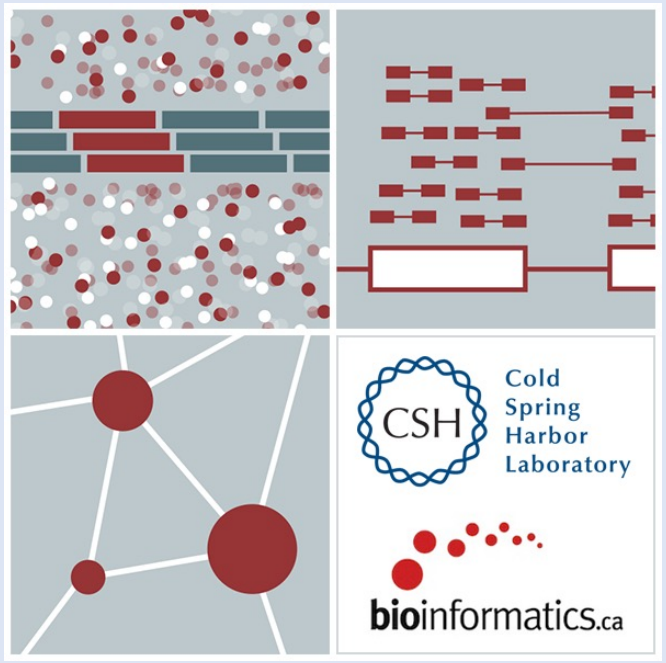
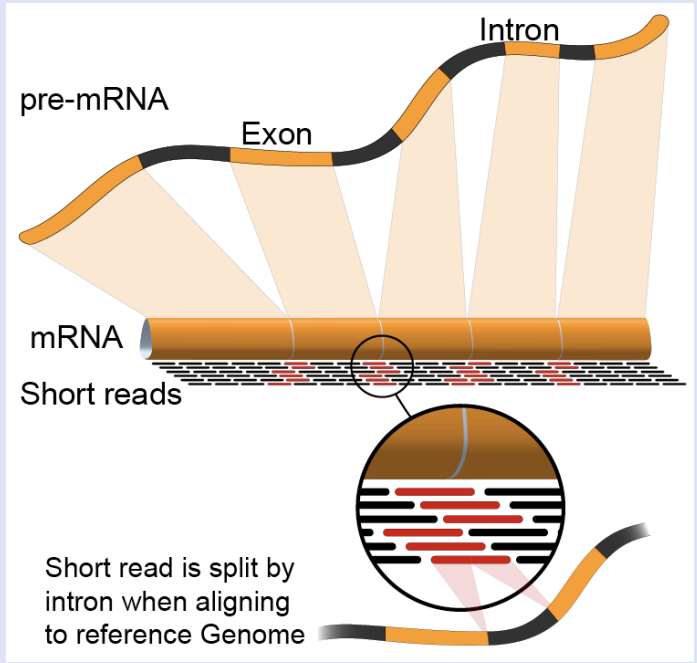
Cold Spring Harbor Laboratory

RNA-Seq Module 3
Abundance Estimation and Differential Expression

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

Alternatives to FPKM

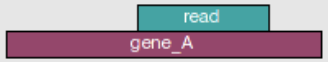
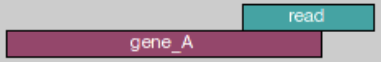
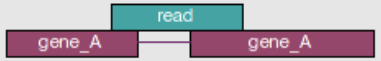
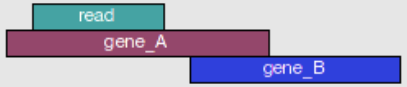
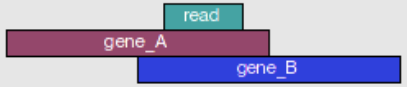
- Raw read counts for differential expression analysis
  - Assign reads/fragments to defined genes/transcripts, get “raw counts”
    - Transcript structures could still be defined by something like Stringtie

- HTSeq (htseq-count)
  - https://htseq.readthedocs.io/

```
htseq-count --mode intersection-strict --stranded no --minaqual 1 --type exon --idattr transcript_id
accepted_hits.sam chr22.gff > transcript_read_counts_table.tsv
```

- Caveats of ‘transcript’ analysis by htseq-count:
  - Designed for genes - ambiguous reads from overlapping transcripts may not be handled!
  - http://seqanswers.com/forums/showthread.php?t=18068

HTSeq-count basically counts reads supporting a feature (exon, gene) by assessing overlapping coordinates

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Note, if gene_A and gene_B on opposite strands, sequence data is stranded, and correct HTSeq parameter set then this read may not be ambiguous

Whether a read is counted depends on the nature of overlap and “mode” selected

We are on a Coffee Break & Networking Session