Introduction to Bioinformatics

RNA-Seq

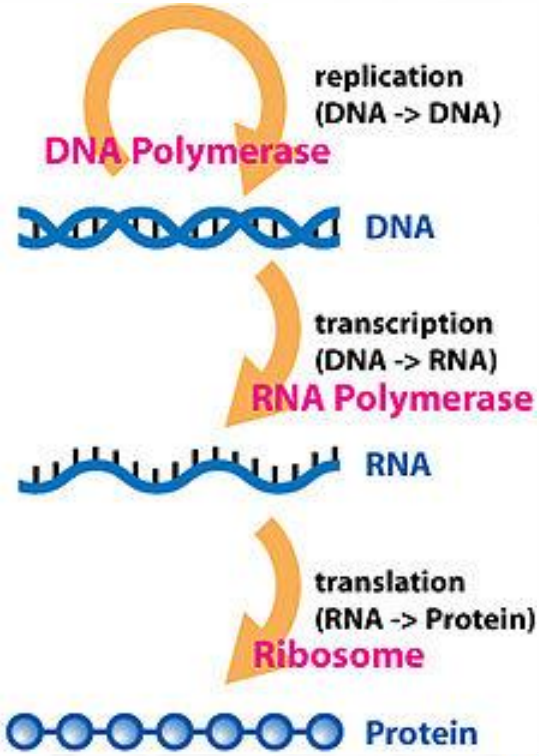
Stevan Radanović

Resources

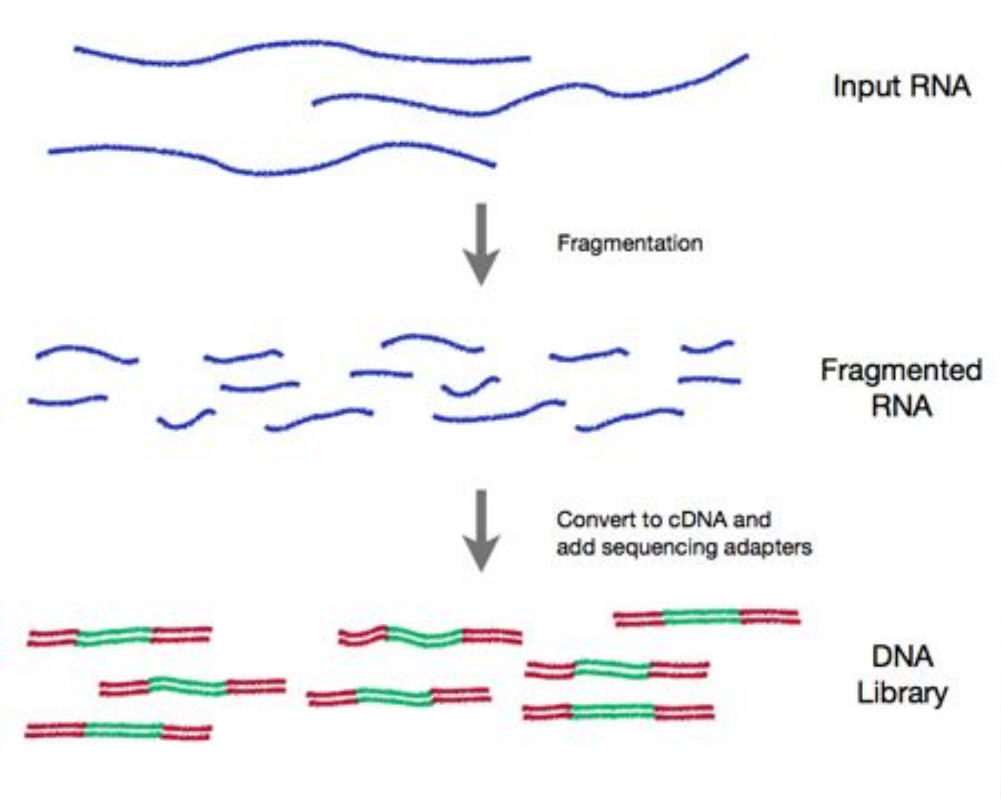
- RNA-seqlopedia, http://rnaseq.uoregon.edu/

Why do we do (m)RNA-Seq?

- Central dogma - DNA is used to make RNA, then RNA is used to make proteins, and proteins “run the show”
- If DNA = cookbook, and proteins = ready meals, than RNA = intermediate stages in this cooking process

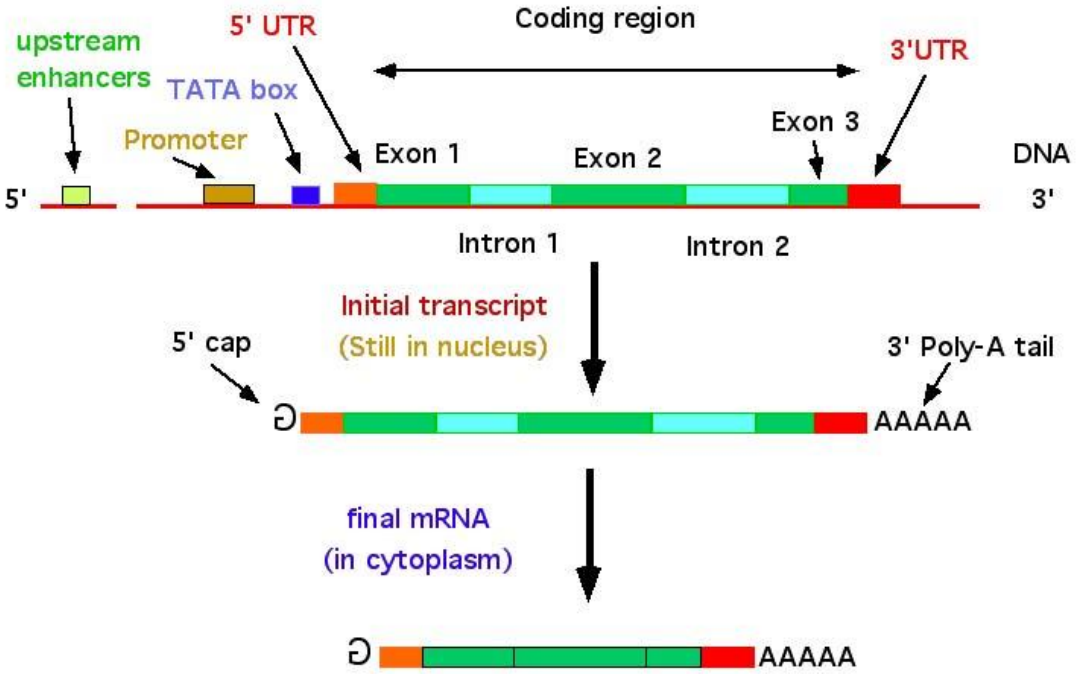


mRNA-Seq

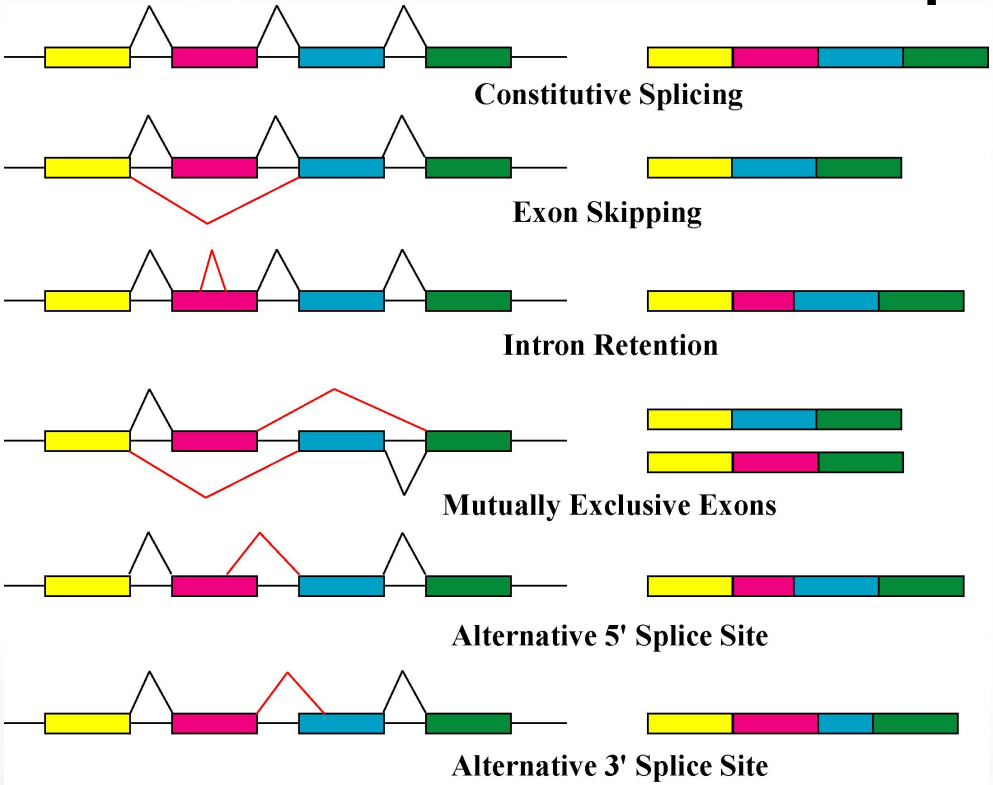


SevenBridges genomics

Complications - splicing (introns)



Complications - alternative splicing



SevenBridges genomics

RNA-Seq analyses

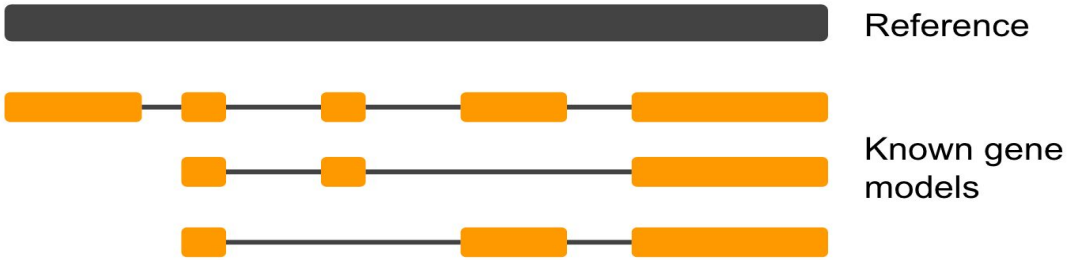
- Qualitative - identifying expressed transcripts, exon-intron boundaries, transcriptional start sites (TSS), poly-A sites
- Transcriptome annotation - GTF files

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..

RNA-Seq analyses

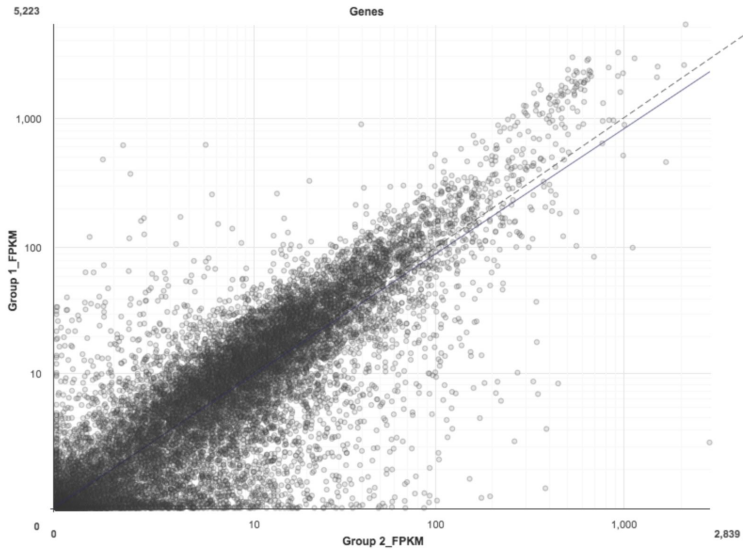
- Transcriptome annotation - GTF files

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..



RNA-Seq analyses

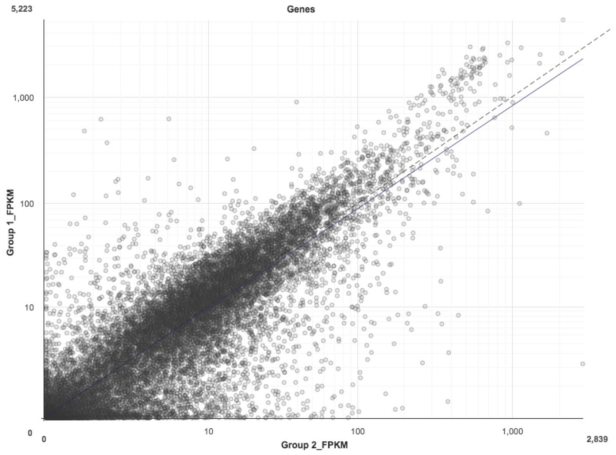
- Quantitative - measuring differences in expression, alternative splicing, TSS, poly-A between two or more treatments or groups

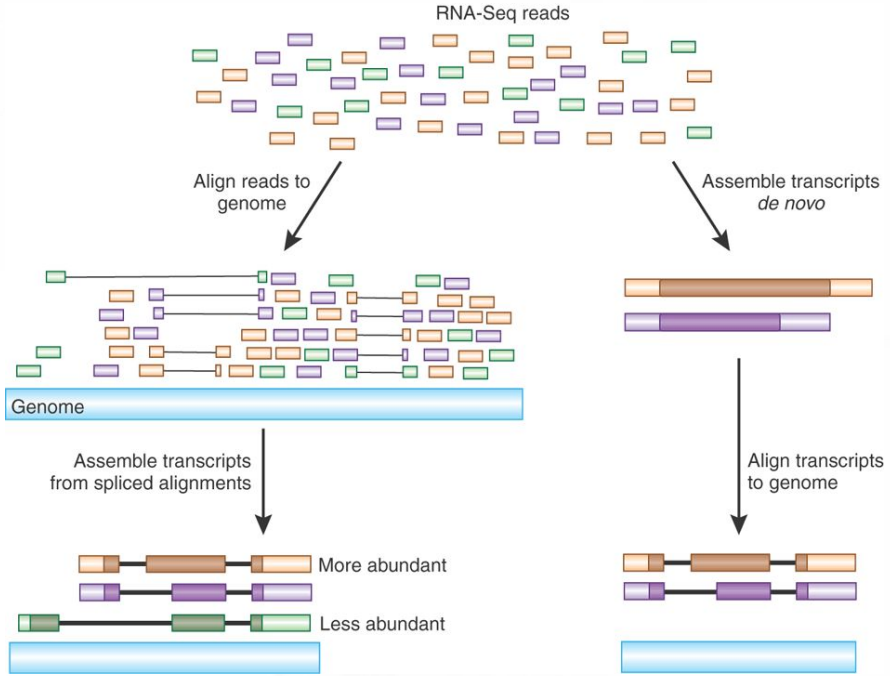


SevenBridges genomics

RNA-Seq data analysis pipeline

- Alignment
- Quantification
- Differential expression





SevenBridges genomics

RNA-Seq alignment - Bowtie

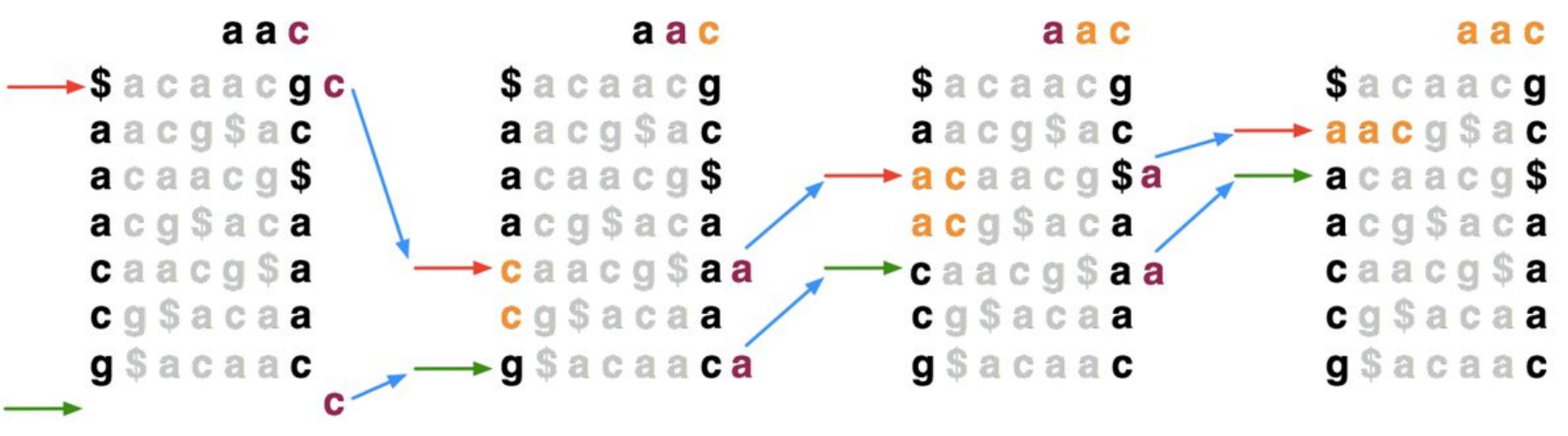
- Burrows-Wheeler Transformation - reversible lossless transformation algorithm which permutes an input string into a new string
- BWT string lends itself to an effective compression
- Rank preserving property - LF mapping

RNA-Seq alignment - Bowtie

- BWT is reversible
- Recreating T from BWT(T) - start in the first row and apply LF repeatedly, accumulating predecessors along the way

RNA-Seq alignment - Bowtie

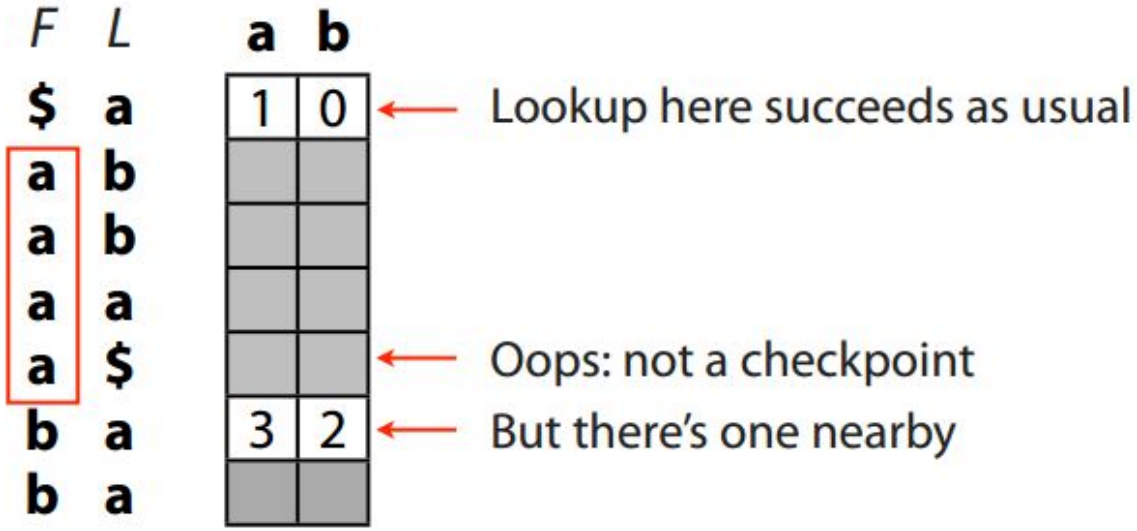
- Exact match, checkpoints, suffix array sample



SevenBridges genomics

RNA-Seq alignment - Bowtie

- Exact match, checkpoints, suffix array sample



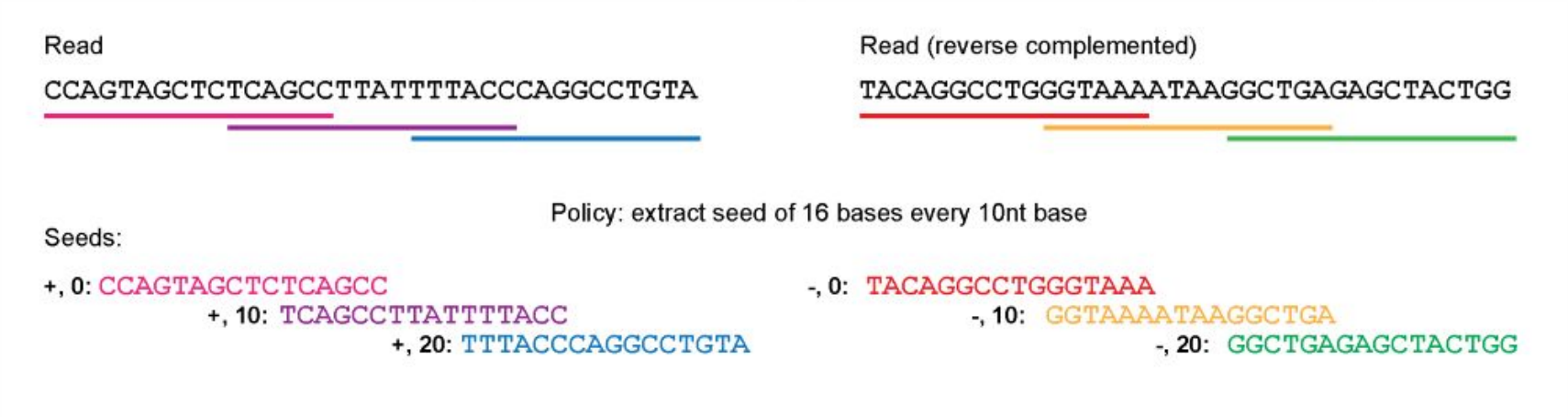
RNA-Seq alignment - Bowtie

- Exact match, checkpoints, suffix array sample

Bowtie marks every 32nd row by default (configurable)

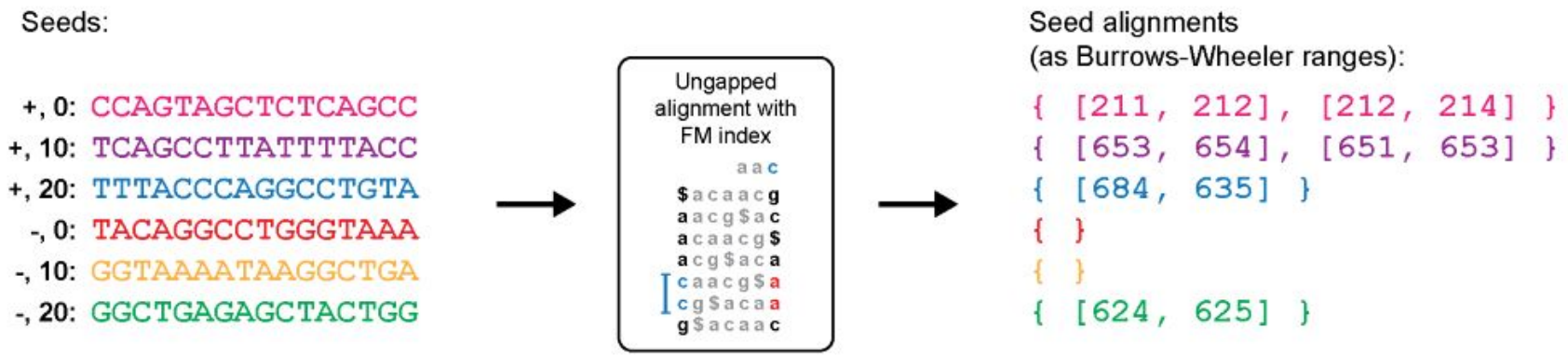
RNA-Seq alignment - Bowtie

- Alignment:
  - step 1: extracting seeds from the read and its complement



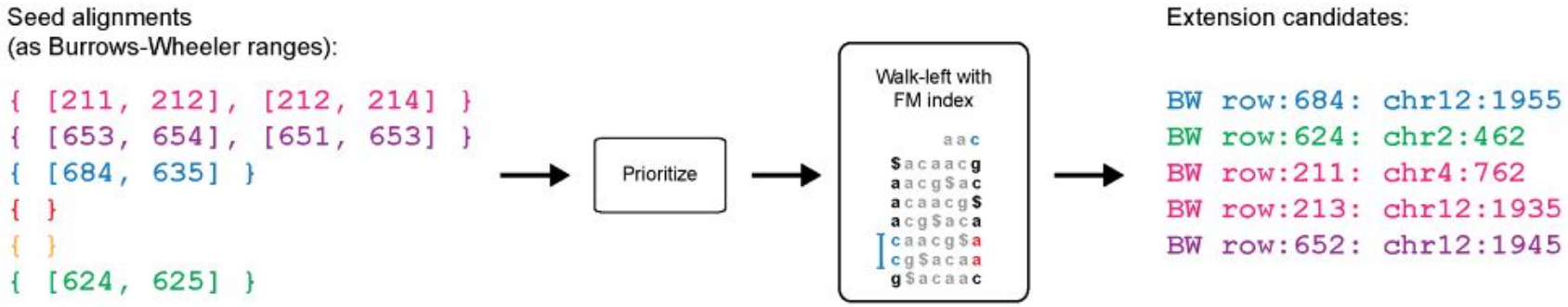
RNA-Seq alignment - Bowtie

- Alignment:
  - step 1: extracting seeds from the read and its complement
  - step 2: seed alignment using exact matching



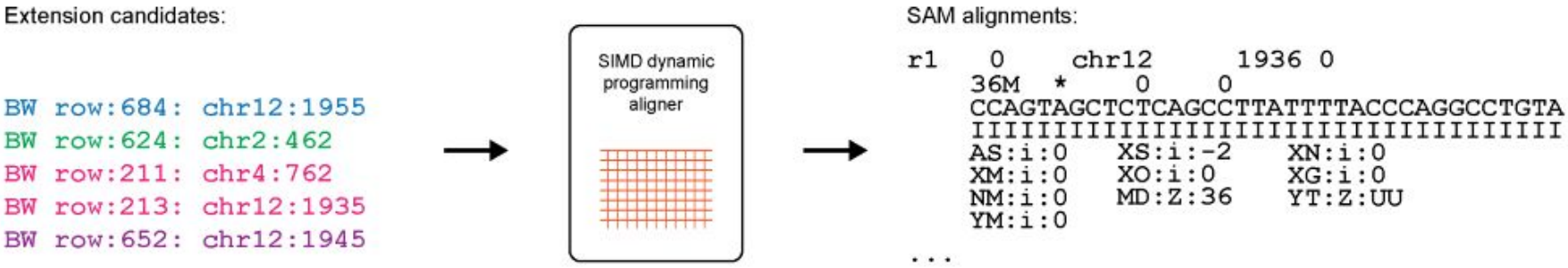
RNA-Seq alignment - Bowtie

- Alignment:
  - step 1: extracting seeds from the read and its complement
  - step 2: seed alignment using exact matching
  - step 3: prioritization and offset resolving



RNA-Seq alignment - Bowtie

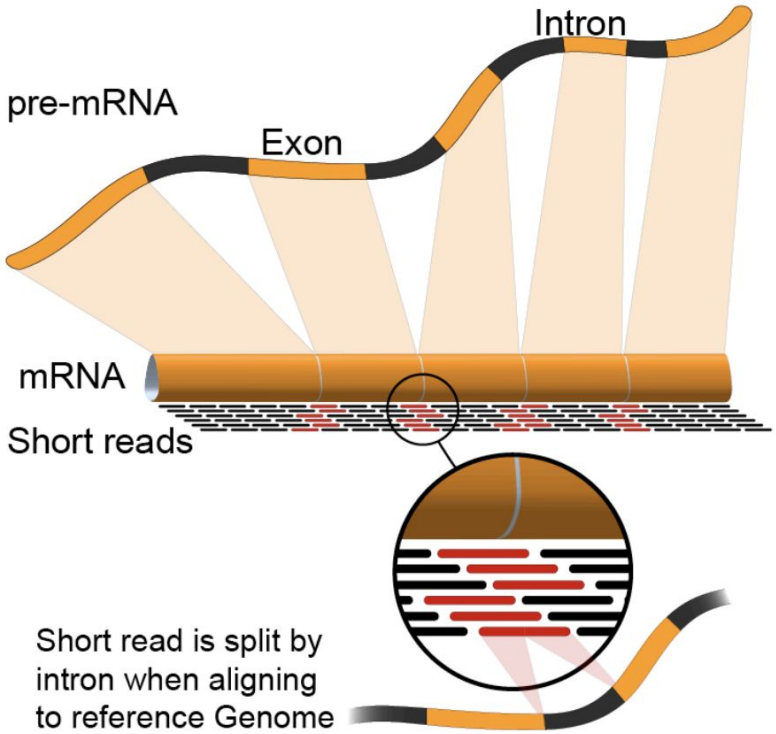
- Alignment:
  - step 1: extracting seeds from the read and its complement
  - step 2: seed alignment using exact matching
  - step 3: prioritization and offset resolving
  - step 4: extending (local alignment) using dynamic programming



SevenBridges genomics

RNA-Seq alignment - TopHat

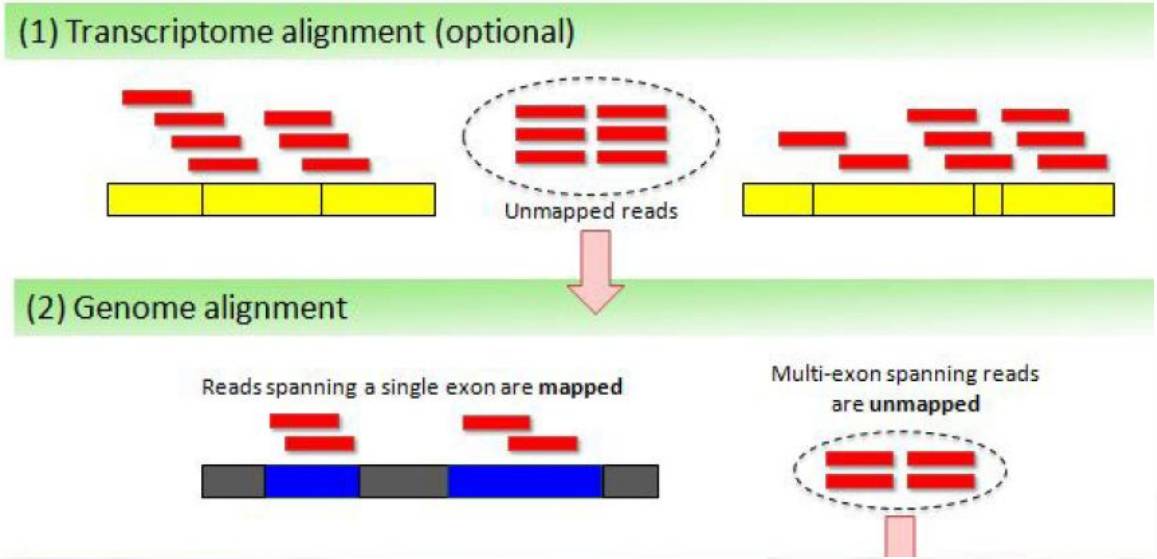
- Average length of mature mRNA transcript - 2,227 bp
- Average exon length - 235 bp
- Average number of exons per transcript: 9.5
- Assuming that 100 bp reads are uniformly distributed along a transcript we would expect ~ 35% of reads to span two or more exons



SevenBridges genomics

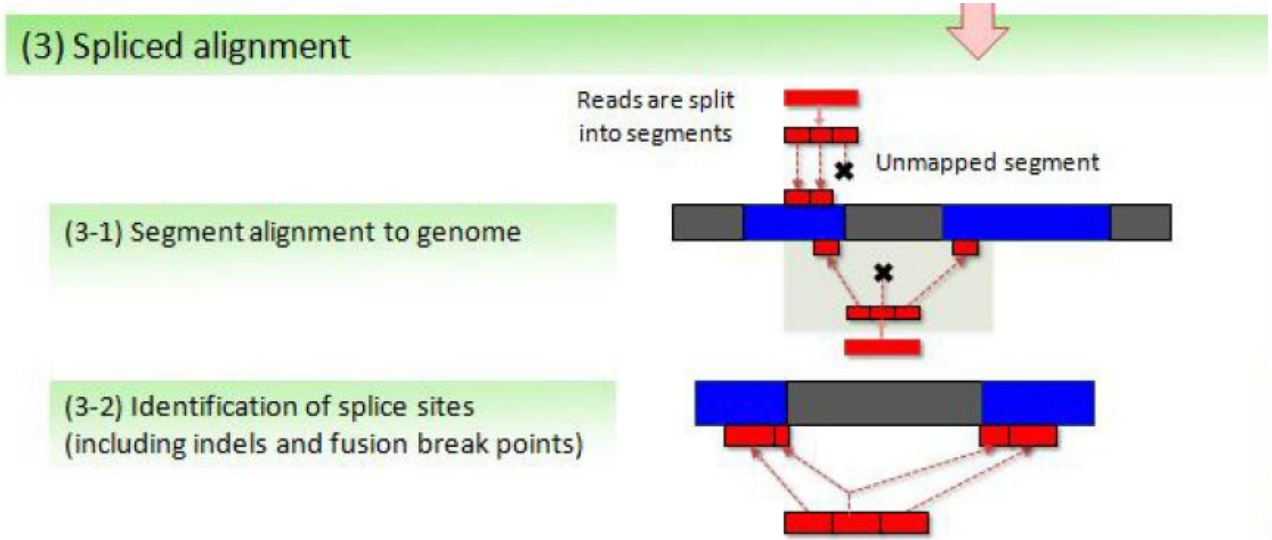
RNA-Seq alignment - TopHat

- Step 1: alignment to transcriptome, if annotation (GTF) provided
- Step 2: alignment to genome



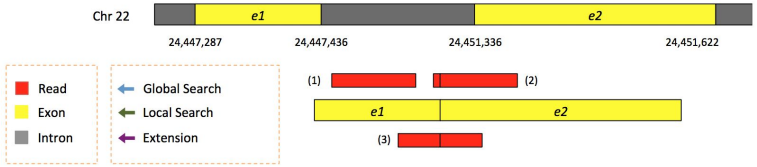
RNA-Seq alignment - TopHat

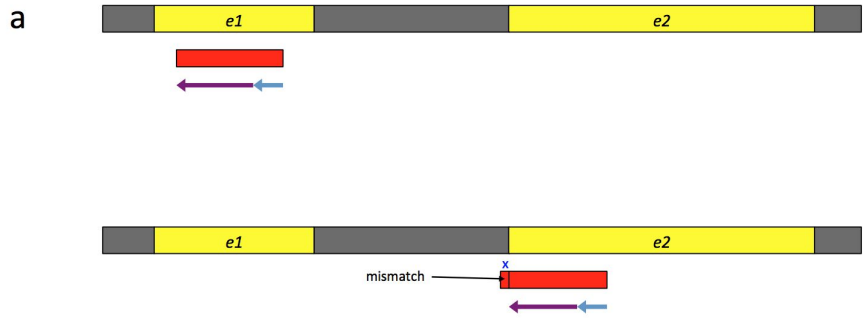
- Step 3: TopHat examines any case in which the left and right segments of the same read are mapped within user-defined maximum intron size

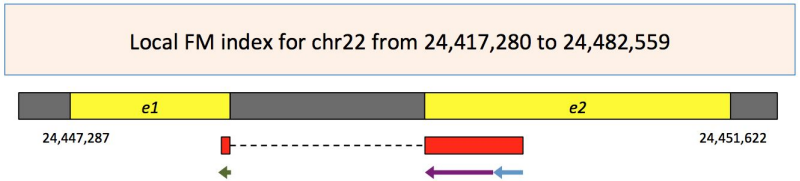


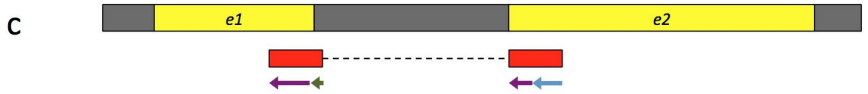
RNA-Seq alignment - HISAT

- Global FM-index, and
- Local FM-indices (~ 48k, 64k bp each, 1k bp overlap)
- Global vs local search





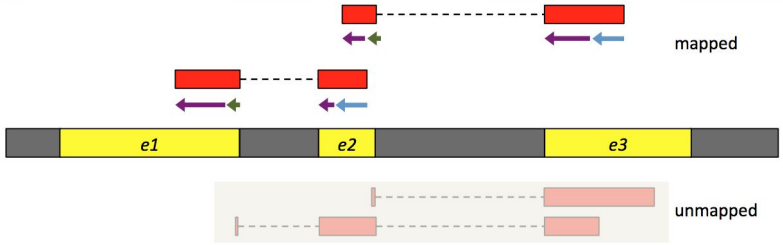




SevenBridges genomics

RNA-Seq alignment - HISAT

1st run of HISAT to discover splice sites



2nd run of HISAT to align reads by making use of the list of splice sites collected above