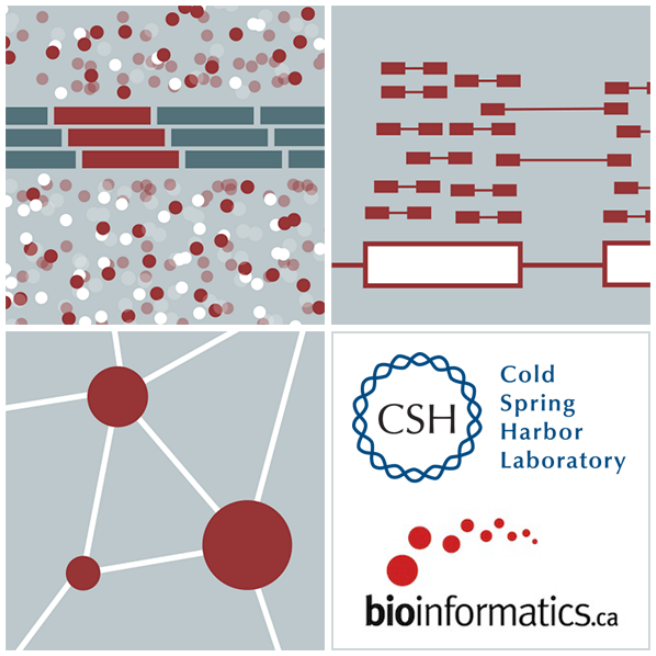
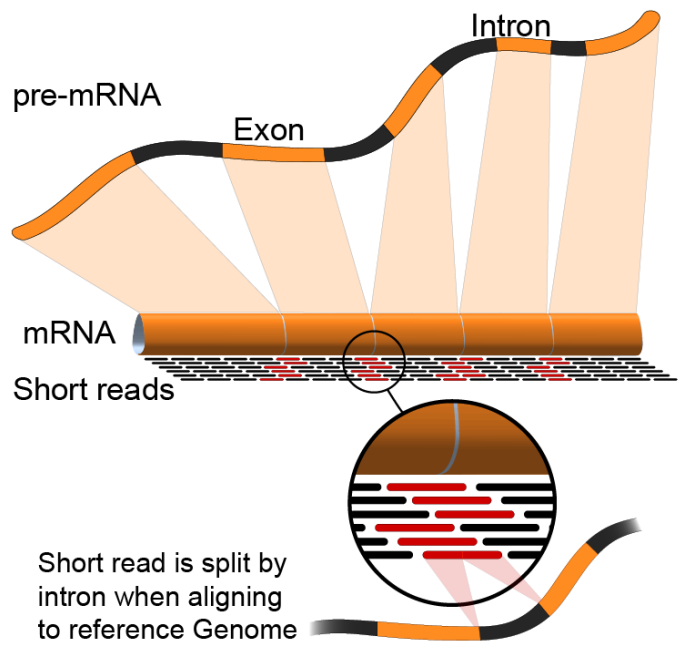
Cold Spring Harbor Laboratory

RNA-Seq Module 3
Abundance Estimation and Differential Expression

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal

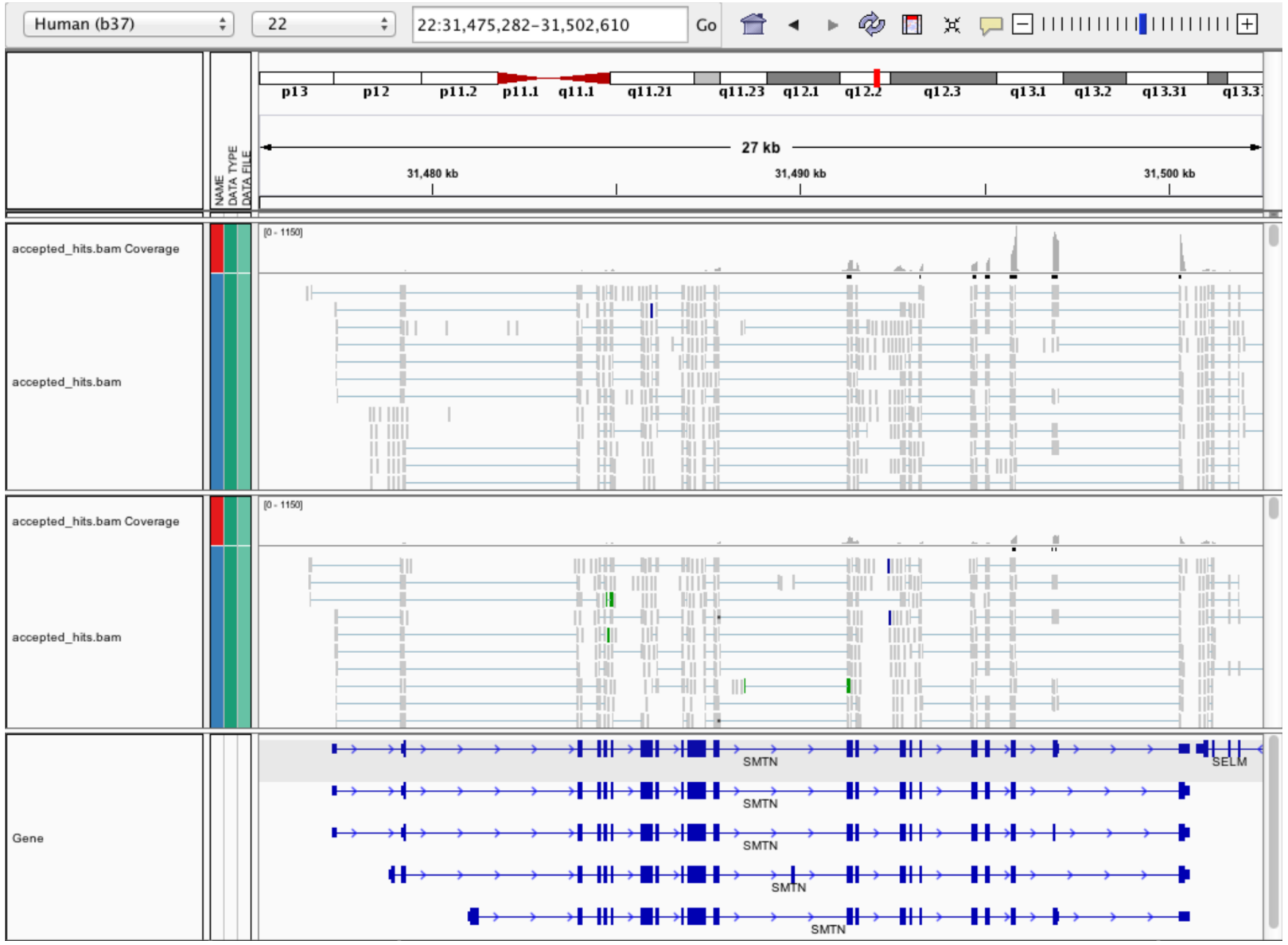
Advanced Sequencing Technologies & Bioinformatics Analysis November 10-23, 2024





Washington University in St. Louis
School of Medicine

Expression estimation for known genes and transcripts



3' bias

Down-regulated

What is FPKM (RPKM)?

- RPKM: **Reads** Per Kilobase of transcript per Million mapped reads.
- FPKM: **Fragments** Per Kilobase of transcript per Million mapped reads.

- No essential difference - Just a terminology change to better describe paired-end reads!

What is FPKM?

- Why not just count reads in my RNAseq data? → **Fragments**

- The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:

  - # fragments is biased towards larger genes → **Per Kilobase of transcript**

  - # fragments is related to total library depth → **per Million mapped reads.**

What is FPKM?

- FPKM attempts to normalize for gene size and library depth
    - remember – RPKM is essentially the same!

- C = number of mappable fragments for a gene (transcript)
- N = total number of mappable fragments in the library
- L = number of base pairs in the gene (transcript)
    - FPKM = (C / (N x L)) x 1,000 x 1,000,000
    - FPKM = (1,000,000,000 x C) / (N x L)
    - FPKM = (C / (N / 1,000,000)) / (L/1000)

- More reading:
    - http://www.biostars.org/p/11378/
    - http://www.biostars.org/p/68126/

How do FPKM and TPM differ?

- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:

FPKM

1) Determine total fragment count, divide by 1,000,000 (per Million)

2) Divide each gene/transcript fragment count by #1 (Fragments Per Million)

3) Divide each FPM by length of each gene/transcript in kilobases (FPKM)

TPM

1) Divide each gene/transcript fragment count by length of the transcript in kilobases (Fragments Per Kilobase)

2) Sum all FPK values for the sample and divide by 1,000,000 (per Million)

3) Divide #1 by #2 (TPM)

- The sum of all TPMs in each sample is the same. Easier to compare across samples!

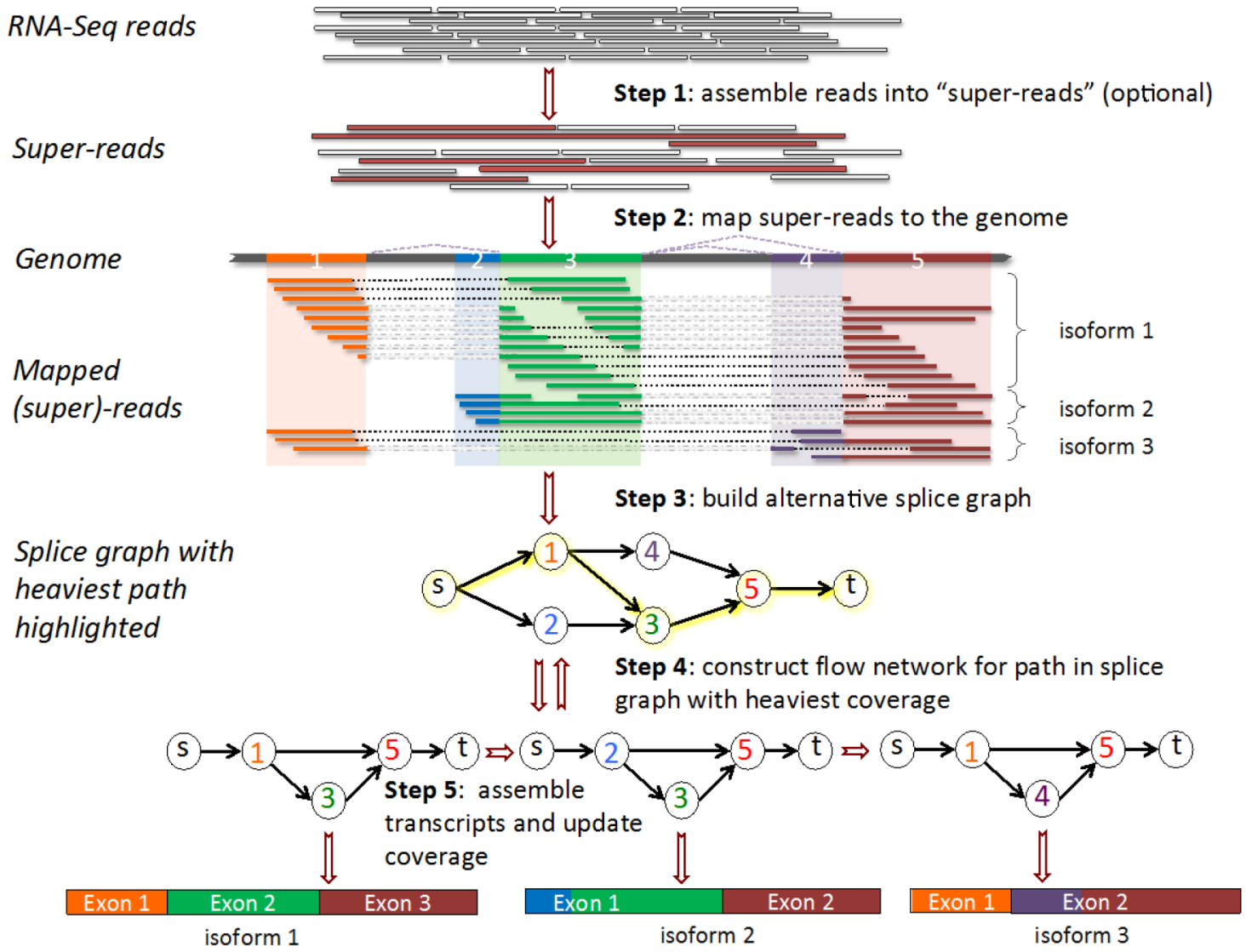
- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- https://www.ncbi.nlm.nih.gov/pubmed/22872506

How does StringTie work?

- Align reads to the genome, optionally assemble super-reads and re-align
- Group reads into clusters

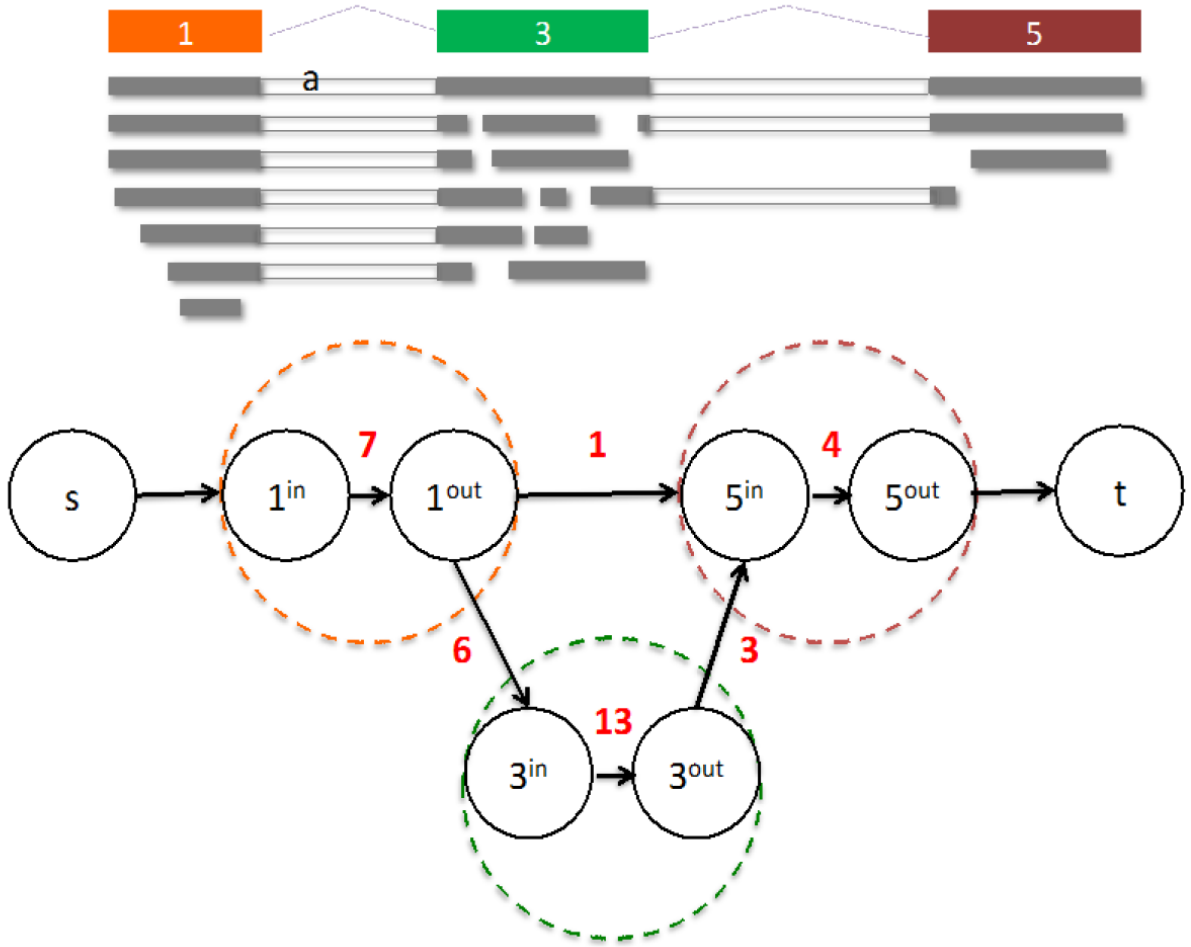
Infer isoforms:

- Build alternative splice graph (ASG)
- Iteratively extract the heaviest path from a splice graph
- construct a flow network
- compute maximum flow to estimate abundance
- update the splice graph by removing reads that were assigned by the flow algorithm
- This process repeats until all reads have been assigned.



Pertea et al. Nature Biotechnology, 2015

From flow network for each transcript, maximum flow is used to assemble transcript and estimate abundance



StringTie uses basic graph theory (splice graph), custom heuristics (heaviest path), more graph theory (flow network) and optimization theory (maximum flow). See StringTie paper for definitions and math.

StringTie Modes

- Expression estimation mode (“Reference Only”)
  - “–G $GTF_File” AND “–e” option
  - no "novel" transcript assemblies (isoforms)
  - read alignments not overlapping reference transcripts ignored
  - Faster, especially when given limited set of reference transcripts
    - Avoids complicated steps of clustering and building alternative splice graph from scratch, skipping straight to abundance estimation
- “Reference guided mode”
  - “–G $GTF_File” WITHOUT “–e” option
  - Both known and novel transcript assemblies
- “De novo” mode
  - NEITHER “–G $GTF_File” NOR “–e” option
  - Novel transcript assemblies only

Pertea et al. Nature Protocols, 2016

StringTie -merge

- Merge together all gene structures from all samples
  - Some samples may only partially represent a gene structure
- Incorporates known transcripts with assembled, potentially novel transcripts
- For de novo or reference guided mode, we will rerun StringTie with the merged transcript assembly.

Pertea et al. Nature Protocols, 2016

gffcompare

- gffcompare will compare a merged transcript GTF with known annotation, also in GTF/GFF3 format
- http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html#cuffcompare-output-files

Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

We are on a Coffee Break & Networking Session