



Introduction to Bioinformatics

RNA-Seq

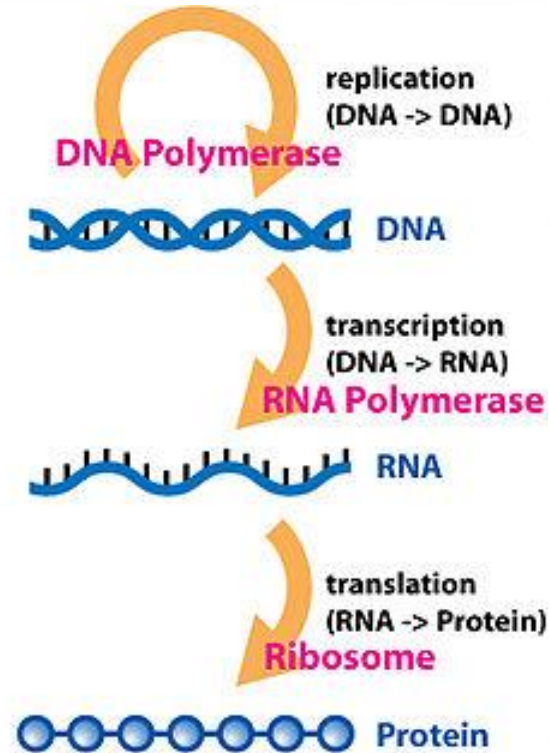
Stevan Radanović

Resources

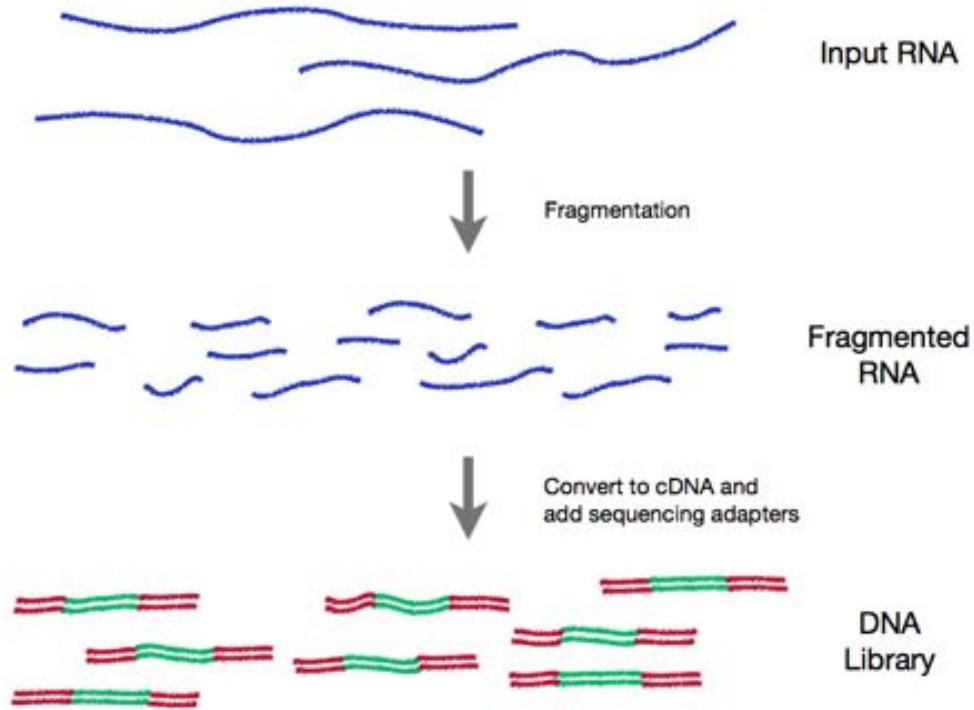
- RNA-seqlopedia, <http://rnaseq.uoregon.edu/>

Why do we do (m)RNA-Seq?

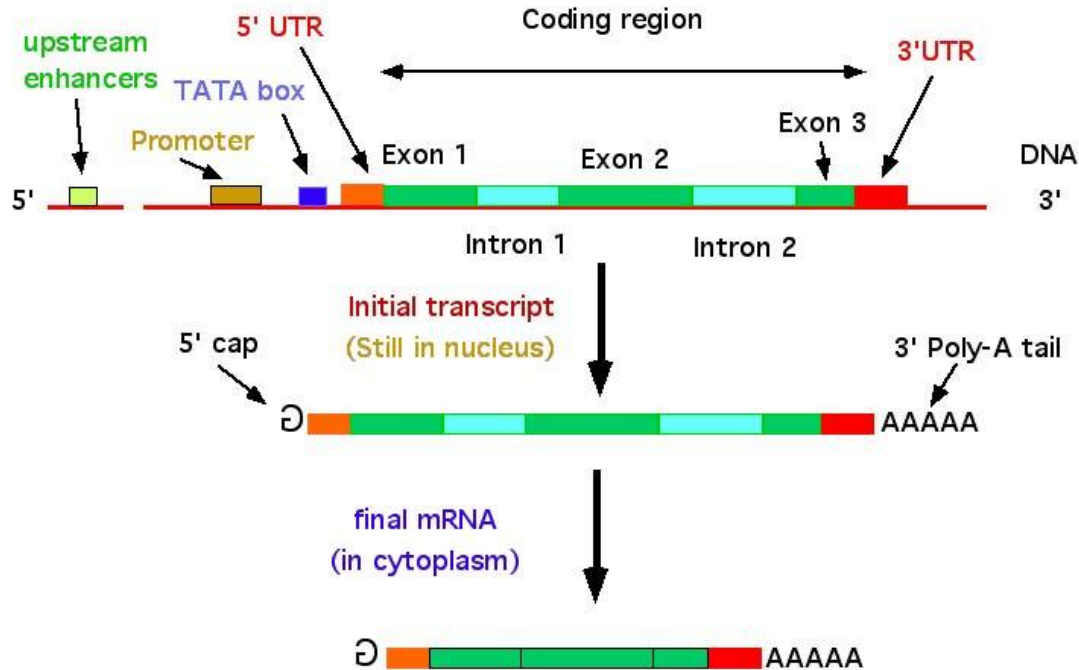
- Central dogma - DNA is used to make RNA, then RNA is used to make proteins, and proteins “run the show”
- If DNA = cookbook, and proteins = ready meals, then RNA = intermediate stages in this cooking process



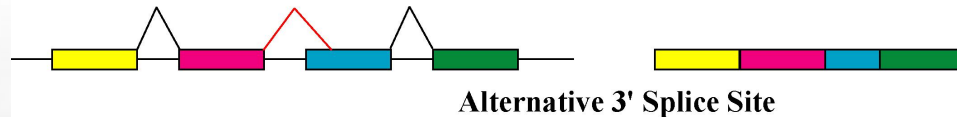
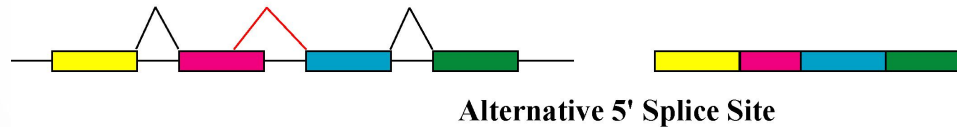
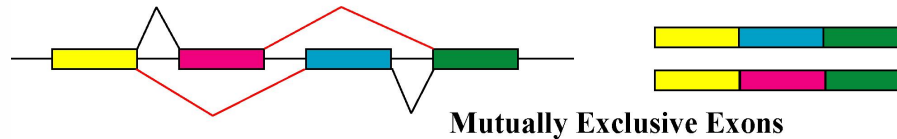
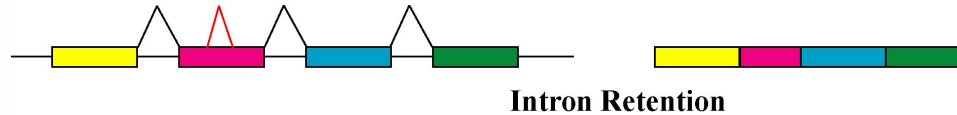
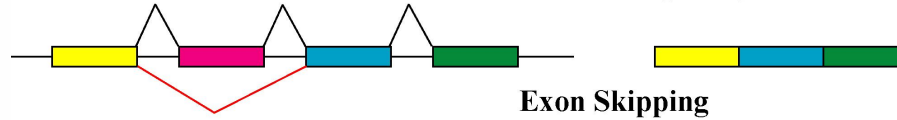
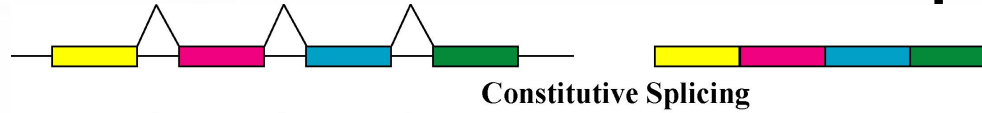
mRNA-Seq



Complications - splicing (introns)



Complications - alternative splicing



RNA-Seq analyses

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

<u>Col 1</u>	<u>Col 2</u>	<u>Col 3</u>	<u>Col 4</u>	<u>Col 5</u>	<u>Col 6</u>	<u>Col 7</u>	<u>Col 8</u>	<u>Col 9</u>
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

<u>Col 1</u>	<u>Col 2</u>	<u>Col 3</u>	<u>Col 4</u>	<u>Col 5</u>	<u>Col 6</u>	<u>Col 7</u>	<u>Col 8</u>	<u>Col 9</u>
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG000000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG000000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG000000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG000000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG000000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG000000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG000000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG000000169..



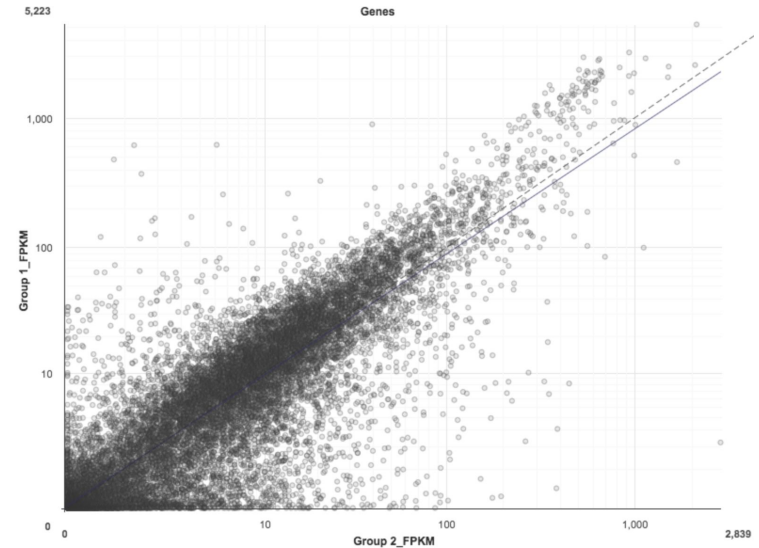
Reference



Known gene models

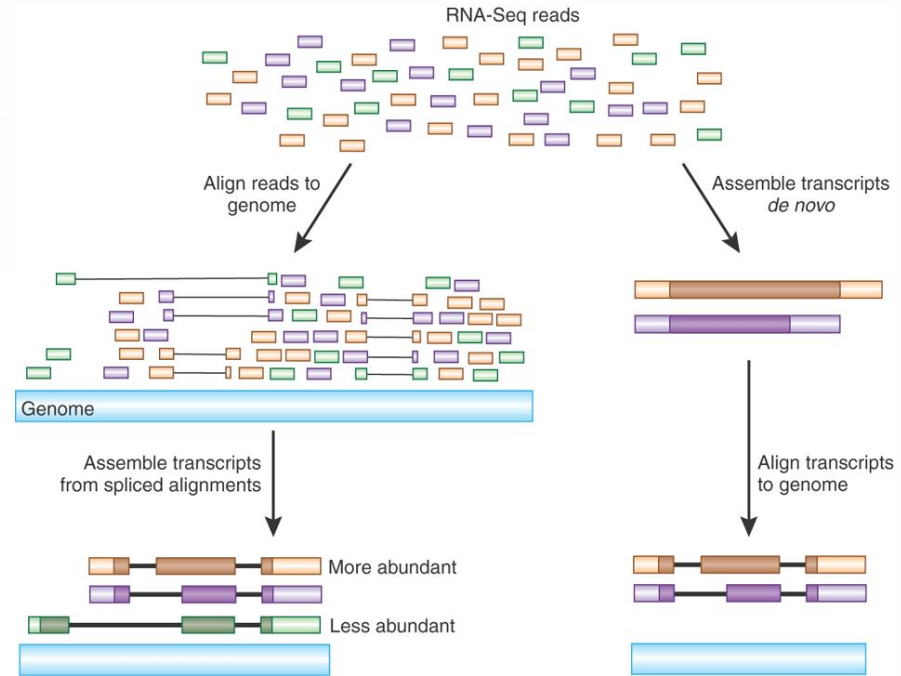
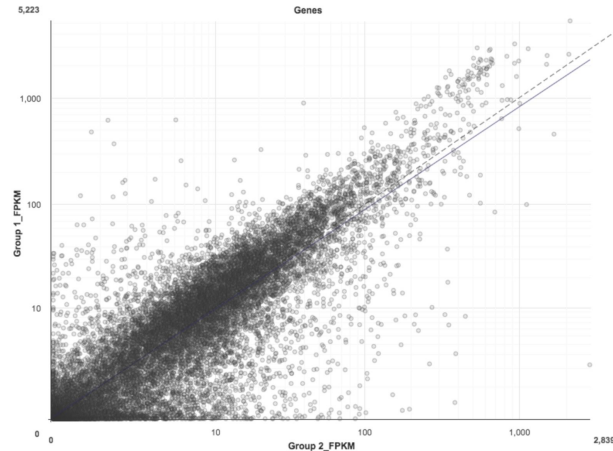
RNA-Seq analyses

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



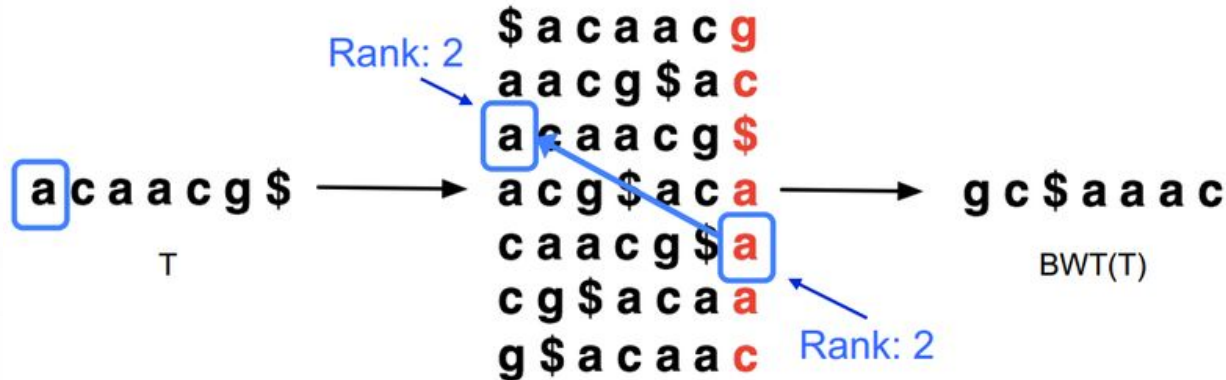
RNA-Seq data analysis pipeline

- Alignment
- Quantification
- Differential expression



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



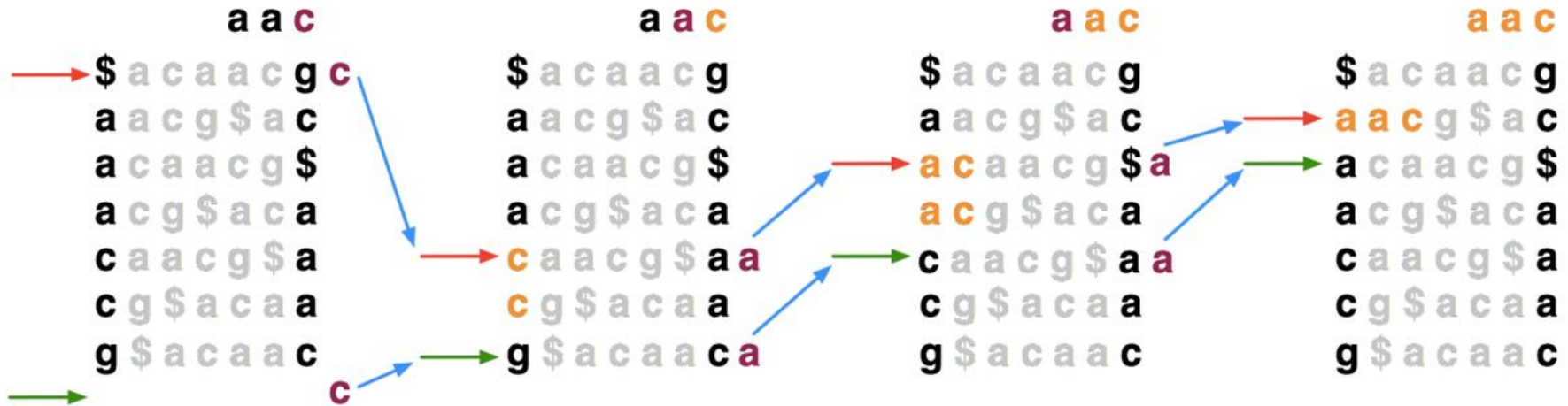
RNA-Seq alignment - Bowtie

- 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



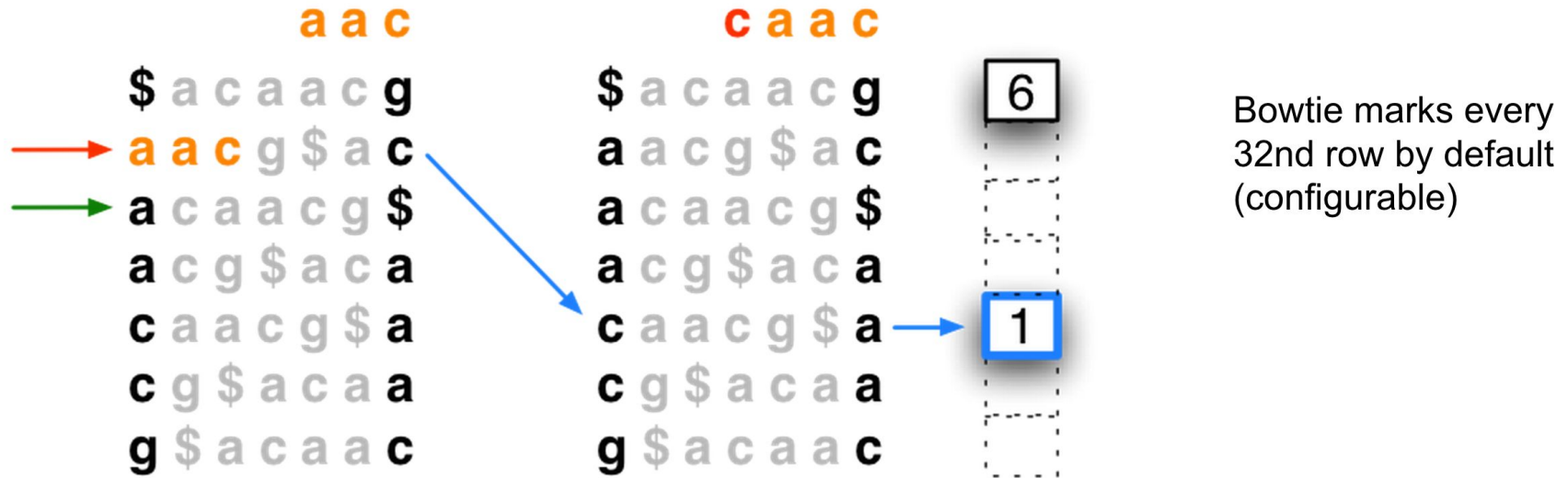
RNA-Seq alignment - Bowtie

- Exact match, checkpoints, suffix array sample

<i>F</i>	<i>L</i>	a	b	
\$	a	1	0	← Lookup here succeeds as usual
a	b			
a	b			
a	a			
a	\$			← Oops: not a checkpoint
b	a	3	2	← But there's one nearby
b	a			

RNA-Seq alignment - Bowtie

- Exact match, checkpoints, suffix array sample



RNA-Seq alignment - Bowtie

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

Read

CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA

Read (reverse complemented)

TACAGGCCTGGGTAAAATAAGGCTGAGAGCTACTGG

Policy: extract seed of 16 bases every 10nt base

Seeds:

+, 0: CCAGTAGCTCTCAGCC

+, 10: TCAGCCTTATTTTACC

+, 20: TTTACCCAGGCCTGTA

-, 0: TACAGGCCTGGGTAAA

-, 10: GGTAAAATAAGGCTGA

-, 20: GGCTGAGAGCTACTGG

RNA-Seq alignment - Bowtie

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

Seeds:

+, 0: CCAGTAGCTCTCAGCC
+, 10: TCAGCCTTATTTTACC
+, 20: TTTACCCAGGCCTGTA
-, 0: TACAGGCCTGGGTAAA
-, 10: GGTAAAATAAGGCTGA
-, 20: GGCTGAGAGCTACTGG



Ungapped
alignment with
FM index

```
      aac
$acaacg
aacg$ac
acaacg$
acg$aca
I caacg$a
cg$acaa
g$acaaac
```



Seed alignments

(as Burrows-Wheeler ranges):

```
{ [211, 212], [212, 214] }
{ [653, 654], [651, 653] }
{ [684, 635] }
{ }
{ }
{ [624, 625] }
```

RNA-Seq alignment - Bowtie

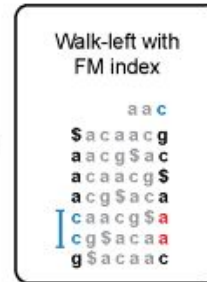
- Alignment:
 - step 1: extracting seeds from the read and its complement
 - step 2: seed alignment using exact matching
 - step 3: prioritization and offset resolving

Seed alignments
(as Burrows-Wheeler ranges):

```
{ [211, 212], [212, 214] }  
{ [653, 654], [651, 653] }  
{ [684, 635] }  
{ }  
{ }  
{ [624, 625] }
```



Prioritize



Extension candidates:

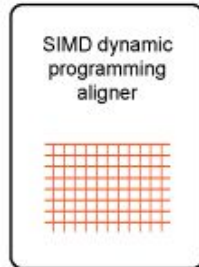
```
BW row:684: chr12:1955  
BW row:624: chr2:462  
BW row:211: chr4:762  
BW row:213: chr12:1935  
BW row:652: chr12:1945
```

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

Extension candidates:

```
BW row:684: chr12:1955
BW row:624: chr2:462
BW row:211: chr4:762
BW row:213: chr12:1935
BW row:652: chr12:1945
```

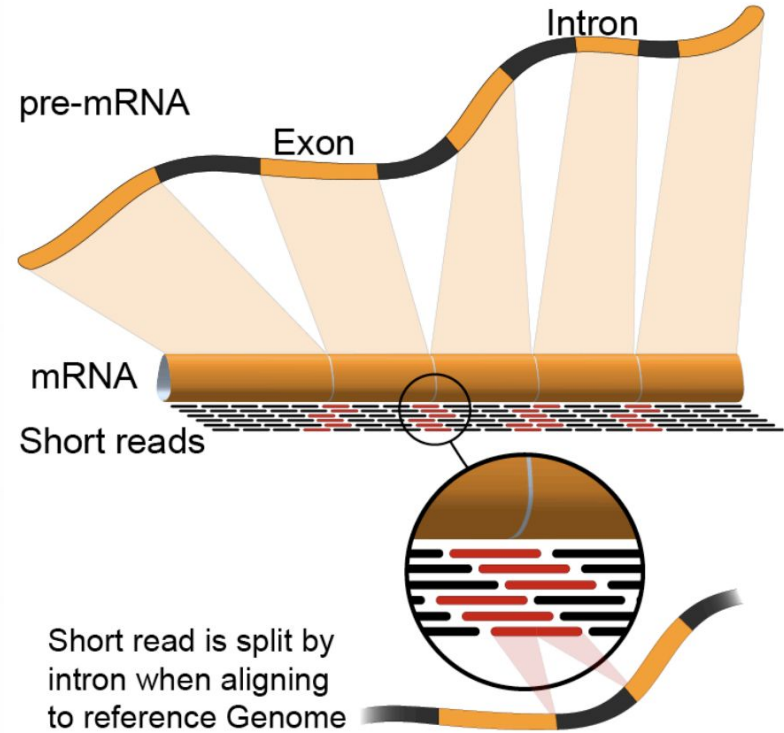


SAM alignments:

```
r1    0      chr12      1936 0
36M  *      0      0
CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
AS:i:0    XS:i:-2    XN:i:0
XM:i:0    XO:i:0    XG:i:0
NM:i:0    MD:Z:36    YT:Z:UU
YM:i:0
...
```

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



RNA-Seq alignment - TopHat

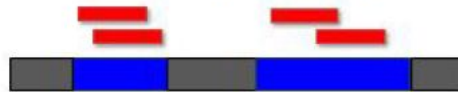
- Step 1: alignment to transcriptome, if annotation (GTF) provided
- Step 2: alignment to genome

(1) Transcriptome alignment (optional)



(2) Genome alignment

Reads spanning a single exon are **mapped**

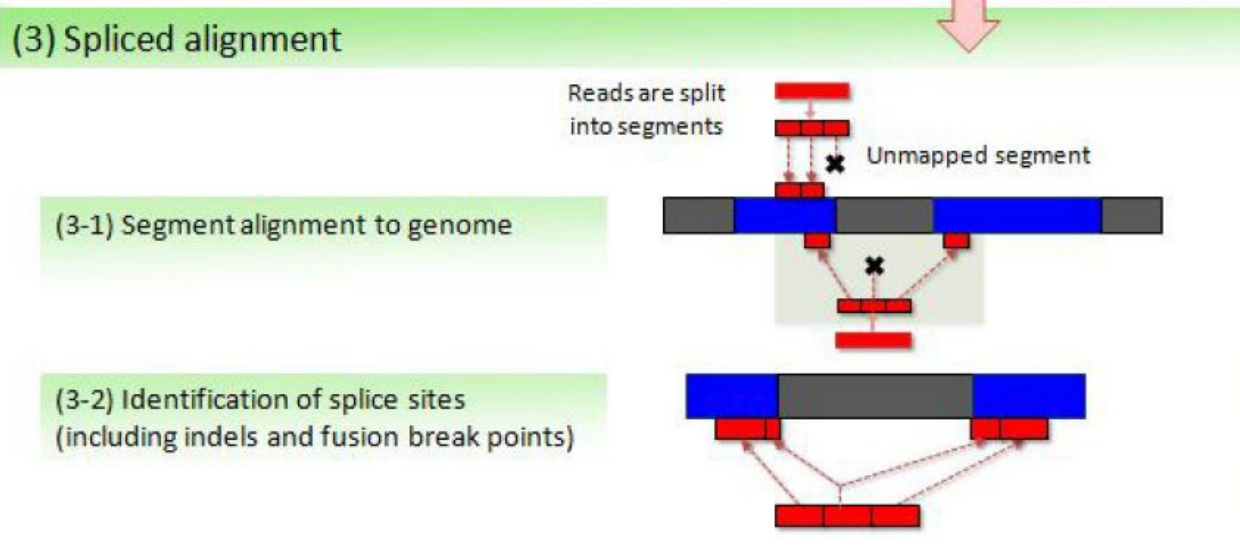


Multi-exon spanning reads are **unmapped**



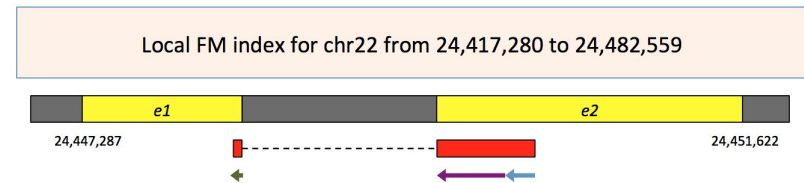
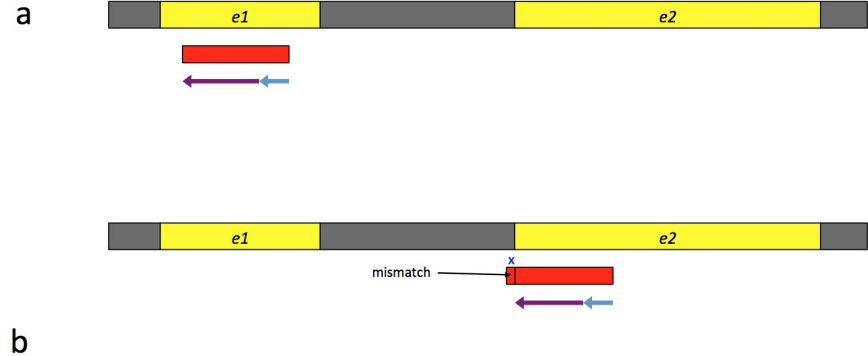
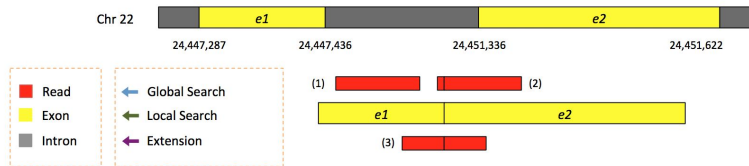
RNA-Seq alignment - TopHat

- 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



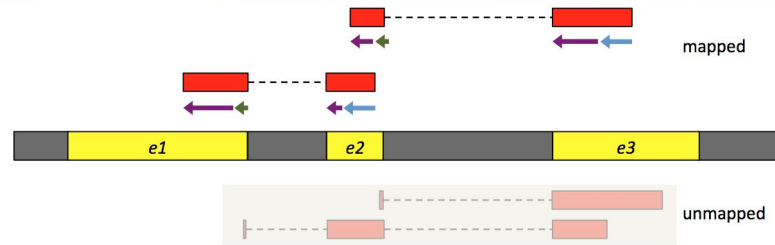
RNA-Seq alignment - HISAT

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



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

