

# Canadian Bioinformatics Workshops

[www.bioinformatics.ca](http://www.bioinformatics.ca)

[bioinformaticsdotca.github.io](https://bioinformaticsdotca.github.io)

Supported by



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 slovenščina jezik srpski (latinica) Sotho svenska  
中文 漢語 (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](#)

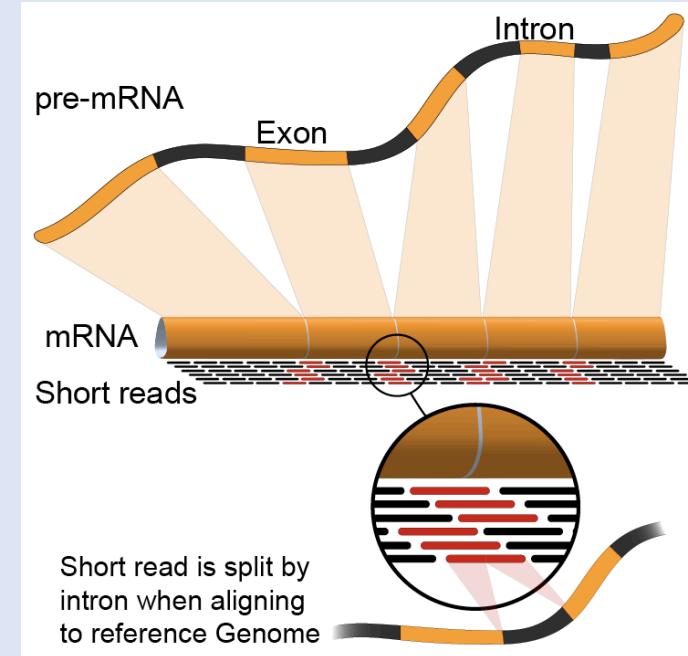
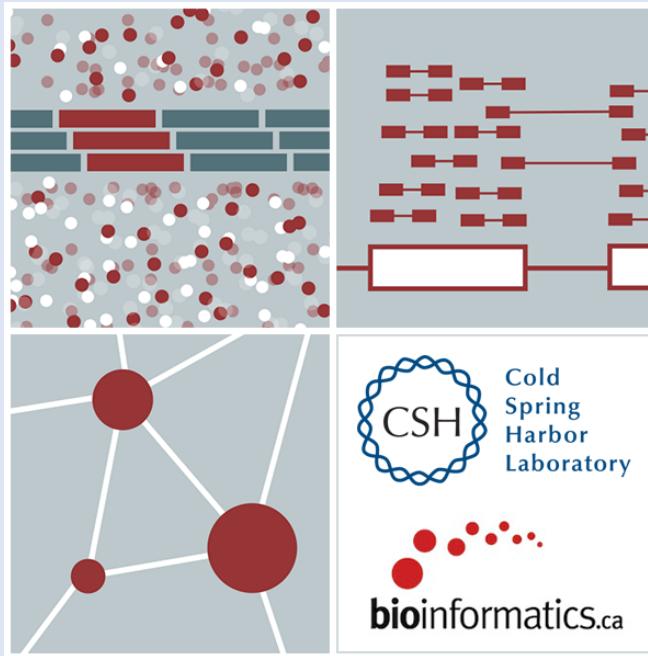
Learn how to distribute your work using this licence

# Indexing

Kelsy Cotto, Obi Griffith, Malachi Griffith, Saad Khan, Allegra Petti, Huiming Xia

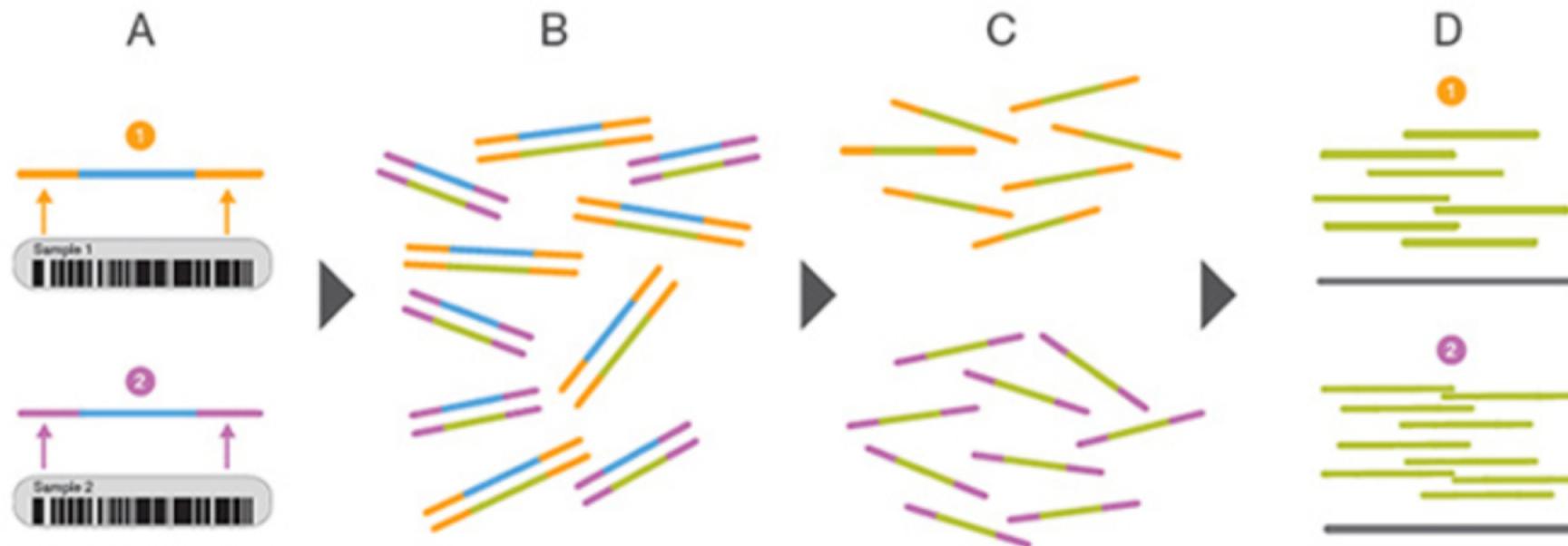
Informatics for RNA-seq Analysis

June 17-19, 2020



# “Index” has many different meanings

- Indexes can refer to unique barcodes used for multiplexing DNA before sequencing



<https://www.illumina.com/science/technology/next-generation-sequencing/multiplex-sequencing.html>

# Indexing in bioinformatics/CS enables rapid access

- Indexing is a recurring theme in genome analysis
  - Files are \*big\* - scanning through them can take a long time
  - Indexing builds a table-of-contents so that we can jump directly to specific positions
- 
- Indexing may require significant compute/time but typically only occurs once
  - Each application may require a different indexing strategy

# What's inside a fasta's index file? (.fai)

contig name	bases in contig	byte index of the file where the contig begins	bases per line	bytes per line
chr1	248956422	6	60	61
chr2	242193529	253105708	60	61
chr3	198295559	499335802	60	61
chr4	190214555	700936293	60	61
chr5	181538259	894321097	60	61
chr6	170805979	1078885000	60	61
chr7	159345973	1252537752	60	61
chr8	145138636	1414539498	60	61
chr9	138394717	1562097118	60	61
chr10	133797422	1702798421	60	61

# Example index applications and associated files

Source file	Indexed file	Indexing tool	Use case
.bam	.bai	samtools index	Visualize bam in IGV
.fasta	.fai	faidx	Extract specific sequences from ref genome
.vcf	vcf.gz.tbi	bgzip/tabix	Pull out specific variants
.bed	.bed.gz.tbi	bgzip/tabix	extract specific genomic regions

# Indexing is also essential for alignment

- Finding out where to place a read in the genome is impractical unless matches can be quickly found
- All read aligners use some kind of indexing
- These indices must be “built” once for a reference genome, but can then be used every time the aligner is run
- Different aligners use different indexing schemes that are not compatible

# We are on a Coffee Break & Networking Session

Workshop Sponsors:

compute | calcul  
canada | canada



Canadian Centre for  
Computational  
Genomics

**MicCM** McGill initiative in  
Computational Medicine