RNA-Seq Primer

Understanding the RNA-Seq evidence tracks on the GEP UCSC Genome Browser

Wilson Leung 08/2016

Introduction to RNA-Seq

- RNA-Seq: Massively parallel **RNA Seq**uencing using second or third generation sequencing technologies
  - Illumina, Ion Torrent, PacBio
- Goal: Identify regions in the genome that are being transcribed in a sample
  - Different tissues, developmental stages, treatments
- Provide more comprehensive and more accurate measurements of gene expression than microarrays
  - RNA-Seq read count corresponds to the expression level

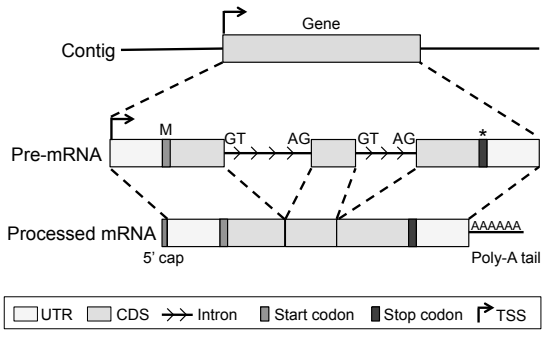
Common applications

- Gene annotation
  - Identify transcribed regions (gene and exon structure)
  - Alternative splice junctions
  - RNA editing
- Differential expression analysis
  - Treatment versus control samples
    - Tumor versus normal cells
- Identify changes in gene structure
  - Gene fusions (cancer genomes)
    - Maher CA *et al*. Transcriptome sequencing to detect gene fusions in cancer. Nature. (2009) Mar 5;458(7234):97-101

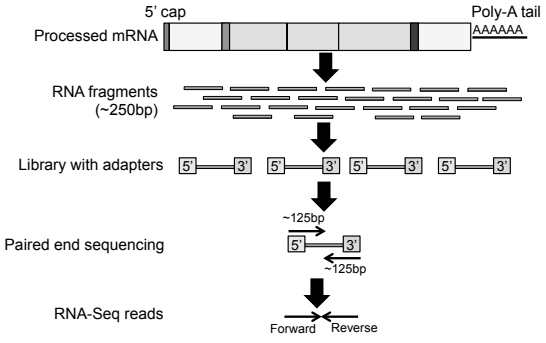
RNA-Seq evidence tracks on the GEP UCSC Genome Browser

- RNA-Seq read alignments
  - *D. biarmipes* RNA-Seq
- Number and quality of mapped reads
  - Read Coverage, Alignment Summary
- Splice junction predictions
  - RNA-Seq TopHat, Spliced RNA-Seq
- Transcripts assembled from RNA-Seq reads
  - Cufflinks, Oases

Pre-mRNA processing



Generating RNA-Seq data (Illumina)



Wang *et al*. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics (10) 57-63

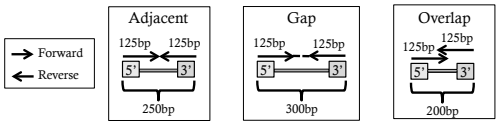
RNA-Seq analysis pipeline
(Reference-guided)

- Map RNA-Seq reads against the reference assembly
  - Bowtie2, BWA, Maq, …
- Use an aligner that recognizes splice sites to try to map the initially unmapped reads (IUM reads)
  - HISAT2, TopHat, TrueSight, MapSplice, …
- Construct transcripts from read coverage and the splice junction predictions
  - StringTie, Cufflinks, Scripture, CEM, …

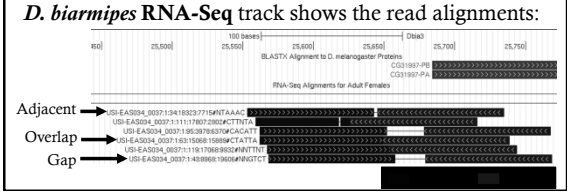
Roberts A, *et al.* Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011 Sep 1;27(17):2325-9

Mapping unspliced RNA-Seq reads

Read placement based on RNA-Seq fragment sizes:

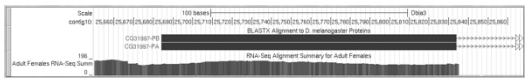


***D. biarmipes* RNA-Seq** track shows the read alignments:

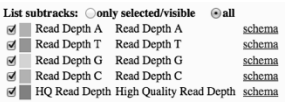


RNA-Seq Alignment Summary track

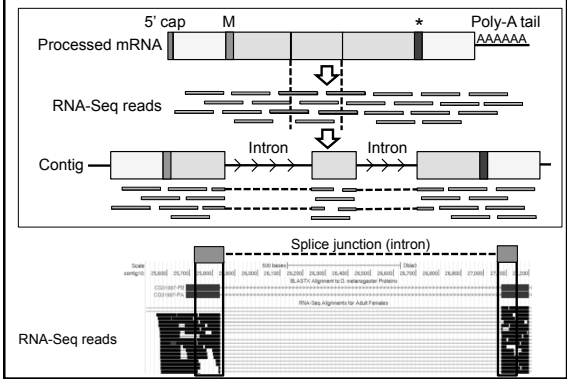
- Shows the number of reads mapped to each position of the genome:



  - Y-axis shows the read depth
  - Color corresponds to the different nucleotides or the mapping quality:



Mapping spliced RNA-Seq reads

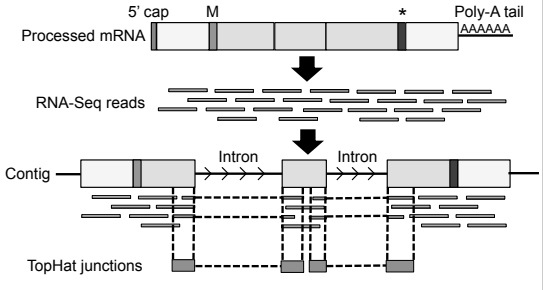


TopHat Splice junction predictions

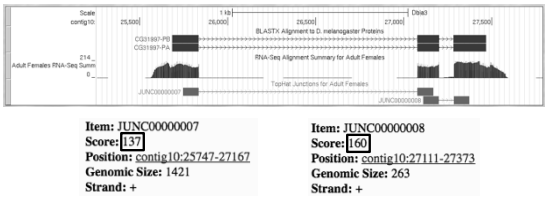
- Spliced RNA-Seq reads have a distinct signature when mapped against the genome
  - Use reads mapped by Bowtie2 to define the region to search for potential splice sites
- Analyze mapped reads in the context of known biological properties of splice sites:
  - Canonical splice donor (GT/GC) and acceptor sites (AG)
  - Minimum intron size

Trapnell C, *et al.* TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11

TopHat splice junction predictions



RNA-Seq TopHat track



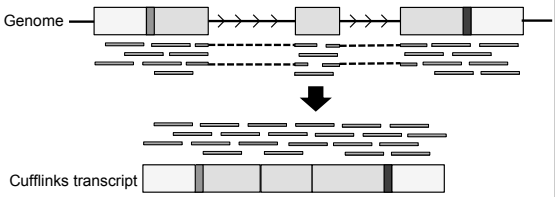
Item: JUNC00000007
Score: 137
Position: contig10:25747-27167
Genomic Size: 1421
Strand: +

Item: JUNC00000008
Score: 160
Position: contig10:27111-27373
Genomic Size: 263
Strand: +

- The **score** of a TopHat prediction corresponds to the number of reads that support the splice junction
- The **width** of the boxes are defined by the extents of the RNA-Seq reads that support the splice junction

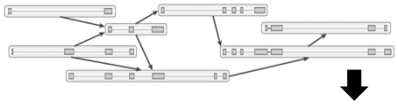
Reference-guided transcriptome assembly (e.g. Cufflinks)

- Predict transcript models and relative abundance based on aligned RNA-Seq reads
  - Create the most parsimonious set of transcripts that explains most of the regions with RNA-Seq coverage

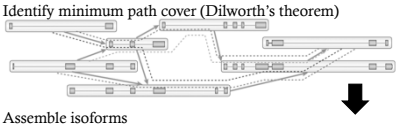


Cufflinks – reference-based transcriptome assembly

1. Build graph of incompatible RNA-Seq fragments



2. Identify minimum path cover (Dilworth's theorem)



3. Assemble isoforms

- Use TransDecoder to identify coding regions within assembled transcripts

Martin JA, Wang Z. *Next-generation transcriptome assembly*. Nat Rev Genet. (2011) Sep 7;12(10):671-82.

RNA-Seq analysis pipeline
(*De novo* transcriptome assembly)

- Create transcriptome assembly based on overlapping RNA-Seq reads
  - Oases, SOAPdenovo-trans, Trinity, …
- Compare assembled transcripts against a database of known proteins or conserved domains (e.g. Pfam)
  - TransDecoder, blastx, hmmer, …
- Map assembled transcripts against a reference genome
  - BLAT, Exonerate, PASA, …

Zhao QY, *et al.* Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics. 2011 Dec 14;12

Limitations of RNA-Seq

- Lack of RNA-Seq read coverage is **a negative result**
  - Transcript might be expressed at low levels or might not be expressed at the developmental stage sampled by RNA-Seq
  - Sequencing and sampling bias (e.g. poly-A selection)
  - Read mapping biases (e.g. simple repeats)
- Difficult to identify splice junctions located within a larger exon
- GEP exercise that illustrates some of the challenges in interpreting RNA-Seq data:
  - **Browser-Based Annotation and RNA-Seq Data**

Use of RNA-Seq data in GEP annotation projects

- Confirm the proposed gene model
- Identify small or weakly conserved exons
- Confirm non-canonical splice sites
  - GC-AG and AT-AC introns

Additional information

- Comprehensive overview on RNA-Seq
  - Garber M, *et al*. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011 Jun;8(6):469-77.
- *Drosophila* transcriptome
  - Daines B, *et al.* The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. Genome Res. 2011 Feb;21(2):315-24.
- *De novo* transcriptome assembly
  - Li B, *et al.* Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014 Dec 21;15(12):553.
- Differential expression analysis
  - Trapnell C, *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562-78

Questions

http://www.flickr.com/photos/horiavarlan/4273168957/sizes/l/in/photostream/