Sequencing technologies
part of “High-Throughput Analyzes of Genome Sequences”

Sonja J. Prohaska

Computational EvoDevo
University of Leipzig

Leipzig, WS 2014/15
## Illumina sequencer

<table>
<thead>
<tr>
<th>Key applications</th>
<th>MiSeq</th>
<th>NextSeq 500</th>
<th>HiSeq 2500</th>
<th>Production power.</th>
<th>Power and efficiency for large-scale genomics.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small genome, amplicon, and targeted gene panel sequencing.</td>
<td>N/A</td>
<td>Mid-Output</td>
<td>High-Output</td>
<td>Rapid Run</td>
<td>High-Output</td>
</tr>
<tr>
<td>Everyday genome, exome, transcriptome sequencing, and more.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1 or 2</td>
<td>1 or 2</td>
</tr>
<tr>
<td>Production-scale genome, exome, transcriptome sequencing, and more.</td>
<td>0.3-15 Gb</td>
<td>20-39 Gb</td>
<td>30-120 Gb</td>
<td>10-300 Gb</td>
<td>50-1000 Gb</td>
</tr>
<tr>
<td>Output range</td>
<td>5-55 hours</td>
<td>15-26 hours</td>
<td>12-30 hours</td>
<td>7-60 hours</td>
<td>&lt; 1 day - 6 days</td>
</tr>
<tr>
<td>Run time</td>
<td>25 Million†</td>
<td>130 Million</td>
<td>400 Million</td>
<td>300 Million</td>
<td>2 Billion</td>
</tr>
<tr>
<td>Maximum read length</td>
<td>2 x 300 bp</td>
<td>2 x 150 bp</td>
<td>2 x 150 bp</td>
<td>2 x 250 bp</td>
<td>2 x 125 bp</td>
</tr>
</tbody>
</table>
Sample Multiplexing – flow cell

- one flow cell – 8 lanes
- 1 lane – 12-fold coverage of the human genome
- human genome size: $3.2 \times 10^9$ bp
- sequenced: $\sim 40$ billion bp
- read length: 200bp
- 200 million reads
Sample Multiplexing – barcoding

What should be put in the other lanes?

- additional samples
- sample-specific indexes/barcodes
- embedded in library adapters
- risk of sample misidentification

Below are the 12 barcodes used in the Illumina TruSeq system, they are base-balanced and work well as In-Line barcodes as well as Multiplex.

ATCACG  CGATGT  TTAGGC  TGACCA  ACATGT  GCCAAT  
CAGATC  ACTTGA  GATCAG  TAGCTT  GGCTAG  CTTGTA
Base calling
Base calling – Quality Scores

Phred quality scores

- Phred quality score $Q$
- base-calling error probabilities $P$

\[ Q = -10 \log_{10} P \] (1)

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
<tr>
<td>60</td>
<td>1 in 1,000,000</td>
<td>99.9999%</td>
</tr>
</tbody>
</table>
FASTQ format

- store nucleotide sequence (read)
- quality score

@SEQ_ID
GATTTGGGGTTCAAGCAAGTATCGATCAAAATAGTAATCTCATTGTCAACTCAGTTT
+
!"*( (((***+))%%%++)(%%%%).1***-+**'))**55CCF>>>>>>>CCCCCCC65

!"#$%&'(()*+,-./0123456789;<=?>@ABCDEFGHIJKLMNOPQRSTUVWXYZ\_%`abcdefghijklmnopqrstuvwxyz{|}~

- line 1: “@” sequence identifier
- line 2: sequence
- line 3: “+” [sequence identifier]
- line 4: quality values in ASCII code
- ASCII character 33(!) to 126(∼) (93 score values)
Mapping – reads back to the genome

**Will we find the locus the read originated from?**

- 23-29Mb of sequence are absent from the human genome
- sequences might not map to the reference genome
- sequences might map to more than one region
- sequences might be erroneous (quality?)
- tremendous amounts of sequences to map (billions of reads)

How long does a sequence need to be to map uniquely to the human genome?
Monarch Butterfly Genome – sequencing

Genome Assembly and Gene Content
We used a whole-genome shotgun approach with next-generation sequencing platforms to generate the draft genome of the monarch butterfly (Table 1 and Table S1 available online). The combined assembly of 14.7 Gb pairs of Illumina reads (equal to 53.3× coverage of the whole genome) and 6.2 Gb Roche 454 reads (22.3×) resulted in 273 megabases (Mb) of genomic sequence (combined total coverage of 74.7×) (Table S1A). This was termed the v1 assembly and was used for all subsequent analyses (Table S1B). Assessment of the completeness and quality of the assembly v1 is described in the Experimental Procedures.

- genome size: 273 Mb on 29-30 chromosomes
- only 13.1% predicted repeats
- 7.51% predicted coding
- 16866 predicted genes