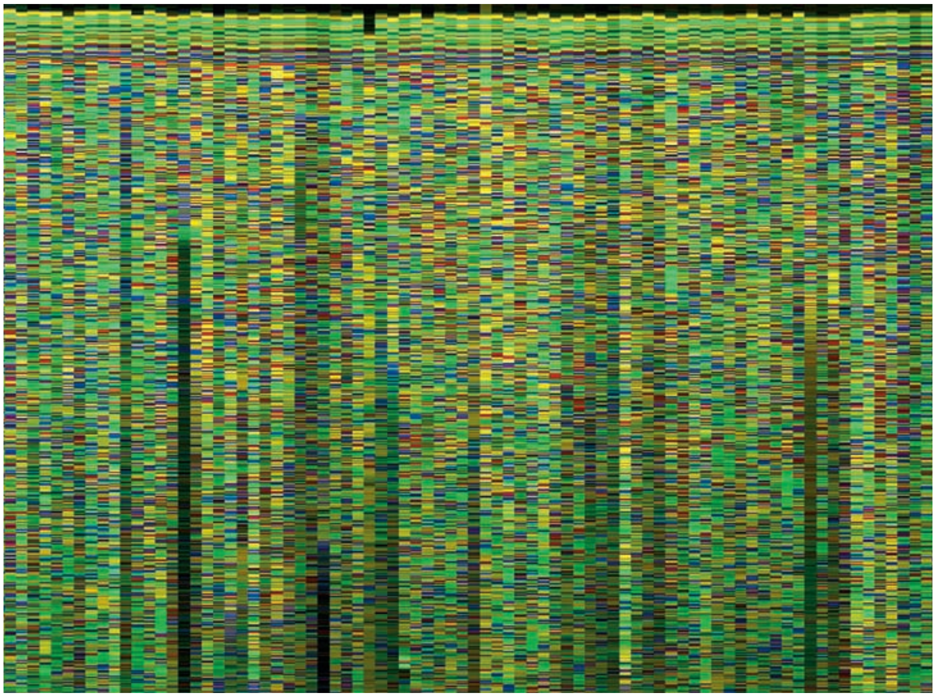
Genomics

20



Output from an automated genome sequencing machine, representing about 48 000 bases from the human genome. Each vertical stripe represents the sequence of a stretch of DNA.

KEY CONCEPTS

- Once a genome has been completely sequenced, researchers use a variety of techniques to identify which sequences code for products and which act as regulatory sites.
- In bacteria and archaea, there is a positive correlation between the number of genes in a species and the species' metabolic capabilities. Gene transfer between species is also common.
- In eukaryotes, genomes are dominated by sequences that have little to no effect on the fitness of the organism.
- Data from genome sequencing projects are now being used in the development of new drugs and vaccines, and to search for alleles associated with inherited diseases.

The first data sets describing the complete DNA sequence, or **genome**, of humans were published in February 2001. These papers were immediately hailed as a landmark in the history of science. In just 50 years, biologists had gone from not understanding the molecular nature of the gene to knowing the molecular makeup of every gene present in our species.

Leading up to the 2001 papers and since then, the multinational effort called the **Human Genome Project** has produced a wide array of data on the locations and functions of genes and other types of DNA sequences found in humans. It's important to recognize, though, that research on *Homo sapiens* is part of a much larger and ongoing effort to sequence genomes from an array of other eukaryotes, hundreds of bacteria, and dozens of archaea. The effort to sequence, interpret, and compare whole genomes is referred to as **genomics**. The pace of research in this field is nothing short of explosive.

The arrival of genomics has also triggered the development of the field called functional genomics. Genomics supplies a list of the genes present in an organism; **functional genomics** asks when those genes are expressed and how their products interact. This effort is important. Besides providing new insights into how gene expression is coordinated as an embryo develops into a university student or an oak tree, functional genomics is helping researchers identify new drug targets in disease-causing species, design new vaccines, and speed efforts to locate the genes responsible for inherited diseases.

As an introductory biology student, you are part of the first generation trained in the genome era. Genomics is revolutionizing several fields within biological science and will almost certainly be an important part of your personal and professional life. Let's look at what genomics is, how and why it's being done, and what is being learned.

20.1 Whole-Genome Sequencing

Genomics has moved to the cutting edge of research in biology largely because a series of technological advances—including the fluorescent markers and gel-filled capillary tubes introduced in Chapter 19—have increased the speed of DNA sequencing and driven down the expense. Thanks largely to increased automation, the cost of sequencing DNA has declined by a factor of two every year and a half since the Human Genome Project began in 1988. Researchers have now established factory-style DNA sequencing centres, each containing dozens of automated sequencing machines, in 18 countries including the United States, Canada, the United Kingdom, Germany, France, Japan, and China. Some of these laboratories employ dozens of biologists and can conduct 100 000 sequencing reactions daily.

As data became less expensive and faster to obtain, the pace of whole-genome sequencing accelerated. The result is that an almost mind-boggling number of sequences are now being generated. As this book goes to press, the primary international repositories for DNA sequence data contain over 146 *billion* nucleotides. (By way of comparison, a haploid human genome contains about 3 billion bases.) The size of this database is increasing by about 20 percent every year (**Figure 20.1**).

How Are Complete Genomes Sequenced?

Genomes range in size from a few million base pairs to several billion. But a single sequencing reaction can analyze only about 1000 base pairs. How do investigators break a genome into sequencing-sized pieces, and then figure out how the thousands or millions of pieces go back together?

Most recent genome sequencing projects answer this question with a whole-genome shotgun approach, or simply **shotgun sequencing**. In shotgun sequencing, a genome is broken up into a set of overlapping fragments that are small enough to be sequenced. Using the regions of overlap, the sequenced fragments are then put back into the correct order.

As step 1 of **Figure 20.2** shows, shotgun sequencing begins by using high-frequency sound waves, or sonication, to break up a genome into pieces about 160 kilobases (kb) long (1 kb = 1000 bases). Next, each 160 kb piece is inserted into a plasmid called a **bacterial artificial chromosome (BAC).** BACs are able to replicate large segments of DNA. Using transformation techniques introduced in Chapter 19, each BAC is then inserted into a different *Escherichia coli* cell, creating what researchers call a BAC library (step 2). A BAC library is a genomic library: a set of all the DNA sequences in a particular genome, split into small segments and inserted into a cloning vector (see Chapter 19). By separating the cells in a BAC library and then allowing each cell to grow into a large colony, researchers can isolate large numbers of each 160 kb fragment.

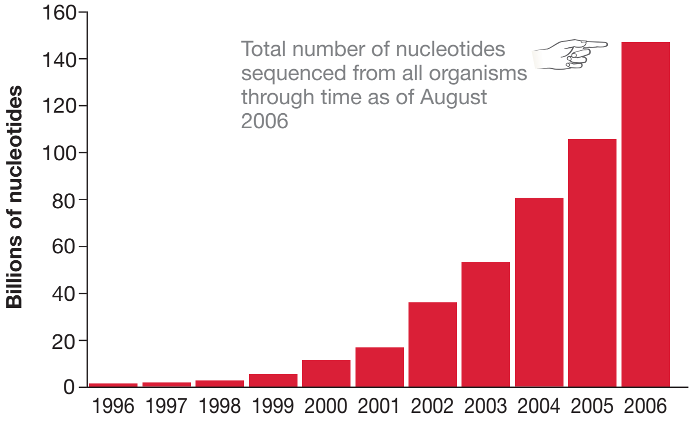


FIGURE 20.1 The Total Number of Bases Sequenced Is Growing Rapidly.

Once a research group has many copies of each 160 kb fragment, the DNA is again broken into fragments—but this time, segments that are about 1000 base pairs long (step 3). These small fragments are then inserted into plasmids and placed inside bacterial cells (step 4). In this way, a genome is broken down into two manageable levels: 160 kb fragments in BACs and 1 kb segments in plasmids. The plasmids are copied many times as the bacterial cells grow into a large population. Large numbers of each 1000-base-pair fragment are then available for sequencing reactions (step 5).

Once the 1000-base-pair fragments from each 160 kb BAC clone are sequenced, computer programs analyze regions where the ends of each 1000-base-pair segment overlap (step 6). Overlaps occur because there were many copies of each 160 kb segment, and each was fragmented randomly by sonication. The computer mixes and matches segments from a single BAC clone until an alignment consistent with all available data is obtained. Then the ends of each BAC are analyzed in a similar way (step 7). The goal is to arrange each 160 kb segment in its correct position along the chromosome, based on regions of overlap.

In essence, the shotgun strategy consists of breaking a genome into tiny fragments, sequencing the fragments, and then putting the sequence data back into the correct order. ● If you understand shotgun sequencing, you should be able to explain why it is essential for regions of overlap to exist between fragments that are adjacent to each other on a chromosome.

Once complete genome sequences became available, databases that could hold completed sequence information had to be created and managed in a way that made the raw data and a variety of annotations available to the international community of researchers. These sequence databases also had to be searchable, so that investigators could evaluate how similar newly discovered genes were to genes that had been studied previously.

Because the amount of data involved is so large, the computational challenges involved in genomics are formidable. Thus far, sophisticated algorithms and continually improving computer hardware have allowed researchers to keep pace with the rate of data acquisition. The vast quantity of data generated by genome sequencing centres has made **bioinformatics**—the effort

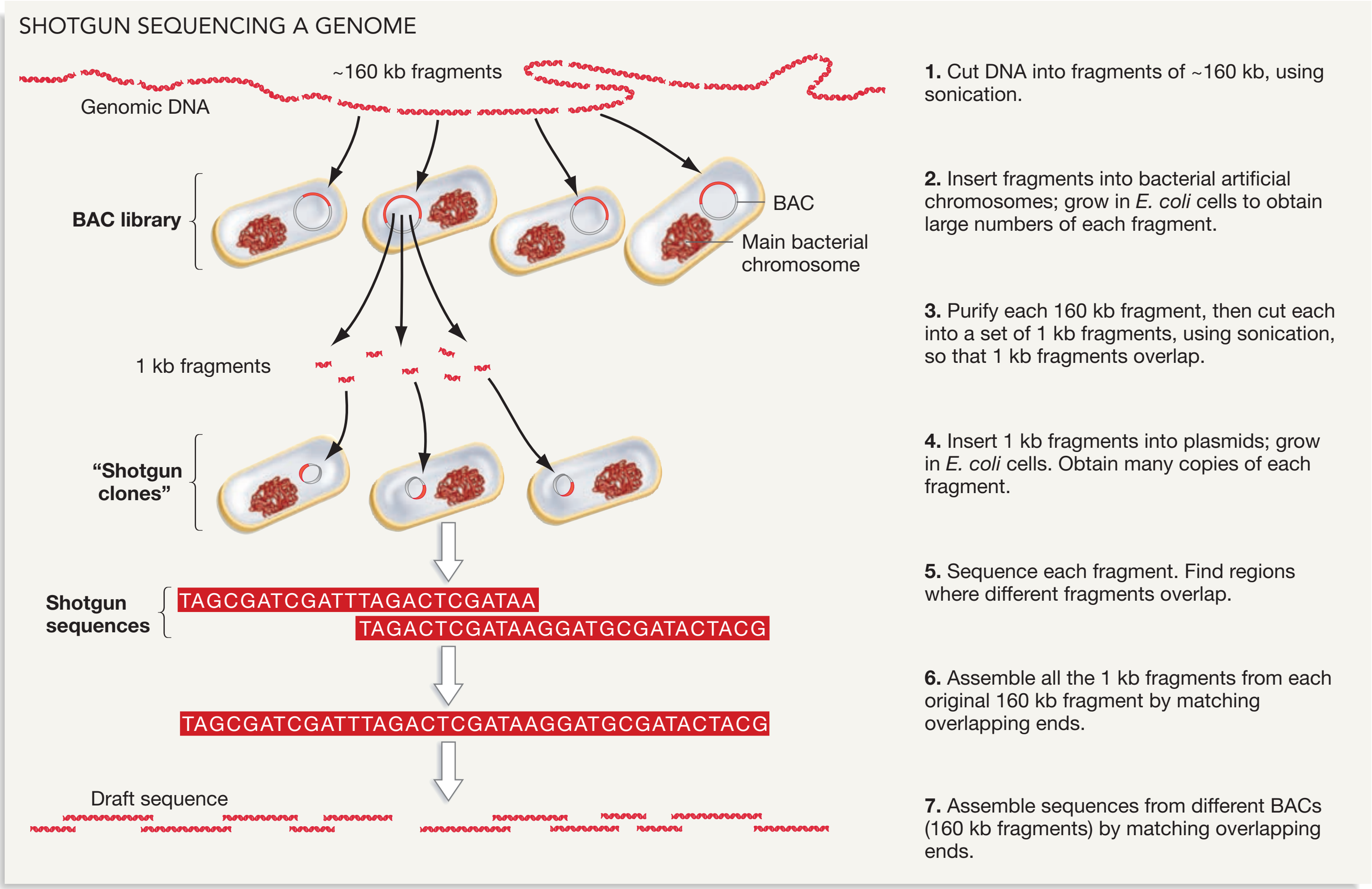

SHOTGUN SEQUENCING A GENOME
~160 kb fragments
Genomic DNA
BAC library
BAC
Main bacterial chromosome
1 kb fragments
"Shotgun clones"
Shotgun sequences
TAGCGATCGATTTAGACTCGATAA
TAGACTCGATAAGGATGCGATACTACG
TAGCGATCGATTTAGACTCGATAAGGATGCGATACTACG
Draft sequence
1. Cut DNA into fragments of ~160 kb, using sonication.
2. Insert fragments into bacterial artificial chromosomes; grow in E. coli cells to obtain large numbers of each fragment.
3. Purify each 160 kb fragment, then cut each into a set of 1 kb fragments, using sonication, so that 1 kb fragments overlap.
4. Insert 1 kb fragments into plasmids; grow in E. coli cells. Obtain many copies of each fragment.
5. Sequence each fragment. Find regions where different fragments overlap.
6. Assemble all the 1 kb fragments from each original 160 kb fragment by matching overlapping ends.
7. Assemble sequences from different BACs (160 kb fragments) by matching overlapping ends.

FIGURE 20.2 Shotgun Sequencing Breaks Large Genomes into Many Short Segments. Shotgun sequencing is a multistep process. A genome is initially fragmented into 160 kb sections via sonication and cloned into bacterial artificial chromosomes (BACs). Each 160 kb section is then cut into 1 kb fragments that can be cloned into plasmids, grown to a large number of copies, and sequenced.

QUESTION Why is "shotgun" an appropriate way to describe this sequencing strategy?

to manage, analyze, and interpret information in biology—a key to continued progress in the field.

Which Genomes Are Being Sequenced, and Why?

The first genome to be sequenced from an organism—not a virus—came from a bacterium that lives in the human upper respiratory tract. This bacterium, *Haemophilus influenzae*, has one circular chromosome and a total of 1 830 138 base pairs of DNA. This genome was small enough to sequence completely with a reasonable amount of time and money, given the technology available in the early 1990s. *H. influenzae* was an important research subject because it causes earaches and respiratory tract infections in children; what's more, one particular strain is capable of infecting the membranes surrounding the brain and spinal cord, causing meningitis.

Publication of the *H. influenzae* genome in 1995 was quickly followed by complete genomes sequenced from an assortment of bacteria and archaea. The first eukaryotic genome, from the yeast *Saccharomyces cerevisiae*, was finished in 1996. After that breakthrough, complete genome sequences were published from a variety of protists, plants, and animals. To date, complete genomes have been sequenced from over 510 bacterial species, 47 archaeal species, and 52 eukaryotic species. Incomplete genome sequence data are available from more than 200 other species.

Most of the organisms that have been selected for whole-genome sequencing cause disease or have other interesting biological properties. For example, genomes of bacteria and archaea that inhabit extremely hot environments have been sequenced in the hopes of discovering enzymes useful for high-temperature industrial applications and understanding how proteins can work under those conditions. Other bacteria and archaea were chosen for sequencing because they do interesting chemistry, such as producing methane (natural gas; CH_4) or other compounds. In some cases, researchers hope that these organisms might act as an important source of commercial products. The rice genome was sequenced because rice is the main food source for most humans. Finally, species such as the fruit fly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans*, the house mouse *Mus musculus*, and the mustard plant *Arabidopsis thaliana* were analyzed because they serve as model organisms in biology and because data from well-studied organisms promised to help researchers interpret the human genome.

Which Sequences Are Genes?

Obtaining raw sequence data is just the beginning of the effort to understand a genome. As researchers point out, raw sequence data are analogous to the parts list for a house. The parts list reads something like "windowwabeborogovestaircasedoorjubjub ...," however, because it has no punctuation. Where do the genes for "window," "staircase," and "door" start and end? Are the segments that read "wabeborogove" and "jubjub" important in gene regulation, or are they simply spacers or other types of sequences that have no function at all?

The most basic task in annotating or interpreting a genome is to identify which bases constitute genes—the segments of DNA that code for an RNA or a protein product and that regulate their production. In bacteria and archaea, identifying genes is relatively straightforward. It is much more difficult, however, in eukaryotes.

Identifying Genes in Bacterial and Archaeal Genomes

Biologists begin with computer programs that scan the sequence of a genome in both directions. These programs identify each reading frame that is possible on the two strands of the DNA. (Recall from Chapter 15 that a reading frame is the sequence in which codons are read.) With codons consisting of three bases, three reading frames are possible on each strand, for a total of six possible reading frames (**Figure 20.3**). Because randomly generated sequences contain a stop codon about one in every 20 codons on average, a long stretch of codons that lacks a stop codon is a good indication of a coding sequence. The computer program highlights any "gene-sized" stretches of sequence that lack a stop codon but are flanked by a stop codon and a start codon. Because polypeptides range in size from a few dozen amino acids to many hundreds of amino acids, gene-sized stretches of sequence range from several hundred bases to thousands of bases. ● In addition, the computer programs look for sequences typical of promoters, operators, or other regulatory sites. DNA segments that are identified in this way are called **open reading frames**, or **ORFs**.

Once an ORF is found, a computer program compares its sequence with the sequences of known genes from well-studied species. If the ORF appears to be a gene that has not yet been described in any other species, further research is required before it can actually be considered a gene. A "hit," in contrast, means that the ORF shares a significant amount of sequence with a known gene from another species. Similarities between genes in different species are usually due to **homology**. If genes are homologous, it means they are similar because they are related by descent from a common ancestor. Homologous genes have similar base sequences and the same or a similar function. For example, consider the genes introduced in Chapter 14 that code for enzymes involved in repairing mismatches in DNA. Recall that the mismatch repair genes in *E. coli*, yeast, and humans are extremely similar in structure, DNA sequence, and function. To explain this similarity, biologists hypothesize that the common ancestor of all cells living today had mismatch repair genes—thus, the descendants of this ancestral species also have versions of these genes.

Based on this logic, researchers can confirm that an ORF is actually a gene by finding that it is homologous to a known gene. They can also analyze the product that would be produced by an ORF, and see if it conforms to a known gene.

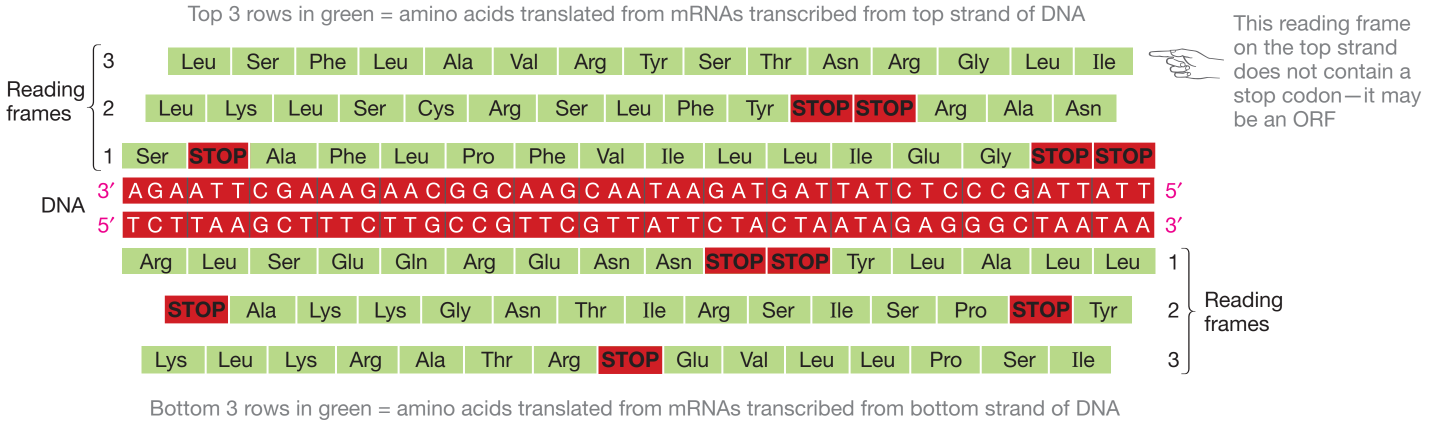

Top 3 rows in green = amino acids translated from mRNAs transcribed from top strand of DNA
Reading frames
3 Leu Ser Phe Leu Ala Val Arg Tyr Ser Thr Asn Arg Gly Leu Ile
2 Leu Lys Leu Ser Cys Arg Ser Leu Phe Tyr STOP STOP Arg Ala Asn
1 Ser STOP Ala Phe Leu Pro Phe Val Ile Leu Leu Ile Glu Gly STOP STOP
This reading frame on the top strand does not contain a stop codon—it may be an ORF
DNA
3′ AGAATTCGAAAGAACGGCAAGCAATAAGATGATTATCTCCCGATTATT 5′
5′ TCTTAAGCTTTCTTGCCGTTCGTTATTCTACTAATAGAGGGCTAATAA 3′
Arg Leu Ser Glu Gln Arg Glu Asn Asn STOP STOP Tyr Leu Ala Leu Leu 1
STOP Ala Lys Lys Gly Asn Thr Ile Arg Ser Ile Ser Pro STOP Tyr 2
Lys Leu Lys Arg Ala Thr Arg STOP Glu Val Leu Leu Pro Ser Ile 3
Reading frames
Bottom 3 rows in green = amino acids translated from mRNAs transcribed from bottom strand of DNA

FIGURE 20.3 Open Reading Frames May Be the Locations of Genes. Computer programs scan the three possible reading frames on each strand of DNA and use the genetic code to translate each codon. A long stretch of codons that contains a start codon but lacks a stop codon may be an open reading frame (ORF)—a possible gene.

● **QUESTION** To predict the mRNA codons that would be produced by a particular reading frame, a computer analyzes the DNA in the 3′ to 5′ direction. Why?

Unfortunately, finding and analyzing genes by identifying ORFs does not work well in eukaryotes.

Identifying Genes in Eukaryotic Genomes Mining eukaryotic sequence data for genes is complicated by two observations: Coding regions are broken up by introns, and the vast majority of eukaryotic DNA does not actually code for a product. In the human genome, for example, it is estimated that less than 2 percent of the DNA present actually codes for proteins, tRNAs, ribosomal RNAs, or other types of products. Finding a coding region in eukaryotic DNA is like finding a diamond in a huge pile of rock. To do so, researchers pursue several strategies:

- Computer programs can be written to search for sequences that are homologous to known genes. If a stretch of bases in the newly sequenced genome is similar to the sequence of a known gene, then researchers infer that it codes for a product whose function is similar to the function of the known gene.
- As Chapter 19 showed, investigators can isolate mRNAs from the organism being studied and then use enzymes to make the complementary DNAs (cDNAs). If the sequence of these cDNAs is determined, then a computer program can scan the genome sequence and pinpoint where each of the cDNAs is located. This approach allows researchers to identify genes that are expressed in certain cell types—the tissues where the original mRNA was found.
- To identify genes that have no known function, computers compare the genomes of closely related species and highlight sequences that are similar. Sequences that are shared by closely related species are hypothesized to be located in the protein-coding or regulatory regions of genes. The logic behind this claim runs as follows: Sequences that are part of a gene are expected to change much more slowly over time than sequences that are not actually part of a gene. Gene sequences change slowly over time because most gene products work less efficiently when they change randomly by mutation. Thus, it is logical to expect that natural selection eliminates most mutations in genes and that genes should change slowly over time. But changes in sequences that do not code for products or regulate gene expression do not affect the organism's phenotype. Mutations in these regions are much less likely to be eliminated by natural selection, so they change relatively rapidly over time.

Although each of these gene-finding strategies has been productive, it will probably be many years before biologists are convinced they have identified all of the coding regions in even a single eukaryotic genome. As that effort continues, though, researchers are analyzing the data and making some remarkable observations. Let's first consider what genome sequencing has revealed about the nature of bacterial and archaeal genomes and then move on to eukaryotes. Is the effort to sequence whole genomes paying off?

Web Animation at www.masteringbio.com
Human Genome Sequencing Strategies

20.2 Bacterial and Archaeal Genomes

By the time you read this paragraph, the genomes of over 600 bacterial and archaeal species will have been sequenced. In addition to this impressive array of different species, complete genome sequences are now available for several strains of the same bacterial species. For example, researchers have sequenced the genome of a laboratory population of *Escherichia coli*—derived from the harmless strain that lives in your gut—as well as the genome of a form that causes severe disease in humans. As a result, researchers can now compare the genomes of closely related cells that have different ways of life.

This section focuses on a simple question: Based on data published between 1995 and 2007, what general observations have biologists been able to make about the nature of bacterial and archaeal genomes?

The Natural History of Prokaryotic Genomes

In a sense, biologists who are working in genomics can be compared to the naturalists of the eighteenth and nineteenth centuries. These early biologists explored the globe, collecting the plants and animals they encountered. Their goal was to describe what existed. Similarly, the first task of a genome sequencer is to catalogue what is in a genome—specifically, the number, type, and organization of genes. Several interesting conclusions can be drawn from relatively straightforward observations about the data obtained thus far:

- In bacteria, there is a general correlation between the size of a genome and the metabolic capabilities of the organism. For example, most parasites have much smaller genomes than nonparasitic organisms do. **Parasites** live off a host and thus reduce the host's fitness—its ability to produce offspring. Some of the smallest genomes are found in parasitic bacteria from the genus *Mycoplasma*. These bacteria live and multiply inside host cells. *Mycoplasma pneumoniae*, for example, parasitizes lung cells and causes pneumonia in humans. *Mycoplasma* lack the enzymes required to manufacture many essential compounds. Instead they acquire almost all of their nutrients from their hosts. In contrast, the genomes of nonparasitic strains of the bacteria *E. coli* and *Pseudomonas aeruginosa* are 8 to 10 times larger. Their genes code for enzymes that synthesize virtually every molecule needed by the cell. Based on this observation, it is not surprising that *E. coli* is able to grow under a wide variety of environmental conditions. Using similar logic, researchers hypothesize that the large genome of *P. aeruginosa* explains why it is able to occupy a wide array of soil types, including marine and marshy habitats, as well as human tissues, where it can cause illness.

- Biologists still do not know the function of many of the genes that have been identified. Although *E. coli* probably qualifies as the most intensively studied of all organisms, the function of over 30 percent of its genes is unknown.
- There is tremendous genetic diversity among bacteria and archaea. About 15 percent of the genes in each species' genome appear to be unique. That is, about one in six genes in one of these species is found nowhere else.
- Redundancy among genes is common. For instance, the genome of *E. coli* has 86 pairs of genes whose DNA sequences are nearly identical—meaning that the proteins they produce are nearly alike in structure and presumably in function. Although the significance of this redundancy is unknown, biologists hypothesize that slightly different forms of the same protein are produced in response to slight changes in environmental conditions.
- Multiple chromosomes are more common than anticipated. Several species of bacteria and archaea have two different circular chromosomes instead of one. And at least some bacteria have linear chromosomes.
- Many species contain the small, extrachromosomal DNA molecules called plasmids. Recall from Chapter 19 that plasmids contain a small number of genes, though not genes that are absolutely essential for growth. In many cases, plasmids can be exchanged between cells of the same species or even of different species (see Chapter 12).

Perhaps the most surprising observation of all is that in many bacterial and archaeal species, a significant proportion of the genome appears to have been acquired from other, often distantly related, species. In some bacteria and archaea, 15–25 percent of the genetic material appears to be "foreign." This is a remarkable claim. What evidence backs up the assertion that prokaryotes acquire DNA from other species? How could this happen, and what are the consequences?

Evidence for Lateral Gene Transfer

Biologists use two general criteria to support the hypothesis that sequences in bacterial or archaeal genomes originated in another species: (1) when stretches of DNA are much more similar to genes in distantly related species than to those in closely related species and (2) when the proportion of G-C base pairs to A-T base pairs in a particular gene or series of genes is markedly different from the base composition of the rest of the genome. In many cases, the proportion of G-C bases in a genome is characteristic of a genus or species.

How can genes move from one species to another? In at least some cases, plasmids appear to be responsible. For example, most of the genes that are responsible for conferring resistance to antibiotics are found on plasmids. Researchers have documented the transfer of plasmid-borne antibiotic-resistance genes between very distantly related species of disease-causing bacteria. In some cases, genes from plasmids become integrated into the main chromosome of a bacterium, resulting in genetic recombination (see Chapter 12). The movement of DNA from one species to another species is called **lateral gene transfer** (**Figure 20.4**).

Some biologists hypothesize that lateral gene transfer also occurs via transformation—when bacteria and archaea take up raw pieces of DNA from the environment, perhaps in the course of acquiring other molecules. This may have occurred in the bacterium *Thermotoga maritima*, which occupies the high-temperature environments near deep-sea vents. Almost 25 percent of the genes in this species are extremely closely related to genes found in archaea that live in the same habitats. The archaea-like genes occur in distinctive clusters within the *T. maritima* genome, which supports the hypothesis that the sequences were transferred in large pieces from an archaean to the bacterium.

Similar types of direct gene transfer are hypothesized to have occurred in the bacterium *Chlamydia trachomatis*. This organism is a major cause of blindness in humans from Africa and Asia; it also causes chlamydia, the most common sexually transmitted bacterial disease in the United States. The *C. trachomatis* genome contains 35 genes that resemble eukaryotic genes in structure. Because *C. trachomatis* lives inside the cells that it parasitizes, the most logical explanation for this observation is that the bacterium occasionally takes up DNA directly from its host cell, resulting in a eukaryote-to-bacterium transfer.

In addition to being transferred between species by means of plasmids or DNA fragments, genes can be transported by viruses. For example, investigators who compared the sequences of laboratory and pathogenic (disease-causing) strains of *E. coli* found that the pathogenic cells have almost 1400 "extra" genes. Compared with the rest of the genome, most of these genes have a distinctive G-C to A-T ratio. Many are also extremely similar to sequences isolated from viruses that infect *E. coli*. Based on these observations, most researchers support the hypothesis that at least some of the disease-causing genes in *E. coli* were brought in by viruses.

To summarize, mutation and genetic recombination within species are not the only source of genetic variation in bacteria and archaea. Over the course of evolution, lateral gene transfer has been an important source of new genes and allelic diversity in these domains. This insight would not have been possible without data from whole-genome sequencing. Have efforts to sequence eukaryotic genomes led to similar types of insights?

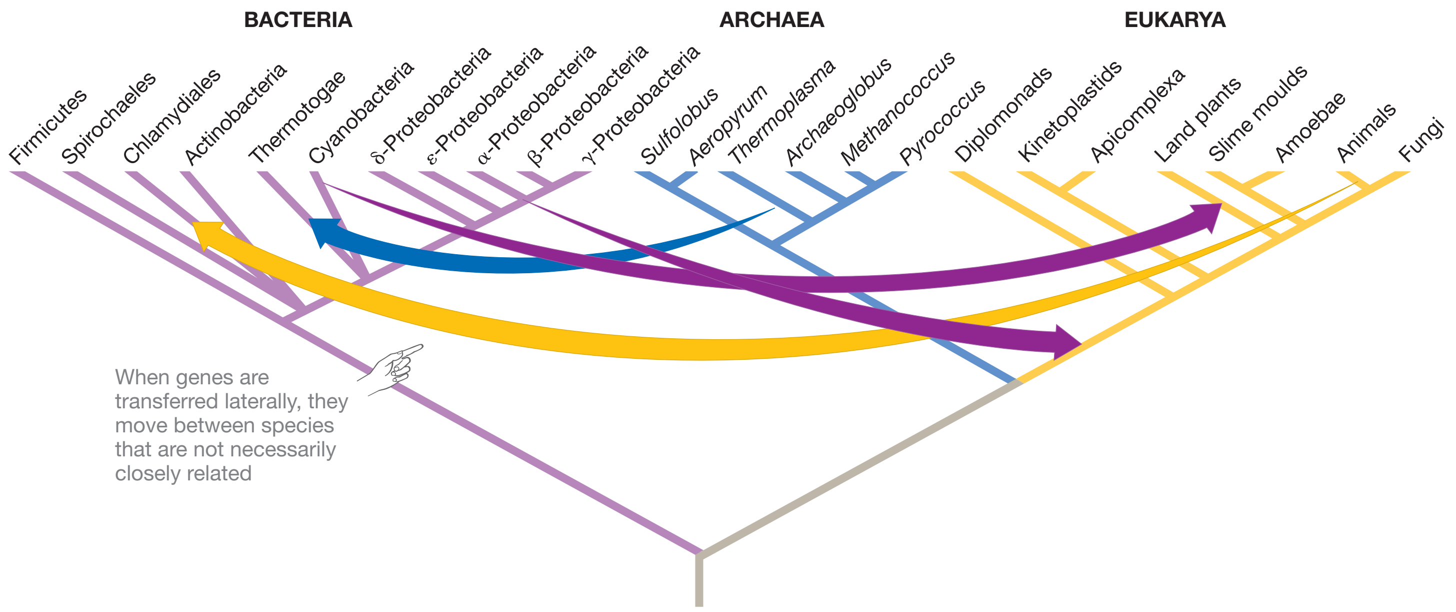


FIGURE 20.4 Lateral Gene Transfer Is Movement of DNA between Species. Lateral gene transfer can occur between very distantly related organisms. For tips on how to read an evolutionary tree like this, see **BioSkills 2**.

EXERCISE Beside this tree, list the mechanisms responsible for lateral gene transfer.

Check Your Understanding

If you understand that…

- The major surprise that came out of genome sequencing projects involving bacteria and archaea was the extent and importance of lateral gene transfer—that is, the movement of DNA from one species to another.

You should be able to…

1) Summarize how researchers infer that lateral gene transfer is responsible for the presence of a particular DNA sequence.
2) Summarize evidence that the size of a prokaryotic genome is correlated with the organism's metabolic capabilities.

20.3 Eukaryotic Genomes

Sequencing eukaryotic genomes presents two daunting challenges. The first is sheer size. Compared with the genomes of bacteria and archaea, which range from 580 070 base pairs in *Mycoplasma genitalium* to over 6.3 million base pairs in *Pseudomonas aeruginosa*, eukaryotic genomes are even larger. The haploid genome of *Saccharomyces cerevisiae* (baker's yeast), a unicellular eukaryote, contains over 12 million base pairs. The roundworm *Caenorhabditis elegans* has a genome of 97 million base pairs; the fruit-fly genome contains 180 million base pairs; the mustard plant *Arabidopsis thaliana*'s genome has 130 million base pairs; and humans, rats, mice, and cattle contain roughly 3 billion base pairs each.

The second great challenge in sequencing eukaryotic genes is coping with noncoding sequences that are repeated many times. Many eukaryotic genomes are dominated by repeated DNA sequences that occur between genes and do not code for products used by the organism. These repeated sequences pose serious problems in aligning and interpreting sequence data. What are they? If such sequences don't code for a product, why do they exist?

Natural History: Types of Sequences

In many eukaryotic genomes, the exons and regulatory sequences associated with genes make up a relatively small percentage of the genome. Recall from Section 20.1 that in humans, protein-coding sequences constitute less than 2 percent of the total genome while repeated sequences account for well over 50 percent. In contrast, over 90 percent of a bacterial or archaeal genome consists of genes—DNA sequences that code for a product needed by the cell and regulate its transcription.

When noncoding and repeated sequences were discovered, they were initially considered "junk DNA" that was nonfunctional and probably unimportant and uninteresting. But subsequent work has shown that many of the repeated sequences observed in eukaryotes are actually derived from sequences known as transposable elements. **Transposable elements** are segments of DNA that are capable of moving from one location to another, or transposing, in a genome. They are similar to viruses, except that viruses leave a host cell that they have infected and find a new cell to infect. In contrast, transposable elements never leave their host cell—they simply make copies of themselves and move to new locations in the genome. Transposable elements are passed from parents to offspring, generation after generation, because they are part of the genome.

Transposable elements are examples of what biologists call selfish genes. A selfish gene is a DNA sequence that survives and reproduces but does not increase the fitness of the host genome. Transposable elements and viruses are classified as parasitic because it takes time and resources to copy them along with the rest of the genome, and because they can disrupt gene function when they move and insert in a new location. As a result, they decrease their host's fitness. Transposable elements are genomic parasites.

How Do Transposable Elements Work? Transposable elements come in a wide variety of types and spread through genomes in a variety of ways. Different species—fruit flies, yeast, and humans, for example—contain distinct types of transposable elements.

As an example of how these selfish genes work, let's consider a well-studied type called a **long interspersed nuclear element (LINE)** that is found in humans and other eukaryotes. Because LINEs are so similar to the retroviruses, introduced in Chapter 19, biologists hypothesize that they are derived from them evolutionarily. Your genome contains tens of thousands of LINEs, each between 1000 and 5000 bases long.

An active LINE contains all the sequences required for it to make copies of itself and insert them into a new location in the genome (**Figure 20.5**, step 1): a gene that codes for the enzyme reverse transcriptase, a gene that codes for the enzyme integrase, and a single promoter that is recognized by RNA polymerase II (step 2). After a LINE is transcribed to an mRNA, reverse transcriptase and integrase are synthesized by ribosomes in the cytoplasm (steps 3 and 4). Reverse transcriptase makes a cDNA version of the LINE mRNA, and integrase inserts the newly synthesized LINE DNA into a new location in the genome (step 5). In this way, the parasitic sequence reproduces (step 6). If the transposition event occurs in reproductive cells that go on to form eggs or sperm, the copied LINE will be passed on to offspring. If the LINE happens to insert itself inside a gene or a regulatory sequence, it causes a mutation that is almost certain to reduce the host's fitness.

Most of the LINEs observed in the human genome do not actually function, however, because they don't contain a promoter or the genes for either reverse transcriptase or integrase. To make sense of this observation, researchers hypothesize that the insertion process illustrated in steps 6 and 7 of Figure 20.5 is usually disrupted in some way. Analyses of the human genome have revealed that only a handful of our LINEs appear to be complete and potentially active.

Virtually every prokaryotic and eukaryotic genome examined to date contains at least some transposable elements. They vary

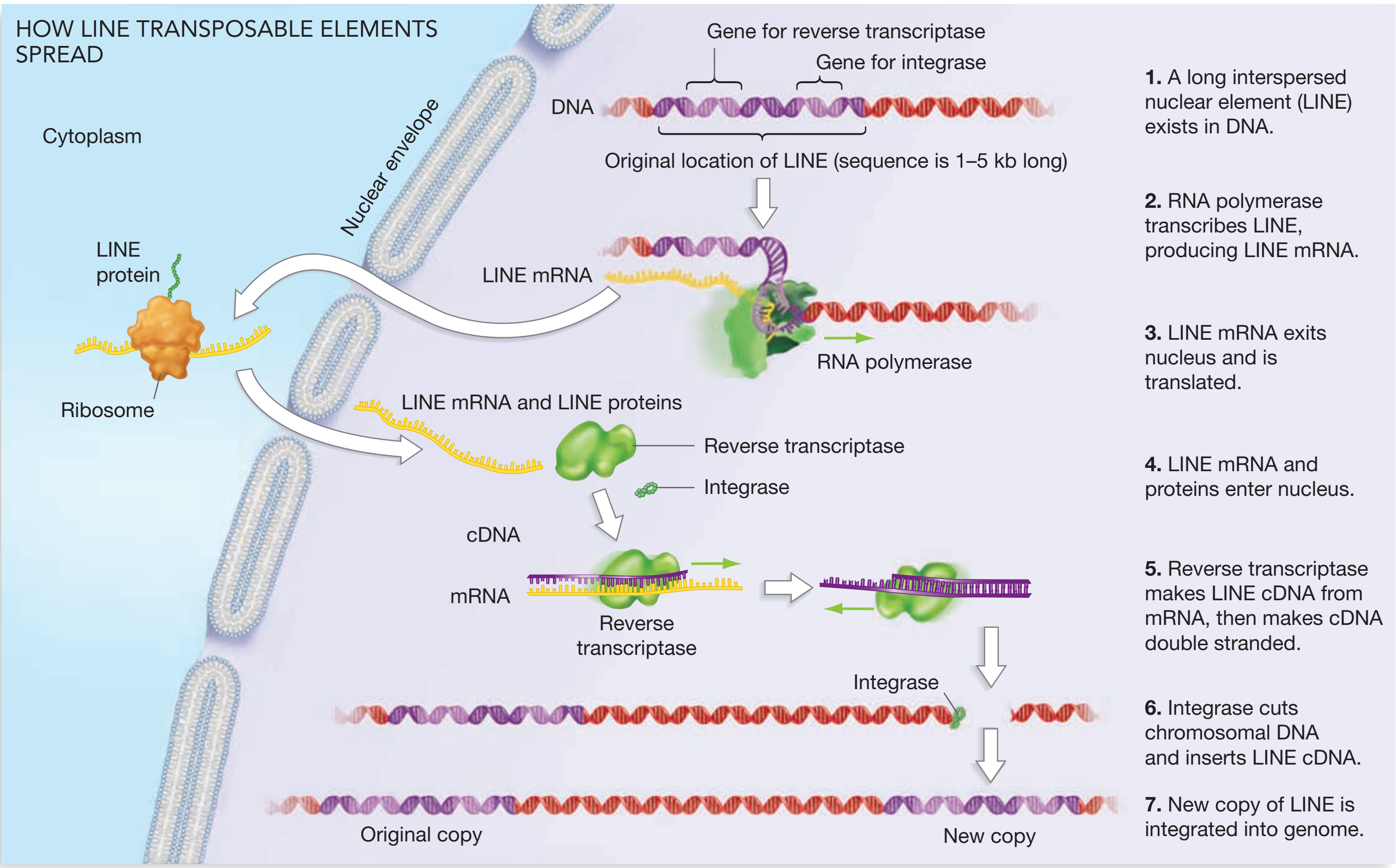

HOW LINE TRANSPOSABLE ELEMENTS SPREAD
Cytoplasm
Nuclear envelope
Gene for reverse transcriptase
Gene for integrase
DNA
Original location of LINE (sequence is 1–5 kb long)
LINE mRNA
RNA polymerase
LINE protein
Ribosome
LINE mRNA and LINE proteins
Reverse transcriptase
Integrase
cDNA
mRNA
Reverse transcriptase
Integrase
Original copy
New copy
1. A long interspersed nuclear element (LINE) exists in DNA.
2. RNA polymerase transcribes LINE, producing LINE mRNA.
3. LINE mRNA exits nucleus and is translated.
4. LINE mRNA and proteins enter nucleus.
5. Reverse transcriptase makes LINE cDNA from mRNA, then makes cDNA double stranded.
6. Integrase cuts chromosomal DNA and inserts LINE cDNA.
7. New copy of LINE is integrated into genome.

FIGURE 20.5 Transposable Elements Spread within a Genome.

widely in type and number, however, and bacterial and archaeal genomes have relatively few transposable elements compared to most eukaryotes studied thus far. This observation has inspired the hypothesis that bacteria and archaea either have efficient means of removing parasitic sequences or can somehow thwart insertion events. To date, however, this hypothesis has yet to be tested rigorously.

Research on transposable elements and lateral gene transfer has revolutionized how biologists view the genome. Many genomes are riddled with parasitic sequences, and others have undergone radical change in response to lateral gene transfer events. Genomes are much more dynamic and complex than previously thought. Their size and composition can change dramatically over time.

Repeated Sequences and DNA Fingerprinting In addition to containing repeated sequences from transposable elements, eukaryotic genomes have several thousand loci called simple tandem repeats (STRs). These are small sequences repeated one after another down the length of a chromosome. There are two major classes of STRs: Repeating units that are just 1 to 5 bases long are known as **microsatellites** or **simple sequence repeats**; repeating units that are 6 to 500 bases long are called **minisatellites** or **variable number tandem repeats** (**VNTRs**). Both types of repeated sequences make up 3 percent of the human genome. The most common type of microsatellite is a repeated stretch of the dinucleotide AC, giving the sequence ACACACAC.... Microsatellite sequences are thought to originate when DNA polymerase skips or mistakenly adds extra bases during replication; the origin of minisatellites is still unclear.

Soon after these sequences were first characterized, Alec Jeffreys and co-workers established that microsatellite and minisatellite loci are "hypervariable," meaning that they vary among individuals much more than any other type of sequence does. **Figure 20.6** illustrates one hypothesis for why microsatellites and minisatellites have so many different alleles: These highly repetitive stretches often align incorrectly when homologous chromosomes synapse and cross over during prophase of meiosis I. Instead of lining up in exactly the same location, the two chromosomes pair in a way that matches up bases in different repeated segments. Due to this misalignment, **unequal crossover** occurs. Chromosomes produced by unequal crossover contain different numbers of repeats. The key observation is that if a particular microsatellite or minisatellite locus has a unique number of repeats, it represents a unique allele. Each allele has a unique length. As with any allele, microsatellite and minisatellite alleles are transmitted from parents to offspring.

Misalignment or errors by DNA polymerase are so common at these loci that, in most eukaryotes, the genome of virtually every individual has at least one new allele. This variation in repeat number among individuals is the basis of DNA fingerprinting. **DNA fingerprinting** refers to any technique for identifying individuals based on the unique features of their genomes. Because microsatellite and minisatellite loci vary so much among individuals, they are now the loci of choice for DNA fingerprinting. To fingerprint an individual, researchers obtain a DNA sample and perform the polymerase chain reaction (PCR), using primers that flank a region containing an STR. Once many copies of the region are available, they can be analyzed to determine the number of repeats present (**Figure 20.7a**). Primers are now available for many different STR loci, so

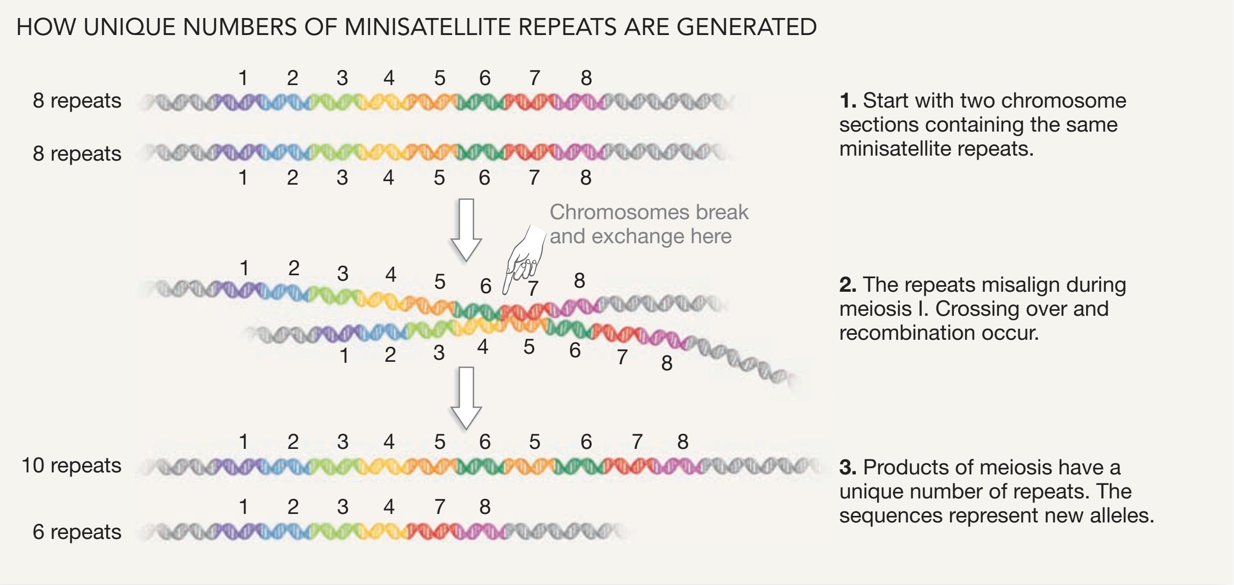
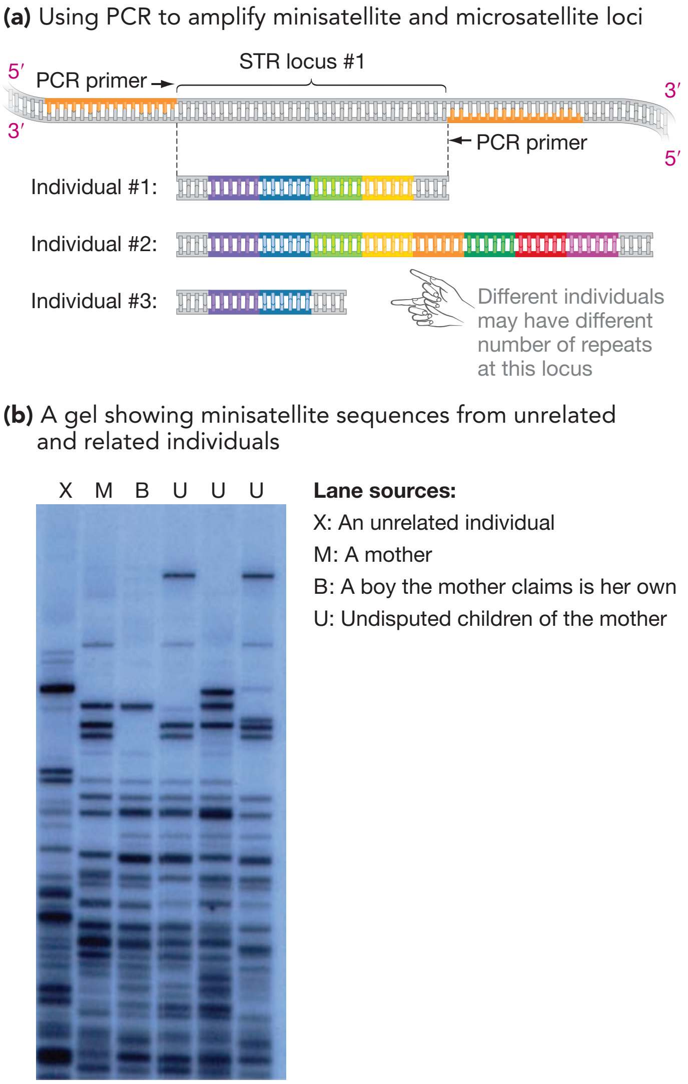

HOW UNIQUE NUMBERS OF MINISATELLITE REPEATS ARE GENERATED
8 repeats
8 repeats
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
1. Start with two chromosome sections containing the same minisatellite repeats.
Chromosomes break and exchange here
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
2. The repeats misalign during meiosis I. Crossing over and recombination occur.
10 repeats
1 2 3 4 5 6 5 6 7 8
6 repeats
1 2 3 4 7 8
3. Products of meiosis have a unique number of repeats. The sequences represent new alleles.

FIGURE 20.6 Unequal Crossover Changes the Numbers of Simple Sequence Repeats. The alignment of homologous chromosomes during prophase of meiosis I is driven by sequence similarity between homologs. Because simple sequence repeats are so similar, they are likely to misalign during synapsis.


(a) Using PCR to amplify minisatellite and microsatellite loci

5′
PCR primer
STR locus #1
3′
3′
PCR primer
5′
Individual #1:
Individual #2:
Individual #3:
Different individuals may have different number of repeats at this locus

(b) A gel showing minisatellite sequences from unrelated and related individuals

X M B U U U
Lane sources:
X: An unrelated individual
M: A mother
B: A boy the mother claims is her own
U: Undisputed children of the mother

FIGURE 20.7 DNA Fingerprinting Can Be Used to Identify Fathers.
(a) If a minisatellite locus contains different numbers of repeats in different individuals, then DNA fragments from those individuals will have different lengths. **(b)** A gel containing DNA fragments from a mother, a purported biological child of the woman, and several of her confirmed biological children was probed with sequences from a minisatellite locus. Related individuals tend to share fragment patterns.

EXERCISE Circle fragments in the M, B, and U lanes that support the hypothesis that the disputed boy actually is M's son.

researchers can analyze the alleles present at many STRs efficiently.

Research on repeated sequences has revealed that the probability of getting a new allele is higher for shorter repeats than for longer repeats. For some two-base-pair repeats, the number of repeats present changes so quickly over time that only very close relatives are likely to share any of the same alleles. This observation has important practical implications. For example, DNA fingerprinting of blood or semen found at crime scenes has been used to show that people who were accused of crimes were actually innocent. DNA fingerprinting has also been used as evidence to convict criminals or assign paternity in birds, humans, and other species that have well-characterized microsatellite or minisatellite sequences (**Figure 20.7b**).

Now that we've reviewed the characteristics of some particularly prominent types of noncoding sequences in eukaryotes, it's time to consider the nature of the coding sequences in these genomes. Let's start with the most basic question of all: Where do eukaryotic genes come from?

Gene Duplication and the Origin of Gene Families

In eukaryotes, the major source of new genes is the duplication of existing genes. Biologists infer that genes have been duplicated recently when they find groups of similar genes clustered along the same chromosome. The genes are usually similar in structural aspects, such as the arrangement of exons and introns, and in their base sequence. The degree of sequence similarity among these clustered genes varies. In the genes that code for ribosomal RNAs (rRNAs) in vertebrates, the sequences are virtually identical—meaning that each individual has many exact copies of the same gene. In other cases, though, the proportion of bases that are identical is 50 percent or less.

Within a species, genes that are extremely similar to each other in structure and function are considered to be part of the same **gene family**. Genes that make up gene families are hypothesized to have arisen from a common ancestral sequence through gene duplication. When **gene duplication** occurs, an extra copy of a gene is added to the genome.

The most common type of gene duplication results from crossover during meiosis. As **Figure 20.8** shows, gene-sized segments of chromosomes can be deleted or duplicated if homologous chromosomes misalign during prophase of meiosis I and an unequal crossover occurs. The duplicated segments resulting from unequal crossover are arranged in tandem—one after the other.

Gene duplication is important because the original gene is still functional and produces a normal product. As a result, the new, duplicated stretches of sequence are redundant. In some cases the duplicated genes retain their original function and provide additional quantities of the same product. But if mutations in the duplicated sequence alter the protein product, and if the altered protein product performs a valuable new function in the cell, then an important new gene has been created. The duplicated gene may also be regulated in a different way, so that it is expressed in novel locations or at novel times. In either case, the duplicated sequences represent new genes and can lead to the evolution of novel traits. Gene duplication produces new genes and creates gene families.

Alternatively, mutations in the duplicated region may make expression of the new gene impossible. For example, a mutation could produce a stop codon in the middle of an exon. A member of a gene family that resembles a working gene but does not code for a functional product, due to early stop codons

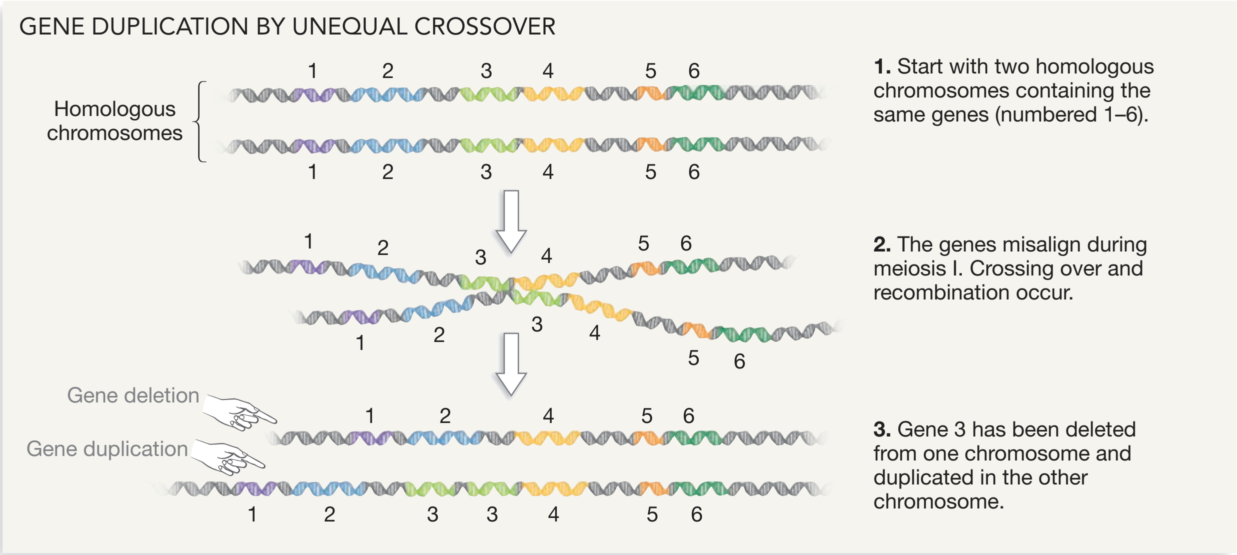

GENE DUPLICATION BY UNEQUAL CROSSOVER
Homologous chromosomes
1 2 3 4 5 6
1 2 3 4 5 6
1. Start with two homologous chromosomes containing the same genes (numbered 1–6).
1 2 3 4 5 6
1 2 3 4 5 6
2. The genes misalign during meiosis I. Crossing over and recombination occur.
Gene deletion
Gene duplication
1 2 4 5 6
1 2 3 3 4 5 6
3. Gene 3 has been deleted from one chromosome and duplicated in the other chromosome.

FIGURE 20.8 Unequal Crossover Changes the Numbers of Genes along a Chromosome. If unequal crossover occurs as diagrammed here, the resulting chromosomes contain one gene fewer than the original chromosome or an additional copy of a gene.

or other defects, is called a **pseudogene**. Pseudogenes have no function.

As an example of a gene family, consider the human globin genes diagrammed in **Figure 20.9**. These genes code for proteins that form part of hemoglobin—the oxygen-carrying molecule in your red blood cells. Analyzing the globin genes illustrates several important points about gene families. In humans, the globin gene family contains several pseudogenes, along with several genes that code for oxygen-transporting proteins. The various coding genes in the family serve slightly different functions. For example, some genes are active only in the fetus or the adult. Follow-up work showed that oxygen is much more likely to bind to the proteins encoded by the fetal genes compared to the proteins expressed in adults. As a result, oxygen is able to move from the mother's blood to the fetus's blood (see Chapter 44).

In addition to the gene duplication events resulting from unequal crossover, the entire complement of chromosomes may be duplicated due to a mistake in either mitosis or meiosis. In this case, the resulting cell contains double the normal complement of chromosomes. Recall from Chapter 12 that species with duplicated chromosome complements are said to be **polyploid**. When polyploidy occurs, every gene in the duplicated genome is redundant. As a result, each gene may experience mutations that lead to new functions or to the loss of function and creation of a pseudogene.

By comparing how many copies of gene families occur in eukaryotes that have completely sequenced genomes, researchers have concluded that a whole-genome duplication occurred early in the evolution of vertebrates. They further conclude that another genome duplication occurred early in the evolution of the ray-finned fish—a lineage of over 24 000 living species, including familiar groups such as the trout, tuna, and guppies. Genome duplication has also been a particularly important source of new genes in plants.

Insights from the Human Genome Project

The human genome is rapidly becoming the most intensively studied of all eukaryotic genomes. In many or even most cases, researchers are gaining insights into how the human genome works by comparing it to the genomes of other species.

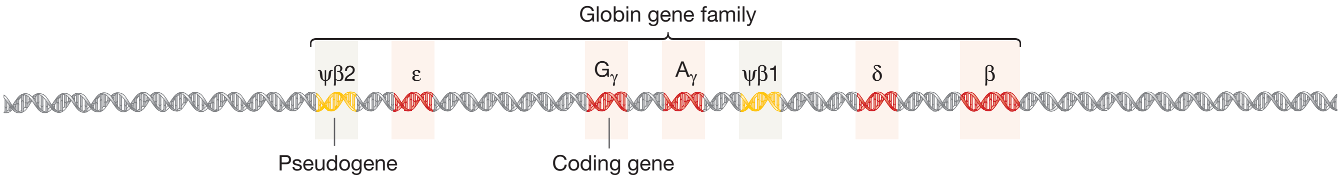


FIGURE 20.9 Gene Families Are Clusters of Closely Related Genes. Genes within the globin family. Red segments represent functioning genes, and yellow segments are pseudogenes. The members of a gene family are arranged one after the other, or in tandem. Most of these genes are expressed at different times during development.

● **EXERCISE** Suppose that during prophase of meiosis I, the β locus on one chromosome aligned with the $\psi\beta2$ locus on another chromosome, and then crossing over occurred just to their left. Draw the chromosomes that would result.

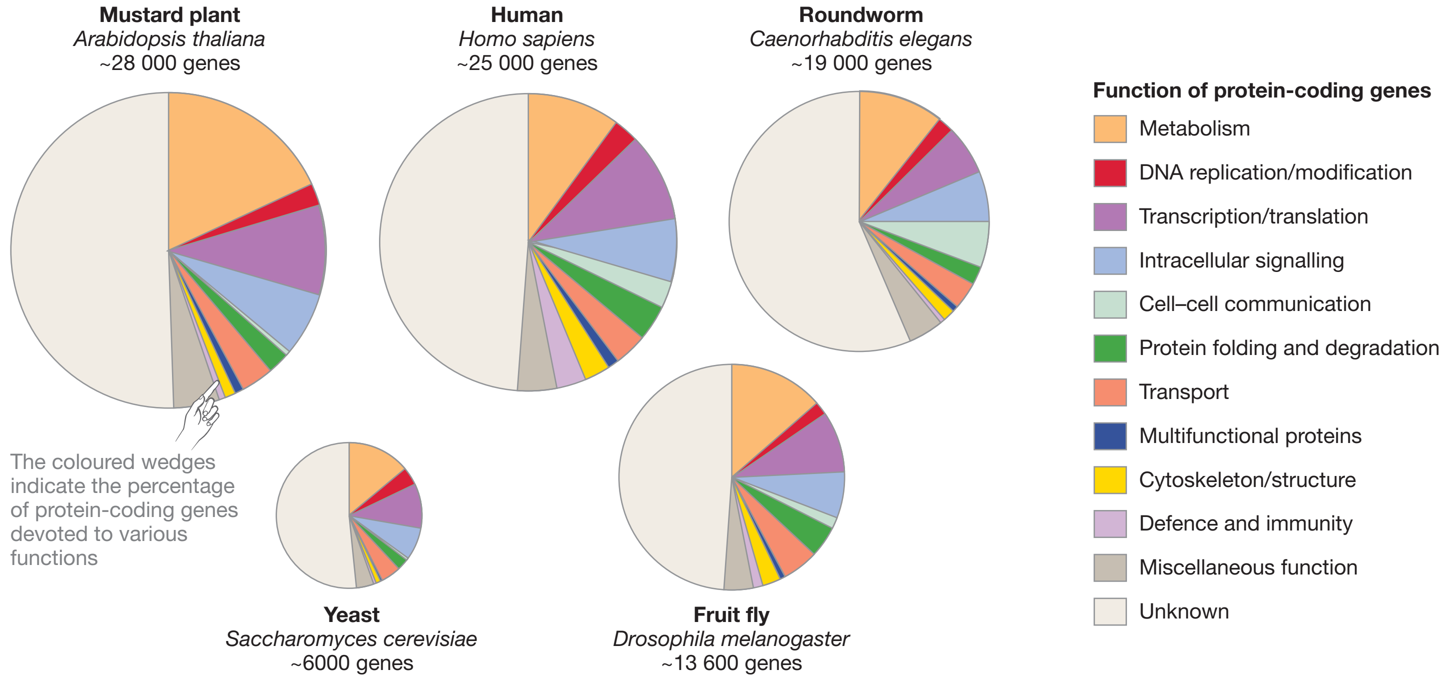


FIGURE 20.10 Comparing Gene Functions in Various Eukaryotes.

QUESTION In the organisms pictured here, about what percentage of the genes present have unknown function?

Figure 20.10, for example, shows the relative proportion of genes devoted to various functions in humans and four other species of eukaryotes. A careful look at this figure should convince you that humans have a particularly large proportion of their genome devoted to immunity (defence against bacteria, viruses, and other parasites), and that humans and the roundworm *Caenorhabditis elegans* have a larger percentage of their genome devoted to cell–cell signalling than do other eukaryotes studied thus far. But even a quick look at these diagrams carries an important message: No one knows the function of most genes found in humans and other eukaryotes.

Although it is clear that a great deal remains to be learned about the human genome, two important questions have emerged from the early studies. Let's consider each of them in turn.

Why Do Humans Have So Few Genes? Of all observations about the nature of eukaryotic genomes, perhaps the most striking is that organisms with complex morphology and behaviour do not appear to have particularly large numbers of genes. **Table 20.1** indicates the estimated number of genes found in selected eukaryotes. Notice that the total number of genes in *Homo sapiens*, which is considered a particularly complex organism, is not that much higher than the total number of genes in fruit flies. Gene number in humans is about the same as in roundworms, mice, rats, puffer fish, and chickens and substantially less than in rice and the weedy mustard plant *Arabidopsis thaliana*. Before the human genome was sequenced, many biologists expected that humans would have at least 100 000 genes. We now know that we have only a fifth of that number—perhaps less.

How can this be? In prokaryotes there is a correlation between genome size, gene number, a cell's metabolic capabilities, and the cell's ability to live in a variety of habitats. Similarly, it is logical to observe that plants have exceptionally large numbers of genes because they synthesize so many different and complex molecules from just carbon dioxide, nitrate ions, phosphate ions, and other simple nutrients. The idea is that large numbers of genes enable plants to produce large numbers of enzymes. But why isn't there a stronger correlation between gene number and morphological and behavioural complexity in animals?

The leading hypothesis focuses on **alternative splicing**. Recall from Chapter 18 that the exons of a particular gene can be spliced in ways that produce distinct mature mRNAs. As a result, a single eukaryotic gene can code for multiple transcripts and thus multiple proteins. The alternative-splicing hypothesis claims that at least certain multicellular eukaryotes do not need enormous numbers of distinct genes. Instead, alternative splicing creates different proteins from the same gene. The alternative forms might be produced at different developmental stages or in response to different environmental conditions.

In support of the alternative-splicing hypothesis, researchers have analyzed the mRNAs produced by human genes and have estimated that each gene produces an average of slightly more than three distinct transcripts. If this result is valid for the rest of the genome, the actual number of different proteins that can be produced is more than triple the gene number. Humans may

TABLE 20.1 **Number of Genes in Selected Eukaryotes**

Species	Description	Genome Size (Millions of Base Pairs)	Estimated Number of Genes
Saccharomyces cerevisiae	Baker's and brewer's yeast; a unicellular fungus; an important model organism in biochemistry and genetics	12	6 000
Plasmodium falciparum	Single-celled, parasitic eukaryote; causes malaria in humans	30	6 500
Drosophila melanogaster	Fruit fly; an important model organism in genetics and developmental biology	180	13 600
Caenorhabditis elegans	A roundworm; an important model organism in developmental biology	97	19 000
Canis familiarus	Domestic dog	2 410	19 300
Gallus gallus	Chicken	1 050	20 000–23 000
Homo sapiens	Humans	3 000	20 000
Rattus norvegicus	Norway rat; an important model organism in physiology and behaviour	2 750	21 000
Mus musculus	House mouse; an important model organism in genetics and developmental biology	2 500	~30 000
Arabidopsis thaliana	A mustard plant; an important model organism in genetics and developmental biology	119	28 000
Oryza sativa	Rice	389	37 500

have fewer than 20 000 genes according to current estimates, but these genes may have the ability to produce 100 000 different transcripts. Researchers are currently trying to assess whether alternative splicing is this frequent in other eukaryotes, as well.

How Can the Human and Chimp Genomes Be So Similar?

Comparing the numbers of genes found in humans and in mice created a paradox that may be resolved by the alternative-splicing hypothesis. Comparing the base sequences of genes in humans and in other species has created an analogous paradox.

Here is the issue: At the level of base sequences, human beings and chimpanzees are 98.8 percent identical on average. Of the homologous genes analyzed in humans and chimps, 29 percent are identical in amino acid sequence; the average difference between homologous proteins is just two amino acids. If humans and chimps are so similar genetically, why do they appear to be so different in their morphology and behaviour?

The leading hypothesis to resolve this paradox focuses on the importance of regulatory genes and regulatory sequences. Recall from Chapter 18 that a **regulatory sequence** is a section of DNA involved in controlling the activity of other genes; it may be a promoter, a promoter proximal element, an enhancer, or a silencer. The term **structural gene**, in contrast, refers to a sequence that codes for a tRNA, rRNA, protein, or other type of product. **Regulatory genes** code for regulatory transcription factors that alter the expression of specific genes.

To resolve the sequence-similarity paradox, biologists propose that even though many structural genes in closely related species, such as humans and chimps, are identical or nearly identical, regulatory sequences and regulatory genes might have important differences between the two species. Suppose that the structural gene for human growth hormone and chimp growth hormone are identical in base sequence. But if changes in transcription factors, enhancers, or promoters change the pattern of expression of that gene—perhaps turning it on later and longer in humans than in chimps—then height and other characteristics will change even though the structural gene is the same. Based on current analyses, biologists suggest that the human genome contains about 3000 different regulatory transcription factors. Subtle mutations in these proteins and the regulatory sites that they bind to could have a significant effect on gene expression and thus on the phenotype.

The regulatory hypothesis is certainly logical, and it is consistent with data suggesting that regulation of alternative splicing underlies phenotypic complexity in *Homo sapiens* and other large vertebrates. It may be true that most of the genetic changes responsible for the rapid evolution of humans over the past 5 million years have been due to changes in regulatory genes and sequences and alternative splicing rather than to changes in structural genes. To date, however, there are no specific examples of changes in the regulatory sequences responsible for the phenotypic differences observed between humans and chimps or other closely related species. The regulatory hypothesis still needs to be tested rigorously.

Check Your Understanding

If you understand that…

- Eukaryotic genomes are riddled with parasitic sequences that do not contribute to the fitness of the organism.
- Simple repeated sequences are also common in eukaryotic genomes.
- In eukaryotes, many of the coding sequences are organized into families of genes with related functions.

You should be able to…

1) Explain why transposable elements are considered selfish genes.
2) Explain why simple sequence repeats make DNA fingerprinting possible.
3) Explain how unequal crossover leads to duplicated sequences.
4) Explain why researchers hypothesize that human evolution has been dominated by alternative splicing and other types of regulatory changes.

20.4 Functional Genomics and Proteomics

To explain the impact of genomics on the future of biological science, Eric Lander has compared the sequencing of the human genome to the establishment of the periodic table of the elements in chemistry. Once the periodic table was established and validated, chemists focused on understanding how the elements combine to form molecules. Similarly, biologists now want to understand how the elements of the human genome combine to produce an individual.

In essence, a genome sequence is a parts list. Once that list is assembled, researchers delve deeper to understand how genes interact to produce an organism. Let's explore some of the ways in which researchers use whole-genome data to answer fundamental questions about how organisms work.

What Is Functional Genomics?

For decades, biologists have worked at understanding how and when individual genes are expressed. Research on the *lac* operon and *trp* operon, reviewed in Chapter 17, is typical of this effort. But now, with complete catalogues of the genes present in a variety of organisms whose genomes have been sequenced, researchers can ask how and when *all* of the genes in an organism are expressed. These types of large-scale analyses of gene expression are sometimes called functional genomics. The research is motivated by the realization that gene products do not exist in a vacuum. Instead, groups of RNAs and proteins act together to respond to environmental challenges such as extreme heat or drought. Similarly, distinct groups of genes are transcribed at different stages as a multicellular eukaryote grows and develops.

One of the most basic tools used in functional genomics is called a microarray. A **DNA microarray** consists of a large number of single-stranded DNAs that are permanently affixed to a glass slide. For example, the slide pictured in **Figure 20.11** contains thousands of spots, each of which contains single-stranded DNA from a unique exon found in the human genome.

To do an experiment with a DNA microarray, researchers follow the protocol outlined in **Figure 20.12**. The first step is to isolate the mRNAs being produced by two contrasting types of cells. In this example, the control cells are functioning at normal temperature. The other cells, in contrast, have been exposed to high temperatures. The goal of this experiment is to compare genes that are expressed during normal cell activity with those expressed under heat stress.

Once they've purified mRNAs from the two populations of cells, investigators use reverse transcriptase to make a single-stranded cDNA version of each RNA in the samples (see Chapter 19). In addition to the four standard dNTPs, one of the DNA building blocks used in this reaction carries a fluorescent label. The label used for normal cells glows green, while the label chosen for the heat-stressed cells glows red.

The labelled cDNAs can then be used to probe the microarray. As Chapter 19 noted, a probe allows an investigator to find a particular molecule in a sample containing many different molecules. In this case, the labelled cDNAs will bind to the single-stranded DNAs on the plate by complementary base pairing. Out of all the exons in the genome, then, only the

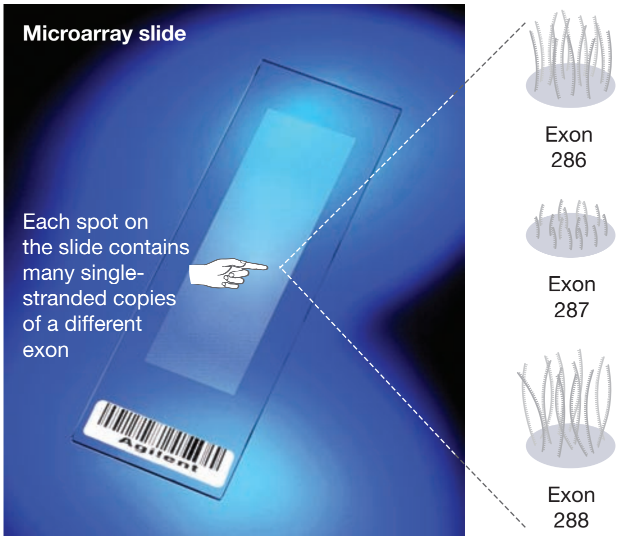

Microarray slide
Each spot on the slide contains many single-stranded copies of a different exon
Exon 286
Exon 287
Exon 288

FIGURE 20.11 DNA Microarrays Represent Every Gene in a Genome. To create a DNA microarray, investigators spot thousands of short, single-stranded DNA sequences from coding sequences onto a glass plate. The DNAs typically represent every exon in the genome of a particular species.

exons that are being expressed will be labelled. In our example, genes that are expressed under normal conditions will be labelled green, while those expressed during heat stress will be labelled red. If one of the exons in the microarray is expressed under both sets of conditions, then both green- and red-labelled cDNAs will bind to that spot and make it appear yellow. In this way, a microarray lets researchers study the expression of thousands of genes at a time. As a result, they can identify which sets of genes are expressed in concert under specific sets of conditions.

Once a microarray has been used, the bound cDNA probes can be removed. The original DNAs remain in place, so the slide may then be reused to assess gene expression in a different type of cell, or in the same cell type under different conditions. Researchers can use microarrays to establish which genes are transcribed in different organs and tissues, during cancerous growth, or in response to changes in environmental conditions, such as starvation, the presence of a toxin, or a viral infection. ● If you understand the concept of how microarrays are used, you should be able to design an experiment that uses a DNA microarray to compare the genes expressed in brain cells versus liver cells of an adult human.

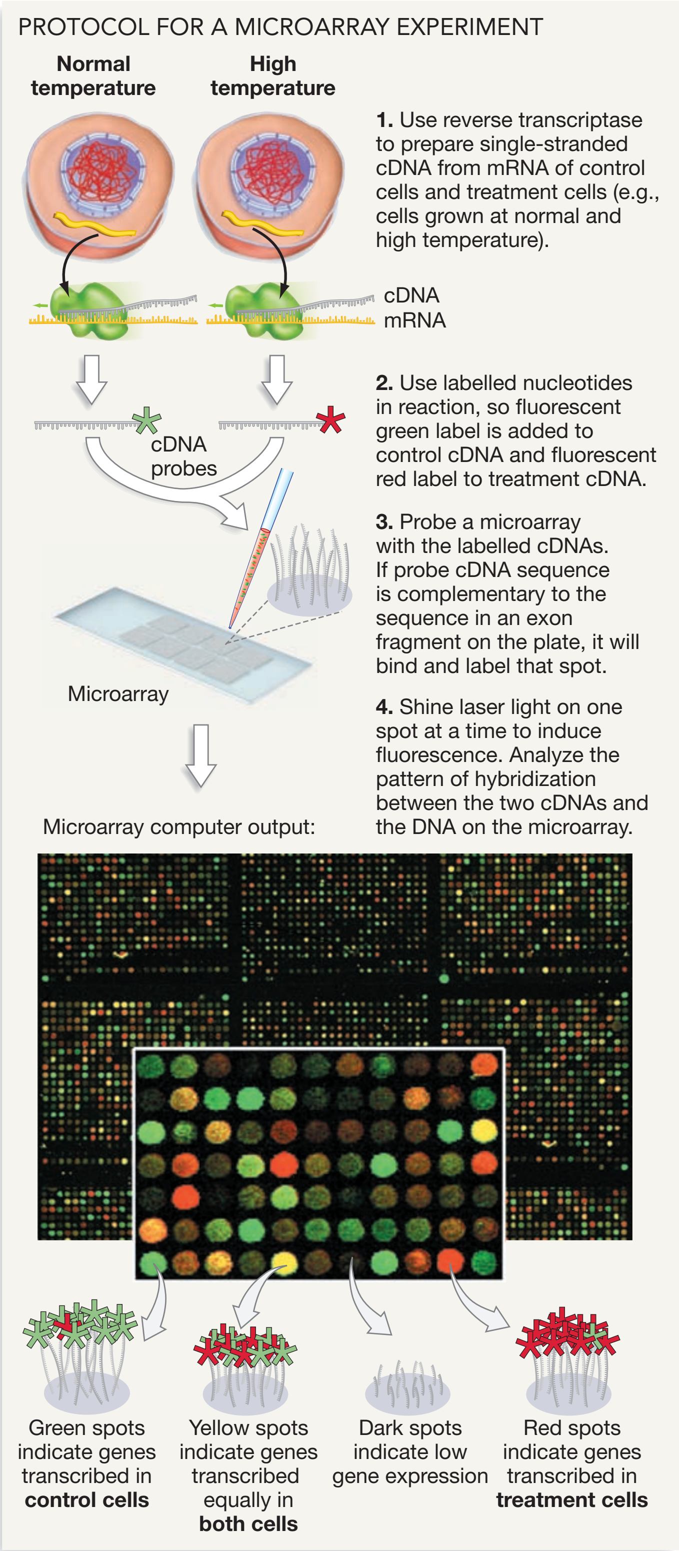

PROTOCOL FOR A MICROARRAY EXPERIMENT
Normal temperature
High temperature
1. Use reverse transcriptase to prepare single-stranded cDNA from mRNA of control cells and treatment cells (e.g., cells grown at normal and high temperature).
cDNA
mRNA
2. Use labelled nucleotides in reaction, so fluorescent green label is added to control cDNA and fluorescent red label to treatment cDNA.
cDNA probes
3. Probe a microarray with the labelled cDNAs. If probe cDNA sequence is complementary to the sequence in an exon fragment on the plate, it will bind and label that spot.
Microarray
4. Shine laser light on one spot at a time to induce fluorescence. Analyze the pattern of hybridization between the two cDNAs and the DNA on the microarray.
Microarray computer output:
Green spots indicate genes transcribed in control cells
Yellow spots indicate genes transcribed equally in both cells
Dark spots indicate low gene expression
Red spots indicate genes transcribed in treatment cells

FIGURE 20.12 DNA Microarrays Are Used to Study Changes in Gene Expression. By probing a microarray with labelled cDNAs synthesized from mRNAs, researchers can identify which coding sequences are being transcribed. Here mRNAs from cells growing at normal temperature are green, while mRNAs from cells growing at high temperature are red.

● **QUESTION** If every spot on a microarray represents a different exon, how could an experiment like this document the existence of alternative splicing?

What Is Proteomics?

The Greek root *–ome*, meaning "all," inspired the term *genome*. Similarly, biologists use the term **transcriptome** in referring to the complete set of genes that are transcribed in a particular cell, and **proteome** in referring to the complete set of proteins that are produced. **Proteomics**, it follows, is the large-scale study of protein function. Proteomic studies begin by identifying the proteins present in a cell or organelle; then, researchers attempt to determine the locations and interactions of proteins and document how they change through time or compare with other cells.

Proteomics can be thought of as a branch of functional genomics. Instead of studying individual proteins or how two proteins might interact, biologists can study all of the proteins present at once. One approach to studying protein–protein interactions is similar to the use of DNA microarrays, except that large numbers of proteins, rather than DNA sequences, are affixed on a glass plate. This microarray of proteins is then treated with an assortment of proteins produced by the same organism. These proteins are labelled with a fluorescent or radioactive tag. If any labelled proteins bind to the proteins in the microarray, the two molecules may also interact in the cell. In this way, researchers hope to identify proteins that physically bind to one another—like the G proteins and associated enzymes introduced in Chapter 8, or the cyclin and Cdk molecules introduced in Chapter 11. Microarray technology is allowing biologists to study protein–protein interactions on a massive scale.

CANADIAN **ISSUES** 20.1

Genome Canada

Most genomics and proteomics research projects have two things in common: They involve several lab groups working together, and they require lots of funding. The Human Genome Project, completed in 2004, took the International Human Genome Sequencing Consortium 13 years and $2.7 billion to complete. In Canada the organization that coordinates regional, national, and international projects—and provides much of their funding—is Genome Canada. Since its inception in 2000, it has overseen 115 projects and almost $1.8 billion in grants.

Most of these projects are concerned with health issues. One of the most promising avenues of genomics research is pharmacogenomics, the study of how a person's genetic makeup affects how he or she responds to pharmaceutical drugs. A drug might be beneficial for one person, inconsequential for a second person, and toxic for a third. Michael Phillips from McGill University and Jean-Claude Tardif from the University of Montreal are studying several drugs used to treat cardiovascular disease. The goal of their work is to be able to match a patient's genetic profile to the appropriate drug and dose that person should receive.

A large part of Genome Canada's mandate is the study and improvement of agriculturally important plants and animals. For example, Brian Fowler at the University of Saskatchewan leads an international team studying the genetics of how crop plants respond to environmental stress. One of their successes has been the identification of genes in cold-resistant strains that can then be introduced into cold-sensitive strains. Cold-tolerant strains of wheat (**Figure A**) and barley are less likely to be damaged or killed by frost, which is of great economic importance.

A third focus of Genome Canada is studying the relationship between genomics and society. These projects are collectively known as genomics ethics, environmental, economic, legal and social issues (GE^{3}LS). Bartha Maria Knoppers from the University of Montreal is the project director of a group studying when someone's personal genetic data should become public. For example, if information on patients is placed into a database, who should be able to access it? What should a physician do if a patient refuses to inform his or her relatives that they are at risk for a genetic disease? These questions and others need to be addressed as genomics research provides us with new information about ourselves and the organisms around us.

Figure A *Wheat is a very important agricultural product for Canada.*

20.5 Can Genomics Help Improve Human Health and Welfare?

With the advent of microarray technology, the "periodic tables" provided by genome projects are having an important impact on research into gene expression and protein–protein interactions. But the governments and corporations that fund genome projects have underwritten the expense primarily because of the potential benefits for improving human health and welfare. In this respect, is genomics living up to its promise?

Although large amounts of genome sequence data have been available for only a few years, early indications are that the investment may indeed pay off with substantial advances in biomedicine. Let's consider how whole-genome data are informing the development of new drugs and vaccines, and then look at a project focused on searching for the alleles associated with inherited diseases.

Identifying Potential Drug Targets

Currently, dozens of whole-genome sequencing projects are focused on species that cause disease in humans or the livestock and crops that we depend on. As each of these projects is completed, biologists begin comparing the genomes of pathogenic strains with closely related species that are harmless. The goal is to achieve a much more detailed understanding of the genetic basis of **virulence**—the tendency for a parasite to harm its host.

One of the first tasks biologists undertake is to identify genes that occur only in pathogens and that may be required for virulence. ● Typically, biologists are finding that "virulence genes" code for proteins that allow parasitic cells to adhere to host cells, produce enzymes that break down host-cell walls or membranes, or secrete toxins that poison host-cell enzymes. Identifying these genes is important because it gives investigators targets for the development of new drugs. If drugs that knock out the protein products of these genes can be formulated, they would inhibit disease-causing cells while leaving closely related but helpful species unharmed. For example, comparing the genomes of benign strains of *E. coli* with *E. coli* strains that cause food poisoning allowed researchers to find virulence genes in the pathogenic strain. These genes code for proteins that poison cells in the human intestine and cause severe diarrhea. If drugs that neutralize these proteins can be developed, they might provide an effective treatment for food poisoning.

Having the complete gene catalogue from pathogenic species is giving biomedical researchers new targets for drug development and new possibilities for therapy. This work is particularly urgent, because disease-causing bacteria continue to evolve resistance to many of the antibiotics currently in use (see Chapter 24).

Designing Vaccines

Although efforts to exploit genome data in drug design are still in their infancy, genomics has already inspired important advances in vaccine development. In essence, researchers are testing proteins that are identified by genome sequencing to see if the molecules stimulate the immune system enough to function as vaccines.

To illustrate how this work is proceeding, consider recent research on the bacterium *Neisseria meningitidis*. This species is a major cause of meningitis and blood infections in children and was one of the first bacteria to have its genome sequenced. Although antibiotics can treat *N. meningitidis* infections effectively, the organism grows so quickly that it often injures or even kills the victim before a diagnosis can be made and drugs administered. As a result, biomedical researchers have been interested in developing a vaccine that would prime the immune system and allow children to ward off infections.

Vaccine development has been difficult in this case, however. As Chapter 49 will explain in detail, the immune system usually responds to molecules on the outer surface of bacteria or viruses. When an immune system cell recognizes one of these surface molecules as foreign, the invading bacterium or viral particle is destroyed. This recognition step by the immune system takes time, however, and is speeded up by vaccination. Vaccines contain surface molecules from a bacterium or virus. Even though the vaccine is harmless, the immune system cells go through their normal recognition sequence. In this way, ingesting a vaccine alerts or primes the immune system. If an actual infection occurs later, the cells are ready to spring into action and destroy the pathogens before they can grow and cause disease.

Unfortunately, *N. meningitidis* is covered with a polysaccharide that is identical to a compound found on the surface of brain cells. Immune system cells normally do not attack compounds found on the body's own cells, so a vaccine composed of the *N. meningitidis* polysaccharide would elicit no response.

To circumvent this problem, biologists analyzed the genome sequence of *N. meningitidis* and tested 600 open reading frames for the ability to encode vaccine components. The researchers inserted the 600 DNA sequences into *E. coli* cells, following the steps shown in **Figure 20.13**. Later they succeeded in isolating 350 different *N. meningitidis* proteins from the transformed cells. The biologists injected these proteins into mice and then analyzed whether an immune response occurred. Their results show that seven of the proteins tested evoked a strong immune response and represent promising vaccine components. Follow-up work is now under way to determine whether one or more of these proteins could act as a safe and effective vaccine in humans.

Finding Genes Associated with Inherited Disease: The HapMap Project

Chapter 19 explained how gene hunters analyzed genetic markers to find the allele responsible for Huntington's disease. Data from human genome sequencing have made this approach to finding disease genes much more powerful.

Because DNA from many individuals was used as source material during the Human Genome Project, and because overlapping segments of genes were routinely sequenced, researchers were able to identify 1.42 million sites where single bases vary

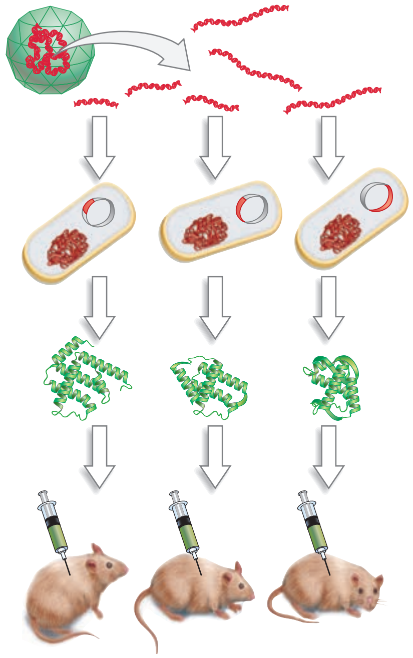
Experiment

Question: Could vaccines be developed from the products of newly discovered genes?

Hypothesis: Some of the genes discovered through genome sequencing code for proteins that can be used in vaccines.

Null hypothesis: None of the genes discovered through genome sequencing code for proteins that can be used in vaccines.

Experimental setup:



1. Isolate open reading frames (ORFs) from pathogen genome sequence.

2. Introduce ORFs into *E. coli* cells.

3. Isolate proteins that result from transcription and translation.

4. Inject proteins into mice. As a control, inject only the solution used to suspend the proteins.

Prediction: Some proteins will elicit an immune response similar to that elicited by vaccines. The control solution will not elicit an immune response.

Prediction of null hypothesis: No injections (control or treatments) will elicit an immune response similar to that elicited by vaccines.

Results:	Number of mice with immune response:	Number of mice with no immune response:
Strong immune responses ☞	7	343

Conclusion: The seven proteins that elicited an immune response are potential vaccine components. Further research is needed to test their safety and effectiveness.

FIGURE 20.13 Newly Discovered Proteins Can Be Tested for Vaccine Development. Because all potential genes are identified after whole-genome sequencing, virtually all proteins produced by a pathogen can be tested for their ability to provoke an immune response and act as a vaccine.

among individuals. Where you might have a "C" at a particular site, others may have a "T." Recall that these variable sites are called **single nucleotide polymorphisms,** or **SNPs,** and that they can serve as genetic markers—mapped sites in the genome that vary among individuals.

By sequencing DNA samples from dozens of people representing four ethnic groups from widely dispersed geographic areas, a recent research effort called the HapMap Project has extended this initial catalogue of SNPs to a current total of 10 million sites that vary among humans. HapMap is short for haplotype mapping. A **haplotype** is the set of alleles found on a single chromosome or chromosome segment. By mapping SNPs, researchers hope to be able to determine the haplotype of any individual at any of their chromosomes.

For disease gene hunters, this new catalogue of SNPs is an enormously powerful resource. The fundamental idea is to compare the haplotypes of individuals who have an inherited disease with the haplotypes of unaffected people. If certain SNPs are extremely common in affected individuals but rare or absent in unaffected individuals, it is likely that those SNPs are near or even within a gene that contributes to the disease.

The possibility of analyzing the inheritance of millions of polymorphic sites all over the genome—instead of just a few hundred widely scattered ones like those used in the Huntington's disease gene hunt—makes it much more likely that researchers can locate genes associated with illnesses such as Alzheimer's disease, bipolar disorder, diabetes, rheumatoid arthritis, and cardiovascular disease. The database promises to be particularly important in understanding the genetic basis of diseases that involve many different genes, instead of a single allele as Huntington's disease does. As biologists continue their efforts to annotate the human genome, it's very likely that they will be able to track down the genes responsible for many or even most inherited diseases.

CANADIAN **RESEARCH** 20.1

Human Genetic Variation

Why are people different from one another? One way to think about this question is to consider what makes your DNA different from another person's (**Figure A**). So far in this chapter you have learned about two types of genetic variation: single nucleotide polymorphisms (SNPs) and polymorphic repeat sequences. A SNP within a gene can affect the protein it makes, and this may influence anything from a person's height to his or her health. Polymorphic microsatellites and minisatellites next to or within a gene can also influence gene expression. However, in just the last few years, a third type of polymorphism has been found that may be as important if not more so than either of these. It was discovered almost by accident by several geneticists around the world who were studying genetic diseases.

One of the geneticists involved in this discovery was Dr. Martin Somerville, a professor in the Department of Medical Genetics at the University of Alberta and the director of the Molecular Diagnostic Laboratory. Some of the diseases his lab tests for are caused by not having the normal two copies of an important gene. This occurs when the chromosome a child inherits from one parent has the gene but the chromosome from the other parent has a small deletion and is missing the gene. These small deletions occur during gamete formation in a parent. Figure 20.8 shows how an unequal crossover event during meiosis produces chromosomes that have fewer or additional genes. A child that received the top chromosome in this figure from one parent would be missing one copy of gene 3.

Dr. Somerville's lab diagnoses gene deletions with a modified form of the polymerase chain reaction called real-time quantitative PCR. The reaction is the same as in Figure 19.7 (page 410), but it takes place within a machine that measures how fast the PCR products accumulate (see **Figure B**, part i). The purpose of the technique is not to obtain a product, but to determine how much template DNA was present initially. For example, let us say that a patient has only a single copy of a particular gene. A sample of the DNA will have half the normal number of pieces of DNA containing this gene. If we use the patient's DNA and try to amplify this gene, the PCR reaction will occur but the reaction will be noticeably slower (**Figure B**, part ii).

While some diseases are known to be caused by gene deletions, Dr. Somerville's team decided to hunt for others. One such candidate was congenital heart disease. About six in 1000 children are born with moderate or severe heart defects. Doctors suspected that some cases might be due to a problem with a gene expressed in the heart called Connexin40. The protein made by this gene is part of the gap junctions (see Figure 8.14) that connect heart cells to each other. These gap junctions help to hold the cells together as the heart forms during embryogenesis. It seemed likely that a problem with Connexin40 would lead to a malformed heart, but at the time of Dr. Somerville's research no one had found a patient with a mutation in this gene.

Figure A *Human chromosomes that have been given false colours as in Figure 12.6.*

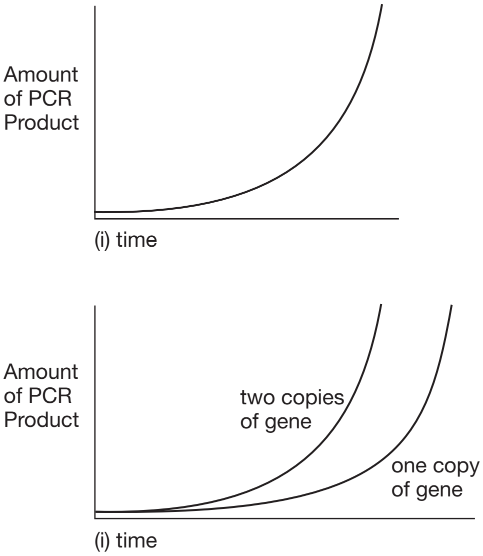


Figure B *The progress of PCR reactions.*
(i) PCR products accumulate exponentially during a PCR reaction.
(ii) When we use PCR to reproduce a human gene, we can tell whether the person who donated the DNA sample has two copies or one copy of the gene by how quickly the PCR products accumulate.

The University of Alberta group tested 505 children with congenital heart defects. **Figure C** on the next page shows how real-time quantitative PCR would give a different result if a child only had one copy of the Connexin40 gene. The researchers found that three of these children did in fact have a deletion of this gene. Their health problems were due to the absence of one copy of one of their 25 000 different genes.

What does this story have to do with overall human genetic variation? Well, what Dr. Somerville and other geneticists have discovered is that *everyone* has deletions of a few of their genes. Conversely, a person may have two copies of the same gene side by side on the chromosome. Previously we assumed that since people had two copies of each chromosome (with the exception

continued

CANADIAN **RESEARCH** 20.1 *(continued)*

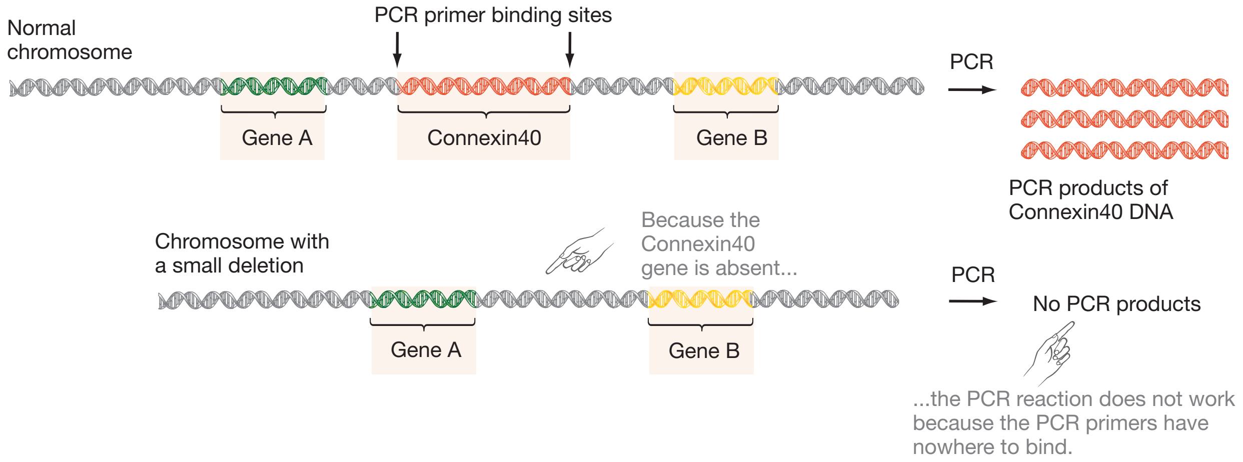
Experiment

Question: Are some cases of congenital heart disease associated with a deletion of the Connexin40 gene?

Hypothesis: A patient with congenital heart disease has only a single copy of the Connexin40 gene.

Null hypothesis: A patient with congenital heart disease has the normal two copies of the Connexin40 gene.

Experimental setup: 1. Isolate DNA from 505 patients and from healthy volunteers; 2. Use real-time quantitative PCR to determine if each person has two normal chromosomes or one normal and one deleted chromosome.


Normal chromosome
PCR primer binding sites
Gene A
Connexin40
Gene B
PCR
PCR products of Connexin40 DNA
Chromosome with a small deletion
Because the Connexin40 gene is absent...
Gene A
Gene B
PCR
No PCR products
...the PCR reaction does not work because the PCR primers have nowhere to bind.

Prediction: As shown in Figure B, the reaction using the patient's DNA will proceed at a slower rate than the reaction using a healthy person's DNA.

Prediction of null hypothesis: The reaction using the patient's DNA will proceed at the same rate as the reaction using a healthy person's DNA.

Results: For three of the 505 patients, the reaction went at a slower rate.

Conclusion: Some patients with congenital heart problems have deletions of Connexin40. Because they only have a single copy of the Connexin40 gene, it is possible that they do not make enough Connexin40 proteins in their heart tissue and that this is the cause of their heart problems.

Figure C *The experiment done by Somerville's group.*

of the X chromosome in males), we would also have two copies of each gene. Now we know that for several genes a person may have one, two, three, or four copies. These are called copy number variations (CNVs) or copy number polymorphisms (CNPs).

Dr. Somerville was part of an international project studying CNVs that included researchers from Toronto's Hospital for Sick Children and the University of Toronto, as well as geneticists from the United Kingdom, the United States, Spain, and Japan. The goal was to identify CNVs by comparing DNA sequences from 270 people from all over the world. Much to everyone's surprise, they found *1447* copy number variable regions (CNVRs). **Figure D** shows their distribution on just one of our chromosomes. CNVRs make up 12 percent of our genome, much more than the 3 percent of our genome included in SNPs. The Connexin40 example demonstrates that CNVs can have a profound influence on our health. Right now geneticists in Canada and other countries are looking for other diseases that are strongly influenced by the copy number of key genes.

References: Christiansen et al. (2004). Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. *Circulation Research* 94:1429–1435; Redon et al. (2006). Global variation in copy number in the human genome. *Nature* 444:444–454.

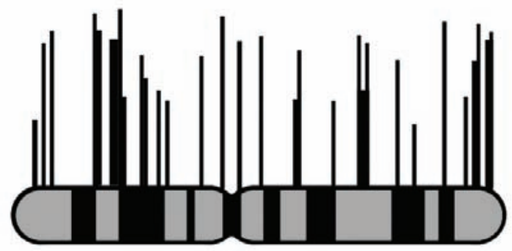


Figure D *The CNVs on human chromosome 20. The height of each line is proportional to the size of the CNV at that location. The longest CNVs are over 1 million base pairs long. (From Redon et al., 2006.)*

Chapter Review

SUMMARY OF KEY CONCEPTS

Once a genome has been completely sequenced, researchers use a variety of techniques to identify which sequences code for products and which act as regulatory sites.

Recent technical advances have allowed investigators to sequence DNA much more rapidly and cheaply than before, resulting in a flood of genome data. Researchers annotate genome sequences by finding genes and determining their function. To identify genes in bacteria and archaea, researchers use computers to scan the genome for start and stop codons that are in the same reading frame and that are separated by gene-sized stretches of sequence. Finding such open reading frames (ORFs) is difficult in eukaryotes, because exons are interrupted by introns and because most eukaryotic DNA does not code for a product. One approach to finding eukaryotic genes is to analyze the sequences of complementary DNAs (cDNAs) synthesized from mRNAs and then match these sequences to DNA found in the genome itself. Sequences that are highly conserved among species are also hypothesized to indicate the locations of genes.

You should be able to describe how a research group that discovered a gene for coat colour in mice would determine whether a homologous gene existed in the human genome. ●

MB **Web Animation** at www.masteringbio.com
Human Genome Sequencing Strategies

In bacteria and archaea, there is a positive correlation between the number of genes in a species and the species' metabolic capabilities. Gene transfer between species is also common.

Species of bacteria and archaea are usually targeted for whole-genome sequencing because they cause disease or have interesting metabolic abilities. In these groups, the size of an organism's genome and its morphological complexity or biochemical capabilities are correlated. Parasites tend to have small genomes; organisms that live in a broad array of habitats or that use a wide variety of nutrients tend to have larger genomes. Many of the genes identified in bacteria and archaea still have no known function, however, and a significant percentage of them are extremely similar to other genes in the same genome. Another generalization about prokaryotic genomes is that genes are frequently transferred laterally, or between species. Lateral gene transfer appears to be common in genes responsible for causing disease.

You should be able to describe two mechanisms responsible for lateral gene transfer in bacteria and archaea.

In eukaryotes, genomes are dominated by sequences that have little to no effect on the fitness of the organism.

Compared with prokaryotic genomes, eukaryotic genomes are large and contain a high percentage of transposable elements, repeated sequences, and other noncoding sequences. There is no obvious correlation between morphological complexity and gene number in eukaryotes, although the number of distinct transcripts produced may be much larger than the actual gene number in certain species as a result of alternative splicing. Gene duplication and polyploidy have been the most important sources of new genes in eukaryotes.

You should be able to explain what biologists mean when they refer to "junk DNA," and whether these sequences lack function and are uninteresting, as originally proposed. ●

Data from genome sequencing projects are now being used in the development of new drugs and vaccines, and to search for alleles associated with inherited diseases.

The availability of whole-genome sequences is inspiring new research programs. Biologists are affixing exons or proteins to microarrays in order to study changes in gene expression or protein–protein interactions. In addition, the availability of whole-genome data has allowed investigators to find new drug targets, new proteins that may serve as vaccine candidates, and new genetic markers that should aid in the hunt for alleles associated with human disease.

You should be able to explain the difference between studies of gene expression in single genes versus microarrays, and expression of individual proteins versus microarrays. ●

QUESTIONS

Test Your Knowledge

1. What is an open reading frame?
   - **a.** a gene whose function is already known
   - **b.** a DNA section that is thought to code for a protein because it is similar to a complementary DNA (cDNA)
   - **c.** a DNA section that is thought to code for a protein because it has a start codon and a stop codon flanking hundreds of base pairs
   - **d.** any member of a gene family

2. What best describes the logic behind shotgun sequencing?
   - **a.** Break the genome into tiny pieces. Sequence each piece. Use overlapping ends to assemble the pieces in the correct order.
   - **b.** Start with one end of each chromosome. Sequence straight through to the other end of the chromosome.
   - **c.** Use a variety of techniques to identify genes and ORFs. Sequence these segments—not the noncoding and repeated sequences.
   - **d.** Break the genome into pieces. Map the location of each piece. Then sequence each piece.

3. What are minisatellites and microsatellites?
   - **a.** small, extrachromosomal loops of DNA that are similar to plasmids
   - **b.** parts of viruses that have become integrated into the genome of an organism

c. incomplete or “dead” remains of transposable elements in a host cell
d. short and simple repeated sequences in DNA

4. What is the leading hypothesis to explain the paradox that large, morphologically complex eukaryotes such as humans have relatively small numbers of genes?
a. lateral transfer of genes from other species
b. alternative splicing of mRNAs
c. polyploidy, or the doubling of the genome’s entire chromosome complement
d. expansion of gene families through gene duplication

5. What evidence do biologists use to infer that a gene is part of a gene family?
a. Its sequence is exactly identical to that of another gene.
b. Its structure—meaning its pattern of exons and introns—is identical to that of a gene found in another species.
c. Its composition, in terms of percentage of A-T and G-C pairs, is unique.
d. Its sequence, structure, and composition are similar to those of another gene in the same genome.

6. What is a pseudogene?
a. a coding sequence that originated in a lateral gene transfer
b. a gene whose function has not yet been established
c. a polymorphic gene—meaning that more than one allele is present in a population
d. a gene whose sequence is similar to that of functioning genes but does not produce a functioning product

Test Your Knowledge answers: 1. c; 2. a; 3. d; 4. b; 5. d; 6. d

Test Your Understanding

Answers are available at www.masteringbio.com

1. Explain how open reading frames are identified in the genomes of bacteria and archaea. Why is it more difficult to find open reading frames in eukaryotes?
2. Why is the observation that parasitic organisms tend to have relatively small genomes logical?
3. Review how a LINE sequence transmits a copy of itself to a new location in the genome. Why are LINEs and other repeated sequences referred to as “genomic parasites”?
4. How does DNA fingerprinting work? Stated another way, how does variation in the size of microsatellite and minisatellite loci allow investigators to identify individuals?
5. Researchers can create microarrays of short, single-stranded DNAs that represent many or all of the exons in a genome. Explain how these microarrays are used to document changes in the transcription of genes over time or in response to environmental challenges.
6. Explain the concept of homology and how identifying homologous genes helps researchers identify the function of unknown genes. Are duplicated sequences that form gene families homologous? Explain.

Applying Concepts to New Situations

Answers are available at www.masteringbio.com

1. Parasites lack genes for many of the enzymes found in their hosts. Most parasites, however, have evolved from free-living ancestors that had large genomes. Based on these observations, W. Ford Doolittle claims that the loss of genes in parasites represents an evolutionary trend. He summarizes his hypothesis with the quip “use it or lose it.” What does he mean?
2. According to eyewitness accounts, communist revolutionaries executed Nicholas II, the last czar of Russia, along with his wife and five children, the family physician, and several servants. Many decades after this event, a grave purporting to hold the remains of the royal family was identified. Biologists were asked to analyze DNA from each adult and juvenile skeleton and determine whether the bodies were indeed those of several young siblings, two parents, and several unrelated adults. If the grave was authentic, describe how similar the DNA fingerprints of each skeleton would be relative to the fingerprints of other individuals in the grave.
3. The human genome contains a gene that encodes a protein called syncytin. This gene is expressed in placental cells during pregnancy. The syncytin gene is nearly identical in DNA sequence to a gene in a virus that infects humans. In this virus, the syncytin-like gene codes for a protein found in the virus’s outer envelope. State a hypothesis to explain the similarity between the two genes.
4. A recent study used microarrays to compare the patterns of expression of genes that are active in the brain, liver, and blood of chimpanzees and humans. Although the overall patterns of gene expression were similar in the liver and blood of the two species, expression patterns were strikingly different in the brain. How does this study relate to the hypothesis that most differences between humans and chimps involve changes in gene regulation?