Clay Davis, 19-9-23

# The human genome, structures, elements and comparisons

*"I would be quite proud to have served on the committee that designed the E. coli genome. There is, however, no way that I would admit to serving on a committee that designed the human genome. Not even a university committee could botch something that badly."*
*(David Penny to Danny Graur)*

The human genome project (2.7 billion dollars over 13 years) made one thing very clear. The human genome is a mess - and our instruction manual has no instruction manual. Nonetheless, this is our starting point. What's in there? How is it organized?

Reading:

**MBoC(6th) Ch1**: *pgs 12-30,* **Ch4**: *pgs 194-196, 210-216,* **Ch14**: *pgs 802-805.*

*Lander et al. (2001) **Initial Sequencing and Analysis of the Human Genome**. Nature 409, 860-921.

Little (2005) **Structure and function of the human genome**. Genome Res, 15:1759-1766.

Nurk and the T2T gang (2022) **The complete sequence of a human genome**[1]. Science, 376: 44-53.

Rhie and the T2T gang (2023) **The complete sequence of a human Y chromosome**. Nature (in press).

In the HGISH topics we often move beyond the material provided in your text book or in any textbook. Additional material listed in the reading are posted online. Ones that are especially useful are marked with (*).

Need to know and understand

Basic material

Classes of DNA: highly repetitive, middle repetitive, single copy

**Nucleolus** -site of transcription of rRNA genes

**Euchromatin**, **Heterochromatin**

ploidy (N) and C-value (C). Number of different copies of chromosomes - (N), multiples of haploid DNA content (C)

sequence homology, homolog, paralog, ortholog

Chromosome structure

Human chromosomes

22 Autosomes - 2 copies each in the human diploid phase

The X and Y sex chromosomes - The Y is special, male-only and one copy. It is very, very different than all the other chromosomes.

telomere

centromere

Autonomously replicating sequences (ARS = ori)

gene families

highly repetitive DNA, centromeric heterochromatin, satellite DNAs

middle repetitive DNA, interspersed repetitive DNA, LINES (L1), SINES (Alu)

[1] Perhaps true, but just a touch disingenuous; they sequenced a woman's genome. The Y is the hardest nut of all.

Minisatellite DNAs = Variable number Tandem Repeats (VNTRs)
microsatellite DNA = dust

Mitochondrial chromosomes

Tiny, 16,570 bp in humans, circular, encode two rRNAs, 22 tRNAs, some of the proteins in the oxidative phosphorylation pathway.

Bacterial chromosomes

haploid
closed circle
nucleoid
Stuck in one or more places to cell membrane
present sometimes in several copies/bacteria

Information Content in the E.coli Genome

Fraction that is transcribed - approximately 90%
Number of protein coding genes - approximately 4,300
Fraction that is protein-coding sequence - approximately 88%

Plasmids

Episome
Tiny, 2-20 kbp
found in bacteria and yeast often contain genes conferring antibiotic resistance

Viral genomes

large variety of different viral genomes, including ssDNA or dsDNA or RNA, closed circle or linear, single chromosome in one or several copies, several chromosomes

Gene organization in eukaryotes vs. prokaryotes, fraction of genome that encodes proteins in each.

Information Content in the Human Genome

Fraction that is transcribed - unclear, perhaps as high as 40%
Number of protein-coding genes - approximately 20,000
Fraction that is protein-coding sequence - approximately 1-2%

Focus illustrations: Source: Molecular Biology of the Cell: Sixth Edition (2015) Alberts et al., Garland Science, NY

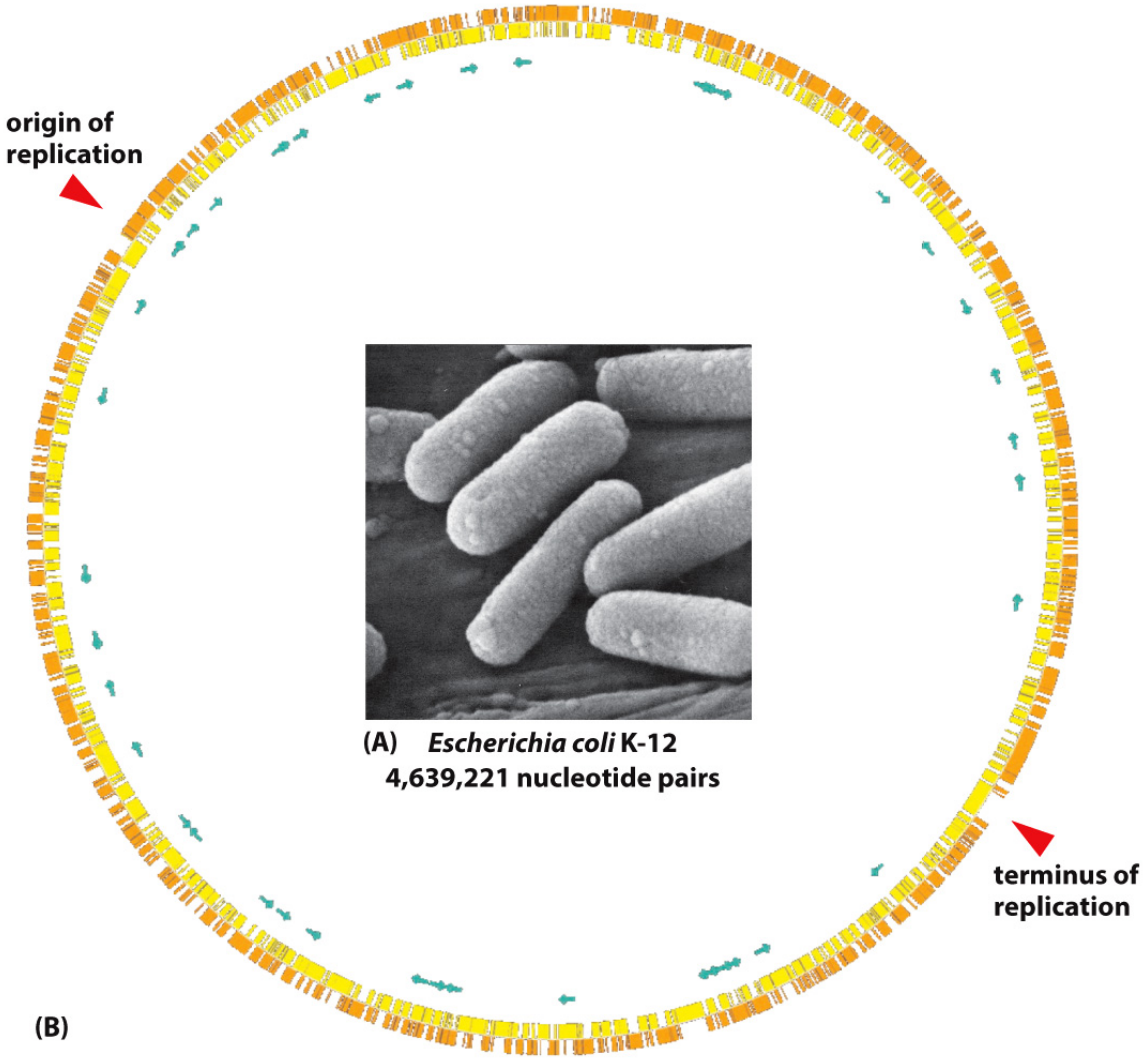


Figure 1-24 Molecular Biology of the Cell 6e (© Garland Science 2015)

Figure 1–24 The genome of E. coli. (A) A cluster of *E. coli* cells. (B) A diagram of the genome of E. coli strain k-12. The diagram is circular because the DNA of *E. coli*, like that of other prokaryotes, forms a single, closed loop. Protein-coding genes are shown as yellow or orange bars, depending on the DNA strand from which they are transcribed; genes encoding only RNA molecules are indicated by green arrows. some genes are transcribed from one strand of the DNA double helix (in a clockwise direction in this diagram), others from the other strand (counterclockwise). (A, courtesy of dr. Tony Brain and david Parker/Photo Researchers; B, adapted from F.R. Blattner et al., Science 277:1453–1462, 1997.)

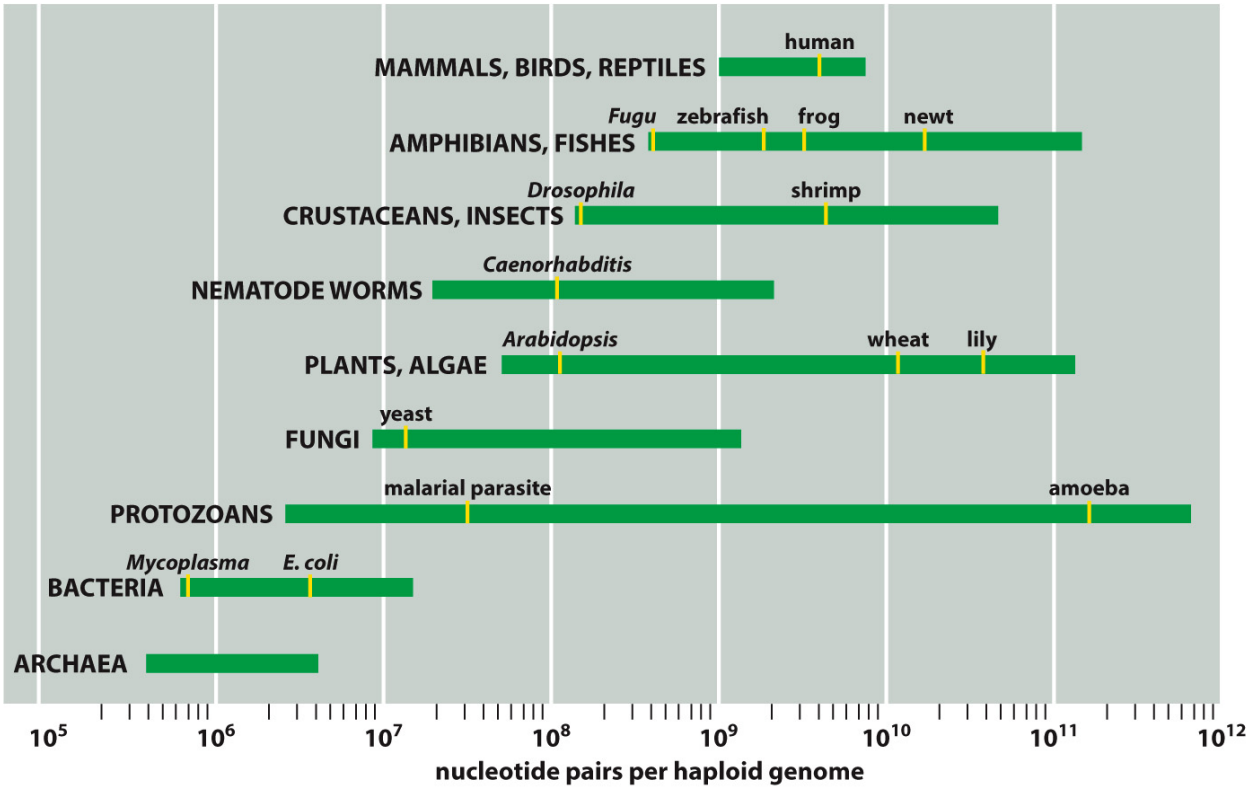


Figure 1-32 Molecular Biology of the Cell 6e (© Garland Science 2015)

Figure 1–32 Genome sizes compared. Genome size is measured in nucleotide pairs of DNA per haploid genome, that is, per single copy of the genome. (The cells of sexually reproducing organisms such as ourselves are generally diploid: they contain two copies of the genome, one inherited from the mother, the other from the father.) Closely related organisms can vary widely in the quantity of DNA in their genomes, even though they contain similar numbers of functionally distinct genes. (data from W.H. Li, molecular evolution, pp. 380–383. Sunderland, MA: Sinauer, 1997.)

**TABLE 1–2 Some Model Organisms and Their Genomes**

| Organism | Genome size* (nucleotide pairs) | Approximate number of genes |
|---|---|---|
| *Escherichia coli* (bacterium) | $4.6 \times 10^6$ | 4300 |
| *Saccharomyces cerevisiae* (yeast) | $13 \times 10^6$ | 6600 |
| *Caenorhabditis elegans* (roundworm) | $130 \times 10^6$ | 21,000 |
| *Arabidopsis thaliana* (plant) | $220 \times 10^6$ | 29,000 |
| *Drosophila melanogaster* (fruit fly) | $200 \times 10^6$ | 15,000 |
| *Danio rerio* (zebrafish) | $1400 \times 10^6$ | 32,000 |
| *Mus musculus* (mouse) | $2800 \times 10^6$ | 30,000 |
| *Homo sapiens* (human) | $3200 \times 10^6$ | 30,000 |

*Genome size includes an estimate for the amount of highly repeated DNA sequence not in genome databases.

Table 1-2 Molecular Biology of the Cell 6e (© Garland Science 2015)

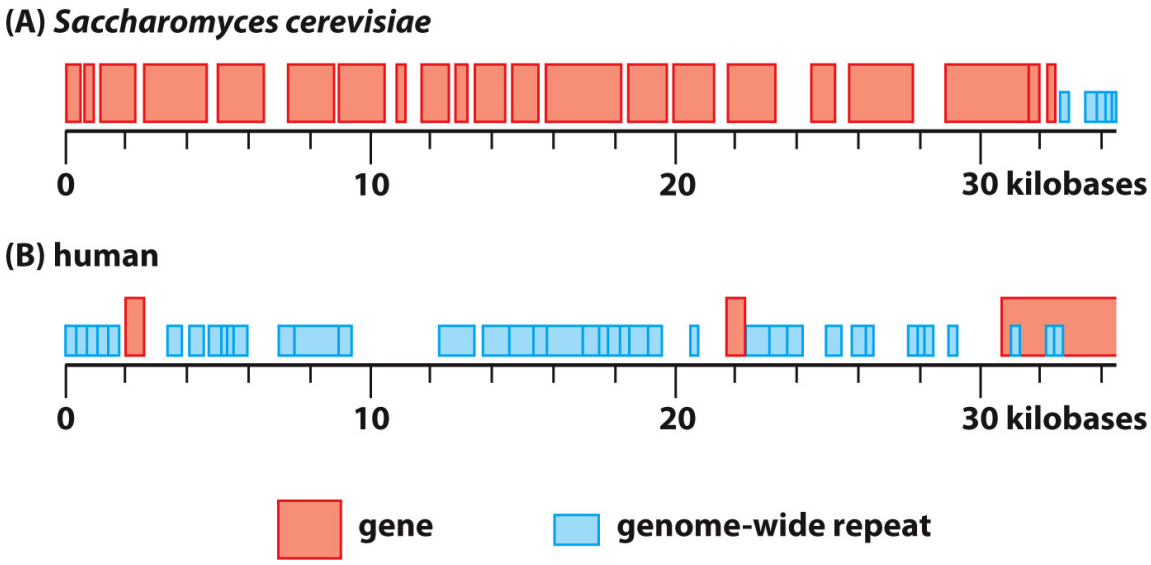


Figure 4-13 Molecular Biology of the Cell 6e (© Garland Science 2015)

Figure 4–13 The arrangement of genes in the genome of *S. cerevisiae* compared to humans. (A) *S. cerevisiae* is a budding yeast widely used for brewing and baking. The genome of this single-celled eukaryote is distributed over 16 chromosomes. A small region of one chromosome has been arbitrarily selected to show its high density of genes. (b) A region of the human genome of equal length to the yeast segment in (A). The human genes are much less densely packed and the amount of interspersed DNA sequence is far greater. Not shown in this sample of human DNA is the fact that most human genes are much longer than yeast genes (see Figure 4–15).

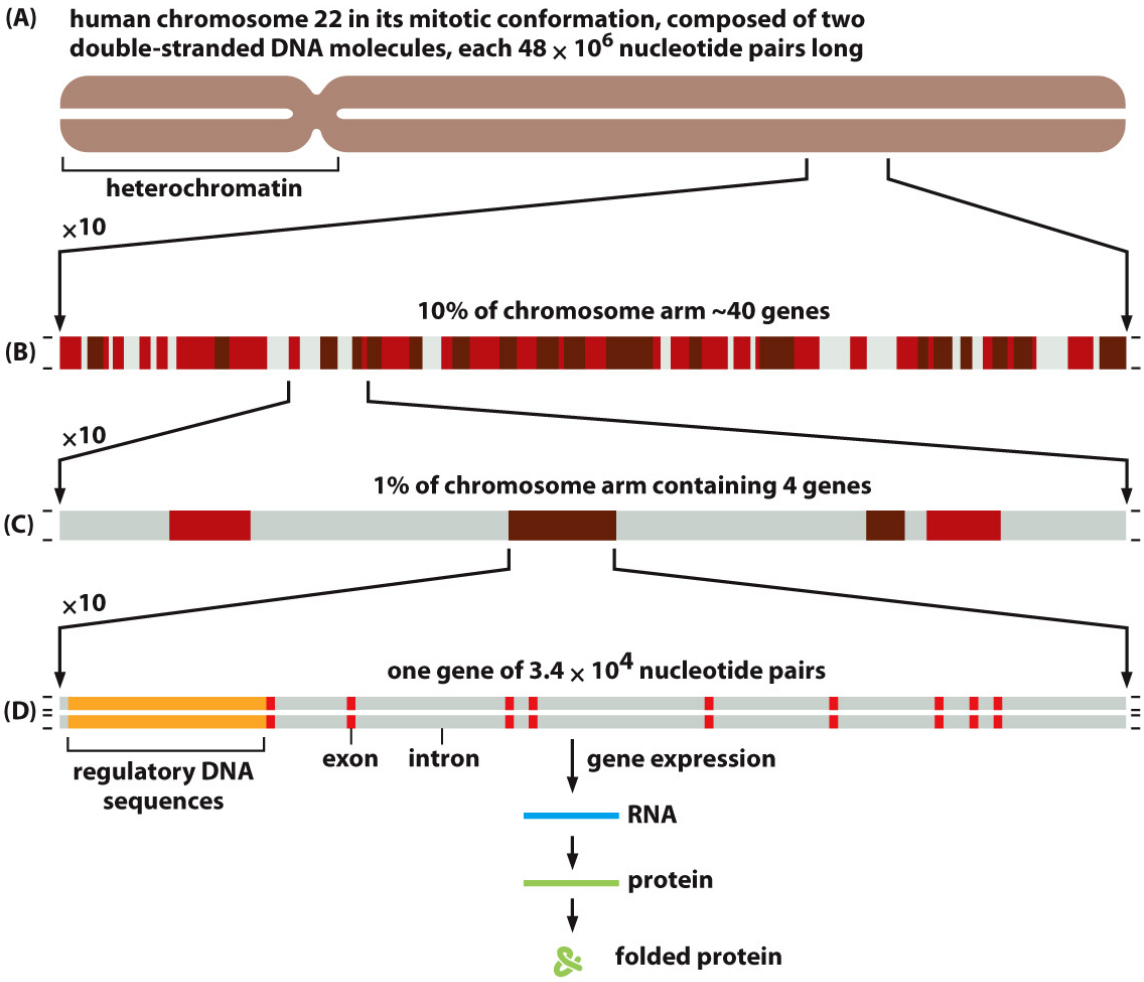


Figure 4-15 Molecular Biology of the Cell 6e (© Garland Science 2015)

Figure 4–15 The organization of genes on a human chromosome. (A) Chromosome 22, one of the smallest human chromosomes, contains 48 × $10^6$ nucleotide pairs and makes up approximately 1.5% of the human genome. Most of the left arm of chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin) discussed later in this chapter. (b) A tenfold expansion of a portion of chromosome 22, with about 40 genes indicated. Those in dark brown are known genes and those in red are predicted genes. (C) An expanded portion of (b) showing four genes. (D) The intron–exon arrangement of a typical gene is shown after a further tenfold expansion. Each exon (red) codes for a portion of the protein, while the DNA sequence of the introns (gray) is relatively unimportant, as discussed in detail in Chapter 6.

The human genome (3.2 × $10^9$ nucleotide pairs) is the totality of genetic information belonging to our species. Almost all of this genome is distributed over the 22 different autosomes and 2 sex chromosomes (see Figures 4–10 and 4–11) found within the nucleus. A minute fraction of the human genome (16,569 nucleotide pairs—in multiple copies per cell) is found in the mitochondria (introduced in Chapter 1, and discussed in detail in Chapter 14). The term human genome sequence refers to the complete nucleotide sequence of DNA in the 24 nuclear chromosomes and the mitochondria. being diploid, a human somatic cell nucleus contains roughly twice the haploid amount of DNA, or 6.4 × $10^9$ nucleotide pairs, when not duplicating its chromosomes in preparation for division. (Adapted from International Human Genome Sequencing Consortium, Nature 409:860–921, 2001. With permission from Macmillan Publishers Ltd.)

| TABLE 4–1 Some Vital Statistics for the Human Genome | |
|---|---|
| **Human genome** | |
| DNA length | $3.2 \times 10^9$ nucleotide pairs* |
| Number of genes coding for proteins | Approximately 21,000 |
| Largest gene coding for protein | $2.4 \times 10^6$ nucleotide pairs |
| Mean size for protein-coding genes | 27,000 nucleotide pairs |
| Smallest number of exons per gene | 1 |
| Largest number of exons per gene | 178 |
| Mean number of exons per gene | 10.4 |
| Largest exon size | 17,106 nucleotide pairs |
| Mean exon size | 145 nucleotide pairs |
| Number of noncoding RNA genes | Approximately 9000** |
| Number of pseudogenes*** | More than 20,000 |
| Percentage of DNA sequence in exons (protein-coding sequences) | 1.5% |
| Percentage of DNA in other highly conserved sequences**** | 3.5% |
| Percentage of DNA in high-copy-number repetitive elements | Approximately 50% |

* The sequence of 2.85 billion nucleotides is known precisely (error rate of only about 1 in 100,000 nucleotides). The remaining DNA primarily consists of short sequences that are tandemly repeated many times over, with repeat numbers differing from one individual to the next. These highly repetitive blocks are hard to sequence accurately.
** This number is only a very rough estimate.
*** A pseudogene is a DNA sequence closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression or function. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.
**** These conserved functional regions include DNA encoding 5′ and 3′ UTRs (untranslated regions of mRNA), DNA specifying structural and functional RNAs, and DNA with conserved protein-binding sites.

Table 4-1 Molecular Biology of the Cell 6e (© Garland Science 2015)

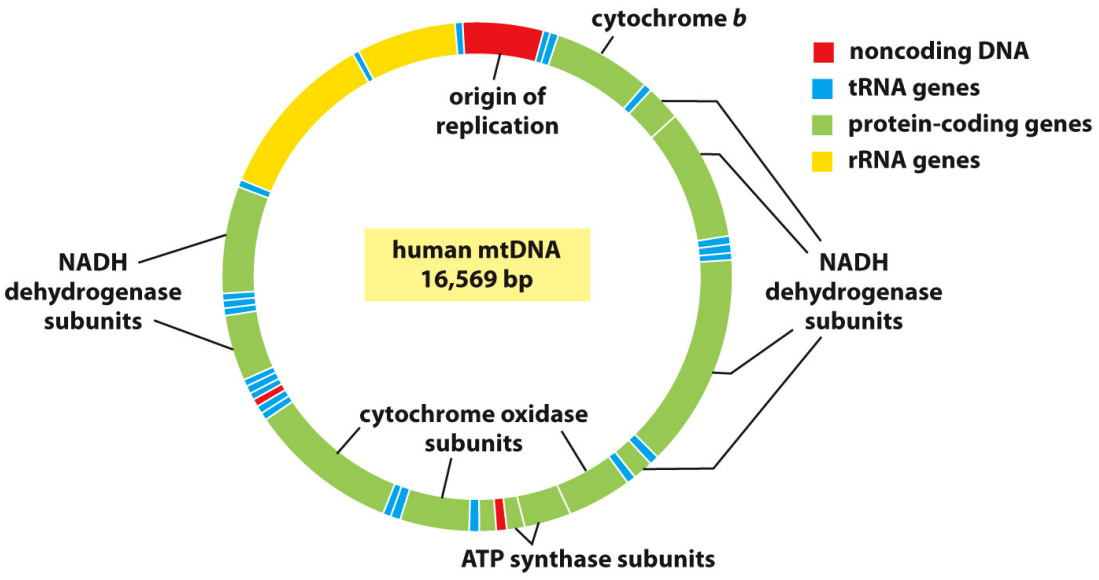


Figure 14-65 Molecular Biology of the Cell 6e (© Garland Science 2015)

Figure 14–65 The organization of the human mitochondrial genome. The human mitochondrial genome of ≈16,600 nucleotide pairs contains 2 rRNA genes, 22 tRNA genes, and 13 protein-coding sequences. There are two transcriptional promoters, one for each strand of the mitochondrial DNA (mtDNA). The DNAs of many other animal mitochondrial genomes have been completely sequenced. Most of these animal mitochondrial DNAs encode precisely the same genes as humans, with the gene order being identical for animals ranging from fish to mammals.