Next, 2nd and 3rd Generation Sequencing

What, How and What

08-01-2013, Gabino Sanchez Perez
Principles of DNA sequencing

[Image of a man holding a piece of paper, with a diagram of DNA structure]

Courtesy of Dr. F. Sanger, MRC, Cambridge. Non-commercial, educational use only.
Principles of DNA termination sequencing

ddNTPs terminate DNA synthesis.

Normal dNTP (extends DNA strand)

ddNTP (terminates synthesis)
Principles of Sanger sequencing
ABI 3730 XL capillary sequencing
No sequence is perfect

DNA Polymerase slipping

Double signals

Chimeric reads
In perspectief:

Tomato genome

Traditional seq 2003-2008

Next gen seq 2009
Choosing technologies

<table>
<thead>
<tr>
<th></th>
<th>Pyrosequencing</th>
<th>Synthesis</th>
<th>Single molecule</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Throughput/day</strong></td>
<td>600 Mb</td>
<td>60 Gb</td>
<td>600 Mb</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>High (~99%)</td>
<td>High (~99%)</td>
<td>Low (~85%)</td>
</tr>
<tr>
<td><strong>Read length</strong></td>
<td>Medium (~700 bp)</td>
<td>Short (~100 x 2 bp)</td>
<td>Long (10+ Kbp)</td>
</tr>
<tr>
<td><strong>Running time</strong></td>
<td>Short (1 day)</td>
<td>Long (11 days)</td>
<td>Short (16 hours)</td>
</tr>
<tr>
<td><strong>Mate pairs</strong></td>
<td>3, 8 &amp; 20 Kb</td>
<td>2, 6 &amp; other</td>
<td>-</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td>0.5-5 μg</td>
<td>50ng-1μg</td>
<td>2-5μg</td>
</tr>
<tr>
<td><strong>Price</strong></td>
<td>5800 € / plate</td>
<td>3400 € / lane</td>
<td>1600 € / SMRT</td>
</tr>
</tbody>
</table>
Impact of read length on assembly

Short Reads

Long Reads
Assembly: prior knowledge or not?

- De-novo
- Reference
De-novo assembly process:

Genome:
‘Draft’ – ‘HQ draft’ – ‘complete’

Overlapping reads assembled to larger contiguous pieces: ‘contigs’

Assembly ‘gaps’ between contigs

Contigs ordered to ‘scaffolds’
Assembly problems, incompleteness & mistakes
<table>
<thead>
<tr>
<th>Plant</th>
<th>Genome Size (Mb)</th>
<th>Genes (K)</th>
<th>X-coverage</th>
<th>Assembled (%)</th>
<th>Assembly</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>125</td>
<td>31</td>
<td>100</td>
<td>chromosomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rice</td>
<td>466</td>
<td>37</td>
<td>6</td>
<td>42,109 Contigs</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Poplar</td>
<td>485</td>
<td>&gt;45</td>
<td>7.5</td>
<td>2,447 major scaffolds</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Grapevine</td>
<td>504</td>
<td>29.5</td>
<td>10.7</td>
<td>2,093 meta ctgs</td>
<td>S+454</td>
<td></td>
</tr>
<tr>
<td>Papaya</td>
<td>372</td>
<td>28</td>
<td>3</td>
<td>17,700 scaffolds</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Apple</td>
<td>742</td>
<td>57</td>
<td>16.9</td>
<td>162 meta ctgs, 122,149 ctgs</td>
<td>S+454</td>
<td></td>
</tr>
<tr>
<td>Strawberry</td>
<td>240</td>
<td>34.8</td>
<td>39</td>
<td>3,200 scaffolds, 18,577 ctgs</td>
<td>454+i+Solid</td>
<td></td>
</tr>
<tr>
<td>Sorghum</td>
<td>730</td>
<td>34</td>
<td>8.5</td>
<td>221 scaffolds</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Potato</td>
<td>850</td>
<td>39</td>
<td>70</td>
<td>66,253 scaffolds</td>
<td>S+i+454</td>
<td></td>
</tr>
<tr>
<td>Tomato</td>
<td>900</td>
<td>34.7</td>
<td>140</td>
<td>3,232 scaffolds</td>
<td>454+S+i</td>
<td></td>
</tr>
<tr>
<td>Canabis</td>
<td>400</td>
<td>ND</td>
<td>379</td>
<td>170,000 ctgs</td>
<td>i</td>
<td></td>
</tr>
<tr>
<td>Banana</td>
<td>523</td>
<td>36.5</td>
<td>70</td>
<td>7,513 scaffolds</td>
<td>454+S+i</td>
<td></td>
</tr>
</tbody>
</table>
three NGS technologies

454

HiSeq

PacBio
Next generation sequencing: 454

- 700 base pairs long reads
- $1 \times 10^6$ HQ reads / run
- Up to 1 Giga bases / run
- De-novo sequencing
  - Genomes, transcriptomes
  - Amplicon Variant Analysis
  - Metagenome sequencing
  - Large insert Paired End Sequencing
NGS – 454 workflow
NGS - 454 pyrosequencing raw read

![NGS-454 pyrosequencing raw read graph]
## NGS - 454 WGS run output

<table>
<thead>
<tr>
<th></th>
<th>GACT (Library)</th>
<th>Region 1</th>
<th>Region 2</th>
<th>Region Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Wells</td>
<td>1,022,754</td>
<td>1,016,241</td>
<td></td>
<td>2,038,995</td>
</tr>
<tr>
<td>Key Pass Wells</td>
<td>970,172</td>
<td>964,489</td>
<td></td>
<td>1,934,661</td>
</tr>
<tr>
<td>Passed Filter Wells</td>
<td>573,163</td>
<td>603,050</td>
<td></td>
<td>1,176,213</td>
</tr>
<tr>
<td>Total Bases</td>
<td>274,895,922</td>
<td>296,609,089</td>
<td></td>
<td>571,505,011</td>
</tr>
</tbody>
</table>

|                        |                |          |          |
| Length Average         | 479.61         | 491.85   | 485.89   |
| Length Std Deviation   | 153.68         | 156.45   |          |
| Longest Reads Length   | 1,594          | 1,594    | 1,594    |
| Shortest Reads Length  | 40             | 40       | 40       |
| Median Reads Length    | 505.0          | 519.0    | 512.0    |
2nd Gen Sequencing Illumina HiSeq2000

- **Highest Output**
  - <600 Gb per run

- **Fastest Data Rate**
  - 11 days for 2 x 100 bp

- **Highest Number of Reads**
  - 1 billion single-end reads
  - 2 billion paired-end reads
  - 50 or 100 bp reads
2nd Gen Sequencing: Illumina HiSeq2000

- Optics
- Flow cell access door
- Syringe pumps
- Reagents compartment
- Flow cell
  - 8 channels
DNA (0.1-5.0 μg)

Library Preparation

Single molecule array

Cluster Growth

Sequencing

Image Acquisition

Base Calling

TGCTACGAT...
Illumina HiSeq DNA library fragment

STRUCTURE DETAILS

Rd1 Seq Primer

Sequence of Interest

Index Seq Primer

P5

P7

RD2 Seq Primer

INDEX

PLANT RESEARCH INTERNATIONAL WAGENINGEN UR
3rd Gen Sequencing: PacBio

Three Key Innovations of SMRT Technology

- Phospholinked Nucleotides
- PacBio RS
- Real-Time Detection
- SMRT™ Cell
Phospholinked Nucleotides

- Fluorophore clipped off by polymerase
- DNA synthesized is natural
- No steric hindrance or accumulation of background signal
DNA Polymerase as Sequencing Engine

ZMW Zero Mode Waveguide
High Performance Single-Molecule Detection
Longer Sequencing Readlength

Greater than 1 base per second
2.6 kb single continuous read
SMRTbell™ Template prep

Key Advantages:

- Structurally linear
- Topologically circular
- Structural homogeneity of templates
- Provides sequences of both forward and reverse strands in the same trace
SMRT bell template sequencing
Universal SMRTbell Template

Standard Sequencing

Large Insert Sizes

Generates one pass on each molecule sequenced

Circular Consensus Sequencing

Small Insert Sizes

Continued generation of reads per insert size

Generates multiple passes on each molecule sequenced
PacBio output & circular sequencing

- Kb read length
- >100,000 reads/run
- Read quality 0.8 to 0.85
- 3 passes base quality >95%
- fragment size ~ 1kb
### Readlength Histogram

**Post-filter**

![Readlength Histogram](image)

- # of SMRT Cells: 4
- # of Movies: 8
- Pre-Filter # of Bases: 859133249 bp
- Post-Filter # of Bases: 765147831 bp
- Pre-Filter # of Reads: 801168
- Post-Filter # of Reads: 283125
- Pre-Filter Mean Readlength: 1429 bp
- Post-Filter Mean Readlength: 2908 bp
- Pre-Filter Mean Read Quality: 0.383
- Post-Filter Mean Read Quality: 0.846

### Read Quality Histogram

**Post-filter**

![Read Quality Histogram](image)
Error Correction and SMRT® Hybrid Assembly

- Pre-process reads
- Map short to long reads
- Separate repeats
- Compute layout
- Trim at coverage gaps
- Compute consensus
- Assemble corrected reads with Overlap Layout Consensus (OLC)
NGS conclusions for data production and analysis

- Many possible sequencing technologies and data sources
- Be aware of technical limitations
- Use longest read length if possible
- Use paired end reads if possible
- Combinations of technologies > more reliability