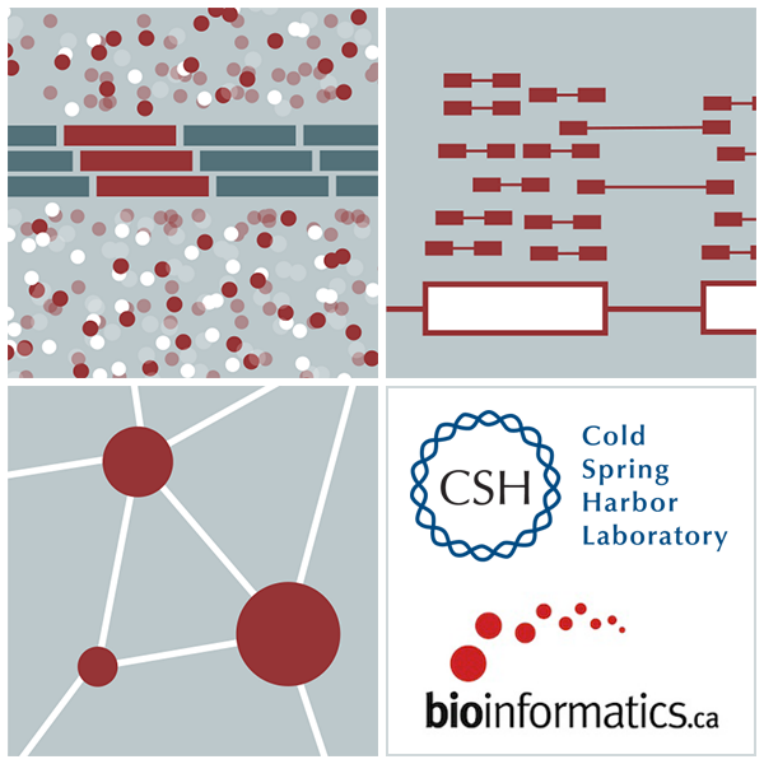
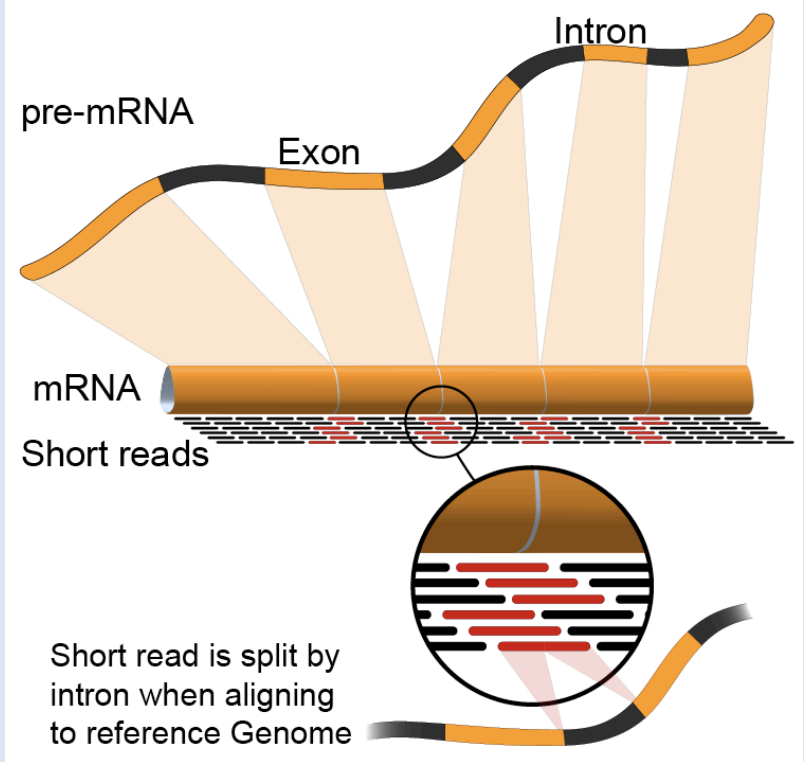
RNA-Seq Module 1 Indexing

Kelsy Cotto, Felicia Gomez,
Obi Griffith, Malachi Griffith, Huiming Xia

Advanced Sequencing Technologies & Applications

November 5- 16, 2019

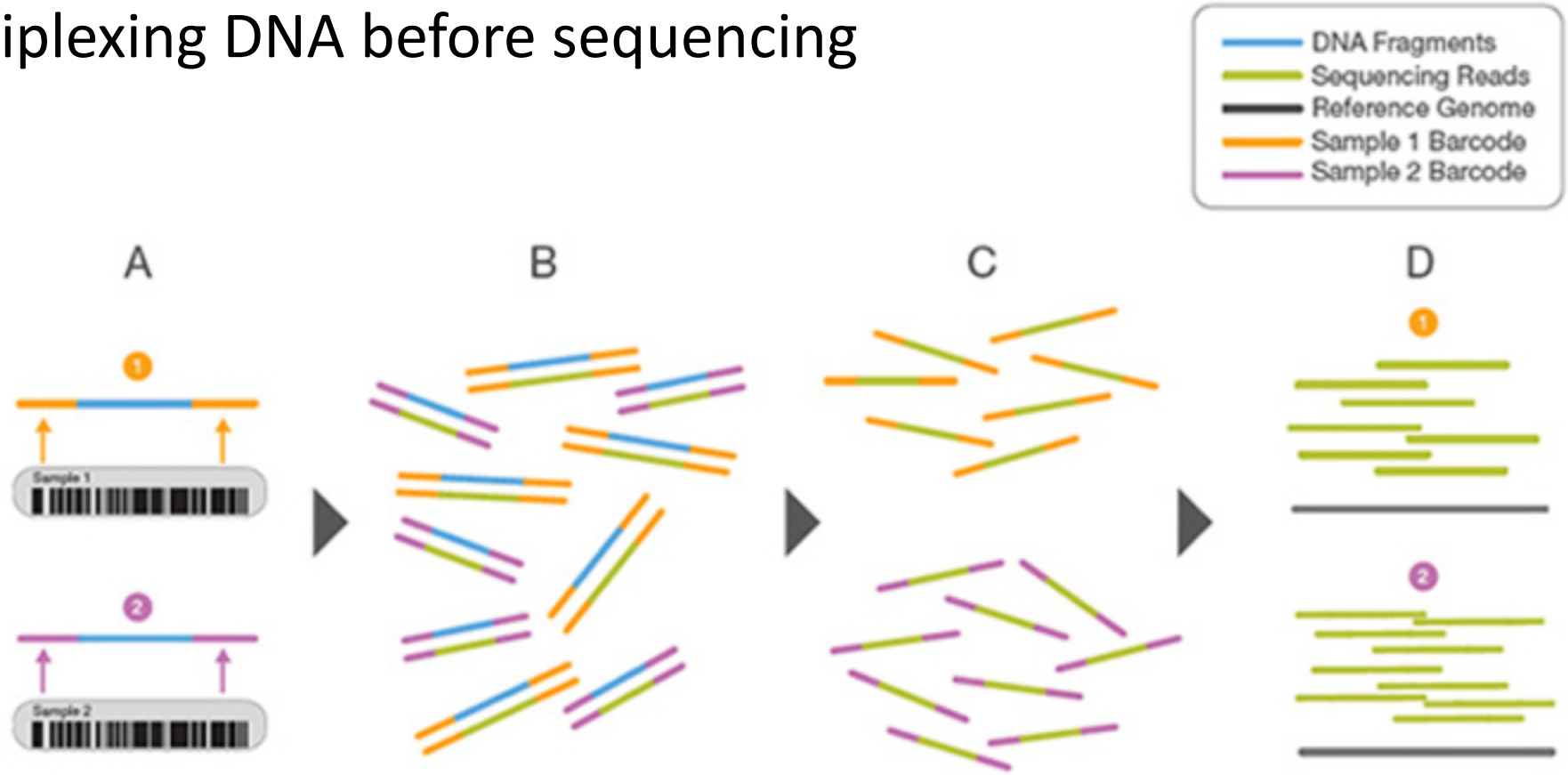




THE GENOME INSTITUTE
at Washington University

“Index” has many different meanings

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



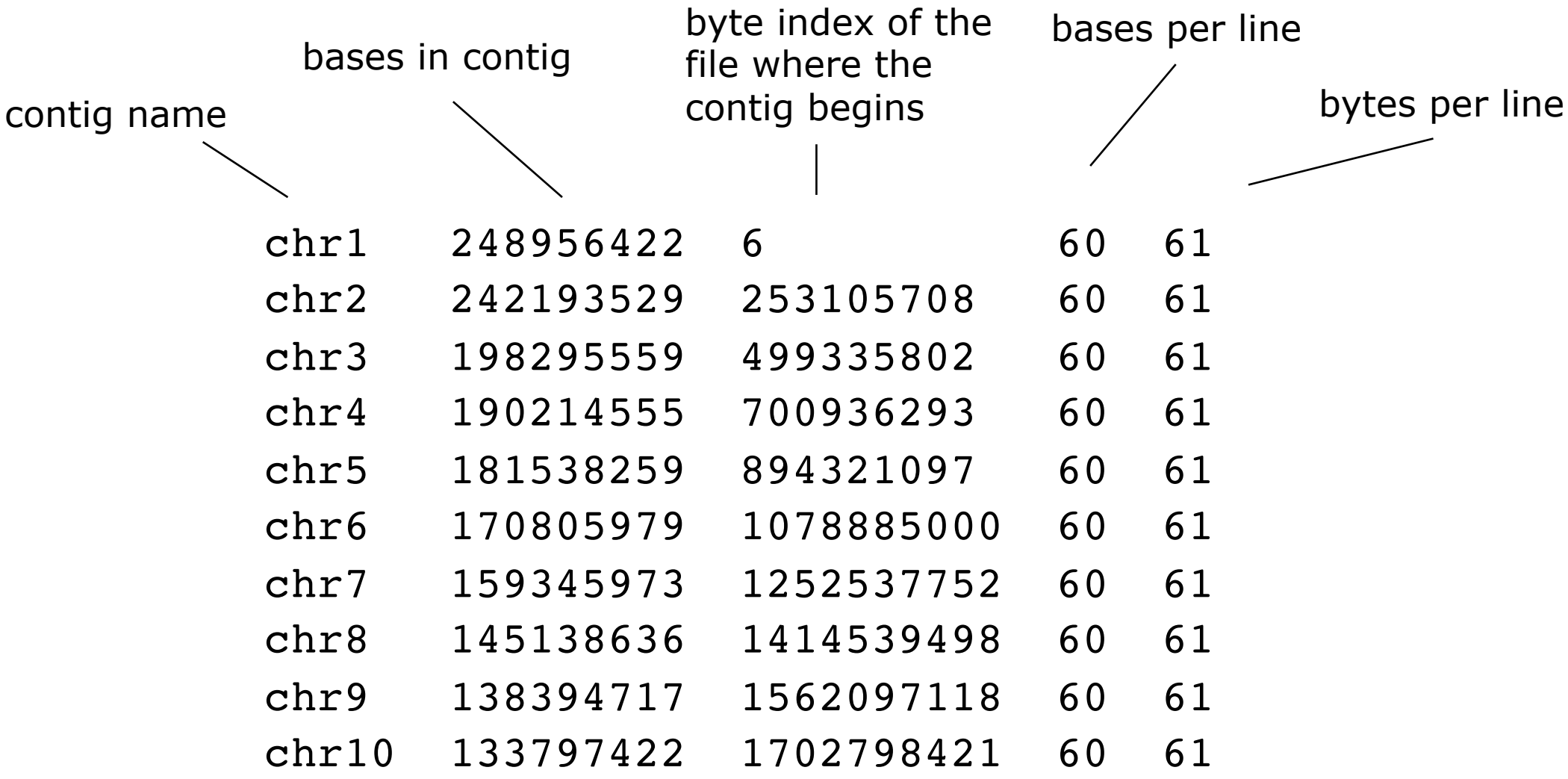
https://www.illumina.com/science/technology/next-generation-sequencing/multiplex-sequencing.html

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)



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