

Canadian Bioinformatics Workshops

www.bioinformatics.ca
bioinformaticsdotca.github.io

This page is available in the following languages:

Afrikaans Ӣunrapoai Català Dansk Deutsch Eλλnvwά English English (CA) English (GB) English (US) Esperanto
 Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
 Euskara Suomiεki français français (CA) Galego ગુજરાતi hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
 Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik срpski (latinica) Sotho svenska
 中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

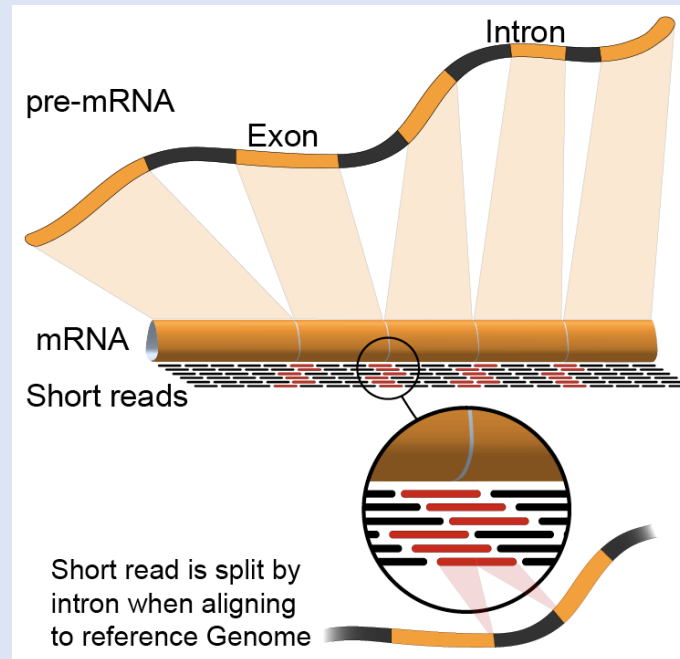
[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.
 This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

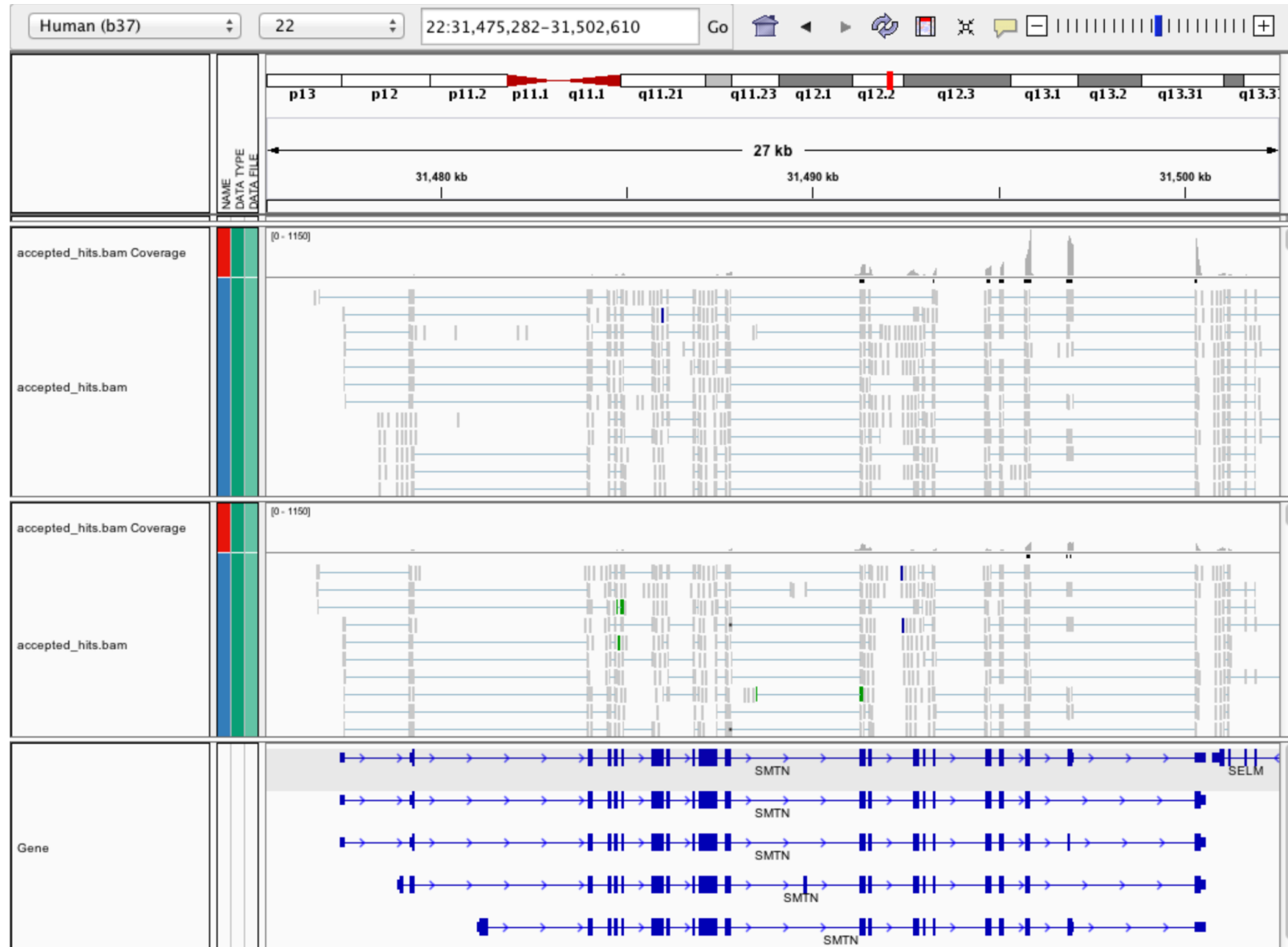
[Learn how to distribute your work using this licence](#)

RNA-Seq Module 3: Abundance Estimation

Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian
RNA-seq Analysis 2024. June 17-19, 2024



Expression estimation for known genes and transcripts



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)
 - $FPKM = (C / (N \times L)) \times 1,000 \times 1,000,000$
 - $FPKM = (1,000,000,000 \times C) / (N \times L)$
 - $FPKM = (C / (N / 1,000,000)) / (L/1000)$
- 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)

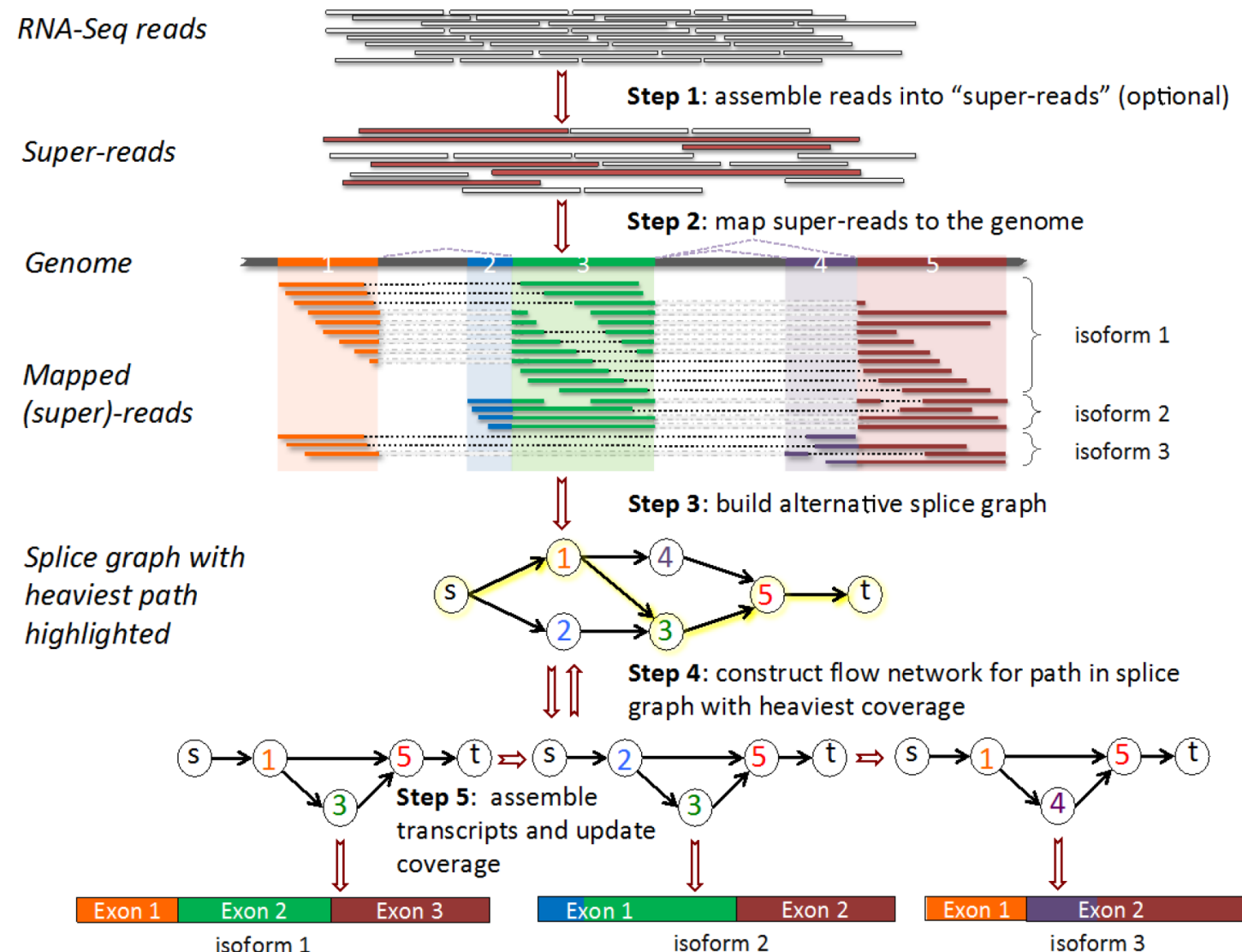
- The sum of all TPMs in each sample is the same. Easier to compare across samples!
- <http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/>
- <https://www.ncbi.nlm.nih.gov/pubmed/22872506>

How does StringTie work?

Map reads to the genome

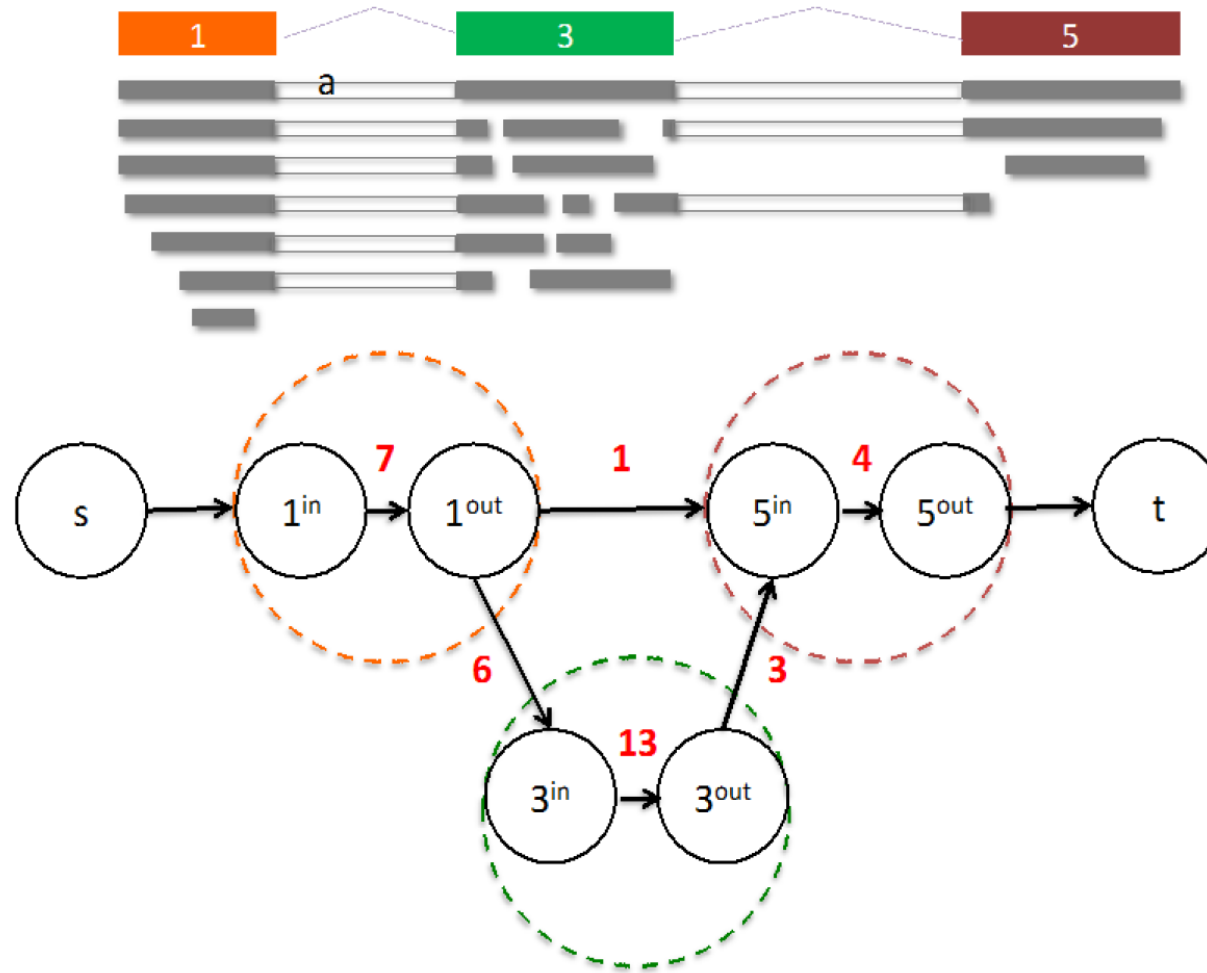
Infer isoforms:

- 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 -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

Workshop Sponsors:



Canadian Centre for
Computational
Genomics



HPC4Health



OICR
Ontario Institute
for Cancer Research



GenomeCanada