21

Genomes and Their Evolution

▲ **Figure 21.1 What genomic information distinguishes a human from a chimpanzee?**

KEY CONCEPTS

21.1 New approaches have accelerated the pace of genome sequencing

21.2 Scientists use bioinformatics to analyze genomes and their functions

21.3 Genomes vary in size, number of genes, and gene density

21.4 Multicellular eukaryotes have much noncoding DNA and many multigene families

21.5 Duplication, rearrangement, and mutation of DNA contribute to genome evolution

21.6 Comparing genome sequences provides clues to evolution and development

OVERVIEW

Reading the Leaves from the Tree of Life

The chimpanzee (*Pan troglodytes*) is our closest living relative on the evolutionary tree of life. The boy in **Figure 21.1** and his chimpanzee companion are intently studying the same leaf, but only one of them is able to talk about it. What accounts for this difference between two primates that share so much of their evolutionary history? With the advent of recent techniques for rapidly sequencing complete genomes, we can now start to address the genetic basis of intriguing questions like this.

The chimpanzee genome was sequenced in 2005, two years after sequencing of the human genome was largely completed. Now that we can compare our genome with that of the chimpanzee base by base, we can tackle the more general issue of what differences in the genetic information account for the distinct characteristics of these two species of primates.

In addition to determining the sequences of the human and chimpanzee genomes, researchers have obtained complete genome sequences for *E. coli* and numerous other prokaryotes, as well as many eukaryotes, including *Zea mays* (corn), *Drosophila melanogaster* (fruit fly), *Mus musculus* (house mouse), and *Macaca mulatta* (rhesus macaque). In 2010, a draft sequence was announced for the genome of *Homo neanderthalensis,* an extinct species closely related to present-day humans. These whole and partial genomes are of great interest in their own right and are also providing important insights into evolution and other biological processes. Broadening the human-chimpanzee comparison to the genomes of other primates and more distantly related animals should reveal the sets of genes that control group-defining characteristics. Beyond that, comparisons with the genomes of bacteria, archaea, fungi, protists, and plants should enlighten us about the long evolutionary history of shared ancient genes and their products.

With the genomes of many species fully sequenced, scientists can study whole sets of genes and their interactions, an approach called **genomics**. The sequencing efforts that feed this approach have generated, and continue to generate, enormous volumes of data. The need to deal with this ever-increasing flood of information has spawned the field of **bioinformatics**, the application of computational methods to the storage and analysis of biological data.

We will begin this chapter by discussing two approaches to genome sequencing and some of the advances in bioinformatics and its applications. We will then summarize what has been learned from the genomes that have been sequenced thus far. Next, we will describe the composition of the human genome as a representative genome of a complex multicellular eukaryote. Finally, we will explore current ideas about how genomes evolve and about how the evolution of developmental mechanisms could have generated the great diversity of life on Earth today.

CONCEPT 21.1

New approaches have accelerated the pace of genome sequencing

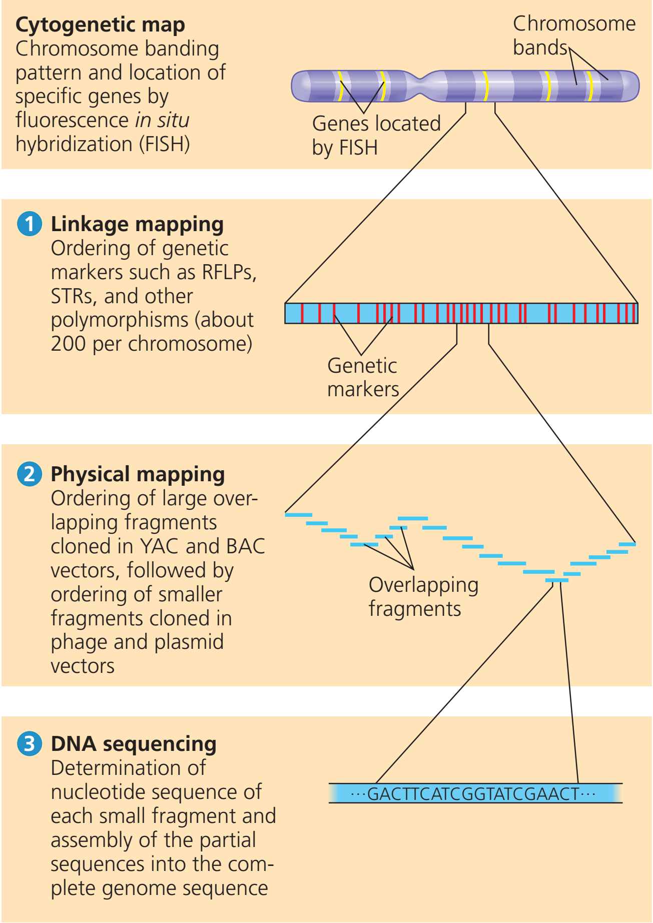
Sequencing of the human genome, an ambitious undertaking, officially began as the **Human Genome Project** in 1990. Organized by an international, publicly funded consortium of scientists at universities and research institutes, the project involved 20 large sequencing centers in six countries plus a host of other labs working on small projects.

After sequencing of the human genome was largely completed in 2003, the sequence of each chromosome was carefully analyzed and described in a series of papers, the last of which covered chromosome 1 and was published in 2006. With this refinement, researchers termed the sequencing "virtually complete." To reach these milestones, the project proceeded through three stages that provided progressively more detailed views of the human genome: linkage mapping, physical mapping, and DNA sequencing.

Three-Stage Approach to Genome Sequencing

Even before the Human Genome Project began, earlier research had sketched a rough picture of the organization of the genomes of many organisms. For instance, the karyotyping of many species had revealed their chromosome numbers and banding patterns (see Figure 13.3). And some human genes had already been located on a particular region of a chromosome by fluorescence *in situ* hybridization (FISH), a method in which fluorescently labeled nucleic acid probes are allowed to hybridize to an immobilized array of whole chromosomes (see Figure 15.1). Cytogenetic maps based on this type of information provided the starting point for more detailed mapping of the human genome.

With these cytogenetic maps of the chromosomes in hand, the initial stage in sequencing the human genome was to construct a **linkage map** (a type of genetic map; see Figure 15.11) of several thousand genetic markers spaced throughout the chromosomes (**Figure 21.2**, stage ①). The order of the markers and the relative distances between them on such a map are based on recombination frequencies. The markers can be genes or any other identifiable sequences in the DNA, such as RFLPs or short tandem repeats (STRs), both discussed in Chapter 20. By 1992, researchers had compiled a human linkage map with some 5,000 markers. Such a map enabled them to locate other markers, including genes, by testing for genetic linkage to the known markers. It was also valuable as a framework for organizing more detailed maps of particular regions. Remember from Chapter 15, however, that absolute distances between genes cannot be determined using this approach.


Cytogenetic map
Chromosome banding pattern and location of specific genes by fluorescence *in situ* hybridization (FISH)
Chromosome bands
Genes located by FISH
1 **Linkage mapping**
Ordering of genetic markers such as RFLPs, STRs, and other polymorphisms (about 200 per chromosome)
Genetic markers
2 **Physical mapping**
Ordering of large overlapping fragments cloned in YAC and BAC vectors, followed by ordering of smaller fragments cloned in phage and plasmid vectors
Overlapping fragments
3 **DNA sequencing**
Determination of nucleotide sequence of each small fragment and assembly of the partial sequences into the complete genome sequence
···GACTTCATCGGTATCGAACT···

▲ **Figure 21.2 Three-stage approach to sequencing an entire genome.** Starting with a cytogenetic map of each chromosome, researchers with the Human Genome Project proceeded through three stages to reach the ultimate goal, the virtually complete nucleotide sequence of every chromosome.

The next stage was the physical mapping of the human genome. In a **physical map**, the distances between markers are expressed by some physical measure, usually the number of base pairs along the DNA. For whole-genome mapping, a physical map is made by cutting the DNA of each chromosome into a number of restriction fragments and then determining the original order of the fragments in the chromosomal DNA. The key is to make fragments that overlap and then use probes or automated nucleotide sequencing of the ends to find the overlaps (see Figure 21.2, stage ②). In this way, fragments can be assigned to a sequential order that corresponds to their order in a chromosome.

The DNA fragments used for physical mapping were prepared by DNA cloning. With such a large genome, researchers had to carry out several rounds of DNA cutting, cloning, and physical mapping. In this approach, the first cloning vector was often a yeast artificial chromosome (YAC), which can carry inserted fragments a million base pairs long, or a bacterial artificial chromosome (BAC), which typically carries inserts of 100,000–300,000 base pairs. After such long fragments were put in order, each fragment was cut into smaller pieces, which were cloned in plasmids or phages, ordered in turn, and finally sequenced.

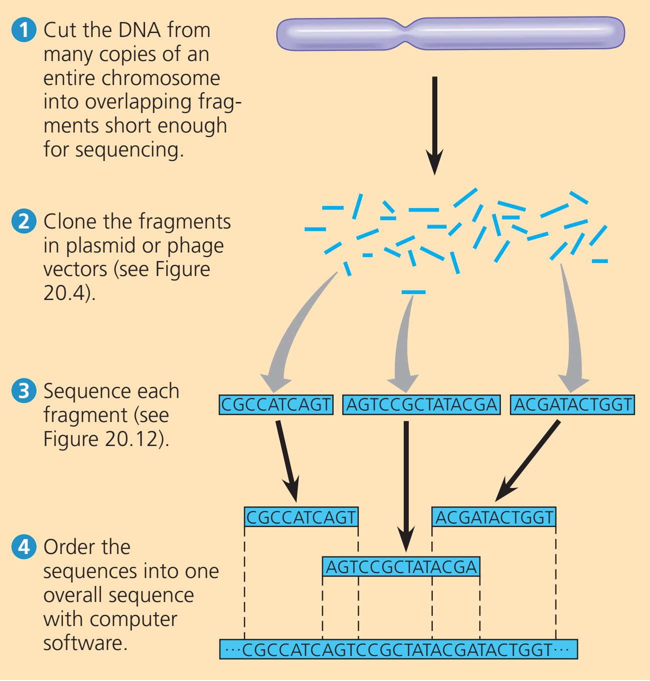
The ultimate goal in mapping any genome is to determine the complete nucleotide sequence of each chromosome (see Figure 21.2, stage 3). For the human genome, this was accomplished by sequencing machines, using the dideoxy chain termination method described in Figure 20.12. Even with automation, the sequencing of all 3 billion base pairs in a haploid set of human chromosomes presented a formidable challenge. In fact, a major thrust of the Human Genome Project was the development of technology for faster sequencing. Improvements over the years chipped away at each time-consuming step, enabling the rate of sequencing to accelerate impressively: Whereas a productive lab could typically sequence 1,000 base pairs a day in the 1980s, by the year 2000 each research center working on the Human Genome Project was sequencing 1,000 base pairs *per second*, 24 hours a day, seven days a week. Methods like this that can analyze biological materials very rapidly and produce enormous volumes of data are said to be "high-throughput." Sequencing machines are an example of high-throughput devices.

In practice, the three stages shown in Figure 21.2 overlapped in a way that our simplified version does not portray, but they accurately represent the overarching strategy employed in the Human Genome Project. During the project, however, an alternative strategy for genome sequencing emerged that was extremely efficient and became widely adopted.

Whole-Genome Shotgun Approach to Genome Sequencing

In 1992, emboldened by advances in sequencing and computer technology, molecular biologist J. Craig Venter devised an alternative approach to the sequencing of whole genomes. Called the *whole-genome shotgun approach*, it essentially skips the linkage mapping and physical mapping stages and starts directly with the sequencing of DNA fragments from randomly cut DNA. Powerful computer programs then assemble the resulting very large number of overlapping short sequences into a single continuous sequence **(Figure 21.3)**. In 1998, despite the skepticism of many scientists, Venter set up a company (Celera Genomics) and declared his intention to sequence the entire human genome. Five years later, and 13 years after the Human Genome Project began, Celera Genomics and the public consortium jointly announced that sequencing of the human genome was largely complete.

Representatives of the public consortium point out that Celera's accomplishment relied heavily on the consortium's maps and sequence data and that the infrastructure established by their approach was a tremendous aid to Celera's efforts. Venter, on the other hand, has argued for the efficiency and economy of Celera's methods, and indeed, the public consortium made some use of them as well. Evidently, both approaches made valuable contributions.


1 Cut the DNA from many copies of an entire chromosome into overlapping fragments short enough for sequencing.

2 Clone the fragments in plasmid or phage vectors (see Figure 20.4).

3 Sequence each fragment (see Figure 20.12).

4 Order the sequences into one overall sequence with computer software.

▲ **Figure 21.3 Whole-genome shotgun approach to sequencing.** In this approach, developed by Craig Venter and colleagues at the company he founded, Celera Genomics, random DNA fragments are sequenced and then ordered relative to each other. Compare this approach with the hierarchical, three-stage approach shown in Figure 21.2.

? *The fragments in stage 2 of this figure are depicted as scattered, whereas those in stage 2 of Figure 21.2 are drawn in a much more orderly fashion. How do these depictions reflect the two approaches?*

Today, the whole-genome shotgun approach is widely used. Also, the development of newer sequencing techniques, generally called *sequencing by synthesis* (see Chapter 20), has resulted in massive increases in speed and decreases in the cost of sequencing entire genomes. In these new techniques, many very small fragments (fewer than 100 base pairs) are sequenced at the same time, and computer software rapidly assembles the complete sequence. Because of the sensitivity of these techniques, the fragments can be sequenced directly; the cloning step (stage 2 in Figure 21.3) is unnecessary. Whereas sequencing the first human genome took 13 years and cost $100 million, James Watson's genome was sequenced during four months in 2007 for about $1 million, and a group of researchers reported in 2010 that they had rapidly sequenced three human genomes for approximately $4,400 each!

These technological advances have also facilitated an approach called **metagenomics** (from the Greek *meta*, beyond), in which DNA from a group of species (a *metagenome*) is collected from an environmental sample and sequenced. Again, computer software accomplishes the task of sorting out the partial sequences and assembling them into specific genomes. So far, this approach has been applied to microbial communities found in environments as diverse as the Sargasso Sea and the human intestine. The ability to sequence the DNA of mixed populations eliminates the need to culture each species separately in the lab, a difficulty that has limited the study of many microbial species.

At first glance, genome sequences of humans and other organisms are simply dry lists of nucleotide bases—millions of A's, T's, C's, and G's in mind-numbing succession. Crucial to making sense of this massive amount of data have been new analytical approaches, which we discuss next.

CONCEPT CHECK 21.1

1. What is the major difference between a linkage map and a physical map of a chromosome?
2. In general, how does the approach to genome mapping used in the Human Genome Project differ from the whole-genome shotgun approach?

For suggested answers, see Appendix A.

CONCEPT 21.2

Scientists use bioinformatics to analyze genomes and their functions

Each of the 20 or so sequencing centers around the world working on the Human Genome Project churned out voluminous amounts of DNA sequence day after day. As the data began to accumulate, the need to coordinate efforts to keep track of all the sequences became clear. Thanks to the foresight of research scientists and government officials involved in the Human Genome Project, its goals included the establishment of banks of data, or databases, and the refining of analytical software. These databases and software programs would then be centralized and made readily accessible on the Internet. Accomplishing this aim has accelerated progress in DNA sequence analysis by making bioinformatics resources available to researchers worldwide and by speeding up the dissemination of information.

Centralized Resources for Analyzing Genome Sequences

Government-funded agencies carried out their mandate to establish databases and provide software with which scientists could analyze the sequence data. For example, in the United States, a joint endeavor between the National Library of Medicine and the National Institutes of Health (NIH) created the National Center for Biotechnology Information (NCBI), which maintains a website (www.ncbi.nlm.nih.gov) with extensive bioinformatics resources. On this site are links to databases, software, and a wealth of information about genomics and related topics. Similar websites have also been established by the European Molecular Biology Laboratory, the DNA Data Bank of Japan, and BGI (formerly known as the Beijing Genome Institute) in Shenzhen, China, three genome centers with which the NCBI collaborates. These large, comprehensive websites are complemented by others maintained by individual or small groups of laboratories. Smaller websites often provide databases and software designed for a narrower purpose, such as studying genetic and genomic changes in one particular type of cancer.

The NCBI database of sequences is called GenBank. As of May 2010, it included the sequences of 119 million fragments of genomic DNA, totaling 114 billion base pairs! GenBank is constantly updated, and the amount of data it contains is estimated to double approximately every 18 months. Any sequence in the database can be retrieved and analyzed using software from the NCBI website or elsewhere.

One software program available on the NCBI website, called BLAST, allows the visitor to compare a DNA sequence with every sequence in GenBank, base by base, to look for similar regions. Another program allows comparison of predicted protein sequences. Yet a third can search any protein sequence for common stretches of amino acids (domains) for which a function is known or suspected, and it can show a three-dimensional model of the domain alongside other relevant information (**Figure 21.4**, on the next page). There is even a software program that can compare a collection of sequences, either nucleic acids or polypeptides, and diagram them in the form of an evolutionary tree based on the sequence relationships. (One such diagram is shown in Figure 21.16.)

Two research institutions, Rutgers University and the University of California, San Diego, also maintain a worldwide Protein Data Bank, a database of all three-dimensional protein structures that have been determined. (The database is accessible at www.wwpdb.org.) These structures can be rotated by the viewer to show all sides of the protein.

There is a vast array of resources available for researchers anywhere in the world to use. Let us now consider the types of questions scientists can address using these resources.

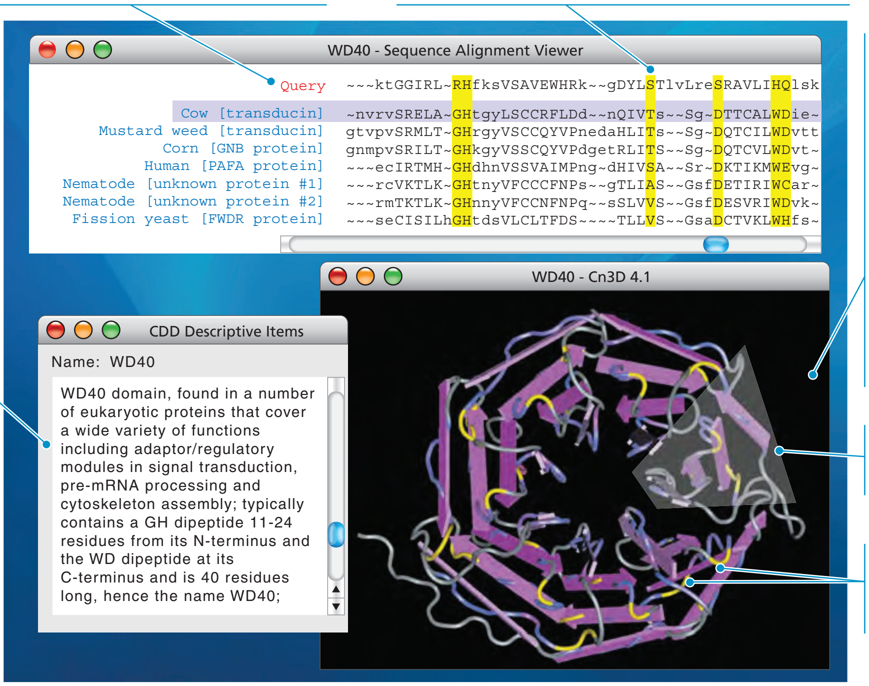
Identifying Protein-Coding Genes and Understanding Their Functions

Using available DNA sequences, geneticists can study genes directly, without having to infer genotype from phenotype as in classical genetics. But this approach, called *reverse genetics*, poses a new challenge: determining the phenotype from the genotype. Given a long DNA sequence from a database such as GenBank, the aim of scientists is to identify all protein-coding genes in the sequence and ultimately their functions. This process is called **gene annotation**.

In the past, gene annotation was carried out laboriously by individual scientists interested in particular genes, but the process has now been largely automated. The usual approach is to use software to scan the stored sequences for transcriptional and translational start and stop signals, for RNA-splicing sites, and for other telltale signs of protein-coding genes. The software also looks for certain short sequences that specify known mRNAs. Thousands of such sequences, called

In this window, a partial amino acid sequence from an unknown muskmelon protein ("Query") is aligned with sequences from other proteins that the computer program found to be similar. Each sequence represents a domain called WD40.

Four hallmarks of the WD40 domain are highlighted in yellow. (Sequence similarity is based on chemical aspects of the amino acids, so the amino acids in each hallmark region are not always identical.)



```
                        Query   ~~~ktGGIRL~RHfksVSAVEWHRk~~gDYLSTlvLreSRAVLIHQlsk

              Cow [transducin]  ~nvrvSRELA~GHtgyLSCCRFLDd~~nQIVTs~~Sg~DTTCALWDie~
     Mustard weed [transducin]  gtvpvSRMLT~GHrgyVSCCQYVPnedaHLITs~~Sg~DQTCILWDvtt
           Corn [GNB protein]   gnmpvSRILT~GHkgyVSSCQYVPdgetRLITS~~Sg~DQTCVLWDvt~
         Human [PAFA protein]   ~~~ecIRTMH~GHdhnVSSVAIMPng~dHIVSA~~Sr~DKTIKMWEvg~
Nematode [unknown protein #1]   ~~~rcVKTLK~GHtnyVFCCCFNPs~~gTLIAS~~GsfDETIRIWCar~
Nematode [unknown protein #2]   ~~~rmTKTLK~GHnnyVFCCNFNPq~~sSLVVS~~GsfDESVRIWDvk~
  Fission yeast [FWDR protein]  ~~~seCISILhGHtdsVLCLTFDS~~~~TLLVS~~GsaDCTVKLWHfs~
```

Name: WD40

WD40 domain, found in a number of eukaryotic proteins that cover a wide variety of functions including adaptor/regulatory modules in signal transduction, pre-mRNA processing and cytoskeleton assembly; typically contains a GH dipeptide 11-24 residues from its N-terminus and the WD dipeptide at its C-terminus and is 40 residues long, hence the name WD40;

This window displays information about the WD40 domain from the Conserved Domain Database.

The Cn3D program displays a three-dimensional ribbon model of cow transducin (the protein highlighted in purple in the Sequence Alignment Viewer). This protein is the only one of those shown for which a structure has been determined. The sequence similarity of the other proteins to cow transducin suggests that their structures are likely to be similar.

Cow transducin contains seven WD40 domains, one of which is highlighted here in gray.

The yellow segments correspond to the WD40 hallmarks highlighted in yellow in the window above.

▲ **Figure 21.4 Bioinformatics tools available on the Internet.** A website maintained by the National Center for Biotechnology Information allows scientists and the public to access DNA and protein sequences and other stored data. The site includes a link to a protein structure database (Conserved Domain Database, CDD) that can find and describe similar domains in related proteins, as well as software (Cn3D, "See in 3D") that displays three-dimensional models of domains for which the structure has been determined. Some results are shown from a search for regions of proteins similar to an amino acid sequence in a muskmelon protein.

expressed sequence tags, or *ESTs*, have been collected from cDNA sequences and are cataloged in computer databases. This type of analysis identifies sequences that may be previously unknown protein-coding genes.

The identities of about half of the human genes were known before the Human Genome Project began. But what about the others, the previously unknown genes revealed by analysis of DNA sequences? Clues about their identities and functions come from comparing sequences that might be genes with known genes from other organisms, using the software described previously. Due to redundancy in the genetic code, the DNA sequence itself may vary more than the protein sequence does. Thus, scientists interested in proteins often compare the predicted amino acid sequence of a protein to that of other proteins.

Sometimes a newly identified sequence will match, at least partially, the sequence of a gene or protein whose function is well known. For example, part of a new gene may match a known gene that encodes an important signaling pathway protein such as a protein kinase (see Chapter 11), suggesting that the new gene does, too. Alternatively, the new gene sequence may be similar to a previously encountered sequence whose function is still unknown. Another possibility is that the sequence is entirely unlike anything ever seen before. This was true for about a third of the genes of *E. coli* when its genome was sequenced. In the last case, protein function is usually deduced through a combination of biochemical and functional studies. The biochemical approach aims to determine the three-dimensional structure of the protein as well as other attributes, such as potential binding sites for other molecules. Functional studies usually involve blocking or disabling the gene to see how the phenotype is affected. RNAi, described in Chapter 20, is an example of an experimental technique used to block gene function.

Understanding Genes and Gene Expression at the Systems Level

The impressive computational power provided by the tools of bioinformatics allows the study of whole sets of genes and their interactions, as well as the comparison of genomes from different species. Genomics is a rich source of new insights into fundamental questions about genome organization, regulation of gene expression, growth and development, and evolution.

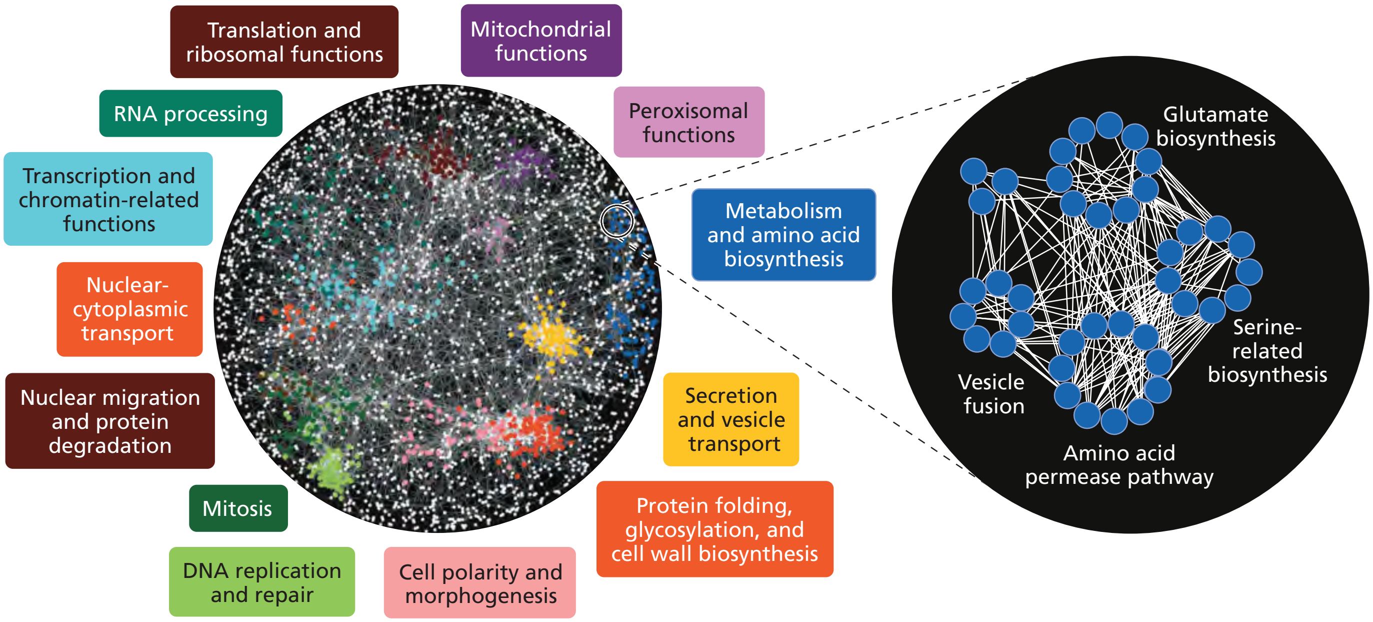
One informative approach has been taken by a research project called ENCODE (Encyclopedia of DNA Elements), which began in 2003. First, researchers focused intensively on 1% of the human genome and attempted to learn all they could about the functionally important elements in that sequence. They looked for protein-coding genes and genes for noncoding RNAs as well as sequences that regulate DNA replication, gene expression (such as enhancers and promoters), and chromatin modifications. The pilot project was completed in 2007, yielding a wealth of information. One big surprise, discussed in Concept 18.3, was that over 90% of the region was transcribed into RNA, even though less than 2% codes for proteins. The success of this approach has led to two follow-up studies, one extending the analysis to the entire human genome and the other analyzing in a similar fashion the genomes of two model organisms, the soil nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. Because genetic and molecular biological experiments can be performed on these species, testing the activities of potentially functional DNA elements in their genomes will reveal much about how the human genome works.

The success in sequencing genomes and studying entire sets of genes has encouraged scientists to attempt similar systematic study of the full protein sets (*proteomes*) encoded by genomes, an approach called **proteomics**. Proteins, not the genes that encode them, actually carry out most of the activities of the cell. Therefore, we must study when and where proteins are produced in an organism, as well as how they interact in networks, if we are to understand the functioning of cells and organisms.

How Systems Are Studied: An Example

Genomics and proteomics are enabling molecular biologists to approach the study of life from an increasingly global perspective. Using the tools we have described, biologists have begun to compile catalogs of genes and proteins—listings of all the "parts" that contribute to the operation of cells, tissues, and organisms. With such catalogs in hand, researchers have shifted their attention from the individual parts to their functional integration in biological systems. As you may recall, in Chapter 1 we discussed this systems biology approach, which aims to model the dynamic behavior of whole biological systems.

One important use of the systems biology approach is to define gene circuits and protein interaction networks. To map the protein interaction network in the yeast *Saccharomyces cerevisiae*, for instance, researchers used sophisticated techniques to knock out (disable) pairs of genes, one pair at a time, creating doubly mutant cells. They then compared the fitness of each double mutant (based in part on the size of the cell colony it formed) to that predicted from the fitnesses of the two single mutants. The researchers reasoned that if the observed fitness matched the prediction, then the products of the two genes didn't interact with each other, but if the observed fitness was greater or less than predicted, then the gene products interacted in the cell. Computer software then mapped genes based on the similarity of their interactions; a network-like "functional map" of these genetic interactions is displayed in **Figure 21.5**. To process the vast number of protein-protein interactions generated by this experiment and integrate them


Translation and ribosomal functions
Mitochondrial functions
RNA processing
Peroxisomal functions
Transcription and chromatin-related functions
Metabolism and amino acid biosynthesis
Nuclear-cytoplasmic transport
Nuclear migration and protein degradation
Secretion and vesicle transport
Mitosis
Protein folding, glycosylation, and cell wall biosynthesis
DNA replication and repair
Cell polarity and morphogenesis
Glutamate biosynthesis
Serine-related biosynthesis
Vesicle fusion
Amino acid permease pathway

▲ **Figure 21.5 The systems biology approach to protein interactions.** This global protein interaction map shows the likely interactions (lines) among about 4,500 gene products (circles) in the yeast *Saccharomyces cerevisiae*. Circles of the same color represent gene products involved in one of the 13 cellular functions listed around the map. The blowup shows additional details of one map region where the gene products (blue circles) carry out amino acid biosynthesis, uptake, and related functions.

into the completed map required powerful computers, mathematical tools, and newly developed software. Thus, the systems biology approach has really been made possible by advances in computer technology and bioinformatics.

Application of Systems Biology to Medicine

The Cancer Genome Atlas is another example of systems biology in which a large group of interacting genes and gene products are analyzed together. This project, under the joint leadership of the National Cancer Institute and the NIH, aims to determine how changes in biological systems lead to cancer. A three-year pilot project beginning in 2007 set out to find all the common mutations in three types of cancer—lung cancer, ovarian cancer, and glioblastoma of the brain—by comparing gene sequences and patterns of gene expression in cancer cells with those in normal cells. Work on glioblastoma has confirmed the role of several suspected genes and identified a few unknown ones, suggesting possible new targets for therapies. The approach has proved so fruitful for these three types of cancer that it has been extended to ten other types, chosen because they are common and often lethal in humans.

Systems biology has tremendous potential in human medicine that is just starting to be explored. Silicon and glass "chips" have been developed that hold a microarray of most of the known human genes **(Figure 21.6)**. Such chips are being used to analyze gene expression patterns in patients suffering from various cancers and other diseases, with the eventual aim of tailoring their treatment to their unique genetic makeup and the specifics of their cancers. This approach has had modest success in characterizing subsets of several cancers.

Ultimately, people may carry with their medical records a catalog of their DNA sequence, a sort of genetic bar code, with regions highlighted that predispose them to specific diseases. The use of such sequences for personalized medicine—disease prevention and treatment—has great potential.

Systems biology is a very efficient way to study emergent properties at the molecular level. Recall from Chapter 1 that according to the theme of emergent properties, novel properties arise at each successive level of biological complexity as a result of the arrangement of building blocks at the underlying level. The more we can learn about the arrangement and interactions of the components of genetic systems, the deeper will be our understanding of whole organisms. The rest of this chapter will survey what we've learned from genomic studies thus far.

◀ **Figure 21.6 A human gene microarray chip.** Tiny spots of DNA arranged in a grid on this silicon wafer represent almost all of the genes in the human genome. Using this chip, researchers can analyze expression patterns for all these genes at the same time.

CONCEPT CHECK 21.2

1. What role does the Internet play in current genomics and proteomics research?
2. Explain the advantage of the systems biology approach to studying cancer versus the approach of studying a single gene at a time.
3. **MAKE CONNECTIONS** The ENCODE pilot project found that more than 90% of the genomic region being studied was transcribed into RNAs, far more than could be accounted for by protein-coding genes. Review Concept 18.3 (pp. 364–366) and suggest some roles that these RNAs might play.
4. **MAKE CONNECTIONS** In Concept 20.2 (p. 411), you learned about genome-wide association studies. Explain how these studies use the systems biology approach.

For suggested answers, see Appendix A.

CONCEPT **21.3**

Genomes vary in size, number of genes, and gene density

By early 2010, the sequencing of about 1,200 genomes had been completed and that of over 5,500 genomes and over 200 metagenomes was in progress. In the completely sequenced group, about 1,000 are genomes of bacteria, and 80 are archaeal genomes. Among the 124 eukaryotic species in the group are vertebrates, invertebrates, protists, fungi, and plants. The accumulated genome sequences contain a wealth of information that we are now beginning to mine. What have we learned so far by comparing the genomes that have been sequenced? In this section, we will examine the characteristics of genome size, number of genes, and gene density. Because these characteristics are so broad, we will focus on general trends, for which there are often exceptions.

Genome Size

Comparing the three domains (Bacteria, Archaea, and Eukarya), we find a general difference in genome size between prokaryotes and eukaryotes **(Table 21.1)**. While there are some exceptions, most bacterial genomes have between 1 and 6 million base pairs (Mb); the genome of *E. coli*, for instance, has 4.6 Mb. Genomes of archaea are, for the most part, within the size range of bacterial genomes. (Keep in mind, however, that many fewer archaeal genomes have

Table 21.1 Genome Sizes and Estimated Numbers of Genes*

Organism	Haploid Genome Size (Mb)	Number of Genes	Genes per Mb
Bacteria			
Haemophilus influenzae	1.8	1,700	940
Escherichia coli	4.6	4,400	950
Archaea			
Archaeoglobus fulgidus	2.2	2,500	1,130
Methanosarcina barkeri	4.8	3,600	750
Eukaryotes			
Saccharomyces cerevisiae (yeast, a fungus)	12	6,300	525
Caenorhabditis elegans (nematode)	100	20,100	200
Arabidopsis thaliana (mustard family plant)	120	27,000	225
Drosophila melanogaster (fruit fly)	165	13,700	83
Oryza sativa (rice)	430	42,000	98
Zea mays (corn)	2,300	32,000	14
Mus musculus (house mouse)	2,600	22,000	11
Ailuropoda melanoleuca (giant panda)	2,400	21,000	9
Homo sapiens (human)	3,000	<21,000	7
Fritillaria assyriaca (lily family plant)	124,000	ND	ND

*Some values given here are likely to be revised as genome analysis continues. Mb = million base pairs. ND = not determined.

been completely sequenced, so this picture may change.) Eukaryotic genomes tend to be larger: The genome of the single-celled yeast *Saccharomyces cerevisiae* (a fungus) has about 12 Mb, while most animals and plants, which are multicellular, have genomes of at least 100 Mb. There are 165 Mb in the fruit fly genome, while humans have 3,000 Mb, about 500 to 3,000 times as many as a typical bacterium.

Aside from this general difference between prokaryotes and eukaryotes, a comparison of genome sizes among eukaryotes fails to reveal any systematic relationship between genome size and the organism's phenotype. For instance, the genome of *Fritillaria assyriaca*, a flowering plant in the lily family, contains 124 billion base pairs (124,000 Mb), about 40 times the size of the human genome. Even more striking, there is a single-celled amoeba, *Polychaos dubia*, whose genome size has been estimated at 670,000 Mb. (This genome has not yet been sequenced.) On a finer scale, comparing two insect species, the cricket (*Anabrus simplex*) genome turns out to have 11 times as many base pairs as the *Drosophila melanogaster* genome. There is a wide range of genome sizes within the groups of protists, insects, amphibians, and plants and less of a range within mammals and reptiles.

Number of Genes

The number of genes also varies between prokaryotes and eukaryotes: Bacteria and archaea, in general, have fewer genes than eukaryotes. Free-living bacteria and archaea have from 1,500 to 7,500 genes, while the number of genes in eukaryotes ranges from about 5,000 for unicellular fungi to at least 40,000 for some multicellular eukaryotes (see Table 21.1).

Within the eukaryotes, the number of genes in a species is often lower than expected from simply considering the size of its genome. Looking at Table 21.1, you can see that the genome of the nematode *C. elegans* is 100 Mb in size and contains roughly 20,000 genes. The *Drosophila* genome, in comparison, is much bigger (165 Mb) but has about two-thirds the number of genes—only 13,700 genes.

Considering an example closer to home, we noted that the human genome contains 3,000 Mb, well over ten times the size of either the *Drosophila* or *C. elegans* genome. At the outset of the Human Genome Project, biologists expected somewhere between 50,000 and 100,000 genes to be identified in the completed sequence, based on the number of known human proteins. As the project progressed, the estimate was revised downward several times, and in 2010, the most reliable count placed the number at fewer than 21,000. This relatively low number, similar to the number of genes in the nematode *C. elegans*, has surprised biologists, who had clearly expected many more human genes.

What genetic attributes allow humans (and other vertebrates) to get by with no more genes than nematodes? An important factor is that vertebrate genomes "get more bang for the buck" from their coding sequences because of extensive alternative splicing of RNA transcripts. Recall that this process generates more than one functional protein from a single gene (see Figure 18.13). A typical human gene contains about ten exons, and an estimated 93% or so of these multi-exon genes are spliced in at least two different ways. Some genes are expressed in hundreds of alternatively spliced forms, others in just two. It is not yet possible to catalog all of the different forms, but it is clear that the number of different proteins encoded in the human genome far exceeds the proposed number of genes.

Additional polypeptide diversity could result from post-translational modifications such as cleavage or the addition of carbohydrate groups in different cell types or at different developmental stages. Finally, the discovery of miRNAs and other small RNAs that play regulatory roles have added a new variable to the mix (see Concept 18.3). Some scientists think that this added level of regulation, when present, may contribute to greater organismal complexity for a given number of genes.

Gene Density and Noncoding DNA

In addition to genome size and number of genes, we can compare gene density in different species—in other words, how many genes there are in a given length of DNA. When we compare the genomes of bacteria, archaea, and eukaryotes, we see that eukaryotes generally have larger genomes but fewer genes in a given number of base pairs. Humans have hundreds or thousands of times as many base pairs in their genome as most bacteria, as we already noted, but only 5 to 15 times as many genes; thus, gene density is lower in humans (see Table 21.1). Even unicellular eukaryotes, such as yeasts, have fewer genes per million base pairs than bacteria and archaea. Among the genomes that have been sequenced completely thus far, humans and other mammals have the lowest gene density.

In all bacterial genomes studied so far, most of the DNA consists of genes for protein, tRNA, or rRNA; the small amount remaining consists mainly of nontranscribed regulatory sequences, such as promoters. The sequence of nucleotides along a bacterial protein-coding gene proceeds from start to finish without interruption by noncoding sequences (introns). In eukaryotic genomes, by contrast, most of the DNA neither encodes protein nor is transcribed into RNA molecules of known function, and the DNA includes more complex regulatory sequences. In fact, humans have 10,000 times as much noncoding DNA as bacteria. Some of this DNA in multicellular eukaryotes is present as introns within genes. Indeed, introns account for most of the difference in average length between human genes (27,000 base pairs) and bacterial genes (1,000 base pairs).

In addition to introns, multicellular eukaryotes have a vast amount of non-protein-coding DNA between genes. In the next section, we will describe the composition and arrangement of these great stretches of DNA in the human genome.

CONCEPT CHECK 21.3

1. According to the best current estimate, the human genome contains fewer than 21,000 genes. However, there is evidence that human cells produce many more than 21,000 different polypeptides. What processes might account for this discrepancy?
2. The number of sequenced genomes is constantly being updated. Go to www.genomesonline.org to find the current number of completed genomes for each domain as well as the number of genomes whose sequencing is in progress. (*Hint*: Click on "Enter GOLD," and then click on "Published Complete Genomes" for extra information.)
3. **WHAT IF?** What evolutionary processes might account for prokaryotes having smaller genomes than eukaryotes?

For suggested answers, see Appendix A.

CONCEPT 21.4

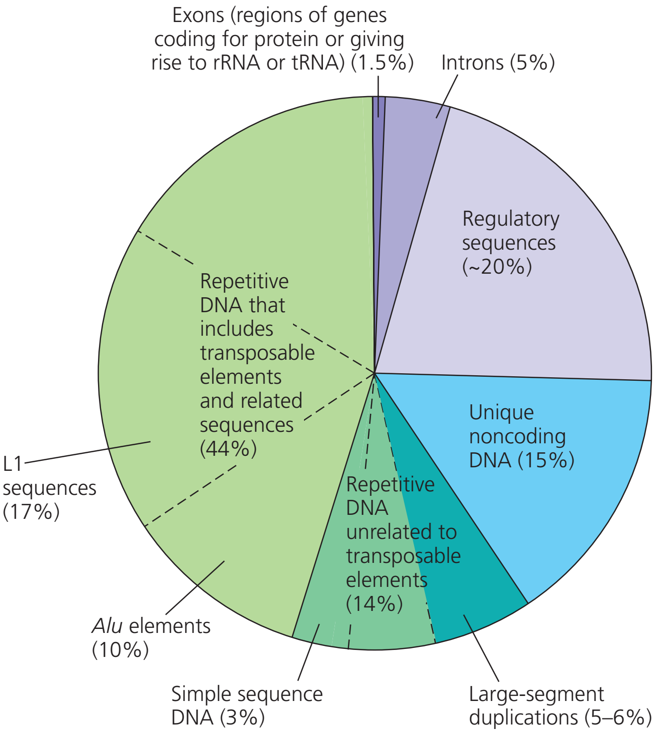
Multicellular eukaryotes have much noncoding DNA and many multigene families

We have spent most of this chapter, and indeed this unit, focusing on genes that code for proteins. Yet the coding regions of these genes and the genes for RNA products such as rRNA, tRNA, and miRNA make up only a small portion of the genomes of most multicellular eukaryotes. The bulk of many eukaryotic genomes consists of DNA sequences that neither code for proteins nor are transcribed to produce RNAs with known functions; this noncoding DNA was often described in the past as "junk DNA." However, much evidence is accumulating that this DNA plays important roles in the cell, an idea supported by its persistence in diverse genomes over many hundreds of generations. For example, comparison of the genomes of humans, rats, and mice has revealed the presence of almost 500 regions of noncoding DNA that are identical in sequence in all three species. This is a higher level of sequence conservation than is seen for protein-coding regions in these species, strongly suggesting that the noncoding regions have important functions. In this section, we examine how genes and noncoding DNA sequences are organized within genomes of multicellular eukaryotes, using the human genome as our main example. Genome organization tells us much about how genomes have evolved and continue to evolve, the next subject we'll consider.

Once the sequencing of the human genome was completed, it became clear that only a tiny part—1.5%—codes for proteins or is transcribed into rRNAs or tRNAs. **Figure 21.7** shows what is known about the makeup of the remaining 98.5%. Gene-related regulatory sequences and introns account, respectively, for 5% and about 20% of the human genome. The rest, located between functional genes, includes some unique noncoding DNA, such as gene fragments and **pseudogenes**, former genes that have accumulated mutations over a long time and no longer produce functional proteins. (The genes that produce small noncoding RNAs are a tiny percentage of the genome, distributed between the 20% introns and the 15% unique noncoding DNA.) Most intergenic DNA, however, is **repetitive DNA**, which consists of sequences that are present in multiple copies in the genome. Somewhat surprisingly, about 75% of this repetitive DNA (44% of the entire human genome) is made up of units called transposable elements and sequences related to them.

Transposable Elements and Related Sequences

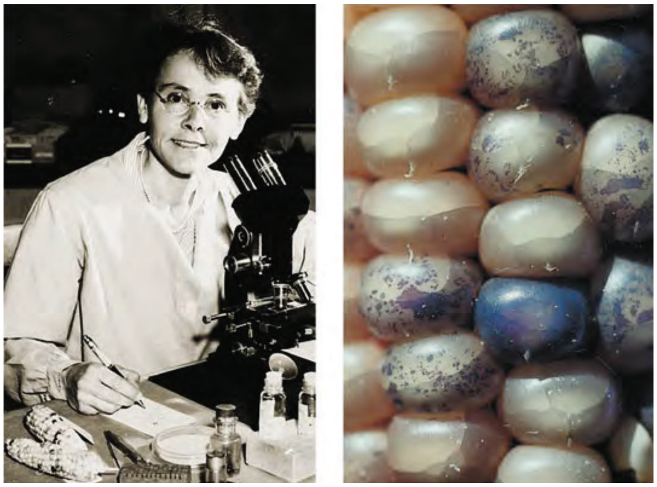
Both prokaryotes and eukaryotes have stretches of DNA that can move from one location to another within the genome. These stretches are known as *transposable genetic elements*, or



▲ **Figure 21.7 Types of DNA sequences in the human genome.** The gene sequences that code for proteins or are transcribed into rRNA or tRNA molecules make up only about 1.5% of the human genome (dark purple in the pie chart), while introns and regulatory sequences associated with genes (light purple) make up about a quarter. The vast majority of the human genome does not code for proteins or give rise to known RNAs, and much of it is repetitive DNA (dark and light green and teal). Because repetitive DNA is the most difficult to sequence and analyze, classification of some portions is tentative, and the percentages given here may shift slightly as genome analysis proceeds. The genes that are transcribed into small noncoding RNAs such as miRNAs, which were recently discovered, are found among unique noncoding DNA sequences and within introns and thus would be included in two segments of this chart.

simply **transposable elements**. During the process called *transposition*, a transposable element moves from one site in a cell's DNA to a different target site by a type of recombination process. Transposable elements are sometimes called "jumping genes," but it should be kept in mind that they never completely detach from the cell's DNA. Instead, the original and new DNA sites are brought together by enzymes and other proteins that bend the DNA.

The first evidence for wandering DNA segments came from American geneticist Barbara McClintock's breeding experiments with Indian corn (maize) in the 1940s and 1950s **(Figure 21.8)**. As she tracked corn plants through multiple generations, McClintock identified changes in the color of corn kernels that made sense only if she postulated the existence of genetic elements capable of moving from other locations in the genome into the genes for kernel color, disrupting the genes so that the kernel color was changed. McClintock's discovery was met with great skepticism and virtually discounted at the time. Her careful work and insightful ideas were finally validated many years later when transposable elements were found in bacteria. In 1983, at the age of 81, McClintock received the Nobel Prize for her pioneering research.

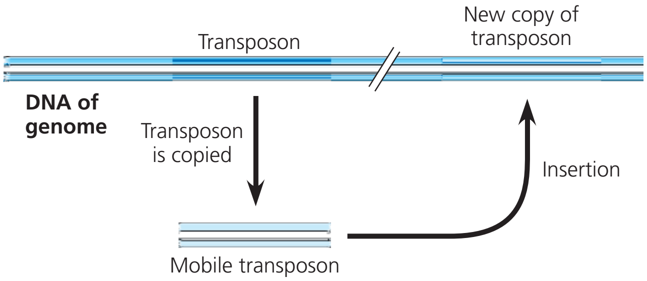


▲ **Figure 21.8 The effect of transposable elements on corn kernel color.** Barbara McClintock first proposed the idea of mobile genetic elements after observing variegations in corn kernel color (right).

Movement of Transposons and Retrotransposons

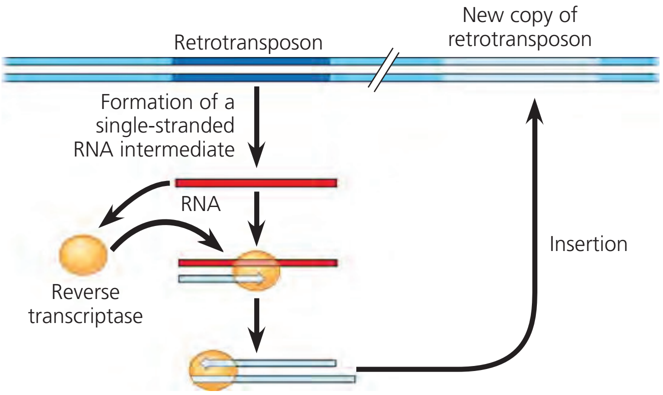
Eukaryotic transposable elements are of two types. The first type are **transposons**, which move within a genome by means of a DNA intermediate. Transposons can move by a "cut-and-paste" mechanism, which removes the element from the original site, or by a "copy-and-paste" mechanism, which leaves a copy behind **(Figure 21.9)**. Both mechanisms require an enzyme called *transposase*, which is generally encoded by the transposon.

Most transposable elements in eukaryotic genomes are of the second type, **retrotransposons**, which move by means of an RNA intermediate that is a transcript of the retrotransposon DNA. Retrotransposons always leave a copy at the original site during transposition, since they are initially transcribed



▲ **Figure 21.9 Transposon movement.** Movement of transposons by either the cut-and-paste mechanism or the copy-and-paste mechanism (shown here) involves a double-stranded DNA intermediate that is inserted into the genome.

? *How would this figure differ if it showed the cut-and-paste mechanism?*


Retrotransposon
New copy of retrotransposon
Formation of a single-stranded RNA intermediate
RNA
Insertion
Reverse transcriptase

▲ **Figure 21.10 Retrotransposon movement.** Movement begins with formation of a single-stranded RNA intermediate. The remaining steps are essentially identical to part of the retrovirus replicative cycle (see Figure 19.8).

into an RNA intermediate **(Figure 21.10)**. To insert at another site, the RNA intermediate is first converted back to DNA by reverse transcriptase, an enzyme encoded by the retrotransposon. (Reverse transcriptase is also encoded by retroviruses, as you learned in Chapter 19. In fact, retroviruses may have evolved from retrotransposons.) Another cellular enzyme catalyzes insertion of the reverse-transcribed DNA at a new site.

Sequences Related to Transposable Elements

Multiple copies of transposable elements and sequences related to them are scattered throughout eukaryotic genomes. A single unit is usually hundreds to thousands of base pairs long, and the dispersed "copies" are similar but usually not identical to each other. Some of these are transposable elements that can move; the enzymes required for this movement may be encoded by any transposable element, including the one that is moving. Others are related sequences that have lost the ability to move altogether. Transposable elements and related sequences make up 25–50% of most mammalian genomes (see Figure 21.7) and even higher percentages in amphibians and many plants. In fact, the very large size of some plant genomes is accounted for not by extra genes, but by extra transposable elements. For example, sequences like these make up 85% of the corn genome!

In humans and other primates, a large portion of transposable element–related DNA consists of a family of similar sequences called *Alu elements*. These sequences alone account for approximately 10% of the human genome. *Alu* elements are about 300 nucleotides long, much shorter than most functional transposable elements, and they do not code for any protein. However, many *Alu* elements are transcribed into RNA; its cellular function, if any, is currently unknown.

An even larger percentage (17%) of the human genome is made up of a type of retrotransposon called *LINE-1*, or *L1*. These sequences are much longer than *Alu* elements—about 6,500 base pairs—and have a low rate of transposition. What might account for this low rate? Recent research has uncovered the presence of sequences within L1 that block the progress of RNA polymerase, which is necessary for transposition. An accompanying genomic analysis found L1 sequences within the introns of nearly 80% of the human genes that were analyzed, suggesting that L1 may help regulate gene expression. Other researchers have proposed that L1 retrotransposons may have differential effects on gene expression in developing neurons, contributing to the great diversity of neuronal cell types (see Chapter 48).

Although many transposable elements encode proteins, these proteins do not carry out normal cellular functions. Therefore, transposable elements are usually included in the "noncoding" DNA category, along with other repetitive sequences.

Other Repetitive DNA, Including Simple Sequence DNA

Repetitive DNA that is not related to transposable elements probably arises due to mistakes during DNA replication or recombination. Such DNA accounts for about 14% of the human genome (see Figure 21.7). About a third of this (5–6% of the human genome) consists of duplications of long stretches of DNA, with each unit ranging from 10,000 to 300,000 base pairs. The large segments seem to have been copied from one chromosomal location to another site on the same or a different chromosome and probably include some functional genes.

In contrast to scattered copies of long sequences, **simple sequence DNA** contains many copies of tandemly repeated short sequences, as in the following example (showing one DNA strand only):

. . . GTTACGTTACGTTACGTTACGTTACGTTAC . . .

In this case, the repeated unit (GTTAC) consists of 5 nucleotides. Repeated units may contain as many as 500 nucleotides, but often contain fewer than 15 nucleotides, as in this example. When the unit contains 2–5 nucleotides, the series of repeats is called a **short tandem repeat**, or **STR**; we discussed the use of STR analysis in preparing genetic profiles in Chapter 20. The number of copies of the repeated unit can vary from site to site within a given genome. There could be as many as several hundred thousand repetitions of the GTTAC unit at one site, but only half that number at another. STR analysis is performed on sites selected because they have relatively few repeats. The repeat number can vary from person to person, and since humans are diploid, each person has two alleles per site, which can differ. This diversity produces the variation represented in the genetic profiles that result from STR analysis. Altogether, simple sequence DNA makes up 3% of the human genome.

Much of a genome's simple sequence DNA is located at chromosomal telomeres and centromeres, suggesting that this DNA plays a structural role for chromosomes. The DNA at centromeres is essential for the separation of chromatids in cell division (see Chapter 12). Centromeric DNA, along with simple sequence DNA located elsewhere, may also help organize the chromatin within the interphase nucleus. The simple sequence DNA located at telomeres, at the tips of chromosomes, prevents genes from being lost as the DNA shortens with each round of replication (see Chapter 16). Telomeric DNA also binds proteins that protect the ends of a chromosome from degradation and from joining to other chromosomes.

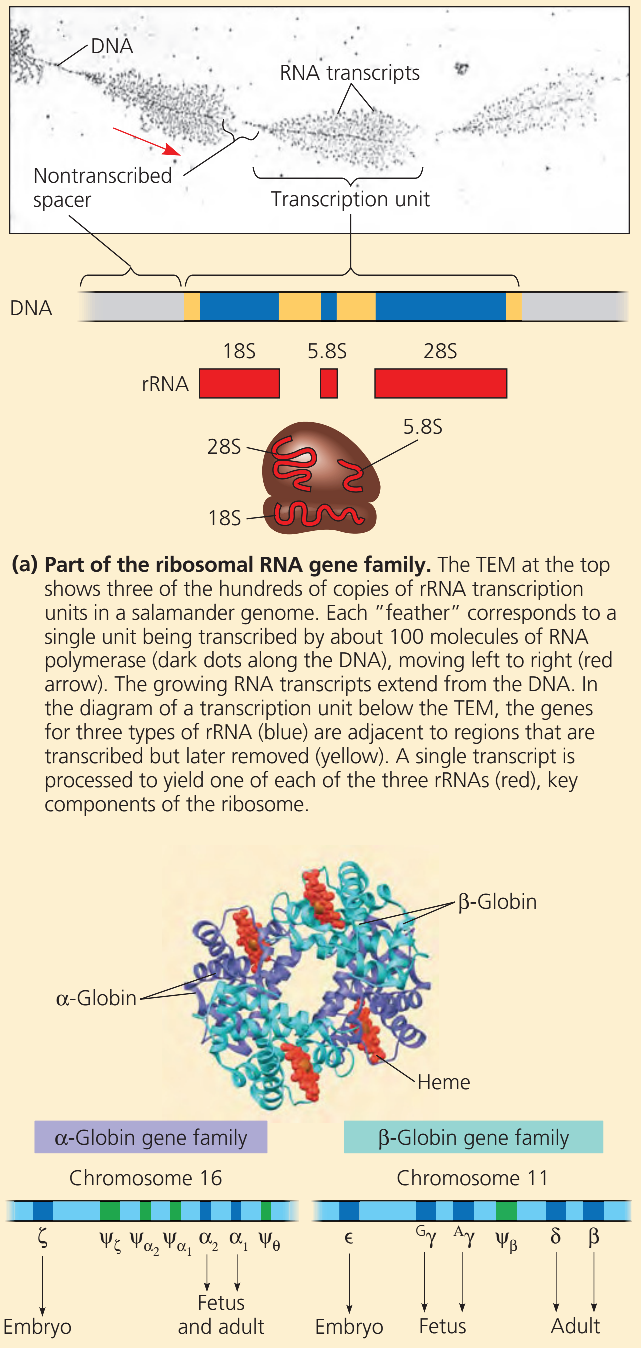
Genes and Multigene Families

We finish our discussion of the various types of DNA sequences in eukaryotic genomes with a closer look at genes. Recall that DNA sequences that code for proteins or give rise to tRNA or rRNA compose a mere 1.5% of the human genome (see Figure 21.7). If we include introns and regulatory sequences associated with genes, the total amount of DNA that is gene-related—coding and noncoding—constitutes about 25% of the human genome. Put another way, only about 6% (1.5% out of 25%) of the length of the average gene is represented in the final gene product.

Like the genes of bacteria, many eukaryotic genes are present as unique sequences, with only one copy per haploid set of chromosomes. But in the human genome and the genomes of many other animals and plants, solitary genes make up less than half of the total gene-related DNA. The rest occur in **multigene families**, collections of two or more identical or very similar genes.

In multigene families that consist of *identical* DNA sequences, those sequences are usually clustered tandemly and, with the notable exception of the genes for histone proteins, have RNAs as their final products. An example is the family of identical DNA sequences that are the genes for the three largest rRNA molecules **(Figure 21.11a)**. These rRNA molecules are transcribed from a single transcription unit that is repeated tandemly hundreds to thousands of times in one or several clusters in the genome of a multicellular eukaryote. The many copies of this rRNA transcription unit help cells to quickly make the millions of ribosomes needed for active protein synthesis. The primary transcript is cleaved to yield the three rRNA molecules, which combine with proteins and one other kind of rRNA (5S rRNA) to form ribosomal subunits.

The classic examples of multigene families of *nonidentical* genes are two related families of genes that encode globins, a group of proteins that include the α and β polypeptide subunits of hemoglobin. One family, located on chromosome 16 in humans, encodes various forms of α-globin; the other, on chromosome 11, encodes forms of β-globin **(Figure 21.11b)**. The different forms of each globin subunit are expressed at different times in development, allowing hemoglobin to function effectively in the changing environment of the developing animal. In humans, for example, the embryonic and fetal forms of hemoglobin have a higher affinity for oxygen



(a) Part of the ribosomal RNA gene family. The TEM at the top shows three of the hundreds of copies of rRNA transcription units in a salamander genome. Each "feather" corresponds to a single unit being transcribed by about 100 molecules of RNA polymerase (dark dots along the DNA), moving left to right (red arrow). The growing RNA transcripts extend from the DNA. In the diagram of a transcription unit below the TEM, the genes for three types of rRNA (blue) are adjacent to regions that are transcribed but later removed (yellow). A single transcript is processed to yield one of each of the three rRNAs (red), key components of the ribosome.

(b) The human α-globin and β-globin gene families. Adult hemoglobin is composed of two α-globin and two β-globin polypeptide subunits, as shown in the molecular model. The genes (dark blue) encoding α- and β-globins are found in two families, organized as shown here. The noncoding DNA separating the functional genes within each family includes pseudogenes (ψ; green), versions of the functional genes that no longer produce functional proteins. Genes and pseudogenes are named with Greek letters. Some genes are expressed only in the embryo or fetus.

▲ **Figure 21.11 Gene families.**

? *In (a), how could you determine the direction of transcription if it wasn't indicated by the red arrow?*

than the adult forms, ensuring the efficient transfer of oxygen from mother to fetus. Also found in the globin gene family clusters are several pseudogenes.

The arrangement of the genes in gene families has given biologists insight into the evolution of genomes. We will consider some of the processes that have shaped the genomes of different species over evolutionary time in the next section.

CONCEPT CHECK 21.4

1. Discuss the characteristics of mammalian genomes that make them larger than prokaryotic genomes.
2. Which of the three mechanisms described in Figures 21.9 and 21.10 result(s) in a copy remaining at the original site as well as appearing in a new location?
3. Contrast the organizations of the rRNA gene family and the globin gene families. For each, explain how the existence of a family of genes benefits the organism.
4. **MAKE CONNECTIONS** Assign each DNA segment at the top of Figure 18.8 (p. 359) to a sector in the pie chart in Figure 21.7.

For suggested answers, see Appendix A.

CONCEPT 21.5

Duplication, rearrangement, and mutation of DNA contribute to genome evolution

EVOLUTION The basis of change at the genomic level is mutation, which underlies much of genome evolution. It seems likely that the earliest forms of life had a minimal number of genes—those necessary for survival and reproduction. If this were indeed the case, one aspect of evolution must have been an increase in the size of the genome, with the extra genetic material providing the raw material for gene diversification. In this section, we will first describe how extra copies of all or part of a genome can arise and then consider subsequent processes that can lead to the evolution of proteins (or RNA products) with slightly different or entirely new functions.

Duplication of Entire Chromosome Sets

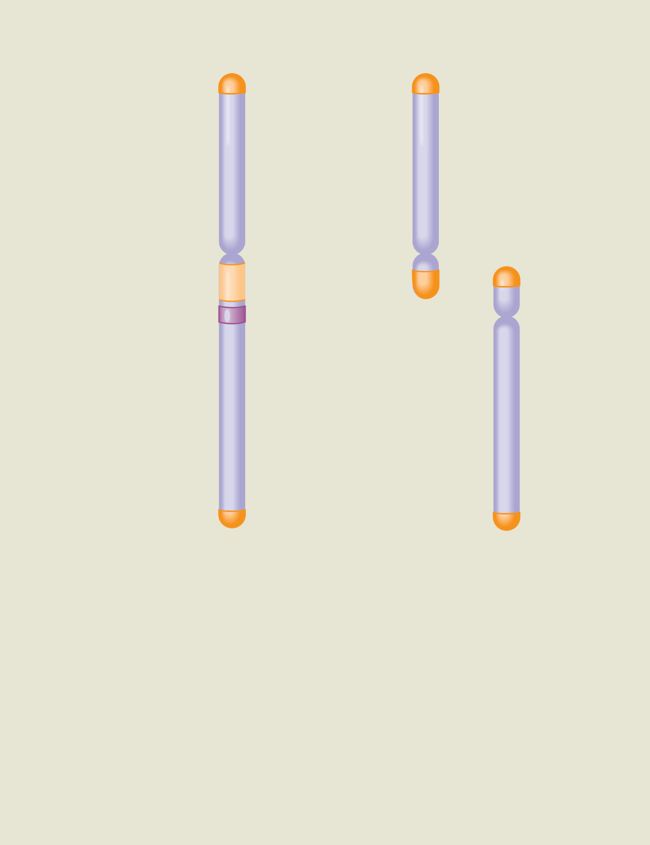
An accident in meiosis can result in one or more extra sets of chromosomes, a condition known as polyploidy. Although such accidents would most often be lethal, in rare cases they could facilitate the evolution of genes. In a polyploid organism, one set of genes can provide essential functions for the organism. The genes in the one or more extra sets can diverge by accumulating mutations; these variations may persist if the organism carrying them survives and reproduces. In this way, genes with novel functions can evolve. As long as one copy of an essential gene is expressed, the divergence of another copy can lead to its encoded protein acting in a novel way, thereby changing the organism's phenotype. The outcome of this accumulation of mutations may be the branching off of a new species, as happens often in flowering plants (see Chapter 24). Polyploid animals also exist, but they are much rarer; the tetraploid model organism *Xenopus laevis*, the African clawed frog, is an example.

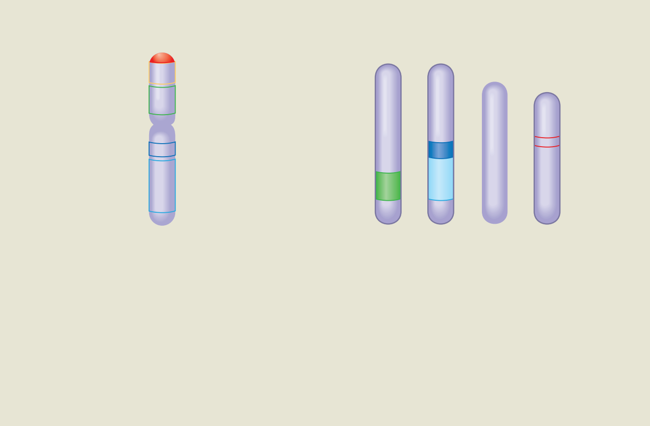
Alterations of Chromosome Structure

Scientists have long known that sometime in the last 6 million years, when the ancestors of humans and chimpanzees diverged as species, the fusion of two ancestral chromosomes in the human line led to different haploid numbers for humans ($n = 23$) and chimpanzees ($n = 24$). The banding patterns in stained chromosomes suggested that the ancestral versions of current chimp chromosomes 12 and 13 fused end to end, forming chromosome 2 in an ancestor of the human lineage. With the recent explosion in genomic sequence information, we can now compare the chromosomal organizations of many different species on a much finer scale. This information allows us to make inferences about the evolutionary processes that shape chromosomes and may drive speciation. Sequencing and analysis of human chromosome 2 in 2005 provided very strong supporting evidence for the model we have just described **(Figure 21.12a)**.

In another study of broader scope, researchers compared the DNA sequence of each human chromosome with the whole-genome sequence of the mouse. **Figure 21.12b** shows the results of this comparison for human chromosome 16: Large blocks of genes on this chromosome are found on four mouse chromosomes, indicating that the genes in each block stayed together during the evolution of the mouse and human lineages.

Performing the same comparative analysis between chromosomes of humans and six other mammalian species allowed the researchers to reconstruct the evolutionary history of chromosomal rearrangements in these eight species. They found many duplications and inversions of large portions of chromosomes, the result of mistakes during meiotic recombination in which the DNA broke and was rejoined incorrectly. The rate of these events seems to have accelerated about 100 million years ago, around the time large dinosaurs became extinct and the number of mammalian species increased rapidly. The apparent coincidence is interesting because chromosomal rearrangements are thought to contribute to the generation of new species. Although two individuals with different arrangements could still mate and produce offspring, the offspring would have two nonequivalent sets of chromosomes, making meiosis inefficient or even impossible. Thus, chromosomal rearrangements would lead to two populations that could not successfully mate with each other, a step on the way to their becoming two separate species. (You'll learn more about this in Chapter 24.)





Duplication and Divergence of Gene-Sized Regions of DNA

Errors during meiosis can also lead to the duplication of chromosomal regions that are smaller than the ones we've just discussed, including segments the length of individual genes. Unequal crossing over during prophase I of meiosis, for instance, can result in one chromosome with a deletion and another with a duplication of a particular gene. As illustrated in **Figure 21.13**, transposable elements can provide homologous sites where nonsister chromatids can cross over, even when other chromatid regions are not correctly aligned.

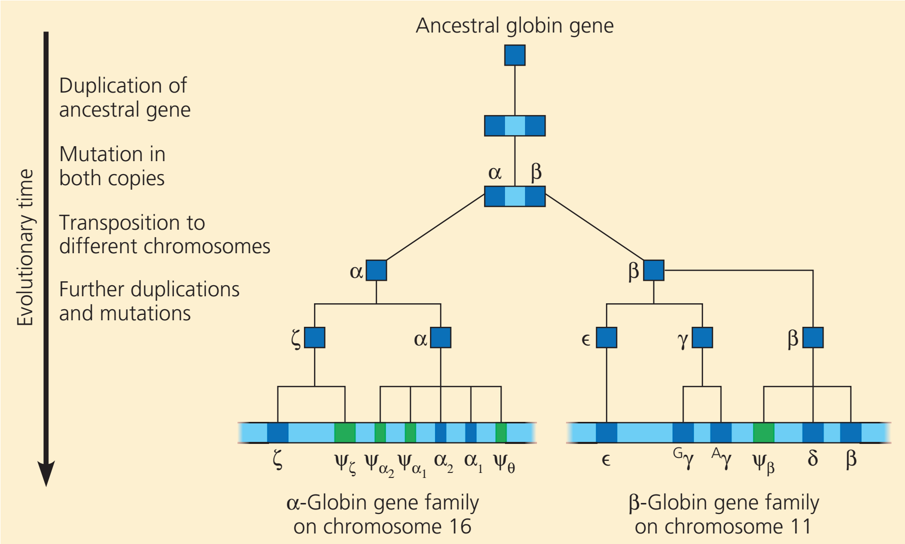
Also, slippage can occur during DNA replication, such that the template shifts with respect to the new complementary strand, and a part of the template strand is either skipped by the replication machinery or used twice as a template. As a result, a segment of DNA is deleted or duplicated. It is easy to imagine how such errors could occur in regions of repeats. The variable number of repeated units of simple sequence DNA at a given site, used for STR analysis, is probably due to errors like these. Evidence that unequal crossing over and template slippage during DNA replication lead to duplication of genes is found in the existence of multigene families, such as the globin family.

Evolution of Genes with Related Functions: The Human Globin Genes

Duplication events can lead to the evolution of genes with related functions, such as those of the α-globin and β-globin gene families (see Figure 21.11b). A comparison of gene sequences within a multigene family can suggest the order in which the genes arose. This approach to re-creating the evolutionary history of the globin genes indicates that they all evolved from one common ancestral globin gene that underwent duplication and divergence into the α-globin and β-globin ancestral genes about 450–500 million years ago **(Figure 21.14)**. Each of these genes was later duplicated several times, and the copies then diverged from each other in sequence, yielding the current family members. In fact, the common ancestral globin gene also gave rise to the oxygen-binding muscle protein myoglobin and to the plant protein leghemoglobin. The latter two proteins function as monomers, and their genes are included in a "globin superfamily."

After the duplication events, the differences between the genes in the globin families undoubtedly arose from mutations that accumulated in the gene copies over many generations. The current model is that the necessary function provided by an α-globin protein, for example, was fulfilled by one gene, while other copies of the α-globin gene accumulated random mutations. Many mutations may have had an adverse effect on the organism and others may have had no effect, but a few mutations must have altered the function of the protein product in a way that was advantageous to the organism at a particular life stage without substantially changing the protein's oxygen-carrying function. Presumably, natural selection acted on these altered genes, maintaining them in the population.

The similarity in the amino acid sequences of the various α-globin and β-globin polypeptides supports this model of gene duplication and mutation **(Table 21.2)**. The amino acid sequences of the β-globins, for instance, are much more similar to each other than to the α-globin sequences. The existence of several pseudogenes among the functional globin genes provides additional evidence for this model (see Figure 21.11b): Random mutations in these "genes" over evolutionary time have destroyed their function.



▲ **Figure 21.14 A model for the evolution of the human α-globin and β-globin gene families from a single ancestral globin gene.**

? *The green elements are pseudogenes. Explain how they could have arisen after gene duplication.*

Table 21.2 Percentage of Similarity in Amino Acid Sequence Between Human Globin Proteins

		α-Globins		β-Globins		
		α	ζ	β	γ	ε
α-Globins	α	—	58	42	39	37
	ζ	58	—	34	38	37
β-Globins	β	42	34	—	73	75
	γ	39	38	73	—	80
	ε	37	37	75	80	—

Evolution of Genes with Novel Functions

In the evolution of the globin gene families, gene duplication and subsequent divergence produced family members whose protein products performed similar functions (oxygen transport). Alternatively, one copy of a duplicated gene can undergo alterations that lead to a completely new function for the protein product. The genes for lysozyme and α-lactalbumin are good examples.

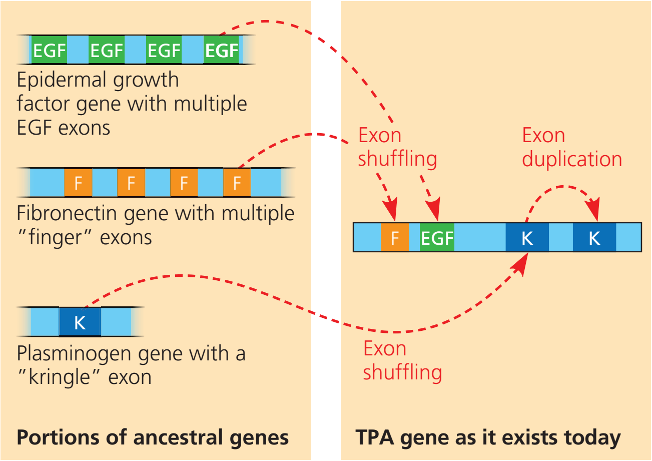
Lysozyme is an enzyme that helps protect animals against bacterial infection by hydrolyzing bacterial cell walls; α-lactalbumin is a nonenzymatic protein that plays a role in milk production in mammals. The two proteins are quite similar in their amino acid sequences and three-dimensional structures. Both genes are found in mammals, whereas only the lysozyme gene is present in birds. These findings suggest that at some time after the lineages leading to mammals and birds had separated, the lysozyme gene was duplicated in the mammalian lineage but not in the avian lineage. Subsequently, one copy of the duplicated lysozyme gene evolved into a gene encoding α-lactalbumin, a protein with a completely different function.

Rearrangements of Parts of Genes: Exon Duplication and Exon Shuffling

Rearrangement of existing DNA sequences within genes has also contributed to genome evolution. The presence of introns in most genes of multicellular eukaryotes may have promoted the evolution of new and potentially useful proteins by facilitating the duplication or repositioning of exons in the genome. Recall from Chapter 17 that an exon often codes for a domain, a distinct structural or functional region of a protein.

We've already seen that unequal crossing over during meiosis can lead to duplication of a gene on one chromosome and its loss from the homologous chromosome (see Figure 21.13). By a similar process, a particular exon within a gene could be duplicated on one chromosome and deleted from the other. The gene with the duplicated exon would code for a protein containing a second copy of the encoded domain. This change in the protein's structure could augment its function by increasing its stability, enhancing its ability to bind a particular ligand, or altering some other property. Quite a few protein-coding genes have multiple copies of related exons, which presumably arose by duplication and then diverged. The gene encoding the extracellular matrix protein collagen is a good example. Collagen is a structural protein with a highly repetitive amino acid sequence, which is reflected in the repetitive pattern of exons in the collagen gene.

Alternatively, we can imagine the occasional mixing and matching of different exons either within a gene or between two different (nonallelic) genes owing to errors in meiotic recombination. This process, termed *exon shuffling*, could lead to new proteins with novel combinations of functions. As an example, let's consider the gene for tissue plasminogen activator (TPA). The TPA protein is an extracellular protein that helps control blood clotting. It has four domains of three types, each encoded by an exon; one exon is present in two copies. Because each type of exon is also found in other proteins, the gene for TPA is thought to have arisen by several instances of exon shuffling and duplication **(Figure 21.15)**.


EGF
EGF
EGF
EGF
Epidermal growth factor gene with multiple EGF exons
F
F
F
F
Fibronectin gene with multiple "finger" exons
K
Plasminogen gene with a "kringle" exon
Portions of ancestral genes
Exon shuffling
Exon duplication
F
EGF
K
K
Exon shuffling
TPA gene as it exists today

▲ **Figure 21.15 Evolution of a new gene by exon shuffling.** Exon shuffling could have moved exons, each encoding a particular domain, from ancestral forms of the genes for epidermal growth factor, fibronectin, and plasminogen (left) into the evolving gene for tissue plasminogen activator, TPA (right). Duplication of the "kringle" exon from the plasminogen gene after its movement could account for the two copies of this exon in the TPA gene.

? *How could the presence of transposable elements in introns have facilitated the exon shuffling shown here?*

How Transposable Elements Contribute to Genome Evolution

The persistence of transposable elements as a large fraction of some eukaryotic genomes is consistent with the idea that they play an important role in shaping a genome over evolutionary time. These elements can contribute to the evolution of the genome in several ways. They can promote recombination, disrupt cellular genes or control elements, and carry entire genes or individual exons to new locations.

Transposable elements of similar sequence scattered throughout the genome facilitate recombination between different chromosomes by providing homologous regions for crossing over. Most such recombination events are probably detrimental, causing chromosomal translocations and other changes in the genome that may be lethal to the organism. But over the course of evolutionary time, an occasional recombination event of this sort may be advantageous to the organism. (For the change to be heritable, of course, it must happen in a cell that will give rise to a gamete.)

The movement of a transposable element can have a variety of consequences. For instance, if a transposable element "jumps" into the middle of a protein-coding sequence, it will prevent the production of a normal transcript of the gene. If a transposable element inserts within a regulatory sequence, the transposition may lead to increased or decreased production of one or more proteins. Transposition caused both types of effects on the genes coding for pigment-synthesizing enzymes in McClintock's corn kernels. Again, while such changes are usually harmful, in the long run some may prove beneficial by providing a survival advantage.

During transposition, a transposable element may carry along a gene or group of genes to a new position in the genome. This mechanism probably accounts for the location of the α-globin and β-globin gene families on different human chromosomes, as well as the dispersion of the genes of certain other gene families. By a similar tag-along process, an exon from one gene may be inserted into another gene in a mechanism similar to that of exon shuffling during recombination. For example, an exon may be inserted by transposition into the intron of a protein-coding gene. If the inserted exon is retained in the RNA transcript during RNA splicing, the protein that is synthesized will have an additional domain, which may confer a new function on the protein.

All the processes discussed in this section most often produce either harmful effects, which may be lethal, or no effect at all. In a few cases, however, small beneficial heritable changes may occur. Over many generations, the resulting genetic diversity provides valuable raw material for natural selection. Diversification of genes and their products is an important factor in the evolution of new species. Thus, the accumulation of changes in the genome of each species provides a record of its evolutionary history. To read this record, we must be able to identify genomic changes. Comparing the genomes of different species allows us to do that and has increased our understanding of how genomes evolve. You will learn more about these topics in the final section.

CONCEPT CHECK 21.5

1. Describe three examples of errors in cellular processes that lead to DNA duplications.
2. Explain how multiple exons might have arisen in the ancestral EGF and fibronectin genes shown in Figure 21.15 (left).
3. What are three ways that transposable elements are thought to contribute to genome evolution?
4. **WHAT IF?** In 2005, Icelandic scientists reported finding a large chromosomal inversion present in 20% of northern Europeans, and they noted that Icelandic women with this inversion had significantly more children than women without it. What would you expect to happen to the frequency of this inversion in the Icelandic population in future generations?

For suggested answers, see Appendix A.

CONCEPT 21.6

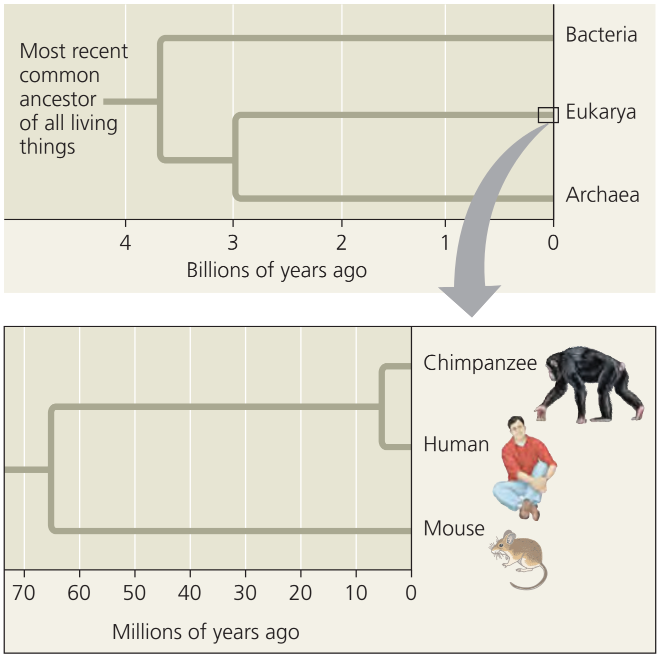
Comparing genome sequences provides clues to evolution and development

EVOLUTION One researcher has likened the current state of biology to the Age of Exploration in the 15th century after major improvements in navigation and the building of faster ships. In the last 25 years, we have seen rapid advances in genome sequencing and data collection, new techniques for assessing gene activity across the whole genome, and refined approaches for understanding how genes and their products work together in complex systems. We are truly poised on the brink of a new world.

Comparisons of genome sequences from different species reveal much about the evolutionary history of life, from very ancient to more recent. Similarly, comparative studies of the genetic programs that direct embryonic development in different species are beginning to clarify the mechanisms that generated the great diversity of life-forms present today. In this final section of the chapter, we will discuss what has been learned from these two approaches.

Comparing Genomes

The more similar in sequence the genes and genomes of two species are, the more closely related those species are in their evolutionary history. Comparing genomes of closely related species sheds light on more recent evolutionary events, whereas comparing genomes of very distantly related species helps us understand ancient evolutionary history. In either case, learning about characteristics that are shared or divergent between groups enhances our picture of the evolution of life-forms and biological processes. As you learned in Chapter 1, the evolutionary relationships between species can be represented by a diagram in the form of a tree (often turned sideways), where each branch point marks the divergence of two lineages. **Figure 21.16** shows the evolutionary relationships of some groups and species we will be discussing. We will consider comparisons between distantly related species first.



▲ **Figure 21.16 Evolutionary relationships of the three domains of life.** This tree diagram shows the ancient divergence of bacteria, archaea, and eukaryotes. A portion of the eukaryote lineage is expanded in the inset to show the more recent divergence of three mammalian species discussed in this chapter.

Comparing Distantly Related Species

Determining which genes have remained similar—that is, are *highly conserved*—in distantly related species can help clarify evolutionary relationships among species that diverged from each other long ago. Indeed, comparisons of the complete genome sequences of bacteria, archaea, and eukaryotes indicate that these three groups diverged between 2 and 4 billion years ago and strongly support the theory that they are the fundamental domains of life (see Figure 21.16).

In addition to their value in evolutionary biology, comparative genomic studies confirm the relevance of research on model organisms to our understanding of biology in general and human biology in particular. Genes that evolved a very long time ago can still be surprisingly similar in disparate species. As a case in point, several genes in yeast are so similar to certain human disease genes that researchers have deduced the functions of the disease genes by studying their yeast counterparts. This striking similarity underscores the common origin of these two distantly related species.

Comparing Closely Related Species

The genomes of two closely related species are likely to be organized similarly because of their relatively recent divergence. As we mentioned earlier, this allows the fully sequenced genome of one species to be used as a scaffold for assembling the genomic sequences of a closely related species, accelerating mapping of the second genome. For instance, using the human genome sequence as a guide, researchers were able to quickly sequence the chimpanzee genome.

The recent divergence of two closely related species also underlies the small number of gene differences that are found when their genomes are compared. The particular genetic differences can therefore be more easily correlated with phenotypic differences between the two species. An exciting application of this type of analysis is seen as researchers compare the human genome with the genomes of the chimpanzee, mouse, rat, and other mammals. Identifying the genes shared by all of these species but not by nonmammals should give clues about what it takes to make a mammal, while finding the genes shared by chimpanzees and humans but not by rodents should tell us something about primates. And, of course, comparing the human genome with that of the chimpanzee should help us answer the tantalizing question we asked at the beginning of the chapter: What genomic information makes a human or a chimpanzee?

An analysis of the overall composition of the human and chimpanzee genomes, which are thought to have diverged only about 6 million years ago (see Figure 21.16), reveals some general differences. Considering single nucleotide substitutions, the two genomes differ by only 1.2%. When researchers looked at longer stretches of DNA, however, they were surprised to find a further 2.7% difference due to insertions or deletions of larger regions in the genome of one or the other species; many of the insertions were duplications or other repetitive DNA. In fact, a third of the human duplications are not present in the chimpanzee genome, and some of these duplications contain regions associated with human diseases. There are more *Alu* elements in the human genome than in the chimpanzee genome, and the latter contains many copies of a retroviral provirus not present in humans. All of these observations provide clues to the forces that might have swept the two genomes along different paths, but we don't have a complete picture yet. We also don't know how these differences might account for the distinct characteristics of each species.

To discover the basis for the phenotypic differences between the two species, biologists are studying specific genes and types of genes that differ between humans and chimpanzees and comparing them with their counterparts in other mammals. This approach has revealed a number of genes that are apparently changing (evolving) faster in the human than in either the chimpanzee or the mouse. Among them are genes involved in defense against malaria and tuberculosis and at least one gene that regulates brain size. When genes are classified by function, the genes that seem to be evolving the fastest are those that code for transcription factors. This discovery makes sense because transcription factors regulate gene expression and thus play a key role in orchestrating the overall genetic program.

One transcription factor whose gene shows evidence of rapid change in the human lineage is called FOXP2. Several lines of evidence suggest that the *FOXP2* gene functions in vocalization in vertebrates. For one thing, mutations in this gene can produce severe speech and language impairment in humans. Moreover, the *FOXP2* gene is expressed in the brains of zebra finches and canaries at the time when these songbirds are learning their songs. But perhaps the strongest evidence comes from a "knock-out" experiment in which researchers disrupted the *FOXP2* gene in mice and analyzed the resulting phenotype (**Figure 21.17**, on the next page). The homozygous mutant mice had malformed brains and failed to emit normal ultrasonic vocalizations, and mice with one faulty copy of the gene also showed significant problems with vocalization. These results support the idea that the *FOXP2* gene product turns on genes involved in vocalization.

Expanding on this analysis, another research group more recently replaced the *FOXP2* gene in mice with a "humanized" copy coding for the human versions of two amino acids that differ between human and chimp; these are the changes potentially responsible for a human's ability to speak. Although the mice were generally healthy, they had subtly different vocalizations and showed changes in brain cells in circuits associated with speech in human brains.

The *FOXP2* story is an excellent example of how different approaches can complement each other in uncovering biological phenomena of widespread importance. The

▼ Figure 21.17 INQUIRY

What is the function of a gene (*FOXP2*) that is rapidly evolving in the human lineage?

EXPERIMENT Several lines of evidence support a role for the *FOXP2* gene in the development of speech and language in humans and of vocalization in other vertebrates. In 2005, Joseph Buxbaum and collaborators at the Mount Sinai School of Medicine and several other institutions tested the function of *FOXP2*. They used the mouse, a model organism in which genes can be easily knocked out, as a representative vertebrate that vocalizes: Mice produce ultrasonic squeaks (whistles) to communicate stress. The researchers used genetic engineering to produce mice in which one or both copies of *FOXP2* were disrupted.

Wild type: two normal copies of *FOXP2*

Heterozygote: one copy of *FOXP2* disrupted

Homozygote: both copies of *FOXP2* disrupted

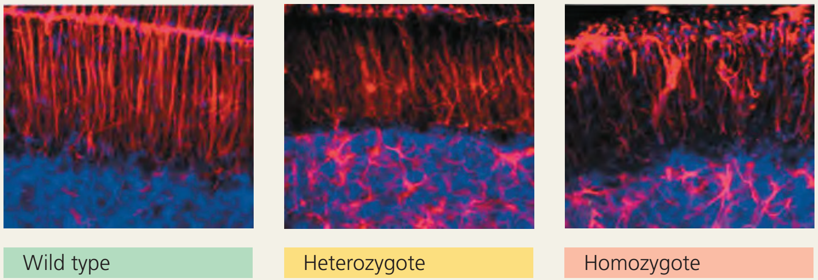
They then compared the phenotypes of these mice. Two of the characters they examined are included here: brain anatomy and vocalization.

Experiment 1: Researchers cut thin sections of brain and stained them with reagents that allow visualization of brain anatomy in a UV fluorescence microscope.

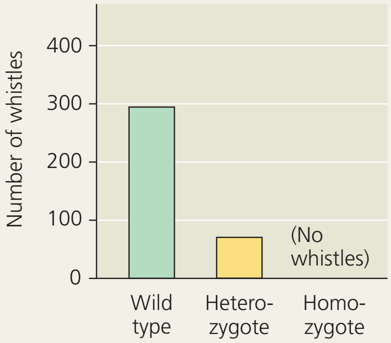
Experiment 2: Researchers separated each newborn pup from its mother and recorded the number of ultrasonic whistles produced by the pup.

RESULTS

Experiment 1: Disruption of both copies of *FOXP2* led to brain abnormalities in which the cells were disorganized. Phenotypic effects on the brain of heterozygotes, with one disrupted copy, were less severe. (Each color reveals a different cell or tissue type.)



Experiment 2: Disruption of both copies of *FOXP2* led to an absence of ultrasonic vocalization in response to stress. The effect on vocalization in the heterozygote was also extreme.



CONCLUSION *FOXP2* plays a significant role in the development of functional communication systems in mice. The results augment evidence from studies of birds and humans, supporting the hypothesis that *FOXP2* may act similarly in diverse organisms.

SOURCE W. Shu et al., Altered ultrasonic vocalization in mice with a disruption in the *Foxp2* gene, *Proceedings of the National Academy of Sciences* 102:9643–9648 (2005).

WHAT IF? Since the results support a role for mouse *FOXP2* in vocalization, you might wonder whether the human FOXP2 protein is a key regulator of speech. If you were given the amino acid sequences of wild-type and mutant human FOXP2 proteins and the wild-type chimpanzee FOXP2 protein, how would you investigate this question? What further clues could you obtain by comparing these sequences to that of the mouse FOXP2 protein?

FOXP2 experiments used mice as a model for humans because it would be unethical (as well as impractical) to carry out such experiments in humans. Mice and humans diverged about 65.5 million years ago (see Figure 21.16) and share about 85% of their genes. This genetic similarity can be exploited in studying human genetic disorders. If researchers know the organ or tissue that is affected by a particular genetic disorder, they can look for genes that are expressed in these locations in mice.

Further research efforts are under way to extend genomic studies to many more microbial species, additional primates, and neglected species from diverse branches of the tree of life.

These studies will advance our understanding of all aspects of biology, including health and ecology as well as evolution.

Comparing Genomes Within a Species

Another exciting consequence of our ability to analyze genomes is our growing understanding of the spectrum of genetic variation in humans. Because the history of the human species is so short—probably about 200,000 years—the amount of DNA variation among humans is small compared to that of many other species. Much of our diversity seems to be in the form of single nucleotide polymorphisms (SNPs, described in Chapter 20), usually detected by DNA sequencing. In the human genome, SNPs occur on average about once in 100–300 base pairs. Scientists have already identified the location of several million SNP sites in the human genome and continue to find more.

In the course of this search, they have also found other variations—including inversions, deletions, and duplications. The most surprising discovery has been the widespread occurrence of *copy-number variants* (*CNVs*), loci where some individuals have one or multiple copies of a particular gene or genetic region, rather than the standard two copies (one on each homolog). CNVs result from regions of the genome being duplicated or deleted inconsistently within the population. A 2010 study of 40 people found more than 8,000 CNVs involving 13% of the genes in the genome, and these CNVs probably represent just a small subset of the total. Since these variants encompass much longer stretches of DNA than the single nucleotides of SNPs, CNVs are more likely to have phenotypic consequences and to play a role in complex diseases and disorders. At the very least, the high incidence of copy-number variation casts doubt on the meaning of the phrase "a normal human genome."

Copy-number variants, SNPs, and variations in repetitive DNA such as short tandem repeats (STRs) will be useful genetic markers for studying human evolution. In 2010, the genomes of two Africans from different communities were sequenced: Archbishop Desmond Tutu, the South African civil rights advocate and a member of the Bantu tribe, the majority population in southern Africa; and !Gubi, a hunter-gatherer from the Khoisan community in Namibia, a minority African population that is probably the human group with the oldest known lineage. The comparison revealed many differences, as you might expect. The analysis was then broadened to compare the protein-coding regions of !Gubi's genome with those of three other Khoisan community members (self-identified Bushmen) living nearby. Remarkably, these four genomes differed more from each other than a European would from an Asian. These data highlight the extensive diversity among African genomes. Extending this approach will help us answer important questions about the differences between human populations and the migratory routes of human populations throughout history.

Comparing Developmental Processes

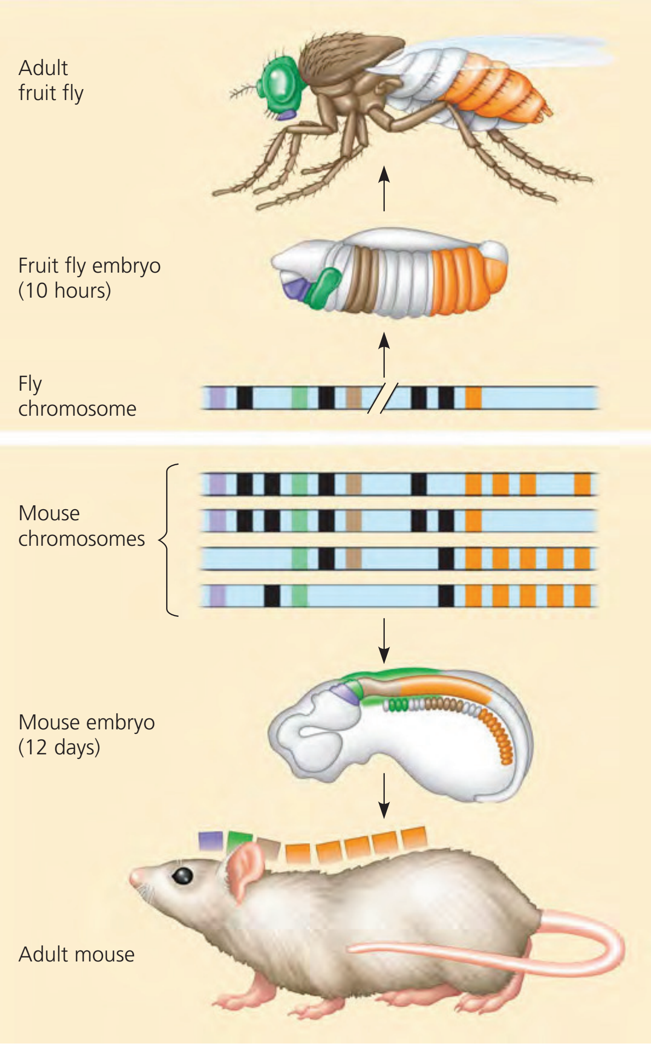
Biologists in the field of evolutionary developmental biology, or **evo-devo** as it is often called, compare developmental processes of different multicellular organisms. Their aim is to understand how these processes have evolved and how changes in them can modify existing organismal features or lead to new ones. With the advent of molecular techniques and the recent flood of genomic information, we are beginning to realize that the genomes of related species with strikingly different forms may have only minor differences in gene sequence or regulation. Discovering the molecular basis of these differences in turn helps us understand the origins of the myriad diverse forms that cohabit this planet, thus informing our study of evolution.

Widespread Conservation of Developmental Genes Among Animals

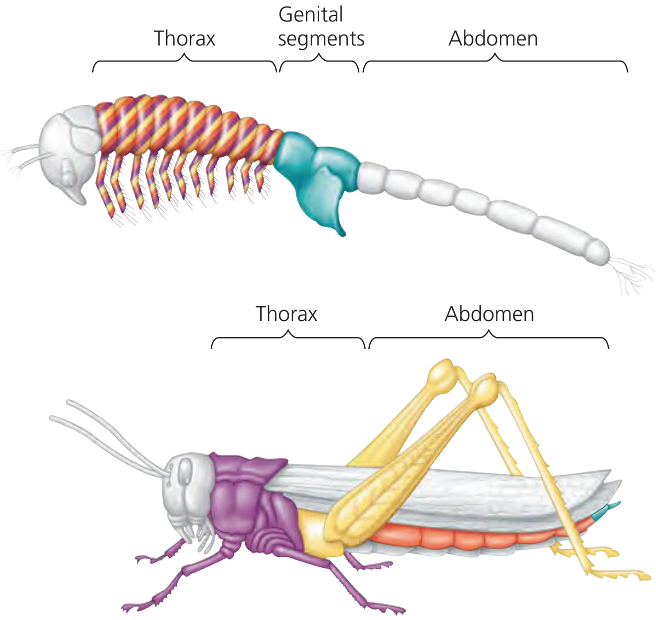
In Chapter 18, you learned about the homeotic genes in *Drosophila*, which specify the identity of body segments in the fruit fly (see Figure 18.20). Molecular analysis of the homeotic genes in *Drosophila* has shown that they all include a 180-nucleotide sequence called a **homeobox**, which specifies a 60-amino-acid *homeodomain* in the encoded proteins. An identical or very similar nucleotide sequence has been discovered in the homeotic genes of many invertebrates and vertebrates. The sequences are so similar between humans and fruit flies, in fact, that one researcher has whimsically referred to flies as "little people with wings." The resemblance even extends to the organization of these genes: The vertebrate genes homologous to the homeotic genes of fruit flies have kept the same chromosomal arrangement (**Figure 21.18**, on the next page). Homeobox-containing sequences have also been found in regulatory genes of much more distantly related eukaryotes, including plants and yeasts. From these similarities, we can deduce that the homeobox DNA sequence evolved very early in the history of life and was sufficiently valuable to organisms to have been conserved in animals and plants virtually unchanged for hundreds of millions of years.

Homeotic genes in animals were named *Hox* genes, short for *h*omeob*ox*-containing genes, because homeotic genes were the first genes found to have this sequence. Other homeobox-containing genes were later found that do not act as homeotic genes; that is, they do not directly control the identity of body parts. However, most of these genes, in animals at least, are associated with development, suggesting their ancient and fundamental importance in that process. In *Drosophila*, for example, homeoboxes are present not only in the homeotic genes but also in the egg-polarity gene *bicoid* (see Figures 18.21 and 18.22), in several of the segmentation genes, and in a master regulatory gene for eye development.

Researchers have discovered that the homeobox-encoded homeodomain is the part of a protein that binds to DNA



▲ **Figure 21.18 Conservation of homeotic genes in a fruit fly and a mouse.** Homeotic genes that control the form of anterior and posterior structures of the body occur in the same linear sequence on chromosomes in *Drosophila* and mice. Each colored band on the chromosomes shown here represents a homeotic gene. In fruit flies, all homeotic genes are found on one chromosome. The mouse and other mammals have the same or similar sets of genes on four chromosomes. The color code indicates the parts of the embryos in which these genes are expressed and the adult body regions that result. All of these genes are essentially identical in flies and mice, except for those represented by black bands, which are less similar in the two animals.



▲ **Figure 21.19 Effect of differences in *Hox* gene expression in crustaceans and insects.** Changes in the expression patterns of *Hox* genes have occurred over evolutionary time. These changes account in part for the different body plans of the brine shrimp *Artemia*, a crustacean (top), and the grasshopper, an insect. Shown here are regions of the adult body color-coded for expression of four *Hox* genes that determine formation of particular body parts during embryonic development. Each color represents a specific *Hox* gene. Colored stripes on the thorax of *Artemia* indicate co-expression of three *Hox* genes.

when the protein functions as a transcriptional regulator. However, the shape of the homeodomain allows it to bind to any DNA segment; its own structure is not specific for a particular sequence. Instead, other, more variable domains in a homeodomain-containing protein determine which genes the protein regulates. Interaction of these variable domains with still other transcription factors helps a homeodomain-containing protein recognize specific enhancers in the DNA. Proteins with homeodomains probably regulate development by coordinating the transcription of batteries of developmental genes, switching them on or off. In embryos of *Drosophila* and other animal species, different combinations of homeobox genes are active in different parts of the embryo. This selective expression of regulatory genes, varying over time and space, is central to pattern formation.

Developmental biologists have found that in addition to homeotic genes, many other genes involved in development are highly conserved from species to species. These include numerous genes encoding components of signaling pathways. The extraordinary similarity among particular developmental genes in different animal species raises a question: How can the same genes be involved in the development of animals whose forms are so very different from each other?

Ongoing studies are suggesting answers to this question. In some cases, small changes in regulatory sequences of particular genes cause changes in gene expression patterns that can lead to major changes in body form. For example, the differing patterns of expression of the *Hox* genes along the body axis in insects and crustaceans can explain the variation in number of leg-bearing segments among these segmented animals **(Figure 21.19)**. Also, recent research suggests that the same *Hox* gene product may have subtly dissimilar effects in different species, turning on new genes or turning on the same genes at higher or lower levels. In other cases, similar genes direct differing developmental processes in different organisms, resulting in diverse body shapes. Several *Hox* genes, for instance, are expressed in the embryonic and larval stages of the sea urchin, a nonsegmented animal that has a body plan quite different from those of insects and mice. Sea urchin adults make the pincushion-shaped shells you may have seen on the beach (see Figure 8.4). They are among the organisms long used in classical embryological studies (see Chapter 47).

Comparison of Animal and Plant Development

The last common ancestor of animals and plants was probably a single-celled eukaryote that lived hundreds of millions of years ago, so the processes of development must have evolved independently in the two multicellular lineages of organisms. Plants evolved with rigid cell walls, which rule out the morphogenetic movements of cells and tissues that are so important in animals. Instead, morphogenesis in plants relies primarily on differing planes of cell division and on selective cell enlargement. (You will learn about these processes in Chapter 35.) But despite the differences between animals and plants, there are similarities in the molecular mechanisms of development, which are legacies of their shared unicellular origin.

In both animals and plants, development relies on a cascade of transcriptional regulators turning on or turning off genes in a finely tuned series. For example, work on the small flowering plant *Arabidopsis thaliana* has shown that establishing the radial pattern of flower parts, like setting up the head-to-tail axis in *Drosophila*, involves a cascade of transcription factors (see Chapter 35). The genes that direct these processes, however, differ considerably in animals and plants. While quite a few of the master regulatory switches in *Drosophila* are homeobox-containing *Hox* genes, those in *Arabidopsis* belong to a completely different family of genes, called the *MADS-box* genes. And although homeobox-containing genes can be found in plants and *MADS-box* genes in animals, in neither case do they perform the same major roles in development that they do in the other group. Thus, molecular evidence supports the supposition that developmental programs evolved separately in animals and plants.

In this final chapter of the genetics unit, you have learned how studying genomic composition and comparing the genomes of different species can disclose much about how genomes evolve. Further, comparing developmental programs, we can see that the unity of life is reflected in the similarity of molecular and cellular mechanisms used to establish body pattern, although the genes directing development may differ among organisms. The similarities between genomes reflect the common ancestry of life on Earth. But the differences are also crucial, for they have created the huge diversity of organisms that have evolved. In the remainder of the book, we expand our perspective beyond the level of molecules, cells, and genes to explore this diversity on the organismal level.

CONCEPT CHECK 21.6

1. Would you expect the genome of the macaque (a monkey) to be more similar to the mouse genome or the human genome? Why?
2. The DNA sequences called homeoboxes, which help homeotic genes in animals direct development, are common to flies and mice. Given this similarity, explain why these animals are so different.
3. WHAT IF? There are three times as many *Alu* elements in the human genome as in the chimpanzee genome. How do you think these extra *Alu* elements arose in the human genome? Propose a role they might have played in the divergence of these two species.

For suggested answers, see Appendix A.

21 CHAPTER REVIEW

SUMMARY OF KEY CONCEPTS

CONCEPT 21.1

New approaches have accelerated the pace of genome sequencing (pp. 427–429)

- The **Human Genome Project** began in 1990, using a three-stage approach. In **linkage mapping**, the order of genes and other inherited markers in the genome and the relative distances between them can be determined from recombination frequencies. Next, **physical mapping** uses overlaps between DNA fragments to order the fragments and determine the distance in base pairs between markers. Finally, the ordered fragments are sequenced, providing the finished genome sequence.
- In the whole-genome shotgun approach, the whole genome is cut into many small, overlapping fragments that are sequenced; computer software then assembles the complete sequence. Correct assembly is made easier when mapping information is also available.

? *Why has the whole-genome shotgun approach been widely adopted for genome-sequencing projects?*

CONCEPT 21.2

Scientists use bioinformatics to analyze genomes and their functions (pp. 429–432)

- Websites on the Internet provide centralized access to genome sequence databases, analytical tools, and genome-related information.
- Computer analysis of genome sequences aids **gene annotation**, the identification of protein-coding sequences and determination of their function. Methods for determining gene function include comparing the sequences of newly discovered genes with those of known genes in other species and observing the phenotypic effects of experimentally inactivating genes of unknown function.
- In systems biology, scientists use the computer-based tools of **bioinformatics** to compare genomes and study sets of genes and proteins as whole systems (**genomics** and **proteomics**). Studies

include large-scale analyses of protein interactions, functional DNA elements, and genes contributing to medical conditions.

? *What was the most significant finding of the ENCODE pilot project? Why has the project been expanded to include other species?*

CONCEPT 21.3

Genomes vary in size, number of genes, and gene density (pp. 432–434)

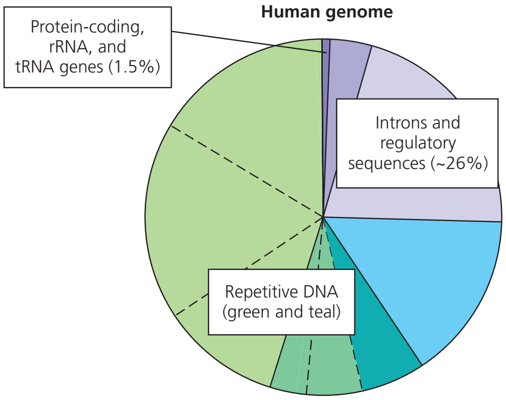
	Bacteria	Archaea	Eukarya
Genome size	Most are 1–6 Mb		Most are 10–4,000 Mb, but a few are much larger
Number of genes	1,500–7,500		5,000–40,000
Gene density	Higher than in eukaryotes		Lower than in prokaryotes (Within eukaryotes, lower density is correlated with larger genomes.)
Introns	None in protein-coding genes	Present in some genes	Unicellular eukaryotes: present, but prevalent only in some species Multicellular eukaryotes: present in most genes
Other noncoding DNA	Very little		Can be large amounts; generally more repetitive noncoding DNA in multicellular eukaryotes

? *Compare genome size, gene number, and gene density (a) in the three domains and (b) among eukaryotes.*

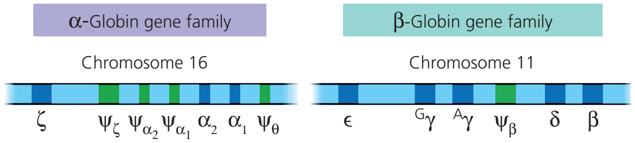
CONCEPT 21.4

Multicellular eukaryotes have much noncoding DNA and many multigene families (pp. 434–438)

- Only 1.5% of the human genome codes for proteins or gives rise to rRNAs or tRNAs; the rest is noncoding DNA, including **pseudogenes** and **repetitive DNA** of unknown function.



- The most abundant type of repetitive DNA in multicellular eukaryotes consists of **transposable elements** and related sequences. In eukaryotes, there are two types of transposable elements: **transposons**, which move via a DNA intermediate, and **retrotransposons**, which are more prevalent and move via an RNA intermediate.
- Other repetitive DNA includes short noncoding sequences that are tandemly repeated thousands of times (**simple sequence DNA**, which includes **STRs**); these sequences are especially prominent in centromeres and telomeres, where they probably play structural roles in the chromosome.
- Though many eukaryotic genes are present in one copy per haploid chromosome set, others (most, in some species) are members of a family of related genes, such as the human globin gene families:



? *Explain how the function of transposable elements might account for their prevalence in human noncoding DNA.*

CONCEPT 21.5

Duplication, rearrangement, and mutation of DNA contribute to genome evolution (pp. 438–442)

- Accidents in cell division can lead to extra copies of all or part of entire chromosome sets, which may then diverge if one set accumulates sequence changes.
- The chromosomal organization of genomes can be compared among species, providing information about evolutionary relationships. Within a given species, rearrangements of chromosomes are thought to contribute to the emergence of new species.
- The genes encoding the various globin proteins evolved from one common ancestral globin gene, which duplicated and diverged into α-globin and β-globin ancestral genes. Subsequent duplication and random mutation gave rise to the present globin genes, all of which code for oxygen-binding proteins. The copies of some duplicated genes have diverged so much that the functions of their encoded proteins (such as lysozyme and α-lactalbumin) are now substantially different.
- Rearrangement of exons within and between genes during evolution has led to genes containing multiple copies of similar exons and/or several different exons derived from other genes.
- Movement of transposable elements or recombination between copies of the same element occasionally generates new sequence combinations that are beneficial to the organism. Such mechanisms can alter the functions of genes or their patterns of expression and regulation.

? *How could chromosomal rearrangements lead to the emergence of new species?*

CONCEPT 21.6

Comparing genome sequences provides clues to evolution and development (pp. 442–447)

- Comparative studies of genomes from widely divergent and closely related species provide valuable information about ancient and more recent evolutionary history, respectively. Human and chimpanzee sequences show about 4% difference, mostly due to insertions, deletions, and duplications in one lineage. Along with nucleotide variations in specific genes (such as *FOXP2*, a gene affecting speech), these differences may account for the distinct characteristics of the two species. Analysis of single nucleotide polymorphisms (SNPs) and copy-number variants (CNVs) among individuals in a species can also yield information about the evolution of that species.

- Evolutionary developmental (**evo-devo**) biologists have shown that homeotic genes and some other genes associated with animal development contain a **homeobox** region whose sequence is highly conserved among diverse species. Related sequences are present in the genes of plants and yeasts. During embryonic development in both plants and animals, a cascade of transcription regulators turns genes on or off in a carefully regulated sequence. However, the genes that direct analogous developmental processes differ in plants and animals as a result of their remote ancestry.

? *What type of information can be obtained by comparing the genomes of closely related species? Of very distantly related species?*

TEST YOUR UNDERSTANDING

LEVEL 1: KNOWLEDGE/COMPREHENSION

1. Bioinformatics includes all of the following *except*
a. using computer programs to align DNA sequences.
b. analyzing protein interactions in a species.
c. using molecular biology to combine DNA from two different sources in a test tube.
d. developing computer-based tools for genome analysis.
e. using mathematical tools to make sense of biological systems.

2. One of the characteristics of retrotransposons is that
a. they code for an enzyme that synthesizes DNA using an RNA template.
b. they are found only in animal cells.
c. they generally move by a cut-and-paste mechanism.
d. they contribute a significant portion of the genetic variability seen within a population of gametes.
e. their amplification is dependent on a retrovirus.

3. Homeotic genes
a. encode transcription factors that control the expression of genes responsible for specific anatomical structures.
b. are found only in *Drosophila* and other arthropods.
c. are the only genes that contain the homeobox domain.
d. encode proteins that form anatomical structures in the fly.
e. are responsible for patterning during plant development.

LEVEL 2: APPLICATION/ANALYSIS

4. Two eukaryotic proteins have one domain in common but are otherwise very different. Which of the following processes is most likely to have contributed to this similarity?
a. gene duplication
b. RNA splicing
c. exon shuffling
d. histone modification
e. random point mutations

5. DRAW IT Below are the amino acid sequences (using the single-letter code; see Figure 5.16) of four short segments of the FOXP2 protein from six species: chimpanzee, orangutan, gorilla, rhesus macaque, mouse, and human. These segments contain all of the amino acid differences between the FOXP2 proteins of these species.

1. ATETI...PKSSD...TSSTT...NARRD
2. ATETI...PKSSE...TSSTT...NARRD
3. ATETI...PKSSD...TSSTT...NARRD
4. ATETI...PKSSD...TSSNT...S ARRD
5. ATETI...PKSSD...TSSTT...NARRD
6. VTETI...PKSSD...TSSTT...NARRD

Use a highlighter to color any amino acid that varies among the species. (Color that amino acid in all sequences.) Then answer the questions at the top of the next column.

(a) The chimpanzee, gorilla, and rhesus macaque (C, G, R) sequences are identical. Which lines correspond to those sequences?
(b) The human sequence differs from that of the C, G, R species at two amino acids. Which line corresponds to the human sequence? Underline the two differences.
(c) The orangutan sequence differs from the C, G, R sequence at one amino acid (having valine instead of alanine) and from the human sequence at three amino acids. Which line corresponds to the orangutan sequence?
(d) How many amino acid differences are there between the mouse and the C, G, R species? Circle the amino acid(s) that differ(s) in the mouse. How many amino acid differences are there between the mouse and the human? Draw a square around the amino acid(s) that differ(s) in the mouse.
(e) Primates and rodents diverged between 60 and 100 million years ago, and chimpanzees and humans diverged about 6 million years ago. Knowing that, what can you conclude by comparing the amino acid differences between the mouse and the C, G, R species with the differences between the human and the C, G, R species?

LEVEL 3: SYNTHESIS/EVALUATION

6. **EVOLUTION CONNECTION**
Genes important in the embryonic development of animals, such as homeobox-containing genes, have been relatively well conserved during evolution; that is, they are more similar among different species than are many other genes. Why is this?

7. **SCIENTIFIC INQUIRY**
The scientists mapping the SNPs in the human genome noticed that groups of SNPs tended to be inherited together, in blocks known as haplotypes, ranging in length from 5,000 to 200,000 base pairs. There are as few as four or five commonly occurring combinations of SNPs per haplotype. Propose an explanation for this observation, integrating what you've learned throughout this chapter and this unit.

8. **WRITE ABOUT A THEME**
The Genetic Basis of Life The continuity of life is based on heritable information in the form of DNA. In a short essay (100–150 words), explain how mutations in protein-coding genes and regulatory DNA contribute to evolution.

For selected answers, see Appendix A.

MasteringBIOLOGY www.masteringbiology.com

1. MasteringBiology® Assignments

Tutorial Shotgun Approach to Whole-Genome Sequencing • Using BLAST: Can You Identify a Pathogen from a Nucleotide Sequence?
Activity The Human Genome Project: Genes on Human Chromosome 17
Questions Student Misconceptions • Reading Quiz • Multiple Choice • End-of-Chapter

2. eText

Read your book online, search, take notes, highlight text, and more.

3. The Study Area

Practice Tests • Cumulative Test • BioFlix 3-D Animations • MP3 Tutor Sessions • Videos • Activities • Investigations • Lab Media • Audio Glossary • Word Study Tools • Art

UNIT 4

Mechanisms of Evolution

An Interview with Geerat J. Vermeij

Born in the Netherlands, Geerat Vermeij (pronounced "ver-may") lost his sight at the age of 3. Undeterred, he went on to earn degrees from Princeton and Yale and is now a Distinguished Professor at the University of California, Davis. A member of the Department of Geology, he nevertheless focuses on biology—the structure, evolution, and ecology of marine molluscs, both living and extinct. He is particularly well known for his work on the evolutionary "arms race" between long-extinct molluscs and their predators and more generally the roles of organismal interactions in evolution, although his many publications reflect much wider-ranging interests. (One of his books, *Nature: An Economic History*, relates the principles of evolution to the principles of economics; he has also written a memoir, *Privileged Hands: A Scientific Life*.) Dr. Vermeij has received numerous awards, including the MacArthur Award and the Daniel Giraud Elliot Medal from the National Academy of Sciences. His office at UC Davis features a large collection of marine shells and fossils and an extensive library. Jane Reece and Michael Cain spoke with him there.

How did you first become interested in biology?

As far back as I can recall, back to my earliest childhood, I've always liked natural history. When I was a child in the Netherlands, I liked pinecones and seeds and shells on the beach and leaves. I liked the whole ambience of being outside! Also, my parents were very good observers, and they spent a lot of time describing the world to me and letting me touch as much as I possibly could. When I moved to the U.S. at the age of 9, I found myself in a completely different environment. In New Jersey where we lived, there were wild forests full of huge vines, noisy crickets, cicadas, and strange birds, and I found this environment so different from the one I had left behind that I began to ask myself why this was.

When I was in the fourth grade, I had a wonderful teacher who brought shells from Florida to her classroom. And I explored these things and fell in love with them. And again, I wondered why these things were so different from anything I had collected in Holland. They were beautiful, with lovely shapes and wonderful contrast between the outside surface and the inside. I was smitten. And from then on I knew I was going to do something scientific.

Much of your work focuses on marine molluscs. Please tell us about these animals.

Molluscs include snails, clams, squids, octopuses, and many lesser-known groups. There are something like 100,000 species living, and we know of fossils of tens of thousands of extinct ones, dating all the way back to at least 540 million years ago. Molluscs are a major animal group on the tree of life. Found on land and in fresh water and the sea, they do just about everything you can imagine—they range from top predators (such as squids) to suspension feeders, herbivores, detritus feeders, and parasites. Originally, all molluscs had some kind of mantle covering the major organs of the body. They probably started off as pretty simple creatures without shells, but shells soon evolved. Most living molluscs have shells, though some have lost their shell in the course of evolution.

How do you identify the shells you are studying?

Entirely by touch. You know, shells differ in size, shape, and texture, all of which are readily discernable by the fingers, and the same is true of fossils. Shells are ideally suited for a blind person like me.

In your research over the years, what are the main questions you have been trying to answer?

The questions have changed over the course of my career, but the overarching ones have been, What are the pathways of adaptation by which all the different lineages of organisms—not just molluscs, but all of life—have gotten here? How have the conditions to which organisms are exposed changed over time? How have organisms affected those conditions over time? I'm very interested in the history of life and how this history has been shaped.

What makes molluscs a good research focus for answering your questions?

Molluscs have several huge advantages. For me personally, of course, they are accessible. Most of them don't run away—squids and octopuses being exceptions. Their shells are extremely easy to observe with the hand, and importantly from a paleontologist's point of view, these hard mineralized objects have left a very good fossil record. That's a gigantic advantage. Not only can we trace molluscs all the way back through time, but because we know so much about how shells work, we can figure out how the extinct animals lived—even those that lived hundreds of millions of years ago.

What kinds of evolutionary insights can fossils provide that cannot be extracted from DNA evidence?

First of all, I should say that my collaborators and I do use DNA sequences ourselves to reconstruct the order of branching in evolutionary trees. But to estimate *when* these evolutionary lineages arose, we need to calibrate the tree with fossils of known age. Moreover, you can only get DNA from living things and a few rather recent fossils; so if you go back far enough, you find many lineages that no longer exist and for which, therefore, DNA evidence is simply unavailable. And yet these animals often have combinations of traits that we never see in living animals. Fossils give us a very good idea of what the ancestral organisms were like, which you couldn't get solely from DNA sequences of living organisms. So if you're trying to reconstruct early branches in the tree of life, fossils are very helpful.

How does your research relate to the mechanisms of evolution—to the principles as opposed to the pattern?

That's an important question. I do distinguish between describing what actually happened and the mechanisms that account for evolutionary events over time. A lot of our work is descriptive, figuring out what happened and what extinct animals were like. But we also try to determine the mechanisms that account for the phenomena. And given that I work on adaptive characteristics and on the fit between animals and plants and their environments, I am particularly interested in the mechanisms by which organisms become adapted to their environments. That's not simply natural selection; it's also the modification of environments by the animals and plants that reside there.

Tell us how you go about your work.

I have done a lot of fieldwork all over the world. In the field I observe molluscs in nature and occasionally do some experiments with them. I want to understand how the molluscs relate to their environments, including their predators. I am interested in how they live—for example, how quickly they move—and how their performance levels compare with the performance of the agents that are out there making the world tough for them.

Recently I've spent more time in museum collections. I also maintain a very large research collection, most of which I've collected myself over the years. All of these collections are critical for learning what the shapes of organisms are in different evolutionary groups. I also visit and learn from other scientists. And I do an enormous amount of reading, because I like to synthesize information and ideas, to put things together. I read hundreds of papers a year about a very wide variety of subjects, everything from biology to geology to economics and history, so that I can place the particular work that I do into a larger context. As a scientist, you can never read enough.

When you do this kind of work, whether in collections of specimens or in the field or library, you always come across wonderful surprises—perhaps a shell with a feature you've never seen or even a book you didn't know about. Every single day for me is like that.

You have written about the "arms race" of evolution. What do you mean by that, and how has it played out in the creatures you've studied?

All living things are exposed to competition for resources and also to predation, where one animal eats another animal or part of another animal. The animals I am working on mostly don't move very fast, and one of the typical results of predation is that armor in the form of shells evolves in the prey; the mollusc shell probably first evolved as armor. But as predators become more powerful thanks to competition among themselves, the performance criteria for an effective shell also escalate. Nowadays, in order to survive in tropical shallow water environments where there are lots of predators, a mollusc needs a very well-armored shell—one that has thick walls, bumps all over the place, a narrow opening, and many other features. In fact, if you look at shell architecture over geological time, keeping habitat as constant as you can, you find that some of these protective features (the narrow opening, for example) are found only in the more recent evolutionary lineages and don't appear at all in the first couple of hundred million years of mollusc history. Meanwhile, all sorts of ways of overcoming mollusc defenses have evolved in predators. They have developed stronger, more powerful jaws or claws. They have "learned" how to drill a hole through the wall of a shell. They swallow or envelop larger prey. That tells us that there has been an arms race, an escalation of improvements in both shell architecture and methods of attack by predators.

"When you do this kind of work . . . you always come across wonderful surprises. . . . Every single day for me is like that."

In addition to arms races, what other kinds of ecological interactions have influenced evolutionary history?

Competition and predation are fundamental and inevitable, but the history of life from the very beginning is also a story of cooperation. The reason for that I think is simple: By cooperating, you can do things that neither party can do by itself. So, cooperation or some other kind of mutually beneficial relationship is a wonderful way to compete. Biology is absolutely filled with examples of social animals, mutually beneficial relationships between individuals of different species, and so forth. Cooperation is in some sense an emergent property of life as a whole. The interactions of organisms with one another give rise to properties that the individual components don't have. For example, a lichen, which is an alga and a fungus living together, has properties different from those of either participant.

How do the things you've been saying about the effects of ecological interactions on evolutionary history fit in with Darwin's main ideas?

Darwin was an incredibly smart guy. One of the many, many things that he got right was that natural selection is often brought about by the interactions of organisms with other organisms, as well as interactions of organisms with their physical environment. Natural selection isn't some nebulous agency out there that's choosing survivors over nonsurvivors.

Why is it important for people to understand evolution?

There are many reasons. An understanding of evolution is certainly of practical importance in medicine and agriculture. But an understanding of evolution also gives us a closer connection to the rest of life. Also, it's very important for people to understand that the theory of evolution, like all scientific theories, is a body of explanation and fact that explains natural phenomena and can predict them. A lot of the resistance to evolution comes down to the idea some people have that somehow evolution makes life meaningless or purposeless. To this I reply that meaning and purpose is an emergent property of evolution! It is our own responsibility to make life meaningful.

Geerat Vermeij (right) with Michael Cain (center) and Jane Reece