



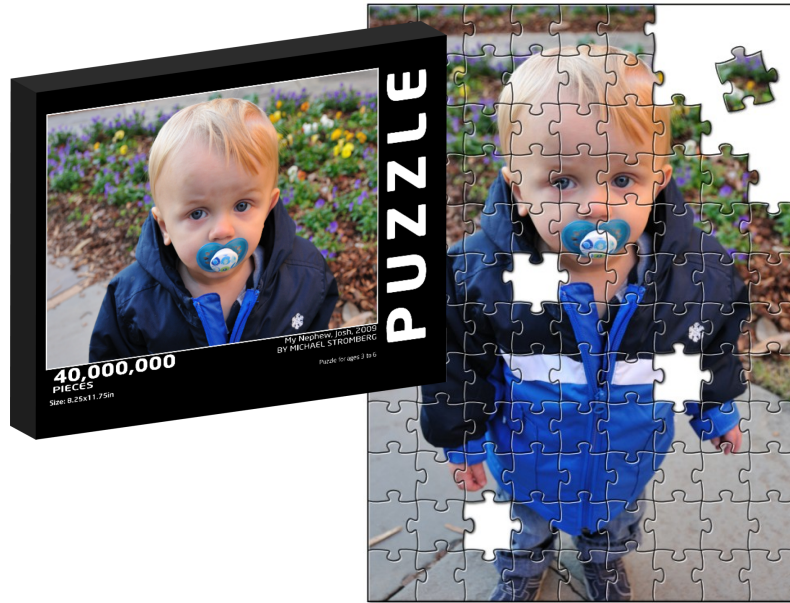
Cold
Spring
Harbor
Laboratory

RNA-Seq Module 2 Alignment

Kelsy Cotto, Felicia Gomez,
Obi Griffith, Malachi Griffith, Huiming Xia
Advanced Sequencing Technologies & Applications
November 5- 16, 2019



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:

GAGACCGTAAAAAA**C**GTC

↑
A variant!

RNA-seq alignment challenges

- Computational cost
 - 100's of millions of reads
- Introns!
 - 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

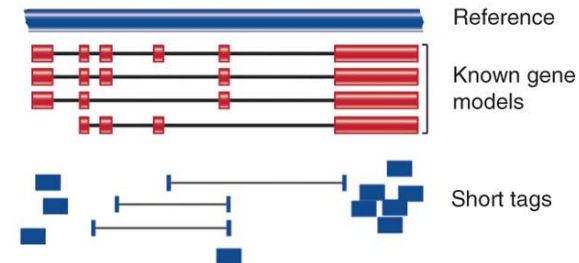


Assemble transcripts from overlapping tags



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



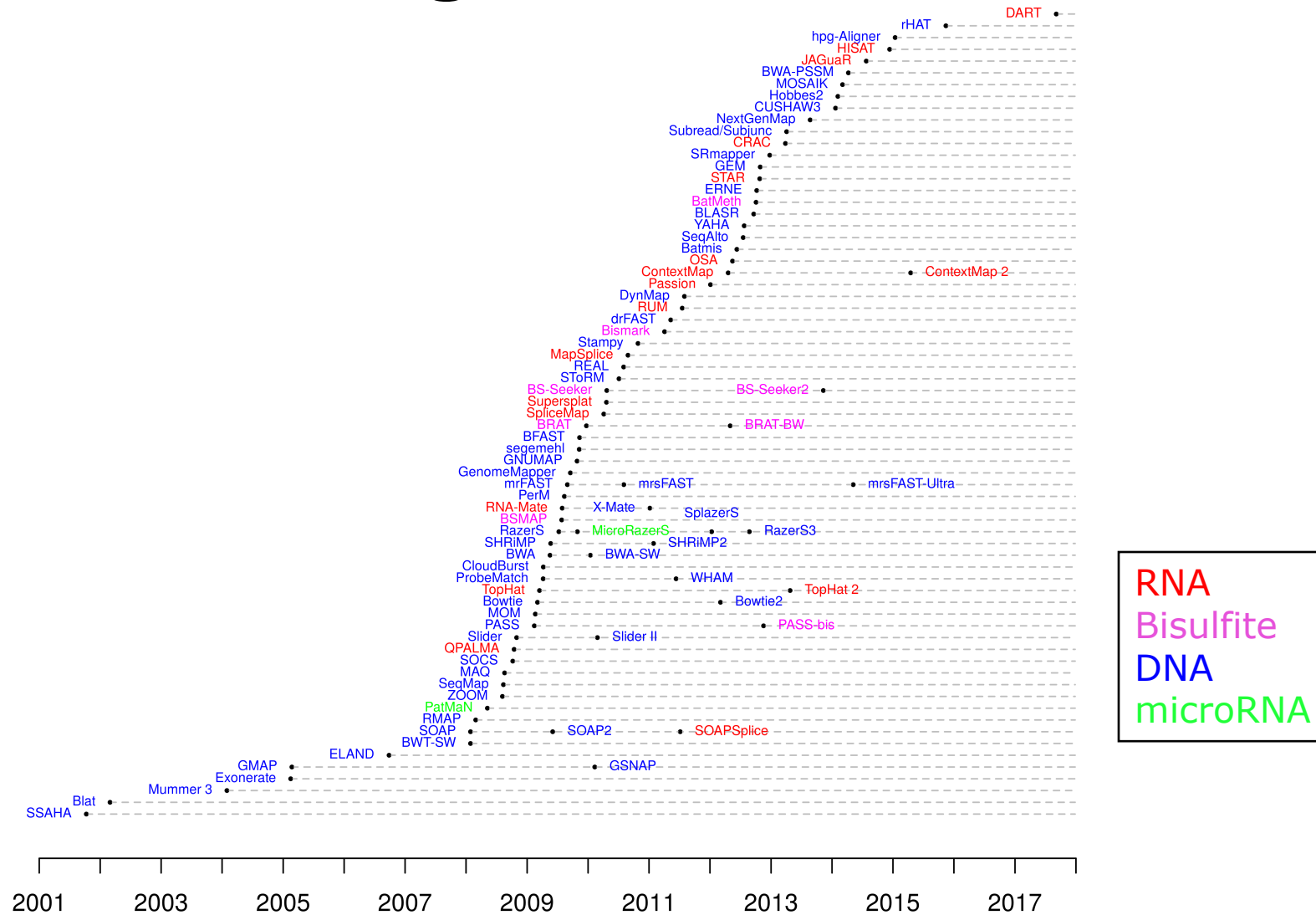
Infer possible transcripts and abundance

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)
- Align to reference genome
 - All other cases
- Each strategy involves different alignment/assembly tools

Which read aligner should I use?

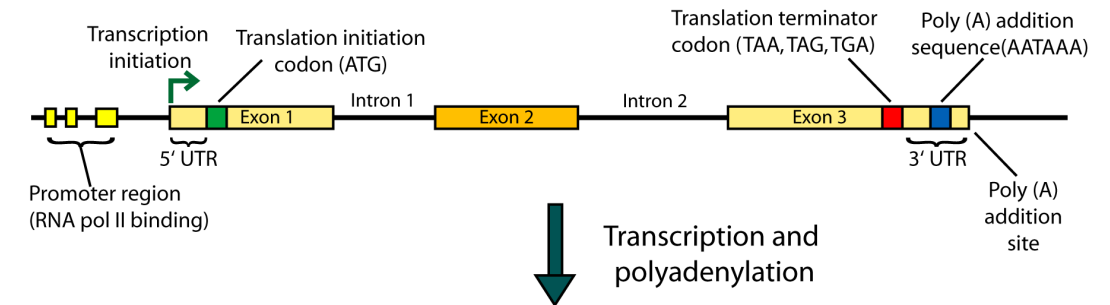


http://wwwdev.ebi.ac.uk/fg/hts_mappers/

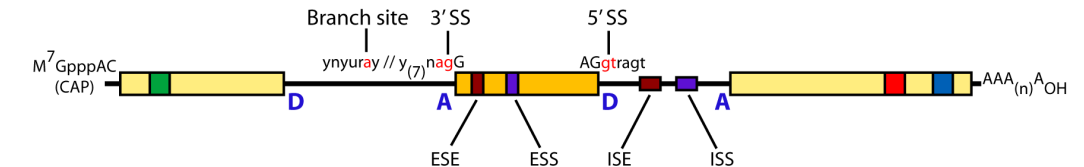
Should I use a splice-aware or unspliced mapper?

- RNA-seq reads may span large introns
- 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.

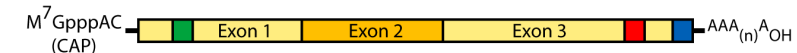
Double-stranded genomic DNA template



Single-stranded pre-mRNA (nuclear RNA)



Mature mRNA



HISAT/HISAT2

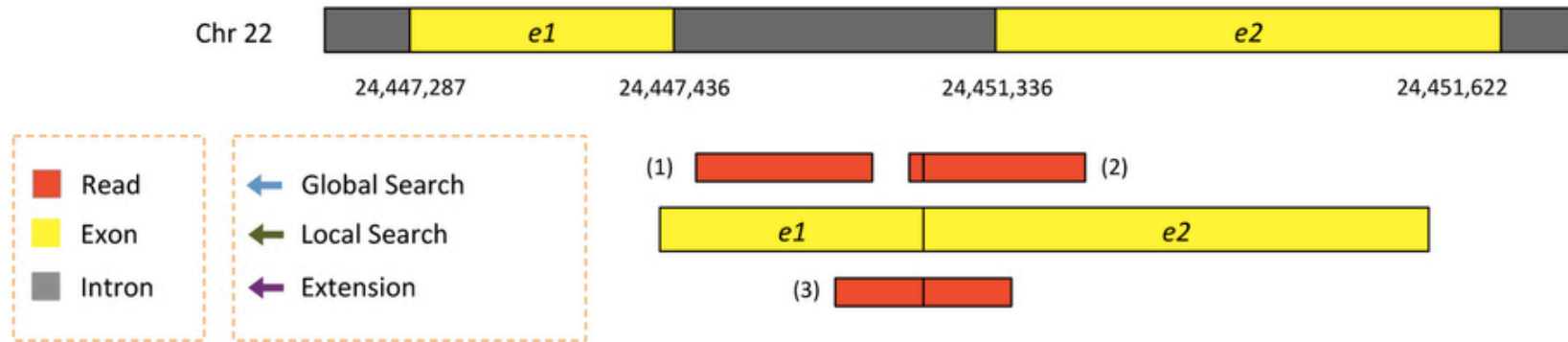
- HISAT is a 'splice-aware' RNA-seq read aligner
- Requires a reference genome
- Very fast
- Uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index
- Multiple types of indexes for alignment
 - a whole-genome FM index to anchor each alignment
 - numerous local 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

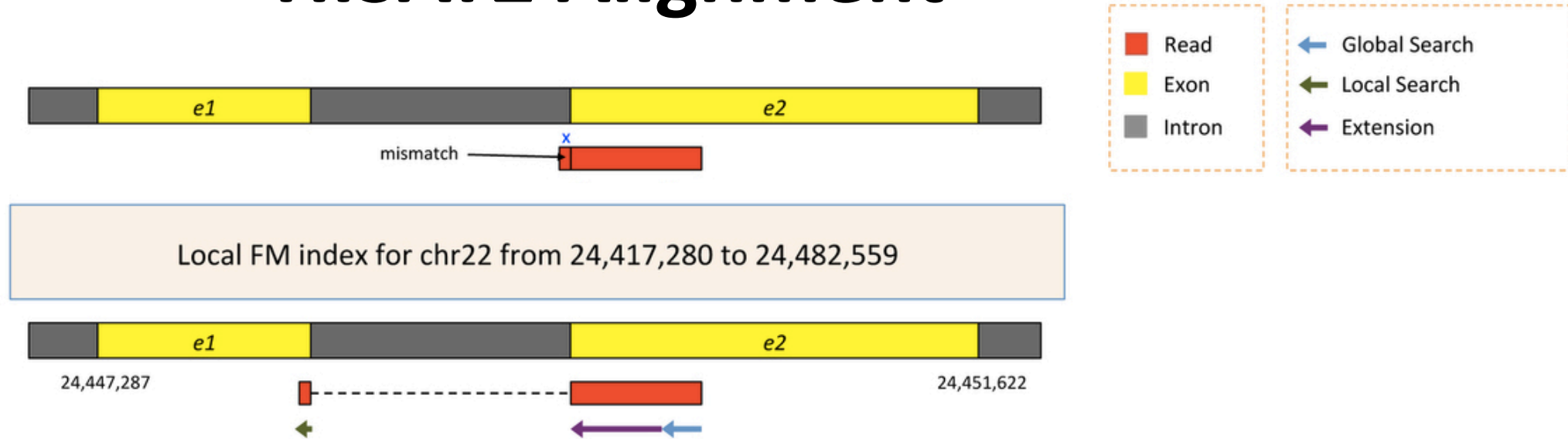
HISAT2 Alignment



- 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

Should I allow 'multi-mapped' reads?

- Depends on the application
- In *DNA* analysis it is common to use a mapper to randomly select alignments from a series of equally good alignments
- In *RNA* analysis this is less common
 - Perhaps disallow multi-mapped reads if you are variant calling
 - Definitely should allow multi-mapped reads for expression analysis with Cufflinks (and StringTie?)
 - Definitely should allow multi-mapped reads for gene fusion discovery

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/>