Sequencing Technologies - perspective

1\textsuperscript{st} generation: Frederic Sanger develops DNA sequencing technology. Latest versions 3 million bases/day, 1500bp long reads

2\textsuperscript{nd} generation: (next-gen) sequencing started 2005 with the release of the 454 sequencing platform. 600 billion bases/week, 150bp long reads

3\textsuperscript{rd} generation: single molecule (no DNA amplification required), these are not replacing but augmenting 2\textsuperscript{nd} generation systems, longer reads, shorter turnarounds)

Random DNA fragment sequencing with Illumina

- Forward sequencing
- Reverse sequencing
- Fragmentation
- For each fragment -> adapter ligation -> separate by strands -> some pieces get sequenced

Sequencing Instruments

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Run time</th>
<th>Million reads/week</th>
<th>Bases/bp</th>
<th>Yield Mb/ran</th>
<th>Reagent cost/million</th>
<th>Reagent cost/sketch</th>
<th>Minimum unit cost/100x</th>
</tr>
</thead>
<tbody>
<tr>
<td>3730xl (capillary)</td>
<td>2 h</td>
<td>0.00056</td>
<td>650</td>
<td>0.06</td>
<td>896</td>
<td>$800</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>Ion Torrent - 314 (chip)</td>
<td>2 h</td>
<td>0.10</td>
<td>300</td>
<td>&gt;10</td>
<td>8000</td>
<td>&lt;50</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>Ion Torrent - 414 (chip)</td>
<td>2 h</td>
<td>0.10</td>
<td>400</td>
<td>50</td>
<td>1300</td>
<td>&lt;50</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>10 h</td>
<td>1</td>
<td>400</td>
<td>50</td>
<td>6200</td>
<td>&lt;50</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>454 FLX Titanium+</td>
<td>16-20 h</td>
<td>1</td>
<td>700</td>
<td>300</td>
<td>6200</td>
<td>&lt;50</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>Ion Torrent - 316/chip</td>
<td>2 h</td>
<td>1</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>770</td>
<td>&lt;50</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>Helicos</td>
<td>NA</td>
<td>NA</td>
<td>600</td>
<td>35</td>
<td>20000</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>2 h</td>
<td>2</td>
<td>500</td>
<td>300</td>
<td>3200</td>
<td>$80</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>2 days</td>
<td>250</td>
<td>500</td>
<td>1000</td>
<td>$100</td>
<td>$20</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>SOLiD</td>
<td>4</td>
<td>14 days</td>
<td>320</td>
<td>300</td>
<td>300</td>
<td>200</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>SOLiD</td>
<td>5</td>
<td>12 days</td>
<td>&gt;800</td>
<td>90</td>
<td>300</td>
<td>$80</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>SOLiD</td>
<td>6</td>
<td>8 days</td>
<td>300</td>
<td>300</td>
<td>1000</td>
<td>$100</td>
<td>$20</td>
</tr>
<tr>
<td>SOLiD</td>
<td>7</td>
<td>200</td>
<td>200</td>
<td>2000</td>
<td>$20</td>
<td>&lt;500 (100%)</td>
<td></td>
</tr>
<tr>
<td>SOLiD</td>
<td>8</td>
<td>6 days</td>
<td>&gt;700</td>
<td>35</td>
<td>150</td>
<td>$100</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>SOLiD</td>
<td>9</td>
<td>8 days</td>
<td>&gt;1400</td>
<td>35</td>
<td>150</td>
<td>$100</td>
<td>&lt;500 (100%)</td>
</tr>
<tr>
<td>SOLiD</td>
<td>10</td>
<td>10 days</td>
<td>1000</td>
<td>50</td>
<td>1500</td>
<td>$25</td>
<td>&lt;500 (100%)</td>
</tr>
</tbody>
</table>
Sequencing Coverage (Depth)
Unequal length reads
Lander/Waterman model (random reads from random genome)

Coverage \( \rightarrow C = \frac{\text{sum}(L_i)}{G} \)

\[ \text{sum}(L_i) \rightarrow \text{total yield} \]

\( L_i = \text{length of read } i, i \rightarrow 1 \ldots N, N=\text{number of reads} \)
\( G = \text{size of genome} \)

Sequencing Coverage (Depth)
Equal length reads
Lander/Waterman model (random reads from random genome)

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

\( N \times L \rightarrow \text{total yield} \)

\( N = \text{number of reads}, L = \text{length of reads} \)
\( G = \text{size of genome} \)

Probability of a base not being sequenced
Lander/Waterman model (random reads from random genome)

Probability of a base \textbf{not being} sequenced

\[ P = \exp(-C) \]

Example: \( N=35 \) million, \( L=35, \ G=250 \) million, \( C=\text{coverage} \)

\( C = 5 \rightarrow 0.6\% \text{ genome not sequenced} \rightarrow 15 \) million bases not covered

\textit{(note: in practice the probabilities are always higher, this shows the lowest possible limit)}

Realistic coverage measures

Neither of the models assumptions are correct

Empirical observation \( \rightarrow \) raise the required coverage at least 10 fold

What part of the genome is \textbf{coverable} to begin with?

What part of the genome is \textbf{uniquely coverable} with a give read size?

Nomenclature: “accessible”, “mappable”, “effective” genome sizes
Sequence duplication

Type:  Natural  Artificial (PCR, Optical)

Source:  Sample  Contaminant

Detection:  Sequence identity  Alignment identity

How to deal with duplicates  no easy or universal answer
SNP calling  typically remove if coverage is not too high
All other methodologies  need careful evaluation, natural duplicates my be important

Quality Control

Quality control is the first step of any data analysis.

This is where you first meet your data.

The first step is to identify problems then improve of them.

Bad data is like poison  it can turn even good data into bad

NGS sequencing read formats

Reads: short sequences produced by the instrument

New instruments produce FastQ format (.fastq or .fq). Current platforms:
- HiSeq/MiSeq (Illumina)
- IonTorrent (Thermo Fisher)
- PacBio (Pacific Biosciences)
- MiniON (Oxford Nanopore)

Obsolete:
- Solid  colortspace fasta (.xsq or .csfasta + .qual)
- 454  standard flowgram format (.sff)

Quality control tools: FastQC (a great tool but with numerous flaws)

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
Per base quality scores

Averaging over columns.
The average quality value at a given position

Per sequence quality score

Computes an average quality for each read.
How many of a given average were found.

Sequence duplication: misleading plot

What does this plot actually mean: See Biostar threads (linked from the course page)

Adapter Content

Checks known contamination.
You can add your own contaminants
Kmer enrichment

Attempts to find frequent subsequences.

The plot can be greatly misleading. A longer motif would be counted multiple times. The vertical scale is wonky.
Use it just as a starting point of investigation. Any behavior seen here will have to be verified via other means.

Kmer – subsequence of length k

A kmer or k-mer is a short DNA sequence consisting of a fixed number (K) of bases. The value of k is usually divisible by 4 so that a kmer can fit compactly into a basevector object. Typical values include 12, 20, 24, 36, and 48; kmers of these sizes are referred to as 12-mers, 20-mers, and so forth.

Homework 7 (part 1)

We want to sequence the pathogenic E. coli strain O157:H7 with the genome size of 5.44 MBp (mega base pairs, 5.44 * 10^6 bp). We have access to an instrument that can be tuned for read lengths and read numbers.

1. How many 150bp long reads will produce a 10x coverage? How many bases of the genome have not been sequenced (in theory)?
2. If the desired coverage is 25x and the instrument produces 1 million reads how long should each read be? How many bases of the genome have not been sequenced?
3. At maximum the instrument can produce 10 million 250 bp long reads on each run. How many samples could be multiplexed (run in parallel) to cover each sample at 10x. How many bases of the genome have not been sequenced?

Hint: the algebra is very simple – the challenge is plugging the values into the right places.

Homework 7 (part 2)

Pick a Fastqc sequencing platform report from homework 7.

What is the most enriched k-mer in that dataset.