Chapter 21
Genomes and Their Evolution

Overview: Reading the Leaves from the Tree of Life

- Complete genome sequences exist for a human, chimpanzee, *E. coli*, brewer's yeast, corn, fruit fly, house mouse, rhesus macaque, and other organisms
- Comparisons of genomes among organisms provide information about the evolutionary history of genes and taxonomic groups

- **Genomics** is the study of whole sets of genes and their interactions
- **Bioinformatics** is the application of computational methods to the storage and analysis of biological data

Concept 21.1: New approaches have accelerated the pace of genome sequencing

- The most ambitious mapping project to date has been the sequencing of the human genome
- Officially begun as the Human Genome Project in 1990, the sequencing was largely completed by 2003
- The project had three stages
  - Genetic (or linkage) mapping
  - Physical mapping
  - DNA sequencing

Three-Stage Approach to Genome Sequencing

- A **linkage map** (genetic map) maps the location of several thousand genetic markers on each chromosome
- A genetic marker is a gene or other identifiable DNA sequence
- Recombination frequencies are used to determine the order and relative distances between genetic markers
• **A physical map** expresses the distance between genetic markers, usually as the number of base pairs along the DNA.
  
  • It is constructed by cutting a DNA molecule into many short fragments and arranging them in order by identifying overlaps.

• Sequencing machines are used to determine the complete nucleotide sequence of each chromosome.
  
  • A complete haploid set of human chromosomes consists of 3.2 billion base pairs.
Whole-Genome Shotgun Approach to Genome Sequencing

- The whole-genome shotgun approach was developed by J. Craig Venter in 1992
- This approach skips genetic and physical mapping and sequences random DNA fragments directly
- Powerful computer programs are used to order fragments into a continuous sequence

Both the three-stage process and the whole-genome shotgun approach were used for the Human Genome Project and for genome sequencing of other organisms
At first many scientists were skeptical about the whole-genome shotgun approach, but it is now widely used as the sequencing method of choice
The development of newer sequencing techniques has resulted in massive increases in speed and decreases in cost

Technological advances have also facilitated metagenomics, in which DNA from a group of species (a metagenome) is collected from an environmental sample and sequenced
This technique has been used on microbial communities, allowing the sequencing of DNA of mixed populations, and eliminating the need to culture species in the lab
Concept 21.2 Scientists use bioinformatics to analyze genomes and their functions
• The Human Genome Project established databases and refined analytical software to make data available on the Internet
• This has accelerated progress in DNA sequence analysis

Centralized Resources for Analyzing Genome Sequences
• Bioinformatics resources are provided by a number of sources
  – National Library of Medicine and the National Institutes of Health (NIH) created the National Center for Biotechnology Information (NCBI)
  – European Molecular Biology Laboratory
  – DNA Data Bank of Japan
  – BGI in Shenzhen, China

Identifying Protein-Coding Genes and Understanding Their Functions
• Using available DNA sequences, geneticists can study genes directly in an approach called reverse genetics
• The identification of protein coding genes within DNA sequences in a database is called gene annotation

• Gene annotation is largely an automated process
• Comparison of sequences of previously unknown genes with those of known genes in other species may help provide clues about their function

Genbank, the NCBI database of sequences, doubles its data approximately every 18 months
• Software is available that allows online visitors to search Genbank for matches to
  – A specific DNA sequence
  – A predicted protein sequence
  – Common stretches of amino acids in a protein
• The NCBI website also provides 3-D views of all protein structures that have been determined
Understanding Genes and Gene Expression at the Systems Level

- **Proteomics** is the systematic study of all proteins encoded by a genome.
- Proteins, not genes, carry out most of the activities of the cell.

How Systems Are Studied: An Example

- A systems biology approach can be applied to define gene circuits and protein interaction networks.
- Researchers working on the yeast *Saccharomyces cerevisiae* used sophisticated techniques to disable pairs of genes one pair at a time, creating double mutants.
- Computer software then mapped genes to produce a network-like “functional map” of their interactions.
- The systems biology approach is possible because of advances in bioinformatics.

Application of Systems Biology to Medicine

- A systems biology approach has several medical applications:
  - The Cancer Genome Atlas project is currently seeking all the common mutations in three types of cancer by comparing gene sequences and expression in cancer versus normal cells.
  - This has been so fruitful, it will be extended to ten other common cancers.
  - Silicon and glass “chips” have been produced that hold a microarray of most known human genes.
Figure 21.6

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

- By early 2010, over 1,200 genomes were completely sequenced, including 1,000 bacteria, 80 archaea, and 124 eukaryotes
- Sequencing of over 5,500 genomes and over 200 metagenomes is currently in progress

Genome Size

- Genomes of most bacteria and archaea range from 1 to 6 million base pairs (Mb); genomes of eukaryotes are usually larger
- Most plants and animals have genomes greater than 100 Mb; humans have 3,000 Mb
- Within each domain there is no systematic relationship between genome size and phenotype

Table 21.1

The number of genes is not correlated to genome size
- For example, it is estimated that the nematode *C. elegans* has 100 Mb and 20,000 genes, while *Drosophila* has 165 Mb and 13,700 genes
- Vertebrate genomes can produce more than one polypeptide per gene because of alternative splicing of RNA transcripts

Number of Genes

- Free-living bacteria and archaea have 1,500 to 7,500 genes
- Unicellular fungi have from about 5,000 genes and multicellular eukaryotes up to at least 40,000 genes
Gene Density and Noncoding DNA

- Humans and other mammals have the lowest gene density, or number of genes, in a given length of DNA
- Multicellular eukaryotes have many introns within genes and noncoding DNA between genes

Concept 21.4: Multicellular eukaryotes have much noncoding DNA and many multigene families

- The bulk of most eukaryotic genomes neither encodes proteins nor functional RNAs
- Much evidence indicates that noncoding DNA (previously called “junk DNA”) plays important roles in the cell
- For example, genomes of humans, rats, and mice show high sequence conservation for about 500 noncoding regions

- Sequencing of the human genome reveals that 98.5% does not code for proteins, rRNAs, or tRNAs
- About a quarter of the human genome codes for introns and gene-related regulatory sequences

Intergenic DNA is noncoding DNA found between genes
- Pseudogenes are former genes that have accumulated mutations and are nonfunctional
- Repetitive DNA is present in multiple copies in the genome
- About three-fourths of repetitive DNA is made up of transposable elements and sequences related to them

Transposable Elements and Related Sequences

- The first evidence for mobile DNA segments came from geneticist Barbara McClintock’s breeding experiments with Indian corn
- McClintock identified changes in the color of corn kernels that made sense only by postulating that some genetic elements move from other genome locations into the genes for kernel color
- These transposable elements move from one site to another in a cell’s DNA; they are present in both prokaryotes and eukaryotes
Movement of Transposons and Retrotransposons

- Eukaryotic transposable elements are of two types
  - Transposons, which move by means of a DNA intermediate
  - Retrotransposons, which move by means of an RNA intermediate
Sequences Related to Transposable Elements

- Multiple copies of transposable elements and related sequences are scattered throughout the eukaryotic genome
- In primates, a large portion of transposable element–related DNA consists of a family of similar sequences called Alu elements
- Many Alu elements are transcribed into RNA molecules; however their function, if any, is unknown

Other Repetitive DNA, Including Simple Sequence DNA

- About 15% of the human genome consists of duplication of long sequences of DNA from one location to another
- In contrast, simple sequence DNA contains many copies of tandemly repeated short sequences

Genes and Multigene Families

- Many eukaryotic genes are present in one copy per haploid set of chromosomes
- The rest of the genes occur in multigene families, collections of identical or very similar genes
- Some multigene families consist of identical DNA sequences, usually clustered tandemly, such as those that code for rRNA products
The classic examples of multigene families of nonidentical genes are two related families of genes that encode globins. α-globins and β-globins are polypeptides of hemoglobin and are coded by genes on different human chromosomes and are expressed at different times in development.

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

- The basis of change at the genomic level is mutation, which underlies much of genome evolution.
- The earliest forms of life likely had a minimal number of genes, including only those necessary for survival and reproduction.
- The size of genomes has increased over evolutionary time, with the extra genetic material providing raw material for gene diversification.

Duplication of Entire Chromosome Sets

- Accidents in meiosis can lead to one or more extra sets of chromosomes, a condition known as polyploidy.
- The genes in one or more of the extra sets can diverge by accumulating mutations; these variations may persist if the organism carrying them survives and reproduces.
Alterations of Chromosome Structure

- Humans have 23 pairs of chromosomes, while chimpanzees have 24 pairs.
- Following the divergence of humans and chimpanzees from a common ancestor, two ancestral chromosomes fused in the human line.
- Duplications and inversions result from mistakes during meiotic recombination.
- Comparative analysis between chromosomes of humans and seven mammalian species paints a hypothetical chromosomal evolutionary history.

The rate of duplications and inversions seems to have accelerated about 100 million years ago.
This coincides with when large dinosaurs went extinct and mammals diversified.
Chromosomal rearrangements are thought to contribute to the generation of new species.
Some of the recombination “hot spots” associated with chromosomal rearrangement are also locations that are associated with diseases.

Duplication and Divergence of Gene-Sized Regions of DNA

- Unequal crossing over during prophase I of meiosis can result in one chromosome with a deletion and another with a duplication of a particular region.
- Transposable elements can provide sites for crossover between nonsister chromatids.
Evolution of Genes with Related Functions: The Human Globin Genes

- The genes encoding the various globin proteins evolved from one common ancestral globin gene, which duplicated and diverged about 450–500 million years ago.
- After the duplication events, differences between the genes in the globin family arose from the accumulation of mutations.

- Subsequent duplications of these genes and random mutations gave rise to the present globin genes, which code for oxygen-binding proteins.
- The similarity in the amino acid sequences of the various globin proteins supports this model of gene duplication and mutation.

Evolution of Genes with Novel Functions

- The copies of some duplicated genes have diverged so much in evolution that the functions of their encoded proteins are now very different.
- For example, the lysozyme gene was duplicated and evolved into the gene that encodes α-lactalbumin in mammals.
- Lysozyme is an enzyme that helps protect animals against bacterial infection.
- α-lactalbumin is a nonenzymatic protein that plays a role in milk production in mammals.
Rearrangements of Parts of Genes: Exon Duplication and Exon Shuffling

- The duplication or repositioning of exons has contributed to genome evolution
- Errors in meiosis can result in an exon being duplicated on one chromosome and deleted from the homologous chromosome
- In exon shuffling, errors in meiotic recombination lead to some mixing and matching of exons, either within a gene or between two nonallelic genes

How Transposable Elements Contribute to Genome Evolution

- Multiple copies of similar transposable elements may facilitate recombination, or crossing over, between different chromosomes
- Insertion of transposable elements within a protein-coding sequence may block protein production
- Insertion of transposable elements within a regulatory sequence may increase or decrease protein production

Concept 21.6: Comparing genome sequences provides clues to evolution and development

- Genome sequencing and data collection has advanced rapidly in the last 25 years
- Comparative studies of genomes
  - Advance our understanding of the evolutionary history of life
  - Help explain how the evolution of development leads to morphological diversity

Comparing Genomes

- Genome comparisons of closely related species help us understand recent evolutionary events
- Genome comparisons of distantly related species help us understand ancient evolutionary events
- Relationships among species can be represented by a tree-shaped diagram
Comparing Distantly Related Species

- Highly conserved genes have changed very little over time
- These help clarify relationships among species that diverged from each other long ago
- Bacteria, archaea, and eukaryotes diverged from each other between 2 and 4 billion years ago
- Highly conserved genes can be studied in one model organism, and the results applied to other organisms

Comparing Closely Related Species

- Genetic differences between closely related species can be correlated with phenotypic differences
- For example, genetic comparison of several mammals with nonmammals helps identify what it takes to make a mammal

Humans and chimpanzees differ in the expression of the **FOXP2** gene, whose product turns on genes involved in vocalization
- Differences in the **FOXP2** gene may explain why humans but not chimpanzees communicate by speech

Experiment

**Wild type:** two normal copies of **FOXP2**

**Heterozygote:** one copy of **FOXP2** disrupted

**Homozygote:** both copies of **FOXP2** disrupted

**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.

<table>
<thead>
<tr>
<th></th>
<th>Wild type</th>
<th>Heterozygote</th>
<th>Homozygote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of whistles</td>
<td>400</td>
<td>300</td>
<td>200</td>
</tr>
</tbody>
</table>

Wild type | Heterozygote | Homozygote
---|-------------|-------------

(No whistles)
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**
- Wild type: two normal copies of FOXP2
- Heterozygote: one copy of FOXP2 disrupted
- Homozygote: both copies of FOXP2 disrupted

**Experiment 2**
- Number of whistles
  - Wild type
  - Heterozygote
  - Homozygote (No whistles)

**Figure 21.17a**
- Wild type
- Heterozygote
- Homozygote

**Figure 21.17b**
- Wild type
- Heterozygote
- Homozygote (No whistles)

**Figure 21.17c**
- Wild type
- Heterozygote
- Homozygote

**Figure 21.17d**
- Wild type
- Heterozygote
- Homozygote (No whistles)

**Figure 21.17e**
- Wild type
- Heterozygote
- Homozygote

**Figure 21.17f**
- Wild type
- Heterozygote
- Homozygote (No whistles)
Comparing Genomes Within a Species

- As a species, humans have only been around about 200,000 years and have low within-species genetic variation
- Variation within humans is due to single nucleotide polymorphisms, inversions, deletions, and duplications
- Most surprising is the large number of copy-number variants
- These variations are useful for studying human evolution and human health

Comparing Developmental Processes

- Evolutionary developmental biology, or evo-devo, is the study of the evolution of developmental processes in multicellular organisms
- Genomic information shows that minor differences in gene sequence or regulation can result in striking differences in form

Widespread Conservation of Developmental Genes Among Animals

- Molecular analysis of the homeotic genes in Drosophila has shown that they all include a sequence called a homeobox
- An identical or very similar nucleotide sequence has been discovered in the homeotic genes of both vertebrates and invertebrates
- Homeobox genes code for a domain that allows a protein to bind to DNA and to function as a transcription regulator
- Homeotic genes in animals are called Hox genes
• Related homeobox sequences have been found in regulatory genes of yeasts, plants, and even prokaryotes
• In addition to homeotic genes, many other developmental genes are highly conserved from species to species

• Sometimes small changes in regulatory sequences of certain genes lead to major changes in body form
• For example, variation in Hox gene expression controls variation in leg-bearing segments of crustaceans and insects
• In other cases, genes with conserved sequences play different roles in different species

Comparison of Animal and Plant Development

• In both plants and animals, development relies on a cascade of transcriptional regulators turning genes on or off in a finely tuned series
• Molecular evidence supports the separate evolution of developmental programs in plants and animals
• Mads-box genes in plants are the regulatory equivalent of Hox genes in animals

<table>
<thead>
<tr>
<th>Genome size</th>
<th>Bacteria</th>
<th>Archaea</th>
<th>Eukarya</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes</td>
<td>Most are 1–6 Mb</td>
<td>Most are 10–4,000 Mb, but a few are much larger</td>
<td></td>
</tr>
<tr>
<td>Gene density</td>
<td>1,500–7,500</td>
<td>5,000–40,000</td>
<td></td>
</tr>
<tr>
<td>Gene density</td>
<td>Higher than in eukaryotes</td>
<td>Lower than in prokaryotes (Within eukaryotes, lower density is correlated with larger genomes.)</td>
<td></td>
</tr>
<tr>
<td>Introns</td>
<td>None in protein-coding genes</td>
<td>Present in some genes</td>
<td></td>
</tr>
<tr>
<td>Other noncoding DNA</td>
<td>Very little</td>
<td>Can be large amounts; generally more repetitive noncoding DNA in multicellular eukaryotes</td>
<td></td>
</tr>
</tbody>
</table>

Human genome

Protein-coding, rRNA, and tRNA genes (1.5%)
Introns and regulatory sequences (~26%)
Repetitive DNA (green and teal)
α-Globin gene family
Chromosome 16
ζ ψ2 ψ1 Δ2 α1 ψ0

β-Globin gene family
Chromosome 11
ε αγ Δγ ψβ δ β

Crossover point

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