GENOMES
Learning objectives

• Describe how is DNA sequenced

• Describe the various features of genomes

• Explain what are model organisms

• Describe the complexities of eukaryotic genomes
DNA SEQUENCING
Why do we need DNA sequencing?

More specifically:

- What are the sequences of our genes?
- Where are our genes?
- What promoter sequences regulate our genes?
- What mutations lead to disease?
- Evolution of genomes/species
Strategies for DNA sequencing

• Only “short” DNA fragments (~500bp) can be sequenced.

• We need to chop our genomes into small fragments and reconstruct the whole sequence from these 500bp bits.

• Two strategies are mainly used:
  – Hierarchical sequencing
  – Shotgun sequencing
Shotgun Sequencing

- Cloned genomes
- Multiple genomes are sheared into variable sized segments
- Unordered sequenced segments
- Computational automated assembly
- Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing)
- Overlapping sequence segments combined to construct the genome consensus.
Hierarchical Sequencing

b)

- Cloned genomes
- Genome divided into large segments of known order.
- Ordered genome segments
- Multiple genome portions are sheared into variable sized segments
- Unordered sequenced segments
- Computational automated assembly
- Resulting overlapping sequenced segments. (The higher the coverage the better the quality of the sequencing.)
- Overlapping sequences segments combined to construct the genome consensus.
Frederick Sanger: modern’s biology daddy
Sanger method

**Tools for Investigating Life**

(A)

Deoxyribonucleoside triphosphate (dNTP) (normal)

Dideoxyribonucleoside triphosphate (ddNTP) (chemically modified)
Sanger method

a. Denatured Template

Add dNTPs and Polymerase

Template/Product

b. Denaturing Gel

Labelled Strands
Sanger method

Different-length strands can be lined up by size to determine DNA sequence.

Smaller fragments

5’ end

G
A
T
C

Larger fragments

3’ end

5’ CAACGACAATCC

3’ GTTGCTGTAGG

5’ Non-template DNA

3’ Template DNA

Figure 19-6c  Biological Science, 2/e
© 2005 Pearson Prentice Hall, Inc.
Sanger method

Template strand

5' ??????????????????CGCA
3' GCGT

3' Primer (sequence known)

5' AATCTGGGCTATTCGGCGT
3' GCGT

3' Electrophoresis

5' T?????????????????CGCA
3' ATCTGGGCTATTCGGCGT

3' Longest fragment

5' T?????????????????CGCA
3' ATCTGGGCTATTCGGCGT

5' Shortest fragment

5' TTAGACCCGATAAGCCCGCA
3'
High Throughput Sequencing

- Sanger method revolutionized biology.

- However, it is a slow, low throughput and inefficient method for sequencing large genomes.

- New methods have been developed that allow for massive amounts of DNA fragments to be sequenced in a short time.
High Throughput Sequencing

- Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)
- Polony sequencing
- **454 pyrosequencing**
- **Illumina (Solexa) sequencing**
- **SOLiD sequencing**
- Ion semiconductor sequencing
- DNA nanoball sequencing
- Helioscope(TM) single molecule sequencing
- Single Molecule SMRT(TM) sequencing
- Single Molecule real time (RNAP) sequencing
- Nanopore DNA sequencing
- VisiGen Biotechnologies approach

**DON’T LEARN THIS BY HEART ➔ JUST AN EXAMPLE!!!**
High Throughput Sequencing

- Single molecule?
- Massively parallel sequencing
- Short-read sequencers
- Microwell pyrosequencing
- Capillary sequencing
- Second-generation capillary sequencer
- First-generation capillary
- Automated slab gel
- Manual slab gel
- Gel-based systems

Kilobases per day per machine

Year

High Throughput Sequencing

Cost per Genome

Moore's Law

National Human Genome Research Institute

genome.gov/sequencingcosts
High Throughput Sequencing

TOOLS FOR INVESTIGATING LIFE

1. A large DNA molecule is cut into fragments of 300–800 bp and denatured to single strands.
2. Each single-stranded DNA fragment is attached to a microbead.
3. PCR amplifies each fragment to 2 million copies per bead.
4. Each bead is put into a microwell on a plate.
5. DNA sequencing is done one fluorescent base at a time and read by a laser scanner.
6. The sequence is analyzed by computer.
Cycle

Scan

Seq/spot

Cycle

Scan

Seq/spot
Recap!

- Sanger sequencing
- HTP sequencing
GENOMIC FEATURES
What have we learned so far

- Chromosome
- Centromere sequences
- Telomere sequences
- Histones
- DNA replication machinery
What have we learned so far

Epigenetic modification of gene: methylation

Promoter of transcription

Open reading frame (protein coding sequence)

mRNA

RNA polymerase

Terminator of transcription

Noncoding sequences

RNA genes

tRNA
If we look at prokaryotes...

• Sequencing of prokaryote genomes has allowed for many crucial discoveries relating to:
  – Diseases
  – Ecology
  – Evolution
  – Genomic features

• Sequencing of prokaryotes also allowed for the “creation” of new research fields, such as:
  – Functional genomics
  – Metagenomics
  – Comparative genomics
Transposons

(A) Transposable element

DNA

Copying and insertion

mRNA

If a transposable element is copied and inserted into the middle of another gene, the original gene is transcribed into an altered mRNA.

(B) Transposable element

Other genes

Transposable element

Transposon

A transposon consists of two transposable elements flanking another gene or genes. The entire transposon is copied and inserted as a unit.
Transposons

**DNA vs RNA Transposons**

- **Genomic DNA**
- **Transposon**

**DNA Transposons**

**RNA Transposons**
The book defines “functional genomics” as the **biological discipline that assigns functions to the products of genes** and shows this image:

On this map, colors denote specific gene functions. For example, red genes regulate cellular processes...

...yellow genes regulate replication...

...and green genes regulate the production of the cell wall.
Functional genomics

- Well, functional genomics also studies how these products (proteins) work together by identifying functional networks and groups
Metagenomics

Also referred to as "community genomics" or "environmental genomics", metagenomics “is the sequencing and analysis of DNA of microorganisms recovered from an environment, without the need for culturing them”.

Craig Venter
GLOBAL OCEAN SAMPLING EXPEDITION

http://www.jcvi.org/cms/research/projects/gos/overview/
Similarly, thanks to genome info...

Eukaryotic genomes are bigger, but...

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>HAPLOID GENOME SIZE (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td></td>
</tr>
<tr>
<td><em>M. genitalium</em></td>
<td>0.58</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>1.8</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4.6</td>
</tr>
<tr>
<td>Yeasts</td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>12.5</td>
</tr>
<tr>
<td><em>S. pombe</em></td>
<td>12.5</td>
</tr>
<tr>
<td>Plants</td>
<td></td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>115</td>
</tr>
<tr>
<td>Rice</td>
<td>390</td>
</tr>
<tr>
<td>Animals</td>
<td></td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>100</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>123</td>
</tr>
<tr>
<td>Pufferfish</td>
<td>342</td>
</tr>
<tr>
<td>Chicken</td>
<td>1,130</td>
</tr>
<tr>
<td>Human</td>
<td>3,300</td>
</tr>
</tbody>
</table>

Mb = millions of base pairs
... it doesn’t mean much!

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>HAPLOID GENOME SIZE (Mb)</th>
<th>NUMBER OF GENES</th>
<th>PROTEIN-CODING SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>M. genitalium</em></td>
<td>0.58</td>
<td>485</td>
<td>88%</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>1.8</td>
<td>1,738</td>
<td>89%</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4.6</td>
<td>4,377</td>
<td>88%</td>
</tr>
<tr>
<td>Yeasts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>12.5</td>
<td>5,770</td>
<td>70%</td>
</tr>
<tr>
<td><em>S. pombe</em></td>
<td>12.5</td>
<td>4,929</td>
<td>60%</td>
</tr>
<tr>
<td>Plants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>115</td>
<td>28,000</td>
<td>25%</td>
</tr>
<tr>
<td>Rice</td>
<td>390</td>
<td>37,544</td>
<td>12%</td>
</tr>
<tr>
<td>Animals</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>100</td>
<td>19,427</td>
<td>25%</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>123</td>
<td>13,379</td>
<td>13%</td>
</tr>
<tr>
<td>Pufferfish</td>
<td>342</td>
<td>27,918</td>
<td>10%</td>
</tr>
<tr>
<td>Chicken</td>
<td>1,130</td>
<td>25,000</td>
<td>3%</td>
</tr>
<tr>
<td>Human</td>
<td>3,300</td>
<td>24,000</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

Mb = millions of base pairs
### TABLE 17.2

Comparison of the Genomes of *E. coli* and Yeast

<table>
<thead>
<tr>
<th></th>
<th>E. COLI</th>
<th>YEAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome length (base pairs)</td>
<td>4,640,000</td>
<td>12,068,000</td>
</tr>
</tbody>
</table>
For example:

<table>
<thead>
<tr>
<th></th>
<th>E. COLI</th>
<th>YEAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome length (base pairs)</td>
<td>4,640,000</td>
<td>12,068,000</td>
</tr>
<tr>
<td>Number of protein-coding genes</td>
<td>4,290</td>
<td>5,770</td>
</tr>
<tr>
<td>Proteins with roles in:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>650</td>
<td>650</td>
</tr>
<tr>
<td>Energy production/storage</td>
<td>240</td>
<td>175</td>
</tr>
<tr>
<td>Membrane transport</td>
<td>280</td>
<td>250</td>
</tr>
<tr>
<td>DNA replication/repair/recombination</td>
<td>120</td>
<td>175</td>
</tr>
<tr>
<td>Transcription</td>
<td>230</td>
<td>400</td>
</tr>
<tr>
<td>Translation</td>
<td>180</td>
<td>350</td>
</tr>
<tr>
<td>Protein targeting/secretion</td>
<td>35</td>
<td>430</td>
</tr>
<tr>
<td>Cell structure</td>
<td>180</td>
<td>250</td>
</tr>
</tbody>
</table>
### Comparison of the Genomes of *E. coli* and Yeast

<table>
<thead>
<tr>
<th></th>
<th>E. COLI</th>
<th>YEAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome length (base pairs)</td>
<td>4,640,000</td>
<td>12,068,000</td>
</tr>
<tr>
<td>Number of protein-coding genes</td>
<td>4,290</td>
<td>5,770</td>
</tr>
<tr>
<td>Proteins with roles in:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>650</td>
<td>650</td>
</tr>
<tr>
<td>Energy production/storage</td>
<td>240</td>
<td>175</td>
</tr>
<tr>
<td>Membrane transport</td>
<td>280</td>
<td>250</td>
</tr>
<tr>
<td>DNA replication/repair/recombination</td>
<td>120</td>
<td>175</td>
</tr>
<tr>
<td>Transcription</td>
<td>230</td>
<td>400</td>
</tr>
<tr>
<td>Translation</td>
<td>180</td>
<td>350</td>
</tr>
<tr>
<td>Protein targeting/secretion</td>
<td>35</td>
<td>430</td>
</tr>
<tr>
<td>Cell structure</td>
<td>180</td>
<td>250</td>
</tr>
</tbody>
</table>

One “role” is really different between *E. coli* and yeast. Why?
Protein targeting/secretion

EUKARYOTIC CELL
- Nucleolus
- Nucleus
- Centriole
- Cytoplasm
- Mitochondrion
- Cell Membrane
- Golgi Complex
- Rough Endoplasmic Reticulum
- Smooth Endoplasmic Reticulum
- Nucleoid

FEROARYOTIC CELL
- Flagellum
- Cell Membrane
- Cytosol
- Cell Wall
- Ribs
- Capsule
Model organisms

A model organism is a non-human species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the organism model will provide insight into the workings of other organisms.

The “classical” model organisms are:
Yeast

- Unicellular
- Simple genome
- Simple to work with
- Allows for genetic tricks that other organisms can’t
- Usually discoveries made in yeast can be used as a guide for more complex organisms
- A lot of what we now in humans comes from experiments in yeast
C. elegans - worm

• Tiny worm
• Transparent → ideal for live microscopy experiments
• Has 998 cells, yet is a complete animal
• All cells can be followed from the initial fertilized egg:
  - ideal organism to study cellular communication and interactions
**D. melanogaster** – fruit fly

- The most widely studied animal in genetics.
- Most of what we know about genetics of multicellular organisms comes from the fly
- Crucial organism in studying “developmental biology”
A. thaliana

• The most widely studied plant.
• Easy to grow, easy to study
• A mix of yeast and fly for the plant kingdom
**M. musculus - mouse**

- Is the closest model to humans when studying highly complex processes such as:
  - disease: there are hundreds (or thousands) of mouse disease models
  - development of mammalian/skeletal organs (fe: eyes)
  - stem cells: easier to obtain from mice and less “ethical”
In general, model organisms:

- Model organisms are in vivo models and are widely used to research human disease, development and other biological processes, when human experimentation would be unfeasible or unethical.

- Model organisms are easy to "grow" and to "manipulate"
- Model organisms are cheap
- Model organisms are good generalisations/representations of their kingdom.
WHAT HAVE WE LEARNED FROM MODEL ORGANISMS?
Splice variants

Human PDE7A Gene
(124 kb)

PDE7A1 mRNA - 4.3 kb
(55-57 kDa protein)

PDE7A2 mRNA - 3.0 kb
(50-52 kDa protein)

PDE7A3 mRNA - 3.8 kb
(50 kDa protein)
Gene families

α-globin gene cluster

DNA

ζ₂ \quad \psi \zeta_1 \quad \psi \alpha_1 \quad \alpha_2 \quad \alpha_1

Nonfunctional pseudogenes

β-globin gene cluster

ε \quad G_\gamma \quad A_\gamma \quad \psi \beta_1 \quad \delta \quad \beta

Chromosome 16

Chromosome 11
Hemoglobins change during development

Stage of Development
Embryonic (<8 weeks)
Fetal (3-9 months)
Adult (from birth)

Human hemoglobins
\(\zeta_2\epsilon_2\), \(\zeta_2\gamma_2\), \(\alpha_2\epsilon_2\), \(\alpha_2\gamma_2\), \(\alpha_2\delta_2\), \(\alpha_2\beta_2\)
Comparative Genomics

- Genome-wide variation from one human being to another can be up to 0.5% (99.5% similarity)
- Chimpanzees are 96% to 98% similar to humans, depending on how it is calculated.
- Cats have 90% of homologous genes with humans, 82% with dogs, 80% with cows, 79% with chimpanzees, 69% with rats and 67% with mice.
- Cows (Bos taurus) are 80% genetically similar to humans
- 75% of mouse genes have equivalents in humans, 90% of the mouse genome could be lined up with a region on the human genome 99% of mouse genes turn out to have analogues in humans
- The fruit fly (Drosophila) shares about 60% of its DNA with humans.
- About 60% of chicken genes correspond to a similar human gene.
Comparative Genomics

- Increasing complexity
- Functions necessary for life
- Cell compartments
- Multicellularity
- Development

Vertebrates only 22%
Vertebrates and animals only 24%
All prokaryotes and eukaryotes 21%
Eukaryotes only 32%
Immune system
Nervous system