**The dog genome**

*Canis lupus familiaris*, the dog, was domesticated by humans from the gray wolf thousands of years ago. While there are many kinds of wolves, they all look more or less the same. Not so with “man’s best friend.” The American Kennel Club recognizes about 155 different breeds. Dog breeds not only look different, they vary greatly in size. For example, an adult Chihuahua weighs just 1.5 kg, while a Scottish deerhound weighs 70 kg. No other mammal shows such large phenotypic variation, and biologists are curious about how this occurs. Also, there are hundreds of genetic diseases in dogs, and many of these diseases have counterparts in humans. To find out about the genes behind the phenotypic variation, and to elucidate the relationships between genes and diseases, the Dog Genome Project began in the late 1990s. Since then the sequences of several dog genomes have been published.

Two dogs—a boxer and a poodle—were the first to have their entire genomes sequenced. The dog genome contains 2.8 billion base pairs of DNA in 39 pairs of chromosomes. There are 19,000 protein-coding genes, most of them with close counterparts in other mammals, including humans. The whole genome sequence made it easy to create a map of genetic markers—specific nucleotides or short sequences of DNA at particular locations on the genome that differ between individual dogs and/or breeds.

Genetic markers are used to map the locations of (and thus identify) genes that control particular traits. For example, Dr. Elaine Ostrander and her colleagues at the National Institutes of Health studied Portuguese water dogs to identify genes that control size. Taking samples of cells for DNA isolation was relatively easy: a cotton swab was swept over the inside of the cheek. As Dr. Ostrander said, the dogs “didn’t care, especially if they thought they were going to get a treat or if there was a tennis ball in our other hand.” It turned out that the gene for *insulin-like growth factor 1* (IGF-1) is important in determining size: large breeds have an allele that codes for an active IGF-1 and small breeds have a different allele that codes for a less active IGF-1.

Another gene important to phenotypic variation was found in whippets, sleek dogs that run fast and are often raced. A mutation in the gene for myostatin, a protein that inhibits overdevelopment of muscles, results in a

**Variation in Dogs**  The Chihuahua (bottom) and the Brazilian mastiff (top) are the same species, *Canis lupus familiaris*, and yet show great variation in size. Genome sequencing has revealed insights into how size is controlled by genes.
17.1 How Are Genomes Sequenced?

As you saw in the opening story on dogs, one reason for sequencing genomes is to compare different organisms. Another is to identify changes in the genome that result in disease. In 1986, the Nobel laureate Renato Dulbecco and others proposed that the world scientific community be mobilized to undertake the sequencing of the entire human genome. One challenge discussed at the time was to detect DNA damage in people who had survived the atomic bomb attacks and been exposed to radiation in Japan during World War II. But in order to detect changes in the human genome, scientists first needed to know its normal sequence.

The result was the publicly funded Human Genome Project, an enormous undertaking that was successfully completed in 2003. This effort was aided and complemented by privately funded groups. The project benefited from the development of many new methods that were first used in the sequencing of smaller genomes—those of prokaryotes and simple eukaryotes.

Two approaches were used to sequence the human genome

Many prokaryotes have a single chromosome, while eukaryotes have several to many. Because of their differing sizes, chromosomes can be separated from one another, identified, and experimentally manipulated. It might seem that the most straightforward approach to sequencing a chromosome would be to start at one end and simply sequence the entire DNA molecule. However, this approach is not practical since only about 700 base pairs can be sequenced at a time using current methods. Prokaryotic chromosomes contain 1–4 million base pairs and human chromosome 1 contains 246 million base pairs.

To sequence an entire genome, chromosomal DNA must be cut into short fragments about 500 base pairs long, which are separated and sequenced. For the haploid human genome, which has about 3.3 billion base pairs, there are more than 6 million such fragments. When all of the fragments have been sequenced, the problem becomes how to put these millions of sequences together. This task can be accomplished using larger, overlapping fragments.

Let’s illustrate this process using a single, 10 base-pair (bp) DNA molecule. (This is a double-stranded molecule, but for convenience we show only the sequence of the noncoding...
strand.) The molecule is cut three ways. The first cut generates the fragments:

TG, ATG, and CCTAC

The second cut of the same molecule generates the fragments:

AT, GCC, and TACTG

The third cut results in:

CTG, CTA, and ATGC

Can you put the fragments into the correct order? (The answer is ATG CCTACTG.) Of course, the problem of ordering 6 million fragments, each about 500 bp long, is more of a challenge! The field of bioinformatics was developed to analyze DNA sequences using complex mathematics and computer programs.

Until recently, two broad approaches were used to analyze DNA fragments for alignment: hierarchical sequencing and shotgun sequencing. These were developed for the Human Genome Project, but have been applied to other organisms as well.

**HIERARCHICAL SEQUENCING** The publicly funded human genome sequencing team developed a method known as hierarchical sequencing. The first step was to systematically identify short marker sequences along the chromosomes, ensuring that every fragment of DNA to be sequenced would contain a marker (Figure 17.1A). Genetic markers can be short tandem repeats (STRs), single nucleotide polymorphisms (SNPs), or the recognition sites for restriction enzymes, which recognize and cut DNA at specific sequences (see Chapter 15).

Some restriction enzymes recognize sequences of 4–6 base pairs and generate many fragments from a large DNA molecule. For example, the enzyme *Sau3A* cuts DNA every time it encounters GATC. Other restriction enzymes recognize sequences of 8–12 base pairs (*NotI* cuts at GCGGCCGC, for example) and generate far fewer, but much larger, fragments.

In hierarchical sequencing, genomic DNA is cut up into a set of relatively large (55,000 to 2 million bp) fragments. If different enzymes are used in separate digests, the fragments will overlap so that some fragments share particular markers. Each fragment is inserted into a bacterial plasmid to create a bacterial artificial chromosome (BAC), which is then inserted into bacteria. Each bacterium gets just one plasmid with its fragment of (for example) the human genome and is allowed to grow into a colony containing millions of genetically identical bacteria (called a clone). Clones differ from one another in that each has a different fragment from the human genome. A collection of clones, containing many different fragments of a genome, is called a genomic library.

The DNA from each clone is then extracted and cut into smaller overlapping pieces, which in turn are cloned, purified, and sequenced. The overlapping parts of the sequences allow researchers (with the aid of computers) to align them to create the complete sequence of the BAC clone. The genetic markers on each BAC clone are used to arrange the larger fragments in the proper order along the chromosome map. This method works, but it is slow. An alternative approach, shotgun sequencing, makes far greater use of use of computers to align the sequences.

**TOOLS FOR INVESTIGATING LIFE**

17.1 Sequencing Genomes Involves Fragment Overlaps

Short fragments of the whole genome can be sequenced, but then the fragments must be correctly aligned. Historically two approaches were used. Both involved the use of bacterial clones to separate and amplify individual DNA fragments.

- **Hierarchical sequencing**
  1. A marker map is made on a large DNA.
  2. The DNA is cut into fragments of 55,000 to 2 million bp. Several cuts are made to create overlapping fragments.
  3. Each fragment is amplified in a bacterial artificial chromosome (BAC).
  4. Marker sequences are identified on the fragments; common ones indicate overlap.
  5. The BAC fragments are cut into small pieces and sequenced from marker to marker, 500 bp at a time.

- **Shotgun sequencing**
  1. DNA is randomly broken into 500 bp fragments. Several cuts are made to create overlapping fragments.
  2. Each fragment is amplified and then sequenced.
  3. A computer finds sequences shared by fragments (overlaps), and aligns the fragments.
17.2 Sequencing DNA

(A) The normal substrates for DNA replication are dNTPs. The chemically modified structure of ddNTPs causes DNA synthesis to stop. (B) When labeled ddNTPs are incorporated into a reaction mixture for replicating a DNA template of unknown sequence, the result is a collection of fragments of varying lengths that can be separated by electrophoresis.

**SHOTGUN SEQUENCING** Instead of mapping the genome and creating a BAC library, the *shotgun sequencing* method involves directly cutting genomic DNA into smaller, overlapping fragments that are cloned and sequenced. Powerful computers align the fragments by finding sequence homologies in the overlapping regions (Figure 17.1B). As sequencing technologies and computers have improved, the shotgun approach has become much faster and cheaper than the hierarchical approach.

As a demonstration, researchers used this method to sequence a 1.8 million-base-pair prokaryotic genome in just a few months. Next came larger genomes. The entire 180 million-base-pair fruit fly genome was sequenced by the shotgun method in little over a year. This success proved that the shotgun method might work for the much larger human genome, and in fact it was used to sequence the human genome rapidly relative to the hierarchical method.

The nucleotide sequence of DNA can be determined

How are the individual DNA fragments generated by the hierarchical or shotgun methods sequenced? Current techniques are variations of a method developed in the late 1970s by Frederick Sanger. This method uses chemically modified nucleotides that were originally developed to stop cell division in cancer. As we discuss in Chapter 13, deoxyribonucleoside triphosphates (dNTPs) are the normal substrates for DNA replication, and contain the sugar deoxyribose. If that sugar is replaced with 2,3-dideoxyribose, the resulting dideoxyribonucleoside triphosphates (ddNTP) will still be added by DNA polymerase to a growing polynucleotide chain. However, because the ddNTP has no hydroxyl group (—OH) at the 3' position, the next nucleotide cannot be added (Figure 17.2A). Thus synthesis stops at the position where ddNTP has been incorporated into the growing end of a DNA strand.
To determine the sequence of a DNA fragment (usually no more than 700 base pairs long), it is isolated and mixed with
- DNA polymerase
- A short primer appropriate for the DNA sequence
- The four dNTPs (dATP, dGTP, dCTP, and dTTP)
- Small amounts of the four ddNTPs, each bonded to a differently colored fluorescent “tag”

In the first step of the reaction, the DNA is heated to denature it (separate it into single strands). Only one of these strands will act as a template for sequencing—the one to which the primer binds. DNA replication proceeds, and the test tube soon contains a mixture of the original DNA strands and shorter, new complementary strands. The new strands, each ending with a fluorescent ddNTP, are of varying lengths. For example, each time a T is reached on the template strand, DNA polymerase adds either a dATP or a ddATP to the growing complementary strand. If dATP is added, the strand continues to grow. If ddATP is added, growth stops (Figure 17.2B).

After DNA replication has been allowed to proceed for a while, the new DNA fragments are denatured and the single-stranded fragments separated by electrophoresis (see Figure 15.8), which sorts the DNA fragments by length. During the electrophoresis run, the fragments pass through a laser beam that excites the fluorescent tags, and the distinctive color of light emitted by each ddNTP is detected. The color indicates which ddNTP is at the end of each strand. A computer processes this information and prints out the DNA sequence of the fragment (see Figure 17.2B).

The delivery of chemical reagents by automated machines, coupled with automated analysis, has made DNA sequencing faster than ever. Huge laboratories often have 80 sequencing machines operating at once, each of which can sequence and analyze up to 70,000 bp in a typical 4-hour run. This may be fast enough for a prokaryotic genome with 1.5 million base pairs (20 runs), but when it comes to routine sequencing of larger genomes (like the 3.3 billion-base-pair human genome), even more speed is needed.

High-throughput sequencing has been developed for large genomes

The first decade of the new millennium has seen rapid development of high-throughput sequencing methods—fast, cheap ways to sequence and analyze large genomes. A variety of different approaches are being used. They generally involve the amplification of DNA templates by the polymerase chain reaction (PCR; see Section 13.5), and the physical binding of template DNA to a solid surface or to tiny beads called microbeads. These techniques are often referred to as massively parallel DNA sequencing, because thousands or millions of sequencing reactions are run at once to greatly speed up the process. One such high-throughput method is illustrated in Figure 17.3. In one 7-hour run, these machines can sequence 50,000,000 base pairs of DNA! How does it work?
For massively parallel sequencing using microbeads, the genomic DNA is first cut into 300- to 800-base-pair fragments. The fragments are denatured to single strands and attached to tiny beads that are less than 20 μm in diameter, one DNA fragment (template) per bead. PCR is used to create several million identical copies of the fragment on each bead. Then each bead is loaded into a tiny (40 μm diameter) well in a multi-well plate, and the sequencing begins.

The automated sequencer adds a reaction mix like the one described above, but containing only one of four fluorescently labeled dNTPs. That nucleotide will become incorporated as the first nucleotide in a complementary strand only in wells where the first nucleotide in the template strand can base-pair with it. For example, if the first nucleotide on the template in well #1 has base T, then a fluorescent nucleotide with base A will bind to that well. Next, the reaction mix is removed and a scanner captures an image of the plate, indicating which wells contain the fluorescent nucleotide. This process is repeated with a different labeled nucleotide. The machine cycles through many repeats using all four dNTPs, and records which wells gain new nucleotides after each cycle. A computer then identifies the sequence of nucleotides that were gained by each well, and aligns the fragments to provide the complete sequence of the genome.

This method was used to sequence the genome of James Watson, codiscoverer of the DNA double helix. It took less than two months and cost less than $1 million. Sequencing methods are being continually refined to increase speed and accuracy and decrease costs.

**Genome sequences yield several kinds of information**

New genome sequences are published more and more frequently, creating a torrent of biological information (Figure 17.4). In general, biologists use sequence information to identify:

- **Open reading frames**, the coding regions of genes. For protein-coding genes, these regions can be recognized by the start and stop codons for translation, and by intron consensus sequences that indicate the locations of introns.
- **Amino acid sequences** of proteins, which can be deduced from the DNA sequences of open reading frames by applying the genetic code (see Figure 14.6).
- **Regulatory sequences**, such as promoters and terminators for transcription.
- **RNA genes**, including rRNA, tRNA, and small nuclear RNA (snRNA) genes.
- **Other noncoding sequences** that can be classified into various categories including centromeric and telomeric regions, nuclear matrix attachment regions, transposons, and repetitive sequences such as short tandem repeats.

Sequence information is also used for **comparative genomics**, the comparison of a newly sequenced genome (or parts thereof) with sequences from other organisms. This can give information about the functions of sequences, and can be used to trace evolutionary relationships among different organisms.

---

**Figure 17.4 The Genomic Book of Life**

Genome sequences contain many features, some of which are summarized in this overview. Sifting through all the information contained in a genome sequence can help us understand how an organism functions and what its evolutionary history might be.
17.1 RECAP
The sequencing of genomes required the development of ways to cut large chromosomes into fragments, sequence the fragments, and then line them up on the chromosome. Two ways to do this are hierarchical sequencing and shotgun sequencing. Today new procedures are being developed that require automation and powerful computers. Actual DNA sequencing involves labeled nucleotides that are detected at the ends of growing polynucleotide chains.

- What are the hierarchical and shotgun approaches to genome analysis? See pp. 367–368 and Figure 17.1
- What is the dideoxy method for DNA sequencing? See pp. 368–369 and Figure 17.2
- Explain how high-throughput sequencing methods work. See pp. 369–370 and Figure 17.3
- How are open reading frames recognized in a genomic sequence? What kind of information can be derived from an open reading frame? See p. 370

We now turn to the first organisms whose sequences were determined, prokaryotes, and the information these sequences provided.

17.2 What Have We Learned from Sequencing Prokaryotic Genomes?
When DNA sequencing became possible in the late 1970s, the first life forms to be sequenced were the simplest viruses with their relatively small genomes. The sequences quickly provided new information on how these viruses infect their hosts and reproduce. But the manual sequencing techniques used on viruses were not up to the task of studying the genomes of prokaryotes and eukaryotes. The newer, automated sequencing techniques we just described made such studies possible. We now have genome sequences for many prokaryotes, to the great benefit of microbiology and medicine.

The sequencing of prokaryotic genomes led to new genomics disciplines
In 1995 a team led by Craig Venter and Hamilton Smith determined the first complete genomic sequence of a free-living cellular organism, the bacterium *Haemophilus influenzae*. Many more prokaryotic sequences have followed, revealing not only how prokaryotes apportion their genes to perform different cellular functions, but also how their specialized functions are carried out. Soon we may even be able to ask the provocative question of what the minimal requirements of a living cell might be.

**FUNCTIONAL GENOMICS** Functional genomics is the biological discipline that assigns functions to the products of genes. This
Transposable elements are DNA sequences that move from one location to another. They can be inserted into different parts of the genome, affecting the expression of genes. Transposons are a type of transposable element that can move about the genome. This movement can lead to changes in gene expression and can be used for research purposes.

**Some sequences of DNA can move about the genome**

Genome sequencing allowed scientists to study more broadly a class of DNA sequences that had been discovered by geneticists decades earlier. Segments of DNA called transposable elements can move from place to place in the genome and can even be inserted into another piece of DNA in the same cell (e.g., a plasmid). A transposable element might be at one location in the genome of one *E. coli* strain, and at a different location in another strain. The insertion of this movable DNA sequence from elsewhere in the genome into the middle of a protein-coding gene disrupts that gene (Figure 17.6A). Any mRNA expressed from the disrupted gene will have the extra sequence and the

17.5 Functional Organization of the Genome of *H. influenzae*

The entire DNA sequence has 1,830,137 base pairs. Different colors reflect different classes of gene function.

On this map, colors denote specific gene functions. For example, red genes regulate cellular processes... 

...yellow genes regulate replication... 

...and green genes regulate the production of the cell wall.

17.6 DNA Sequences that Move

Transposable elements are DNA sequences that move from one location to another. (A) In one method of transposition, the DNA sequence is replicated and the copy inserts elsewhere in the genome. (B) Transposons contain transposable elements and other genes.

**Comparative Genomics** Soon after the sequence of *H. influenzae* was announced, smaller (Mycoplasma genitalium; 580,073 base pairs) and larger (*E. coli*; 4,639,221 base pairs) prokaryotic sequences were completed. Thus began a new era in biology, that of comparative genomics, which compares genome sequences from different organisms. Scientists can identify genes that are present in one bacterium and missing in another, allowing them to relate these genes to bacterial function.

*M. genitalium*, for example, lacks the enzymes needed to synthesize amino acids, which *E. coli* and *H. influenzae* both possess. This finding reveals that *M. genitalium* must obtain all its amino acids from its environment (usually the human urogenital tract). Furthermore, *E. coli* has 55 regulatory genes coding for transcriptional activators and 58 for repressors; *M. genitalium* only has 3 genes for activators. What do such findings tell us about an organism’s lifestyle? For example, is the biochemical flexibility of *M. genitalium* limited by its relative lack of control over gene expression?

*H. influenzae* is humans. It lives in the upper respiratory tract and can cause ear infections or, more seriously, meningitis in children. Its single circular chromosome has 1,830,137 base pairs (Figure 17.5). In addition to its origin of replication and the genes coding for rRNAs and tRNAs, this bacterial chromosome has 1,738 open reading frames with promoters nearby. When this sequence was first announced, only 1,007 (58 percent) of the open reading frames coded for proteins whose functions were unknown. Since then scientists have identified many of these proteins’ roles. For example, they found genes for enzymes of glycolysis, fermentation, and electron transport. Other gene sequences code for membrane proteins, including those involved in active transport. An important finding was that highly infective strains of *H. influenzae*, but not noninfective strains, have genes for surface proteins that attach the bacterium to the human respiratory tract. These surface proteins are now a focus of research on possible treatments for *H. influenzae* infections.
protein will be abnormal. Some transposable elements can produce significant phenotypic effects by inactivating genes.

Transposable elements are often short sequences of 1,000–2,000 base pairs, and are found at many sites in prokaryotic genomes. The mechanisms that allow them to move vary. For example, a transposable element may be replicated, and then the copy inserted into another site in the genome. Or the element might splice out of one location and move to another location.

Longer transposable elements (up to 5,000 bp) carry additional genes and are called transposons (Figure 17.6B). Sometimes these DNA regions contain a gene for antibiotic resistance.

The sequencing of prokaryotic and viral genomes has many potential benefits

Prokaryotic genome sequencing promises to provide insights into microorganisms that cause human diseases. Genome sequencing has revealed unknown genes and proteins that can be targeted for isolation and functional study. Such studies are revealing new methods to combat pathogens and their infections. Sequencing has also revealed surprising relationships between some pathogenic organisms, suggesting that genes may be transferred between different strains.

- Chlamydia trachomatis causes the most common sexually transmitted disease in the United States. Because it is an intracellular parasite, it has been very hard to study. Among its 900 genes are several for ATP synthesis—something scientists used to think this bacterium could not accomplish on its own.
- Rickettsia prowazekii causes typhus; it is carried by lice and infects people bitten by the lice. Of its 634 genes, 6 encode proteins that are essential for virulence. These virulence proteins are being used to develop vaccines.
- Mycobacterium tuberculosis causes tuberculosis. It has a relatively large genome, coding for 4,000 proteins. Over 250 of these are used to metabolize lipids, so this may be the main way that this bacterium gets its energy. Some of its genes code for previously unidentified cell surface proteins; these proteins are targets for potential vaccines.
- Streptomyces coelicolor and its close relatives are the source for the genes for two-thirds of all naturally occurring antibiotics currently in clinical use. These antibiotics include streptomycin, tetracycline, and erythromycin. The genome sequence of S. coelicolor reveals 22 clusters of genes responsible for antibiotic production, of which only four were previously known. This finding may lead to new antibiotics to combat pathogens that have evolved resistance to conventional antibiotics.
- E. coli strain O157:H7 causes illness (sometimes severe) in at least 70,000 people a year in the United States. Its genome has 5,416 genes, of which 1,387 are different from those in the familiar (and harmless) laboratory strains of this bacterium. Many of these unique genes are also present in other pathogenic bacteria, such as Salmonella and Shigella. This finding suggests that there is extensive genetic ex-

change among these species, and that “superbugs” that share genes for antibiotic resistance may be on the horizon.

- Severe acute respiratory syndrome (SARS) was first detected in southern China in 2002 and rapidly spread in 2003. There is no effective treatment and 10 percent of infected people die. Isolation of the causative agent, a virus, and the rapid sequencing of its genome revealed several novel proteins that are possible targets for antiviral drugs or vaccines. Research is underway on both fronts, since another outbreak is anticipated.

Genome sequencing also provides insights into organisms involved in global ecological cycles (see Chapter 58). In addition to the well-known carbon dioxide, another important gas contributing to the atmospheric “greenhouse effect” and global warming is methane (CH₄; see Figure 2.7). Some bacteria, such as Methanococcus, produce methane in the stomachs of cows. Others, such as Methylococcus, remove methane from the air and use it as an energy source. The genomes of both of these bacteria have been sequenced. Understanding the genes involved in methane production and oxidation may help us to slow the progress of global warming.

Metagenomics allows us to describe new organisms and ecosystems

If you take a microbiology laboratory course you will learn how to identify various prokaryotes on the basis of their growth in lab cultures. For example, staphylococci are a group of bacteria that infect skin and nasal passages. When grown on a special medium called blood agar they form round, raised colonies. Microorganisms can also be identified by their nutritional requirements or the conditions under which they will grow (for example, aerobic versus anaerobic). Such culture methods have been the mainstay of microbial identification for over a century and are still useful and important. However, scientists can now use PCR and modern DNA analysis techniques to analyze microbes without culturing them in the laboratory.

In 1985, Norman Pace, then at Indiana University, came up with the idea of isolating DNA directly from environmental samples. He used PCR to amplify specific sequences from the samples to determine whether particular microbes were present. The PCR products were sequenced to explore their diversity. The term metagenomics was coined to describe this approach of analyzing genes without isolating the intact organism. It is now possible to perform shotgun sequencing with samples from almost any environment. The sequences can be used to detect the presence of known microbes and pathogens, and perhaps even the presence of heretofore unidentified organisms (Figure 17.7). For example:

- Shotgun sequencing of DNA from 200 liters of seawater indicated that it contained 5,000 different viruses and 2,000 different bacteria, many of which had not been described previously.
- One kilogram of marine sediment contained a million different viruses, most of them new.
Another way to define the minimal genome is to take an organism with a simple genome and deliberately mutate one gene at a time to see what happens. M. genitalium has one of the smallest known genomes—only 482 protein-coding genes. Even so, some of its genes are dispensable under some circumstances. For example, it has genes for metabolizing both glucose and fructose, but it can survive in the laboratory on a medium containing only one of these sugars.

What about other genes? Researchers have addressed this question with experiments involving the use of transposons as mutagens. When transposons in the bacterium are activated, they insert themselves into genes at random, mutating and inactivating them (Figure 17.8). The mutated bacteria are tested for growth and survival, and DNA from interesting mutants is sequenced to find out which genes contain transposons.

The astonishing result of these studies is that M. genitalium can survive in the laboratory with a minimal genome of only 382 functional genes! Is this really all it takes to make a viable organism? Experiments are underway to make a synthetic genome based on that of M. genitalium, and then insert it into an empty bacterial cell. If the cell starts transcribing mRNA and making proteins—is in fact viable—it may turn out to be the first life created by humans.

In addition to the technical feat of creating artificial life, this technique could have important applications. New microbes could be made with entirely new abilities, such as degrading oil spills, making synthetic fibers, reducing tooth decay, or converting cellulose to ethanol for use as fuel. On the other hand, fears of the misuse or mishandling of this knowledge are not unfounded. For example, it might also be possible to develop synthetic bacteria harmful to people, animals or plants, and use them as agents of biological warfare or bioterrorism. The “genomics genie” is, for better or worse, already out of the bottle. Hopefully human societies will use it to their benefit.

### 17.2 Recap

DNA sequencing is used to study the genomes of prokaryotes that are important to humans and to ecosystems. Functional genomics uses gene sequences to determine the functions of the gene products. Comparative genomics compares gene sequences from different organisms to help identify their functions and evolutionary relationships. Transposable elements and transposons move from one place to another in the genome.

- Give some examples of prokaryotic genomes that have been sequenced. What have the sequences shown? See pp. 371–373
- What is metagenomics and how is it used? See pp. 373–374 and Figure 17.7
- How are selective inactivation studies being used to determine the minimal genome? See p. 374 and Figure 17.8

Will defining the genes required for cellular life lead to artificial life?

When the genomes of prokaryotes and eukaryotes are compared, a striking conclusion arises: certain genes are present in all organisms (universal genes). There are also some (nearly) universal gene segments that are present in many genes in many organisms; for example, the sequence that codes for an ATP binding site. These findings suggest that there is some ancient, minimal set of DNA sequences common to all cells. One way to identify these sequences is to look for them in computer analyses of sequenced genomes.

### 17.7 Metagenomics

Microbial DNA extracted from the environment can be amplified and analyzed. This has led to the description of many new genes and species.

- Water runoff from a mine contained many new species of prokaryotes thriving in this apparently inhospitable environment. Some of these organisms exhibited metabolic pathways that were previously unknown to biologists. These organisms and their capabilities may be useful in cleaning up pollutants from the water.

These and other discoveries are truly extraordinary and potentially very important. It is estimated that 90 percent of the microbial world has been invisible to biologists and is only now being revealed by metagenomics. Entirely new ecosystems of bacteria and viruses are being discovered in which, for example, one species produces a molecule that another metabolizes. It is hard to overemphasize the importance of such an increase in our knowledge of the hidden world of microbes. This knowledge will help us to understand natural ecological processes, and has the potential to help us find better ways to manage environmental catastrophes such as oil spills, or remove toxic heavy metals from soil.
to do those specialized jobs. A typical virus contains enough DNA to code for only a few proteins—about 10,000 base pairs (bp). As we saw above, the simplest prokaryote, Mycoplasma, has several hundred protein-coding genes in a genome of 0.5 million bp. A rice plant, in contrast, has 37,544 genes.

• Eukaryotic genomes have more regulatory sequences—and many more regulatory proteins—than prokaryotic genomes. The greater complexity of eukaryotes requires much more regulation, which is evident in the many points of control associated with the expression of eukaryotic genes (see Figure 16.13).

• Much of eukaryotic DNA is noncoding. Distributed throughout many eukaryotic genomes are various kinds of DNA sequences that are not transcribed into mRNA, most notably introns and gene control sequences. As we discuss in Chapter 16, some noncoding sequences are transcribed into microRNAs. In addition, eukaryotic genomes contain various kinds of repeated sequences. These features are rare in prokaryotes.

• Eukaryotes have multiple chromosomes. The genomic “encyclopedia” of a eukaryote is separated into multiple “volumes.” Each chromosome must have, at a minimum, three defining DNA sequences that we have described in previous chapters: an origin of replication (ori) that is recognized by the DNA replication machinery; a centromere region that holds the replicated chromosomes together before mitosis; and a telomeric sequence at each end of the chromosome that maintains chromosome integrity.

Model organisms reveal many characteristics of eukaryotic genomes

Most of the lessons learned from eukaryotic genomes have come from several simple model organisms that have been studied extensively: the yeast Saccharomyces cerevisiae, the nematode (roundworm) Caenorhabditis elegans, the fruit fly Drosophila melanogaster, and—representing plants—the thale cress, Arabidopsis thaliana. Model organisms have been chosen because they are relatively easy to grow and study in a laboratory, their genetics are well studied, and they exhibit characteristics that represent a larger group of organisms.

YEAST: THE BASIC EUKARYOTIC MODEL

Yeasts are single-celled eukaryotes. Like most eukaryotes, they have membrane-enclosed organelles, such as the nucleus and endoplasmic reticulum, and a life cycle that alternates between haploid and diploid generations (see Figure 11.15).
While the prokaryote *E. coli* has a single circular chromosome with about 4.6 million bp and 4,290 protein-coding genes, budding yeast (*Saccharomyces cerevisiae*) has 16 linear chromosomes and a haploid content of more than 12.5 million bp, with 5,770 protein-coding genes. Gene inactivation studies similar to those carried out for *M. genitalium* (see Figure 17.7) indicate that fewer than 20 percent of these genes are essential to survival.

The most striking difference between the yeast genome and that of *E. coli* is in the number of genes for targeting proteins to organelles (Table 17.2). Both of these single-celled organisms appear to use about the same numbers of genes to perform the basic functions of cell survival. It is the compartmentalization of the eukaryotic yeast cell into organelles that requires it to have many more genes. This finding is direct, quantitative confirmation of something we have known for a century: the eukaryotic cell is structurally more complex than the prokaryotic cell.

**TABLE 17.1**

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>HAPLOID GENOME SIZE (Mb)</th>
<th>NUMBER OF GENES</th>
<th>PROTEIN-CODING SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>M. genitalium</em></td>
<td>0.58</td>
<td>485</td>
<td>88%</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>1.8</td>
<td>1,738</td>
<td>89%</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4.6</td>
<td>4,377</td>
<td>88%</td>
</tr>
<tr>
<td>Yeasts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>12.5</td>
<td>5,770</td>
<td>70%</td>
</tr>
<tr>
<td><em>S. pombe</em></td>
<td>12.5</td>
<td>4,929</td>
<td>60%</td>
</tr>
<tr>
<td>Plants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>115</td>
<td>28,000</td>
<td>25%</td>
</tr>
<tr>
<td>Rice</td>
<td>390</td>
<td>37,544</td>
<td>12%</td>
</tr>
<tr>
<td>Animals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>100</td>
<td>19,427</td>
<td>25%</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>123</td>
<td>13,379</td>
<td>13%</td>
</tr>
<tr>
<td>Pufferfish</td>
<td>342</td>
<td>27,918</td>
<td>10%</td>
</tr>
<tr>
<td>Chicken</td>
<td>1,130</td>
<td>25,000</td>
<td>3%</td>
</tr>
<tr>
<td>Human</td>
<td>3,300</td>
<td>24,000</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

*Mb = millions of base pairs*

Model organism of developmental biologists (see Section 19.4). The nematode has a transparent body that develops over 3 days from a fertilized egg to an adult worm made up of nearly 1,000 cells. In spite of its small number of cells, the nematode has a nervous system, digests food, reproduces sexually, and ages. So it is not surprising that an intense effort was made to sequence the genome of this model organism.

The *C. elegans* genome (100 million bp) is eight times larger than that of yeast and has 3.5 times as many protein-coding genes (19,427). Gene inactivation studies have shown that the worm can survive in laboratory cultures with only 10 percent of these genes. So the “minimum genome” of a worm is about twice the size of that of yeast, which in turn is four times the size of the minimum genome for *Mycoplasma*. What do these extra genes do?

All cells must have genes for survival, growth, and division. In addition, the cells of multicellular organisms must have genes for holding cells together to form tissues, for cell differentiation, and for intercellular communication. Looking at Table 17.3, you will recognize functions that we discussed in earlier chapters.

**TABLE 17.2**

<table>
<thead>
<tr>
<th></th>
<th><em>E. coli</em></th>
<th><em>YEAST</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome length (base pairs)</td>
<td>4,640,000</td>
<td>12,068,000</td>
</tr>
<tr>
<td>Number of protein-coding genes</td>
<td>4,290</td>
<td>5,770</td>
</tr>
<tr>
<td>Proteins with roles in:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>650</td>
<td>650</td>
</tr>
<tr>
<td>Energy production/storage</td>
<td>240</td>
<td>175</td>
</tr>
<tr>
<td>Membrane transport</td>
<td>280</td>
<td>250</td>
</tr>
<tr>
<td>DNA replication/repair/recombination</td>
<td>120</td>
<td>175</td>
</tr>
<tr>
<td>Transcription</td>
<td>230</td>
<td>400</td>
</tr>
<tr>
<td>Translation</td>
<td>180</td>
<td>350</td>
</tr>
<tr>
<td>Protein targeting/secretion</td>
<td>35</td>
<td>430</td>
</tr>
<tr>
<td>Cell structure</td>
<td>180</td>
<td>250</td>
</tr>
</tbody>
</table>

**TABLE 17.3**

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>PROTEIN/DOMAIN</th>
<th>NUMBER OF GENES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription control</td>
<td>Zinc finger; homeobox</td>
<td>540</td>
</tr>
<tr>
<td>RNA processing</td>
<td>RNA binding domains</td>
<td>100</td>
</tr>
<tr>
<td>Nerve impulse transmission</td>
<td>Gated ion channels</td>
<td>80</td>
</tr>
<tr>
<td>Tissue formation</td>
<td>Collagens</td>
<td>170</td>
</tr>
<tr>
<td>Cell interactions</td>
<td>Extracellular domains; glycotransferases</td>
<td>330</td>
</tr>
<tr>
<td>Cell–cell signaling</td>
<td>G protein-linked receptors; protein kinases; protein phosphatases</td>
<td>1,290</td>
</tr>
</tbody>
</table>

THE NEMATODE: UNDERSTANDING EUKARYOTIC DEVELOPMENT

In 1965 Sydney Brenner, fresh from being part of the team that first isolated mRNA, looked for a simple organism in which to study multicellularity. He settled on *Caenorhabditis elegans*, a millimeter-long nematode (roundworm) that normally lives in the soil. It can also live in the laboratory, where it has become a favorite model organism of developmental biologists (see Section 19.4).
including gene regulation (see Chapter 16) and cell communication (see Chapter 7).

**DROSOPHILA MELANOGASTER: RELATING GENETICS TO GENOMICS**

The fruit fly *Drosophila melanogaster* is a famous model organism. Studies of fruit fly genetics resulted in the formulation of many basic principles of genetics (see Section 12.4). Over 2,500 mutations of *D. melanogaster* had been described by the 1990s when genome sequencing began, and this fact alone was a good reason for sequencing the fruit fly’s DNA. The fruit fly is a much larger organism than *C. elegans*, both in size (it has 10 times more cells) and complexity, and it undergoes complicated developmental transformations from egg to larva to pupa to adult.

Not surprisingly, the fly’s genome (about 123 million bp) is larger than that of *C. elegans*. But as we mentioned earlier, genome size does not necessarily correlate with the number of genes encoded. In this case, the larger fruit fly genome contains fewer genes (13,379) than the smaller nematode genome. **Figure 17.9** summarizes the functions of the *Drosophila* genes that have been characterized so far; this distribution is typical of complex eukaryotes.

**ARABIDOPSIS: STUDYING THE GENOMES OF PLANTS**

About 250,000 species of flowering plants dominate the land and fresh water. But in the context of the history of life, the flowering plants are fairly young, having evolved only about 200 million years ago. The genomes of some plants are huge—for example, the genome of corn is about 3 billion bp, and that of wheat is 16 billion bp. So although we are naturally most interested in the genomes of plants we use as food and fiber, it is not surprising that scientists first chose to sequence a simpler flowering plant.

*Arabidopsis thaliana*, thale cress, is a member of the mustard family and has long been a favorite model organism of plant biologists. It is small (hundreds could grow and reproduce in the space occupied by this page) and easy to manipulate, and has a relatively small (115 million bp) genome.

The *Arabidopsis* genome has about 28,000 protein-coding genes but, remarkably, many of these genes are duplicates and probably originated by chromosomal rearrangements. When these duplicate genes are subtracted from the total, about 15,000 unique genes are left—similar to the gene numbers found in fruit flies and nematodes. Indeed, many of the genes found in these animals have homologs (genes with very similar sequences) in *Arabidopsis* and other plants, suggesting that plants and animals have a common ancestor.

But *Arabidopsis* has some genes that distinguish it as a plant (Table 17.4). These include genes involved in photosynthesis, in the transport of water into the root and throughout the plant, in the assembly of the cell wall, in the uptake and metabolism of inorganic substances from the environment, and in the synthesis of specific molecules used for defense against microbes and herbivores (organisms that eat plants). These plant defense molecules may be a major reason why the number of protein-coding genes in plants is higher than in animals. Plants cannot escape their enemies or other adverse conditions as animals can, and so they must cope with the situation where they are. So they make tens of thousands of molecules to fight their enemies and adapt to the environment (see Chapter 39).

These plant-specific genes are also found in the genomes of other plants, including rice, the first major crop plant whose sequence has been determined. Rice (*Oryza sativa*) is the world’s most important crop; it is a staple in the diet of 3 billion people. The larger genome in rice has a set of genes remarkably similar to that of *Arabidopsis*. More recently the genome of the poplar tree, *Populus trichocarpa*, was sequenced to gain insight into the potential for this rapidly growing tree to be used as a source of fixed carbon for making fuel. A comparison of the three genomes shows many genes in common, comprising the basic plant genome (Figure 17.10).

### Table 17.4

**Arabidopsis Genes Unique to Plants**

<table>
<thead>
<tr>
<th>Function</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell wall and growth</td>
<td>42</td>
</tr>
<tr>
<td>Water channels</td>
<td>300</td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>139</td>
</tr>
<tr>
<td>Defense and metabolism</td>
<td>94</td>
</tr>
</tbody>
</table>

**Eukaryotes have gene families**

About half of all eukaryotic protein-coding genes exist as only one copy in the haploid genome (two copies in somatic cells). The rest are present in multiple copies, which arise from gene duplications. Over evolutionary time, different copies of genes have undergone separate mutations, giving rise to groups of closely related genes called **gene families**. Some gene families, such as those encoding the globin proteins that make up hemoglobin, contain only a few members; other families, such as the genes encoding the immunoglobulins that make up antibodies, have hundreds of members. In the human genome,
Three plant genomes share a common set of approximately 21,000 genes that appear to comprise the “minimal” plant genome.

The globin gene family is a good example of the gene families found in vertebrates. These proteins are found in hemoglobin and myoglobin (an oxygen-binding protein present in muscle). The globin genes all arose long ago from a single common ancestral gene. In humans, there are three functional members of the α-globin cluster and five in the β-globin cluster (Figure 17.11). In adults, each hemoglobin molecule is a tetramer containing two identical α-globin subunits, two identical β-globin subunits, and four heme pigments (see Figure 3.10).

During human development, different members of the globin gene cluster are expressed at different times and in different tissues. This differential gene expression has great physiological significance. For example, hemoglobin containing γ-globin, a subunit found in the hemoglobin of the human fetus, binds O₂ more tightly than adult hemoglobin does. This specialized form of hemoglobin ensures that in the placenta, O₂ will be transferred from the mother’s blood to the developing fetus’s blood. Just before birth the liver stops synthesizing fetal hemoglobin and the bone marrow cells take over, making the adult forms (2α and 2β). Thus hemoglobins with different binding affinities for O₂ are provided at different stages of human development.

In addition to genes that encode proteins, many gene families include nonfunctional pseudogenes, which are designated with the Greek letter psi (Ψ) (see Figure 17.11). These pseudogenes result from mutations that cause a loss of function rather than an enhanced or new function. The DNA sequence of a pseudogene may not differ greatly from that of other family members. It may simply lack a promoter, for example, and thus fail to be transcribed. Or it may lack a recognition site needed for the removal of an intron, so that the transcript it makes is not correctly processed into a useful mature mRNA. In some gene families pseudogenes outnumber functional genes. Because some members of the family are functional, there appears to be little selection pressure for the deletion of pseudogenes.

Eukaryotic genomes contain many repetitive sequences

Eukaryotic genomes contain numerous repetitive DNA sequences that do not code for polypeptides. These include highly repetitive sequences, moderately repetitive sequences, and transposons.

Highly repetitive sequences are short (less than 100 bp) sequences that are repeated thousands of times in tandem (side-by-side) arrangements in the genome. They are not transcribed. Their proportion in eukaryotic genomes varies, from 10 percent in humans to about half the genome in some species of fruit flies. Often they are associated with heterochromatin, the densely packed, transcriptionally inactive part of the genome. Other highly repetitive sequences are scattered around the genome. For example, short tandem repeats (STRs) of 1–5 bp can be repeated up to 100 times at a particular chromosomal location. The copy number of an STR at a particular location varies between individuals and is inherited. In Chapter 15 we describe how STRs can be used in the identification of individuals (DNA fingerprinting).

Moderately repetitive sequences are repeated 10–1000 times in the eukaryotic genome. These sequences include the genes that are transcribed to produce tRNAs and rRNAs, which are used in protein synthesis. The cell makes tRNAs and rRNAs constantly, but even at the maximum rate of transcription, single copies of the tRNA and RNA genes would be inadequate to supply the large amounts of these molecules needed by most cells. Thus the genome has multiple copies of these genes.

In mammals, four different rRNA molecules make up the ribosome: the 18S, 5.8S, 28S, and 5S rRNAs. (The S stands for Svedberg unit, which is a measure of size.) The 18S, 5.8S, and...
28S rRNAs are transcribed together as a single precursor RNA molecule (Figure 17.12). As a result of several posttranscriptional steps, the precursor is cut into the final three rRNA products, and the noncoding “spacer” RNA is discarded. The sequence encoding these RNAs is moderately repetitive in humans: a total of 280 copies of the sequence are located in clusters on five different chromosomes.

**TRANSPOSONS** Apart from the RNA genes, most moderately repetitive sequences are not stably integrated into the genome. Instead, these sequences can move from place to place, and are thus called transposable elements or transposons. Prokaryotes also have transposons (see Figure 17.6). Transposons make up over 40 percent of the human genome and about 50 percent of the maize genome, although the percentage is smaller (3–10 percent) in many other eukaryotes.

There are four main types of transposons in eukaryotes:

1. **SINEs** (short interspersed elements) are up to 500 bp long and are transcribed but not translated. There are about 1.5 million of them scattered over the human genome, making up about 15 percent of the total DNA content. A single type, the 300-bp Alu element, accounts for 11 percent of the human genome; it is present in a million copies.

2. **LINEs** (long interspersed elements) are up to 7,000 bp long, and some are transcribed and translated into proteins. They constitute about 17 percent of the human genome.

3. **Retrotransposons** also make RNA copies of themselves when they move about the genome. Some of them encode proteins needed for their own transposition, and others do not. SINEs and LINEs are types of retrotransposons. Non-SINE, non-LINE retrotransposons constitute about 8 percent of the human genome.

4. **DNA transposons** do not use RNA intermediates. Like some prokaryotic transposable elements, they are excised from the original location and become inserted at a new location without being replicated.

What role do these moving sequences play in the cell? The best answer so far seems to be that transposons are simply cellular parasites that can be replicated. The insertion of a transposon at a new location can have important consequences. For example, the insertion of a transposon into the coding region of a gene results in a mutation (see Figure 17.8). This phenomenon accounts for a few rare forms of several genetic diseases in humans, including hemophilia and muscular dystrophy. If the insertion takes place in a somatic cell, cancer may result.

Sometimes an adjacent gene can be replicated along with a transposon, resulting in a gene duplication. A transposon can carry a gene, or a part of it, to a new location in the genome, shuffling the genetic material and creating new genes. Clearly, transposition stirs the genetic pot in the eukaryotic genome and thus contributes to genetic variation.

Section 5.5 describes the theory of **endosymbiosis**, which proposes that chloroplasts and mitochondria are the descendants of once free-living prokaryotes. Transposons may have played a role in endosymbiosis. In living eukaryotes the chloroplasts and mitochondria contain some DNA, but the nucleus contains most of the genes...
17.3 RECAP

The sequencing of the genomes of model organisms demonstrated common features of the eukaryotic genome, including the presence of repetitive sequences and transposons. Some eukaryotic genes are in families, which may include members that are mutated and nonfunctional. Some sequences are transcribed, but others are not.

- What are the major differences between prokaryotic and eukaryotic genomes? See p. 375
- Describe one function of genes found in C. elegans that has no counterpart in the genome of yeast. See p. 376 and Table 17.3
- What is the evolutionary role of eukaryotic gene families? See p. 377
- Why are there multiple copies of sequences coding for rRNA in the mammalian genome? See p. 378
- What effects can transposons have on a genome? See p. 379

The analysis of eukaryotic genomes has resulted in an enormous amount of useful information, as we have seen. In the next section we look more closely at the human genome.

17.4 What Are the Characteristics of the Human Genome?

By the start of 2005 the first human genome sequences were completed, two years ahead of schedule and well under budget. The published sequences, one produced by the publicly funded Human Genome Project, and the other by a private company, were haploid genomes that were composites of several people. Since 2005, the diploid genomes of several individuals have been sequenced and published.

The human genome sequence held some surprises

The following are just some of the interesting facts that we have learned about the human genome:

- Of the 3.3 billion base pairs in the haploid human genome, fewer than 2 percent (about 24,000 genes) make up protein-coding regions. This was a surprise. Before sequencing began, humans were estimated, based on the diversity of their proteins, to have 80,000–150,000 genes. The actual number of genes—not many more than in a fruit fly—means that posttranscriptional mechanisms (such as alternative splicing) must account for the observed number of proteins in humans. That is, the average human gene must code for several different proteins.
- The average gene has 27,000 base pairs. Gene sizes vary greatly, from about 1,000 to 2.4 million base pairs. Variation in gene size is to be expected given that human proteins (and RNAs) vary in size, from 100 to about 5,000 amino acids per polypeptide chain.
- Virtually all human genes have many introns.
- Over 50 percent of the genome is made up of transposons and other highly repetitive sequences. Repetitive sequences near genes are GC-rich, while those farther away from genes are AT-rich.
- Most of the genome (about 97 percent) is the same in all people. Despite this apparent homogeneity, there are, of course, many individual differences. Scientists have mapped over 7 million single nucleotide polymorphisms (SNPs) in humans.
- Genes are not evenly distributed over the genome. Chromosome 19 is packed densely with genes, while chromosome 8 has long stretches without coding regions. The Y chromosome has the fewest genes (231), while chromosome 1 has the most (2,968).
- Comparisons between sequenced genomes from prokaryotes and eukaryotes have revealed some of the evolutionary relationships between genes. Some genes are present in both prokaryotes and eukaryotes; others are only in eukaryotes; still others are only in animals, or only in vertebrates (Figure 17.14).
A comparison of the human and other genomes has revealed how genes with new functions have been added over the course of evolution. Each percentage number refers to genes in the human genome. Thus, 21 percent of human genes have homologs in prokaryotes and other eukaryotes, 32 percent of human genes occur only in other eukaryotes, and so on.

More comparative genomics is possible now that the genomes of two other primates, the chimpanzee and the rhesus macaque, have been sequenced. The chimpanzee is evolutionarily close to humans, and shares 95 percent of the human genome sequence. The more distantly related rhesus macaque shares 91 percent of the human sequence. The search is on for a set of human genes that differ from the other primates and “make humans human.”

**Human genomics has potential benefits in medicine**

Complex phenotypes are determined not by single genes, but by multiple genes interacting with the environment. The single-allele models of phenylketonuria and sickle-cell anemia (see Chapter 15) do not apply to such common disorders as diabetes, heart disease, and Alzheimer’s disease. To understand the genetic bases of these diseases, biologists are now using rapid genotyping technologies to create “haplotype maps,” which are used to identify SNPs (pronounced “snips”) that are linked to genes involved in disease.

**HAPLOTYPE MAPPING** The SNPs that differ between individuals are not inherited as independent alleles. Rather, a set of SNPs that are present on a segment of chromosome are usually inherited as a unit. This linked piece of a chromosome is called a haplotype. You can think of the haplotype as a sentence and the SNP as a word in the sentence. Analyses of haplotypes in humans from all over the world have shown that there are at most 500,000 common variations.

**GENOTYPING TECHNOLOGY AND PERSONAL GENOMICS** New technologies are continually being developed to analyze thousands or millions of SNPs in the genomes of individuals. Such technologies include rapid sequencing methods and DNA microarrays that depend on DNA hybridization to identify specific SNPs. For example, a microarray of 500,000 SNPs has been used to analyze thousands of people to find out which SNPs are associated with specific diseases. The aim is to correlate the SNP-defined haplotype with a disease state. The amount of data is prodigious: 500,000 SNPs, thousands of people, thousands of medical records. With so much natural variation, statistical measures of association between a haplotype and a disease need to be very rigorous.

These association tests have revealed particular haplotypes or alleles that are associated with modestly increased risks for such diseases as breast cancer, diabetes, arthritis, obesity, and coronary heart disease (Figure 17.15 and Table 17.5). Private companies will now scan a human genome for these variants—and the price for this service keeps getting lower. However, at this point it is unclear what a person without symptoms should do with the information, since multiple genes, environmental influences, and epigenetic effects all contribute to the development of these diseases.

Of course, the best way to analyze a person’s genome is by actually sequencing it. Until recently, this was prohibitively expensive. As we mentioned earlier, DNA pioneer James Watson’s genome cost over $1 million, certainly too much for a typical person or insurance company to afford in the context of health care. But with advances in sequencing technologies the cost is decreasing rapidly. One new method automatically sequences protein-coding exons only, for example. Once the cost of genome sequencing is within an affordable range, SNP testing will be superseded.
PHARMACOGENOMICS  Genetic variation can affect how an individual responds to a particular drug. For example, a drug may be chemically modified in the liver to make it more or less active. Consider an enzyme that catalyzes the following reaction:

$$\text{active drug} \rightarrow \text{less active drug}$$

A mutation in the gene that encodes this enzyme may make the enzyme less active. For a given dose of the drug, a person with the mutation would have more active drug in the bloodstream than a person without the mutation. So the effective dose of the drug would be lower in these people.

Now consider a different case, in which the liver enzyme is needed to make the drug active:

$$\text{inactive drug} \rightarrow \text{active drug}$$

A person carrying a mutation in the gene encoding this liver enzyme would not be affected by the drug, since the activating enzyme is not present.

The study of how an individual’s genome affects his or her response to drugs or other outside agents is called pharmacogenomics. This type of analysis makes it possible to predict whether a drug will be effective. The objective is to personalize drug treatment so that a physician can know in advance whether an individual will benefit from a particular drug (Figure 17.16). This approach might also be used to reduce the incidence of adverse drug reactions by identifying individuals that will metabolize a drug slowly, which can lead to a dangerously high level of the drug in the body.

17.4 RECAP

The haploid human genome has 3.3 billion base pairs, but less than 2 percent of the genome codes for proteins. Most human genes are subject to alternative splicing; this may account for the fact that there are more proteins than genes. SNP mapping to find correlations with disease and drug susceptibility holds promise for personalized medicine.

- What are some of the major characteristics of the human genome? See p. 380
- How does SNP mapping work in personalized medicine? See pp. 381–382 and Figures 17.15 and 17.16

Genome sequencing has had great success in advancing biological understanding. High-throughput technologies are now being applied to other components of the cell: proteins and metabolites. We now turn to the results of these studies.

17.5 What Do the New Disciplines Proteomics and Metabolomics Reveal?

“The human genome is the book of life.” Statements like this were common at the time the human genome sequence was first revealed. They reflect “genetic determinism,” that a person’s phenotype is determined by his or her genotype. But is an organism just a product of gene expression? We know that it is not. The proteins and small molecules present in any cell at a given point in time reflect not just gene expression but modifications by the intracellular and extracellular environment. Two new fields have emerged to complement genomics and take a more complete snapshot of a cell and organism—proteomics and metabolomics.

The proteome is more complex than the genome

As mentioned above, many genes encode more than a single protein (Figure 17.17A). Alternative splicing leads to different combinations of exons in the mature mRNAs transcribed from a single gene (see Figure 16.22). Posttranslational modifications also increase the number of proteins that can be derived from one gene (see Figure 14.22). The proteome is the sum total of the proteins produced by an organism, and it is more complex than its genome.
Two methods are commonly used to analyze the proteome:

- Because of their unique amino acid compositions (primary structures), most proteins have unique combinations of electric charge and size. On the basis of these two properties, they can be separated by two-dimensional gel electrophoresis. Thus isolated, individual proteins can be analyzed, sequenced, and studied (Figure 17.17B).

- Mass spectrometry uses electromagnets to identify proteins by the masses of their atoms and displays them as peaks on a graph.

The ultimate aim of proteomics is just as ambitious as that of genomics. While genomics seeks to describe the genome and its expression, proteomics seeks to identify and characterize all of the expressed proteins.

Comparisons of the proteomes of humans and other eukaryotic organisms have revealed a common set of proteins that can be categorized into groups with similar amino acid sequences and similar functions. Forty-six percent of the yeast proteome, 43 percent of the worm proteome, and 61 percent of the fly proteome are shared by the human proteome. Functional analyses indicate that this set of 1,300 proteins provide the basic metabolic functions of an eukaryotic cell, such as glycolysis, the citric acid cycle, membrane transport, protein synthesis, DNA replication, and so on. (Figure 17.18).

Of course, these are not the only human proteins. There are many more, which presumably distinguish us as human eukaryotic organisms. As we have mentioned before, proteins have different functional regions called domains (for example, a domain for binding a substrate, or a domain for spanning a membrane). While a particular organism may have many unique proteins, those proteins are often just unique combinations of domains that exist in other organisms. This reshuffling of the genetic deck is a key to evolution.

**Metabolomics is the study of chemical phenotype**

Studying genes and proteins gives a limited picture of what is going on in a cell. But as we have seen, both gene function and protein function are affected by the internal and external environments of the cell. Many proteins are enzymes and their activities affect the concentrations of their substrates and products. So as the proteome changes, so will the abundances of these often-small molecules, called metabolites. The **metabolome** is the quantitative description of all of the small molecules in a cell or organism. These include:

- **Primary metabolites** involved in normal processes, such as intermediates in pathways like glycolysis. This category also includes hormones and other signaling molecules.

- **Secondary metabolites**, which are often unique to particular organisms or groups of organisms. They are often involved in special responses to the environment. Examples are antibiotics made by microbes, and the many chemicals made by plants that are used in defense against pathogens and herbivores.

Not surprisingly, measuring metabolites involves sophisticated analytical instruments. If you have studied organic or analytical chemistry, you may be familiar with gas chromatography and high-performance liquid chromatography, which separate molecules, and mass spectrometry and nuclear magnetic resonance spectroscopy, which are used to identify them. These measurements result in “chemical snapshots” of cells or organisms, which can be related to physiological states.

There has been some progress in defining the human metabolome. A database created by David Wishart and col-
leagues at the University of Alberta contains over 6,500 metabolite entries. The challenge now is to relate levels of these substances to physiology. For example, you probably know high levels of glucose in the blood are associated with diabetes. But what about early stages of heart disease? There may be a pattern of metabolites that is diagnostic of this disease. This could aid in early diagnosis and treatment.

Plant biologists are far ahead of medical researchers in the study of metabolomics. Over the years, tens of thousands of secondary metabolites have been identified in plants, many of them made in response to environmental challenges. Some of these are discussed in Chapter 39. The metabolome of the model organism Arabidopsis thaliana is being described, and will give insight into how a plant copes with stresses such as drought or pathogen attack. This knowledge could be helpful in optimizing plant growth for agriculture.

### 17.5 RECAP

The proteome is the total of all proteins produced by an organism. There are more proteins than genes in the genome. The metabolome is the total content of small molecules such as intermediates in metabolism, hormones, and secondary metabolites. The proteome and the metabolome can be analyzed using chemical methods that separate and identify molecules.

- How is the proteome analyzed? See p. 383 and Figure 17.17
- Explain the differences between genome, proteome, and metabolome.

### CHAPTER SUMMARY

#### 17.1 How Are Genomes Sequenced?
- The sequencing of genomes required the development of ways to cut large chromosomes into fragments, sequence each of the fragments, and then line them up on the chromosome. Review Figure 17.1, ANIMATED TUTORIAL 17.1
- Hierarchical sequencing involves mapping the genome with genetic markers, cutting the genome into smaller pieces and sequencing them, then lining up the sequences using the markers.
- Shotgun sequencing involves directly cutting the genome into overlapping fragments, sequencing them, and using a computer to line up the sequences.
- DNA sequencing technologies involve labeled nucleotides that terminate the growing polynucleotide chain. Review Figure 17.2
- Rapid, automated methods for high-throughput sequencing are being developed. Review Figure 17.3, ANIMATED TUTORIAL 17.2

#### 17.2 What Have We Learned from Sequencing Prokaryotic Genomes?
- DNA sequencing is used to study the genomes of prokaryotes that are important to humans and ecosystems.
- Functional genomics aims to determine the functions of gene products. Comparative genomics involves comparisons of genes and genomes from different organisms to identify common features and functions.
- Transposable elements and transposons can move about the genome. Review Figure 17.6
- Metagenomics is the identification of DNA sequences without first isolating, growing, and identifying the organisms present in an environmental sample. Many of these sequences are from prokaryotes that were heretofore unknown to biologists. Review Figure 17.7
- Transposon mutagenesis can be used to inactivate genes one by one. Then the organism can be tested for survival. In this way, a minimal genome of less than 350 genes was identified for the bacterium Mycoplasma genitalium. Review Figure 17.8

#### 17.3 What Have We Learned from Sequencing Eukaryotic Genomes?
- Genome sequences from model organisms have demonstrated some common features of the eukaryotic genome. In addition, there are specialized genes for cellular compartmentation, development, and features unique to plants. Review Tables 17.1–17.4 and Figures 17.9 and 17.10
- Some eukaryotic genes exist as members of gene families. Proteins may be made from these closely related genes at different times and in different tissues. Some members of gene families may be nonfunctional pseudogenes.
- Repeated sequences are present in the eukaryotic genome.
- Moderately repeated sequences include those coding for rRNA. Review Figure 17.12

#### 17.4 What Are the Characteristics of the Human Genome?
- The haploid human genome has 3.3 billion base pairs.
- Only 2 percent of the genome codes for proteins; the rest consists of repeated sequences and noncoding DNA.
- Virtually all human genes have introns, and alternative splicing leads to the production of more than one protein per gene.
- SNP genotyping correlates variations in the genome with diseases or drug sensitivity. It may lead to personalized medicine. Review Figure 17.15
- Pharmacogenomics is the analysis of genetics as applied to drug metabolism.

#### 17.5 What Do the New Disciplines of Proteomics and Metabolomics Reveal?
- The proteome is the total protein content of an organism.
- There are more proteins than protein-coding genes in the genome.
- The proteome can be analyzed using chemical methods that separate and identify proteins. These include two-dimensional electrophoresis and mass spectrometry. See Figure 17.17
- The metabolome is the total content of small molecules, such as intermediates in metabolism, hormones, and secondary metabolites.

SEE WEB ACTIVITY 17.1 for a concept review of this chapter.
1. Eukaryotic protein-coding genes differ from their prokaryotic counterparts in that eukaryotic genes
   a. are double-stranded.
   b. are present in only a single copy.
   c. contain introns.
   d. have promoters.
   e. are transcribed into mRNA.
2. A comparison of the genomes of yeast and bacteria shows that only yeast has many genes for
   a. energy metabolism.
   b. cell wall synthesis.
   c. intracellular protein targeting.
   d. DNA-binding proteins.
   e. RNA polymerase.
3. The genomes of the fruit fly and the nematode are similar to that of yeast, except that the former organisms have
   many genes for
   a. intercellular signaling.
   b. synthesis of polysaccharides.
   c. cell cycle regulation.
   d. intracellular protein targeting.
   e. transposable elements.
4. The minimum genome of *Mycoplasma genitalium*
   a. has 100 genes.
   b. has been used to create new species.
   c. has an RNA genome.
   d. is larger than the genome of *E. coli*.
   e. was derived by transposon mutagenesis.
5. Which is not true of metagenomics?
   a. It has been done with bacteria.
   b. It is done on rRNA sequences.
   c. It has revealed many new metabolic capacities.
   d. It involves extracting DNA from the environment.
   e. It can be done on seawater.
6. Transposons
   a. always use RNA for replication.
   b. are approximately 50 bp long.
   c. are made up of either DNA or RNA.
   d. do not contain genes coding for proteins.
   e. make up about 40 percent of the human genome.
7. Vertebrate gene families
   a. have mostly inactive genes.
   b. include the globins.
   c. are not produced by gene duplications.
   d. increase the number of unique genes in the genome.
   e. are not transcribed.
8. The DNA sequences that code for eukaryotic rRNA
   a. are transcribed only at the ribosome.
   b. are repeated hundreds of times.
   c. contain all the genes clustered directly beside one another.
   d. are on only one human chromosome.
   e. are identical to the sequences that code for miRNA.
9. The human genome
   a. contains very few repeated sequences.
   b. has 3.3 billion base pairs.
   c. was sequenced by hierarchical sequencing only.
   d. has genes evenly distributed along chromosomes.
   e. has few genes with introns.
10. Which of the following about genome sequencing is true?
    a. In hierarchical sequencing, but not high-throughput sequencing, DNA is amplified in BAC vectors.
    b. In hierarchical sequencing, a genetic map is made after the DNA is sequenced.
    c. Shotgun sequencing is considerably slower than hierarchical sequencing.
    d. The human genome was first sequenced by the high-throughput methods.
    e. DNA sequence determination by chain termination is the basis of shotgun sequencing only.

FOR DISCUSSION
1. In rats, a protein-coding gene 1,440 bp long codes for an enzyme made up of 192 amino acids. Discuss this apparent discrepancy. How long would the initial and final mRNA transcripts be?
2. The genomes of rice, wheat, and corn are similar to one another and to that of *Arabidopsis* in many ways. Discuss how these plants might nevertheless have very different proteins.
3. Why are the proteome and the metabolome more complex than the genome?

ADDITIONAL INVESTIGATION
It is the year 2025. You are taking care of a patient who is concerned about having an early stage of kidney cancer. His mother died from this disease.

a. Assume that the SNPs linked to genes involved in the development of this type of cancer have been identified. How would you determine if this man has a genetic predisposition for developing kidney cancer? Explain how you would do the analysis.

b. How might you develop a metabolomic profile for kidney cancer and then use it to determine whether your patient has kidney cancer?

c. If the patient was diagnosed with cancer by the methods in (a) and (b), how would you use pharmacogenomics to choose the right medications to treat the tumor in this patient?