Genomes summary

1. >930 bacterial genomes sequenced.
3. 2-10 Mbases, 470 – 7,000 genes
4. Genomes of >200 eukaryotes (45 “higher”) sequenced.
5. Linear chromosomes
6. On average, ~50% of gene functions “known”.
7. Human genome: <40,000 genes code for >120,000 proteins.
   Large gene families (e.g. 500 protein kinases)
   98% of human DNA is noncoding.
   ~3% of human DNA = simple repeats (satellites, minisatellites, microsatellites)
   ~50% of DNA = mobile elements (DNA transposons, retrotransposons (LTR and nonLTR) & pseudogenes)

Bacterial genome sizes

<table>
<thead>
<tr>
<th>Predicted genes in bacterial species</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>470</td>
</tr>
<tr>
<td>Mycoplasma mycoides</td>
<td>985</td>
</tr>
<tr>
<td>E. coli</td>
<td>4,288</td>
</tr>
<tr>
<td>B. anthracis</td>
<td>5,508</td>
</tr>
<tr>
<td>P. aeruginosa</td>
<td>5,570</td>
</tr>
<tr>
<td>Mycobacterium leprae</td>
<td>1,604</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>3,995</td>
</tr>
</tbody>
</table>

+ ~930 sequenced microbial genomes
  (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi)

Small and large
### Genome sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
<th>Estimated Number of Genes</th>
<th>Genes per Mb*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. influenza</em> (bacterium)</td>
<td>1.8 Mb*</td>
<td>1,200</td>
<td>950</td>
</tr>
<tr>
<td><em>S. cerevisiae</em> (yeast)</td>
<td>12 Mb</td>
<td>6,000</td>
<td>500</td>
</tr>
<tr>
<td><em>C. elegans</em> (nematode)</td>
<td>97 Mb</td>
<td>19,000</td>
<td>200</td>
</tr>
<tr>
<td><em>A. thaliana</em> (plant)</td>
<td>100 Mb</td>
<td>28,000</td>
<td>200</td>
</tr>
<tr>
<td><em>D. melanogaster</em> (fruit fly)</td>
<td>180 Mb</td>
<td>13,000</td>
<td>100</td>
</tr>
<tr>
<td><em>H. sapiens</em> (human)</td>
<td>3,200 Mb</td>
<td>20,000–40,000</td>
<td>10</td>
</tr>
</tbody>
</table>

*Mb = million base pairs

**Gene density down in mammals**

**Bacterial genomes are circular and densely packed with genes - 1**

*E. coli.* Genes (circles 1 & 2).  
*B. anthracis.* Genes (circles 1 & 2).
Bacterial genomes are circular and densely packed with genes - 2

**M. tuberculosis** (4.41 MB).
Genes (circles 1 & 2).

**M. leprae** (4.41 MB).
Genes (circles 1 & 2),
1116 pseudogenes (circles 3 & 4).

Representative gene arrangements in 50 kb segments of yeast, fly and human DNA.

Few yeast genes contain introns (exons are blue). Genes above and below the line are transcribed in opposite directions.
Numbers and types of genes in different eukaryotes

About half the genes encode proteins of unknown function.

Human genome: <2% ORFs & 48% repeats

<table>
<thead>
<tr>
<th>Class</th>
<th>Length</th>
<th>Copy Number in Human Genome</th>
<th>Fraction of Human Genome, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuclear genes</td>
<td>Variable</td>
<td>1</td>
<td>0.0%</td>
</tr>
<tr>
<td>Unspliced or spliced</td>
<td>Variable</td>
<td>2–10 kb</td>
<td>0.0%</td>
</tr>
<tr>
<td>Transcribed/recorded</td>
<td>Variable</td>
<td>0.3–5 kb</td>
<td>0.0%</td>
</tr>
<tr>
<td>RNA transcripts</td>
<td>Variable</td>
<td>&gt;1 kb</td>
<td>0.0%</td>
</tr>
<tr>
<td>Small RNA sequences</td>
<td>Variable</td>
<td>110–220 bp</td>
<td>0.0%</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>Variable</td>
<td>1–10 bp</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

| 481 sequences >200 bp that are absolutely conserved in mouse. Large gene families (E.g. ~500 Ser/Thr protein kinases many Zn²⁺ fingers, etc.) |
Human genome: individuals 99.9% identical

For every 1000 people . . .

Sequencing revealed one major allele for most genes in populations.

Human populations have not been genetically isolated for very long (~2-3 M years).

Many variations have not had time to spread throughout populations.

Human genome: individuals 0.1% different

For every person . . .

Lots of variation!

\[ 3.2 \times 10^9 \text{ bp/genome} \times 0.001 \text{ changes/bp} = \]
Human genome: individuals 0.1% different

For every person . . .

Lots of variation!
$3.2 \times 10^9 \text{ bp/genome} \times 0.001 \text{ changes/bp} = 3.2 \times 10^6 \text{ changes/genome}$

Human genome: individuals 0.1% different

For every person . . .

Lots of variation!
$3.2 \times 10^9 \text{ bp/genome} \times 0.001 \text{ changes/bp} = 3.2 \times 10^6 \text{ changes/genome}$

Two major types of variation
SNPs
Repeated DNA - short to long repeats

Variations produce RFLPs (Restriction Fragment Length Polymorphisms)!
SNPs

Single Nucleotide Polymorphisms (Changes of a single base)

Some are neutral
Some alter gene function

Identifying SNPs
Phenotype (disease), e.g. Sickle cell anemia
Sequencing genes/cDNAs
Restriction digest

RFLPs

Restriction Fragment Length Polymorphisms (Changes of restriction enzyme sites)
RFLPs

Restriction Fragment Length Polymorphisms (Changes of restriction enzyme sites)

For every random $3 \times 10^6$ SNPs:
- $\sim 1/256$ will be in 4-base restriction sites
- $\sim 10^4$ RFLPs for EACH four-base cutter!
- $\sim 1/4096$ will be in 6-base restriction sites
- $\sim 7.5 \times 10^2$ RFLPs for EACH six-base cutter!

Lots of markers (RFLPs) to map genes by linkage to RFLPs

---

**Human genome: 48% repeats**

**Human genome:**
- <40,000 genes
- Average ~3 proteins/gene
- 95% of DNA is noncoding
- Individuals 99.9% identical
- (1 difference/1000 bp means many markers for mapping)

**Large families of repeats.**
- Satellites (micro, mini and conventional)
- Transposons
- Retrotransposons

---

**TABLE 10.1** Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

<table>
<thead>
<tr>
<th>Class</th>
<th>Length</th>
<th>Copy Number in Human Genome</th>
<th>Fraction of Human Genome, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supernumerous genes</td>
<td>Variable</td>
<td>1</td>
<td>$\sim 10^{-5}$ (0.07)</td>
</tr>
<tr>
<td>Untranslated regions</td>
<td>Variable</td>
<td>$1-10^3$</td>
<td>$\sim 10^{-13}$ (0.07)</td>
</tr>
<tr>
<td>Retrotransposons</td>
<td>Variable</td>
<td>$20-300$</td>
<td>0.3</td>
</tr>
<tr>
<td>Retrotransposons</td>
<td>Variable</td>
<td>$100-1000$</td>
<td>1</td>
</tr>
<tr>
<td>Transposons</td>
<td>Variable</td>
<td>$1000-100,000$</td>
<td>1.3</td>
</tr>
</tbody>
</table>

*Genomic repeats consist of noncoding DNA.
**Repeat-encoding genes for DMU-encoding genes are rare (<1000), but this number is not well known due to their many markers for identifying genes in the human genome sequence.*
Satellites

Microsatellites: 1 - 13 bps in ~150 bp arrays
Minisatellites: 15-100 bps in 1-5 kb arrays
Satellites: 14 - 500 bps in 20-100 kb arrays

Origins of length polymorphisms in simple-sequence repeats.

Generation of length differences by unequal crossing over in meiosis
“Southern” blotting detects DNA sequences by hybridization

1. Digest DNA using restriction enzyme(s)
2. Run gel
3. Transfer DNA from gel to (nitrocellulose) paper.
4. Denature DNA, hybridize probe DNA, and wash off excess probe.
5. Detect the probe on the paper. E.g. by autoradiography.

Different distributions of minisatellites

Three repeats (a, b, c) in 3 people (1, 2, 3)

Southern blot of Hinfl-digested DNA
RFLPs -- DNA "finger print" in a murder case

Southern blot of DNA samples digested with a restriction enzyme

Human genome: 48% repeats

Human genome:
<40,000 genes
Average ~3 proteins/gene
95% of DNA is noncoding
Individuals 99.9% identical
(1 difference/1000 bp means many markers for mapping).
Large families of repeats.
Satellites (micro, mini and conventional)
Transposons
Retrotransposons

<table>
<thead>
<tr>
<th>Class</th>
<th>Length</th>
<th>Copy Number</th>
<th>Human Genomic</th>
<th>Fraction of Human Genome, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding genes</td>
<td></td>
<td>Variable</td>
<td>1</td>
<td>~10^10 (0.07)</td>
</tr>
<tr>
<td>De novo genes</td>
<td></td>
<td>Variable</td>
<td>2000</td>
<td>~10^10 (0.07)</td>
</tr>
<tr>
<td>Randomly scattered genes</td>
<td></td>
<td>Variable</td>
<td>20-200</td>
<td>3</td>
</tr>
<tr>
<td>Regulator DNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable-region DNA</td>
<td>1-400 bp</td>
<td>Variable</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Transpositional DNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA transposons</td>
<td>2-4 kb</td>
<td>500,000</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>LTR transposons</td>
<td>&gt;100 kb</td>
<td>400,000</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Non-LTR transposons</td>
<td>~4 kb</td>
<td>400,000</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>SINE</td>
<td>2-100 bp</td>
<td>2,400,000</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Pseudogene</td>
<td>Variable</td>
<td>4-100</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Embedded spacer DNA</td>
<td>Variable</td>
<td>4-100</td>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>

*Copy number estimates include introns.
Numbers in parentheses refer to number of human proteins expected to be 4000-5000, but this number is uncertain. The number of human genes is probably lower than the number of protein-coding genes due to a number of transposable elements.

Two major classes of mobile elements

Proks and euks
DNA intermediate

Eukaryotes
RNA intermediate

Some consequences of repeat sequences in eukaryotes

**Genomic diversity** in individuals and species. The most common retrotransposon sequences in the human genome are derived from endogenous retroviruses (ERVs). Most of these >440,000 sequences consist only of isolated LTRs, which arise from recombination between the ends.

**Gene families** arise by duplication and divergence.

“**Pseudogenes**” arise from RT acting on mRNAs.

**New genes** arise by “exon shuffling”. 
Exon shuffling may create new proteins in eukaryotes

Mechanism 1: Recombination between homologous interspersed repeats in the introns of separate genes would produce a new combination of exons.

Mechanism 2:
- Transposition of an exon
  (a) DNA hopping of flanking transposons
  (b) Reverse transcription of a LINE RNA extending into the 3’ exon of gene 1 can produce a DNA that gives gene 2 a new 3’ exon upon integration.
Possible results of exon shuffling

1. Modular proteins (with alternate splicing patterns). E.g. Fibronectin gene and mRNA.

2. Separate proteins that form a complex in one organism are sometimes fused into a single polypeptide chain in another organism.

Genomes summary

1. >930 bacterial genomes sequenced.
3. 2-10 Mbases, 470 – 7,000 genes
4. Genomes of >200 eukaryotes (45 “higher”) sequenced.
5. Linear chromosomes
6. On average, ~50% of gene functions “known”.
7. Human genome: <40,000 genes code for >120,000 proteins.
   Large gene families (e.g. 500 protein kinases)
   98% of human DNA is noncoding.
   ~3% of human DNA = simple repeats (satellites, minisatellites, microsatellites)
   ~50% of DNA = mobile elements (DNA transposons, retrotransposons (LTR and nonLTR) & pseudogenes)
Model for DNA transposition in bacteria

Structure of a eukaryotic LTR retrotransposon

Two kinds: LINEs and SINEs

Long Interspersed Elements: encode proteins including RT
Short Interspersed Elements: deletion of protein-coding region

ORF1=RNA binding protein; ORF2=RT and endonuclease.
Experiments with yeast Ty elements demonstrated an RNA intermediate

Summary: Two major classes of mobile elements

Proks and euks
DNA intermediate

Eukaryotes
RNA intermediate
LINEs and SINEs