Chapter 21

Genomes and Their Evolution

Lecture Outline

Overview: Reading the Leaves from the Tree of Life

- The advent of techniques for mapping genomes by rapid, complete genome sequencing enabled scientists to sequence the human genome by 2003 and the genome of the chimp, *Pan troglodytes*, by 2005.
  - Scientists can now ask what differences in genetic information account for the distinct characteristics of humans and chimps.
- Researchers have also completed genome sequences for *Escherichia coli* and numerous other prokaryotes, *Saccharomyces cerevisiae* (brewer's yeast), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), *Macaca mulatta* (macaque), and others.
- Fragments of DNA have been sequenced from extinct species, including the woolly mammoth.
- Comparing the genomes of more distantly related animals should reveal the sets of genes that control group-defining characteristics.
- Comparing the genomes of plants and prokaryotes provides information 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 contribute to this approach generate enormous volumes of data.
  - The need to deal with this information has spawned the field of **bioinformatics**, 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 **Human Genome Project** (HGP) began in 1990, with the goal of sequencing the human genome.

The Human Genome Project used a three-stage approach to mapping the human genome.

- The starting point for the HGP was an incomplete picture of the organization of many genomes.
  - Geneticists had karyotypes for many species, showing the number and banding pattern of chromosomes.
  - The locations of some genes had been identified by fluorescence *in situ* hybridization (FISH), in which fluorescently labeled probes hybridize to an immobilized array of chromosomes.
- The initial stage in the three-stage approach to mapping the human genome was to construct a **linkage map** of several thousand genetic markers spaced throughout each of the chromosomes.
  - 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

    - By 1992, researchers had compiled a human genetic map with about 5,000 markers, enabling them to locate genes by testing for genetic linkage to known markers.
- The second 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.
- 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, to identify the overlaps, and then to assign fragments to a sequential order that corresponds to their order in a chromosome.
- Supplies of the DNA fragments used for physical mapping are prepared by DNA cloning.
    - The first cloning vector is often a yeast artificial chromosome (YAC), which can carry inserted fragments a million base pairs long, or a bacterial artificial chromosome (BAC), which carries inserts of 100,000–300,000 base pairs.
    - After these long fragments are ordered, each fragment is cut into smaller pieces, which are cloned in plasmids or phages, ordered in turn, and finally sequenced.
- The third stage in mapping a genome was to determine the complete nucleotide sequence of each chromosome.
- The sequencing of all 3.2 billion base pairs in a haploid set of human chromosomes presented a formidable challenge.
- This challenge was met by sequencing machines, using the dideoxy chain-termination method.
    - The development of technology for faster sequencing has accelerated the rate of sequencing dramatically—from 1,000 base pairs a day in the 1980s to 1,000 base pairs *per second* in 2000.
    - Methods 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.

The whole-genome shotgun method was adopted in the 1990s.

- In 1992, molecular biologist J. Craig Venter proposed that the sequencing of whole genomes should start directly with the sequencing of random DNA fragments, skipping the genetic mapping and physical mapping stages.
    - Powerful computer programs would then assemble the resulting very large number of overlapping short sequences into a single continuous sequence.
- In 1995, Venter and his colleagues reported the first complete genome sequence of an organism, the bacterium *Haemophilus influenzae*.
- In May 1998, Venter set up a company, Celera Genomics, and declared his intention to complete the human genome sequence.
- In March 2000, Celera Genomics completed the genome sequence of *D. melanogaster*.
- In April 2003, two years before the original target date of the Human Genome Project, the human genome sequence was announced jointly by Celera and the public consortium.
- Celera had kept its data private, whereas the data of the public consortium were made available to all.
    - Celera's accomplishment relied heavily on the consortium's maps and sequence data.

  - Nevertheless, Venter argues for the efficiency and economy of Celera's methods.
  - Both approaches led to the rapid completion of genome sequencing for a number of species.

- Today, the whole-genome shotgun method is widely used
  - The DNA fragments are cloned into three different vectors, each of which takes a defined size of insert.
  - The computer uses the known distance between the ends of the inserted DNA, along with other information, to assemble the sequences.
- A recent study comparing the whole-genome shotgun method with the three-stage approach found that the whole-genome shotgun method can miss some duplicated sequences, thus underestimating the size of the genome and missing some genes in those regions.
- The hybrid approach that ended up being used for the human genome, with the more rapid shotgun sequencing augmented by some mapping of clones, may be the most useful.
- Some gaps still remained with the publication of the detailed sequence of the final chromosome in 2006.
  - Because of the presence of repetitive DNA, certain parts of the chromosomes of multicellular organisms resist detailed mapping by the usual methods.

Concept 21.2 Scientists use bioinformatics to analyze genomes and their functions.

- The goals of the Human Genome Project included establishing databases and refining analytical software, both of which are centralized and readily accessible on the Internet.
- Bioinformatics resources are available to researchers worldwide, speeding up the dissemination of information.

Centralized resources are available for analyzing genome sequences.

- In the United States, the National Library of Medicine and the National Institutes of Health jointly created the National Center for Biotechnology Information (NCBI), which maintains a website with extensive bioinformatics resources.
  - Similar websites have been established by the European Molecular Biology Laboratory and the DNA Data Bank of Japan.
  - Smaller websites maintained by individual labs or groups of labs provide databases and software designed for narrower purposes, such as studying genetic and genomic changes in one particular type of cancer.
- The NCBI database of sequences is called Genbank.
  - As of June 2007, Genbank contained the sequences of 73 million fragments of genomic DNA, totaling 77 billion base pairs.
  - The amount of data in Genbank is estimated to double every 18 months.
- BLAST, a software program available on the NCBI website, allows visitors to compare a DNA sequence to every sequence in Genbank, in order to locate similar regions.
- Another program allows the comparison of predicted protein sequences.
- A third program searches protein sequences for common stretches of amino acids (domains) and generates a three-dimensional model of the domain.
- There is even a software program that can compare a collection of sequences of nucleic acids or polypeptides, and diagram them in the form of an evolutionary tree based on the sequence relationships!

- The NCBI website maintains a database of all three-dimensional protein structures that have been determined.

Protein-coding genes can be identified within DNA sequences.

- How can geneticists recognize protein-coding genes from DNA sequences and determine their function?
- Software is used to scan DNA sequences for transcriptional and translational start and stop signals, for RNA-splicing sites, and for other signs of protein-coding genes.
- Software also looks for certain short sequences that correspond to sequences present in known mRNAs.
  - Thousands of such sequences, called *expressed sequence tags*, or *ESTs*, have been collected from cDNA sequences and are cataloged in computer databases.
- The identities of about half of the human genes were known before the Human Genome Project began.
- Clues about the identities of previously unknown genes come from comparing the sequences of gene candidates with those of known genes from other organisms.
  - Due to redundancy in the genetic code, the DNA sequence may vary more than the protein sequence does.
- Scientists compare the predicted amino acid sequence of a protein with that of other proteins.
- Sometimes a newly identified sequence matches, at least partially, the sequence of a gene or protein whose function is well known.
  - If part of a new gene matches a known gene that encodes an important signaling pathway protein such as a protein kinase, then the new gene may, too.
- Some sequences are 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 these genes, function was 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 binding sites for other molecules.
  - Functional studies disable the gene to see what effect that has on the phenotype.

Genes and their products can be understood at the systems level.

- Genomics is a rich source of new insights into fundamental questions about genome organization, regulation of gene expression, growth and development, and evolution.
- 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**.
- 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.
- Using these catalogs, researchers have shifted their attention from the individual parts to their functional integration in biological systems.
- One basic application of the systems biology approach is to define gene circuits and protein interaction networks.

  - To map a protein interaction network in *D. melanogaster*, researchers started with more than 10,000 predicted RNA transcripts.
  - Researchers used molecular techniques to test interactions between the whole or partial protein products of these transcripts.
  - Using statistical tests to select the interactions for which the data were strongest, researchers ended up with roughly 4,700 proteins that appeared to participate in more than 4,000 interactions.

- The Cancer Genome Atlas is another example of systems biology in which a large group of interacting genes and gene products is analyzed together.
  - The National Cancer Institute and the National Institutes of Health aim to understand how changes in biological systems lead to cancer.
  - In a three-year pilot project running from 2007 to 2010, researchers are analyzing three types of cancer—lung, ovarian, and glioblastoma of the brain—by comparing gene sequences and patterns of gene expression in cancer cells with those in normal cells.
  - A set of approximately 2,000 genes from the cancer cells will be sequenced during the progression of the disease, to monitor changes due to mutations and rearrangements.
- The GeneChip is a microarray containing most of the known human genes.
  - The GeneChip is 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.
  - Ultimately, all of us may carry with our medical records a catalog of our DNA sequence, a sort of genetic bar code, with regions highlighted that predispose us to specific diseases.

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

- By the summer of 2007, the sequencing of more than 600 genomes had been completed and the sequencing of more than 2,100 genomes was in progress.
  - Of the completely sequenced group, about 500 are genomes of bacteria and 45 are archaeal genomes.
  - Among the 65 eukaryotic species in the group are vertebrates, invertebrates, and plants.
  - The accumulated genome sequences contain a wealth of information that we are just beginning to mine.

Comparing bacteria, archaea, and eukaryotes shows a general progression from smaller to larger genomes.

- 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 generally within the size range of bacterial genomes.
- Eukaryotic genomes tend to be larger: The genome of the single-celled yeast *S. cerevisiae* has about 13 Mb, whereas most multicellular animals and plants have genomes with at least 100 Mb.
  - There are 180 Mb in the fruit fly genome, whereas human genomes have 3,200 Mb.
- A comparison of genome sizes among eukaryotes does not show any systematic relationship between genome size and phenotype.
  - The genome of *Fritillaria assyriaca*, a flowering plant in the lily family, contains 120 billion base pairs (120,000 Mb), about 40 times more than the human genome.

  - A single-celled amoeba, *Amoeba dubia*, has a genome with 670 billion bases. (It has not yet been sequenced.)
  - The cricket genome has 11 times as many base pairs as *D. melanogaster*.
  - There is a wide range of genome sizes within the groups of protozoans, insects, amphibians, and plants and less of a range within mammals and reptiles.

Bacteria and archaea have fewer genes than eukaryotes.

- Free-living bacteria and archaea have 1,500–7,500 genes, whereas the number of genes in eukaryotes ranges from about 5,000 for unicellular fungi to at least 40,000 for multicellular eukaryotes.
- Within eukaryotes, the number of genes in a species is often lower than expected, considering the size of the genome.
  - The genome of the nematode *C. elegans* has 100 Mb and contains roughly 20,000 genes.
  - The *D. melanogaster* genome, in contrast, is almost twice as big (180 Mb) but has about two-thirds the number of genes—only 13,700 genes.
- At the outset of the Human Genome Project, biologists expected to identify between 50,000 and 100,000 genes based on the number of known human proteins.
- As the project progressed, the estimate was revised downward several times, and, as of 2007, the most reliable count is 20,488.
- This lower number, similar to the number of genes in the nematode *C. elegans*, has surprised biologists.
- How do humans (and other vertebrates) get by with no more genes than a nematode?
- Vertebrate genomes use extensive alternative splicing of RNA transcripts.
  - This process generates more than one functional protein from a single gene.
- Nearly all human genes contain multiple exons, and an estimated 75% of these multi-exon genes are spliced in at least two different ways.
  - If each alternatively spliced human gene on average specifies three different polypeptides, then the total number of different human polypeptides is about 75,000.
  - Additional polypeptide diversity can result from post-translational modifications.

Gene densities vary.

- Gene density is the number of genes present in a given length of DNA.
- Generally, eukaryotes have larger genomes but lower gene density than prokaryotes.
- Humans have hundreds or thousands of times as many base pairs in their genome as most bacteria, but only 5–15 times as many genes—thus, the gene density is lower.
- 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, mammals such as humans seem to have the lowest gene density.
- In bacterial genomes, most of the DNA consists of genes for protein, tRNA, or rRNA.
  - Nontranscribed regulatory sequences, such as promoters, make up only a small amount of the DNA.
  - Bacterial genes lack introns.

- Most eukaryotic DNA does not code for protein and is not transcribed into functional RNA molecules (such as tRNAs).
- Eukaryotic DNA includes more complex regulatory sequences. In fact, humans have 10,000 times as much noncoding DNA as bacteria.
  - Some of the DNA in multicellular eukaryotes is present as introns within genes.
  - Introns account for most of the difference in average length between human genes (27,000 base pairs) and bacterial genes (1,000 base pairs).

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

- The coding regions of protein-coding 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.
- Most of a eukaryotic genome consists of DNA sequences that don't code for proteins or produce known RNAs; this noncoding DNA has often been described in the past as "junk DNA."
- Far from junk, however, this DNA plays important roles in the cell, explaining why it has persisted in diverse genomes over many hundreds of generations.
- Comparisons of the genomes of humans, rats, and mice revealed almost 500 regions of noncoding DNA that were identical in sequence in all three species.
  - These sequences are more highly conserved than protein-coding genes in these species, supporting the view that the noncoding regions play important roles.
- Only 1.5% of the human genome codes for proteins or produces rRNAs and tRNAs.
- Gene-related regulatory sequences and introns account for 24% of the human genome.
- The rest, located between functional genes, includes unique noncoding DNA such as gene fragments and **pseudogenes**, nonfunctional former genes that have accumulated mutations over a long time.
- Most intergenic DNA is **repetitive DNA**, sequences that are present in multiple copies in the genome.
- Three-quarters 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 can move from one location to another within the genome.

- 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 simply **transposable elements**.
- During *transposition*, a transposable element moves from one site in a cell's DNA to another by a recombination process.
  - Transposable elements were called "jumping genes," but in fact, they never completely detach from the cell's DNA.
  - The original and new DNA sites are brought together by DNA bending.
- The first evidence for transposable elements came from American geneticist Barbara McClintock's breeding experiments with Indian corn (maize) in the 1940s and 1950s.

- ○ As she tracked corn plants through multiple generations, McClintock found color changes in corn that could be explained only by the existence of movable genetic elements.
- ○ The elements moved into genes for kernel color, disrupting the genes so that they could no longer produce color.

- McClintock's discovery was met with great skepticism. Her work was validated many years later, however, when transposable elements were found in bacteria and microbial geneticists learned more about the molecular basis of transposition.
- Eukaryotic transposable elements are of two types: transposons and retrotransposons.
- **Transposons** move within a genome by means of a DNA intermediate 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**.**
- Most transposable elements are **retrotransposons**, which move by means of an RNA intermediate.
  - ○ Retrotransposons always leave a copy at the original site during transposition because they are initially transcribed into an RNA intermediate.
  - ○ To insert at another site, the RNA intermediate is first converted back to DNA by reverse transcriptase, an enzyme encoded in the retrotransposon itself.
  - ○ A cellular enzyme catalyzes insertion of the reverse-transcribed DNA at a new site.
- Retroviruses, which use reverse transcriptase to produce their DNA, may have evolved from retrotransposons.

Multiple copies of transposable elements and sequences related to them are scattered throughout eukaryotic genomes.

- Transposable elements and related sequences are usually hundreds to thousands of base pairs long, and the dispersed "copies" are similar but not identical.
  - ○ These sequences make up 25–50% of most mammalian genomes and even higher percentages of the genomes of amphibians and many plants.
- 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 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.
  - ○ Many *Alu* elements are transcribed into RNA molecules, although their cellular function, if any, is 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 about 6,500 base pairs long and have a slow rate of transposition.
  - ○ Sequences within L1 block RNA polymerase, which is necessary for transposition.
- L1 sequences have been found within the introns of nearly 80% of the human genes analyzed, suggesting that L1 may help regulate gene expression.
  - ○ Researchers have proposed that L1 retrotransposons may affect the genome differently in separate developing neurons, thus contributing to the great diversity of neuronal cell types.
- Although many transposable elements encode proteins, these proteins do not carry out normal cellular functions.

Some repetitive DNA is not related to transposable elements.

- Repetitive DNA that is not related to transposable elements probably arose by mistakes during DNA replication or recombination.
- Such DNA accounts for about 15% of the human genome.
- Another 5% of the human genome consists of duplications of long stretches of DNA, ranging from 10,000 to 300,000 base pairs.
  - The large segments have been copied from one chromosomal location to another site on the same or a different chromosome.
- **Simple-sequence DNA** contains many copies of tandemly repeated short sequences.
  - Repeated units may contain as many as 500 or as few as 15 nucleotides.
  - When the unit contains 2–5 nucleotides, the series of repeats is called a **short tandem repeat**, or **STR**.
- The number of copies of the repeated unit can vary from site to site within a given genome.
- The repeat number can also vary from person to person, producing the variation used for genetic profiling by STR analysis.
- Altogether, simple-sequence DNA makes up 3% of the human genome.
- Simple-sequence DNA has an intrinsically different density from the rest of the cell's DNA.
  - If genomic DNA is cut into pieces and centrifuged at high speed, segments of different density migrate to different positions in the centrifuge tube.
  - Repetitive DNA isolated in this way was originally called *satellite DNA* because it appeared as a "satellite" band in the centrifuge tube, separate from the rest of the DNA. This term is often used interchangeably with *simple-sequence DNA*.
- 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.
  - 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.
  - Telomeric DNA also binds proteins that protect the ends of a chromosome from degradation and from joining to other chromosomes.

Gene-related DNA makes up about 25% of the human genome.

- DNA sequences that code for proteins or produce tRNA or rRNA make up 1.5% of the human genome.
- If introns and regulatory sequences associated with genes are included, the total amount of gene-related DNA—coding and noncoding—constitutes about 25% of the human genome.
- Many eukaryotic genes are present as unique sequences, with only one copy per haploid set of chromosomes.
- In most eukaryotic genomes, such solitary genes make up less than half the total transcribed DNA.
- The rest of the transcribed DNA occurs in **multigene families**, collections of two or more identical or very similar genes.

- In multigene families consisting of *identical* DNA sequences, the sequences are usually clustered tandemly.
- Except for genes that code for histone proteins, these families have RNAs as final products.
- The genes for the three largest rRNA molecules are a family of identical DNA sequences.
- These rRNA molecules are transcribed from a single transcription unit, 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.
- These molecules are then combined 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.
    - 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, the embryonic and fetal forms of hemoglobin have a higher affinity for oxygen than the adult forms, thus ensuring the efficient transfer of oxygen from mother to fetus.

Concept 21.5 Duplication, rearrangement, and mutation of DNA contribute to 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.
- An accident in meiosis can result in one or more extra sets of chromosomes, a condition known as *polyploidy*.
- In rare cases, the polyploidy condition can facilitate the evolution of genes.
    - In a polyploid organism, one set of genes can provide essential functions for the organism.
    - The genes in the extra set may diverge by accumulating mutations.
    - These variations may persist if the organism carrying them survives and reproduces.
    - In this way, genes with novel functions may evolve.
- The accumulation of mutations may lead to the branching off of a new species, as happens often in plants.
- Scientists can compare the chromosomal organizations of many different species to make inferences about the evolutionary processes shaping chromosomes and possibly leading to speciation.
- Researchers performed a computer analysis of DNA sequences to reconstruct the evolutionary history of chromosomal rearrangements in eight mammalian species.

- The researchers found many duplications and inversions of large portions of chromosomes.
  - 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.
- Such 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.
  - Due to chromosome rearrangements, the two populations could not successfully mate with each other, a step on their way to becoming two separate species.
- After 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).
- Another pattern with medical relevance was noted: The chromosomal breakage points associated with the rearrangements were not randomly distributed; specific sites were used over and over again.
  - A number of these recombination "hotspots" correspond to locations of chromosomal rearrangements within the human genome that are associated with congenital diseases.
- Errors during meiosis can lead to the duplication of smaller chromosomal regions, including segments that are about the length of individual genes.
- Unequal crossing over during prophase I can result in one chromosome with a deletion and another with a duplication of a particular gene.
- Transposable elements in the genome can provide sites where nonsister chromatids can cross over, even when their homologous gene sequences are not correctly aligned.
- Slippage during DNA replication can result in the deletion or duplication of DNA regions.
  - Such errors can lead to regions of repeats, such as simple-sequence DNA.
- Evidence that unequal crossing over and template slippage during DNA replication lead to duplication of genes is found in the existence of multigene families.
- Duplication events have led to the evolution of genes with related functions, such as the α-globin and β-globin gene families.
  - A comparison of gene sequences within a multigene family indicates that they all evolved from one common ancestral globin gene, which was duplicated and diverged about 450–500 million years ago.
  - Each of these genes was later duplicated several times, and the copies then diverged from each other in sequence, yielding the current family members.
  - The 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 family arose from mutations that accumulated in the gene copies over many generations.
  - The necessary function provided by an α-globin protein was fulfilled by one gene, while other copies of the α-globin gene accumulated random mutations.
  - Some mutations may have altered the function of the protein product in ways that were beneficial to the organism without changing its oxygen-carrying function.

- The similarity in the amino acid sequences of the various α-globin and β-globin proteins supports this model of gene duplication and mutation.
  - The existence of several pseudogenes among the functional globin genes provides additional evidence for this model.
  - Random mutations accumulating over time in the pseudogenes have destroyed their function.
- In other gene families, 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 prevent infection by hydrolyzing bacterial cell walls; α-lactalbumin is a nonenzymatic protein that plays a role in mammalian milk production.
- Both genes are found in mammals, but only lysozyme is found in birds.
  - The two proteins are similar in their amino acids sequences and three-dimensional structures.
- Findings suggest that at some time after the bird and mammalian lineages had separated, the lysozyme gene underwent a duplication event 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.
- Rearrangement of existing DNA sequences within genes has also contributed to genome evolution.
  - The presence of introns in eukaryotic genes may have promoted the evolution of new and potentially useful proteins by facilitating the duplication or repositioning of exons in the genome.
  - A particular exon within a gene could be duplicated on one chromosome and deleted from the homologous chromosome.
  - The gene with the duplicated exon would code for a protein with a second copy of the encoded domain.
  - This change in the protein's structure could augment its function by increasing its stability or altering its ability to bind a particular ligand.
- A number of protein-coding genes have multiple copies of related exons, which presumably arose by duplication and then diverged.
- The gene coding for 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.
- The mixing and matching of different exons within or between genes owing to errors in meiotic recombination is called *exon shuffling* and could lead to new proteins with novel combinations of functions.
  - The gene for tissue plasminogen activator (TPA), an extracellular protein that helps control blood clotting, 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.
- The persistence of transposable elements as a large percentage of eukaryotic genomes suggests that they play an important role in shaping a genome over evolutionary time.

- Transposable elements can contribute to the evolution of the genome by promoting recombination, disrupting cellular genes or control elements, and carrying entire genes or individual exons to new locations.
- The presence of transposable elements with similar sequence scattered throughout the genome allows recombination to take place between different chromosomes with homologous regions.
  - Most of these alterations are likely detrimental, causing chromosomal translocations and other changes in the genome that may be lethal to the organism.
  - Over the course of evolutionary time, however, an occasional recombination may be advantageous.
- The movement of transposable elements around the genome can have direct consequences.
  - If a transposable element "jumps" into the middle of a coding sequence of a protein-coding gene, it may prevent the normal functioning of that gene.
  - If a transposable element inserts within a regulatory sequence, it may increase or decrease protein production.
- During transposition, a transposable element may transfer genes to a new position on the genome.
  - This process probably accounts for the location of the α-globin and β-globin gene families on different human chromosomes.
- A similar mechanism may insert an exon from one gene into another 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.
- Transposable elements can lead to new coding sequences when an *Alu* element hops into introns to create a weak alternative splice site in the RNA transcript.
  - Splicing usually occurs at the regular splice sites, producing the original protein.
  - Occasionally, splicing occurs at the new weak site.
  - In this way, alternative genetic combinations can be "tried out" while the function of the original gene product is retained.
- These processes produce either no effect or harmful effects in most individual cases.
- Over long periods of time, however, the generation of genetic diversity provides more raw material for natural selection to work on during evolution.
- The accumulation of changes in the genome of each species provides a record of its evolutionary history.
- Comparing the genomes of different species enables scientists to identify genomic changes and has increased our understanding of how genomes evolve.

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

Comparisons of genome sequences from different species tell about the evolutionary history of life.

- The more similar in sequence the genes and genomes of two species, the more closely related those species are in their evolutionary history.

- Comparing the genomes of closely related species provides information about recent evolutionary events; comparing the genomes of distantly related species sheds light on ancient evolutionary history.
- Analyzing *highly conserved* genes in distantly related species can help clarify evolutionary relationships among species that diverged long ago.
  - Comparisons of the complete genome sequences of bacteria, archaea, and eukaryotes strongly support the theory that these groups are the three fundamental domains of life.
- Genes that evolved a very long time ago can still be surprisingly similar in disparate species.
  - Several protein-coding genes in yeast are so similar to some human disease genes that researchers deduced the functions of the disease genes by studying their yeast counterparts.
- The genomes of closely related species are likely to be organized similarly because of their relatively recent divergence.
  - The fully sequenced genome of one species can thus be used as a scaffold for assembly of genomic sequences from a closely related species, accelerating mapping of the second genome.
  - For instance, using the human genome sequence as a guide, researchers were able to sequence the mouse genome very quickly.
- The genetic differences between closely related species can be correlated with phenotypic differences.
- Researchers have compared the human genome with the genomes of the chimpanzee, mouse, rat, and other mammals.
- Identifying the genes shared by these species but not by nonmammals provides clues about what it takes to make a mammal. Identifying the genes shared by chimpanzees and humans but not by rodents gives information about primates.
- Comparing the human genome with that of the chimpanzee helps answer the question: What genomic information makes a human or a chimpanzee?
- In single-base substitutions, chimp and human genomes differ by only 1.2%.
- Longer stretches of DNA show a 2.7% difference due to insertions or deletions of larger regions in the genome of one or the other species.
  - Many of the insertions are duplications or other repetitive DNA.
- A third of the human duplications are not present in the chimpanzee genome.
  - 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.
- Biologists have identified a number of genes that are apparently evolving faster in humans than in either the chimpanzee or the mouse.
  - These include genes involved in defense against malaria and tuberculosis and at least one gene regulating brain size.
  - The genes evolving the fastest in humans are those that code for transcription factors.
  - 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 this gene functions in vocalization in vertebrates.

- ○ Mutations in this gene can produce severe speech and language impairment in humans.
- ○ The *FOXP2* gene is expressed in the brains of zebra finches and canaries at the time when these songbirds are learning their songs.

- In a "knock-out" experiment, Joseph Buxbaum and colleagues disrupted the *FOXP2* gene in mice and analyzed the resulting phenotype.
- Homozygous mutant mice had malformed brains and failed to emit normal ultrasonic vocalizations, while mice with one faulty copy of the gene had significant problems with vocalization.
- Researchers are exploring whether differences between the human and chimpanzee FOXP2 proteins account for the ability of humans, but not chimpanzees, to communicate by speech.
  - ○ There are only two amino acid differences between the human and chimpanzee FOXP2 proteins; the effect of these differences on the function of the human protein is not yet known.
- Analysis of genomes is increasing our understanding 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 human genetic diversity seems to be in the form of single nucleotide polymorphisms (SNPs), usually detected by DNA sequencing.
  - ○ In the human genome, SNPs occur on average about once in 100–300 base pairs.
  - ○ Scientists have identified the location of several million SNP sites in the human genome.
- Other variations—including inversions, deletions, and duplications—seem to occur without ill effect on the individual carrying them.
- These variations will be useful genetic markers for studying human evolution, the differences between human populations, and the migratory routes of human populations throughout history.
- Polymorphisms in human DNA will also be valuable markers for identifying genes that cause diseases or affect our health in more subtle ways.
- Analysis of the differences in individual genomes is likely to change the practice of medicine in the 21st century.

Comparative studies of the genetic programs that direct embryonic development clarify the mechanisms that generated the great diversity of life.

- Biologists in the field of evolutionary developmental biology, or **evo-devo**, compare the developmental processes of multicellular organisms with the goal of understanding how these processes have evolved and how changes in them can modify existing organismal features or lead to new ones.
- The genomes of related species with strikingly different forms may have only minor differences in gene sequence or regulation.
- For example, homeotic genes in *Drosophila* specify the identity of body segments in the fruit fly.
- 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 their chromosomal arrangement.
- Homeobox-containing sequences have been found in regulatory genes of much more distantly related eukaryotes, including plants, yeasts, and even prokaryotes.
    - Clearly, 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 and do not directly control the identity of body parts.
    - Most of these genes 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*, in several segmentation genes, and in a master regulatory gene for eye development.
- The homeobox-encoded homeodomain is the part of a protein that binds to DNA when the protein functions as a transcriptional regulator.
    - However, the shape of the homeodomain allows it to bind to any DNA segment; by itself it cannot select a specific sequence.
    - Other, more variable domains in a homeodomain-containing protein determine which genes the protein regulates.
    - Interaction of these latter 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 *Drosophila* embryos, different combinations of homeobox genes are active in different parts of the embryo.
    - Selective expression of regulatory genes, varying over time and space, is central to pattern formation.
- Many other genes involved in development are highly conserved from species to species.
    - These include numerous genes that encode components of signaling pathways.
- How can the same genes be involved in the development of animals whose forms are so different?
- In some cases, small changes in the regulatory sequences of particular genes cause changes in gene expression patterns that can lead to major changes in body form.
    - The differing patterns of expression of the *Hox* genes along the body axis in insects and crustaceans explain the variation in the number of leg-bearing segments among these segmented animals.
    - The same *Hox* gene product may have different effects in different species, turning on new genes or turning on the same genes at higher or lower concentrations.
- Similar genes direct distinct developmental processes in specific organisms, resulting in different body shapes.
    - Several *Hox* genes 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.

Similarities in the molecular mechanisms of development in plants and animals reflect their shared cellular origin.

- The last common ancestor of animals and plants was 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 and do not show the morphogenetic movements of cells and tissues that are so important in animals.
    - Morphogenesis in plants relies primarily on differing planes of cell division and on selective cell enlargement.
- Despite the differences between animals and plants, there are similarities in their molecular mechanisms of development, which are legacies of their shared cellular 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.
    - Research on a small flowering plant in the mustard family, *Arabidopsis thaliana*, has shown that establishing the radial pattern of flower parts, like setting up the head-to-tail axis in *Drosophila*, uses a cascade of transcription factors.
- The genes that direct these processes differ considerably in plants and animals.
    - Many of the master regulatory switches in *Drosophila* are homeobox-containing *Hox* genes, whereas the switches in *Arabidopsis* belong to a completely different family of genes, called the *Mads-box* genes.
    - Although homeobox-containing genes can be found in plants and *Mads-box* genes in animals, the genes do not perform the same major roles in development in both groups.
- Thus, molecular evidence supports the supposition that developmental programs evolved separately in animals and plants.