



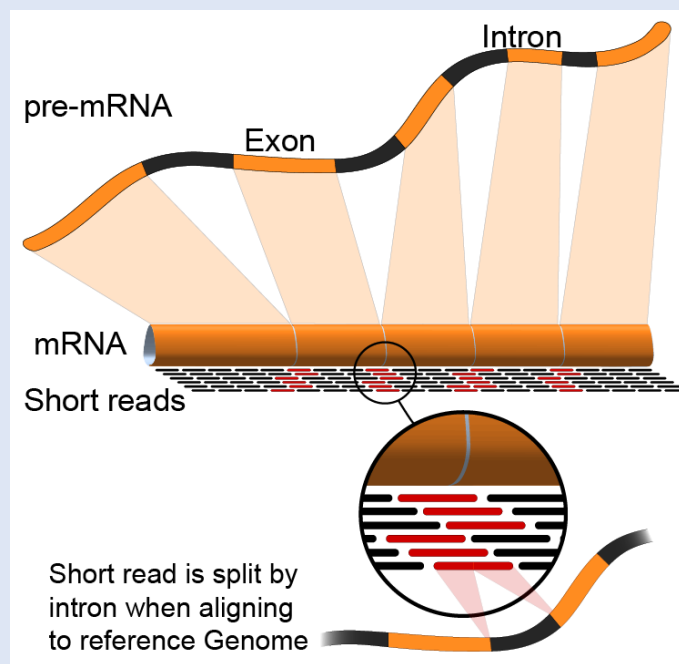
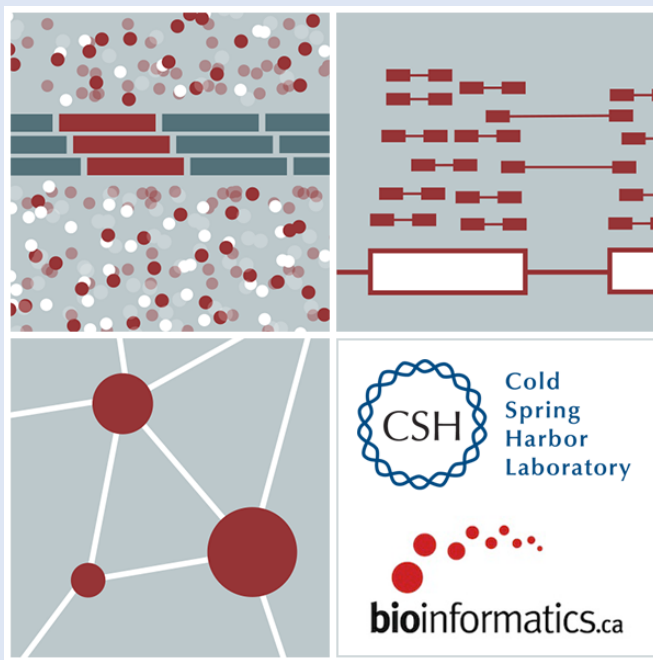
Cold
Spring
Harbor
Laboratory

RNA-Seq Module 4

Alignment Free Expression Estimation (Kallisto)

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

What is a k-mer?

- A fixed sized (**K**) sequence
- A string of length **N** contains **$N-K+1$** k-mers

1-mer

A
C
G
T

2-mer

AA	AC	AG	AT
CA	CC	CG	CT
GA	GC	GG	GT
TA	TC	TG	TT

ATTCGACAGTAGCCATGACTGG

- One can build K -mer index to represent a string

7-mer	iD	N
ATTCGAC	1	1
TTCGACA	2	1
TCGACAG	3	1
...		

Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms Rob Patro, Stephen M. Mount, and Carl Kingsford. *Manuscript Submitted* (2013) <http://www.cs.cmu.edu/~ckingsf/class/02714-f13/Lec05-sailfish.pdf>

<https://www.slideshare.net/duruofei/cmsc702-project-final-presentation>

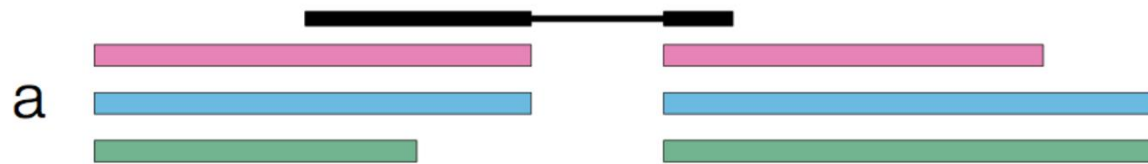
Alignment free approaches for transcript abundance

1. Obtain reference transcript sequences
 - e.g. Ensembl, Refseq, or GENCODE
2. Build a **k-mer index** of all of the k-mers in each transcript sequence
 - Store each k-mer and its position within the transcript. “hashing”

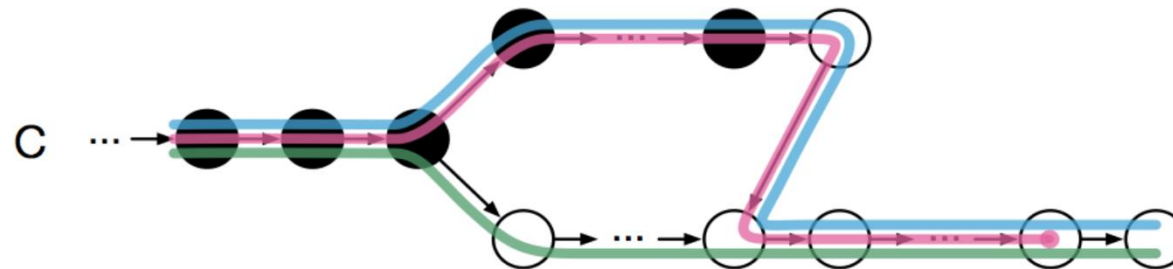
Alignment free approaches for transcript abundance

3. Count number of times each k-mer occurs within each RNAseq read

- Model relationship between RNA-seq read k-mers and the transcript k-mer index.
- What transcript is the most likely source for each read?
- Called “pseudoalignment”, “quasi-mapping”, etc.



Bray, 2016 doi:10.1038/nbt.3519



<https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html>

4. Handle sequencing errors, isoforms, ambiguity, and determine abundance estimates

- Transcriptome de Bruijn graphs, likelihood function, expectation maximization, etc.

Advantages/disadvantages of alignment free approaches

- Advantages
 - Very fast and efficient
 - Similar accuracy to alignment based approach but with much, much shorter run time.
 - Do not need a reference genome, only a reference transcriptome
- Disadvantages
 - You don't get a proper BAM file (though a pseudo-bam can be created)
 - Information in reads with sequence errors may be ignored
 - Limited potential for transcript discovery, variant calling, fusion detection, etc.

Common alignment free tools

- Sailfish
 - “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.” 2014
 - <https://www.ncbi.nlm.nih.gov/pubmed/24752080>
- RNA-Skim
 - “RNA-Skim: a rapid method for RNA-Seq quantification at transcript level.” 2014
 - <https://www.ncbi.nlm.nih.gov/pubmed/24931995>
- Kallisto
 - “Near-optimal probabilistic RNA-seq quantification.” 2016
 - <https://www.ncbi.nlm.nih.gov/pubmed/27043002>
- Salmon
 - “Salmon provides fast and bias-aware quantification of transcript expression.” 2017
 - <https://www.ncbi.nlm.nih.gov/pubmed/28263959>

Which is best?

- Somewhat controversial ...
- <https://liorpachter.wordpress.com/2017/08/02/how-not-to-perform-a-differential-expression-analysis-or-science/>
- Various sources suggest that Salmon, Kallisto, and Sailfish results are quite comparable
- Usability, documentation, and supporting downstream tools could be used to decide

We are on a Coffee Break & Networking
Session