Week 8: Next-Gen Sequencing

RNA-seq Data Analysis

Babu Guda, Ph.D.
Professor, Genetics, Cell Biology & Anatomy
Director, Bioinformatics and Systems Biology Core
University of Nebraska Medical Center

Sanger vs Next-Gen Sequencing

Next-Gen Sequencing

NGS technologies

- No in vivo cloning

Cost of Human Genome Sequencing

<table>
<thead>
<tr>
<th>Year</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>$3,000,000,000</td>
</tr>
<tr>
<td>2006</td>
<td>$20,000,000</td>
</tr>
<tr>
<td>2007</td>
<td>$2,000,000</td>
</tr>
<tr>
<td>2008</td>
<td>$200,000</td>
</tr>
<tr>
<td>2010</td>
<td>$10,000</td>
</tr>
<tr>
<td>2014</td>
<td>$1,000</td>
</tr>
<tr>
<td>2017</td>
<td>$100</td>
</tr>
</tbody>
</table>

Source: https://bloggenohub.files.wordpress.com/2015/01/slide1.jpg

Applications of NGS

- Genome
  - Whole genome sequencing
  - Whole exome sequencing
  - Targeted gene panels (cancer, newborns, autism, etc.)
- Transcriptome
  - Whole RNA sequencing
  - mRNA transcriptome (poly-A selection)
  - Small RNA analysis (siRNA, snoRNA, lincRNA, etc.)
  - Gene expression profiling for selected target genes
- Metagenome
  - Bulk sequencing of many types of bacteria
  - Examples: human gut microbiome, soil samples, food contamination, extremophiles, etc.
- Epigenome
  - Chromatin Immunoprecipitation Sequencing (ChIP-Seq)
  - Methylation Sequencing (Methyl-Seq)
Different Sequencing Libraries

- Single-End Reads - 5' or 3' (random)
- Paired-End Reads - 5' and 3'
  - 200-500 bp
- Mate-Pair Reads - 5' and 3'
  - 2-5 kbp

Source: http://slideplayer.com/7447747/35/images/7/Types+of+Sequencing+Libraries.jpg

Paired-end Sequencing

- Paired-End Reads
- Alignment to the Reference Sequence

Source: https://assets.illumina.com/content/dam/illumina-marketing/images/science/v2/web-graphic/paired-end-vs-single-read-sseq-web-graphic.jpg
FASTQ Files from Paired-end Sequencing

Demultiplexing Mixed Samples
Different File Types in NGS analysis

- **Fastq file** – generated by the sequencer, contains NGS reads
- **SAM file** – Sequence Alignment/Map (generated by aligning the NGS reads with the reference genome)
- **BAM file** – Binary version of the SAM file (SAMtools are used to manipulate SAM/BAM files)
- **GFF file** – General Feature Format used to hold genome annotation (chromosome, strand, frame, exon, CDS, etc.)
- **GTF file** – Gene Transfer Format (Also contains all the info as in GFF and in addition contains gene annotation information)
- **VCF file** – Variant Call Format (used to store variant data such as SNPs, InDels, short structural rearrangements)

**Fastq**

```
@SRR098401.11403008/1
GAGCTATAGCATGGTCAGAACAAGAAGATCACTGGACTGCCCTCGCTCAGCCCTCAGCTACTG
+
>>?>?@>?>@>?>@>?>@@>?@@=@@@@@?>@??@?@?@A?><@@@?><@@?@A@>@A@@A@@AAB@@BB
```

Row 1: Information from the sequencer about the location of this read on the plate
Row 2: The Sequence
Row 3: Metadata provided by the sequencing team
Row 4: Quality scores pertaining to each nucleotide in the sequence
FASTQ format:

FASTQ is based on the popular FASTA format for sequences

FASTA format
>sequence_ID; header in one line
AGTTGATGCTGATGATGGG

FASTQ format provides additional information that includes the quality score
@20FUKAAXX100202:1:64:10634:114560/1
TTGTATTTTTAGTAGAGACGGAGTTTCGCCATGTTGGTCAGGCTGGCCTCGAATTCCTGACCTCAAGTGATCCGCCCGCCTCGGCCTCCCAACGTTTGG
+
?=@7=>B==;;BB?<B?=8539<6?6>8>=BB<<B=08:9@5;:A@@?@9:BAAA<?;8;@AC@BBBBBA?<9

ASCII code for Quality score (Phred score, ranges from 0-50)

ASCII code for Quality score (in the increasing order; ! is the worst and ~ is the best)

Sequence Alignment / Map (SAM / BAM)

Similar to the Fastq file in that it contains the raw sequence and its quality scores.
It also tells you where the sequence aligned to the genome, and how well (this score is also phred-scaled).
In this case, this read aligned to chromosome 22, position 17445857, and has a quality score of 60 (or a 1 in 1,000,000 chance of being placed incorrectly).
**Variant Call Format (VCF)**

(a) VCF example

```
#fileformat=VCFv4.1
#filedate=2010/04/13
#source=VCFtools
#reference=file:///refs/human.NCBI36.fasta
#contig=chr1, length=249250621, md5=1b22b09dedeb4e9304b5d48b26a85128, species="Homo Sapiens"
#contig=chrX, length=155270560, md5=7b82c8892979b776e31dbb8c25460d, species="Homo Sapiens"
/INFO:ID=H2, Number=0, Type=Flag, Description="HapMap2 membership"
/INFO:ID=GT, Number=1, Type=Integer, Description="Genotype"
/INFO:ID=GQ, Number=1, Type=Integer, Description="Genotype Quality"
/INFO:ID=DP, Number=1, Type=Integer, Description="Read Depth"
/INFO:ID=DEL, Description="Deletion"
/INFO:ID=SVTYPE, Number=1, Type=String, Description="Type of structural variant"
/INFO:ID=END, Number=1, Type=Integer, Description="End position of the variant"

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 ACG A AT 40 PASS . GT:DP 1/1:3 2/2:29
1 2 C T CT PASS . GT:DP 0/1:3 2/2
X 85 rs12 A G 67 PASS . GT:DP 1/0:16 2/2:20
X 188 T <DEL> PASS . GT:DP SVTYPE=DEL;END=299 0/1:3 9/0:20:36
```

(b) SNP Alignment

<table>
<thead>
<tr>
<th>SNP</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>12345</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>ATGT</td>
<td>POS REF ALT</td>
</tr>
</tbody>
</table>

(c) Insertion

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>12345</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>ATGT</td>
<td>POS REF ALT</td>
</tr>
</tbody>
</table>

(d) Deletion

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>12345</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>ATGT</td>
<td>POS REF ALT</td>
</tr>
</tbody>
</table>

(e) Replacement

<table>
<thead>
<tr>
<th>Alignment</th>
<th>VCF representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>12345</td>
<td>POS REF ALT</td>
</tr>
<tr>
<td>ATGT</td>
<td>POS REF ALT</td>
</tr>
</tbody>
</table>

---

**RNA-Seq Data Analysis**
Computational Analysis of RNA-Seq Data

(a) Pre-analysis
- Experimental design
- Seabing design
- Quality control
- Library type
- Sequencing length
- Read identity and sequencing depth
- Spike-in
- RNA-Seq Data Analysis Workflow

(b) Core-analysis
- Transcription profiling
- Differential expression
- Interpretation
- Read alignment
- Transcript quantification
- Quantification measure
- Preprocessing
- Differential expression
- Alternative splicing analysis
- Functional profiling

(c) Advanced-analysis
- Visualization
- Other RNA-seq
- Integration
- Genomic browser, R scripts, etc.
- Small and other non-coding RNAs
- Gene fusion detection
- Long-read
- Single-cell analysis
- eQTL analysis
- ChIP-Seq (e.g., ATAC-seq)
- TF binding (e.g., ChIP-seq)
- Post-translational modifications

RNA-Seq Data Analysis Workflow

1. Raw sequence Reads (FASTQ)
2. QC and cleaning
   - FastQC, FQTrim
3. Mapping/Alignment
   - STAR, HISAT, TopHat, Sailfish, Salmon
4. Assembly and Quantification
   - Cufflinks, EdgeR, DESeq
5. Differential Exp. Analysis
   - CuffDiff, DESeq, DegeR, Limma
6. Functional Analysis/Annotation
   - GSEA, IPA, DAVID, GO, etc.
Input Files for RNA-seq Analysis

Galaxy Server
https://usegalaxy.org/

- A large compilation of open-source NGS data analysis tools that are accessible to users on web-based platforms
- Data can be uploaded from a PC/Mac and computing can be done on the cloud
- No need to install tools and maintain servers locally
- In-depth tutorials are available to use Galaxy services
- A list of Public Galaxy Servers can be found at
  - https://galaxyproject.org/public-galaxy-servers/
- Today’s RNA-seq analysis will be performed from the following link
  - https://bioinf-galaxian.erasmusmc.nl/galaxy/
Phred Score (Q) explained

Phred score (Q) vs Error probability (P)

\[ Q = -10 \log_{10} P \]

Phred quality scores are logarithmically linked to error probabilities

<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 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>99.3%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

Base Sequence Quality Interpretation

- **Bad Quality**: Quality drops at the tail end
- **Excellent Quality**: Base call accuracy is high
- **Bad Quality**: Error probability is high
Read Mapping and Assembly

Downstream Analysis of RNA-seq Results