NGS Overview Part I:
A Comparison of Next-Generation Sequencing Platforms.

MIN SOO KIM
APRIL 1, 2013
QUANTITATIVE BIOMEDICAL RESEARCH CENTER
DEPARTMENT OF CLINICAL SCIENCES
UT SOUTHWESTERN
• Background.

• Common Pipeline.
  ○ Library Prep.
  ○ Sequencing – Massively Parallel Sequencing.
  ○ Bioinformatics - Data Analysis.

• Popular Platforms:
  ○ Roche 454, AB SOLiD, Illumina(HiSeq, MiSeq).

• Newer Platforms (Third Generation):
  ○ Ion Torrent, PacBio RS, Oxford Nanopore.
Background

Library Preparation

- DNA samples are randomly fragmented and platform-specific adaptors are added to the flanking ends to produce a “library”.
- Library is then amplified through PCR. (Platform-specific amplification e.g. beads or glass)
- Amplification Introduces Bias:
  - Amplification bias against AT, GC rich regions. (corrected by adding PCR additives.)
  - Alteration of representational abundances(duplicates). Important for quantitative applications like RNA-seq.
    - Overrepresentation of smaller fragments. (corrected by running fewer PCR cycles.)
Massively Parallel Sequencing

- In Sanger-sequencing the DNA synthesis and detection steps are two separate procedures (slow).
- Next generation sequencing relies on coupling the DNA synthesis and detection (sequencing by synthesis) and multiple sequencing reactions are run simultaneously (Massively Parallel Sequencing).
- For most NGS platforms desynchronization of reads during the sequencing and detection cycle is the main cause of sequencing errors and shorter reads.
<table>
<thead>
<tr>
<th>Platform</th>
<th>Chemistry</th>
<th>Read</th>
<th>Run Time</th>
<th>Gb/Run</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 GS Junior (Roche)</td>
<td>Pyro-sequencing</td>
<td>500</td>
<td>8 hrs.</td>
<td>0.04</td>
<td>Long Read Length</td>
<td>High error rate in homopolymer</td>
</tr>
<tr>
<td>454 GS FLX+ (Roche)</td>
<td>Pyro-sequencing</td>
<td>700</td>
<td>23 hrs.</td>
<td>0.7</td>
<td>Long Read Length</td>
<td>High error rate in homopolymer</td>
</tr>
<tr>
<td>HiSeq (Illumina)</td>
<td>Reversible Terminator</td>
<td>2*100</td>
<td>2 days (rapid mode)</td>
<td>120 (rapid mode)</td>
<td>High-throughput / cost</td>
<td>Short reads Long run time (normal mode)</td>
</tr>
<tr>
<td>SOLiD (Life)</td>
<td>Ligation</td>
<td>85</td>
<td>8 days</td>
<td>150</td>
<td>Low Error Rate</td>
<td>Short reads Long run time</td>
</tr>
<tr>
<td>Ion Proton (Life)</td>
<td>Proton Detection</td>
<td>200</td>
<td>2 hrs.</td>
<td>100</td>
<td>Short Run times</td>
<td>New*</td>
</tr>
<tr>
<td>PacBio RS</td>
<td>Real-time Sequencing</td>
<td>3000 (up to 15,000)</td>
<td>20 min</td>
<td>3</td>
<td>No PCR Longest Read Length</td>
<td>High Error Rate</td>
</tr>
</tbody>
</table>
Emulsion PCR (emPCR)

1. DNA fragmentations and adaptor ligation.
2. DNA fragments are added to an oil mixture containing millions of beads.
3. Emulsion PCR results in multiple copies of the fragment.
4. Beads are deposited on plate wells ready for sequencing.
Roche 454 – Pyrosequencing

Illumina – Library Prep.

1. PREPARE GENOMIC DNA SAMPLE
   - Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
   - Fragments become double-stranded.

2. ATTACH DNA TO SURFACE
   - Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
   - Denature the double-stranded molecules.

3. BRIDGE AMPLIFICATION
   - Add unlabeled nucleotides and enzymes to initiate solid-phase bridge amplification.
   - Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

4. DENATURE THE DOUBLE-STRANDED MOLECULES
   - Attached terminus自由末端

5. COMPLETE AMPLIFICATION
   - Attached

The enzymes incorporate nucleotides to build double-stranded bridges on the solid-phase substrate.

Denaturation leaves single-stranded templates anchored to the substrate.
Illumina – Sequence by Synthesis

1. Add dye-labeled nucleotides.
2. Scan and detect nucleotide specific fluorescence.
3. Remove 3’ – blocking group (Reversible termination).
4. Cleave fluorescent group.
5. Rinse and Repeat.

http://hmg.oxfordjournals.org
## Illumina – Sequencing by Synthesis

<table>
<thead>
<tr>
<th>Platform</th>
<th>Run Time</th>
<th>Yield (GB)</th>
<th>Error Type</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAIIx</td>
<td>14 days</td>
<td>96</td>
<td>Sub</td>
<td>0.1</td>
</tr>
<tr>
<td>HiSeq 2000/2500</td>
<td>10/2 days</td>
<td>600/120</td>
<td>Sub</td>
<td>0.03*</td>
</tr>
<tr>
<td>MiSeq</td>
<td>1 day</td>
<td>2</td>
<td>Sub</td>
<td>0.03*</td>
</tr>
</tbody>
</table>
Illumina

Advantages

- High throughput / cost.
- Suitable for a wide range of applications most notably whole genome sequencing.

Disadvantages

- Substitution error rates (recently improved).
- Lagging strand dephasing causes sequence quality deterioration towards the end of read.
SOLiD (Life/AB)

- SOLiD: Sequencing by Oligonucleotide Ligation and Detection.

- Library Prep: emPCR

- “2-base encoding” – Instead of the typical single dNTP addition, two base matching probes are used. (possible 16 probes).

- **Color Space** – four color sequencing encoding further increases accuracy.
1. Anneal primer and hybridize probe (8-mer).
2. Ligation and Detection.
3. Cleave fluorescent tail (3-mer).
4. Repeat ligation cycle.

Repeat steps 1-4 with (n-1) primer.
SOLiD – Color Space

- 16 possible base combination are represented by 4 colors.
- All possible sequencing combination need to be decoded.

Cannot determine any of the bases
SOLiD – SNP Detection

Summary: SOLiD has one of the lowest error-rates (~0.01) due to 2-base encoding. It is however still limited by short read lengths (35 bp / 85 bp for PE).
Ion Torrent (Life Technologies)

- Similar to pyrosequencing but uses semiconducting chip to detect dNTP incorporation.
- The chip measure differences in pH.
- Shown to have problems with homopolymer reads and coverage bias with GC-rich regions.
- Ion Proton™ promises higher output and longer reads.

PacBio RS

- Single Molecule Sequencing – instead of sequencing clonally amplified templates from beads (Pyro) or clusters (Illumina) DNA synthesis is detected on a single DNA strand.

Zero-mode waveguide (ZMW)
- DNA polymerase is affixed to the bottom of a tiny hole (∼70nm).
- Only the bottom portion of the hole is illuminated allowing for detection of incorporation of dye-labeled nucleotide.
PacBio RS

- **Real-time Sequencing** – Unlike reversible termination methods (Illumina) the DNA synthesis process is never halted. Detection occurs in real-time.

**Library Prep.**
- DNA template is circularized by the use of “bell” shaped adapters.
- As long as the polymerase is stable this allows for continuous sequencing of both strands.
PacBio RS

Advantages
- No amplification required.
- Extremely long read lengths.
- Average 2500 bp. Longest 15,000bp.

Disadvantages
- High error rates.
- Error rate of ~15% for Indels. 1% Substitutions.
Oxford Nanopore

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

- Measure changes in ion flow through nanopore.
- Potential for long read lengths and short sequencing times.

http://www2.technologyreview.com
Applications
- Whole Genome Sequencing.
- Exome-sequencing
- Target Resequencing
- RNA-seq
- Chip-seq
- SNP/Indel/Structural Variation Discovery.

Experimental Design.
Acknowledgements

Dr. Tae Hyun Hwang

References