

Canadian Bioinformatics Workshops

www.bioinformatics.ca
bioinformaticsdotca.github.io



CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

Canonical URL: <https://creativecommons.org/licenses/by-sa/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

 **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

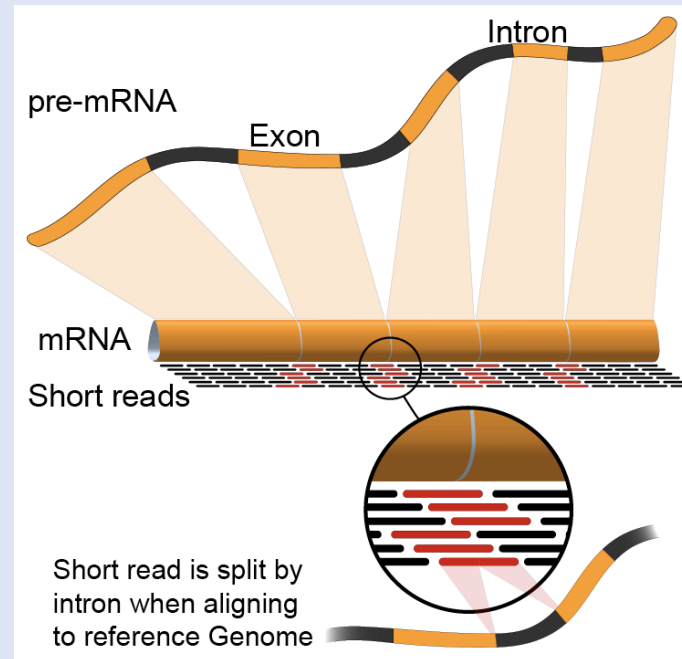
Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

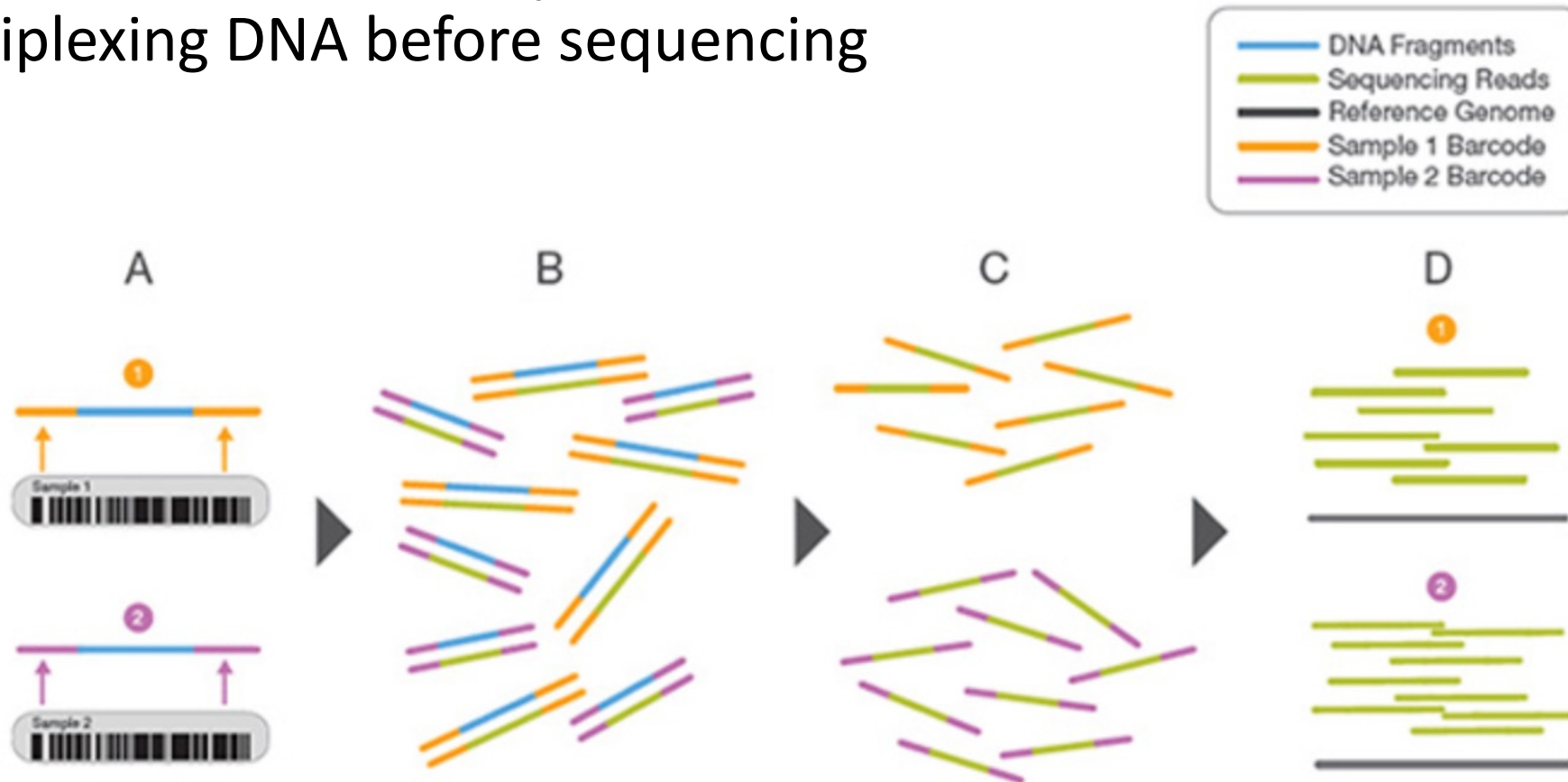
RNA-Seq Module 1: Indexing

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



“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:



Canadian Centre for
Computational
Genomics



HPC4Health



GenomeCanada