Eukaryotic Genomes Complexity (Outline)

- Compare and contrast size and complexity of prokaryotic and eukaryotic genomes
- Eukaryotic genomes vary in size, number of genes, and gene density
- Two basic mechanisms for generating larger complex genomes with repeat sequences
- Transposition and retroposition in eukaryotic cells
- Single copy, repetitive DNA sequences, and pseudogenes in multicellular eukaryotic genomes
- Human Genome composition
Prokaryotic & Eukaryotic Genomes

Prokaryotes
- Small genomes
- Mostly DNA coding for protein or functional RNA (tRNA and rRNA)
- Small amount of non-coding DNA, primarily regulatory sequences.

Eukaryotes
- Large genomes
- Minority of the DNA sequences code for protein or functional RNA
- Most of the genomic DNA is non-coding
Eukaryotic Genomes vary in size, number of genes, and gene density

- Over 4,300 genomes sequenced (4,000 bacteria, 186 archaea, and 183 eukaryotes)

- Sequencing of over 9,600 genomes and over 370 **metagenomes*** is currently in progress

(*) Metagenomics analysis of microbial DNA that is extracted directly from communities in environmental samples.
Genome Size

• Genomes of most bacteria and archaea: 1 to 6 million base pairs (Mb); genomes of eukaryotes are usually larger

• Eukaryotes have larger genomes:
  – Most plants and animals have genomes greater than 100 Mb; humans have 3,000 Mb

• No relationship between genome size and phenotype within the three domains
Gene Density and Noncoding DNA

• Mammals including humans have the lowest gene density (number of genes, in a given length of DNA)

• Multicellular eukaryotes have many introns within genes and a large amount of noncoding DNA between genes
### Table 21.1 Genome Sizes and Estimated Numbers of Genes*

<table>
<thead>
<tr>
<th>Organism</th>
<th>Haploid Genome Size (Mb)</th>
<th>Number of Genes</th>
<th>Genes per Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>1.8</td>
<td>1,700</td>
<td>940</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>4.6</td>
<td>4,400</td>
<td>950</td>
</tr>
<tr>
<td><strong>Archaea</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Archaeoglobus fulgidus</em></td>
<td>2.2</td>
<td>2,500</td>
<td>1,130</td>
</tr>
<tr>
<td><em>Methanosarcina barkeri</em></td>
<td>4.8</td>
<td>3,600</td>
<td>750</td>
</tr>
<tr>
<td><strong>Eukaryotes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em> (yeast, a fungus)</td>
<td>12</td>
<td>6,300</td>
<td>525</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em>  (nematode)</td>
<td>100</td>
<td>20,100</td>
<td>200</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em>    (mustard family plant)</td>
<td>120</td>
<td>27,000</td>
<td>225</td>
</tr>
<tr>
<td><em>Daphnia pulex</em>           (water flea)</td>
<td>200</td>
<td>31,000</td>
<td>155</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> (fruit fly)</td>
<td>165</td>
<td>14,000</td>
<td>85</td>
</tr>
<tr>
<td><em>Oryza sativa</em>            (rice)</td>
<td>430</td>
<td>42,000</td>
<td>98</td>
</tr>
<tr>
<td><em>Zea mays</em>                (corn)</td>
<td>2,300</td>
<td>32,000</td>
<td>14</td>
</tr>
<tr>
<td><em>Ailuropopoda melanoleuca</em> (giant panda)</td>
<td>2,400</td>
<td>21,000</td>
<td>9</td>
</tr>
<tr>
<td><em>Homo sapiens</em>            (human)</td>
<td>3,000</td>
<td>&lt;21,000</td>
<td>7</td>
</tr>
<tr>
<td><em>Paris japonica</em>          (Japanese canopy plant)</td>
<td>149,000</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

*Some values given here are likely to be revised as genome analysis continues. Mb = million base pairs. ND = not determined.
From simple to complex large genomes

Mechanisms of adding more sequence

• errors of replication
• insertion of foreign sequences
From simple to complex large genomes: DNA Replication Errors

Single copy sequence or gene

Gene Duplication

Independent mutations

Identical repeat sequences present in tandem

Homologous repeat sequences present in tandem
Identical Tandem Repeats

**Short** stretches
- within and outside genes

**Long** stretches
- Satellite DNA - structural role at telomeres and centromeres
From simple to complex large genomes:
Insertion of viral sequences

Separate genes

Insertion of Viral DNA

Viral DNA

Integrated viral sequences can interact with regulatory proteins

Integrated viral sequences can add coding sequence to genes and new proteins with new functions
Transposons and Retrotransposons in Eukaryotic genomes

- Two types of transposable elements:
  - **Transposons** - move by means of a DNA intermediate and require a transposase enzyme
  
  - **Retrotransposons**, which move by means of an RNA intermediate, using a reverse transcriptase
Figure 21.8

Transposon

DNA of genome

Transposon is copied

Mobile copy of transposon

New copy of transposon

Insertion
Figure 21.9

Retrotransposon → New copy of retrotransposon

Synthesis of a single-stranded RNA intermediate

RNA → Reverse transcriptase

DNA strand

Mobile copy of retrotransposon
Eukaryotic Genome Organization
(Coding Sequences-genes)

- Genes may be present once or in multiple copies per genome
- Families of genes present as tandem or interspersed repeats
Origin of tandem repeat families in eukaryotic genomes

Single copy sequence or gene

Gene Duplication

Independent mutations

Family of Identical Tandem repeats (same function)

Family of Homologous Tandem repeats (related functions)
Origin of Interspersed Repeat Families in Eukaryotic Genomes

Single copy sequence

Gene Duplication

Transposition or retroposition followed by Independent mutations

Family of Homologous Interspersed repeats
Multicellular eukaryotes genomes

• 98.5% of the human genome does not code for proteins, rRNAs, or tRNAs

  – One quarter of the genome
    • regulatory sequences
    • introns
  – Three-fourths of the genome:
    • repetitive DNA of transposable elements and their related 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
Interspersed repetitive sequence elements

Several families including retroposons and Alu elements

• Members of Alu family are present within introns, coding regions of genes, or non-coding regions of genes