Introduction to High-Throughput Sequencing and RNA-seq

L Collado-Torres

March 6th, 2013
1 High Throughput Sequencing

2 Sources of variation
DNA

1Wikipedia. DNA. URL: http://en.wikipedia.org/wiki/DNA (visited on 03/05/2013).
Human mRNA

The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)

---

Panorama³

³Wendy Weijia Soon, Manoj Hariharan, and Michael P. Snyder. “High-throughput sequencing for biology and medicine”. In: Molecular Systems Biology 9.1 (). URL http://www.nature.com/msb/journal/v9/n1/fig_tab/msb201261_F2.html (visited on 03/05/2013).
Prepare DNA

- Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

---

Amplify\textsuperscript{5}

\begin{itemize}
  \item \textbf{Illumina/Solexa Solid-phase amplification}
    One DNA molecule per cluster
  \item Sample preparation DNA (5 μg)
  \item Template dNTPs and polymerase
  \item Bridge amplification
  \item Cluster growth
  \item 100–200 million molecular clusters
\end{itemize}

\textsuperscript{5}Michael L. Metzker. “Sequencing technologies — the next generation”. In: Nat Rev Genet 11.1 (2010).
High Throughput Sequencing

PCR

PCR: Uses, Steps, Purpose. URL: http://schoolworkhelper.net/pcr-uses-steps-purpose/ (visited on 03/05/2013).
Sequencing by synthesis

Analyze cluster images

---

HiSeq 2000

Figure 3: Next-Generation Sequencing Simplified

From simplified sample preparation kits, to automated cluster generation, to streamlined sequencing by synthesis, to complete data analysis, Illumina’s HiSeq 2000 sequencing system offers the industry’s simplest next-generation sequencing workflow.

HiSeq 2000

## Other 2nd generation sequencers

<table>
<thead>
<tr>
<th>Sequencer</th>
<th>454 GS FLX</th>
<th>HiSeq 2000</th>
<th>SOLiDv4</th>
<th>Sanger 3730xl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing mechanism</td>
<td>Pyrosequencing</td>
<td>Sequencing by synthesis</td>
<td>Ligation and two-base coding</td>
<td>Dideoxy chain termination</td>
</tr>
<tr>
<td>Read length</td>
<td>700 bp</td>
<td>50SE, 50PE, 101PE</td>
<td>50 + 35 bp or 50 + 50 bp</td>
<td>400~900 bp</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99.99%*</td>
<td>98%, (100PE)</td>
<td>99.94% *raw data</td>
<td>99.999%</td>
</tr>
<tr>
<td>Reads</td>
<td>1 M</td>
<td>3 G</td>
<td>1200~1400 M</td>
<td>—</td>
</tr>
<tr>
<td>Output data/run</td>
<td>0.7 Gb</td>
<td>600 Gb</td>
<td>120 Gb</td>
<td>1.9~84 Kb</td>
</tr>
<tr>
<td>Time/run</td>
<td>24 Hours</td>
<td>3~10 Days</td>
<td>7 Days for SE</td>
<td>20 Mins~3 Hours</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14 Days for PE</td>
<td></td>
</tr>
<tr>
<td>Advantage</td>
<td>Read length, fast</td>
<td>High throughput</td>
<td>Accuracy</td>
<td>High quality, long read length</td>
</tr>
<tr>
<td>Disadvantage</td>
<td>Error rate with polybase more than 6, high cost, low throughput</td>
<td>Short read assembly</td>
<td>Short read assembly</td>
<td>High cost low throughput</td>
</tr>
</tbody>
</table>

---

1. High Throughput Sequencing

2. Sources of variation
Sources of variation

Cross-talk\textsuperscript{11}

Phasing and pre-phasing

Figure 3

Sources of variation

Phasing and pre-phasing

12Illumina. Pipeline CASAVA User Guide 15003807 (Pipeline V. 1.4 and Casava V.1.0).
Sources of variation

Phasing example\textsuperscript{13}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example.png}
\end{figure}

Source of variation

Sources of variation

GC bias\textsuperscript{15}

---

\textsuperscript{15} Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry. “Overcoming bias and systematic errors in next generation sequencing data”. In: Genome Medicine 2.12 (2010). PMID: 21144010.
Random primers bias\textsuperscript{16}

Library type\textsuperscript{17}

Batch effects\textsuperscript{18}

\textsuperscript{18}Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry. “Overcoming bias and systematic errors in next generation sequencing data”. In: Genome Medicine 2.12 (2010). PMID: 21144010.
Biological variability\textsuperscript{19}

The future

- Further improvements in library preparation
- Single cell sequencing
- Third generation sequencers like Pacific Biosciences

And biostatistical methods =)
Thanks!

- Google Calendar
  https://www.google.com/calendar/embed?src=7hprep991i5prdi5l5ftksbsfb8%
  40group.calendar.google.com&ctz=America/New_York

- Slides at http://www.biostat.jhsph.edu/~lcollado/misc/HTSintro.pdf