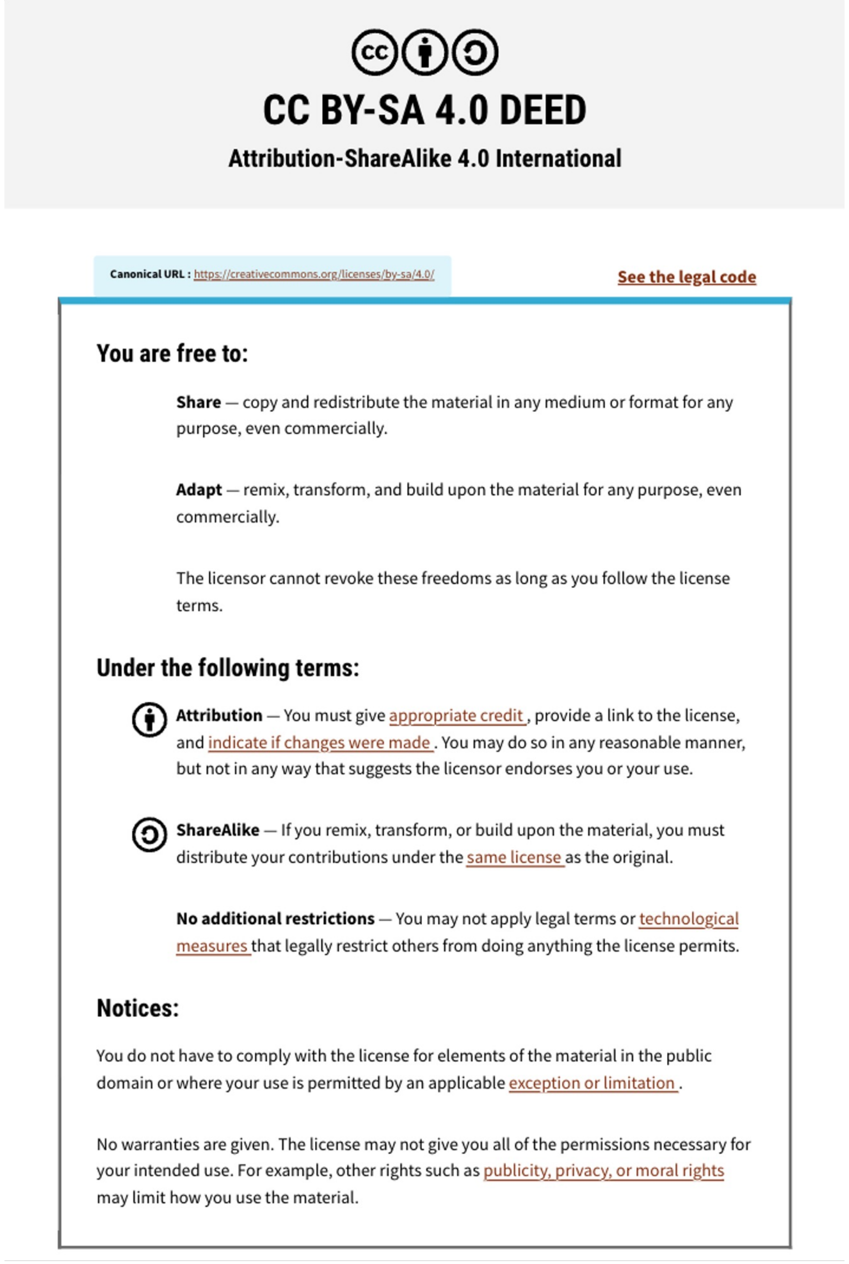
bioinformatics.ca

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 .

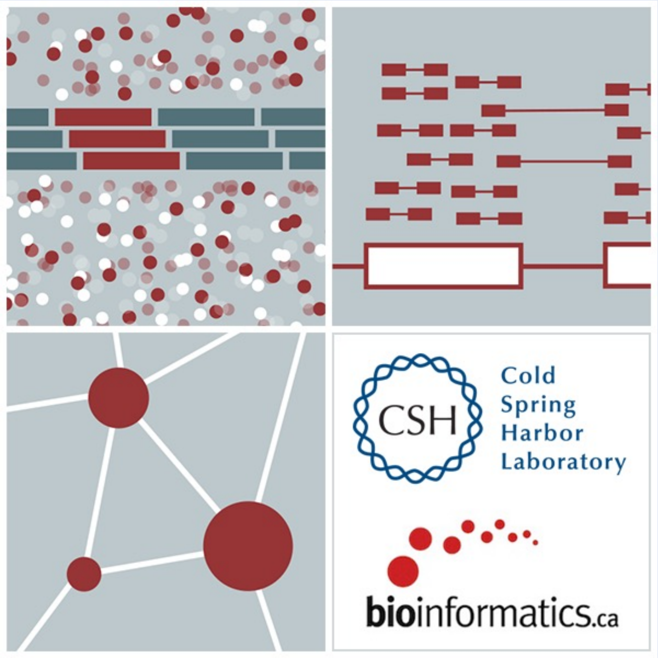
No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

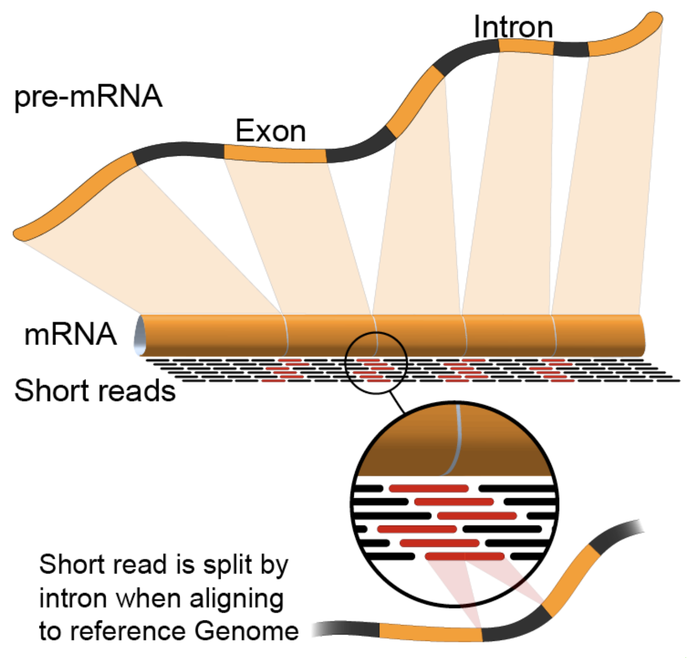
RNA-Seq Module 2: Alignment

Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian

RNA-seq Analysis 2024. June 17-19, 2024

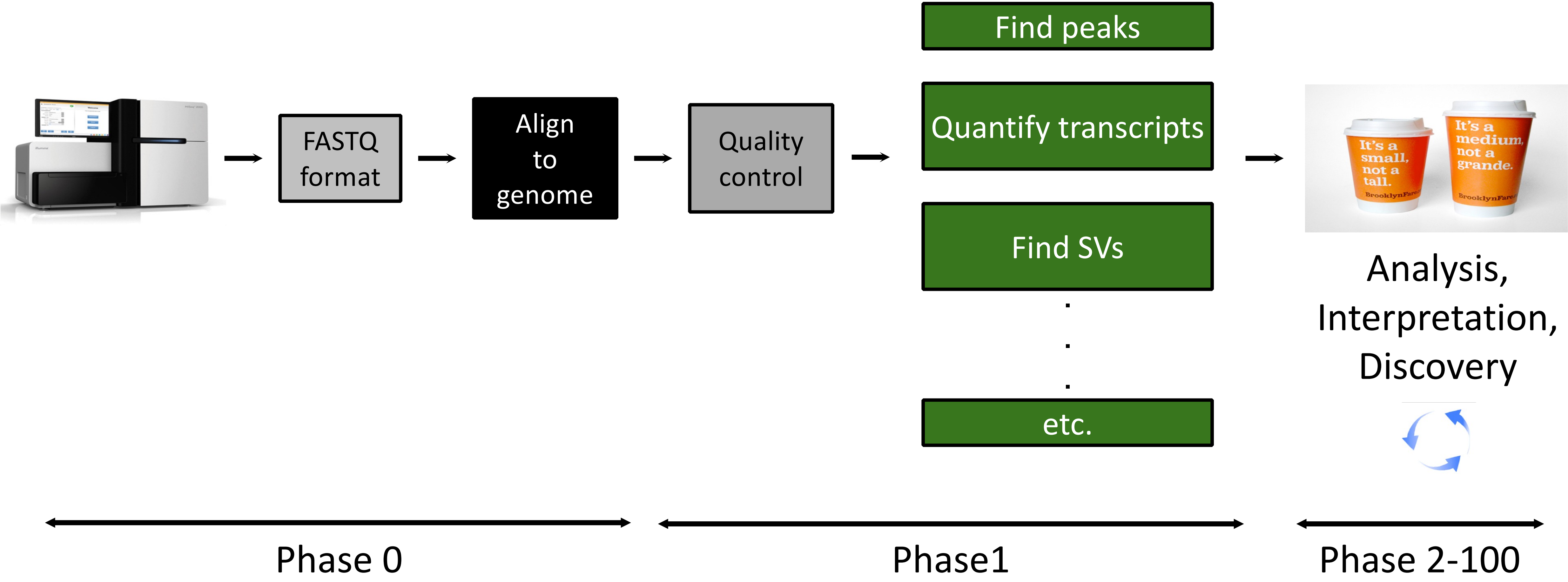
bioinformatics.ca




Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Washington University in St. Louis
School of Medicine

Alignment is central to most genomic research



FASTQ format → Align to genome → Quality control →

Find peaks

Quantify transcripts

Find SVs

. . .

etc.

→ Analysis, Interpretation, Discovery

Phase 0

Phase1

Phase 2-100

Slide courtesy of Andrew Farrell, Obi and Malachi Griffith

Alignment - How does it work?

- Alignment is about fitting individual pieces (reads) into the correct part of the puzzle
- The human genome project gave us the picture on the box cover (the reference genome)
- Imperfections in how the pieces fit can indicate changes to a copy of the picture

Reference: AGCCTGAGACCGTAAAAAA**A**GTCAAG

| | | | | | | | | | | | | | | | |

A read sequence: GAGACCGTAAAAA**C**GTC

↑

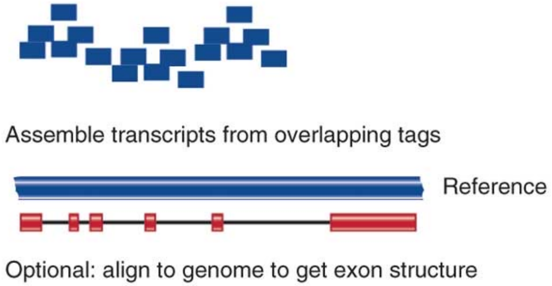
A variant!

RNA-seq alignment challenges

- Computational cost
    - 100’s of millions of reads
- Introns!
    - Align to a transcriptome or align to a genome?
        - Spliced vs. unspliced alignments
- Can I just align my data once using one approach and be done with it?
    - Unfortunately, probably not

Three RNA-seq mapping strategies

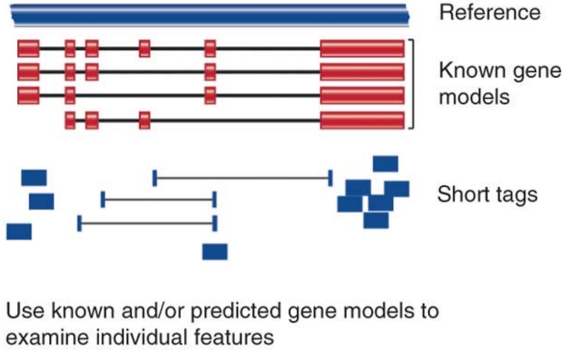
De novo assembly



Assemble transcripts from overlapping tags

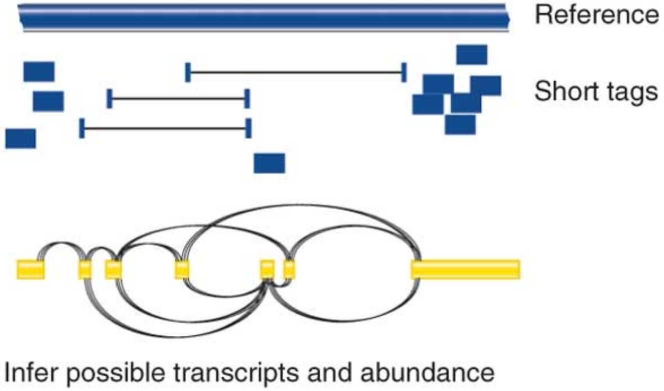
Optional: align to genome to get exon structure

Align to transcriptome



Use known and/or predicted gene models to examine individual features

Align to reference genome



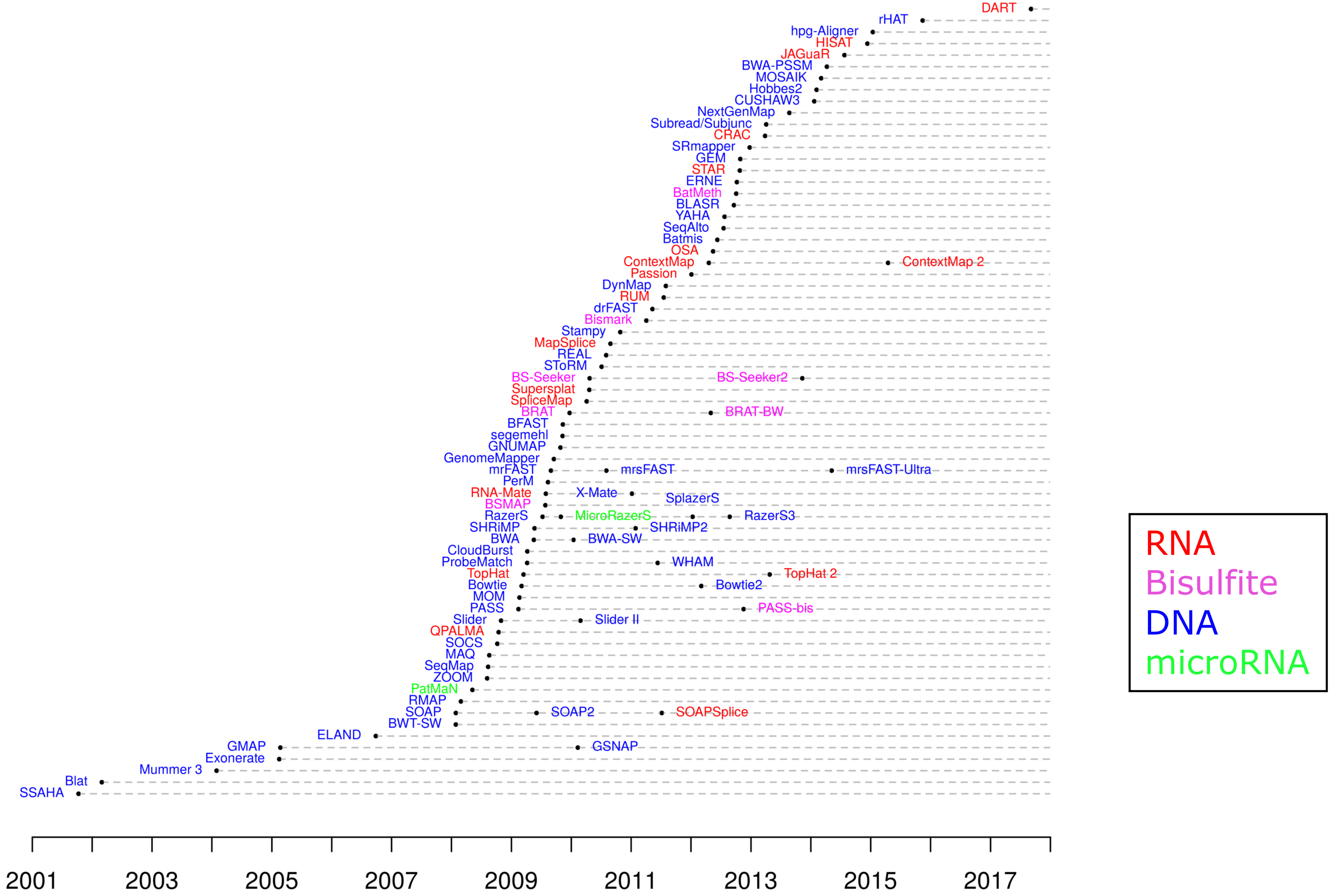
Infer possible transcripts and abundance

Diagrams from Cloonan & Grimmond, Nature Methods 2010

Which alignment strategy is best?

- De novo assembly
  - If a reference genome does not exist for the species being studied
  - If complex polymorphisms/mutations/haplotypes might be missed by comparing to the reference genome
- Align to transcriptome
  - If you have short reads (< 50bp)
  - Relies on known transcripts
- Align to reference genome
  - All other cases
  - Does not rely on known transcripts – allows for discovery
- Each strategy involves different alignment/assembly tools

Which read aligner should I use?



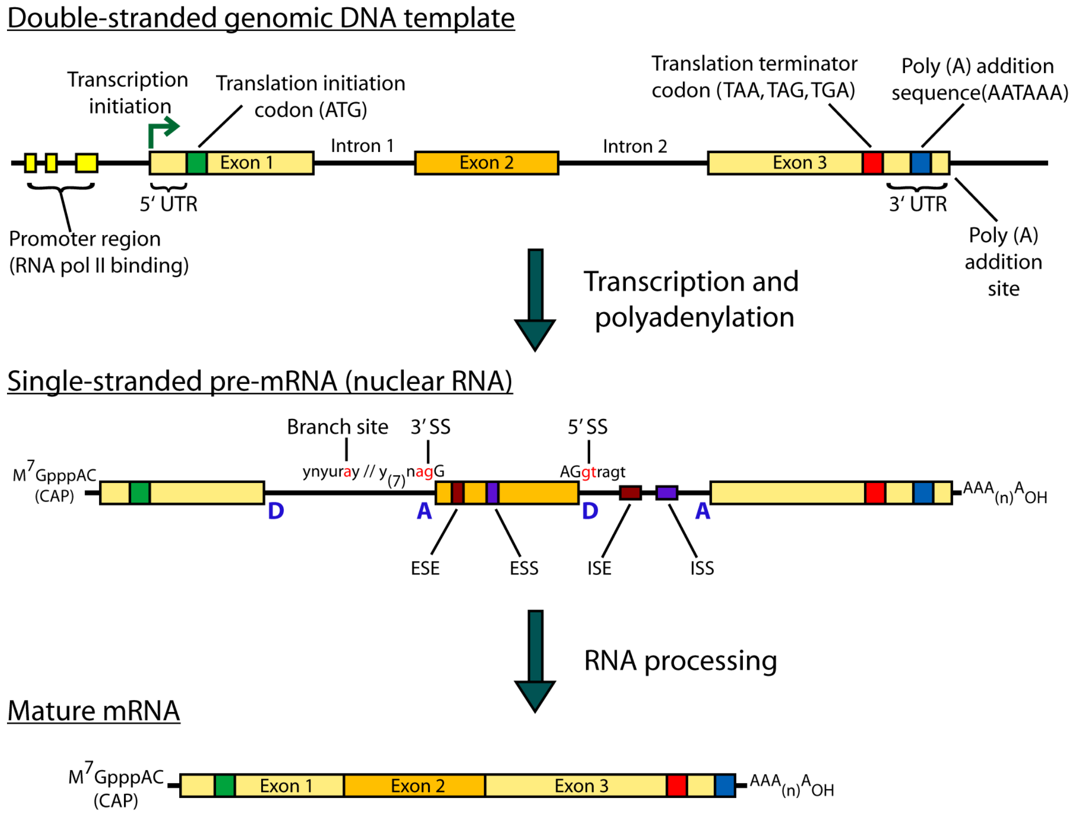
http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Should I use a splice-aware or unspliced mapper?

- The fragments being sequenced in RNA-seq represent mRNA - introns are removed

- But we are usually aligning these reads back to the reference genome

- Unless your reads are short (<50bp) you should use a splice-aware aligner
  - HISAT2, STAR, MapSplice, etc.



Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3′ SS
5′ SS
ynyuray // y(7)nagG
AGgtragt
M7GpppAC (CAP)
AAA(n)AOH
D
A
D
A
ESE
ESS
ISE
ISS
RNA processing
Mature mRNA
M7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
AAA(n)AOH

HISAT/HISAT2

- HISAT is a ‘splice-aware’ RNA-seq read aligner
  - HISAT = **H**ierarchical **I**ndexing for **S**pliced **A**lignments of **T**ranscripts
- Requires a reference genome
- Very fast

- Uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index
- Multiple types of indexes for alignment
  - a whole-genome FM index to anchor each alignment
  - numerous local FM indexes for very rapid extensions of these alignments.
  - Whole-genome indices with SNPs and known transcript structures accounted for

Kim et al. 2015. Nat Methods 12:357–360

HISAT/HISAT2 algorithm

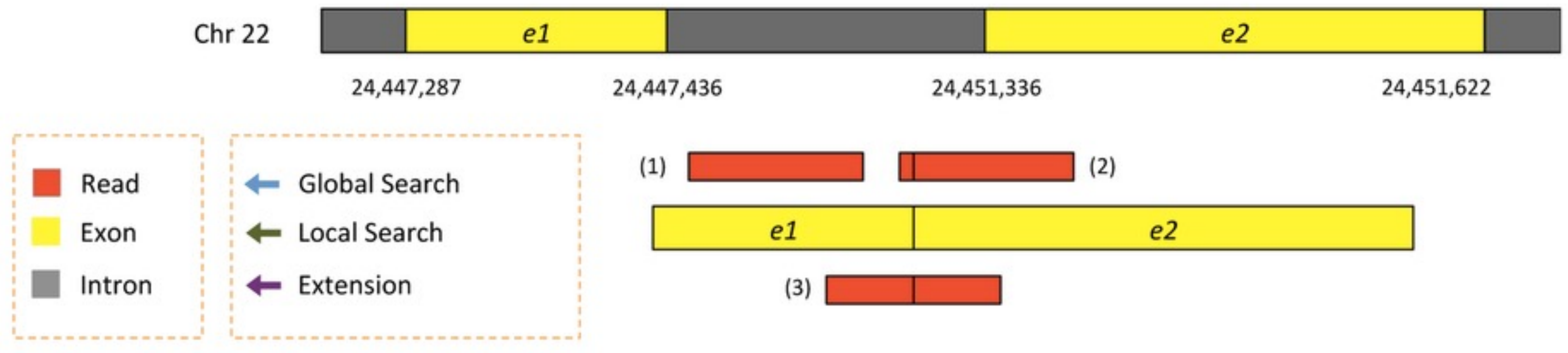
- Uses a hierarchical indexing algorithm + several adaptive strategies
  - based on the position of a read with respect to splice sites

1) Find candidate locations across the whole genome first
   - mapping part of each read using the global FM index
   - Generally identifies one or a small number of candidates.

2) Do local alignment
   - selects one of ~48,000 local indexes for each candidate
   - uses it to align the remainder of the read.

- For paired reads, each mate is separately aligned
  - If a read fails to align, then the alignments of its mate are used as anchors to map the unaligned mate

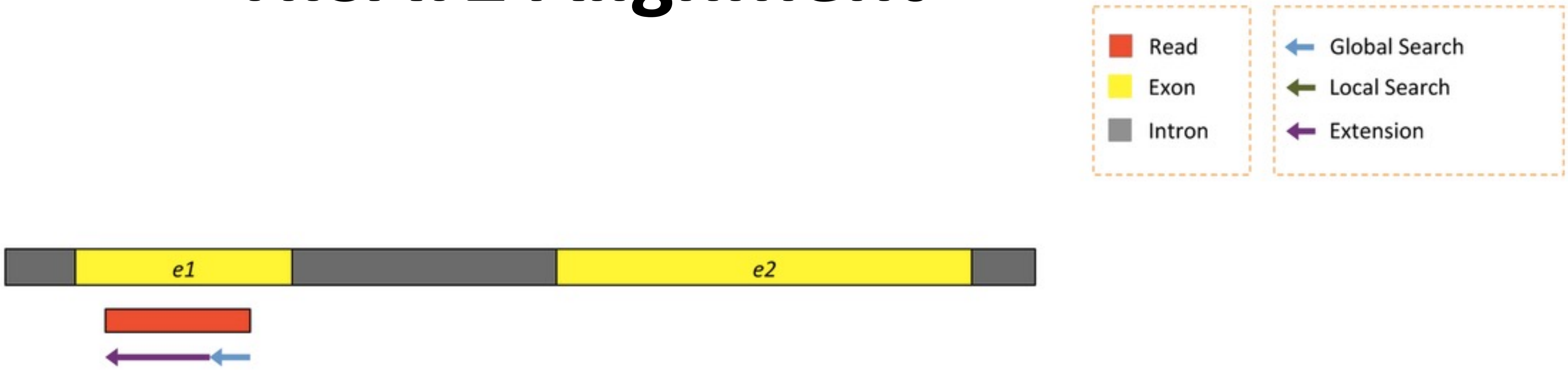
HISAT2 Alignment



- Two exons from chr22
- Three reads

Kim et al. 2015. Nat Methods 12:357–360

HISAT2 Alignment

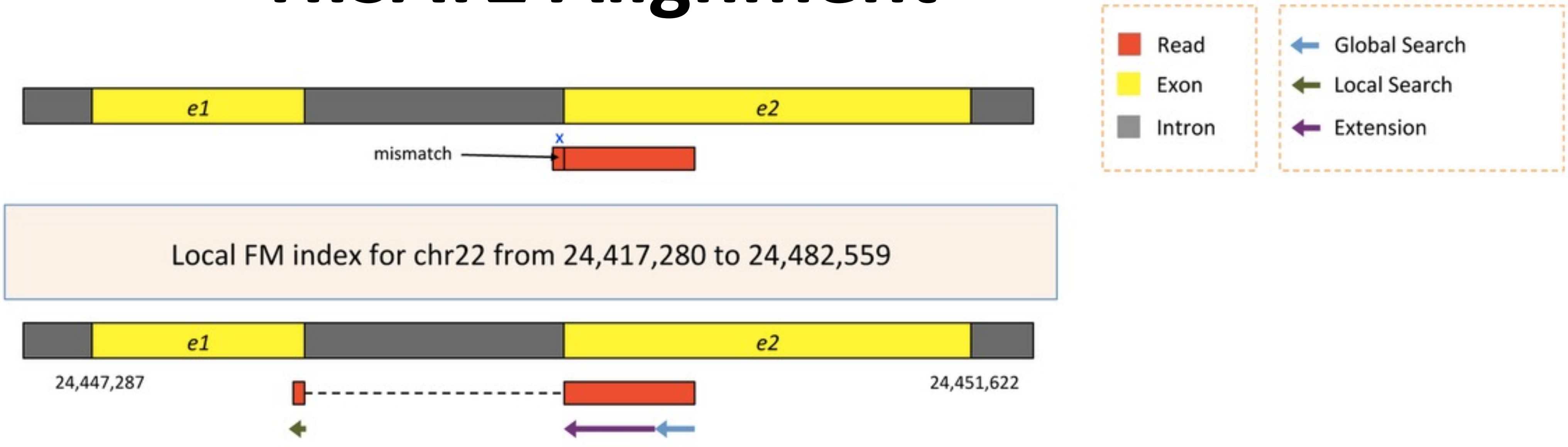


1) Search for read position with global FM index (slower)

2) Once at least 28bp and exactly one location switch to extension mode against reference genome (faster)

Kim et al. 2015. Nat Methods 12:357–360

HISAT2 Alignment



1) Search for read position with global FM index (slower)

2) Extend until mismatch at 93bp (faster)

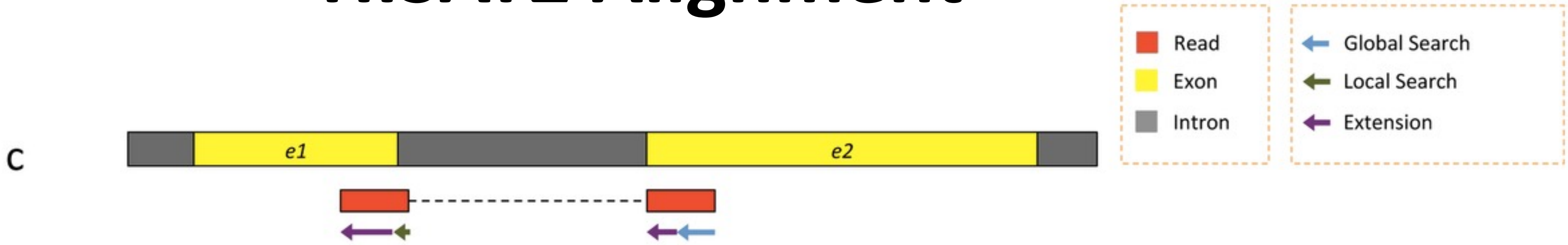
3) Switch to local FM index to align remaining 8bp

- index covers only a small region, so we find just one match

4) Check for compatibility and combine into single spliced alignment

Kim et al. 2015. Nat Methods 12:357–360

HISAT2 Alignment



1) global search until exactly one match of at least 28bp (slower)

2) Extend until mismatch at 51bp (faster)

3) switch to local FM index to align first 8bp of remaining read

- If too many matches increase prefix size

4) Extend again

5) Check for compatibility and combine into single spliced alignment

Kim et al. 2015. Nat Methods 12:357–360

What is the output of HISAT2?

- A SAM/BAM file
  - SAM stands for Sequence Alignment/Map format
  - BAM is the binary version of a SAM file

- Remember, compressed files require special handling compared to plain text files

- How can I convert BAM to SAM?
  - http://www.biostars.org/p/1701/

- Is HISAT2 the only mapper to consider for RNA-seq data?
  - http://www.biostars.org/p/60478/

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics

HPC4Health

OICR Ontario Institute for Cancer Research

Ontario Genomics

GenomeCanada