Sequencing techniques and applications

Yuzhen Ye (yye@indiana.edu)
School of Informatics & Computing, IUB

Contents

- Sequencing techniques
  - Sanger sequencing
  - Next generation sequencing (NGS)
  - Third generation sequencing
- Base calling and PHRED quality score
- FASTQ format

DNA sequencing technology

- Sanger sequencing
  - The main method for sequencing DNA for the past thirty years!
- 2nd generation sequencing techniques (next generation sequencing)
  - Differ from Sanger sequencing in their basic chemistry
  - Massively increased throughput
  - Smaller DNA concentration
  - 454 pyrosequencing, Illumina/Solexa, SOLiD
- 3rd generation (single-molecule)

Next-generation sequencing transforms today’s biology

- Genome sequencing
- Genome re-sequencing
- Metagenomics
- Transcriptomics (RNA-seq)
- Personal genomics ($1000 for sequencing a person’s genome)

DNA sequencing: history

**Sanger method** (1977): labeled ddNTPs terminate DNA copying at random points.


Both methods generate labeled fragments of varying lengths that are further electrophoresed (electrophoretic separation).

10/6/13

**Radioactive sequencing versus dye-terminator sequencing**

dNTPs (chain terminators) are labeled with different fluorescent dyes, each fluorescing at a different wavelength.

**New sequencing techniques**

- Next Generation Sequencing (NGS) (Second Generation)
  - Pyrosequencing
  - Illumina
  - SOLID
- Third generation sequencing
- Single-molecule sequencing technologies
- NHGRI funds development of third generation DNA sequencing technologies
  - "More than $18 million in grants to spur the development of a third generation of DNA sequencing technologies was announced today by the National Human Genome Research Institute (NHGRI). The cost to sequence a human genome has now dipped below $40,000. Ultimately, NHGRI's vision is to cut the cost of whole-genome sequencing of an individual's genome to $1,000 or less, which will enable sequencing to be a part of routine medical care."

2013 NGS field guide


<table>
<thead>
<tr>
<th>Instrument</th>
<th>Frugent Cost</th>
<th>Report Cost A</th>
<th>Report Cost B</th>
<th>Minimum Unit Cost ($/100k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3730xl (capillary)</td>
<td>$144</td>
<td>$2.308</td>
<td></td>
<td>$6 (1%)</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>$6,200</td>
<td>$12</td>
<td>$2,000 (12%)</td>
<td></td>
</tr>
<tr>
<td>PacBio RS</td>
<td>≥ $300</td>
<td>$2-17</td>
<td>$500 (100%)</td>
<td></td>
</tr>
<tr>
<td>Ion Torrent – '316'</td>
<td>$739</td>
<td>$1.20</td>
<td>$-1,000 (100%)</td>
<td></td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>$1,049</td>
<td>$0.70</td>
<td>$-1,400 (100%)</td>
<td></td>
</tr>
<tr>
<td>Ion Torrent – '318'</td>
<td>$939</td>
<td>$0.60</td>
<td>$-1,200 (100%)</td>
<td></td>
</tr>
<tr>
<td>Ion Torrent – Proton I</td>
<td>$1,059</td>
<td>$0.09</td>
<td>? (100%)</td>
<td></td>
</tr>
<tr>
<td>SOLiD – 5500x</td>
<td>$10,553*</td>
<td>&lt; $0.07</td>
<td>$2,000 (12%)</td>
<td></td>
</tr>
<tr>
<td>Illumina HiSeq 2500 – rapid</td>
<td>$6,145*</td>
<td>$0.05</td>
<td>? (55%)</td>
<td></td>
</tr>
</tbody>
</table>
Pyrosequencing

- Pyrosequencing principles
  - the polymerase reaction is modified to emit light as each base gets incorporated.

Roche (454) GS FLX sequencer

Solexa/Illumina sequencing

- Ultrahigh-throughput sequencing
- Keys
  - attachment of randomly fragmented genomic DNA to a planar, optically transparent surface
  - solid phase amplification to create an ultra-high density sequencing flow cell with > 10 million clusters, each containing ~1,000 copies of template per sq. cm.
- Short reads
- Used for gene expression, small RNA discovery etc

Solexa/Illumina sequencing


Applied Biosystems SOLiD sequencer

- Commercial release in October 2007
- Sequencing by Oligo Ligation and Detection
- ~5 days to run / produces 3-4Gb
- The chemistry is based on template-directed ligation of short, “dinucleotide-encoding”, 8-mer oligonucleotides. Dinucleotide-encoding permits discrimination of SNP’s from most chemistry and imaging errors, and subsequent in silico correction of those errors.

Ref: http://appliedbiosystems.cnpg.com/Video/flatFiles/699/index.aspx

3rd generation sequencing

- PacBio “SMRT” (Single-Molecule sequencing in Real Time)
- Helicos tSMS
**Single-cell sequencing**

- "These sequencing-based technologies are increasingly being targeted to individual cells, which will allow many new and longstanding questions to be addressed."
- single-cell genomics to uncover cell lineage relationships
- single-cell transcriptomics to supplant the coarse notion of marker-based cell types
- single-cell epigenomics and proteomics to allow the functional states of individual cells to be analysed.


**Some video clips**

- Ion Torrent (http://www.youtube.com/watch?v=MxkYa9XCyBQ)
- Illumina MiSeq (http://www.youtube.com/watch?v=t0akxx8DwsK)
- 1st, 2nd, and 3rd generation genome sequencing technologies (http://www.youtube.com/watch?v=_ApDinCBt8g)

**Base calling**

- Determine the sequence of nucleotides from chromatograms or flowgram (trace files often in SCF format)
- Peak detection
- Phred quality score

**Phred quality score**

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

<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>1/10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1/100</td>
<td>99%</td>
</tr>
</tbody>
</table>

\[ Q_{\text{max}} = -10 \log_{10}\left(\frac{P_e}{1-P_e}\right) \]

\[ Q_{\text{min}} = -10 \log_{10}(10^{Q_{\text{max}}-10} + 1) \]

\[ Q_{\text{min}} = -10 \log_{10}(10^{Q_{\text{max}}-10} + 1) \]

(for high values the two scores are asymptotically equal)

**FASTQ format**

- FASTQ format is a text-based format for storing both a DNA sequence and its corresponding quality scores.

- The quality score for each base is encoded with a single ASCII character
- Phred quality score from 0 to 93 is encoded with ASCII 33 to 126

**Next generation sequencing as a public health tool**

- ~0.25% of US women (375,000) carry a mutation in BRCA1/2 – at very high risk of breast and ovarian cancer
  - 85% lifetime breast cancer risk
  - 25-50% lifetime ovarian cancer risk

- Knowledge of risk allows prevention
  - Currently we only can identify such women once several family members have developed cancer

- NGS allows population screening for high risk preventable disorders
  - Cancer predisposition, cardiac disease, etc.
  - ~1-2% of population carry such mutations

- ~3-6 million individuals in the US with preventable disorders if identified

**Doctors Urge Women To Test For Breast Cancer Gene After Jolie's Mastectomy**

- preventive double mastectomy---surgery to remove both breasts without a cancer diagnosis
Challenges to harnessing NGS in clinical medicine & public health

- Accuracy
  - 99.99% accuracy x 3 billion nucleotides
  - ~300,000 errors per patient
- Interpretation of the variants we find
- Storage and access in the medical record
- Education of patients and public
- Issues of consent and reporting
- Education of providers

Readings

- The next-generation sequencing revolution and its impact on genomics.
  - Cell. 2013 Sep 26;155(1):27-38
- Next-generation sequencing in the clinic: are we ready?