Introduction to RNA-Seq

Monica Britton, Ph.D.
Bioinformatics Analyst

June 2014 Workshop
Overview of Today’s Activities

Morning
• RNA-Seq Concepts, Terminology, and Work Flows
• Simple Differential Expression with a Model Organism (Single and Paired-End)
• Guest Speaker: Dr. Walter Leal

Afternoon
• Gene Construction with Paired-End Reads
• RNA-Seq Statistics (Blythe Durbin-Johnson)
• Alignment to a Reference Transcriptome

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

A cell contains many types of RNA (rRNA, tRNA, mRNA, miRNA, lncRNA, snoRNA, etc.) – Only ~2% is mRNA

Koning, Plant Physiology Information Website

http://bioinformatics.ucdavis.edu
Gene Structure and Alternative Splicing

**Diagram:**

1. DNA → Transcription → Pre-mRNA → RNA processing (introns removed) → mRNA → Translation → Polypeptide

2. α-tropomyosin gene → Transcription, splicing, and 3' cleavage/polyadenylation

   - striated muscle mRNA
   - smooth muscle mRNA
   - fibroblast mRNA
   - brain mRNA

**Text:**

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 (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

- 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

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:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Differentiation within gene families/paralogs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transcriptome assembly</td>
</tr>
<tr>
<td>200-300</td>
<td>Ion Torrent Sanger (454)</td>
<td>Splice variant identification</td>
</tr>
<tr>
<td>400-600</td>
<td></td>
<td>Transcriptome assembly</td>
</tr>
<tr>
<td>400-800</td>
<td>PacBio (Oxford Nanopore)</td>
<td>Resolve haplotypes (phasing)</td>
</tr>
<tr>
<td>5000 avg</td>
<td></td>
<td>Not recommended for gene expression quantitation</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 are better (up to a point)
• 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, 50-60 million reads.

Control: 3 biological replicates

Treated: 3 biological replicates

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

Illumina HiSeq Flow Cell Lanes
mRNA-Seq Protocol Overview

TOTAL RNA

mRNA

FRAGMENTED mRNA/cDNA

FINISHED LIBRARY

Sample RNAs

poly(A) selection (with magnetic beads)

mRNA

m^7G cap

Fragmented mRNA

Add adapters

Convert to cDNA, PCR

Sequencing of fragment ends by NGS

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

- 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

INSERT SIZE (TLEN)

INNER DISTANCE (+ or -)
**Adapter Contamination**

- 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 reads per fragment

• 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

- Exon A - 21
- Exon B - 12
- Exon C - 22

Alignment to Transcriptome

(66)
Read Alignment and Counting

Alignment to Genome – two splice variants

Alignment to Transcriptome (Gene Sequences)

(Exon A - 21)  (Exon B - 6)  (Exon C - 22)
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. 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:
  – Shared exons (if reference is transcriptome)
  – Common domains, gene families
  – Paralogs, pseudogenes, etc.

• 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
- Genome reference
  - Aligner must be splicing-aware to account for reads that cross intron-exon boundaries
  - TopHat (Bowtie)/TopHat2 (Bowtie2) (tophat.cbcb.umd.edu)
  - GSNAP (research-pub.gene.com/gmap/)
  - STAR (http://gingeraslab.cshl.edu/STAR/) – newest, fastest, 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!
How well did your data align to the reference?

Calculate percentage of reads mapped per sample

- Great!
- Good
- Incomplete Reference?
- Sample Contamination?
RNA Quality Assessment

- rRNA contamination

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

Charlotte Soneson* and Mauro Delorenzi

This paper compares eleven methods
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 (or next year)
Differential Gene Expression Generalized Workflow

- 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>
<th>Gene ID</th>
<th>Gene Name</th>
<th>Transcript ID</th>
<th>TSS ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4382902</td>
<td>4383401</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4383207</td>
<td>4383401</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown start_codon</td>
<td>4383207</td>
<td>4383209</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4385171</td>
<td>4385386</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4385171</td>
<td>4385386</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4387926</td>
<td>4388085</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4387926</td>
<td>4388085</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4398008</td>
<td>4398156</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4398008</td>
<td>4398156</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown stop_codon</td>
<td>4409026</td>
<td>4409172</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown exon</td>
<td>4409026</td>
<td>4414522</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
<tr>
<td>chr12</td>
<td>unknown CDS</td>
<td>4409173</td>
<td>4409175</td>
<td>+</td>
<td>gene_id</td>
<td>gene_name</td>
<td>transcript_id</td>
<td>tss_id</td>
</tr>
</tbody>
</table>

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

The right column includes attributes, including gene ID, etc.
Fields in the GTF File

Sequence Name (i.e., chromosome, scaffold, etc.)
chr12

Source (program that generated the gtf file or feature)
unknown

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

Start (starting location on sequence)
3677872

End (end position on sequence)
3678014

Score
.

Strand (+ or -)
+

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

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>gene_id</th>
<th>transcript_id</th>
<th>exon_number</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;AAEL005599&quot;</td>
<td>&quot;AAEL005599-RA&quot;</td>
<td>&quot;1 of 4&quot;</td>
</tr>
<tr>
<td>&quot;AAEL005599&quot;</td>
<td>&quot;AAEL005599-RA&quot;</td>
<td>&quot;2 of 4&quot;</td>
</tr>
<tr>
<td>&quot;AAEL005599&quot;</td>
<td>&quot;AAEL005599-RA&quot;</td>
<td>&quot;3 of 4&quot;</td>
</tr>
<tr>
<td>&quot;AAEL005599&quot;</td>
<td>&quot;AAEL005599-RA&quot;</td>
<td>&quot;4 of 4&quot;</td>
</tr>
<tr>
<td>&quot;AAEL005599&quot;</td>
<td>&quot;AAEL005599-RA&quot;</td>
<td>&quot;5 of 4&quot;</td>
</tr>
<tr>
<td>&quot;AAEL005599&quot;</td>
<td>&quot;AAEL005599-RA&quot;</td>
<td>&quot;6 of 4&quot;</td>
</tr>
</tbody>
</table>

**Note:**
- Gene IDs: "AAEL005599" and "AAEL016380" are repeated for some entries.
- Transcript IDs: "AAEL005599-RA" are used multiple times with different exon numbers.

### Contact Information

http://bioinformatics.ucdavis.edu

**Bioinformatics Core**

Part of the UC Davis Genome Center
Differential Gene Expression Generalized Workflow

Software

- scythe
- sickle
- tophat2/bowtie2
- (cufflinks2)
- cuffdiff
- Or
- R packages (edgeR)

Diagram:
- RAW READS -> ADAPTER CONTAMINATION REMOVAL -> READ QUALITY TRIMMING -> PROCESSED READS
- ALIGNMENT TO REFERENCE(S) -> ALIGNMENT RESULTS (SAM/BAM) -> rRNA CONTAM., INSERT SIZE
- FRAGMENT COUNTING -> RAW COUNTS
- STATISTICAL ANALYSIS -> DIFFERENTIAL EXPRESSION TABLES
- DIAGNOSTIC PLOTS

http://bioinformatics.ucdavis.edu
Today, we’ll analyze the same human RNA-Seq data in a few ways:

1. “Aligning SE RNA-Seq Reads to a Reference Genome” is typical of a model-organism analysis with TopHat/Cuffdiff, looking for differential expression in known genes.
2. “Aligning PE RNA-Seq Reads to a Genome” is a typical TopHat/Cuffdiff analysis with paired-end reads.
3. “Differential Expression Analysis with edgeR from Genome Alignments” uses the TopHat alignments to generate input for the edgeR statistical package.
4. “Gene Construction” uses Cufflinks to find novel transcripts in the PE data.
5. “Alignment to a Reference Transcriptome” explores differential expression without using a genome as reference.

Let’s get started!