Lecture 5 - Addendum
Sequencing Technology and Applications

Outline
• High-Throughput Sequencing Methods
• Genome Sequencing and Lander-Waterman
• Other Applications: Chip-Seq

DNA Sequencing
How we obtain the sequence of nucleotides of a DNA molecule?

Sequencing Challenges
• Only short DNA molecules
  – No machine can take entire genome
  – Need to fragment DNA into manageable chunks
• Read out only portions of DNA
  – Sequencing technologies produce at most 800bp of sequence at a time
  – Often much less
• Amplification
  – No machine (yet) can take single molecules of DNA
  – Need many identical copies

Sequencing Overview
• Preparation of DNA
  – Library Creation
• Sequencing
  – Sequencing Chemistry
  – Readout
• Analysis
  – Base Calling and Quality Scoring
  – Assembly, etc..

Library Creation
The first step with most sequencing is to create a sequencing library

Serves two purposes:
1. Generate DNA fragments of restricted lengths
2. Amplify each fragment to create many identical copies
Shearing DNA

Library Construction

- Insert sheared fragments into cloning vectors
  - Plasmids
  - Fosmids
  - Cosmids
- Transform E. coli with vectors
  - Insert vectors into bacteria

Different types of vectors

<table>
<thead>
<tr>
<th>VECTOR</th>
<th>Size of insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmid</td>
<td>2,000-10,000 Can control the size</td>
</tr>
<tr>
<td>Cosmid</td>
<td>40,000</td>
</tr>
<tr>
<td>BAC (Bacterial Artificial Chromosome)</td>
<td>70,000-300,000</td>
</tr>
<tr>
<td>YAC (Yeast Artificial Chromosome)</td>
<td>&gt; 300,000 Not used much recently</td>
</tr>
</tbody>
</table>

Plating

All plate(s) with fragments from original DNA are called a library

Library Construction Overview

Cloning Bias

Some sequences are toxic to E. coli
  e.g. AT-rich region

Billions of copies of ea fragment
Sequencing and Detection

We now have lots of *E. coli* with vectors (i.e. plasmids) with many fragments from the original DNA.

We want the sequence for each fragment.

Isolating Vectors

- **Lysing of Cells**
  - Isolate plasmid from *E. coli* DNA for each clone

- **Amplify Plasmid**
  - Rolling sequence amplification

DNA Replication

Two important properties of DNA polymerase:

- **Needs primer**
  - Always adds to existing sequence

- **Requires 3’–OH**
  - Adds 5’->3’
  - Bonds 5’ PPP to 3’ OH

DNA Synthesis for Sequencing

- **Need to prime the sequencing reaction**
  - Need a sequence already hybridized to template to start the sequencing reaction

- **Need to know what nucleotides are added during synthesis**

Chain Terminating Nucleotides

- **Dideoxynucleotide phosphates** (ddNTP) lack the 3’ OH group

- If a ddNTP is added during synthesis, the reaction is halted

If you add ddATP, and you get a synthesis product that is only 18 bases long, you know the 18th base is A!

Sanger Sequencing

The Key Idea

Add just enough ddATP to stop synthesis at every A

Similarly for T, G, C

Lengths of fragments tell you where the bases are
**DNA Gel Readout**

Label primer and run each termination reaction in separate lane

![DNA Gel Readout Diagram](image)

**Capillary Readout**

We can label each ddNTP with a different color fluorescence. Then we can read out all four bases in one capillary.

![Capillary Readout Diagram](image)

**Comparing Technologies**

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cost</th>
<th>Read Length</th>
<th>Days per Gb</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI 3730</td>
<td>1x</td>
<td>750b</td>
<td>650</td>
</tr>
<tr>
<td>454 FLX</td>
<td>0.25-0.1x</td>
<td>230b</td>
<td>3</td>
</tr>
<tr>
<td>Solexa 1G</td>
<td>0.01x</td>
<td>35-50b</td>
<td>3</td>
</tr>
</tbody>
</table>

**ABI 3730**

- **Cost**: 1x
- **Read Length**: 750b
- **Days per Gb**: 650

**Chromatogram**

![Chromatogram](image)

**Challenging to read answer**

Slide credit: Serafim Batzoglou
Reading an electropherogram

1. Filtering
2. Smoothening
3. Correction for length compressions
4. A method for calling the letters – PHRED

PHRED = PHil’s REad EDitor (by Phil Green)

Several better methods exist, but labs are reluctant to change

Output of PHRED: a read

A read: 500-700 nucleotides

A C G A A T C A G A
16 18 21 23 25 28 30 32 21

Quality scores: -10·log_{10}\text{Prob(Error)}

Reads can be obtained from leftmost, rightmost ends of the insert

Double-barreled sequencing: (1990)
Both leftmost & rightmost ends are sequenced, reads are paired

Pyrosequencing

\[(\text{NA})_i + \text{Nucleotide} \rightarrow \text{Polymerase} \rightarrow (\text{NA})_{i+1} + \text{PPI}\]

\[\text{PPI} + \text{APS} \rightarrow \text{ATP sulfurylase} \rightarrow \text{ATP} + \text{SO}_4^{2-}\]

\[\text{ATP} + \text{Luciferin} + \text{O}_2 \rightarrow \text{Luciferase} \rightarrow \text{AMP} + \text{PPI} + \text{Oxyluciferin} + \text{CO}_2 + \text{Light}\]

Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Res 2001
Ronagh M. Pyrosequencing sheds light on DNA sequencing. Genome Res 2001

454 LifeSciences Sequencer

No Cloning – Emulsion PCR

Does not introduce cloning bias

Depositing DNA Beads into the PicoTiter™ Plate
454 LifeSciences Sequencer

Short Reads But Lots of Them
• Read lengths 100-200bp
• 30+ million bases per run
• 0.25x to 0.1x the cost of Sanger sequencing

Solexa

Solexa Sequencing
Universal primers immobilized to glass surface

DNA Preparation – No Cloning

Solexa Sequencing
Hybridize DNA to primers
Solexa Sequencing
Identify Template-Primer Duplexes

Add a single fluorescently labeled nucleotide
Incorporate only one base at a time

Wash out unincorporated nucleotides

Identify those fragments with A at the 1st position

Remove fluorescent label

Repeat by stepping through all 4 nucleotides until max read length reached
Even Shorter Reads – Much More of Them

- Read lengths 30-50bp
- 600 million bases per run
- 0.01x the cost of Sanger sequencing

Comparing Technologies

<table>
<thead>
<tr>
<th></th>
<th>ABI 3730</th>
<th>454 FLX</th>
<th>Solexa 1G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>1x</td>
<td>0.25-0.1x</td>
<td>0.01x</td>
</tr>
<tr>
<td>Length</td>
<td>750bp</td>
<td>230bp</td>
<td>35-50bp</td>
</tr>
<tr>
<td>Days/Gb</td>
<td>650</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Outline

- High-Throughput Sequencing Methods
- Genome Sequencing and Lander-Waterman
- Other Applications: Chip-Seq

How Do We Sequence Genomes?

How do we go from <700 bp reads to complete genomes?

If we know where reads come from we just have to generate enough reads to cover each base once

Definition of Coverage

Length of genomic segment: \( G \)
Number of reads: \( N \)
Length of each read: \( L \)

Definition: Coverage \( C = \frac{L \times N}{G} \)

= Average number of times a base is sequenced

How much coverage do you need to read each base?

Typical contig coverage

Slide credit: Serafim Batzoglou
Lander-Waterman Model

Assumptions
• Sequencing reads will be randomly distributed in the genome
• The ability to detect an overlap between two truly overlapping reads does not vary from clone to clone


Poisson Model

Model number of reads starting in an interval I as Poisson with rate parameter \( \lambda \).

According to LW assumptions, average number of reads starting within any interval is \((N/G)L = \text{Coverage} (C) = \lambda\)

\[ P\{X = i\} = \frac{e^{-\lambda} \lambda^i}{i!} \]

Lander-Waterman Prediction

Then, the probability that a base will not be sequenced is:

\[ P_0 = P(\text{no read starts in I}) = \frac{e^{-a} a^0}{0!} = e^{-a} \]

Amount of genome not sequenced

\[ \approx P_0 G = e^{-a} G \]

Sequence Coverage

The probability any base is NOT sequenced is given by:

\[ P_0 = e^{-a} \]

where \( c = \text{fold sequence coverage} \), \( L = \text{read length} \), and \( N = \# \text{ reads} \).

\[ a = -\frac{N}{G} \]

http://www.genome.ou.edu/poisson_calc.html

Contigs and Gaps

\[ \frac{H}{\text{Contig}} \quad \frac{G}{\text{Gap}} \quad \frac{H}{\text{Contig}} \quad \frac{G}{\text{Gap}} \]

Overlapping reads

Expected #Contigs = #Gaps+1

Number of reads * probability that no other read starts in that read:

\[ \approx N P_0 = N e^{-a} = N e^{-N/L} = (aG/L) e^{-a} \]

http://www.genome.ou.edu/poisson_calc.html

Number of Gaps

Number of Gaps = \( Ne^{-a} \), where \( N = (G/c/L) \)

150kb Target Clone:

<table>
<thead>
<tr>
<th>500</th>
<th>650</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>( e^- \text{ Fingerprint} )</td>
</tr>
<tr>
<td>1</td>
<td>300</td>
</tr>
<tr>
<td>2</td>
<td>600</td>
</tr>
<tr>
<td>3</td>
<td>900</td>
</tr>
<tr>
<td>4</td>
<td>1200</td>
</tr>
<tr>
<td>5</td>
<td>1500</td>
</tr>
<tr>
<td>6</td>
<td>1800</td>
</tr>
<tr>
<td>7</td>
<td>2100</td>
</tr>
<tr>
<td>8</td>
<td>2400</td>
</tr>
<tr>
<td>9</td>
<td>2700</td>
</tr>
<tr>
<td>10</td>
<td>3000</td>
</tr>
</tbody>
</table>

http://www.genome.ou.edu/poisson_calc.html
In practice…

Lander-Waterman is almost always an underestimate
- cloning biases in shotgun libraries
- repeats
- GC/AT rich regions
- other low complexity regions

Genome Sequencing

We don’t typically know where reads come from

Use overlap to create contigs

This is Genome Assembly – the topic of the next lecture

BAC-by-BAC Sequencing

Physical mapping → Minimal tiling path

Mapping/Ordering BACs

What is a marker?
A way to uniquely locate a position in a genome

What is mapping?
Statistical association between markers, ordering markers in linear sequence.

How do we map?
“Shatter” genome and observe how often two markers travel together on the same piece of DNA

What does it mean for two markers to be linked?
P(M1|M2) > P(M1)

What does it mean to order BACs?
Create a minimal tiling path.

Whole Genome Shotgun Sequencing

Genomic DNA

Combined Approach

Shotgun clone

BAC Library/BAC end sequencing

Physical Mapping

BAC shotgun sequencing
Outline

• High-Throughput Sequencing Methods
• Genome Sequencing and Lander-Waterman
• Other Applications: Chip-Seq

Other Applications

Advances in sequencing technology change the way we think about applications

Challenge ➞ Commodity ➞ Assay
(10 yrs ago) (last year) (today)

Sequencing can be considered a general way to determine what sequences are present and how many times

DNA Binding Site Identification

DNA-binding proteins control transcription, replication, DNA repair, and chromosome segregation.

Example: Chromatin and Histones

Chromatin Modification

• Histones have tails that can be modified
  – Methylation, acetylation, etc
• Histone modification associated with cellular state

Immunoprecipitation

ChIP-chip

Circulate DNA and proteins (bound) and acidize chromatin

Sequinate or digest chromatin

Immunoprecipitate, reverse crosslinking, purify DNA

Immunoprecipitation (IP)

Control (no IP)
Chip-Seq

- Instead of hybridizing DNA fragments to a chip, we just sequence them
- Then align to genome sequence

Advantages of Chip-Seq

- Whole genome assayed by a sequencing approach, rather just regions on a microarray.
- Avoids the complications of array hybridization
  - probes with different optimal temperatures
  - probes that hybridize to more than one DNA sequence,
  - interference of hybridization by DNA secondary structure.
- Only half the cost of human whole genome tiling arrays.
- Can immediately be applied even if microarray not available

Coverage – Poisson Model

- Assume genome divided into N bins
- A fraction $f$, of these bins have desired modification
- Collect a total of $R$ reads
- Immunoprecipitation enriches bins with desired modification by factor of $e$

Then number of reads per bin approximately poisson with mean:

$$eM \quad \text{bins with modification} \quad \text{Where} \quad R \quad eM \quad \text{bins w/o modification} \quad \text{M} = \frac{N(ef - (1-f))}{M}$$

Sensitivity and Specificity

- Predicting a particular modification means selecting a threshold number of reads in a bin
- Can estimate accuracy from Poisson models
  - Overlap between Poisson for empty bins versus bins with modification at that read threshold

Example

Assume
- Mouse genome
- 1% of genome has modification
- 20 fold enrichment

Then
- 95% specificity and sensitivity
- 200 bp bins (resolution)
- Requires 5 million reads

Many Other Applications

- Expression Analysis
- Immunosequencing
- Re-sequencing
- Etc…