Introduction to Next Generation Sequencing
BMI7830

Gulcin Ozer, PhD

gulcin.ozer@osumc.edu

Department of Biomedical Informatics

The Ohio State University

October 29th, 2015
What is NGS?

- **Next Generation Sequencing**
- Massively parallel high-throughput sequencing
- Refer to a number of different modern sequencing technologies including:
  - Illumina (Solexa) sequencing
  - SOLiD sequencing
  - Ion torrent: Proton / PGM sequencing
  - Roche 454 sequencing
Sequencing Platforms

Illumina HiSeq
- Leading platform
- Bridge amplification
- Sequencing-by-synthesis

Illumina MiSeq
Launched in 2011
Benchtop sequencer
Lower throughput
Fast turnaround time
Cheap

A sequencing room at Hong Kong
Sequencing Platforms

SOLiD
- Emulsion PCR (beads)
- Sequence by ligation
- Color space

Ion Torrent
Launched in Feb 2010
Ion-sensitive semi-conductor based technology
Cheap, lower throughput
Higher error rates

Ion PGM (Personal Genome Analyzer)

Ion Proton
Sequencing Platforms

Roche 454
Pyrosequencing
Longer reads 400-1Kb
First parallel sequencing platform
Cost per Raw Megabase of DNA Sequence

Moore's Law

National Human Genome Research Institute

genome.gov/sequencingcosts
What to Sequence?

Sequence DNA
- *De novo* sequencing
- Reference-based re-sequencing
  - Whole genome
  - Exome
  - Targeted
- Metagenomics
  - Identify “who is there?” in a mixture of microbes

Study Protein-DNA/RNA interaction
- ChIP-seq (for TF, Pol II binding)
- CLIP-seq (for RNA binding proteins)

Sequence RNA
- RNA-seq (transcriptome-wide sequencing)
- smRNA-seq
- novel ncRNAs

Epigenetics
- DNA methylation
- Histone modification (ChIP-seq)
- Nucleosome positioning
- Chromosome looping
NGS data analysis

High-throughput Sequencers

Gene Expression: RNA-seq
- Expression quantification
- Differential expression
- Alternative splicing
- Non-coding RNA
- Variant detection
- Allelic expression imbalance

Epigenome: ChIP-seq, Methyl-seq
- Alignment
- Peak detection
- Comparative analysis
- Visualization

Resequecing: Exome-seq, Whole Genome Sequencing
- Alignment
- SNP calling
- Small and large INDELs
- Copy number variation
- Gene fusion/translocation

Small RNA-seq
- Alignment
- Expression and Counts
- Known miRNAs
- Novel miRNAs

Integration • Visualization • Hypothesis Discovery

Mining and integration of public datasets and annotations
Illumina Sequencing

1- Sample preparation
2- Cluster generation
3- Sequencing by synthesis
4- Image processing and base calling
4-Illumina Sequencing – Image processing and base calling

TGCTACGAT...
Illumina Sequencing Data

- Two flow cells (HiSeq 2500)
- 8 lanes per flow cell
- ~250 million sequences (reads) per lane
  - 50-150 nt each sequence
- Single reads or paired reads
- Multiplexing is often used to allow multiple samples on the same lane of Illumina sequencer
  - Up to 96 samples per lane
FASTQ is the most commonly used format

Line 1-Identifier

Line 2-Sequence read

Line 3-Continue with quality information. Identifier is optional

Line 4-Quality string
**Sequence Data Quality**

### Phred Score

$$ Q = -10 \log_{10} P $$

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of Incorrect Base Call (P)</th>
<th>Base Call 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>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
</tbody>
</table>

### Encoding Quality Scores

ASCII character encoding system

Illumina v1.8 and later uses ASCII base of 33

<table>
<thead>
<tr>
<th>Character</th>
<th>Quality Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>'A'</td>
<td>32</td>
</tr>
<tr>
<td>'&lt;'</td>
<td>27</td>
</tr>
<tr>
<td>'I'</td>
<td>40</td>
</tr>
</tbody>
</table>

Quality = ord(<character>) - 33
Sequencing Data Size

Size of typical sequencing datasets

- Whole exome sequencing
  >60 million reads for 50X coverage

- Whole genome re-sequencing
  >1.5 billion reads for 50X coverage

- Transcriptome (RNA) sequencing
  ~30-60 million reads (expression profile)
Quality Control
Per base sequence quality

High quality reads
Quality Control
Per base sequence quality

Low quality reads
Quality Control
Per base sequence quality

High quality at the beginning, lower quality towards the end of the reads
Trimming would improve results
Sequencing Data Alignment

Reference Genome
(3 billion bases for human)

Sequence Reads
(millions of short reads)

Alignment/mapping
Sequence reads/short reads/reads

Reference sequence
Sequence read 1
Sequence read 2
Sequence read 3
Sequence read 4
Sequence read 5

CGTAACCTTGGCTGATGTGCGCCGCCTACTTCGGTGGTGAAGGTG
CTGATGTGCCGCCTCACTTCGGTGGT
TGATGTGCCGCCTCACTACGGTGGT
GATGTCGCCGCCTCACTTCGGTGGTG
GCTGATGTGCCGCCTCACTACGGT
GCTGATGTGCCGCCTCACTACGGTG
Sequencing Data Alignment

- **Optimal local alignment**
  - Optimal alignment between a query sequence and a database
  - Smith-Waterman dynamic programming

- The reference sequence, the short reads, or both, are pre-processed into an indexed form for rapid searching

**Slow**
Sequencing Data Alignment

- Hash table based alignment methods (similar to BLAST)
  - Eland (Illumina)
  - SOAP (Li et al, Bioinformatics 2008)
  - Maq (Li et al, Genome Research 2008)

- Burrows-Wheeler Transformation (BWT) based alignment methods
  - BWA (Li and Durbin, Bioinformatics 2009)
  - Bowtie (Langmead et al, Genome Biology 2009)
  - SOAP2 (Li et al, Bioinformatics 2009)
Burrows-Wheeler Transform

**BWT to be a useful pre-processor for lossless data compression**

Allows an efficient inverse transformation that can fully recover the original sequence from the transformed output.

Flicek and Birney, Nature Methods 2009
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.
Visualization in IGV
Variant Detection

Mismathces in alignment results

- **Potential error**
  - PCR artifacts (errors introduced in early PCR rounds)
  - PCR duplicates (same read occurs multiple times)
  - Sequencing errors (erroneous base calls)
  - Mapping errors (mapping to a wrong location in the reference)

- **Variant**
  - Single nucleotide variant
  -Allele frequency
  -Population frequency
    - If common in population → SNP
    - dbSNP database
  - Functional variant or not
Whole Exome Sequencing (WES)

- Utilize a protocol to fish out only coding DNA sequences
- Create sequencing libraries from enriched DNA
- Reduced cost and analysis time compared to whole genome sequencing
WES – Quality Control

- Picard HSmetrics
  - PF unique reads
  - PF unique bases aligned
  - On target bases
  - Percent selected bases
    - Percentage of the PF aligned bases on or near the target
    - > 80%
WES Analysis

**GATK** *(Genome Analysis Tool Kit from Broad Institute)*

Best practice workflow

1. FASTQC: Quality Control
2. BWA: Map to reference genome
3. Picard: Sort and mark duplicates
4. GATK: Indel Realignment
5. GATK: Base Recalibration
WES – Variant Calling

Tumor/Normal Pair

N FASTQ File

GATK Best Practice Workflow

N BAM File

Somatic Variant Calling

MuTect VarScan2

VCF File

T FASTQ File

GATK Best Practice Workflow

T BAM File
Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib

Jennifer A. Woyach, M.D., Richard R. Furman, M.D., Ta-Ming Liu, M.S., Hatice Gulcin Ozer, Ph.D., Marc Zapatka, Ph.D., Amy S. Ruppert, M.A.S., Ling Xue, Ph.D., Daniel Hsieh-Hsin Li, Ph.D., Susanne M. Steggerda, Ph.D., Matthias Versele, Ph.D., Sandeep S. Dave, M.D., Jenny Zhang, B.S., Ayse Selen Yilmaz, M.S., Samantha M. Jagilowski, M.D., M.P.H., Kristie A. Blum, M.D., Arletta Lozanski, M.S., Gerard Lozanski, M.D., Danelle F. James, M.D., Jacqueline C. Barrientos, M.D., Peter Lichter, Ph.D., Stephan Stilgenbauer, M.D., Joseph J. Buggy, Ph.D., Betty Y. Chang, Ph.D., Amy J. Johnson, Ph.D., and John C. Byrd, M.D.

Normal B cell signaling

Extracellular

Intracellular

Antigen

B cell receptor

Mediators

BTK

PLCγ2

PIP3

IP3

DAG

Intracellular Calcium store

Ca

Ca

Nucleus

Proliferation

Apoptosis

Anergy

P

ERK

Akt
CLL B cell signaling

Antigen

B cell receptor

Mediators

PIP3

IP3

Intracellular Calcium store

Ca

BTK

PLCγ2

DAG

Nucleus

Proliferation

Pro-survival

ERK

Akt
• Ibrutinib is an irreversible molecular inhibitor of BTK
• It is an effective and safe treatment option for CLL patients
• Imbruvica FDA approved in February 2014
CLL B cell signaling with Ibrutinib

- Antigen
- B cell receptor
- PLCγ2
- IP3
- DAG
- Intracellular Calcium store
- Ca
- Nucleus
- Apoptosis
- PIP3
- ERK
- Akt
- BTK
- Mediators

Ibrutinib (Ib) inhibits BTK activity.
Resistance to Ibrutinib Treatment

- Only a small portion of patients have had a relapse during ibrutinib treatment.
- We carried out whole exome sequencing (WES) at the baseline and the time of relapse.
- We identified two functional mutations that result in acquired resistance to ibrutinib therapy.
- Functional characteristics of both mutations are studied in detail.
Whole-Exome Sequencing Experiments

- Whole exome-sequencing
  - 6 patients at the baseline and the time of relapse
- 100bp PE Illumina sequencing
  - ~ 91 million reads per sample
  - ~ 63X exonic coverage
Exome-Seq Data Analysis Pipeline

1. Merge lanes per Primary/Relapse Samples & QC
2. Alignment (bwa 0.7.5) Genome Reference Consortium Build37
3. Duplicate read removal - Picard 1.94
4. Candidate Indel Search
5. Aligned reads for Primary samples
6. Realign around indels (gatk 2.8.1)
7. Candidate Indels
8. Known Indels
9. Variant Calling - indel: gatk somatic indel caller - snp: mutect
10. Post-Call Filtration
11. Relapse Specific Variants
12. Aligned reads for Relapse samples
13. Realign around indels (gatk 2.8.1)
Variant Analysis (Variant Prioritization)

- Variant Calling
  - Identification of relapse specific variants
    - ~300,000 variants
    - ~9,500 relapse specific variants per sample
  - Novel variants
    - ~610 novel variants
  - Variants with functional consequence
    - 89 variants across all samples
  - Common variants
    - 2 common gene variants
Ibrutinib resistance

Antigen

B cell receptor

Mediators

C481S

5 patients

R665W
L845F
S707Y

BTK

PLCγ2

PIP3

IP3

DAG

Intracellular Calcium store

Ca

Nucleus

Pro-survival

Proliferation

ERK

Akt

P
What is next?
Create a Galaxy Account
https://usegalaxy.org/
Install IGV
{requires Java}

https://www.broadinstitute.org/software/igv/home
Gulcin Ozer, PhD
340C Lincoln Tower
1800 Cannon Dr.

Gulcin.Ozer@osumc.edu

Wexner Medical Center