Next-Generation Sequencing Technologies

Nicholas E. Navin, Ph.D.
MD Anderson Cancer Center
Dept. Genetics
Dept. Bioinformatics

Introduction to Bioinformatics
GS011062
Generations of Sequencing Technologies

**Zero-Generation**
- **1970s-80s**
  - Sanger Chain Termination Radiolabelling
  - Output: 700bp / day

**First-Generation**
- **1980s-90s**
  - Automated Fluorescent Capillary Sequencers
  - ABI3700
  - Output: 384kb / day

**Second-Generation**
- **2000s**
  - Massively Parallel Sequencing
  - 454 Pyrosequencer
  - Output: 1 gigabase / day
Next-Generation Sequencing Technologies

**Second-Generation**
- 2005-2012
- Illumina HiSeq2000
- Output: 150gb / week
- Read Length: 100bp

**Third-Generation**
- 2010
- PacBio Real-time sequencing
- Output: 1gb /day
- Read Length: 10kb

**Third-Generation**
- 2012
- Oxford Nanopore Sequencing
- Output: ?
- Read Length: 40kb
Today we will begin to learn how to do the $100,000 analysis
Initially, all genome sequencing was performed at Genome Sequencing Centers.
Today NGS can be performed at most Universities

Many Universities (including MD Anderson) have sequencing core facilities that can sequence DNA or RNA samples provided by users

Machines such as the Illumina HiSeq 2000 can generate:

~150 gigabases per week

Very low error rates: 0.01%

~2 human genomes at 40-50X coverage

**Cost of sequencing**
Genome: $5000
Exome: $1000
Transcriptome: $500
Next-Generation Sequencing Platforms

Which sequencing platform to choose?
Technical Considerations

Run Time

Sequencing Error Rates

Amount of Data Output

Library Input Requirements

Hardware Access

Sample Multiplexing

Oxford Nanopore gridION

Life Sciences Ion Proton

Complete Genomics

Illumina HiSeq2000

PacBio RS
### Technical Considerations

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Primary Errors</th>
<th>Final Error Rate %</th>
<th>Data Output Per Run</th>
<th>Read Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB 3730 (Capillary)</td>
<td>Substitutions</td>
<td>0.1-1</td>
<td>48 Mb</td>
<td>700</td>
</tr>
<tr>
<td>454 GS FLX+</td>
<td>Indels</td>
<td>1</td>
<td>0.7 Gb</td>
<td>300-500</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>Substitutions</td>
<td>0.1</td>
<td>300 Gb</td>
<td>150</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>Indels</td>
<td>1</td>
<td>10 Gb</td>
<td>100-200</td>
</tr>
<tr>
<td>SOLID 5500xl</td>
<td>A-T Bias</td>
<td>0.1</td>
<td>180 Gb</td>
<td>50</td>
</tr>
<tr>
<td>Oxford Nanopore</td>
<td>Deletions</td>
<td>4</td>
<td>variable</td>
<td>10-40kb</td>
</tr>
<tr>
<td>PacBio RS</td>
<td>CG Deletions</td>
<td>15</td>
<td>0.1 Gb</td>
<td>10,000</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>Indels</td>
<td>0.0001</td>
<td>60 Gb</td>
<td>35</td>
</tr>
</tbody>
</table>
Pyro Sequencing

**Systems:** GS Junior, GS FLX by Roche 454
**Advantages:** Long read length (300-500bp), short run time (10 hours)
**Disadvantages:** Cannot resolve homopolymers, resulting in false indels; Data output is limited to ~700mb per run; emulsion pcr library prep
Semi-Conductor Sequencing

**Advantages:**
- Highly Scalable
- Short run time: 2 hours
- 10gb data output

**Disadvantages:**
- Cannot resolve homopolymers (resulting in false indels)

**Systems:** Ion Torrent, Ion Proton (Life Technologies)

**Advantages:**
- Highly Scalable
- Short run time 2 hours
- 10gb data output

**Disadvantages:**
- Cannot resolve homopolymers (resulting in false indels),
- Emulsion PCR library prep
Sequencing by Synthesis (Illumina)

**Systems:** HiSeq2000, MiSeq, GA2 (Illumina)

**Advantages:**
- High data output (150gb per run),
- Paired-end sequencing, inexpensive,
- Resolves homopolymers well

**Disadvantages:**
- Long runtime (1 hour per base), must run 8 lanes in parallel (hiseq)
Real-time Sequencing

Zeptoliter (1e-20) chamber uses Zero Mode Wavelength (ZMW) camera

**Advantageous:** Run time occurs in real time; several hours; Read length of 5-10kb; Can detect epigenetic modification to DNA; single molecule sequencing.

**Disadvantageous:** High error rate 15%; low data output: 50mb
**Advantageous:** Very long read lengths (40kb), direct single molecule sequencing, low-throughput USB device will be released

**Disadvantageous:** High error rate 5%, Probably many other problems, but no papers have been published yet

[Oxford Nanopore Video]
NGS Can Detect Different Classes of DNA Mutations

Sequencing throughput has increased from 700 bases per run to 150 gigabases per run on the Illumina HiSeq2000. (Meyerson et al., Nature 2011)
Applications of NGS

DNA
• Copy Number Amplifications and Deletions
• Structural Variations (translocations, inversions)
• Indels (insertions and deletions)
• Point mutations (nonsense, missense, splice-site)

RNA
• RNA expression levels (mRNA, microRNA, lincRNA)
• Alternative Transcripts
• Fusion transcripts

Protein Binding
• Chip-Seq

Epigenetic
• Bisulfite Sequencing
• Third Generation Sequencing Platforms
Illumina Next-Generation Sequencing
Illumina Sequencing Approaches

67% of the market uses Illumina sequencing platforms for NGS experiments.

Workflow for Illumina Sequencing:
- Sample Preparation
- Cluster Generation
- Sequencing
- Run Monitoring
- Data Processing
Single-End, Paired-End or Mate-Pairs?

**Single-End**
- DNA seq
- RNA seq
- Chip-Seq

**Paired-End**
- Genome Sequencing
- Genome Assembly
- Detection of Chromosome rearrangements
- Useful when more data is required; pooled exomes

**Mate-Pairs**
- Genome Assembly
- Resolving genomic regions with complex repetitive elements
Library Construction

Single or Paired-End Libraries

1) Genomic DNA
2) Fragment (200–500bp)
3) Ligate Adaptors

Insert size = 200-500bp

Mate-Pair Sequencing

1) Mate Pair Library Sequencing for Long Inserts
   Genomic DNA
2) Fragment (2–5 kb)
3) Bio
   Biotinylated ends
4) Circularize
5) Fragment (400–600 bp)
6) Enrich biotinylated fragments
7) Ligate adaptors

Insert size = 2000-5000bp
Considerations: Read Length

**Short Reads (36-76bp)**
Useful for ‘Counting Applications’:

- **RNA-Seq** for determining expression levels from read counts
- **Copy Number Profiling** measuring copy number from read count density
- **Chip-Seq** determining protein binding by read density
- Also, **targeted sequencing** of small genomic regions

**Long Reads (100-150bp)**
Useful for:

- **RNA-Seq** to detect alternative splice forms
- **Exome sequencing** in which more data is needed
- **Genome Assembly**
- **Genome Sequencing**
- **Chromosome Structural Rearrangement**

*Important: The read length must be shorter than the library insert size, otherwise many reads will sequence into the adapters*
During library construction the insert size of the DNA libraries can be selected by gel extraction.

The insert size has important implications for sequencing, and determining which read length to use, 200-400bp is common.

Note: the insert size must be larger than the read length, otherwise the adapters will be sequenced – this will cause errors in alignment.
Library Construction

DNA Fragmentation

8bp Indexes allow samples to be pooled together in a single lane
Indexing (Barcoding) Libraries for Pooled Sequencing Reactions

- Index = 8bp unique sequences that is added to each library
- Each library contains a unique index sequence
- Typically samples are pooled together using 4-12 samples

**advantages:** major cost and time savings
For many research projects it may be advantageous to sequence only the coding regions of the human genome (the exome).

This can be achieved using exome capture kits (Illumina, Agilent, Nimblegen) to capture DNA in exonic regions by hybridizing to biotinylated probes before sequencing.
Flowcell surface is coated with a lawn of primers that hybridize to the library adapters.

During the sequencing run, reagents are flown through the chambers after each cycle.

Each Flowcell has 8 lanes for sequencing 8 different samples.
Sequencing libraries are hybridized to a flowcell surface and clonally amplified so that they can be imaged by fluorescent microscopy on the HiSeq2000 Illumina.
Reversible Terminator Chemistry

All 4 labeled nucleotides in 1 reaction
Very High accuracy/fidelity
Base by base incorporation
No problems with homopolymer repeats
- Modified bases with reversible terminators and fluorophores are incorporated during each cycle.
- The fluorophore is excited and TIF images are collected across the surface of the flowcell tiles for base-calling.
- The fluorophore and terminator are then cleaved and the next base is incorporated.
Paired-End Sequencing Run

1. Genomic DNA
2. Fragment (200–500bp)
3. Ligate Adaptors
4. Generate Clusters
5. Sequence First End
6. Regenerate Clusters and Sequence Paired End
Run Metrics: Cluster Density

Excessive cluster density makes it difficult for the image analysis algorithm (RTA) to detect clusters for base calling.

Important: to quantify the DNA library concentration accurately before sequencing (qPCR, Bioanalyzer)
Illumina Sequencers Are Fluorescent Microscopes
Base Quality Scores

Q = -10 log 10 (e)

e = estimated probability base call is wrong

<table>
<thead>
<tr>
<th>Q Score</th>
<th>Prob Incorrect</th>
<th>Base 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>
</tbody>
</table>

Base quality scores are encoded in ASCII characters representing numbers 0-41 (QSCORE +33)

ACTCAGCATCATCATTTCATTATTTAGTA
DDCCBABA@@A?A?>>+++<; ;; :: 98765
Quality scores decrease at the ends of reads due to fluorescence quenching after each cycle.
**Coverage depth** – mean number of read counts across all the bases of a sequenced sample. Usually denoted by $X$:

- ex. 30X coverage depth

**Coverage breadth** – % of the genome sequence at $\geq 1X$ read coverage. Usually denoted by %:

- ex. 80% coverage breadth

*These metrics are calculated using reads with unique mappings only*
Coverage depth normally follows a Poisson distribution.
Coverage Depth

Using a Poisson probability distribution it is possible to calculate how many reads will have X amount of target coverage:

**Example:** what is the probability that a site will be covered exactly ten times if the sample has 5X mean coverage depth?

\[
f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},
\]

\[
5^{10} \times e^{-5} / 10! = 0.018
\]

1.8% of the bases will have exactly 10X coverage.
How Deep Should I Sequence?

Depends largely on how rare the expect alleles are in the sample

For a normal genome, most allele frequencies are 1.0 (homozygous) and 0.5 (heterozygous), so 30X coverage will provide 100% power of detection

\[ P(d) = 1 - (1 - a)^c \]

However in tumors (where heterogeneity is common) alleles may occur at 10% or 1% frequencies, which requires 50X or 500X coverage to detect with 100% power
Coverage Breadth

- Normally coverage breadth will approach 90% in a human genome sample, when reads are mapped with unique coordinates.
- About 10% of the genome cannot be mapped uniquely due to repetitive elements.
- Alternatively coverage breadth can be calculated using multiple mapped reads, in which case it approaches 100%.

Good coverage breadth of the genome calculated using uniquely mapped reads.

Poor coverage breadth.
Data Processing
Pipelines
**Sequencing & Base Calling**

- Library Construction
- Barcoding
- Cluster Generation
- Sequencing Run
- Generate TIF Images
- RTA Cluster Detection
- Base Calling
- Deconvolution of Indexes
- Generate FASTQ files

**Data Processing**

- Align FASTQ files to Human Genome
- Convert to SAM file
- Convert to BAM file (binary file)
- Sort Reads by Chromosome Position
- Remove PCR Duplicates
- Variant Calling (GATK, Varscan)
- Variant Annotation (Annovar, COSMIC)
- Prediction Algorithms (SIFT, Polyphen)
Read Alignment

• A sequencing run generates millions of reads that need to be aligned to the human genome (assembly HG18, HG19) in order to determine their chromosome positions

• In a typical experiment, between 80-90% of the reads will map uniquely to the human genome

• However, some samples, such as mouth swabs may contain significant contamination from bacterial DNA, which can decrease the number of reads that align to the human genome

Popular Short-Read Alignment Software:

BWA  http://bio-bwa.sourceforge.net
Bowtie  http://bowtie-bio.sourceforge.net

**MAPPING ERRORS**: short read aligners have a higher number of mapping errors compared to pair-wise alignment methods (BLAST, Smith-Waterman). Poorly mapped reads can be filtered by mapping quality score before variant detection to eliminate false-positives
Removing PCR Duplicates

During library construction, PCR is used to enrich the DNA before sequencing, which can generate duplicate reads with false-positive mutations.

- PCR duplicates usually constitute about 10% of the total sequence reads.
- They can be removed informatically in post-processing, because they have identical start and stop coordinates.
Sequencing Errors and Stand Bias

- Sequencing errors can be randomly distributed (1) in single reads or expanded across multiple reads (2,3)
- Expanded errors are dangerous because they can be called as real mutations by variant calling algorithms

One way to mitigate these errors is to look at **strand bias**
- Normal mutations should occur on both + and – strands with equal frequencies
- If the mutation occurs on only one strand (2,3) then it is likely an error
• In most genomes GC content follows a normal distribution
• However, PCR and sequencing can generate GC bias in which reads with lower GC content are amplified or sequenced more frequently (right panel)
Indel Alignment Errors

- False-positive mutations frequently occur in regions with indels.
- To remove these errors, it is necessary to realign the sequences around the indels using a more accurate alignment algorithm.
Summary

- Select the appropriate sequencing platform for your project and consider what type of biological information you are interested in measuring (DNA, RNA, CHIP, Epigenetics).

- Sequencing technologies are constantly evolving, understand the caveats and advantageous of each system. The newest platform is not always best.

- Consider the appropriate read length, single or paired end sequencing and whether indexing or exome capture is advantageous for your study.

- Determine what target sequence depth is appropriate for your sample and calculate the number of lanes necessary to achieve that depth.

- Be aware of false-positive technical errors (PCR amplification, sequencing errors, mapping errors) and how to mitigate them in your processing pipeline.
Acknowledgements

Several slides in this lecture were provided by:

Jude Kendall (Cold Spring Harbor Laboratory)
Michael Schatz (Cold Spring Harbor Laboratory)
Yong Wang (MD Anderson)
Illumina Inc. (San Diego)