Introduction to RNA-Seq

Monica Britton, Ph.D.
Sr. Bioinformatics Analyst

June 2015 Workshop
Overview of RNA-Seq Activities

• RNA-Seq Concepts, Terminology, and Work Flows
• Using Single-End Reads and a Reference Genome
• Gene Construction with Paired-End Reads
• Alignment to a Reference Transcriptome
• RNA-Seq Statistics: Blythe Durbin-Johnson

Now that you’re adept at running bioinformatics software, you’ll be doing the exercises “on your own”. Don’t worry if you can’t finish them all today.
RNA Transcription and Processing

A cell contains many types of RNA (rRNA, tRNA, mRNA, miRNA, IncRNA, snoRNA, etc.) – Only ~2% is mRNA
Gene Structure and Alternative Splicing

RNA-coding sequence

Transcription

Pre-mRNA

RNA processing: introns removed

mRNA

Protein-coding sequence

Translation

Polypeptide

α-tropomyosin gene

TRANSCRIPTION, SPLICING, AND 3' CLEAVAGE/POLYADENYLATION

striated muscle mRNA

smooth muscle mRNA

fibroblast mRNA

brain mRNA

Molecular Biology of the Cell, 4th ed.
Some mRNA-Seq Applications

- Differential gene expression analysis
- Transcriptional profiling

**Assumption:**
Changes in transcription/mRNA levels correlate with phenotype (protein expression)

- Identification of splice variants
- Novel gene identification
- Transcriptome assembly
- SNP finding
- RNA editing

http://bioinformatics.ucdavis.edu
Experimental Design

• What biological question am I trying to answer?
• What types of samples (tissue, timepoints, etc.)?
• How much sequence do I need?
• Length of read?
• Platform?
• Single-end or paired-end?
• Barcoding?
• Pooling?
• Biological replicates: how many?
• Technical replicates: how many?
• Protocol considerations?
What Is the Goal of the Experiment?

Many biological questions, such as...

“Characterize the differences between the wild-type and mutant” are broad and open-ended.

Such RNA-Seq experiments can be used to generate hypotheses, help form a more-focused question for the next experiment.

Make sure your experimental approach is suitable for the question you’re asking. (You will not find mutations in non-transcribed regions with RNA-Seq.)
Influence of the Organism

- **Novel** – little/no previous sequencing
- **Non-Model** – some sequence available (ESTs, Unigene set)
- **Genome-Sequenced** – draft genome
  - Thousands of scaffolds, maybe tens of chromosomes
  - Some annotation (\textit{ab initio}, EST-based, etc.)
- **Model** – genome fully sequenced and annotated
  - Multiple genomes available for comparison
  - Well-annotated transcriptome based on experimental evidence
  - Genetic maps with markers available
  - Basic research can be conducted to verify annotations (mutants available)
Amount of Sequence

**ENCODE Guidelines** (v.1 2011, never updated?)

- Depth determined by the goals of the experiment and type of sample
- 20-25M mappable PE reads (~30M raw PE reads) of length >30NT for polyA samples
- 100-200M PE76+ reads needed to reliably detect low copy transcripts/isoforms from a typical mammalian tissue

**Various studies** (often species and tissue/cell line specific):

- 10-20 million reads sufficient for differential gene expression
- But, additional unique transcripts still being found at 1 billion reads
Amount of Sequence

- Differential gene expression, reads/sample
  - Eukaryotes: 30+ million recommended
  - Bacteria: 10+ million recommended
- More sequence is needed to detect rare transcripts

Measures of Robustness of Expression Levels vs. Sequencing Depth (hg19)

Ramskold, et al., 2012
# Platform and Read Length Options

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Platform</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>40+ SE</td>
<td>Illumina (SOLiD)</td>
<td>Gene expression quantitation, SNP-finding</td>
</tr>
<tr>
<td>40+ PE</td>
<td>Illumina Ion Proton</td>
<td>Better specificity for the above Splice variant identification</td>
</tr>
<tr>
<td>100+ PE</td>
<td>Illumina Ion Proton</td>
<td>All the above and: Differentiation within gene families/paralogs Transcriptome assembly</td>
</tr>
<tr>
<td>200-300</td>
<td>Ion Torrent Sanger (454)</td>
<td>Splice variant identification Transcriptome assembly</td>
</tr>
<tr>
<td>400-600</td>
<td>PacBio (Oxford Nanopore)</td>
<td>Resolve haplotypes (phasing)</td>
</tr>
<tr>
<td>400-800</td>
<td></td>
<td>Not recommended for gene expression quantitation</td>
</tr>
<tr>
<td>8kb+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10kb+?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multiplexing

- Short (6-8 nt), unique barcodes (index) introduced as part of adapters
- Provide unique identifier for each sample
- Barcodes should be tolerant of 1-2 sequencing errors
- Barcodes allow deconvolution of samples
- Allows pooling samples to mitigate lane effects
- Allows sequencing capacity to be used efficiently
- Dual barcodes allow deep multiplexing (e.g., 96 samples)
Biological Replicates

• Allow measurement of variation between individuals/samples
• More replicates increase statistical power
• Some studies support increasing replicates over deeper sequencing (as long as you can detect genes of interest)
• Genetic Variation/Heterozygosity:
  – Is each individual a different genotype?
  – Are individuals highly inbred or clonal?
  – Haploid or diploid or polyploid?
• Pooling with barcodes – each sample is a replicate
• Pooling without barcodes – each pool is a replicate
  – Validation on individual samples
Technical Replicates

- Account for variation in preparation
- Cost can be prohibitive
- Better to do more biological replicates
- Barcoding/pooling samples across multiple lanes
  - Recommended to even out lane effects
  - Allow data processing even if one lane fails
Example

- This experimental design has biological replicates and is multiplexed to mitigate lane effects.
- Each sample will generate, on average, 30-40 million reads.

Control: 6 biological replicates

Treated: 6 biological replicates

Each sample is individually barcoded; all samples are pooled and run in two lanes of HiSeq2500

Illumina HiSeq Flow Cell Lanes
mRNA-Seq Protocol Overview

Adapted from Simon et al., 2009, Ann. Rev. Plant Biol. 60:305

Sample RNAs

- poly(A) selection (with magnetic beads)

TOTAL RNA

↓

mRNA

↓

FRAGMENTED mRNA/cDNA

↓

FINISHED LIBRARY

1. Fragmented mRNA
2. Add adapters
3. Convert to cDNA, PCR

Random prime to convert to cDNA

Sequencing of fragment ends by NGS

http://bioinformatics.ucdavis.edu
RNA Processing

- **PolyA Selection**
  - Oligo-dT, often using magnetic beads
  - Isolates mRNA very efficiently *unless total RNA is very dilute*
  - Can’t be used to sequence non-polyA RNA

- **rRNA Depletion**
  - RiboZero, RiboMinus
  - Non-polyA RNAs preserved (non-coding, bacterial RNA, etc.)
  - Can be less effective at removing all rRNA
Strand-Specific (Directional) RNA-Seq

- Now the default for Illumina TruSeq kits
- Preserves orientation of RNA after reverse transcription to cDNA
- Inform alignments to genome
  - Determine which genomic DNA strand is transcribed
  - Identify anti-sense transcription (e.g., IncRNAs)
  - Quantify expression levels more precisely
  - Demarcate coding sequences in microbes with overlapping genes
- Very useful in transcriptome assemblies
  - Allows precise construction of sense and anti-sense transcripts
The “insert” is the cDNA (or RNA) ligated between the adapters. Typical insert size is 160-200 bases, but can be larger. Insert size distribution depends on library prep method.
Paired-End Reads

Read 1 →

INSERT SIZE (TLEN)

INNER DISTANCE (+ or -)

Read 1 →

Read 2 ←

Read 2 ←
• cDNA inserts are a distribution of sizes
• There will be some read-through with adapter sequence at 3’ end
• Removal of adapter contamination can improve fraction of reads that align to the reference
• Very important for de novo assemblies
Alignment - Choosing a Reference

• Fully sequenced and annotated genome
  – Provides exon information to find splice variants

• Predicted/validated transcriptome
  – Simple to use
  – Comprehensive for all but the most novel genes

• NCBI Unigene Sets
  – Often incomplete
  – Good for medium to highly expressed genes

• No Genome? No Problem!
  – Transcriptome assembly
  – Useful for organisms with little or no sequence available
  – But, expect some redundancy and collapsing of gene families
Read Alignment and Counting

- Align reads to genome or transcriptome (output sam/bam)
- Convert alignments to read-counts per gene
  - May need to parse genomic intervals from gene models
  - Output is table of raw counts per gene for each sample
- Simple Normalization
  - RPKM (Reads Per Kilobase per Million reads mapped)
  - FPKM (Fragments Per Kilobase per Million reads mapped)
    - Fragment = cDNA insert
    - Ideally, there are two mappable PE reads per fragment
  - CPM (counts per million)
  - TMM (Trimmed Mean of M-Values) – used in edgeR, presumes most genes are not differentially expressed
- Statistical Analysis (Blythe’s talk)
  - Compare expression between samples, tissues, etc.
  - Use appropriate statistical model for your experiment.
Read Alignment and Counting

Alignment to Genome – one splice variant

Alignment to Transcriptome

[Diagram showing read alignment to genome and transcriptome with exon regions labeled (Exon A - 21), (Exon B - 12), (Exon C - 22).]
Read Alignment and Counting

Alignment to Genome – two splice variants

Alignment to Transcriptome (Gene Sequences)
Splicing-Aware Alignment

A splicing-aware aligner will recognize the difference between a short insert and a read that aligns across exon-intron boundaries.
Transcript Reference vs. Genome Reference

Some reads will align uniquely to an exon in the genome. But how can transcript abundance be determined?
Multiple Mapping within the Genome

Some reads will align to more than one location in the genome. Which gene/transcript should this read be assigned to?
Multiple-Mapping Reads ("Multireads")

• Some reads will align to more than one place in the reference, because:
  – Common domains, gene families
  – Paralogs, pseudogenes, etc.
  – Shared exons (if reference is transcriptome)
• This can distort counts, leading to misleading expression levels
• If a read can’t be uniquely mapped, how should it be counted?
• Should it be ignored (not counted at all)?
• Should it be randomly assigned to one location among all the locations to which it aligns equally well?
• This may depend on the question you’re asking...
• ...and also depends on the software you use.
Choosing an Aligner

- Transcriptome reference – BWA, Bowtie2
  - Splicing-aware not needed

- Genome reference
  - Aligner must be splicing-aware to account for reads that cross intron-exon boundaries
  - TopHat2 (Bowtie2) ([http://ccb.jhu.edu/software/tophat/index.shtml](http://ccb.jhu.edu/software/tophat/index.shtml))
  - GSNAP ([research-pub.gene.com/gmap/](http://research-pub.gene.com/gmap/))
  - STAR ([http://gingeraslab.cshl.edu/STAR/](http://gingeraslab.cshl.edu/STAR/)) – fast, but uses most memory

- Each aligner has multiple parameters that can be tweaked, affecting read mapping results

- Most software is updated regularly, to improve performance and accommodate new technologies

- GET ON THE MAILING LISTS/GOOGLE GROUPS!
How Well Did Your Reads Align to the Reference?

Calculate percentage of reads mapped per sample

<table>
<thead>
<tr>
<th>% Reads Mapped</th>
<th>Great!</th>
<th>Good</th>
<th>Incomplete Reference?</th>
<th>Sample Contamination?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Great!</td>
<td>90</td>
<td>80</td>
<td>60</td>
<td>40</td>
</tr>
<tr>
<td>Good</td>
<td>80</td>
<td>70</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>Incomplete Reference?</td>
<td>60</td>
<td>50</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Sample Contamination?</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>
RNA Quality Assessment

- rRNA contamination

  ![Diagram showing RNA composition]

  - Total RNA sample
  - After PolyA isolation
  - After rRNA depletion

- Align reads to rRNA sequences from organism or relatives
- Generally, don’t need to remove rRNA reads
Checking Your Results

• Key genes that may confirm sample ID
  – Knock-out or knock-down genes
  – Genes identified in previous research

• Specific genes of interest
  – Hypothesis testing
  – Important pathways

• Experimental validation (e.g., qRT-PCR)
  – Generally required for publication
  – The best way to determine if your analysis protocol accurately models your organism/experiment
  – Ideally, validation should be conducted on a different set of samples
Analysis Choices

This paper compares seven methods

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson¹ and Mauro Delorenzi¹²

This paper compares eleven methods

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies¹, Andrea Rau¹, Julie Aubert¹, Christelle Hennequet-Antier¹, Marine Jeanmougin¹, Nicolas Servant¹, Céline Keime¹, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guerme, Bernd Jagla, Luc Jouneau, Denis Laloe, Caroline Le Gall, Brigitte Schaeffer, Stéphane Le Crom¹, Mickaël Guedj¹, Florence Jaffrezic¹ and on behalf of The French StatOmique Consortium

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

http://bioinformatics.ucdavis.edu
Analysis Choices

Evaluated 6 differential gene expression analysis software packages (did not investigate differential isoform expression)

Increasing replicates is more important than increasing sequencing depth

Transcript length bias reduces the ability to find differential expression in shorter genes.

limma and baySeq most closely model “reality”.

limma and edgeR had the fewest number of false positives.

BUT, 5 of 6 packages were out-of-date by publication date; at least two changed substantially, so this analysis might be different today.
Where to find some guidance?

• The RNA Sequencing Quality Control (SEQC) Consortium and the Association of Bimolecular Resource Facilities (ABRF) conducted several systematic large-scale assessments

• RNA-Seq replicates run in ABRF study:
  – 15 lab sites
  – 4 protocols (polyA-select, ribo-depleted, size selected, degraded)
  – 5 platforms (HiSeq, Ion Torrent PGM & Proton, PacBio, 454)

• SEQC generated over 100 billion reads across three platforms

• More than 10Tb data available for analysis
Bioinformatics analyses are *in silico* experiments

The tools and parameters you choose will be influenced by factors including:

- Available reference/annotation
- Experimental design (e.g., pairwise vs. multi-factor)

The “right” tools are the ones that best inform on your experiment

Don’t just shop for methods that give you the answer you want
The GTF (Gene Transfer Format) File

<table>
<thead>
<tr>
<th>chr</th>
<th>source</th>
<th>start</th>
<th>end</th>
<th>strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4382902</td>
<td>4383401</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4383207</td>
<td>4383401</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown start_codon</td>
<td>4383207</td>
<td>4383209</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4385171</td>
<td>4385386</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4385171</td>
<td>4385386</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4387926</td>
<td>4388085</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4387926</td>
<td>4388085</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4398008</td>
<td>4398156</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4398008</td>
<td>4398156</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4409026</td>
<td>4409172</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4409026</td>
<td>4414522</td>
<td>+</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown stop_codon</td>
<td>4409173</td>
<td>4409175</td>
<td>+</td>
</tr>
</tbody>
</table>

The left columns list source, feature type, and genomic coordinates.

The right column includes attributes, including gene ID, etc.

http://bioinformatics.ucdavis.edu
Fields in the GTF File

Sequence Name (i.e., chromosome, scaffold, etc.)

Source (program that generated the gtf file or feature)

Feature (i.e., gene, exon, CDS, start codon, stop codon)

Start (starting location on sequence)

End (end position on sequence)

Score

Strand (+ or -)

Frame (0, 1, or 2: which is first base in codon, zero-based)

Attribute (“;”-delimited list of tags with additional info)

This attribute provides info to Tophat/Cufflinks

gene_id "PRMT8"; gene_name "PRMT8"; p_id "P10933"; transcript_id "NM_019854"; tss_id "TSS4368";
### An Unusual GTF File

<table>
<thead>
<tr>
<th>Supercont 1.64</th>
<th>VectorBase</th>
<th>Type</th>
<th>Gene ID</th>
<th>Transcript ID</th>
<th>Exon Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>gene</td>
<td>834185</td>
<td>AAEL005599</td>
<td></td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>mRNA</td>
<td>834185</td>
<td>AAEL005599</td>
<td></td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>exon</td>
<td>839892</td>
<td>AAEL005599</td>
<td>1 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>exon</td>
<td>839892</td>
<td>AAEL005599</td>
<td>2 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>exon</td>
<td>839190</td>
<td>AAEL005599</td>
<td>3 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>exon</td>
<td>834185</td>
<td>AAEL005599</td>
<td>4 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>five_prime_utr</td>
<td>840615</td>
<td>AAEL005599</td>
<td></td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>start_codon</td>
<td>840612</td>
<td>AAEL005599</td>
<td></td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>CDS</td>
<td>839892</td>
<td>AAEL005599</td>
<td>1 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>CDS</td>
<td>839892</td>
<td>AAEL005599</td>
<td>2 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>CDS</td>
<td>839190</td>
<td>AAEL005599</td>
<td>3 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>stop_codon</td>
<td>834185</td>
<td>AAEL005599</td>
<td></td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>exon</td>
<td>1334853</td>
<td>AAEL005599</td>
<td>5 of 4</td>
</tr>
<tr>
<td>supercont1.64</td>
<td>VectorBase</td>
<td>exon</td>
<td>1497675</td>
<td>AAEL005599</td>
<td>6 of 4</td>
</tr>
</tbody>
</table>
Differential Gene Expression Generalized Workflow

Software
- scythe
- sickle
- tophat2/bowtie2
- (cufflinks2)
- cuffdiff
- Or
- R packages (edgeR)
Today’s Exercises – Differential Gene Expression

Today, we’ll use a few different types of data typically encountered:

1. Single-end reads (“tag counting” with well-annotated genomes)
2. Paired-end reads (finding novel transcripts in a genome with incomplete annotation)
3. Paired-end reads for gene expression when only a transcriptome is available (such as after a transcriptome assembly)

And we’ll be using a few different software:

1. Tophat to align spliced reads to a genome
2. Cuffdiff for differential expression of transcripts/genes from tophat alignments
3. cummeRbund to generate diagnostic plots from cuffdiff output
4. htseq-count to generate raw counts tables for...
5. edgeR, which can also handle more complex experimental designs
6. Bowtie2 to align reads to a transcriptome reference
7. sam2counts to extract raw counts from bowtie2 alignments
Today’s Exercises – Differential Gene Expression

Later, we’ll talk about other RNA-Seq topics, including
• The ever-increasing software packages in the Tuxedo Suite
• Finding novel transcripts (“Gene Construction”)
• And, what happens after the analysis? (Is there an “after”?)

But for now, let’s get some analysis going!