Genomes and Metagenomes
Whole Genome Sequencing and Metagenomics

**Whole Genome Sequencing**

- Culture microbe
  - Extract DNA and Enzyme Digest
    - Shot-gun clone library
      - Randomly sequence clones
        - Fragment Analysis and Gap Closure
          - Editing and Annotation

**Metagenomics**

- Environmental Sample
  - Extract DNA and Enzyme Digest
    - Shot-gun clone library
      - Screen for genes, expression
        - Or randomly select clones
          - Sequence
            - Assign sequences to genomes
              - Edit and Annotate
Whole Genome Sequencing

Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

Shotgun sequence

Assembly
Microbial chromosome

Sonication

DNA fragments

Agarose gel electrophoresis of fragments and DNA size markers

Fragment purification from gel

DNA fragments

Clonal library preparation

Sequence the clonal inserts, particularly the end sequences.

Assembly of a Contig

Overlap

Clone A

Overlap

Clone C

Overlap

Construct sequence contigs and align using overlaps; fill in gaps.

Sequence the clonal inserts, particularly the end sequences.
Shot-gun Clone Libraries

1. Break DNA into pieces and purify
2. Ligate into plasmid, cosmid (30-45kb insert) vectors, or BAC (bacterial artificial chromosome)
   – Isolate vectors with only one insert
3. Transformed into competent *E.coli*
Sequencing

- Thousands of DNA fragments sequenced
- Automated
- Thousands of sequence “reads”
  - All parts of the genome are sequenced multiple times
    - Increases accuracy
    - Allows overlap to make alignment and assembly easier
Sequencing Technology

• Sanger method

• 454 Pyrosequencing

• Illumina sequencing
454 Pyrosequencing
Illumina

- Pyrosequencing technology
- Amplification takes place on strands on a plate instead of on a bead.
1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE-STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE

The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. IMAGE FIRST BASE

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.
Sequencing speed

- 454 and Illumina are faster than Sanger,
- Shorter reads, but many many more reads

Sequence information generated at JGI

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Total Q20* Bases (Billions)</th>
<th>Q20* Bases (Billions) by Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Goal</td>
<td>Actual Total</td>
</tr>
<tr>
<td>Q1 2009</td>
<td>39.9</td>
<td>124.21</td>
</tr>
<tr>
<td>Q2 2009</td>
<td>60.1</td>
<td>196.829</td>
</tr>
<tr>
<td>Q3 2009</td>
<td>71.2</td>
<td>321.039</td>
</tr>
</tbody>
</table>
Fragment Analysis

- Overlapping sequences are lined up and put in order
- Computer assisted
- Assemble **contigs** – continuous nucleotide sequences (when fragments with the same sequence overlap)
- **Contigs** are assembled in the correct order
  - by overlapping the end sequences from different contigs
- Fill in gaps
Assembly of a Contig

Clone A

Clone C

Overlap

Overlap

Construct sequence contigs and align using overlaps; fill in gaps.
• Shotgun library creation can be likened to taking the text from 100 copies of an unknown book and randomly cutting that text at various points in each of the copies.

• Fragment analysis is putting it back together so you have the complete text of the book.
Annotation

• Identify the protein-coding regions, rRNA and tRNA genes

• Open Reading Frame (ORF) – putative gene
  – At least 100 codons that
    • Are not interrupted by a stop codon
    • Apparent ribosomal binding site at 5’ end
    • Terminator sequence at 3’ end

• ORFs compared to known genes in databases
  – Can tentatively identify function of gene

• No genome has more than 80% of ORFs identified
Whole Genome Sequencing

• 1\textsuperscript{st} completed genome *Haemophilus influenzae*, 1995

• Genomes on-line database (GOLD)
  
  [www.genomesonline.org](http://www.genomesonline.org)
  
  – 762 completed genomes,

  – Ongoing Projects
    • 89 Archaea genomes
    • 1749 Bacterial genomes
    • 935 Eukarya genomes

  – Searchable database
• Sequencing centers world wide
  – J. Craig Venter Institute
  – U.S. Dept. of Energy Joint Genome Institute
    • Environmental organisms
    • GEBA project
  – Wellcome Trust Sanger Center (UK)
    • Pathogens
  – Celera Genomics
    • Human Genome
Whole Genome Sequencing

• Related technologies
  – Microarrays – Gene expression
    • Put known genes on a chip, add mRNA or cDNA from organism
    • See where they match, shows which genes are expressed under experimental conditions
  – Proteomics
    • Studies protein expression
Whole Genome Sequencing

• Discover benefits and applications in:
  – Medicine – new pharmaceuticals, virulence factors
    • How antibiotic resistance genes are shared
  – Bioremediation – catabolic pathways
    • *Anthrax* genomes
  – Industrial processes – new biocatalytic enzymes
  – Biosecurity – disease detection
  – Evolution – horizontal gene transfer

- Genomics:GTL, Dept. of Energy
Anthrax investigations

- Ancestral "Ames"
  - B
    - Florida isolate
      - C
      - D
    - 1997 TX goat
      - A
        - 1925 IA cow
          - 2001 CA cow
### Whole Genome Sequencing

#### Example Soil Microorganisms with Completed, Published Genome Sequences

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size (Kb)</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacillus anthracis</td>
<td>5227</td>
<td>Investigate/prevent bioterrorism</td>
</tr>
<tr>
<td>Agrobacterium tumefaciens</td>
<td>4915</td>
<td>Plant pathogen</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>6264</td>
<td>Human pathogen</td>
</tr>
<tr>
<td>Azoarcus st. EBN1</td>
<td>4727</td>
<td>Nitrate-reducing, aromatic hydrocarbon degrader</td>
</tr>
<tr>
<td>Methylococcus capsulatus</td>
<td>3304</td>
<td>CH$_4$-oxidation, cometabolic dechlorination of TCE</td>
</tr>
</tbody>
</table>

GOLD; www.genomesonline.org
• http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
Whole Genome Sequencing

• Benefits
  – Publicly available databases of genome sequences
  – Source of novel microbial products and processes
    • Industrial, medical, ecological
  – Organisms in culture facilitates proteomics experiments

• Limitations
  – Many open reading frames identified, but difficult to identify function
    • No genome sequence is more than 80% decoded
  – What organisms should be sequenced?
Metagenomics

• Also called Environmental genomics or Microbial ecogenomics

• “Culture independent analysis of a mixture of microbial genomes using an approach based either on expression or sequencing”
  – Schloss and Handlesmann, 2005

• “Biopropecting” microbial habitats for novel products and processes

• Determine ecological/biogeochemical role of microbes in unique habitats
Metagenomics

• Putting together a microbial ecosystem:
  - Acid Mine Drainage Biofilm
  - Low Diversity
    - 6 species identified with 16S
  - 10X coverage of dominant species
    - *Leptospirillum* group II
    - *Ferroplasma* group II
  - Identified genes
    - ion transport
    - iron-oxidation
    - carbon fixation
  - \( \text{N}_2 \)-fixation genes found only in a minor community member
    - *Leptospirillum* group III
  - Confirmed genomics with Proteomics
    - Linked 49% of ORF with peptides

Metagenomics

• Scope of diversity: Sargasso Sea
  – Oligotrophic environment
  – More diverse than expected
    • Sequenced $1 \times 10^9$ bases
    • Found 1.2 million new genes
    • 794,061 open reading frames with no known function
    • 69,718 open reading frames for energy transduction
      – 782 rhodopsin-like photoreceptors
    • 1412 rRNA genes, 148 previously unknown phylotypes
      (97% similarity cut off)
      – $\alpha$- and $\gamma$- Proteobacteria dominant groups

Metagenomics

Possible for soil ecosystems?

- MN soil metagenome
  - Only 1% of genome could be assembled into contiguous sequences
  - Est. 3000 – 5000 species
  - 150 K sequence reads, 100 Mbp

  - Too much diversity
    - Need 2-5 Gbp of sequence for enough coverage to identify dominant species

  - Used metagenomes to compare community structure and functions of divergent environments without linking organisms with functional open reading frames

Metagenomics
Possible for soil ecosystems?

• Bioprospecting
  – Express genes from metagenomic library in suitable host
  – Successful products
    • Antibiotics
    • Antibiotic resistance pathways
    • Anti-cancer drugs
    • Degradation pathways
      – Lipases, amylases, nucleases, hemolytic
    • Transport proteins

  - Gillespie, et. al. 2002. AEM 68:4301

• Link functional genes with uncultivated microbes
  – Functional gene on same clone insert as 16S rRNA operon
  – Identified several genes for uncultivated Acidobacterium
    • Insights on physiology and environmental role
    • May improve cultivation efforts

  - Liles, M.R. et al. AEM 69:2684
Metagenomics

• Limitations
  – Too much data?
    • Most genes are not identifiable
  – Contamination, chimeric clone sequences
  – Extraction biases
  – Requires proteomics or expression studies to demonstrate phenotypic characteristics
  – Need a standard method for annotating genomes
  – Requires high throughput instrumentation – not readily available to most institutions