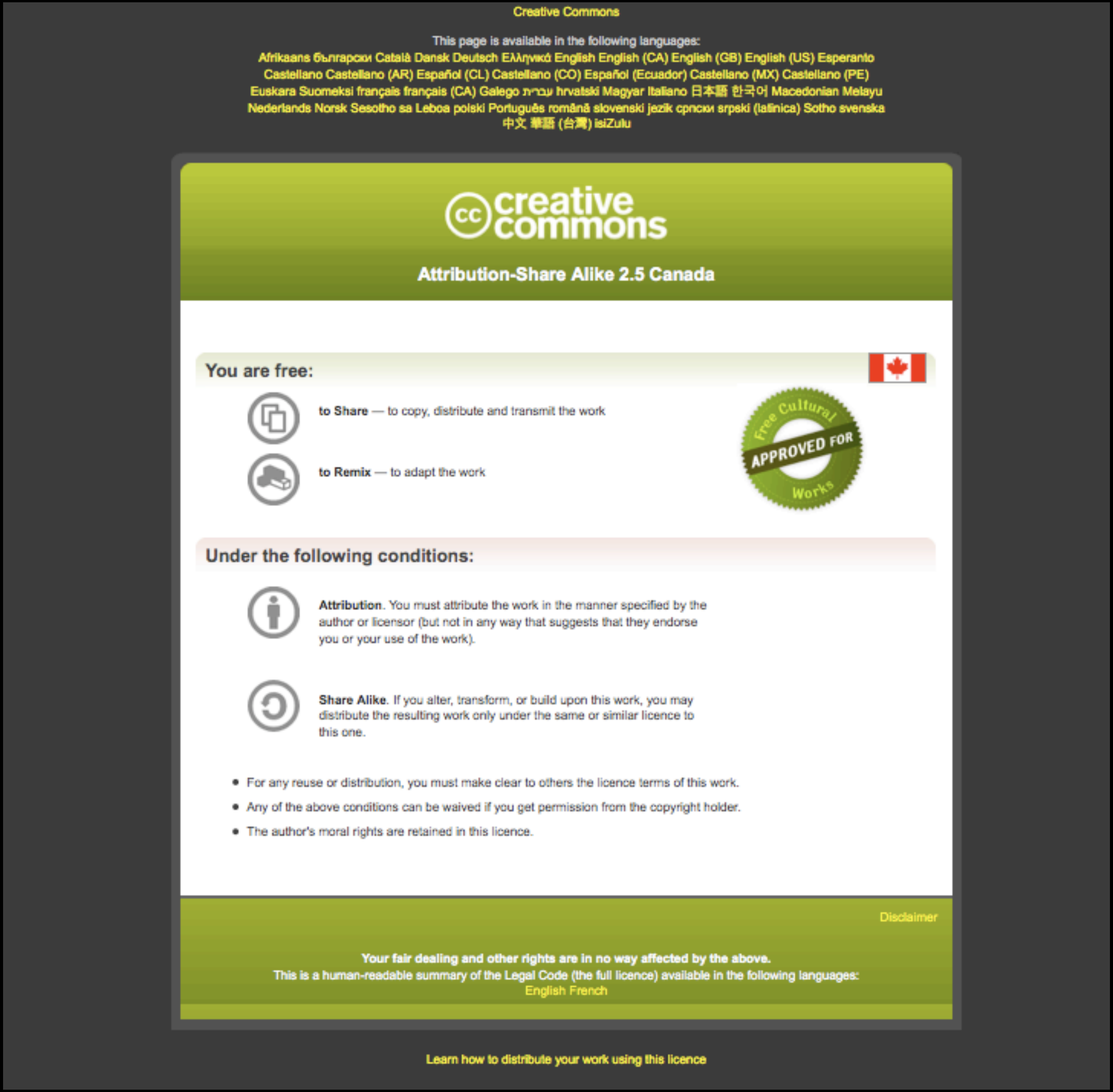
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca



Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

creative commons

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

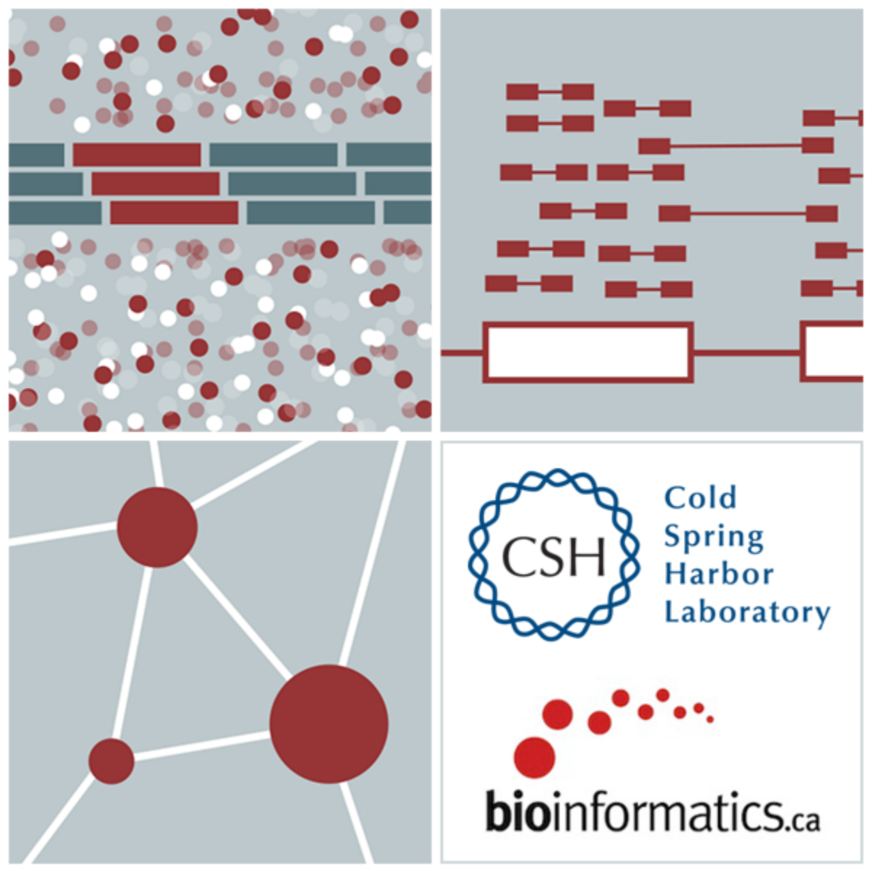
RNA-Seq Module 4
Alignment Free Expression Estimation (lecture)

Malachi Griffith, Obi Griffith

Informatics for RNA-seq Analysis

May 28-30, 2018

bioinformatics.ca



THE GENOME INSTITUTE at Washington University

Learning objectives of the course

- Module 0: Introduction to Cloud Computing
- Module 1: Introduction to RNA Sequencing
- Module 2: Alignment and Visualization
- Module 3: Expression and Differential Expression
- **Module 4: Alignment Free Expression Estimation**
- Module 5: Isoform Discovery and Alternative Expression

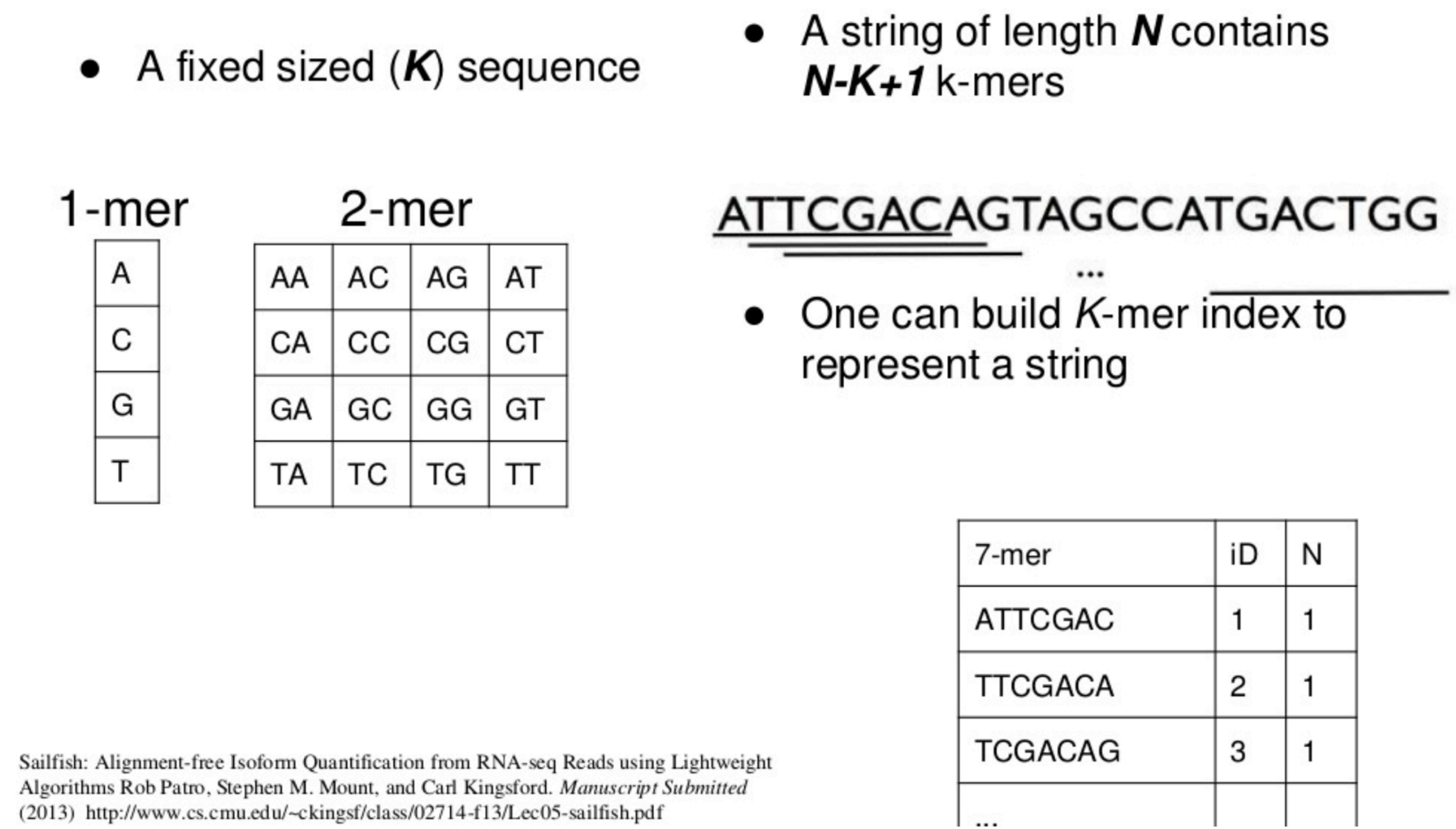
- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a 'reasonable' amount of time with modest computer resources
  - Self contained, self explanatory, portable

Learning objectives of module 4

- Alignment free estimation of transcript abundance

- Introduction to k-mers
- Alignment free tools
  - Sailfish, RNA-Skim, Kallisto, Salmon
- Abundance estimation and differential expression analysis with Kallisto and Sleuth

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

Basic concept of 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 occurring in each transcript sequence
   - Store each k-mer and its position within the transcript. "hashing"
3. Parse all RNA-seq reads and count how many times each k-mer occurs within each 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.
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
  - 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