Overview: Reading the leaves from the tree of life

- The chimpanzee genome was sequenced by 2005, two years after the sequencing of the human genome was completed.
- Comparing the genomes of bacteria, archaea, fungi, protists, and plants 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 need to deal with this volume of information has spawned the field of bioinformatics, the application of computational methods to the storage and analysis of biological data.

Concept 21.1 The Human Genome Project fostered development of faster, less expensive sequencing techniques

- In 1990, the Human Genome Project began the task of sequencing the human genome.
  - Organized by an international, publicly funded consortium of scientists at universities and research institutes, the project involved 20 large sequencing centers in six countries plus many labs working on small projects. The United States portion was headed by Dr. Francis Collins.

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

- The project proceeded through three stages that provided progressively more detailed views of the human genome: linkage mapping, physical mapping, and DNA sequencing.
- The ultimate goal in mapping any genome is to determine the complete nucleotide sequence of each chromosome.
- 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 May 1998, Venter set up a company, Celera Genomics, and declared his intention to complete the human genome sequence using this whole-genome shotgun approach.
- In April 2003, the human genome sequence was announced jointly by Celera and the public consortium.
  - Both approaches have made valuable contributions.
  - Today, the whole-genome shotgun method is widely used.
- The development of newer sequencing techniques, generally called sequencing by synthesis has increased the speed and decreased the cost of sequencing entire genomes.
  - In these new techniques, many very small fragments (fewer than 100 base pairs) are sequenced at the same time, and computer software rapidly assembles the complete sequence.

Concept 21.2 Scientists use bioinformatics to analyze genomes and their functions

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.
- The NCBI database of sequences is called Genbank.
  - As of July 2013, Genbank contained the sequences of 165 million fragments of genomic DNA, totaling 153 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.

• Two research institutions, Rutgers University and the University of California, San Diego, maintain a worldwide Protein Data Bank, a database of all known three-dimensional protein structures.

**Genes and their expression can be understood at the systems level.**

• 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**.

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

**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.

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

• The bulk of eukaryotic genomes consists of DNA sequences that don’t code for proteins or produce known RNAs. This noncoding DNA has been described as “junk DNA.”
  - Far from junk, this DNA plays important roles in the cell, explaining why it has persisted in diverse genomes over hundreds of generations.

• Only 1.5% of the human genome codes for proteins or produces rRNAs and tRNAs.
  - Gene-related regulatory sequences and introns account for, respectively 5% and 20% 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 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.

• 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.
  - In 1983, at the age of 81, McClintock received the Nobel Prize for her research.

• **Transposons** move within a genome by a “cut-and-paste” mechanism, which removes the element from the original site or a “copy-and-paste” mechanism, which leaves a copy behind.
Most transposable elements are **retrotransposons**, which move by an RNA intermediate.

- Retrotransposons always leave a copy at the original site during transposition because they are initially transcribed into an RNA intermediate.

- Retroviruses, which use reverse transcriptase to produce their DNA, may have evolved from retrotransposons.

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

- In most eukaryotic genomes, 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.

- 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, 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 21.6 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 fundamental domains of life.

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

- Biologists in the field of evolutionary developmental biology, or **evo-devo**, compare the developmental processes of multicellular organisms.
  - Their goal is to understand how these processes have evolved and how changes in them can modify existing organisinal features or lead to new ones.
- **Homeotic genes** in *Drosophila* specify the identity of body segments in the fruit fly.
  - Molecular analysis has shown that these genes all include a 180-nucleotide sequence called a **homeobox**, specifying 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 (therefore they are “highly conserved” in evolution.)**
  - The sequences are so similar between humans and fruit flies that one researcher has whimsically referred to flies as “little people with wings.”
- Homeotic genes in animals were named **Hox genes**, short for homeobox-containing genes, because homeotic genes were the first genes found to have this sequence.
  - Most of these genes are associated with development, suggesting their ancient and fundamental importance in that process.
- 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.