Chapter 18

Genomes and Their Evolution

Chapter Focus

This chapter introduces genomics and bioinformatics, new approaches to analyzing and comparing the genomes of life's diverse organisms. Researchers in these fields address questions about genome organization, gene expression, growth and development, and evolution. The various types of noncoding DNA in the human genome are described. The chapter also covers the processes that contribute to genome evolution.

Chapter Review

The complete genome sequences of humans, chimpanzees, and numerous other eukaryotes and prokaryotes are enabling the study of whole sets of genes and their interactions, called **genomics**. The new field of **bioinformatics** is applying computational methods to the analysis of the ever-growing volume of biological data.

18.1 The Human Genome Project fostered development of faster, less expensive sequencing techniques

The international effort to sequence the human genome, called the **Human Genome Project**, was begun in 1990 and declared "virtually completed" in 2006. Technological advances during this project have tremendously accelerated DNA sequencing. Today's sequencing machines are "high-throughput" devices, able to analyze biological materials rapidly and produce enormous volumes of data.

J. C. Venter, founder of the company Celera Genomics, developed a **whole-genome shotgun approach** that relies on powerful computer programs to order the large number of sequenced, overlapping short fragments cut from a chromosome. This approach is now widely used. Newer sequencing techniques, called *sequencing by synthesis*, have both increased the speed and decreased the cost of sequencing genomes.

Metagenomics is an approach in which an environmental sample that includes DNA from many species is sequenced, and then computers sort the sequences into specific genomes.

FOCUS QUESTION 18.1

What types of organisms have been best studied using metagenomics?

18.2 Scientists use bioinformatics to analyze genomes and their functions

Centralized Resources for Analyzing Genome Sequences The National Center for Biotechnology Information (NCBI) and other genome centers maintain websites with databases of DNA sequences and protein sequences and structures, as well as software for analyzing and comparing data. The ever-expanding NCBI database of sequences, called GenBank, makes the resources of bioinformatics available to researchers worldwide. A Protein Data Bank contains all the three-dimensional protein structures determined thus far.

Understanding the Functions of Protein-Coding Genes The sequences of newly identified genes are compared with sequences of known genes of other species to look for similarities that might indicate the gene's function. A combination of biochemical techniques (identifying the three-dimensional structure and potential binding sites of the protein) and functional studies (blocking or disabling the gene to determine the effect on phenotype) helps to reveal protein function.

Understanding Genes and Gene Expression at the Systems Level In addition to comparing genomes of different species, genomics also considers the interactions of genes in a genome. A research project called ENCODE (Encyclopedia of DNA Elements) analyzed 1% of the human genome to identify protein-coding and noncoding RNA genes, regulatory sequences, and chromatin modifications. Researchers found that only 2% of the region they studied codes for proteins, but over 90% of the region was transcribed into RNA. This approach has now been extended to the entire human genome and to those of the nematode *C. elegans* and the fruit fly *D. melangaster*.

Proteomics is the identification and study of entire protein sets *(proteomes)* coded for by a genome. With compiled lists of DNA sequences and proteins now available, researchers are studying the functional integration of these components in biological systems. This **systems biology** approach seeks to model the dynamic behavior of whole systems, using bioinformatics to process and integrate huge amounts of data.

The Cancer Genome Atlas is taking a systems biology approach to analyzing changes in genes and patterns of gene expression in cancer cells. The project catalogues the mutations found in several common cancers and identifies genes of known and unknown functions that may provide new targets for therapies. Silicon "chips" holding arrays of most known human genes are used to analyze gene expression in patients with various diseases. Such approaches may help to match treatments with a person's unique genetic makeup.

FOCUS QUESTION 18.2

The software program BLAST, available on the NCBI website, allows a researcher to compare a DNA sequence to every sequence in GenBank to identify similar regions. What are some uses of this function?

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

Genome Size Of the 3,700 genomes that had been sequenced as of August 2012, most are of bacteria, whose genomes have between 1 and 6 million base pairs (Mb). Archaeal genomes appear to be of a similar size, whereas most animals and plants have genomes of at least 100 Mb. Among eukaryotes, there does not appear to be a correlation between genome size and an organism's phenotype.

Number of Genes Bacteria and archaea have from 1,500 to 7,500 genes; eukaryotes range from about 5,000 genes for unicellular fungi to 40,000 for some multicellular organisms. The human genome contains fewer than 21,000 genes. Alternative splicing of the ten or so exons in most human genes can yield many different proteins for each gene. Small RNAs (such as miRNAs) that regulate gene expression may contribute to greater organismal complexity.

Gene Density and Noncoding DNA Eukaryotes have fewer genes per million base pairs than bacteria or archaea. Mammals appear to have the lowest gene density. Most of the DNA of eukaryotic genomes is noncoding DNA that is located within and between genes (such as introns and complex regulatory sequences) as well as non-protein-coding DNA between genes.

FOCUS QUESTION 18.3

Refer to the organisms listed in Table 18.1 in your text to answer the following questions.

a. Which organism has the highest gene density? __________ the lowest gene density? __________

b. Which organism has the largest number of genes? __________ the smallest number? __________

c. Which organism has the largest haploid genome size? __________ the smallest genome size? __________

d. What is the estimated number of genes in the human genome? __________ Explain the fact that there are many more different polypeptides than genes.

18.4 Multicellular eukaryotes have much noncoding DNA and many multigene families

About 1.5% of the human genome consists of exons that code for proteins, rRNA, or tRNA. The rest includes gene-related regulatory sequences (5%), introns (20%), gene fragments and nonfunctional former genes called **pseudogenes** (15%), and sequences present in many copies, called **repetitive DNA**. Much of this repetitive DNA (44% of the human genome) is either made up of or related to transposable elements.

Noncoding DNA, which was previously referred to as "junk DNA," may turn out to have important functions. Almost 500 identical regions of noncoding DNA have been identified in humans, rats, and mice, a higher level of sequence conservation than for protein-coding regions in these species.

Transposable Elements and Related Sequences Stretches of DNA that can move about within a genome through a process called *transposition* are called *transposable genetic elements* or **transposable elements.**

What are the two types of eukaryotic transposable elements? **Transposons** move about a genome as a DNA intermediate, either by a "cut-and-paste" mechanism or a "copy-and-paste" mechanism. The enzyme *transposase,* generally encoded by the transposon, is required for both mechanisms. **Retrotransposons,** which make up the majority of transposable elements, are first transcribed into an RNA intermediate. This RNA transcript is converted back to DNA by reverse transcriptase, which is coded for by the retrotransposon itself.

Transposable elements may be represented as multiple (although not identical) copies of transposons or as related sequences that have lost the ability to move. In humans, about 10% of the genome is made up of *Alu elements.* Many of these 300-nucleotide-long sequences are transcribed into RNA, which is of unknown function.

About 17% of the human genome consists of *LINE-1,* or *L1,* retrotransposons. The introns of about 80% of analyzed human genes contain L1 sequences, suggesting that L1 may help regulate gene expression.

FOCUS QUESTION 18.4

Why do retrotransposons always move by the "copy-and-paste" mechanism?

Other Repetitive DNA, Including Simple Sequence DNA About 14% of the human genome is repetitive DNA that appears to have arisen from mistakes in DNA replication. Scattered large-segment duplications account for 5–6% of the human genome. **Simple-sequence DNA,** by contrast, makes up 3% of the human genome and consists of multiple copies of tandemly repeated sequences. When the repeat consists of two to five nucleotides, the unit is called a **short tandem repeat, or STR.** The variation in repeat numbers between genomes is the basis for determining genetic profiles, which are used by forensic scientists. Much of a genome's simple sequence DNA is located at centromeres, where it functions in cell division and chromatin organization, and at telomeres, which protect the tips of chromosomes.

Genes and Multigene Families More than half of the gene-related DNA occurs in **multigene families,** collections of similar or identical genes.

With the exception of the genes for histone proteins, *identical* multigene families code for RNA products. The genes coding for the three largest rRNA molecules are arranged in a single transcription unit repeated in huge tandem arrays, enabling cells to produce the millions of ribosomes needed for protein synthesis.

Examples of multigene families of *nonidentical* genes are the two families of genes that code for globins, including the α and β polypeptide subunits of hemoglobin. Different versions of each globin subunit are clustered together on two different chromosomes and are expressed at the appropriate times during development. The families also include several pseudogenes.

FOCUS QUESTION 18.5

For each of the following types of DNA sequences found in the human genome, write the letter of the correct description and the percentage of the genome (listed beneath the descriptions) in the blanks provided.

Types of DNA Description %

1. Exons or rRNA/tRNA-coding ______ ______
2. Introns ______ ______
3. Regulatory sequences ______ ______
4. Transposable elements and related sequences ______ ______
5. *Alu* elements ______ ______
6. L1 sequences ______ ______
7. Unique noncoding DNA ______ ______
8. Large-segment duplications ______ ______
9. Simple sequence DNA ______ ______

Descriptions

A. DNA in centromeres and telomeres, also STRs
B. multiple copies of mostly movable sequences
C. gene fragments and pseudogenes
D. protein- and RNA-coding sequences
E. family of short sequences related to transposable elements
F. multiple copies of large sequences
G. retrotransposons found in introns of most genes
H. enhancers, promoters, and other such sequences
I. noncoding sequences within genes

Choices of percentages: 1.5, 3, 5, 5–6, 10, 15, 17, 20, and 44. (These percentages do not add up to 100 because some of these types of DNA are subsets of other categories, and some types are not listed.)

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

Duplication of Entire Chromosome Sets Extra sets of chromosomes may arise by accidents in meiosis. The resulting extra genes might diverge through mutation, leading to genes with novel functions. Polyploidy is fairly common in plants.

Alterations of Chromosome Structure Using genomic sequence information, researchers can compare the locations of DNA sequences on chromosomes among different species and reconstruct the evolutionary history of chromosomal rearrangements. Duplications and inversions of chromosomes are thought to contribute to speciation, in that matings between individuals from populations with differing chromosomal rearrangements would be less successful.

Duplication and Divergence of Gene-Sized Regions of DNA Errors such as unequal crossing over during meiosis (as may occur between copies of a transposable element on misaligned nonsister chromatids) and slippage of template strands during DNA replication might lead to the duplication of genes.

The α-globin and β-globin gene families appear to have evolved from a common ancestral globin gene, which was duplicated and then diverged. Multiple duplications and mutations within each family have led to the current family of genes with related functions along with several intervening pseudogenes.

In other cases, mutation of a duplicated gene may lead to a protein product with a new function.

FOCUS QUESTION 18.6

Lysozyme and α-lactalbumin have similar amino acid sequences but different functions. The genes for both proteins are found in mammals, but birds have only the gene for lysozyme. What does this observation suggest about the evolution of these genes?

Rearrangements of Parts of Genes: Exon Duplication and Exon Shuffling Unequal crossing over can lead to a gene with a duplicated exon. Exons often code for structural or functional regions called **domains**, and their duplication could provide a protein with enhanced properties. Errors in meiotic recombination could also lead to *exon shuffling* within a gene or between nonallelic genes.

How Transposable Elements Contribute to Genome Evolution Recombination events can take place between homologous transposable element sequences that are scattered throughout the genome, causing chromosomal mutations that may occasionally be beneficial to the organism. Transposable elements that insert within a gene may disrupt its functioning; those that insert within regulatory sequences may increase or decrease gene expression. A transposable element can also move a copy of a gene or an exon to a new location. The increased genetic diversity provided by these mechanisms provides raw material for natural selection.

FOCUS QUESTION 18.7

a. Explain two ways in which exon shuffling could occur.

b. What is a potential benefit of exon shuffling?

18.6 Comparing genome sequences provides clues to evolution and development

Comparing Genomes Comparisons of *highly conserved* genes illuminate the evolutionary relationships among species that are distantly related. Such analyses support the theory that bacteria, archaea, and eukaryotes represent the three domains of life, and also demonstrate the advantages of using model organisms to study both basic biological processes and human biology.

The similarity between genomes of two closely related species allows researchers to use one genome sequence as a framework for mapping the other genome. Also, the identified small differences between genomes can be correlated with the phenotypic divergence of the species. The human and chimpanzee genomes differ in single nucleotide substitutions by only 1.2%. Insertions or deletions of larger regions in the genome result in an additional 2.7% difference. There are more *Alu* elements in the human genome, and a third of the human duplications are not present in the chimpanzee genome.

Comparisons of genetic changes since species diverged show that some genes are changing faster in humans than in the chimpanzee or mouse. Many of these more-quickly evolving genes code for transcription factors; one example is the *FOXP2* gene, which appears to function in vocalization in vertebrates and in speech and language in humans. Mutations in this gene cause verbal impairment in humans; it is expressed in the brains of songbirds during the period they are learning their songs; and knock-out experiments with mice have shown that homozygous mutant mice had malformed brains and they, along with heterozygous mice, did not produce their normal vocalizations.

Comparisons of human genomes have revealed several million **single nucleotide polymorphisms (SNPs)**, single base-pair sites where variation is found in at least 1% of the population. Comparisons have also revealed inversions, deletions, duplications, and a surprisingly high number of *copy-number variants* (*CNVs*) in which some individuals have one or multiple copies of a gene or genetic region rather than the normal two. Such CNVs likely have phenotypic effects. Genetic markers such as SNPs, CNVs, and variations in repetitive DNA (such as STRs) will contribute to the study of human evolution.

Comparing Developmental Processes Biologists in the field of evolutionary developmental biology (**evo-devo**) compare developmental processes to understand how they have evolved and how minor changes in gene sequence or regulation may lead to diverse forms of life.

A sequence of 180 nucleotides called a **homeobox**, which codes for a *homeodomain*, has been found in *Drosophila* homeotic genes. The same or very similar homeobox nucleotide sequences have been identified in homeotic genes of many animals. Homeotic genes in the fruit fly and mouse are found in the same linear sequence on chromosomes. Related sequences are found in regulatory genes of yeast and plants. These similarities indicate that the homeobox sequence must have arisen early and been conserved through evolution as part of the genes involved in the regulation of gene expression and development.

Homeotic genes are often called *Hox* genes in animals. Proteins with homeodomains probably coordinate the transcription of groups of developmental genes.

Many other genes involved in development, such as those coding for components of signaling pathways, are highly conserved. The differing patterns of expression of these genes in different body areas may explain the development of animals with different body plans.

FOCUS QUESTION 18.8

If all *Hox* genes contain the same or very similar homeobox, how can they control different developmental sequences?

Word Roots

pseudo- = false (*pseudogene:* a DNA segment that is very similar to a real gene but does not yield a functional product)

retro- = backward (*retrotransposon:* a transposable element that moves within a genome by means of an RNA intermediate, a transcript of the retrotransposon DNA)

Structure Your Knowledge

1. About 25% of the human genome relates to the production of proteins or RNA products (exons, introns, or regulatory sequences). Is the remaining 75% just "junk"? Describe the following types of noncoding DNA, including some of their possible functions.
   a. transposable elements
   b. *Alu* elements
   c. L1 sequences
   d. simple sequence DNA
   e. pseudogenes
2. Describe some of the processes that contribute to genome evolution.

Test Your Knowledge

MULTIPLE CHOICE: *Choose the one best answer.*

1. Why is the whole-genome shotgun approach now widely used to sequence genomes?
   a. It uses only one, very efficient restriction enzyme to create fragments.
   b. It makes use of linkage and physical maps to order sections of a chromosome.
   c. Newer sequencing techniques, such as sequencing by synthesis, and enhanced computer software can rapidly assemble overlapping fragments into complete sequences.
   d. It uses multiple research labs, which share their results on the Internet.
   e. All of the above contribute to its widespread use.
2. Metagenomics is a new approach that
   a. identifies proteomes and protein interaction networks.
   b. analyzes genomes for all functionally important elements.
   c. provides sequence data and software programs on Internet websites.

d. sequences all the DNA in an environmental sample and uses computer software to assemble the sequences into specific genomes.

e. applies genome-wide association studies to the identification of human genes of medical importance.

3. Why is proteomics important in the systems biology approach?

a. The interactions of networks of proteins are central to the functioning of cells and organisms.

b. This bioinformatics field allows for the mathematical modeling of biological systems.

c. Determining the proteins expressed in a cell identifies the genes more accurately than can be done through genomics.

d. The three-dimensional structure of a protein can be used to predict its function.

e. Comparing the proteins produced by a normal allele and the allele associated with a disease can facilitate improved treatments.

4. Bacterial genes have an average length of 1,000 base pairs; human genes average about 27,000 base pairs. Which of the following statements is the best explanation for that difference?

a. Prokaryotes have smaller, but many more, individual genes.

b. Prokaryotes are more ancient organisms; longer genes arose later in evolution.

c. Prokaryotes are unicellular; humans have many types of differentiated cells.

d. Prokaryotic genes do not have introns; human genes have multiple introns.

e. Prokaryotic proteins are not as large and complex as human proteins.

5. Which of the following statements best explains the discovery that a complex human has roughly the same number of genes as the simple nematode *C. elegans*?

a. The unusually long introns in human genes are involved in regulation of gene expression.

b. More than one polypeptide can be produced from a human gene by alternative splicing.

c. Human genes code for many more types of domains.

d. The human genome has a high proportion of noncoding DNA.

e. The large number of SNPs (single nucleotide polymorphisms) in the human genome provides a great deal of genetic variability.

6. Which of the following statements *best* describes what pseudogenes and introns have in common?

a. They do not result in a functional product.

b. They are DNA segments that lack a promoter but have other control regions.

c. They are transcribed but their translation is blocked by miRNAs.

d. They code for RNA products, not proteins.

e. They appear to have arisen from retrotransposons.

7. Which of the following techniques can be used to determine the function of a newly identified gene?

a. comparisons with genes of known functions that have similar sequences

b. blockage of gene function to see the effect on the phenotype

c. searches for similar sequences for domains of known function in other proteins.

d. both a and b

e. a, b, and c

8. Which of the following statements is *not* descriptive of transposable elements?

a. Barbara McClintock's work with maize provided the first evidence of such DNA segments.

b. Transposable elements or related sequences make up 85% of the corn genome.

c. Retrotransposons called *LINE-1* are found within the introns of many human genes and may help regulate gene expression.

d. Transposable elements often encode the enzymes, such as transposase or reverse transcriptase, necessary for their movement.

e. Each transposable element is present as multiple identical copies, often clustered in the centromere or telomere regions of a chromosome.

9. The protein tissue plasminogen activator (TPA) has three types of domains. Which of the following statements *best* explains why one of each of these types of domains is found in three different proteins (epidermal growth factor, fibronectin, and plasminogen)?

a. The genes for all four proteins are members of a multigene family involved in cell signaling.

b. The gene for TPA was the first gene to evolve; the other three genes each lost two of the domains from the ancestral TPA gene.

c. The gene for TPA arose by exon shuffling involving the other three genes.

d. The gene for TPA has many *Alu* elements that provide alternative splice sites to incorporate these exons.

e. Several duplication events led to the evolution of the TPA gene.

10. Genes that are highly conserved are useful for
    a. identifying genes that led to new species.
    b. determining the function of newly discovered genes.
    c. establishing the sequence of divergence of closely related species.
    d. tracing the relationships of groups that diverged early in the evolution of life.
    e. both a and c.

11. A highly conserved nucleotide sequence that has been found in developmental regulatory genes in many diverse organisms is called
    a. a homeodomain.
    b. a homeobox.
    c. a retrotransposon.
    d. a homeotic gene.
    e. an L1 sequence.

12. Which of the following approaches would be most useful in tracing human evolution?
    a. evo-devo and the comparison of developmental genes in plants and animals
    b. metagenomics and proteomics
    c. systems biology and the use of "knock-out" experiments
    d. analysis of single nucleotide polymorphisms and copy-number variants across individuals from the same and different populations
    e. All of the above make important contributions to studying the evolution of human populations.