BIOL 343
Applied Bioinformatics I

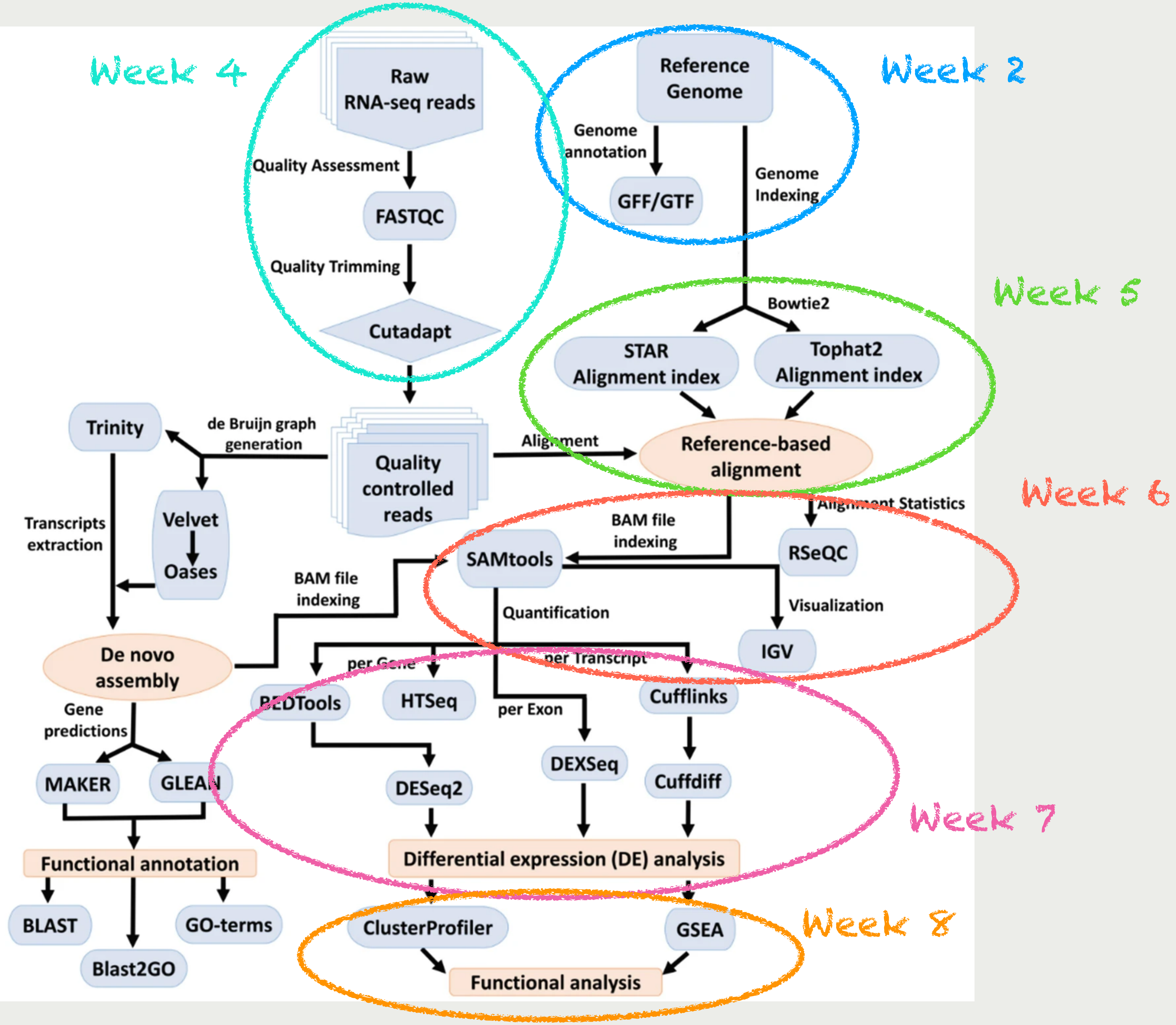
Alignment/Mapping

Dr. Nic Wheeler

Learning Objectives

You will be able to:

1.



Week 4

Raw RNA-seq reads

Quality Assessment

FASTQC

Quality Trimming

Cutadapt

Week 2

Reference Genome

Genome annotation

GFF/GTF

Genome Indexing

Week 5

Bowtie2

STAR Alignment index

Tophat2 Alignment index

Alignment

Reference-based alignment

Week 6

Alignment Statistics

RSeQC

BAM file indexing

SAMtools

Visualization

IGV

Quantification

Trinity

de Bruijn graph generation

Quality controlled reads

Velvet

Oases

Transcripts extraction

BAM file indexing

De novo assembly

per Gene

per Transcript

BEDTools

HTSeq

per Exon

Cufflinks

Week 7

DEXSeq

Cuffdiff

DESeq2

Differential expression (DE) analysis

Gene predictions

MAKER

GLEAN

Functional annotation

BLAST

GO-terms

Blast2GO

Week 8

ClusterProfiler

GSEA

Functional analysis

Bharti 2021, Next Generation Sequencing and Analysis

Alignment is the most important step in RNA-seq analysis

Counting (also important) and DEG ID relies on high-confidence mapping

- Recall the goal of our RNA-seq experiments…
  - Treatment vs Control
  - Mutant vs Wild type
  - ***Identify differentially expressed genes (DEGs)***
- DEGs will be identified using statistical tests comparing ***expression values*** of transcripts/genes
- Expression values will be calculated based on the number of reads that ***align/ map*** to a given genomic locus

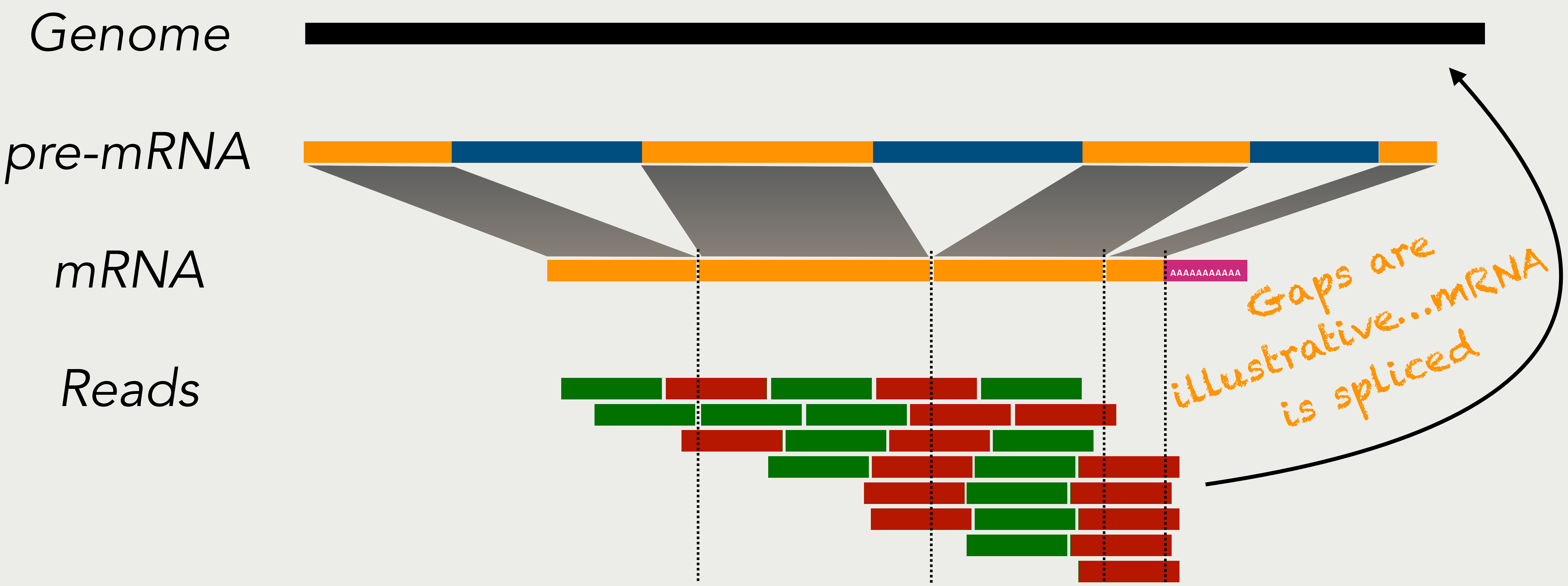
Types of alignment algorithms

Needleman-Wunsch and…

- Needleman-Wunsch (global alignment)
  - Dynamic programming
  - Mismatch penalty (transitions or transversions)
  - Gap penalty
- Problems - not global alignment, reference genomes are *huge* strings with lots of repetition, reads are likely to align many locations, and reads will align with massive gaps if spanning an intron
- Solution - Suffix array (STAR) *or* Burrow-Wheelers transform and FM-index (HISAT)

Splice-aware alignment

Gaps are large and encouraged

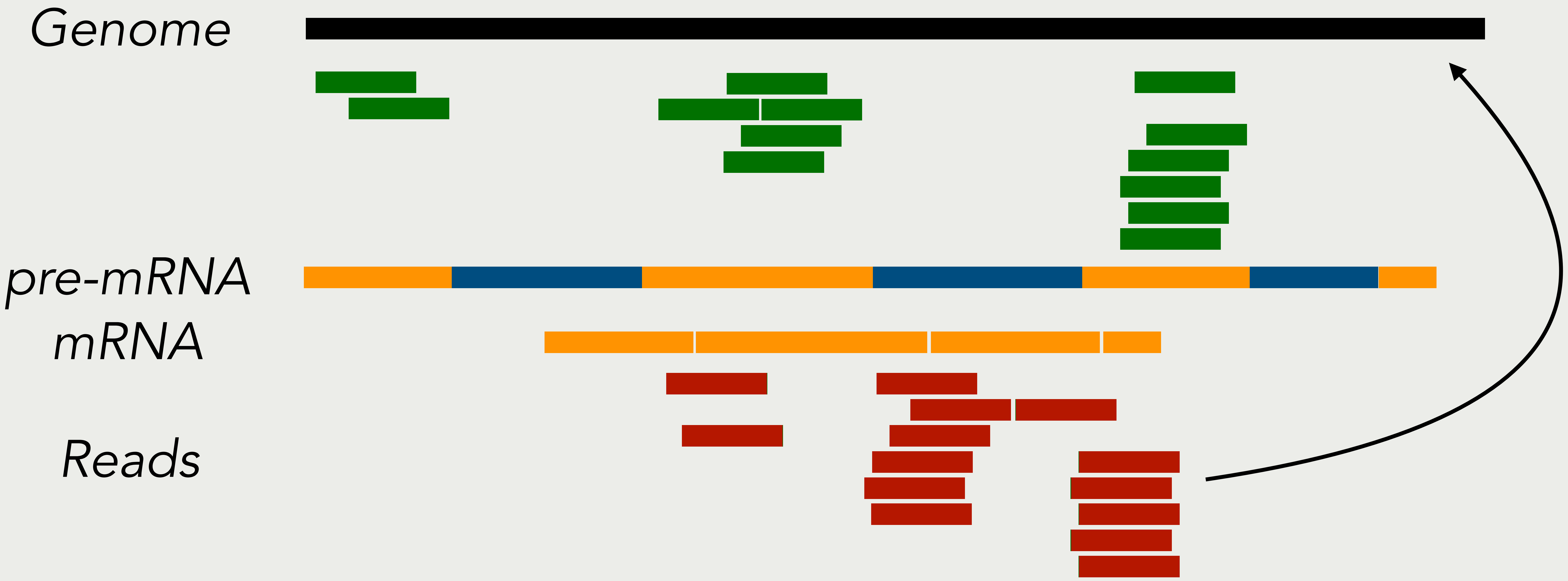


Gaps are illustrative...mRNA is spliced

But, reads aren’t aligned to a transcriptome (mRNAs), but a genome

Splice-aware alignment

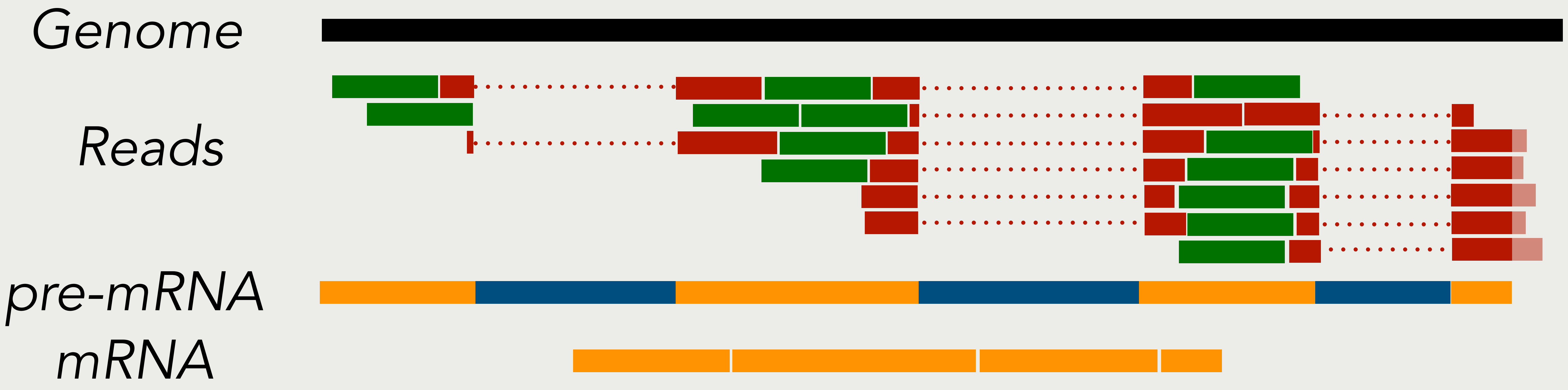
Gaps are large and encouraged



But, reads aren't aligned to a transcriptome (mRNAs), but a genome

Splice-aware alignment

Gaps are large and encouraged



Large gaps - representing introns - and reads from poly(A) tails don't align

Two (main) approaches to splice-aware alignment

STAR and HISAT

STAR

Spliced Transcripts Alignment to a Reference

Published in 2013

40574 citations

Requires a lot of RAM; ultra fast

HISAT

Hierarchical Indexing for Spliced Alignment of Transcripts

Published in 2015

17667 citations

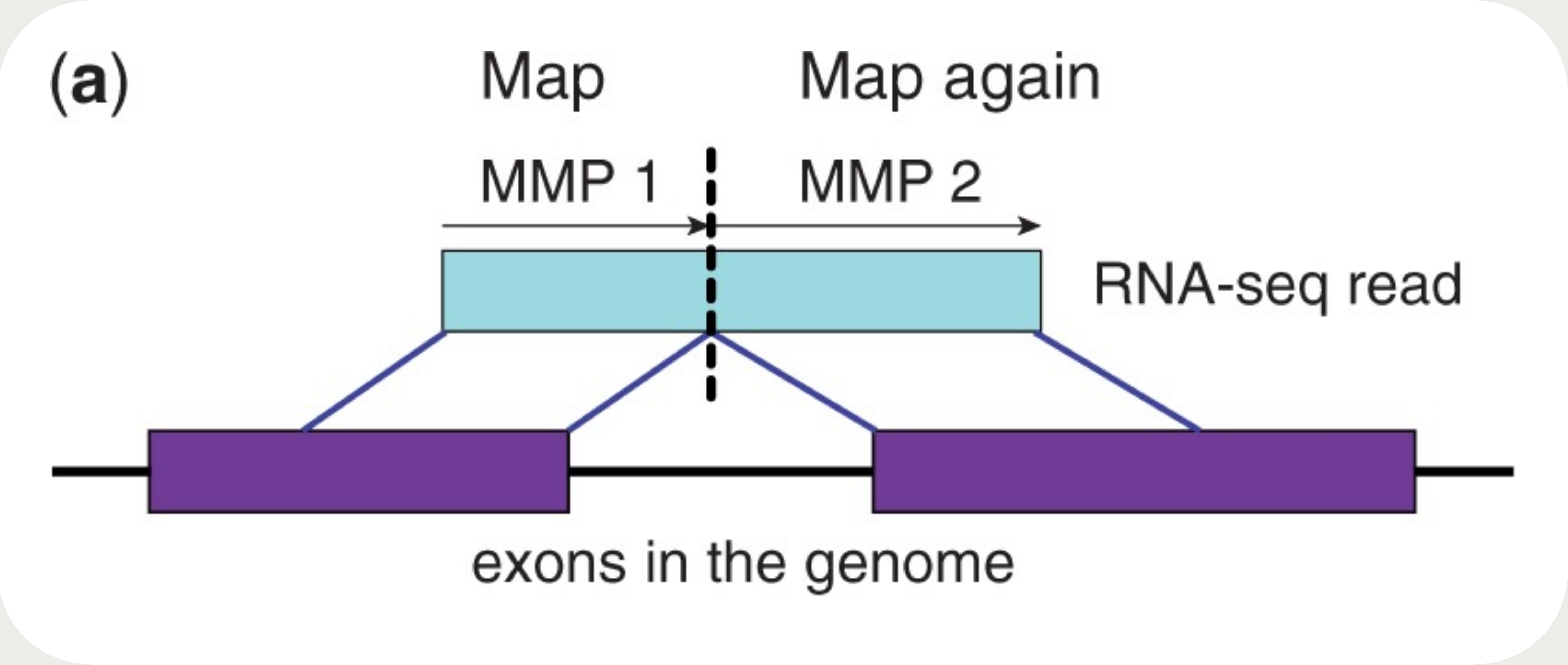
Less RAM needed; still fast

Two (main) approaches to splice-aware alignment

STAR

1. Find the Maximal Mappable Prefix of the read
    - MMP 1 will map to a splice donor
2. Find the MMP of the remainder of the read
    - MMP 2 will map to a splice acceptor

- Uses a suffix array of the reference genome

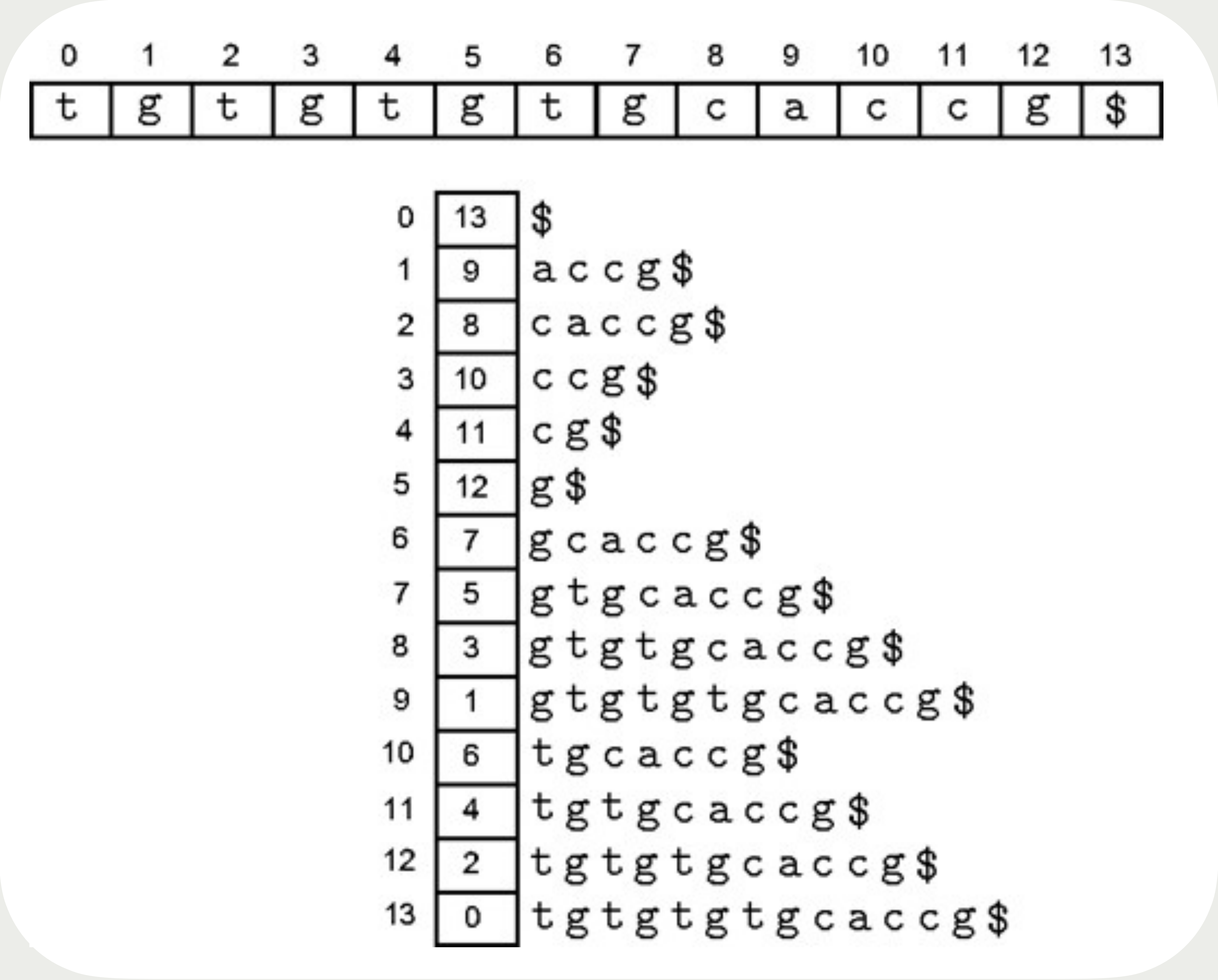


Dobin 2013, Bioinformatics

Two (main) approaches to splice-aware alignment

STAR

- Uses a suffix array of the reference genome
  - Every substring of the genome sorted lexicographically
  - Given a search string *P*, two binary searches to find the boundaries
    - *gtg* - binary search to find boundary 1 at index 5, binary search to find boundary 2 at index 9
- Many developments (ongoing) in 1) generating the SA and 2) searching the SA



0	1	2	3	4	5	6	7	8	9	10	11	12	13
t	g	t	g	t	g	t	g	c	a	c	c	g	$

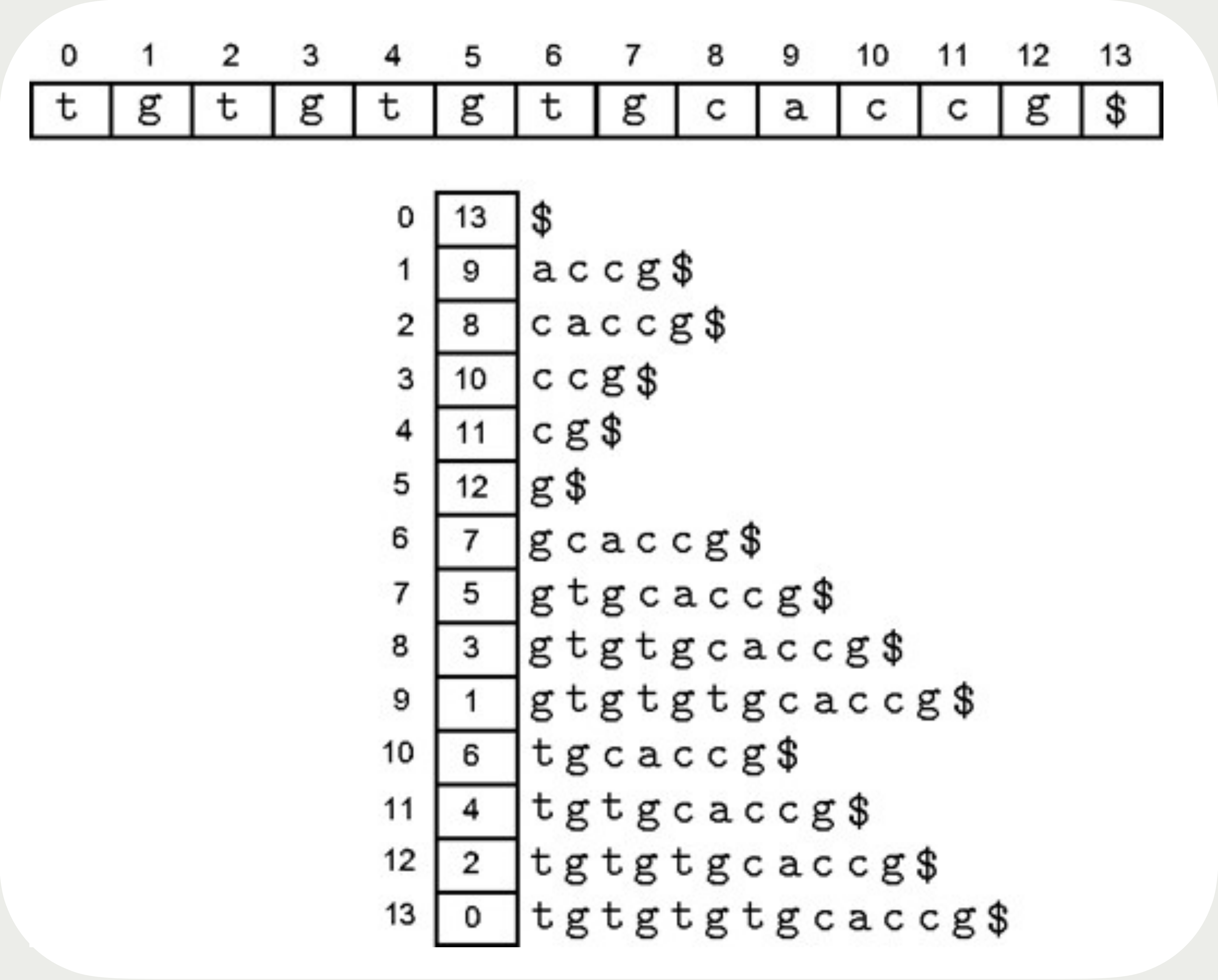
0	13	$
1	9	accg$
2	8	caccg$
3	10	ccg$
4	11	cg$
5	12	g$
6	7	gcaccg$
7	5	gtgcaccg$
8	3	gtgtgcaccg$
9	1	gtgtgtgcaccg$
10	6	tgcaccg$
11	4	tgtgcaccg$
12	2	tgtgtgcaccg$
13	0	tgtgtgtgcaccg$

Shrestha 2014, Briefings in Bioinformatics

Two (main) approaches to splice-aware alignment

STAR

- Only two binary searches - ultrafast!
- Suffix array of a large genome - very big RAM!
- Generating suffix array - kinda slow!



0	1	2	3	4	5	6	7	8	9	10	11	12	13
t	g	t	g	t	g	t	g	c	a	c	c	g	$

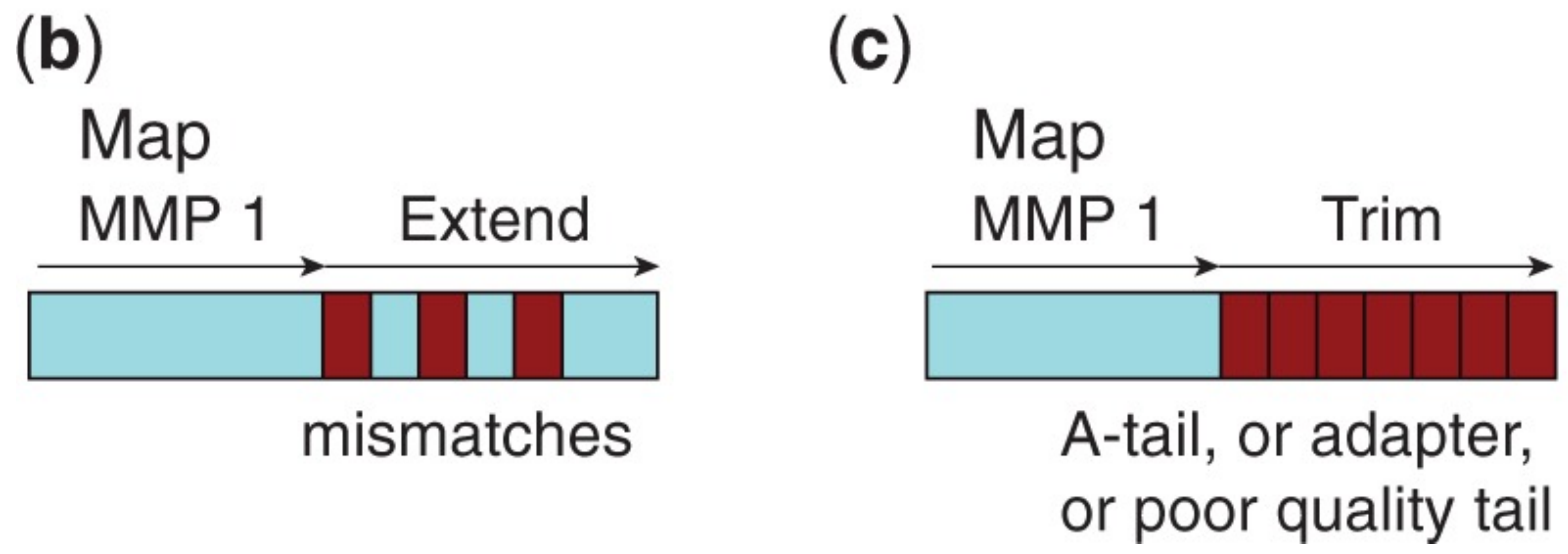
0	13	$
1	9	accg$
2	8	caccg$
3	10	ccg$
4	11	cg$
5	12	g$
6	7	gcaccg$
7	5	gtgcaccg$
8	3	gtgtgcaccg$
9	1	gtgtgtgcaccg$
10	6	tgcaccg$
11	4	tgtgcaccg$
12	2	tgtgtgcaccg$
13	0	tgtgtgtgcaccg$

Shrestha 2014, Briefings in Bioinformatics

Two (main) approaches to splice-aware alignment

STAR

- Other advantages to this approach:
  - Robust to mismatches - MMPs can be extended
  - Can trim (***soft clip***) if extension of MMP results in many mismatches



Dobin 2013, Bioinformatics