Strategies for the systematic sequencing of complex genomes
Genomics - is the study of the structure and function of the Genome of an Organism

**Structural Genomics** - the study of DNA sequence, chromatin structure, and DNA physical interactions in the cell

- Where are the dynamic elements located on the Genome, and what are they?

**Functional Genomics** - how does the structure of the genome:

- Give rise to particular cell types, tissues, organisms?
- Respond to environmental and developmental requirements?
- Become diseased?
Comparative Genomics compares the sequence of diverse genomes in order to determine the functions of their genes, mechanisms of regulation and evolution.
Genome sequencing efficiencies per person

- 1980 - 0.1-1 kb per year
- 1985 - 1-5 kb per year
- 1990 - 25-50 kb per year
- 1996 - 100-200 kb per year
- 2000 - 500-1000 kb per year

~350 kb/dia máquina

© Sam Ogden
Finished genomes or in progress

2001
45 BACTERIA
6 EUCARYA

2005
302 BACTERIA
57 EUCARYA

In progress
989
588
Examples of sequenced genomes

- Homo sapiens
- Mus musculus
- Fugu rubripes
- Arabidopsis thaliana
- Caenorhabditis elegans
- Drosophila melanogaster
- Schizosaccharomyces pombe
- Plasmodium falciparum
- Saccharomyces cerevisiae
- Dictyostelium discoideum
- Encephalitozoon cuniculi
Determining DNA Sequences

- **Sanger Method**
  - uses dideoxynucleoside triphosphates (ddNTP)
- **automated systems**
  - use dideoxynucleotides labeled with fluorescent dyes

![Dideoxynucleotide structure]

used as substrate by DNA polymerase

terminates DNA synthesis
Cycle Sequencing  
Chain Termination

Template

1 Primer + Taq Polymerase w/ Buffer + Cycles

Polymerization until Taq hits ddNTP, linear amplification.
Capillary-based sequencing instruments capable of analysing 96 samples in parallel.
Capillary electrophoresis

Filled with a polymer (resin)
The brown wire-like loops are the 96 individual capillary tubes. They are made of glass and coated with brown plastic. The samples are loaded from below this level on the left side and the DNA is electrophoresed towards the right.
A close up of a few of the 96 capillary tubes. They continue down below this level and are submerged into the 96 well plate so DNA can enter the tubes.
The scanner passes a laser light through a short section of each capillary tube where there is no plastic coating. The laser excites the fluorophores and the emitted light (four colors, one for each base) is detected by a photomultiplier tube.
500-800 bases per reaction
Various robotic systems have been designed to automate specific steps in the sequencing process:

• Systems that facilitate subclone library construction
• That do the picking and arraying of bacterial subclones
• That purify template DNA from individual subclones
• That prepare sequencing reactions and load the samples on slab gels or capillaries before electrophoresis

The estimated cost have been reduced >100-fold over the last past decade!
**Bacterial Artificial Chromosomes (BACs)**

- **F plasmid ancestry**
  - maintain bacterial replication system and copy number control system

- **Universal Priming Sites**
  - On the vector, flanking the genomic insert
  - Fragments from 100-400 kb can be cloned
Computer technology

Important software systems have been developed for analysing:

- **Primary sequence data**
  (calling the nucleotide base at each position and assigning a corresponding quality score that indicates that the base call is correct)
  
  **Ex:** Phred

- **For carrying out sequence assembly**
  
  **Ex:** Phrad, GAP and Consed (viewing of sequence assemblies)
Clone-by-Clone shotgun sequencing

Whole Genome Shotgun sequencing

- Construct clone map
- Select mapped clones
- Generate several thousand sequence reads per clone
- Assemble
Clone-by-clone shotgun: Map construction
Only the clones outlined in red were selected for sequencing.
Subclone library construction
To shotgun sequence 1000 bp, you'd need 8000 bp of sequence, or ~16 sequencing reactions.
Clone-by-clone shotgun sequencing

**Advantages**

- Availability of a map
- The assembly process is facilitated
- Allow sequenced fragments to be anchored to the map before the complete sequence is available
- Very successful on many big genomes

**Disadvantages**

- Cost
- Time
- The maps often have poor continuity
Whole genome shotgun

- developed in 1995 by J. Craig Venter and Hamilton Smith
Sequence assembly
Whole genome shotgun

- No up front large insert clone preparation
- No up front mapping
- Generates data immediately
- Very successful on many small genomes (bacteria)
- Difficulties seem to be repeat based, not size based
- Virtually all large scale projects have needed a clones based finishing step
- Dependent on advances in computer analysis and sequencing technologies
- Dependent on automated labor.
Hybrid approach
<table>
<thead>
<tr>
<th>Organism</th>
<th>Size (Mb)</th>
<th>General strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. elegans</td>
<td>~100</td>
<td>Clone-by clone</td>
</tr>
<tr>
<td>A. taliana</td>
<td>~125</td>
<td>Clone-by clone</td>
</tr>
<tr>
<td>Drosophila</td>
<td>~120</td>
<td>Hybrid</td>
</tr>
<tr>
<td>Human</td>
<td>~3200</td>
<td>Clone-by clone</td>
</tr>
<tr>
<td>Human</td>
<td>~3200</td>
<td>Whole shotgun</td>
</tr>
<tr>
<td>Mouse</td>
<td>~3200</td>
<td>Hybrid</td>
</tr>
<tr>
<td>Mouse</td>
<td>~3200</td>
<td>Whole shotgun</td>
</tr>
<tr>
<td>Rat</td>
<td>~3200</td>
<td>Hybrid</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>~1700</td>
<td>Hybrid</td>
</tr>
<tr>
<td>Pufferfish</td>
<td>~400</td>
<td>Whole shotgun</td>
</tr>
<tr>
<td>Rice</td>
<td>~435</td>
<td>Clone-by clone</td>
</tr>
</tbody>
</table>
Steps in current large-scale sequencing

- Subclone generation
- Production sequencing
  - Template isolation
  - Sequencing reactions
  - Fragment separation
  - Data acquisition
  - Base calling
- Finishing
  - Assembly
  - Gap filling
  - Verification
- Analysis
  - Gene predictions
  - Homology searches
  - Annotation
Finishing

- Assembly
- Finishing
  - Gap filling
  - Verification

**Gap**

**Single clone area**

**Multiple clone coverage on both strands**

**Gap**

**Multiple clone coverage on one strand only**
Gap closure strategies

End of contig1  Small insert  End of contig2

Sequence with forward primer  Sequence with reverse primer

Contig1  Large insert  Contig2

Sequence with forward primer  Sequence with reverse primer

Other genome already sequenced

End of contig1  End of contig2

Rare-cutting endonuclease site  Rare-cutting endonuclease site

Large genomic DNA fragment

Contig1  Contig2  Contig3
Analysis: Gene prediction

Why's it So Hard to Find Genes?

• Exons/Introns
• Alternative Splicing
• Alternate transcription/stop sites
Analysis: Homology searches and Annotation

- process that locates genes in the genome map
- identifies each open reading frame in genome
  - a reading frame > 100 codons that is not interrupted by a stop codon
- uses databases to assign tentative function of genes