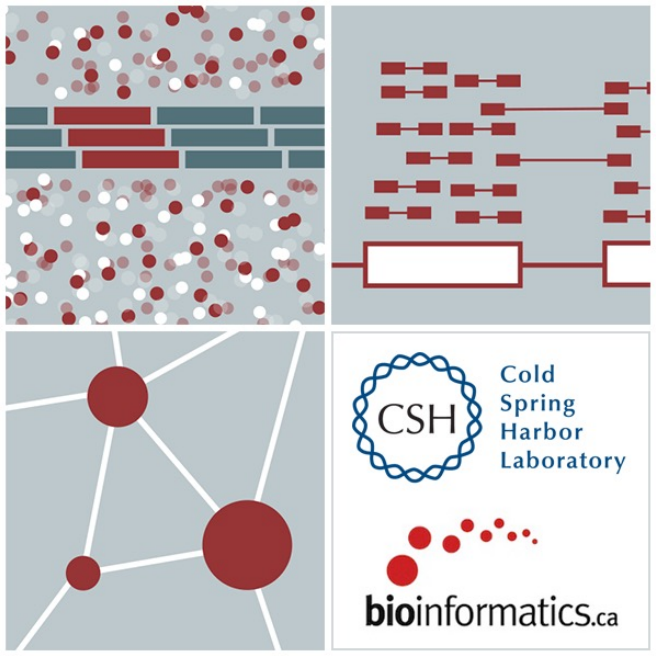
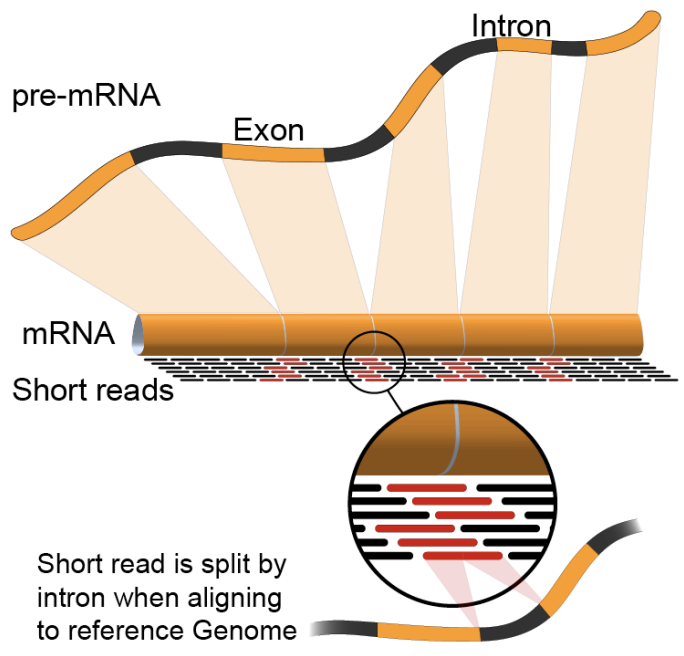
Cold Spring Harbor Laboratory

RNA-Seq Module 1: **Indexing**

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal

Advanced Sequencing Technologies & Bioinformatics Analysis November 5-19, 2023

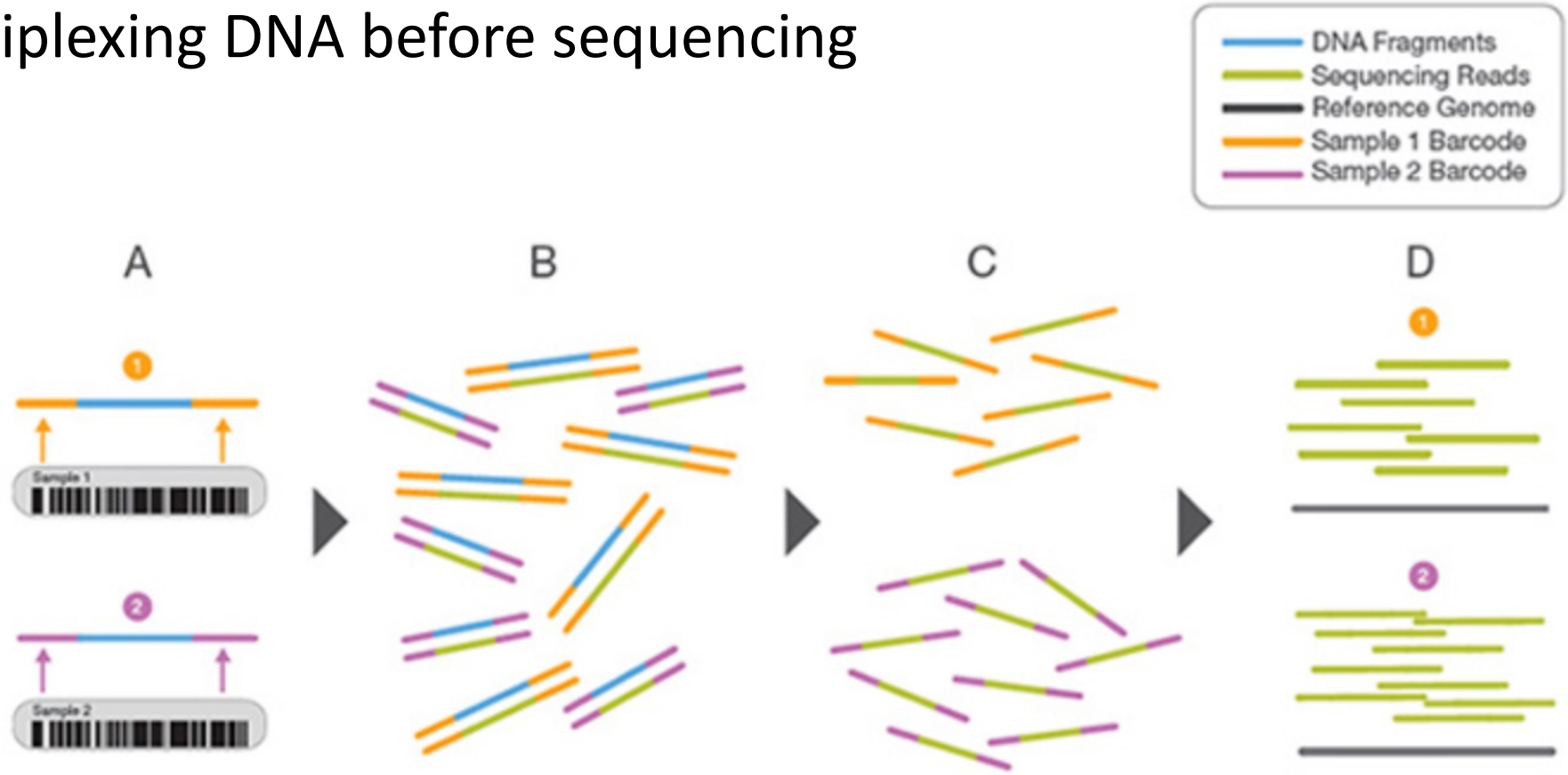




Washington University in St. Louis
School of Medicine

“Index” has many different meanings

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



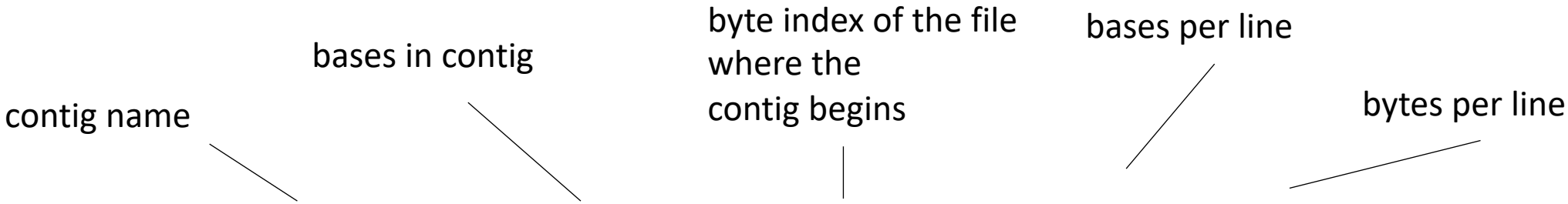
https://www.illumina.com/science/technology/next-generation-sequencing/multiplex-sequencing.html

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)



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