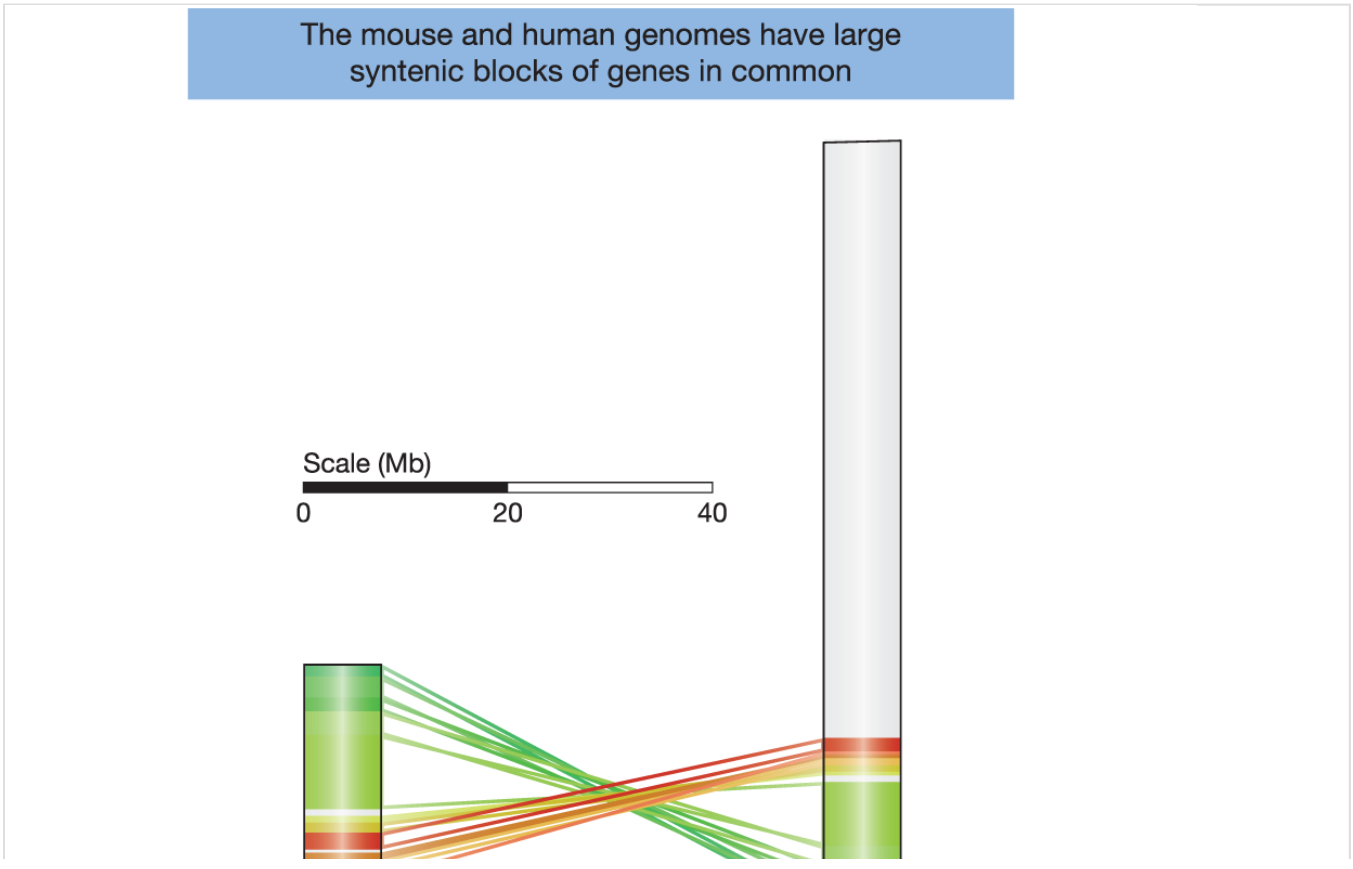
Of mice and humans

The sequence of the mouse genome has been particularly informative for understanding the human genome because of the mouse's long-standing role as a model genetic species, the vast knowledge of its classical genetics, and the mouse's evolutionary relationship to humans. The mouse and human lineages diverged approximately 90 million years ago, which is sufficient time

for mutations to cause their genomes to differ, on average, at about one of every two nucleotides. Thus, sequences common to the mouse and human genomes are likely to indicate common functions.

Homologs are identified because they have similar DNA sequences. Analysis of the mouse genome indicates that the number of protein-coding genes that it contains is similar to that of the human genome. Further inspection of the mouse genes reveals that at least 99 percent of all mouse genes have some homolog in the human genome and that at least 99 percent of all human genes have some homolog in the mouse genome. Thus, the kinds of proteins encoded in each genome are essentially the same. Furthermore, about 80 percent of all mouse and human genes are clearly identifiable orthologs.

The similarities between the genomes extend well beyond the inventory of protein-coding genes to overall genome organization. More than 90 percent of the mouse and human genomes can be partitioned into corresponding regions of conserved **synteny**, where the order of genes within variously sized blocks is the same as their order in the most recent common ancestor of the two species. This synteny is very helpful in relating the maps of the two genomes. For example, human chromosome 17 is orthologous to a single mouse chromosome (chromosome 11). Although there have been extensive intrachromosomal rearrangements in the human chromosome, there are 26 segments of collinear sequences more than 100 kb in size (**Figure 14-20**).



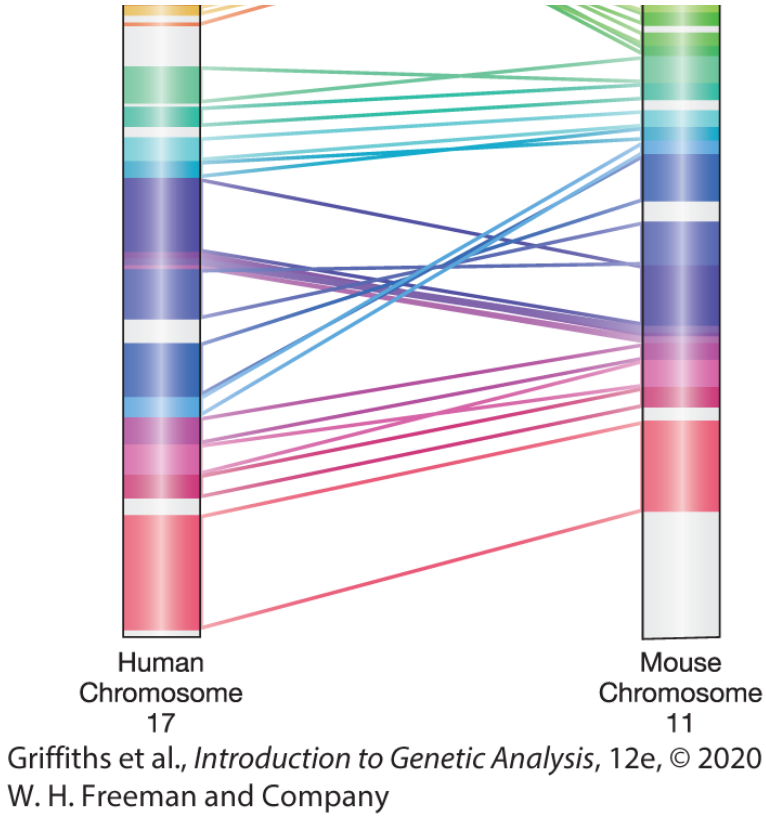


FIGURE 14-20 Synteny between human chromosome 17 and mouse chromosome 11. Large conserved syntenic blocks 100 kb or greater in size are shown between human chromosome 17 and mouse chromosome 11.

KEY CONCEPT The mouse and human genomes contain similar sets of genes, often arranged in similar order. This conserved gene order between species is known as synteny.

There are some detectable differences between the inventories of mouse and human genes. In one family of genes involved in color vision, the opsins, humans possess one additional paralog. The presence of this opsin has equipped humans with so-called trichromatic vision, so that we can perceive colors across the entire spectrum of visible light—violet, blue, green, red—whereas mice cannot. But again, the presence of this additional paralog in humans and its absence in mice does not alone tell us whether it was gained in the human lineage or lost in the mouse lineage. Analysis of other primate and mammalian genomes has revealed that Old World primates such as chimpanzees, gorillas, and the colobus monkey possess this gene, but that all nonprimate mammals lack it. We can safely infer from this phylogenetic distribution of the additional opsin gene that it evolved in an ancestor of Old World primates (that includes humans).

On the other hand, the mouse genome contains more functional copies of some genes that reflect its lifestyle. Mice have about 1400 genes involved in olfaction—this is the largest single functional category of genes in its genome. Dogs, too, have a large number of olfactory genes.

This certainly makes sense for the species' lifestyles. Mice and dogs rely heavily on their sense of smell, and they encounter different odors from those encountered by humans. And the set of human olfactory genes, compared to that of mice and dogs, is strikingly inferior. We have a lot of olfactory genes, but a very large fraction of them are pseudogenes that bear inactivating mutations. For example, in just one class of olfactory genes called *V1r* genes, mice have about 160 functional genes, but just 5 out of the 200 or so *V1r* genes in the human genome are functional.

Still, these differences in gene content are relatively modest in light of the vast differences in anatomy and behavior. The overall similarity in the mouse and human genomes corresponds to the picture we get from examining the genetic toolkit controlling development in different taxa (see Chapter 13)—that great differences can evolve from genomes containing similar sets of genes. This same theme is illustrated by comparing our genome with that of our closest living relative, the chimpanzee.

Comparative genomics of chimpanzees and humans

Chimpanzees and humans last had a common ancestor about 5–7 million years ago. Since that time, genetic differences have accumulated by mutations that have occurred in each lineage. Genome sequencing has revealed that there are about 35 million single-nucleotide differences between chimpanzees and humans, corresponding to about a 1.06 percent degree of divergence. In addition, about 5 million insertions and deletions, ranging in length from just a single nucleotide to more than 15 kb, contribute a total of about 90 Mb of divergent DNA sequence (about 3 percent of the overall genome). Most of these insertions or deletions lie outside of coding regions.

Overall, the proteins encoded by the human and chimpanzee genomes are extremely similar. Twenty-nine percent of all orthologous proteins are *identical* in sequence. Most proteins that differ do so by only about two amino acid replacements. There are a few detectable differences between chimpanzees and humans in the sets of functional genes. About 80 or so genes that were functional in their common ancestor are no longer functional in humans, owing to their deletion or to the accumulation of mutations. Some of these changes may contribute to differences in physiology.

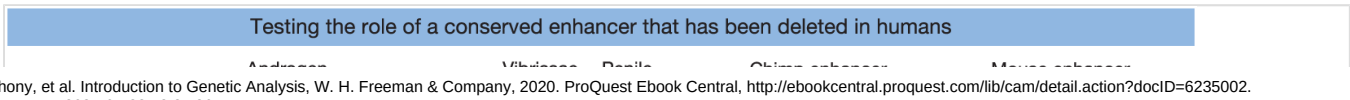
In addition to changes in particular genes, duplications of chromosome segments in a single lineage have contributed to genome divergence. More than 170 genes in the human genome and

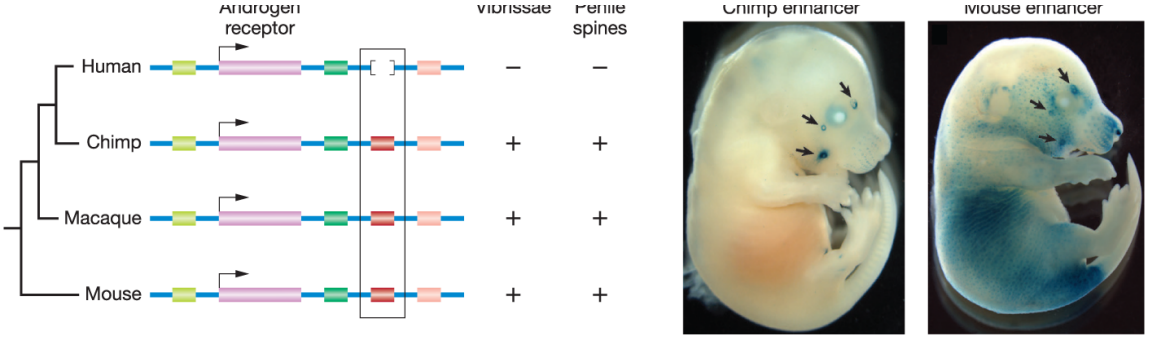
more than 90 genes in the chimpanzee genome are present in large duplicated segments. These duplications are responsible for a greater amount of the total genome divergence than all single-nucleotide mutations combined. Intriguingly, duplications unique to the human genome are enriched for genes that are predicted to play a role in brain development. It has been suggested that at least some of these gene duplications were involved in the expansion of the neocortex in humans relative to other primates. However, whether these duplicated genes contribute to major phenotypic differences between humans and our closest relatives is not yet clear.

Despite the existence of these few differences in gene content between chimpanzees and humans, we have seen that the vast majority of genes are highly conserved, with very few changes in protein-coding regions. How, then, can we explain the dramatic differences in morphology, behavior, and physiology between chimpanzees and humans? In 1975, well before the advent of whole-genome sequencing, Mary-Claire King and Allan Wilson boldly proposed that most of the phenotypic differences between humans and chimpanzees result from mutations that affect gene regulation. Comparative genomics has now provided a tool to identify regulatory mutations that might be responsible for the phenotypic differences between chimpanzees and humans.

KEY CONCEPT Great phenotypic differences can evolve from genomes containing similar sets of genes. Many of the phenotypic differences between species are likely due to genetic changes that affect gene regulation.

Here, we will discuss just one example of the many approaches used to identify putative cis-acting regulatory elements that differ between chimpanzees and humans. In this case, researchers searched for noncoding sequences that were highly conserved in the genomes of chimpanzee, macaque, and other mammals but missing in the human genome. There were 510 such deletions in the human genome, and these deletions were enriched near genes with neural function as well as steroid hormone signaling. One of these deletions is near the androgen receptor gene, which encodes a protein necessary for responses to circulating androgens such as testosterone. Using the previously introduced reporter gene assays in transgenic mice (see Chapter 10), the researchers showed that both the mouse and chimpanzee sequence drove expression in the developing sensory vibrissae (or whiskers) as well as the penile spines, which are both androgen-responsive structures that are present in most mammals but have been lost in humans (**Figure 14-21**). Testing the functions of other putative cis-acting regulatory elements missing in the human genome will likely uncover additional insight into the genetic changes that underlie differences between humans and our closest relatives.





Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company
Reprinted by permission from Macmillan Publishers Ltd. from McLean et al., "Human-specific loss of regulatory DNA and the evolution of human-specific traits," Nature, *2011, March 10; 471, 216–219, Figure 2. Permission conveyed through Copyright Clearance Center, Inc.*

FIGURE 14-21 The androgen receptor coding sequence (pink) is present in the human, chimp, macaque, and mouse genomes. Although some conserved, noncoding sequences near the androgen receptor gene are present in the genomes of all four species (yellow, green, and light pink rectangles), one conserved, noncoding sequence is present in the chimp, macaque, and mouse genomes but absent from the human genome (red rectangle). Using either the chimp or mouse sequence in a reporter gene assay in mice shows that this sequence is a cis-acting enhancer that drives expression in the sensory vibrissae (arrows) and penile spines (not shown).

KEY CONCEPT Genetic changes that underlie phenotypic differences between humans and our closest relatives can be identified using a combination of computational approaches and reporter gene assays.

Of course, all genetic differences between species originate as variations within species. The sequencing of the human genome and the advent of faster and less expensive high-throughput sequencing methods have opened the door to the detailed analysis of human genetic variation.

14.6 COMPARATIVE GENOMICS AND HUMAN MEDICINE

LO 14.4 Compare genomic methods used to identify mutations that have been associated with human disease thus far.

The human species, *Homo sapiens*, originated in Africa approximately 200,000 years ago. Sometime between 50,000 and 100,000 years ago, populations left Africa and migrated across the world, eventually populating five additional continents. These migrating populations encountered different climates, adopted different diets, and combated different pathogens in different parts of the world. Much of the recent evolutionary history of our species is recorded in our genomes, as are the genetic differences that make individuals or populations more or less susceptible to disease.

Overall, any two unrelated humans' genomes are 99.9 percent identical. That difference of just 0.1 percent still corresponds to roughly 3 million bases. The challenge today is to decipher which of those base differences are meaningful with respect to physiology, development, or disease.

Once the sequence of the first human genome was advanced, that accomplishment opened the door to much more rapid and less costly analysis of other individuals. The reason is that with a known genome assembly as a reference, it is much easier to align the raw sequence reads of additional individuals, and to design approaches to studying and comparing parts of the genome.

One of the first and greatest surprises that has emerged from comparing individual human genomes is that humans differ not merely at one base in a thousand, but also in the number of copies of parts of individual genes, entire genes, or sets of genes. These **copy number variations (CNVs)** include repeats and duplications that increase copy number and deletions that reduce copy number. Between any two unrelated individuals, there may be hundreds of segments of DNA greater than 1000 bp in length that differ in copy number. Some CNVs can be quite large and span up to 5 million base pairs. Together, CNVs account for more sequence variation among humans than all the 3 million single base pair changes combined.

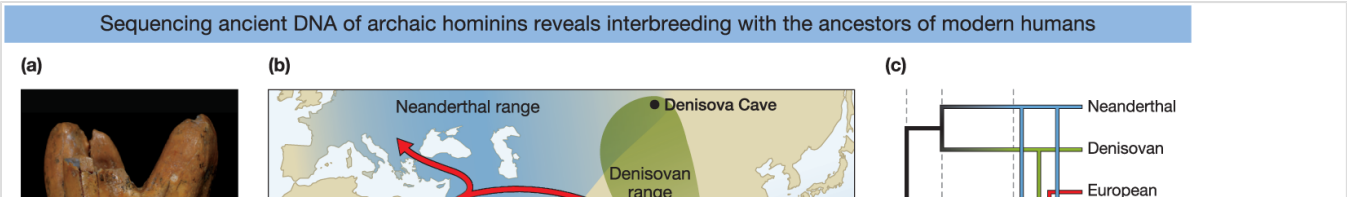
How such copy numbers may play a role in human evolution and disease is of intense interest. One case where increased copy number appears to have been adaptive concerns diet. People with high-starch diets have, on average, more copies of a salivary amylase (an enzyme that breaks

down starch) gene than people with traditionally low-starch diets. In other cases, copy number variations have been associated with human diseases. For example, it now appears that at least 15 percent of human neurodevelopmental diseases are due to changes in copy number that are found at a very low frequency in human populations. Copy number polymorphisms that are relatively common in human populations have also been associated with immune-related diseases such as Crohn's disease, psoriasis, and lupus.

The evolutionary history of human disease genes

One might ask when and where the mutations that cause human disease originated, and why some of these disease alleles are maintained at a relatively high frequency in human populations. Although we are still a long way from answering these questions, some insight has come from analyzing the genome sequences of ancient hominins, including those of our own species *Homo sapiens* as well as now-extinct, archaic hominin lineages like Neanderthals. Advances in sequencing and other technologies have made it possible to extract and sequence whole genomes from ancient DNA samples, even when very small amounts of tissue are found. For example, whole genome sequencing of ancient DNA from a single finger bone and three teeth found in the Denisova Cave in Siberia revealed the existence of an archaic hominin lineage, now called Denisovans, which are genetically very distinct from Neanderthals, diverging approximately 640,000 years ago.

Analyses of these archaic genomes has revealed that as anatomically modern humans left Africa and spread around the globe, they interbred with other hominin species that had already been living in Eurasia for over 200,000 years. Traces of these hybridization events can be seen in the genomes of humans living today. Reflecting the migratory paths of modern humans, all non-African individuals sequenced to date have between 1 percent and 4 percent Neanderthal ancestry, while indigenous Australians and Melanesians also have up to 6 percent Denisovan ancestry (**Figure 14-22**). Many direct-to-consumer genetic testing services will now report what percentage of a person's DNA has been inherited from their archaic human ancestors. **Box 14-1** discusses the various types of direct-to-consumer genetic testing options available today, as well as some important ethical and social implications of such services.



Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company
Courtesy of the Max Planck Institute for Evolutionary Anthropology.

FIGURE 14-22 (a) Side view of a molar found in the Denisova Cave in Siberia. (b) Anatomically modern humans migrated out of Africa into Eurasia through the ranges of both Neanderthals and Denisovans. (c) Genome sequencing provides evidence for early hybridization between Neanderthals and the ancestors of modern Melanesians, East Asians, and Europeans, as well as later hybridization between Neanderthals and the ancestors of modern East Asians (blue arrows). There is also evidence for hybridization between Denisovans and the ancestors of modern Melanesians (green arrow).

BOX 14-1
DIRECT-TO-CONSUMER GENETIC TESTING

The genomics revolution has also led to the democratization of access to personal genetic information. The Human Genome Project was started in part due to the promise of **personal genomics**, exemplified by the case of Nicholas Volker at the beginning of this chapter. Thus, shortly after the first draft of the human genome was completed, a number of so-called "direct-to-consumer genetic testing" companies began to emerge with the goal of fulfilling that promise. Currently, a handful of companies offer direct-to-consumer genetic testing. For the cost of approximately $100 to $200, a consumer can provide a saliva sample or a cheek swab to a company. Their DNA will then be genotyped at roughly 700,000 of the 3 million sites in the genome that are known to vary among humans. The consumer can then retrieve the results of their genome analyses via a Web site or an app. Currently, the services provided by these companies fall into three main categories: medical testing, genetic genealogy, and personal ancestry.

Medical testing The first direct-to-consumer genetic testing companies popped up quickly around 2005–2006, promising to provide individuals with personalized information about their genetic risk for common diseases like diabetes or cancer. Just as quickly, concerns over these tests emerged. For example, it was unknown whether consumers would understand their personal genetic risks without the help of a health care professional or whether they would stop taking preventative health measures based on the results of these tests. Furthermore, there were concerns over privacy and the potential for misuse of the data. Based on these and other concerns, at the end of 2013 the U.S. Food and Drug Administration (FDA) served "cease and desist" letters to these companies, requiring them to obtain FDA authorization for their tests. As of early 2018, only one company, 23andMe, has been authorized by the FDA to provide direct-to-consumer testing for genetic risk factors associated with a limited number of diseases, such as breast cancer, Parkinson's, and Alzheimer's. Consumers can currently also use their services to assess carrier status for over 40 inherited diseases, such as cystic fibrosis and sickle cell anemia.

Genetic genealogy The second most common hobby in the United States is genealogy, or the tracing of family lineages and history. This popularity is reflected in the fact that the companies that offer direct-to-consumer genetic testing for the purpose of genealogy have now collected genetic data for over 15 million people. When an individual submits their DNA sample, their relationship to every other individual in the database is estimated

from genetic data. For example, if that individual had a monozygotic twin in the database, they would show up as a perfect match, while a parent, child, or full sibling would show up as a first-degree relative. Most matches in the database comprise second, third, or fourth cousins. Customers can use this information to identify and contact possible relatives to fill in their family tree. Of course, these genetic matches may reveal unexpected relatives or relationships that were not previously known, and customers must be aware of the repercussions, both positive and negative, of this knowledge. The International Society of Genetic Genealogy has compiled a chart comparing features of the top five companies that offer autosomal DNA testing; this can be found at https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart.

Personal ancestry Where did we come from? Humans have been asking this universal question for millennia. Direct-to-consumer genetic testing promises to answer this question by providing consumers with information about their genetic ancestry, including the percentage of ancestry derived from archaic humans such as Neanderthals and Denisovans. It is important to note, however, that the ability to assign ancestry is reliant upon the other data in the database. For example, if the database of a company is comprised mostly of people of European descent, it is more difficult to determine the ancestry of a person of Asian or African descent. Thus, the results provided by any one company about an individual's ancestry should be interpreted as a rough estimate that is likely to evolve over time as more and more people decide to submit their own DNA samples for ancestry testing.

Ethical, legal, and social implications The ethical, legal, and social implications of direct-to-consumer genetic testing are far-reaching and need to be carefully considered. Thus, the future of direct-to-consumer genetic testing is not currently clear. However, discussions among a wide variety of stakeholders, including geneticists, ethicists, medical providers, companies, regulators, and consumers, are ongoing to ensure that the avalanche of personal genetic information that is now upon us is used for the benefit of individuals and society.

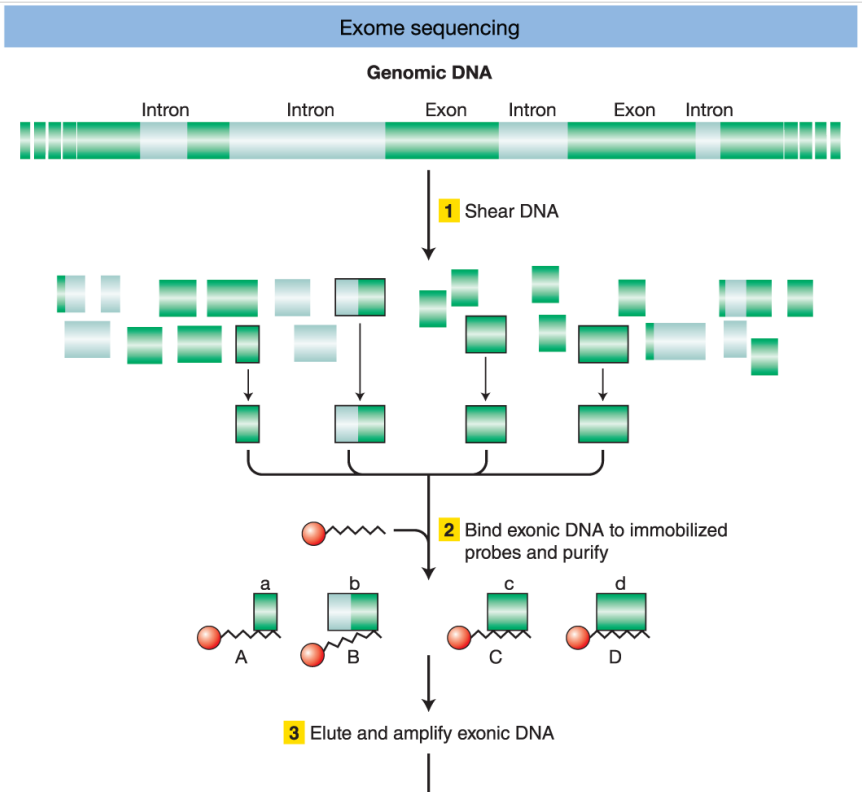
Remarkably, some of the Neanderthal and Denisovan alleles present in modern humans have effects on physiology. For example, gene variants that cause lighter skin in northern Eurasians were present in Neanderthals, and one of the gene variants that has enabled Tibetans to live at high altitudes (see Chapter 1) is Denisovan in origin. However, Neanderthal-derived alleles of some genes are associated with the risk of diseases in modern humans. For example, modern humans with a Neanderthal allele at a gene involved in blood coagulation have a higher risk of blood clots and stroke. Rapid clotting might have been an advantage in early hominids, who were hunting dangerous animals and at risk of excessive bleeding during childbirth. However, in our modern times, these risks are lessened, and humans live much longer. Thus, fast clotting is no longer an advantage and leads to an increased risk of stroke and blood clots. Genetic variants of Neanderthal origin are also linked to increased risk of neurological, immunological, and skin diseases in modern humans. These examples serve as a reminder that our genetic susceptibility to disease has been shaped by our evolutionary history. Genomics has provided a tool that allows us to explore this evolutionary history in ways that could not be imagined before.

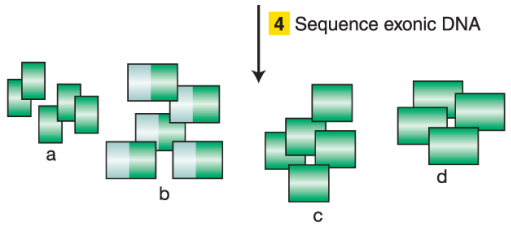
KEY CONCEPT The ability to sequence whole genomes of modern and archaic humans provides a tool to uncover the evolutionary history of humans and to identify mutations associated with disease.

The exome and personalized genomics

Advances in sequencing technologies have reduced the cost of sequencing individual genomes from about $300 million in 2000, to $1 million in 2008, to about $1000 in 2017. But for many large-scale studies, that figure is still prohibitive. For some applications, it is more practical and cost effective, and can be just as informative, to sequence only part of the genome. For example, since many disease-causing mutations occur in coding sequences, strategies have been designed to sequence all of the exons, or the **exome**, of individuals, as was done in the case of Nicholas Volker.

The strategy for exome sequencing involves generating a library of genomic DNA that is enriched for exon sequences (**Figure 14-23**). The DNA is prepared by (1) shearing genomic DNA into short, single-stranded pieces, (2) hybridizing the single-stranded pieces to biotin-labeled probes complementary to the known exonic regions of the human genome and purifying the biotin-labeled duplexes, (3) amplifying the exon-rich duplexes, and (4) sequencing the exon-rich duplexes. In this manner, 30–60 megabases of the human genome is targeted for sequencing, as opposed to the 3200 megabases of total sequence.





Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

FIGURE 14-23 In order to sequence just the exon fraction of the genome, genomic DNA is fragmented and denatured, and exon-containing fragments are hybridized with biotin-labeled probes that are complementary to the known exon sequences in the genome. Duplexes containing annealed probes are then purified and prepared for sequencing.

As of late 2017, the exomes of more than 120,000 individuals have been sequenced, at the current cost of only a few hundred dollars per exome. One particularly important power of exome sequencing is to identify de novo mutations in individuals (mutations that are not present in either parent). Such mutations are responsible for many spontaneously appearing genetic diseases whose origins would not be revealed by traditional pedigree-based studies. As such, whole-exome sequencing is now a rapidly spreading clinical diagnostic tool, particularly for neurodevelopmental and other disorders in pediatric populations.

And just as exome sequencing can be used to identify genetic differences between individuals, it can also be used to identify differences between normal and abnormal cells, such as cancer cells. Cancer is a suite of genetic diseases in which combinations of gene mutations typically contribute to the loss of growth control and metastasis. Understanding what genetic changes are common to particular cancers, or to subsets of cancers, will not only further our understanding of cancer, but also promises to impact diagnosis and treatment in powerful ways. Researchers across the world have recently completed an “atlas” of cancer genomes that has uncovered the extraordinary genetic heterogeneity present in cancer cells and provided a framework for classifying tumor subtypes based on the underlying genomic alterations. This knowledge opens up new opportunities to develop therapies that specifically target the genetic changes found in a particular tumor rather than treating cancer as a homogeneous disease. (See http://cancergenome.nih.gov/ for further information.)

KEY CONCEPT Exome sequencing is a powerful approach to cheaply and rapidly identify mutations associated with human disease.

The ability to rapidly analyze organisms’ genomes is also impacting other dimensions of medicine. We will look at one such case next.

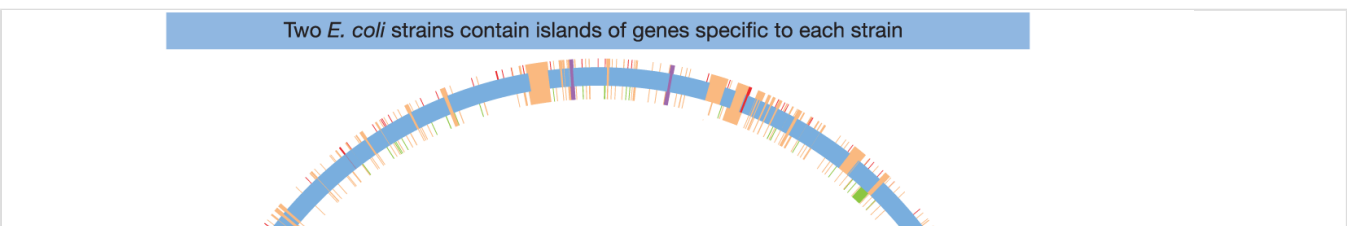
Comparative genomics of nonpathogenic and pathogenic *E. coli*

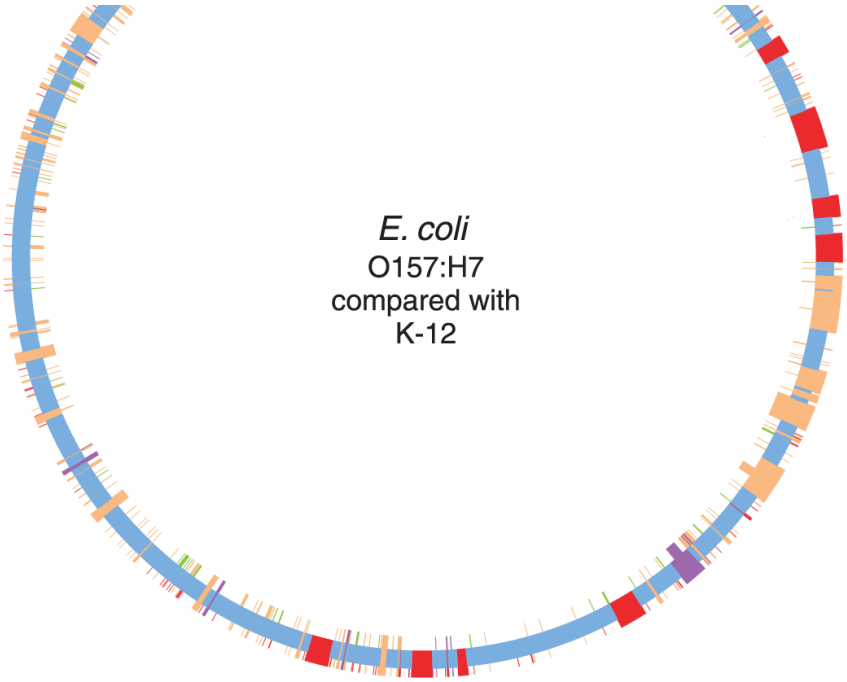
Escherichia coli are found in our mouths and intestinal tracts in vast numbers, and this species is generally a benign symbiont. Because of its central role in genetics research, it was one of the first bacterial genomes sequenced. The *E. coli* genome is about 4.6 Mb in size and contains 4405 genes. However, calling it "the *E. coli* genome" is really not accurate. The first genome sequenced was derived from the common laboratory *E. coli* strain K-12. Many other *E. coli* strains exist, including several important to human health.

In 1982, a multistate outbreak of human disease was traced to the consumption of undercooked ground beef. The *E. coli* strain O157:H7 was identified as the culprit, and it has since been associated with a number of large-scale outbreaks of infection. In fact, there are an estimated 75,000 cases of *E. coli* infection annually in the United States. Although most people recover from the infection, a fraction develop hemolytic uremia syndrome, a potentially life-threatening kidney disease.

To understand the genetic bases of pathogenicity, the genome of an *E. coli* O157:H7 strain has been sequenced. The O157 and K-12 strains have a backbone of 3574 protein-coding genes in common, and the average nucleotide identity among orthologous genes is 98.4 percent, comparable to that of human and chimpanzee orthologs. About 25 percent of the *E. coli* orthologs encode identical proteins, similar to the 29 percent for human and chimpanzee orthologs.

Despite the similarities in many proteins, the genomes and proteomes differ enormously in content. The *E. coli* O157 genome encodes 5416 genes, whereas the *E. coli* K-12 genome encodes 4405 genes. The *E. coli* O157 genome contains 1387 genes that are not found in the K-12 genome, and the K-12 genome contains 528 genes not found in the O157 genome. Comparison of the genome maps reveals that the backbones common to the two strains are interspersed with islands of genes specific to either K-12 or O157 (**Figure 14-24**).





Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

FIGURE 14-24 The circular genome maps of *E. coli* strains K-12 and O157:H7. The circle depicts the distribution of sequences specific to each strain. The colinear backbone common to both strains is shown in blue. The positions of O157:H7-specific sequences are shown in red. The positions of K-12-specific sequences are shown in green. The positions of O157:H7- and K-12-specific sequences at the same location are shown in tan. Hypervariable sequences are shown in purple. [*Data from N. T. Perna et al., "Genome Sequence of Enterohaemorrhagic Escherichia coli O157:H7," Nature 409, 2001, 529–533. Courtesy of Guy Plunkett III and Frederick Blattner.*]

Among the 1387 genes specific to *E. coli* O157 are many genes that are suspected to encode virulence factors, including toxins, cell-invasion proteins, adherence proteins, and secretion systems for toxins, as well as possible metabolic genes that may be required for nutrient transport, antibiotic resistance, and other activities that may confer the ability to survive in different hosts. Most of these genes were not known before sequencing and would not be known today had researchers relied solely on *E. coli* K-12 as a guide to all *E. coli*.

The surprising level of diversity between two members of the same species shows how dynamic genome evolution can be. Most new genes in *E. coli* strains are thought to have been introduced by horizontal transfer from the genomes of viruses and other bacteria (see Chapter 6). Differences can also evolve owing to gene deletion. Other pathogenic *E. coli* and bacterial species also exhibit many differences in gene content from their nonpathogenic cousins. The identification of genes that may contribute directly to pathogenicity opens new avenues to the understanding, prevention, and treatment of infectious disease.

14.7 FUNCTIONAL GENOMICS AND REVERSE GENETICS

LO 14.2 Explain the role of various functional elements within genomes, and differentiate between computational and experimental methods used to identify these elements.
LO 14.5 Outline reverse genetic approaches to analyze the function of genes and genetic elements identified by genome sequencing and comparative genomics.

Geneticists have been studying the expression and interactions of individual gene products for the past several decades. With the advent of genomics, we have an opportunity to expand these studies to a global level by using genome-wide approaches to study most or all gene products systematically and simultaneously, and in species that are not previously established experimental models (see the *Beyond Model Organisms* section of *A Brief Guide to Model Organisms*, at the back of this book). This global approach to the study of the function, expression, and interaction of gene products is termed **functional genomics**.

" 'Omics"

In addition to the genome, other global data sets are of interest. Following the example of the term *genome*, for which "gene" plus "-ome" becomes a word for "all genes," genomics researchers have coined a number of terms to describe other global data sets on which they are working. This *'ome* wish list includes

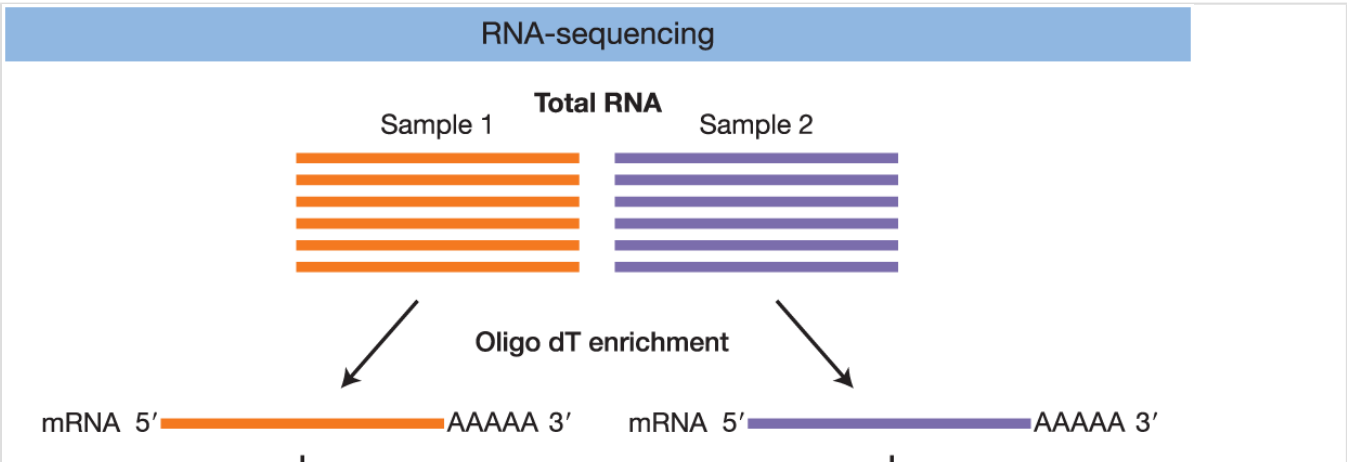
The **transcriptome**. The sequence and expression patterns of all RNA transcripts (which kinds, where in tissues, when, how much).
The **proteome**. The sequence and expression patterns of all proteins (where, when, how much).
The **interactome**. The complete set of physical interactions between proteins and DNA segments, between proteins and RNA segments, and between proteins.

We will not consider all of these *'omes* in this section but will focus on some of the global techniques that are beginning to be exploited to obtain these data sets.

Using RNA-seq to study the transcriptome

Suppose we want to answer the question, what genes are active in a particular cell under certain conditions? Those conditions could be one or more stages in development, or they could be the presence or absence of a pathogen or a hormone. Active genes are transcribed into RNA, and so the set of RNA transcripts present in the cell can tell us what genes are active. Here, the application of next-generation sequencing technologies has been extremely powerful by permitting the assay of RNA transcripts for all genes simultaneously in a single experiment. Let's see how this process works in more detail.

The first step is to isolate the total set of RNA molecules from cells of interest. For example, one set might be extracted from a particular cell type grown under typical conditions. A second set might be made from RNA extracted from cells grown under some experimental condition. Although methods exist for capturing and sequencing different types of RNAs in the cell, we will focus here on the sequencing of mRNA, which is the fraction of the RNA that encodes proteins. The mRNA can be captured from total RNA using an oligo-dT primer, which is complementary to the 3′ poly(A) tail of the mRNA. Then, the mRNA is subjected to reverse transcription to transform it into cDNA (see Chapter 10), which can then be used as a substrate for next-generation sequencing libraries just as for genomic DNA (see Figure 14-6). The sequencing reads are then mapped to the genome, where they align to the transcribed regions of genes. The number of reads present for a particular transcript should reflect its levels of expression in the cell; genes expressed at a low level in a particular cell type will have few reads, and genes expressed at a high level in a particular cell type will have many reads (**Figure 14-25**). In this manner, genes whose levels of expression are increased or decreased under the given experimental condition are identified. Similarly, genes that are active in a given cell type or at a given stage of development can be identified.



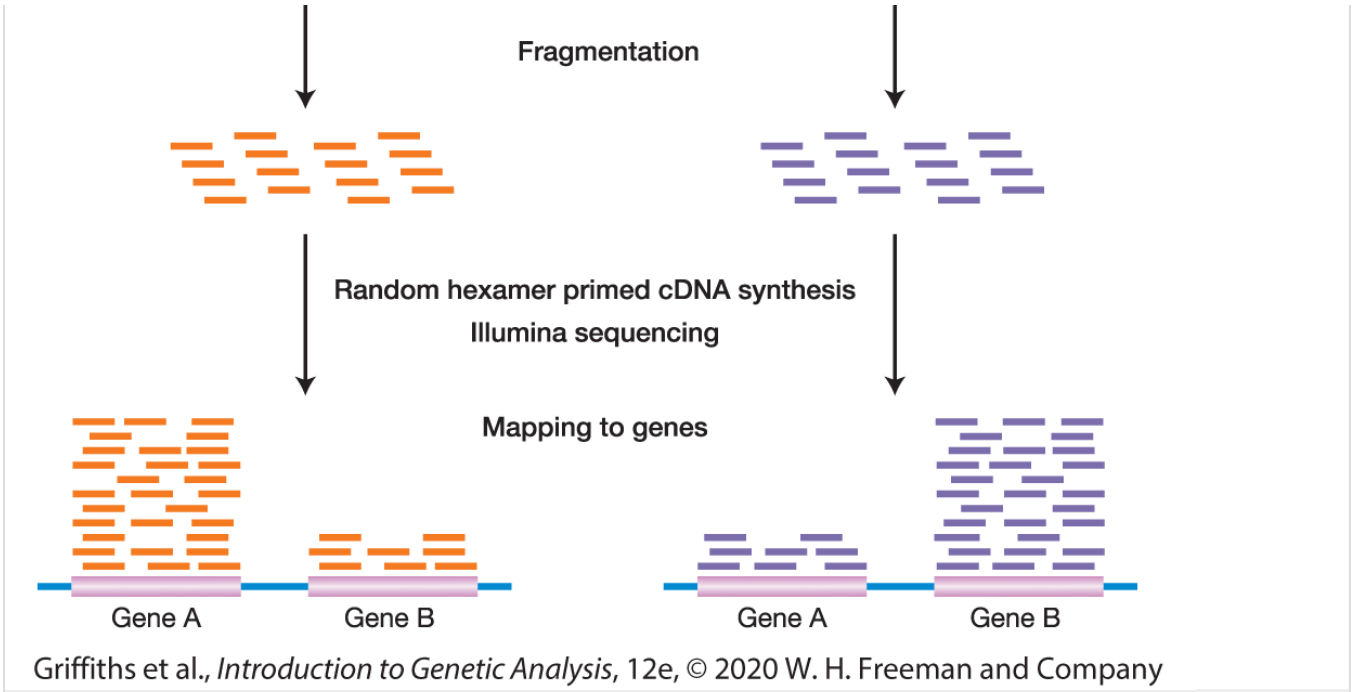


FIGURE 14-25 Total RNA is isolated from cells in two different conditions, followed by mRNA enrichment and cDNA synthesis. The resulting cDNA is sequenced using a next-generation sequencing method. The resulting sequencing reads are aligned to the exonic sequences in the genome, and the number of reads mapping to genes in the different conditions is compared.

ANIMATED ART SaplingPlus

RNA-seq

With an understanding of which genes are active or inactive at a given developmental stage, in a particular cell type, or in various environmental conditions, the sets of genes that may respond to similar regulatory inputs can be identified. Furthermore, gene-expression profiles can paint a picture of the differences between normal and diseased cells. By identifying genes whose expression is altered by mutations, in cancer cells, or by a pathogen, researchers may be able to devise new therapeutic strategies.

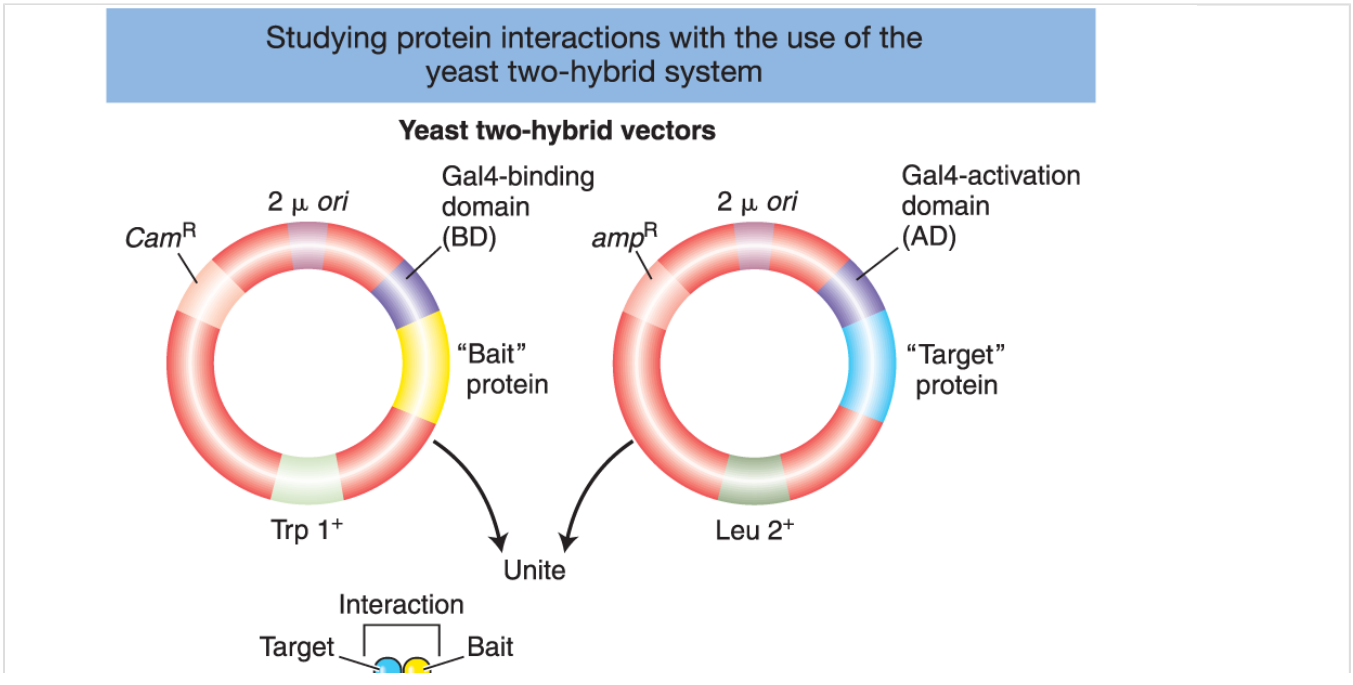
Using the two-hybrid test to study the protein–protein interactome

One of the most important activities of proteins is their interaction with other proteins. Because of the large number of proteins in any cell, biologists have sought ways of systematically studying all of the interactions of individual proteins in a cell. One of the most common ways of studying the interactome uses an engineered system in yeast cells called the **two-hybrid test**, which detects

physical interactions between two proteins. The basis for the test is the transcriptional activator encoded by the yeast *GAL4* gene (see Chapter 12).

Recall that this protein has two domains: (1) a DNA-binding domain that binds to the transcriptional start site and (2) an activation domain that will activate transcription but cannot itself bind to DNA. Thus, the two domains must be in close proximity in order for transcriptional activation to take place. Suppose that you are investigating whether two proteins interact. The strategy of the two-hybrid system is to separate the two domains of the activator encoded by *GAL4*, making activation of a reporter gene impossible. Each domain is connected to a different protein. If the two proteins interact, they will join the two domains together. The activator will become active and start transcription of the reporter gene.

How is this scheme implemented in practice? The *GAL4* gene is divided between two plasmids so that one plasmid contains the part encoding the DNA-binding domain and the other plasmid contains the part encoding the activation domain. On one plasmid, a gene for one protein under investigation is spliced next to the DNA-binding domain, and this fusion protein acts as "bait." On the other plasmid, a gene for another protein under investigation is spliced next to the activation domain, and this fusion protein is said to be the "target" (**Figure 14-26**). The two hybrid plasmids are then introduced into the same yeast cell—perhaps by mating haploid cells containing bait and target plasmids. The final step is to look for activation of transcription by a *GAL4*-regulated reporter gene construct, which would be proof that bait and target bind to each other. The two-hybrid system can be automated to make it possible to hunt for protein interactions throughout the proteome.



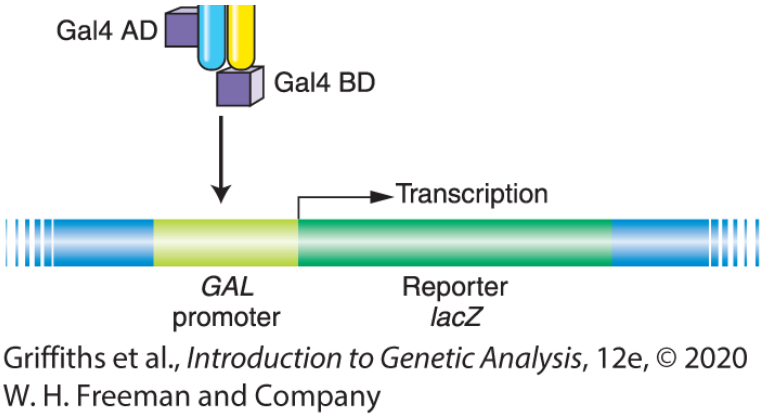


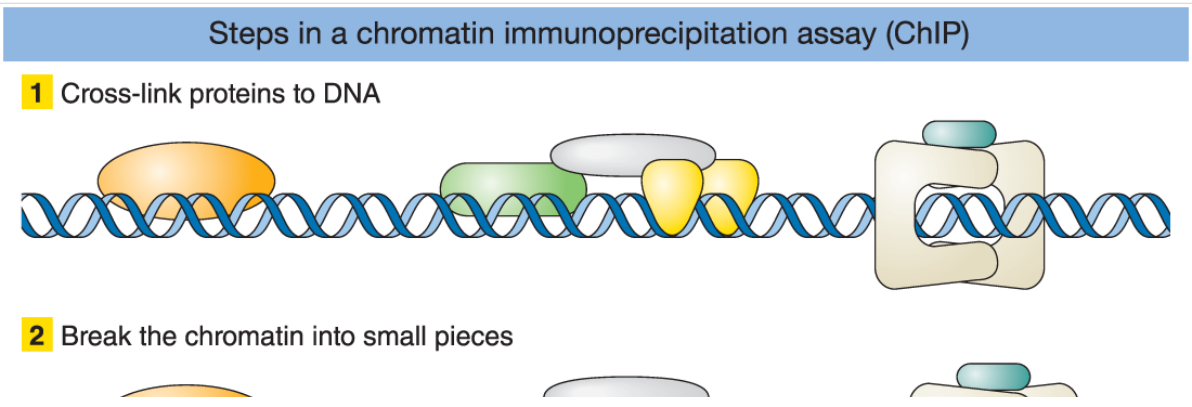
FIGURE 14-26 The system uses the binding of two proteins, a "bait" protein and a "target" protein, to restore the function of the Gal4 protein, which activates a reporter gene. *Cam*, Trp, and Leu are components of the selection systems for moving the plasmids around between cells. The reporter gene is *lacZ*, which resides on a yeast chromosome (shown in blue).

ANIMATED ART SaplingPlus

Yeast two-hybrid systems

Studying the protein–DNA interactome using chromatin immunoprecipitation assay (ChIP)

The sequence-specific binding of proteins to DNA is critical for correct gene expression. For example, regulatory proteins bind to promoters and activate or repress transcription in both bacteria and eukaryotes (see Chapters 11, 12, and 13). In the case of eukaryotes, chromosomes are organized into chromatin, in which the fundamental unit, the nucleosome, contains DNA wrapped around histones. Post-translational modification of histones often dictates what proteins bind and where (see Chapter 12). A variety of technologies have been developed that allow researchers to isolate specific regions of chromatin so that DNA and its associated proteins can be analyzed together. The most widely used method is called **ChIP** (for **chromatin immunoprecipitation**), and its application is described below (**Figure 14-27**).



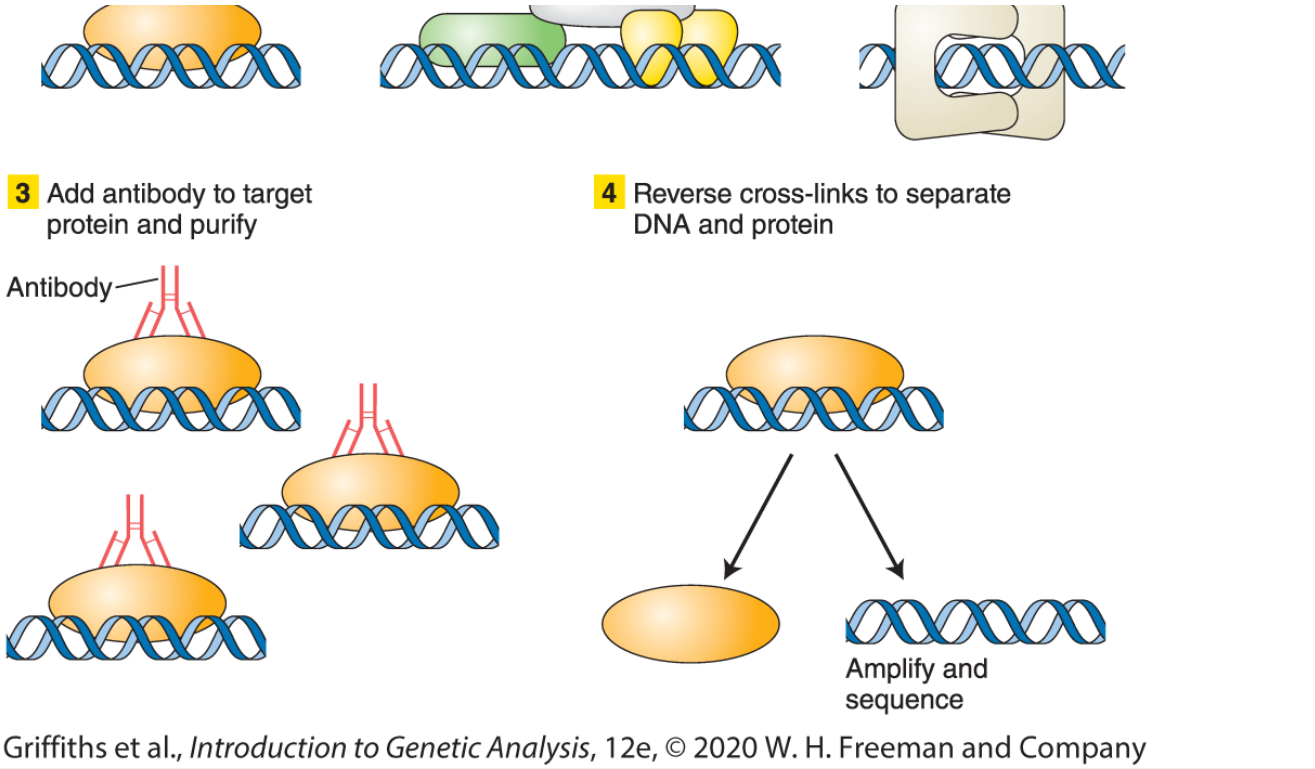


FIGURE 14-27 ChIP is a technique for isolating the DNA and its associated proteins in a specific region of chromatin so that both can be analyzed together.

1 **ANIMATED ART** SaplingPlus

ChIP

Let’s say that you have isolated a gene from yeast and suspect that it encodes a protein that binds to DNA when yeast is grown at high temperature. You want to know whether this protein binds to DNA and, if so, to what yeast sequence. One way to address this question is first to treat yeast cells that have been grown at high temperature with a chemical that will cross-link proteins to the DNA. In this way proteins bound to the DNA at the time of chromatin isolation will remain bound through subsequent treatments. The next step is to break the chromatin into small pieces. To separate the fragment containing your protein–DNA complex from others, you use an antibody that reacts specifically with the encoded protein. You add your antibody to the mixture so that it forms an immune complex that can be purified. The DNA bound in the immune complex can be analyzed after cross-linking is reversed. DNA bound by the protein may be amplified into many copies by PCR to prepare for DNA sequencing, or the DNA may be sequenced directly.

As we saw in Chapter 12, regulatory proteins often activate transcription of many genes simultaneously by binding to several promoter regions. A variation of the ChIP procedure, called **ChIP-seq**, has been devised to identify all the binding sites of a protein in a sequenced genome.

Proteins that bind to many genomic regions are immunoprecipitated as described previously. Then, after cross-linking is reversed, the DNA fragments are subjected to DNA sequencing using a next-generation method such as Illumina sequencing. The sequencing reads are mapped to the genome to reveal the locations where the regulatory protein binds in a particular cell type, environmental condition, or disease state.

KEY CONCEPT Advances in genomic technologies have made it possible to catalog the transcripts and proteins as well and protein–DNA and protein–protein interactions found in normal and diseased cells.

Reverse genetics

The kinds of data obtained from RNA-seq, ChIP-seq, and protein-interaction screens are suggestive of interactions within the genome and proteome, but they do not allow one to draw firm conclusions about gene functions and interactions in vivo. For example, finding out that the expression of certain genes is lost in some cancers is not proof of cause and effect. The gold standard for establishing the function of a gene or genetic element is to disrupt its function and to understand phenotypes in native conditions. Starting from available gene sequences, researchers can now use a variety of methods to disrupt the function of a specific gene. These methods are referred to as reverse genetics. Reverse-genetic analysis starts with a known molecule—a DNA sequence, an mRNA, or a protein—and then attempts to disrupt this molecule to assess the role of the normal gene product in the biology of the organism (see Figure 14-2).

There are several approaches to reverse genetics, and new technologies are constantly being developed and refined. One approach is to introduce random mutations into the genome, but then to hone in on the gene of interest by molecular identification of mutations in the gene. A second approach is to conduct a targeted mutagenesis that produces mutations directly in the gene of interest. A third approach is to create *phenocopies*—effects comparable to mutant phenotypes—usually by treatment with agents that interfere with the mRNA transcript of the gene.

Each approach has its advantages. Random mutagenesis is well established, but it requires that one sift through all the mutations to find those that include the gene of interest. Targeted mutagenesis can also be labor intensive, but, after the targeted mutation has been obtained, its characterization is more straightforward. Creating phenocopies can be very efficient, especially as

libraries of tools have been developed for particular model species. The technical details of these methods are covered in Chapters 8 and 10, so we will here consider examples of each of these approaches.

Reverse genetics through random mutagenesis

Random mutagenesis for reverse genetics employs the same kinds of general mutagens that are used for forward genetics: chemical agents, radiation, or transposable genetic elements (see Figure 6-38). However, instead of screening the genome at large for mutations that exert a particular phenotypic effect, reverse genetics focuses on the gene in question, which can be done in one of two general ways.

One approach is to focus on the map location of the gene. Only mutations falling in the region of the genome where the gene is located are retained for further detailed molecular analysis. Thus, in this approach, the recovered mutations must be mapped. One straightforward way is to cross a new mutant with a mutant containing a known deletion or mutation of the gene of interest (see Figure 17-21). Only the pairings that result in progeny with a mutant phenotype (showing lack of complementation) are saved for study.

In another approach, the gene of interest is identified in the mutagenized genome and checked for the presence of mutations. For example, if the mutagen causes small deletions, then, after PCR amplification of gene fragments, genes from the parental and mutagenized genomes can be compared, looking for a mutagenized genome in which the gene of interest is reduced in size. Similarly, transposable-element insertions into the gene of interest can be readily detected because they increase its size. As the ability to rapidly and cheaply sequence whole genomes improves, it is becoming feasible to identify mutations in genes of interest, including single–base-pair substitutions, by simply sequencing the parental and mutagenized genomes. In these ways, a set of genomes containing random mutations can be effectively screened to identify the small fraction of mutations in a gene of interest to a researcher.

Reverse genetics by targeted mutagenesis

For most of the twentieth century, researchers viewed the ability to direct mutations to a specific gene as the unattainable “holy grail” of genetics. However, now several such techniques are available. After a gene has been inactivated in an individual, geneticists can evaluate the

phenotype exhibited for clues to the gene's function. While the tools for targeted gene mutations were first developed using genetic techniques for model organisms, new technologies, particularly those that are CRISPR-based (see Chapter 10), are revolutionizing the ability to disrupt and manipulate genes in both model and nonmodel species.

Gene-specific mutagenesis usually requires the replacement of a resident wild-type copy of an entire gene by a mutated version of that gene. The mutated gene inserts into the chromosome by a mechanism resembling homologous recombination, replacing the normal sequence with the mutant (**Figure 14-28**). This approach can be used for targeted gene knockout, in which a null allele replaces the wild-type copy. Some techniques are so efficient that in *E. coli* and *S. cerevisiae* , for example, it has been possible to mutate every gene in the genome to try to ascertain its biological function.

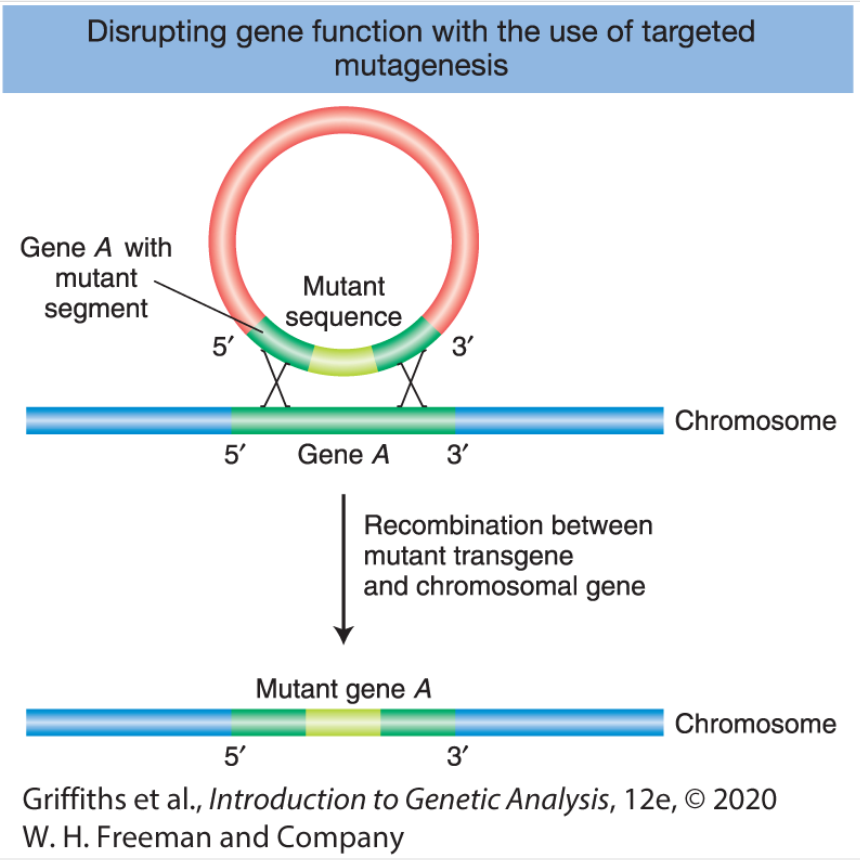


FIGURE 14-28 The basic molecular event in targeted gene replacement. A transgene containing sequences from two ends of a gene but with a selectable segment of DNA in between is introduced into a cell. Double recombination between the transgene and a normal chromosomal gene produces a recombinant chromosomal gene that has incorporated the abnormal segment.

KEY CONCEPT Targeted mutagenesis is the most precise means of obtaining mutations in a specific gene and can now be practiced in a variety of model systems, including mice and flies.

Reverse genetics by phenocopying

The advantage of inactivating a gene itself is that mutations will be passed on from one generation to the next, and so, once obtained, a line of mutants is always available for future study. On the other hand, phenocopying can be applied to a great many organisms regardless of how well developed the genetic technology is for a given species.

One of the most exciting discoveries of the past decade or so has been the discovery of a widespread mechanism whose natural function seems to be to protect a cell from foreign DNA. This mechanism is called **RNA interference (RNAi)**, described in Chapter 8. Researchers have capitalized on this cellular mechanism to make a powerful method for inactivating specific genes. The inactivation is achieved as follows. A double-stranded RNA is made with sequences homologous to part of the gene under study and is introduced into a cell (**Figure 14-29**). The RNA-induced silencing complex, or RISC, then degrades native mRNA that is complementary to the double-stranded RNA. The net result is a complete or considerable reduction of mRNA levels that lasts for hours or days, thereby nullifying expression of that gene. Because the RISC complex is found in most eukaryotes, the technique has been widely applied in model systems such as *C. elegans*, *Drosophila*, zebrafish, and several plant species.

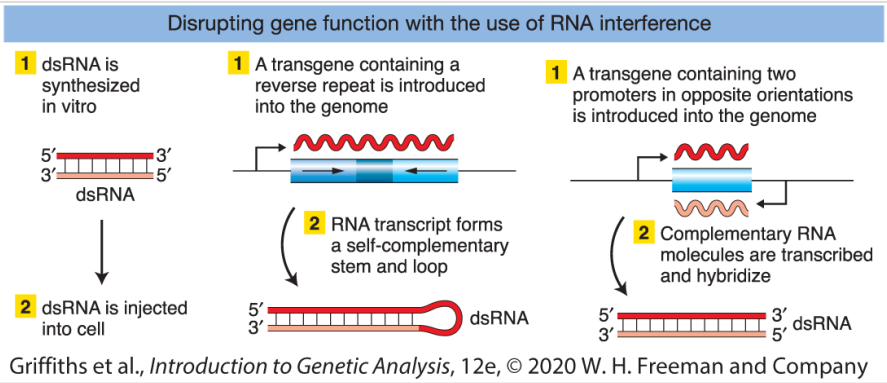


FIGURE 14-29 Three ways to create and introduce double-stranded RNA (dsRNA) into a cell. The dsRNA will then stimulate RNAi, degrading sequences that match those in the dsRNA.

On

But what makes RNAi especially powerful is that it can be applied to nonmodel organisms. First, target genes of interest can be identified by comparative genomics. Then RNAi sequences are produced to target the inhibition of the specific target genes. This technique has been applied, for example, to a mosquito that carries malaria (*Anopheles gambiae*). Using these techniques, scientists can better understand the biological mechanisms relating to the medical or economic effect of such species. The genes that control the complicated life cycle of the malaria parasite,

partly inside a mosquito host and partly inside the human body, can be better understood, revealing new ways to control the single most common infectious disease in the world.

KEY CONCEPT RNAi-based methods provide general ways of experimentally interfering with the function of a specific gene without changing its DNA sequence (generally called *phenocopying*).

Functional genomics with nonmodel organisms

Much of our consideration of mutational dissection and phenocopying has focused on genetic model organisms. One current focus of many geneticists is the broader application of these techniques to other species, including those that have negative effects on human society, such as parasites, disease carriers, or agricultural pests, or those species that are of interest to evolutionary biologists and ecologists (see Chapter 20). Classical genetic techniques are not readily applicable to most of these species, but whole-genome sequencing and functional genomics can now be conducted in any species for which DNA and tissue can be obtained. Furthermore, the roles of specific genes can be assessed through the generation of phenocopies by RNAi and targeted mutagenesis. In particular, the recently developed CRISPR-based methods for genome engineering are already being used in a number of nonmodel systems and promise to enable reverse genetic approaches in nearly any species (see Chapter 10).

KEY CONCEPT Reverse genetic methods are the gold standard to test the functions of genes and genetic elements discovered through genomic approaches. Recent technological advances mean that these methods can now be practiced in a variety of model and nonmodel systems.

SUMMARY

Genomic analysis takes the approaches of genetic analysis and applies them to the collection of global data sets to fulfill goals such as the mapping and sequencing of whole genomes and the characterization of all transcripts and proteins. Genomic techniques require the rapid processing of large sets of experimental material, all dependent on extensive automation.

The key problem in compiling an accurate sequence of a genome is to take short sequence reads and relate them to one another by sequence identity to build up a consensus sequence of an entire genome. This can be done in a straightforward way for bacterial or archaeal genomes by aligning overlapping sequences from different sequence reads to compile the entire genome, because few or no DNA segments are present in more than one copy in such organisms. The problem is that complex genomes of plants and animals are replete with such repetitive sequences. These repetitive sequences interfere with accurate sequence-contig production. The problem is resolved in whole-genome shotgun (WGS) sequencing with the use of paired-end reads.

Having a genomic sequence map provides the raw, encrypted text of the genome. The job of bioinformatics is to interpret this encrypted information. For the analysis of gene products, computational techniques are used to identify ORFs and noncoding RNAs, then to integrate these results with available experimental evidence for mRNA transcript structures (cDNA sequences), protein similarities, and knowledge of characteristic sequence motifs.

One of the most powerful means to advance the analysis and annotation of genomes is by comparing with the genomes of related species. Conservation of sequences among species is a reliable guide to identifying functional sequences in the complex genomes of many animals and plants. Comparative genomics can also reveal how genomes have changed in the course of evolution and how these changes may relate to differences in physiology, anatomy, or behavior among species. Comparisons of modern and archaic human genomes are accelerating the discovery of rare disease mutations. In bacterial genomics, comparisons of pathogenic and nonpathogenic strains have revealed many differences in gene content that contribute to pathogenicity.

Functional genomics attempts to understand the working of the genome as a whole system. Two key elements are the transcriptome, the set of all transcripts produced, and the interactome, the set

of interacting gene products and other molecules that together enable a cell to be produced and to function. The function of individual genes and gene products for which classical mutations are not available can be tested through reverse genetics—by targeted mutation or phenocopying.

KEY TERMS

annotation
bioinformatics
ChIP (chromatin immunoprecipitation)
ChIP-seq
comparative genomics
consensus sequence
copy number variation (CNV)
DNA sequencing library
exome
expressed sequence tag (EST)
forward genetics
functional genomics
genome project
genomics
homologous gene
interactome
open reading frame (ORF)
ortholog
outgroup
paired-end read
paralog
parsimony
personal genomics
phylogeny
phylogenetic inference
processed pseudogene
proteome
pseudogene
reverse genetics
RNA interference (RNAi)
RNA sequencing (RNA-seq)

scaffold
sequence assembly
sequence contig
supercontig
synteny
transcriptome
two-hybrid test
whole-genome shotgun (WGS) sequencing

SOLVED PROBLEMS

SOLVED PROBLEM 1

You want to study the development of the olfactory (smell-reception) system in the mouse. You know that the cells that sense specific chemical odors (odorants) are located in the lining of the nasal passages of the mouse. Describe some approaches for using functional genomics and reverse genetics to study olfaction.

SOLUTION

Many approaches can be imagined. For reverse genetics, you would want to first identify candidate genes that are expressed in the lining of the nasal passages. Given the techniques of functional genomics, this identification could be accomplished by purifying RNA from isolated nasal-passage-lining cells and using this RNA for an RNA-seq experiment. For example, you may choose to first examine mRNAs that are expressed in the nasal-passage lining but nowhere else in the mouse as important candidates for a specific role in olfaction. (Many of the important molecules may also have other jobs elsewhere in the body, but you have to start somewhere.) Alternatively, you may choose to start with those genes whose protein products are candidate proteins for binding the odorants themselves. Regardless of your choice, the next step would be to engineer a targeted knockout of the gene that encodes each mRNA or protein of interest or to use RNA interference to attempt to phenocopy the loss-of-function phenotype of each of the candidate genes.

PROBLEMS

Visit SaplingPlus for supplemental content. Problems with the  icon are available for review/grading.

WORKING WITH THE FIGURES

(The first 18 questions require inspection of text figures.)

1. You have identified a noncoding sequence that is conserved across all mammals, except for primates. You decided to engineer a targeted knockout of this sequence in mice. Based on Figure 14-2, is this a forward or reverse genetic experiment?

2. Based on the projection shown in Figure 14-3, what is the approximate number of human genomes that will be sequenced by 2025? How many basepairs will this represent?

3. Based on Figure 14-4, why must the DNA fragments sequenced overlap in order to obtain a genome sequence?

4. In Figure 14-6, the color pink indicates the base T, the color orange indicates the base A, the color yellow indicates the base G, and the color purple indicates the base C. What is the scanned sequence of the middle cluster in this figure? What is the scanned sequence of the cluster on the left?

5. Filling gaps in draft genome sequences is a major challenge. Based on Figures 14-8 and 14-9, can paired-end reads from a library of 2-kb fragments fill a 10-kb gap?

6. In Figure 14-11, how are the positions of codons determined?

7. In Figure 14-11, how are the positions of transcriptional regulatory elements determined?

8. In Figure 14-12, expressed sequence tags (ESTs) are aligned with genomic sequence. How are ESTs helpful in genome annotation?

9. In Figure 14-12, cDNA sequences are aligned with genomic sequence. How are cDNA sequences helpful in genome annotation? Are cDNAs more important for bacterial or eukaryotic genome annotations?

10. Based on Figure 14-16 and the features of ultraconserved elements, what would you predict you would observe if you injected a reporter-gene construct of the rat ortholog of the *ISL1* ultraconserved element into fertilized mouse oocytes and examined reporter gene expression in the developing embryo?

11. Based on Figure 14-17, did the duplication that created the A and B genes occur before or after speciation of the common ancestor of frogs, humans, and mice?

12. Based on Figure 14-18, are humans more closely related to mice or to dogs?

13. Figure 14-20 shows syntenic regions of mouse chromosome 11 and human chromosome 17. What do these syntenic regions reveal about the genome of the last common ancestor of mice and humans?

14. Based on Figure 14-22, what percent of Denisovan ancestry do you predict would be found in modern Western Europeans?

15. In Figure 14-23, what key step enables exome sequencing and distinguishes it from whole-genome sequencing?

16. The genomes of two *E. coli* strains are compared in Figure 14-24. Would you expect any third strain to contain more of the blue, tan, or red regions shown in Figure 14-24? Explain.

17. In Figure 14-25, why do the mRNA-sequencing reads map only to parts of the genome? Which gene is more highly expressed in sample 1?

18. Figure 14-26 depicts the Gal4-based two-hybrid system. Why do the "bait" proteins fused to the Gal4 DNA-binding protein not activate reporter-gene expression?

BASIC PROBLEMS

19. Explain the approach that you would apply to sequencing the genome of a newly discovered bacterial species.

20. Terminal-sequencing reads of clone inserts are a routine part of genome sequencing. How is the central part of the clone insert ever obtained?

21. What is the difference between a contig and a scaffold?

22. Two particular contigs are suspected to be adjacent, possibly separated by repetitive DNA. In an attempt to link them, end sequences are used as primers to try to bridge the gap. Is this approach reasonable? In what situation will it not work?

23. In a genomic analysis looking for a specific disease gene, one candidate gene was found to have a single–base-pair substitution resulting in a nonsynonymous amino acid change. What would you have to check before concluding that you had identified the disease-causing gene?

24. Is a bacterial operator a binding site?

25. A sequenced fragment of DNA in *Drosophila* was used in a BLAST search. The best (closest) match was to a kinase gene from *Neurospora*. Does this match mean that the *Drosophila* sequence contains a kinase gene?

26. In a two-hybrid test, a certain gene A gave positive results with two clones, M and N. When M was used, it gave positives with three clones, A, S, and Q. Clone N gave only one positive (with A). Develop a tentative interpretation of these results.

27. You have the following sequence reads from a genomic clone of the *Drosophila melanogaster* genome:
    Read 1: TGGCCGTGATGGGCAGTTCCGGTG
    Read 2: TTCCGGTGCCGGAAAGA
    Read 3: CTATCCGGGCGAACTTTTGGCCG
    Read 4: CGTGATGGGCAGTTCCGGTG
    Read 5: TTGGCCGTGATGGGCAGTT
    Read 6: CGAACTTTTGGCCGTGATGGGCAGTTCC

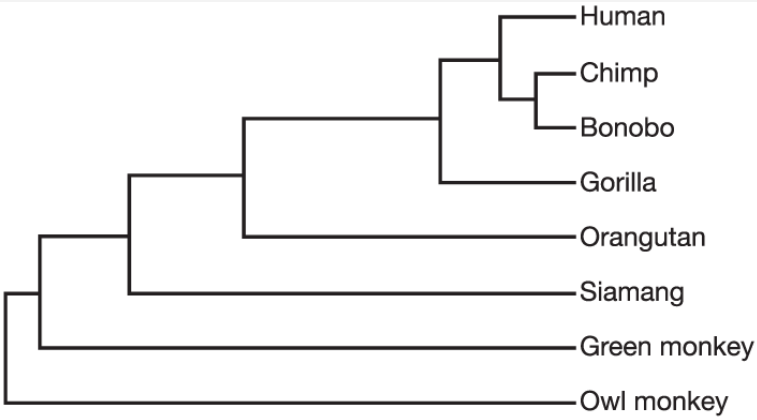
    Use these six sequence reads to create a sequence contig of this part of the *D. melanogaster* genome.

28. Sometimes, cDNAs turn out to be “chimeras”; that is, fusions of DNA copies of two different mRNAs accidentally inserted adjacently to each other in the same clone. You suspect that a cDNA clone from the nematode *Caenorhabditis elegans* is such a chimera because the sequence of the cDNA insert predicts a protein with two structural domains not normally observed in the same protein. How would you use the availability of the entire genomic sequence to assess if this cDNA clone is a chimera or not?

29. In browsing through the human genome sequence, you identify a gene that has an apparently long coding region, but there is a two–base-pair deletion that disrupts the reading frame.

    a. How would you determine whether the deletion was correct or an error in the sequencing?

    b. You find that the exact same deletion exists in the chimpanzee homolog of the gene but that the gorilla gene reading frame is intact. Given the phylogeny of great apes in the figure below, what can you conclude about when in ape evolution the mutation occurred?



Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

30. In browsing through the chimpanzee genome, you find that it has three homologs of a particular gene, whereas humans have only two.

    a. What are two alternative explanations for this observation?

b. How could you distinguish between these two possibilities?

31. The platypus is one of the few venomous mammals. The male platypus has a spur on the hind foot through which it can deliver a mixture of venom proteins. Looking at the phylogeny in Figure 14-18, how would you go about determining whether these venom proteins are unique to the platypus?

32. You have sequenced the genome of the bacterium *Salmonella typhimurium*, and you are using BLAST analysis to identify similarities within the *S. typhimurium* genome to known proteins. You find a protein that is 100 percent identical in the bacterium *Escherichia coli*. When you compare nucleotide sequences of the *S. typhimurium* and *E. coli* genes, you find that their nucleotide sequences are only 87 percent identical.

    a. Explain this observation.

    b. What do these observations tell you about the merits of nucleotide- versus protein-similarity searches in identifying related genes?

33. If you sequenced the genomes of any two unrelated humans, what types of sequence changes would you expect to find, and how many total base pairs would be affected by each type of sequence change?

34. You have access to both normal cells and cancerous cells taken from a biopsy from a patient with liver cancer. Describe the genomic approaches you would use to characterize the differences between these cells.

35. To inactivate a gene by RNAi, what information do you need? Do you need the map position of the target gene?

36. What is the purpose of generating a phenocopy?

37. What is the difference between forward and reverse genetics?

38. Why might exome sequencing fail to identify a disease-causing mutation in an affected person?

39. You have identified a noncoding sequence that is conserved in all mammals. Can you conclude that it is functional?

CHALLENGING PROBLEMS

40. You have the following sequence reads from a genomic clone of the *Homo sapiens* genome:
    Read 1: ATGCGATCTGTGAGCCGAGTCTTTA
    Read 2: AACAAAAATGTTGTTATTTTTATTTCAGATG
    Read 3: TTCAGATGCGATCTGTGAGCCGAG

Read 4: TGTCTGCCATTCTTAAAAACAAAAATGT
Read 5: TGTTATTTTTATTTCAGATGCGA
Read 6: AACAAAAATGTTGTTATT

a. Use these six sequence reads to create a sequence contig of this part of the *H. sapiens* genome.

b. Translate the sequence contig in all possible reading frames.

c. Go to the BLAST page of the National Center for Biotechnology Information, or NCBI (http://www.ncbi.nlm.nih.gov/BLAST/, Appendix B) and see if you can identify the gene of which this sequence is a part by using each of the reading frames as a query for protein–protein comparison (BLASTp).

41. Some sizable regions of different chromosomes of the human genome are more than 99 percent nucleotide identical with one another. These regions were overlooked in the production of the draft genome sequence of the human genome because of their high level of similarity. Of the techniques discussed in this chapter, which would allow genome researchers to identify the existence of such duplicate regions?

42. Some exons in the human genome are quite small (less than 75 bp long). Identification of such "microexons" is difficult because these distances are too short to reliably use ORF identification or codon bias to determine if small genomic sequences are truly part of an mRNA and a polypeptide. What techniques of "gene finding" can be used to try to assess if a given region of 75 bp constitutes an exon?

43. A certain cDNA of size 2 kb hybridized to eight genomic fragments of total size 30 kb and contained two short ESTs. The ESTs were also found in two of the genomic fragments each of size 2 kb. Sketch a possible explanation for these results.

44. You are studying proteins having roles in translation in the mouse. By BLAST analysis of the predicted proteins of the mouse genome, you identify a set of mouse genes that encode proteins with sequences similar to those of known eukaryotic translation-initiation factors. You are interested in determining the phenotypes associated with loss-of-function mutations of these genes.

    a. Would you use forward- or reverse-genetics approaches to identify these mutations?

    b. Briefly outline two different approaches that you might use to look for loss-of-function phenotypes in one of these genes.

45. You are interested in identifying genetic changes that might contribute to behavioral differences between two species of mice: one is promiscuous and the other is monogamous.

    a. Would you conduct whole-genome sequencing or exome sequencing of these two species? Defend your decision.

b. What additional functional genomics experiments would you do to identify differences between these two species?

c. How would you show that the genetic differences you identify actually contribute to behavioral differences between the species?

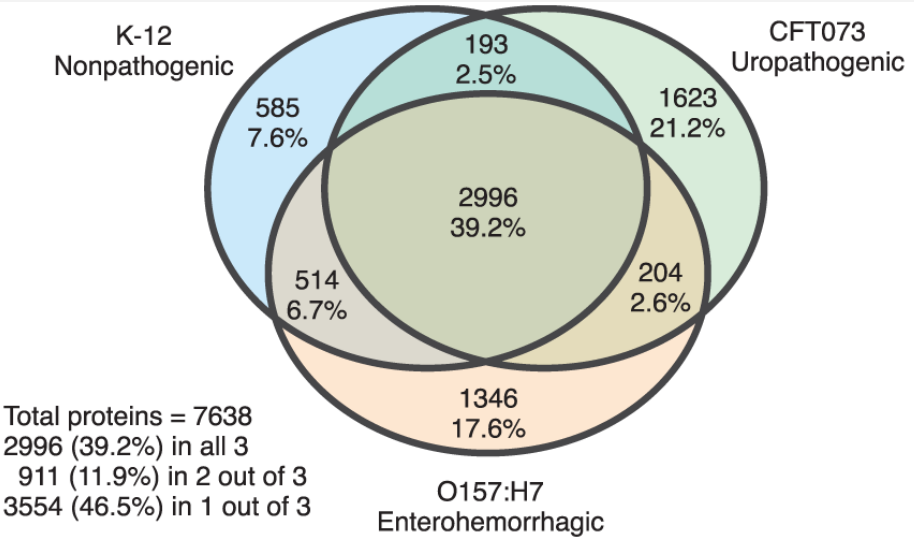
46. Different strains of *E. coli* are responsible for enterohemorrhagic and urinary tract infections. Based on the differences between the benign K-12 strain and the enterohemorrhagic O157:H7 strain, would you predict that there are obvious genomic differences

a. between K-12 and uropathogenic strains?

b. between O157:H7 and uropathogenic strains?

c. What might explain the observed pair-by-pair differences in genome content?

d. How might the function of strain-specific genes be tested?



Griffiths et al., *Introduction to Genetic Analysis*, 12e, © 2020 W. H. Freeman and Company

GENETICS AND SOCIETY

1. You decide to submit samples to two different “direct-to-consumer genetic testing” companies to learn more about your genetic ancestry. However, the results provided by the two companies give you very different estimates of your genetic origins. What might explain these differences?

2. What advice would you give to a friend who was considering doing “direct-to-consumer genetic testing” to learn more about their genealogy?