Introduction to NGS and RNA-Seq

Dhivya Arasappan

(With some slides borrowed from Scott Hunicke-Smith and Jeff Barrick)
Some background

• Research scientist-bioinformatician at CBRS.
  – RNA-Seq
  – Genome Assembly
  – Exome data analysis
  – Benchmarking of tools

• Training
  – Grad students, post-docs.
  – Undergraduate- FRI
Goals of the Class

• When considering an RNA-Seq experiment
  – What kind of options are available for library prep?

• When you have an RNA-Seq dataset
  – What kind of options are available for analysis?

• Hands-on experience running typical RNA-Seq workflows on TACC
  – Some unix, R, TACC skills

• Learn the terminology
Setting General Expectations

• Lots of background and basics to provide comfort with terminology and key concepts.

• Exposure to commands and typically used analysis tools using an example RNA-Seq dataset.
  – No one ‘best’ or ‘standard’ tool.

• A starting point for you to design your RNA-Seq study or analyze your dataset.
Resources

• BioIteam Wiki- Bookmark it!
  https://wikis.utexas.edu/display/bioiteam

• Summer School course materials:
  https://wikis.utexas.edu/display/bioiteam/Introduction+to+RNA+Seq+Course

• Byte Club: Meets Third Wednesday of every month
  https://wikis.utexas.edu/display/bioiteam/Byte+Club

• CCBB Bioinformatics consultants
What is Next Generation(or) Second Generation Sequencing?

- Massively parallel sequencing
- The template DNA is attached to a cluster.
- Billions of clusters sequenced in parallel.
- 3-10 billion independent DNA fragments sequenced in one run.
So, what’s so great about second generation sequencing?

- + Sequence lots more, faster!
- + More cost effective.
So, what’s NOT so great about second generation sequencing?

- Data deluge
- Bioinformaticians and computational biologists to the rescue!
Who are the players?
How do next generation sequencers work?

- Attach a short DNA template on a chip.
- Flood with polymerase, fluorescent labeled nucleobases.
- When a complementary base is generated, take a picture of the fluorescence.
- Do this for millions/billions of templates at the same time.

Sequencers simply observe DNA Replication

http://www.cegat.de/
How do next generation sequencers work?
How does the sequencer work?

- Library prep
- Cluster generation/amplification
- Sequencing by synthesis
- Done in parallel for billions clusters at once.

Let’s watch the official Illumina video. 

http://www.cegat.de/
## Different Types of Illumina Sequencers

### Illumina Specifications Table

<table>
<thead>
<tr>
<th></th>
<th>HiSeq X Ten*</th>
<th>Hi Seq 2500</th>
<th>NextSeq 500</th>
<th>MiSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total output</td>
<td>1.8 Tb</td>
<td>1 Tb</td>
<td>129 Gb</td>
<td>15 Gb</td>
</tr>
<tr>
<td>Run time</td>
<td>3 days</td>
<td>6 days</td>
<td>29 hrs</td>
<td>~65 hrs</td>
</tr>
<tr>
<td>Output/day</td>
<td>600 Gb</td>
<td>167 Gb</td>
<td>~100 Gb</td>
<td>~5.5 Gb</td>
</tr>
<tr>
<td>Read length</td>
<td>2 X 150</td>
<td>2 X 125</td>
<td>2 X 150</td>
<td>2 X 300</td>
</tr>
<tr>
<td># of single reads</td>
<td>6B</td>
<td>4B</td>
<td>600M</td>
<td>25M</td>
</tr>
<tr>
<td>Instrument price</td>
<td>$1M*</td>
<td>$740K</td>
<td>$250K</td>
<td>$125K</td>
</tr>
<tr>
<td>Run price</td>
<td>~$12k</td>
<td>~$29k</td>
<td>~$4k</td>
<td>~$1.4k</td>
</tr>
<tr>
<td>$/Gb</td>
<td>$7</td>
<td>$29</td>
<td>$33</td>
<td>$93</td>
</tr>
</tbody>
</table>

[link](allseq.com/knowledgebank/sequencing-platforms/illumina)
Multiplexing

- Sample specific Indexes/Barcodes are attached to the DNA template.
- 6-8bp indexes/barcodes
- Data off the sequencer must first be demultiplexed to identify which reads belong to which sample.

https://doi.org/10.2147/BLCTT.S51503
What are the Limitations/Challenges?

- Amplification can cause problems.
  - Clusters are made by using PCR amplification.

- Reads are short
  - difficult to align, assemble.
  - too short to span long repeat regions.
  - Difficult to detect large structural variations like inversions.
What is RNA-Seq?

• Examine the state of the transcriptome.

• Genes expression patterns vary in:
  – Tissue types
  – Cell types
  – Development stages
  – Disease conditions
  – Time points

• RNA-Seq measures these expression variations using high-throughput sequencing technologies.
What is RNA-Seq?

• RNA-Seq measures these expression variations
  – At gene level
  – At isoform level
Other Uses of RNA-Seq

- Assembling and annotating a transcriptome
- Characterization of alternative splicing patterns
- Gene fusion detection
- Small RNA profiling
- Targeted approaches using RNA-Seq
- Direct RNA sequencing
Advantages of RNA-Seq

<table>
<thead>
<tr>
<th>Technology</th>
<th>Tiling microarray</th>
<th>RNA-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology specifications</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principle</td>
<td>Hybridization</td>
<td>High-throughput sequencing</td>
</tr>
<tr>
<td>Resolution</td>
<td>From several to 100 bp</td>
<td>Single base</td>
</tr>
<tr>
<td>Throughput</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Reliance on genomic sequence</td>
<td>Yes</td>
<td>In some cases</td>
</tr>
<tr>
<td>Background noise</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Application</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simultaneously map transcribed regions and gene expression</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Dynamic range to quantify gene expression level</td>
<td>Up to a few-hundredfold</td>
<td>&gt;8,000-fold</td>
</tr>
<tr>
<td>Ability to distinguish different isoforms</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Ability to distinguish allelic expression</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Practical issues</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Required amount of RNA</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Cost for mapping transcriptomes of large genomes</td>
<td>High</td>
<td>Relatively low</td>
</tr>
</tbody>
</table>

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein, and Michael Snyder

What are your questions?

• This determines how you set up your experiment and how you analyze the data.

• What are you looking for?
  – Annotating a transcriptome?
  – Differential expression?
    • Novel transcripts/isoforms, junctions?
    • Differential gene expression?
    • Differential exon level counts?
    • Differential regulation?
  – Small RNA?
RNA-Seq... at it’s Most Basic Form

Samples from two conditions

Isolate RNA

Generate cDNA

Create sequencing library by fragmenting, size selection and adding adaptors

Run sequencer

Generate short reads

Identify differentially expressed genes

negative regulation of programmed cell death
negative regulation of cell death
chromosome organization
regulation of cell proliferation
response to DNA damage stimulus
cell cycle process
programmed cell death
response to organic substance
response to cellular response to stress
RNA Illumina Tru-Seq library prep

**Figure 2: Optimized TruSeq RNA Sample Preparation**

A. Poly-A selection, fragmentation and random priming

B. First and second strand synthesis

Starting with total RNA, mRNA is polyA-selected and fragmented. It then undergoes first- and second-strand synthesis to produce products ready for library construction (Figure 4).

Size selection step

Adaptor ligation and standard library preparation

2 days for 8 samples
RNA-Seq Libraries... with More Details

Isolate RNA
Generate cDNA
Fragment, ligate adapters to create seq library

A. rRNA Depletion
Ribominus kit

B. Normalized library

C. Size selection
Reserved for miRNA, siRNA profiling

Image from: www.genxpro.info
RNA-Seq Libraries... with More Details

Isolate RNA

Generate cDNA

Fragment, ligate adapters to create seq library

Second Strand Synthesis-
Many Strand Specific Methods.

Strand-specific libraries for high throughput RNA sequencing prepared without poly(A) selection, Zhang et al.
RNA-Seq... at it’s Most Basic Form

Samples from two conditions

Isolate RNA

Generate short reads

Run sequencer

Create sequencing library by size selection and adding adaptors

Identify differentially expressed genes
What is an adaptor?

Adaptor:

- Allows the template DNA to attach to the flowcell/cluster
- Has primer sequences to start synthesis of
- Has barcodes/indexes for multiplexing
Types of Illumina Fragment Libraries

- **single-end**
  - independent reads

- **paired-end**
  - two inwardly oriented reads separated by ~200 nt

- **mate-paired**
  - two outwardly oriented reads separated by ~3000 nt
<table>
<thead>
<tr>
<th>Criteria</th>
<th>Annotation</th>
<th>Differential Gene Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological replicates</td>
<td>Not necessary but can be useful</td>
<td>Essential</td>
</tr>
<tr>
<td>Coverage across the transcript</td>
<td>Important for de Novo transcript assembly and identifying transcriptional isoforms</td>
<td>Not as important; however the only reads that can be used are those that are uniquely mappable.</td>
</tr>
<tr>
<td>Depth of sequencing</td>
<td>High enough to maximize coverage of rare transcripts and transcriptional isoforms</td>
<td>High enough to infer accurate statistics</td>
</tr>
<tr>
<td>Role of sequencing depth</td>
<td>Obtain reads that overlap along the length of the transcript</td>
<td>Get enough counts of each transcript such that statistical inferences can be made</td>
</tr>
<tr>
<td>DSN</td>
<td>Useful for removing abundant transcripts so that more reads come from rarer transcripts</td>
<td>Not recommended since it can skew counts</td>
</tr>
<tr>
<td>Stranded library prep</td>
<td>Important for de Novo transcript assembly and identifying true anti-sense transcripts</td>
<td>Not generally required especially if there is a reference genome Actually important!</td>
</tr>
<tr>
<td>Long reads (&gt;80 bp)</td>
<td>Important for de Novo transcript assembly and identifying transcriptional isoforms</td>
<td>Not generally required especially if there is a reference genome</td>
</tr>
<tr>
<td>Paired-end reads</td>
<td>Important for de Novo transcript assembly and identifying transcriptional isoforms</td>
<td>Not important Actually important!</td>
</tr>
</tbody>
</table>

From RNA-seqlopedia
3’ TAGSEQ- An Alternative to Whole RNA-Seq

Fig. 1 Overview of the protocol used to prepare 3’ cDNA tag libraries from total RNA. RNA was fragmented at the beginning to eliminate biases resulting from differences in transcript lengths. First-strand cDNA was primed with a modified oligo-dT containing primer to target 3’ ends. Each sample was prepared with a sample-specific oligonucleotide barcode, then quantified and pooled prior to sequencing.
WHY TAGSEQ?

- Cheaper to sequence 3’ end instead of the entire RNA.
- Amount of input RNA required is less.
- You can still identify differential expression.
WHY NOT TAGSEQ?

- If you want to look at differential splicing
- If you want to identify polymorphisms in gene sequences
Whole RNA-Seq vs TagSeq

Figure 2. Key criteria for evaluation of strand-specific RNA sequencing libraries.

Four categories of quality assessment:
1. **Complexity**: Evaluation of how complex the sample is.
2. **Strand Specificity**: Assessment of the strand orientation of the reads.
3. **Evenness of coverage**: Measurement of the evenness of the coverage across the genome.
4. **Comparison to known transcript structure**: Comparison of the sequencing data to the known transcript structure.

Legend:
- Green line: RNA-Seq
- Red line: Tag-Seq

**Note:** The graphs show the normalized coverage of reads through the gene. The top graph represents RNA-Seq, and the bottom graph represents Tag-Seq. The x-axis represents the percent through the gene, and the y-axis represents the normalized coverage.
Whole RNA-Seq vs TagSeq

TagSeq recovers known concentrations of mRNA (ERCC controls) with more accuracy than whole mRNASeq.

Fig. 1 Regression of observed vs. expected ERCC transcripts shows TagSeq has higher adjusted $R^2$ values for four different biological samples prepared with both methods (paired $t$-test, $t = 18.63$, d.f. = 3, $P < 0.001$).

Lohman et al, Molecular Ecology Resources, 2016
Comparing Stranded RNA-Seq Library Protocols

Figure 2. Key criteria for evaluation of strand-specific RNAseq libraries
Four categories of quality assessment. Double stranded genome (black parallel lines), with Gene ORF orientation (thick blue arrow) and UTRs (thin blue line), along with mapped reads (short black arrows – reads mapped to sense strand; red – reads mapped to antisense strand). (a) Complexity. (b) Strand Specificity. (c) Evenness of coverage. (d) Comparison to known transcript structure.

Comprehensive comparative analysis of strand-specific RNA sequencing methods, Levin et al, 2010
Why is RNA-Seq Difficult?

• Biases may mean what we are seeing is not reflective of true state of the transcriptome.
• Ugh, splicing!
• Gene level, exon level?
• Multimapping, partial mapping, not mapping.
• Normalization issues
  – some datasets are larger than others, some genes are larger than others

From Wikipedia- alternative splicing

Exon skipping
Mutually exclusive exons
Alternative 5' donor sites
Alternative 3' acceptor sites
Intron retention
Illumina Fastq file

FASTQ Format
@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
B@BBBBBB@BBBBAAAA>@AABA?BBBAAB??A?

Line 1: @read name
Line 2: called base sequence
Line 3: +read name (optional after +)
Line 4: base quality scores
Illumina Base Quality Scores

http://www.asciitable.com/

<table>
<thead>
<tr>
<th>Quality character</th>
<th>!&quot;#$%&amp;'()*+,-./0123456789:;&lt;=&gt;?@ABCDEFGHIJKLMNOPQRSTUVWXYZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII Value</td>
<td>33 43 53 63 73</td>
</tr>
<tr>
<td>Base Quality (Q)</td>
<td>0 10 20 30 40</td>
</tr>
</tbody>
</table>

Probability of Error = $10^{-\frac{Q}{10}}$

(This is a **Phred** score, also used for other types of qualities.)

<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>99.9%</td>
</tr>
</tbody>
</table>

Quality scores are ASCII encoded in fastq files. Different platforms/older sequencing data can have different encoding! Illumina HiSeq 2500 produces Sanger encoded data. Phred +33 = ASCII
How do we analyze RNA-Seq data?

- **STEP 1**: EVALUATE AND MANIPULATE RAW DATA
- **STEP 2**: MAP TO REFERENCE, ASSESS RESULTS
- **STEP 3**: ASSEMBLE TRANSCRIPTS
- **STEP 4**: QUANTIFY TRANSCRIPTS
- **STEP 5**: TEST FOR DIFFERENTIAL EXPRESSION
- **STEP 6**: VISUALIZE AND PERFORM OTHER DOWNSTREAM ANALYSIS
The Big Picture

MAP TO REFERENCE

BWA/Bowtie2/Kallisto

DAY 2

HISAT2/STAR

Looking for changes in annotated genes

ASSEMBLE TRANSCRIPTS

No novel transcripts

DESeq2/DEXSeq/edgeR/Cuffdiff

ID DEGs

Stringtied Stringtied Merge

BallGown

ASSEMBLE TRANSCRIPTS

DESeq2/DEXSeq/edgeR/Cuffdiff

ID DEGs

Stringtied Stringtied Merge

BallGown

ASSEMBLE TRANSCRIPTS

DESeq2/DEXSeq/edgeR/Cuffdiff

ID DEGs

Stringtied Stringtied Merge

BallGown

ASSEMBLE TRANSCRIPTS

DESeq2/DEXSeq/edgeR/Cuffdiff

ID DEGs

Stringtied Stringtied Merge

BallGown
<table>
<thead>
<tr>
<th>Class</th>
<th>Category</th>
<th>Package</th>
<th>Notes</th>
<th>Uses</th>
<th>Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read mapping</td>
<td>Unspliced aligners*</td>
<td>Short-read mapping package</td>
<td>Smith–Waterman extension</td>
<td>Aligning reads to a reference transcriptome</td>
<td>Reads and reference transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(SHRIMP)*</td>
<td>Probabilistic model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stampy**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bowtie**</td>
<td>Incorporates quality scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BWA**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spliced aligners</td>
<td>Exon-first methods</td>
<td>MapSplice</td>
<td>Aligning reads to a reference genome. Allows for the identification of novel splice junctions</td>
<td>Reads and reference genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SpliceMap**</td>
<td>Works with multiple unspliced aligners</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopHat*</td>
<td>Uses Bowtie alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSNAP**</td>
<td>Can use SNP databases</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>QPALMA*</td>
<td>Smith–Waterman for large gaps</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptome reconstruction</td>
<td>Genome-guided reconstruction</td>
<td>Genome-guided assembly</td>
<td>G.Mor.Se</td>
<td>Identifying novel transcripts using a known reference genome</td>
<td>Alignments to reference genome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Genome-independent assembly</td>
<td>Scripture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cufflinks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expression quantification</td>
<td>Expression quantification</td>
<td>Alexa-seq</td>
<td>Quantifying gene expression models</td>
<td>Reads and transcript models</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Enhanced read analysis of gene expression (ERANGE)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Normalization by expected uniquely mappability (NEUMA)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Isoform quantification</td>
<td>Cufflinks*</td>
<td>Quantifying transcript isoform expression levels</td>
<td>Read alignments to isoforms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MISO**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RNA-seq by expectation maximization (RSEM)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Differential expression</td>
<td>Cuffdiff*</td>
<td>Uses isoform levels in analysis</td>
<td>Identifying differentially expressed genes or transcript isoforms</td>
<td>Read alignments and transcript models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DegSeq**</td>
<td>Uses a normal distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EdgeR**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Differential Expression analysis of count data (DESeq)**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Myrna**</td>
<td>Cloud-based permutation method</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure 1. Methods for strand-specific RNA-Seq**

Salient details for seven protocols for strand-specific RNA-Seq, differential adaptor methods (a) and differential marking methods (b). mRNA is shown in grey, and cDNA in black. For differential adaptor methods, 5’ adaptors are shown in blue, and 3’ adaptors in red.
Third generation sequencing

• Next, next generation sequencing?

• Single molecule sequencing- takes care of all above mentioned issues

• Much longer reads (1-100kb)

• Many issues- high error rate and expensive

• Two categories:
  
  • Sequencing by synthesis (pacbio)
    
    • WATCH DNA as it is sequenced in realtime
    
    • ZMW technology lets smallest amount of light to be detected.
  
  • Direct sequencing
    
    • Oxford nanopore
    
    • Hydrogen ion changes ph in well. Change in ph indicates base has been incorporated.