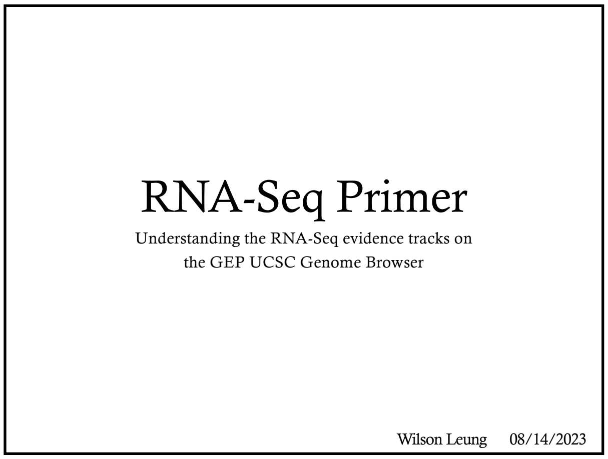
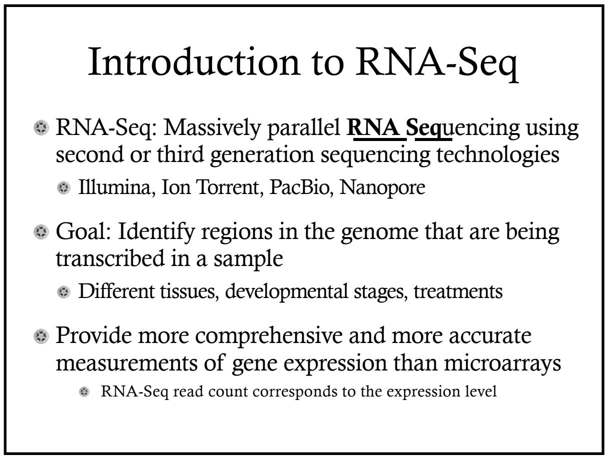


RNA-Seq Primer

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

Wilson Leung 08/14/2023

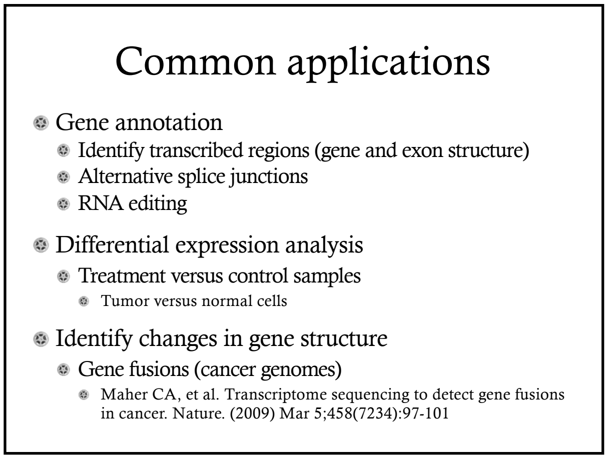
1



Introduction to RNA-Seq

- RNA-Seq: Massively parallel **RNA Seq**uencing using second or third generation sequencing technologies
  - Illumina, Ion Torrent, PacBio, Nanopore
- 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

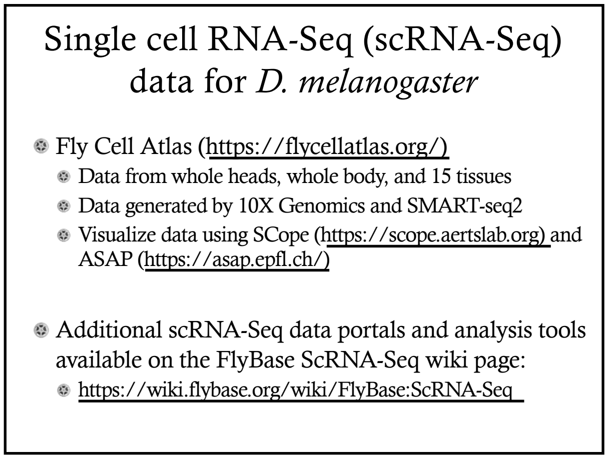
2



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

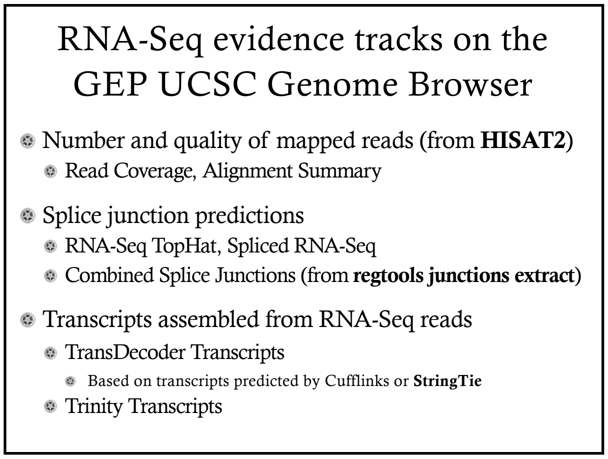
3



Single cell RNA-Seq (scRNA-Seq) data for *D. melanogaster*

- Fly Cell Atlas (https://flycellatlas.org/)
  - Data from whole heads, whole body, and 15 tissues
  - Data generated by 10X Genomics and SMART-seq2
  - Visualize data using SCope (https://scope.aertslab.org) and ASAP (https://asap.epfl.ch/)
- Additional scRNA-Seq data portals and analysis tools available on the FlyBase ScRNA-Seq wiki page:
  - https://wiki.flybase.org/wiki/FlyBase:ScRNA-Seq

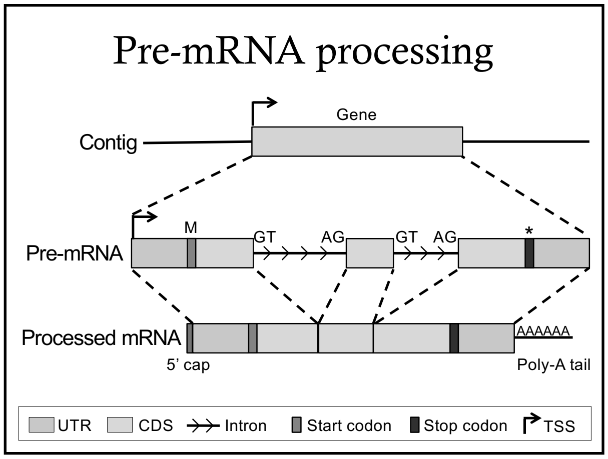
4



RNA-Seq evidence tracks on the GEP UCSC Genome Browser

- Number and quality of mapped reads (from **HISAT2**)
  - Read Coverage, Alignment Summary
- Splice junction predictions
  - RNA-Seq TopHat, Spliced RNA-Seq
  - Combined Splice Junctions (from **regtools junctions extract**)
- Transcripts assembled from RNA-Seq reads
  - TransDecoder Transcripts
    - Based on transcripts predicted by Cufflinks or **StringTie**
  - Trinity Transcripts

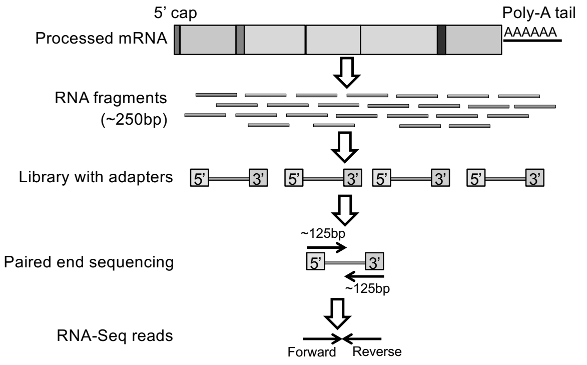
5



Pre-mRNA processing

6

Generating RNA-Seq data (Illumina)



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

7

RNA-Seq analysis pipeline

(Reference-guided)

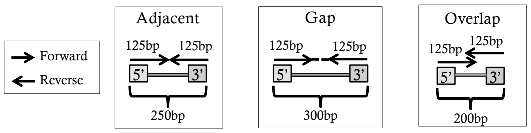
- Map RNA-Seq reads against the reference assembly
  - Bowtie2, BWA, Maq, ...
- Use an aligner that recognizes splice sites to try to map the initially unmapped reads (IUM reads)
  - HISAT2, TopHat, TrueSight, MapSplice, ...
- Construct transcripts from read coverage and the splice junction predictions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

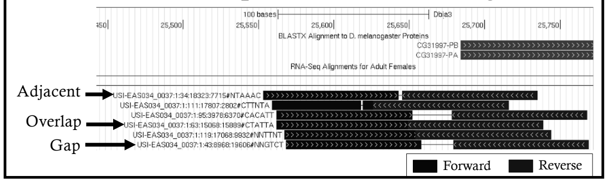
8

Mapping unspliced RNA-Seq reads

Read placement based on RNA-Seq fragment sizes:



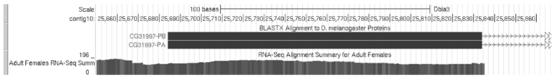
D. biarmipes **RNA-Seq** track shows the read alignments:


100 bases
Dbia3
450
25,500
25,550
25,600
25,650
25,700
25,750
BLASTX Alignment to D. melanogaster Proteins
CG31997-PB
CG31997-PA
RNA-Seq Alignments for Adult Females
Adjacent
USI-EAS034_0037:1:34:18323:7715#NTAAAC
USI-EAS034_0037:1:111:17807:2802#CTTNTA
USI-EAS034_0037:1:95:3978:6370#CACATT
Overlap
USI-EAS034_0037:1:63:15068:15889#CTATTA
USI-EAS034_0037:1:119:17068:9932#NNTTNT
Gap
USI-EAS034_0037:1:43:8968:19606#NNGTCT
Forward
Reverse

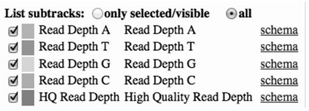
9

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:

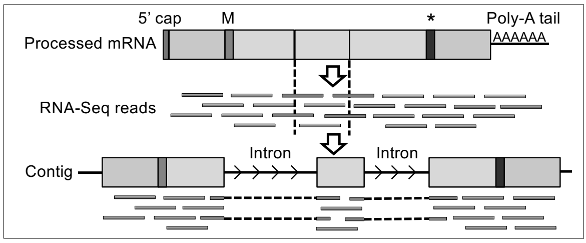


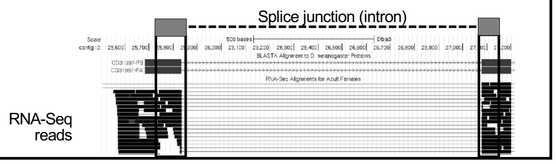
List subtracks: ○ only selected/visible ● all

☑	▨	Read Depth A	Read Depth A	schema
☑	▨	Read Depth T	Read Depth T	schema
☑	▨	Read Depth G	Read Depth G	schema
☑	▨	Read Depth C	Read Depth C	schema
☑	▨	HQ Read Depth	High Quality Read Depth	schema

10

Mapping spliced RNA-Seq reads




Splice junction (intron)
Scale
500 bases
Dbia3
BLASTX Alignment to D. melanogaster Proteins
CG31997-PB
CG31997-PA
RNA-Seq Alignments for Adult Females
RNA-Seq
reads

11

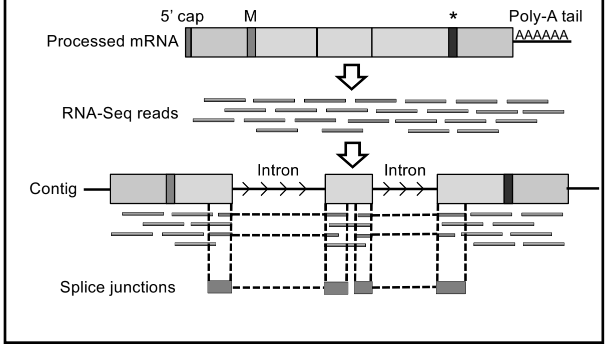
TopHat Splice junction predictions

- 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

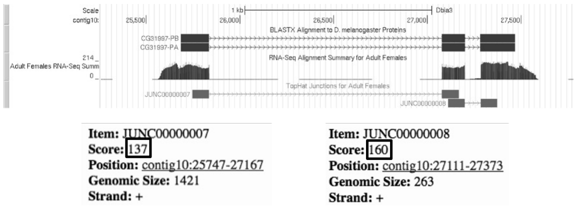
12

TopHat splice junction predictions



13

RNA-Seq TopHat track


Scale
contig10:
1 kb
Dbia3
25,500
26,000
26,500
27,000
27,500
BLASTX Alignment to D. melanogaster Proteins
CG31997-PB
CG31997-PA
214
Adult Females RNA-Seq Summ
0
RNA-Seq Alignment Summary for Adult Females
TopHat Junctions for Adult Females
JUNC00000007
JUNC00000008

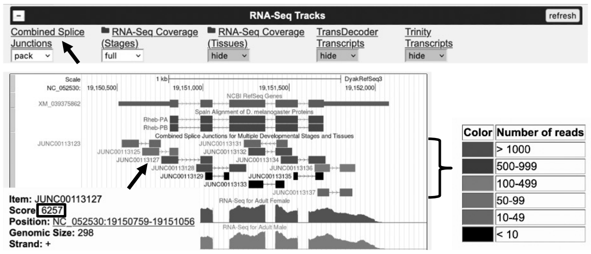
Item: JUNC00000007
Score: 137
Position: contig10:25747-27167
Genomic Size: 1421
Strand: +

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

14

Combined Splice Junctions track (regtools junctions extract)



Item: JUNC00113127
Score 6257
Position: NC_052530:19150759-19151056
Genomic Size: 298
Strand: +

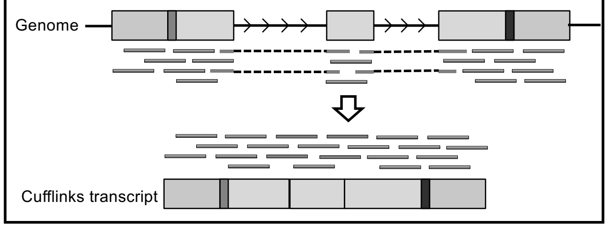
Color	Number of reads
■	> 1000
■	500-999
■	100-499
■	50-99
■	10-49
■	< 10

- Color indicates the number of RNA-Seq reads from multiple stages and tissues that support the splice junction

15

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

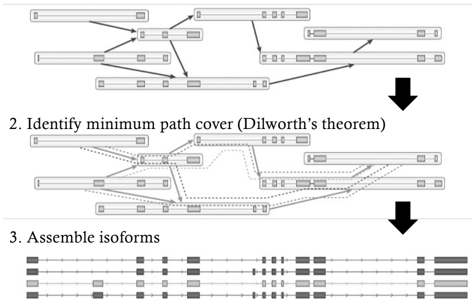
- Predict transcript models and relative abundance based on aligned RNA-Seq reads
  - Create the most parsimonious set of transcripts that explains most of the regions with RNA-Seq coverage



16

Cufflinks — reference-based transcriptome assembly

1. Build graph of incompatible RNA-Seq fragments
2. Identify minimum path cover (Dilworth's theorem)
3. Assemble isoforms

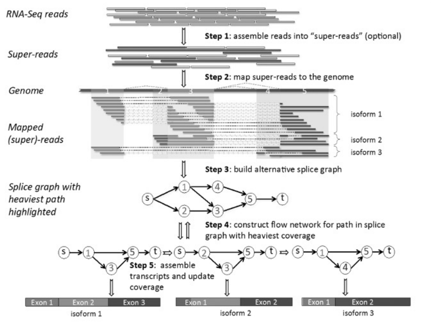


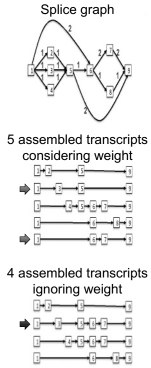
- Use TransDecoder to identify coding regions within assembled transcripts

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

17

StringTie — use flow networks for reference-based transcriptome assembly


RNA-Seq reads
Step 1: assemble reads into "super-reads" (optional)
Super-reads
Step 2: map super-reads to the genome
Genome
1
2
3
4
5
Mapped (super)-reads
isoform 1
isoform 2
isoform 3
Step 3: build alternative splice graph
Splice graph with heaviest path highlighted
s
1
2
3
4
5
t
Step 4: construct flow network for path in splice graph with heaviest coverage
Step 5: assemble transcripts and update coverage
Exon 1
Exon 2
Exon 3
isoform 1
Exon 1
Exon 2
isoform 2
Exon 1
Exon 2
isoform 3


Splice graph
5 assembled transcripts considering weight
4 assembled transcripts ignoring weight
1
2
3
4
5
6
7
8
9

Pertea M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar;33(3):290-5.

18

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

19

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**

20

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

21

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

22

Questions

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

23