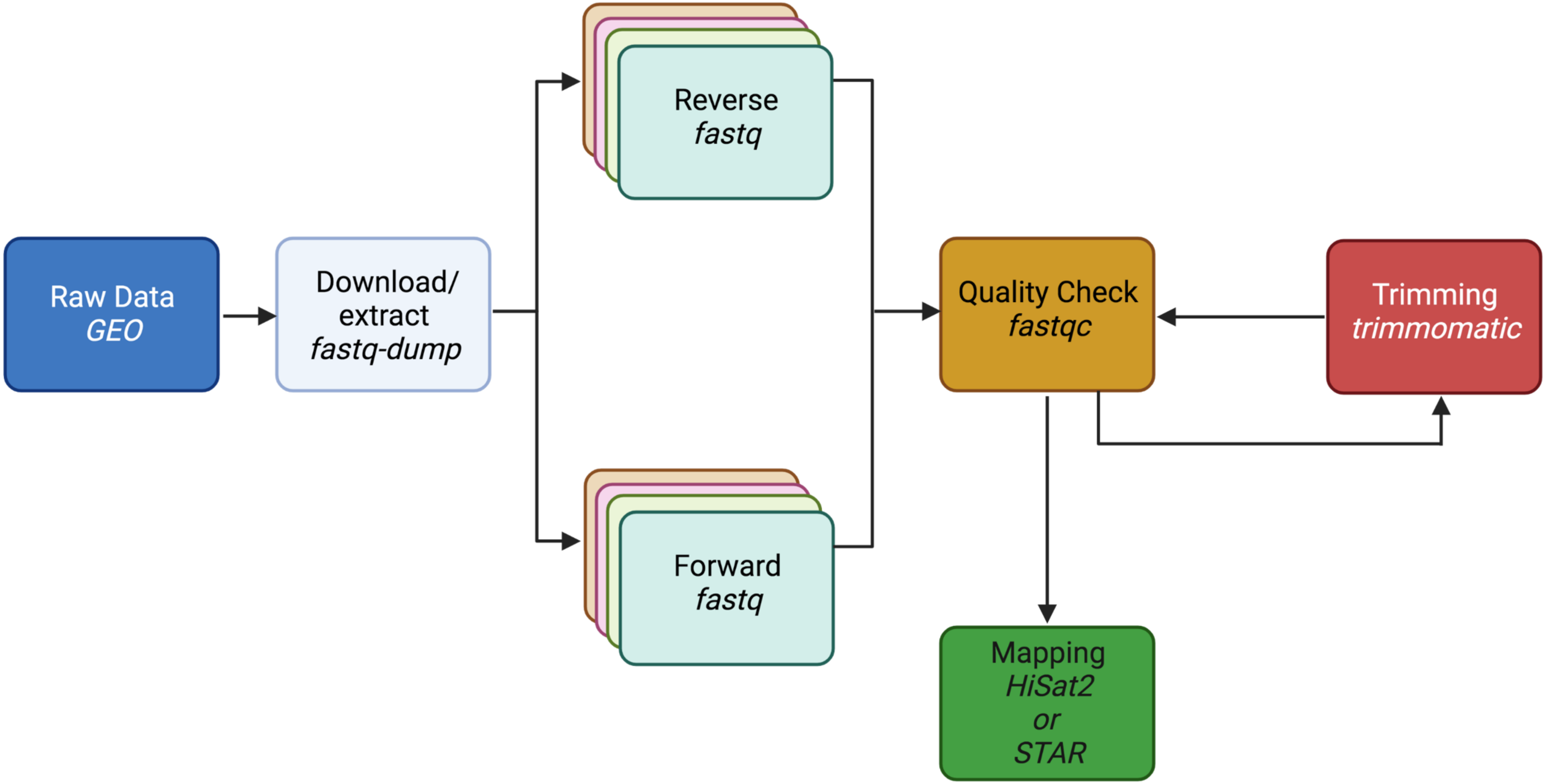


Raw Data
GEO

Download/
extract
fastq-dump

Reverse
fastq

Forward
fastq

Quality Check
fastqc

Trimming
trimmomatic

Mapping
HiSat2
or
STAR

Read Mapping
February 18, 2026

Learning objectives

- ❖ Describe common file formats encountered during read alignment (FASTQ, SAM, BAM, BED, GTF)
- ❖ Identify major challenges in read alignment and evaluate strategies to address them
- ❖Explain why genome indexing is required and outline computational steps involved
- ❖ Analyze the key features of the splice—aware aligner STAR & HISAT2

Outline

- Class Activity #1 = HISAT2_example1 = 10 minutes
  - *This script will take a minimum of 20 minutes to complete*

- Lecture for ~20 mins

- Class Activity #2 = indexed_genomes_example = 10 minutes

- Lecture for ~5 minutes

- Class Activity #3 = HISAT2_example2 = 20 minutes

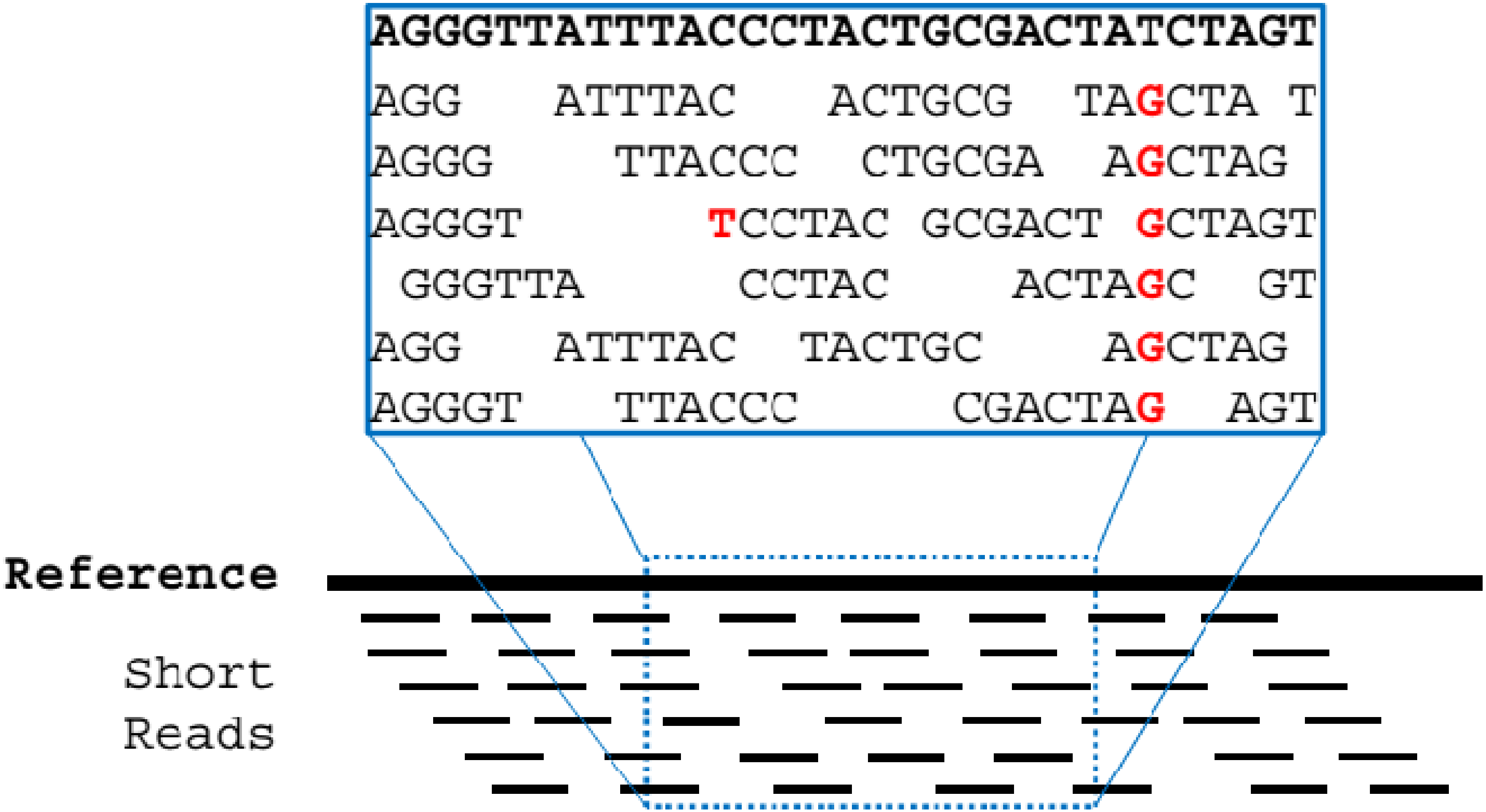
Class activity
Script Submission

HISAT2_example1

If I hand you 40 million short RNA reads (150bp), how would you figure out where they came from in the genome?

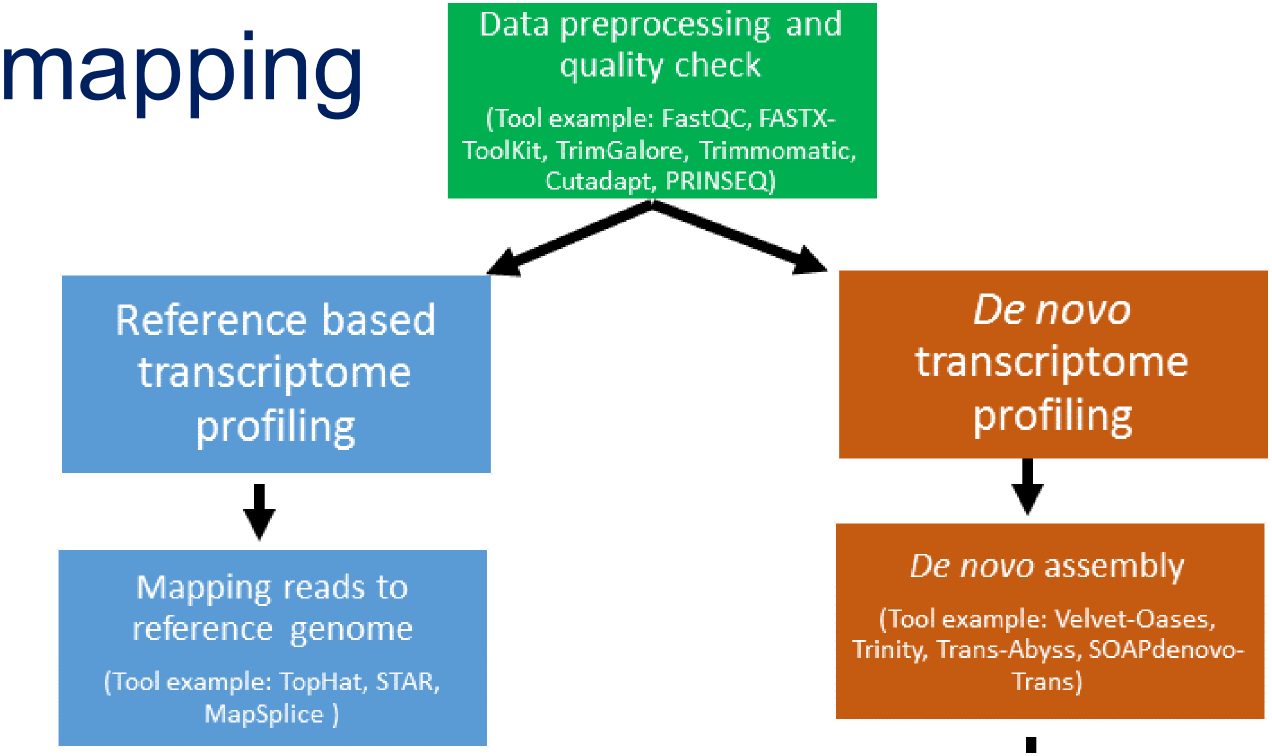
Alignment

Read alignment / “mapping”


AGGGTTATTTACCCTACTGCGACTATCTAGT
AGG ATTTAC ACTGCG TAGCTA T
AGGG TTACCC CTGCGA AGCTAG
AGGGT TCCTAC GCGACT GCTAGT
GGGTTA CCTAC ACTAGC GT
AGG ATTTAC TACTGC AGCTAG
AGGGT TTACCC CGACTAG AGT
Reference
Short
Reads

we are identifying the genomic origin of the sequenced cDNA fragment

RNA-Seq mapping strategies



Data preprocessing and quality check

(Tool example: FastQC, FASTX-ToolKit, TrimGalore, Trimmomatic, Cutadapt, PRINSEQ)

Reference based transcriptome profiling

Mapping reads to reference genome

(Tool example: TopHat, STAR, MapSplice)

De novo transcriptome profiling

De novo assembly

(Tool example: Velvet-Oases, Trinity, Trans-Abyss, SOAPdenovo-Trans)

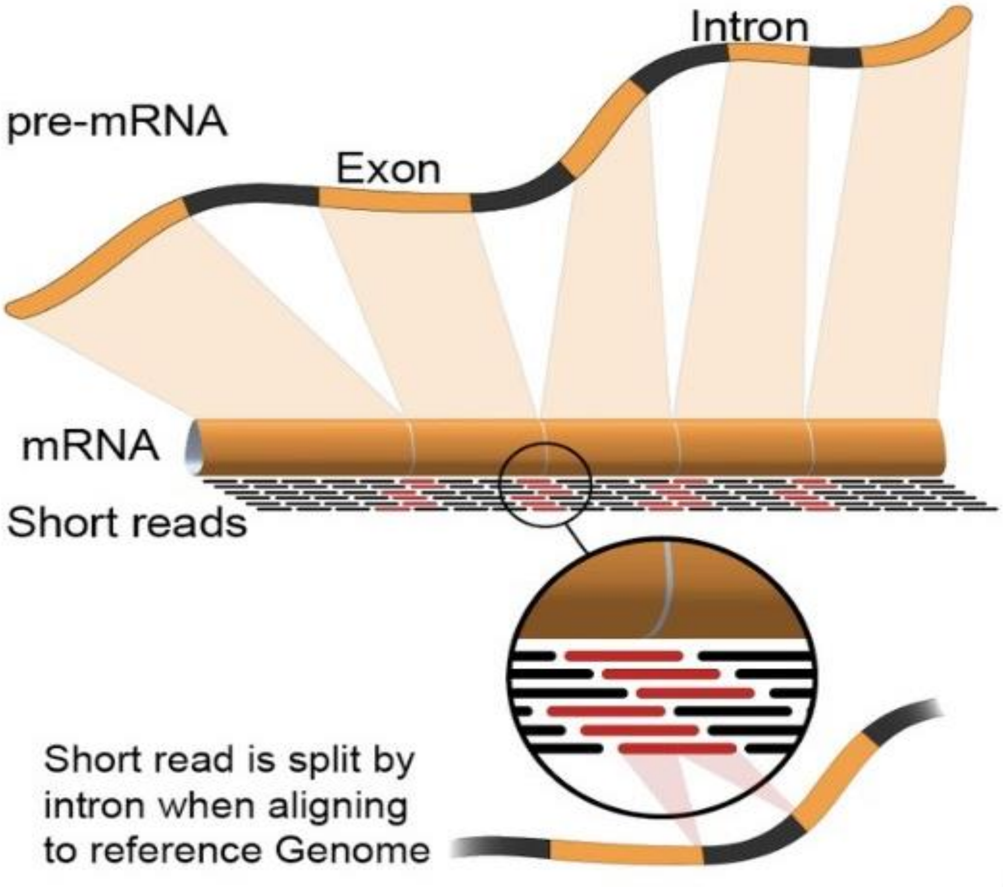
Reference based *	**De novo**
Reference is set of transcripts or genomic DNA that contains introns and exons	No reference genome exists
	Poor genome annotations

Challenges in Read Alignment

1. Intron/Exon Boundaries
2. Genome vs Transcriptome
3. Computational Expense
4. *Sometimes you need to align using multiple methods….hopefully by the end of today's lecture you will understand why*

1

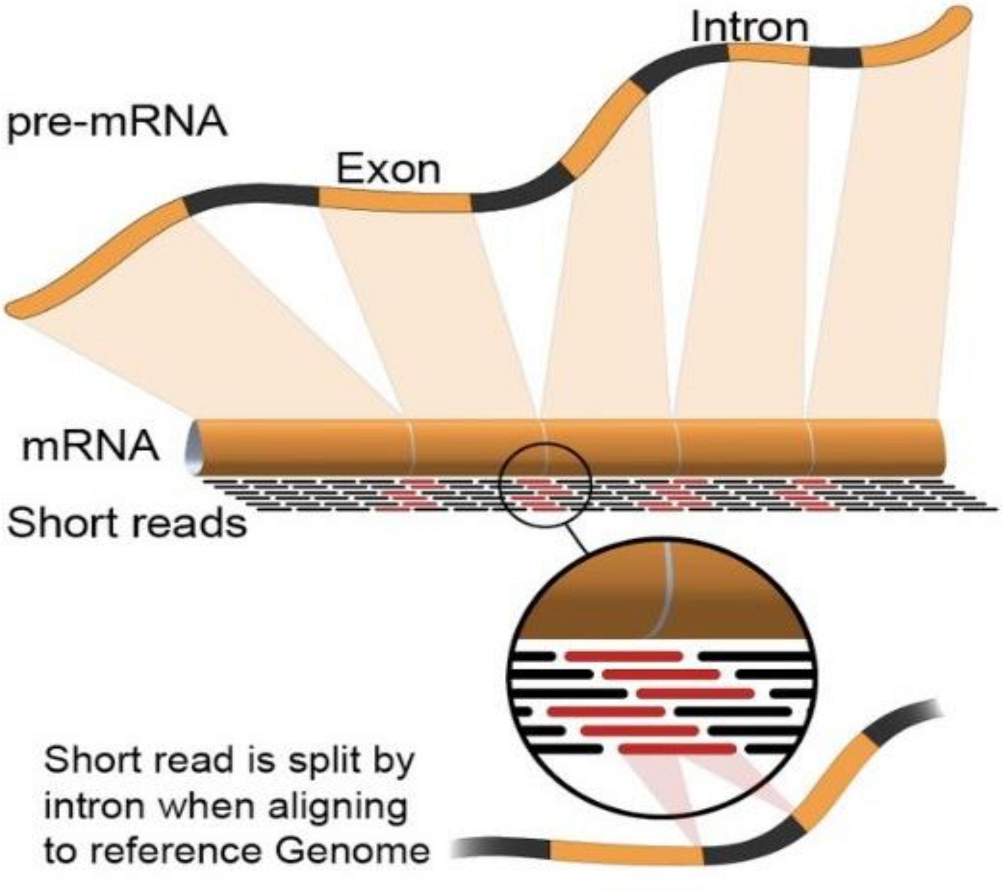
RNASeq Mapping Challenges: Intron/Exon Boundaries


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

Introns
Exons

Genome is contiguous
RNA is not

1 RNASeq Mapping Challenges: Intron/Exon Boundaries

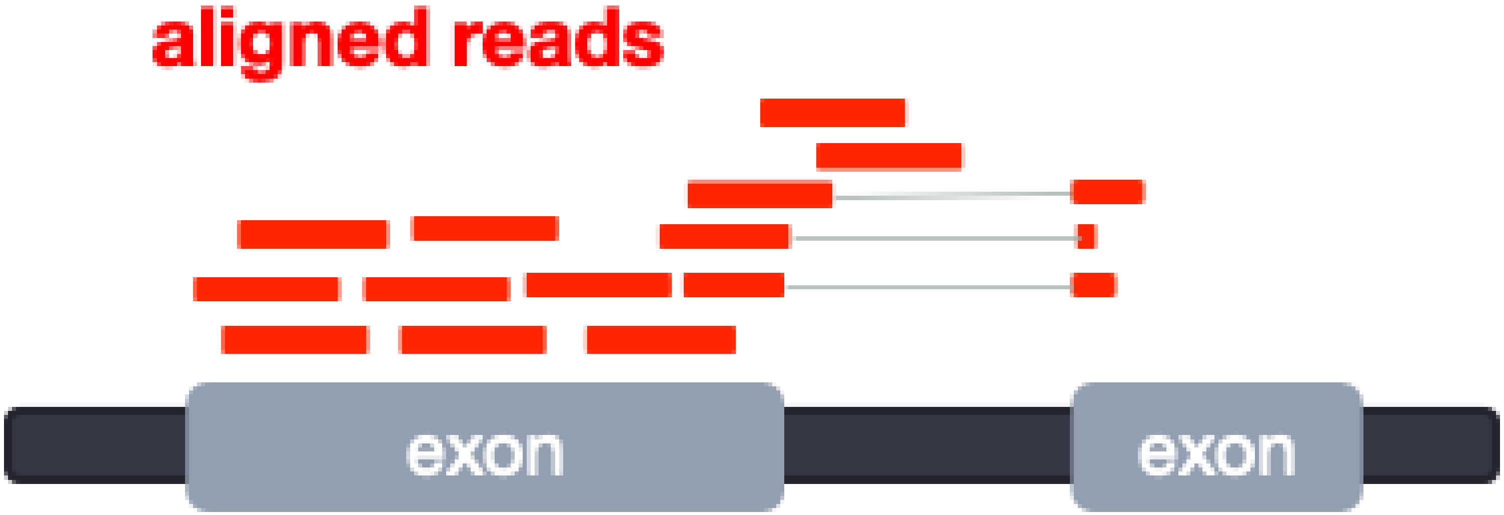

Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Introns
Exons

We have to account for reads that may be split by potentially thousands of bases of intronic sequences

Two categories of reads:

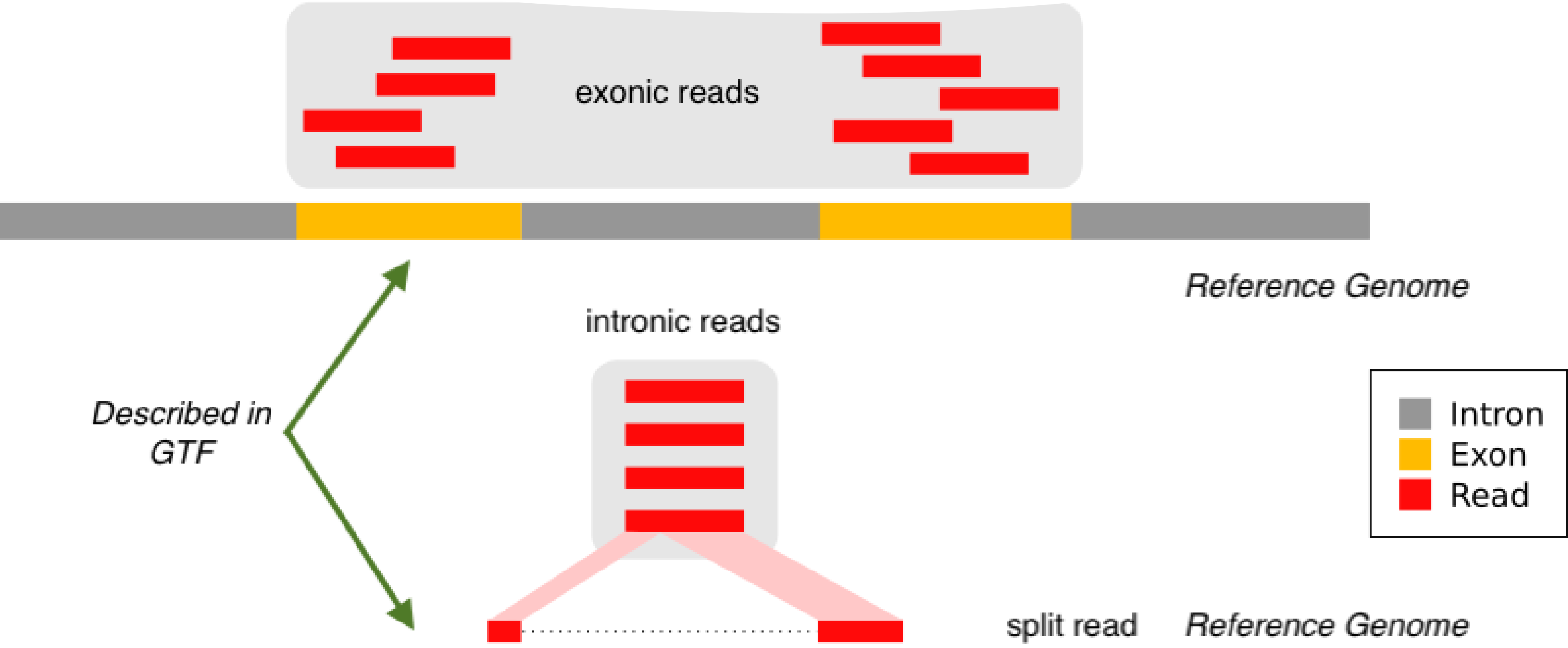
1. Reads that map entirely within exons
2. Reads that span two or more exons



Splice-aware Alignment Tools

- **Splice junction mapping** is critical for mapping reads across splice junctions and understanding alternative transcript usage.
- **Splice aware** aligners will map to splice junctions described in annotation file

greatest downside: it can be resource-intensive!

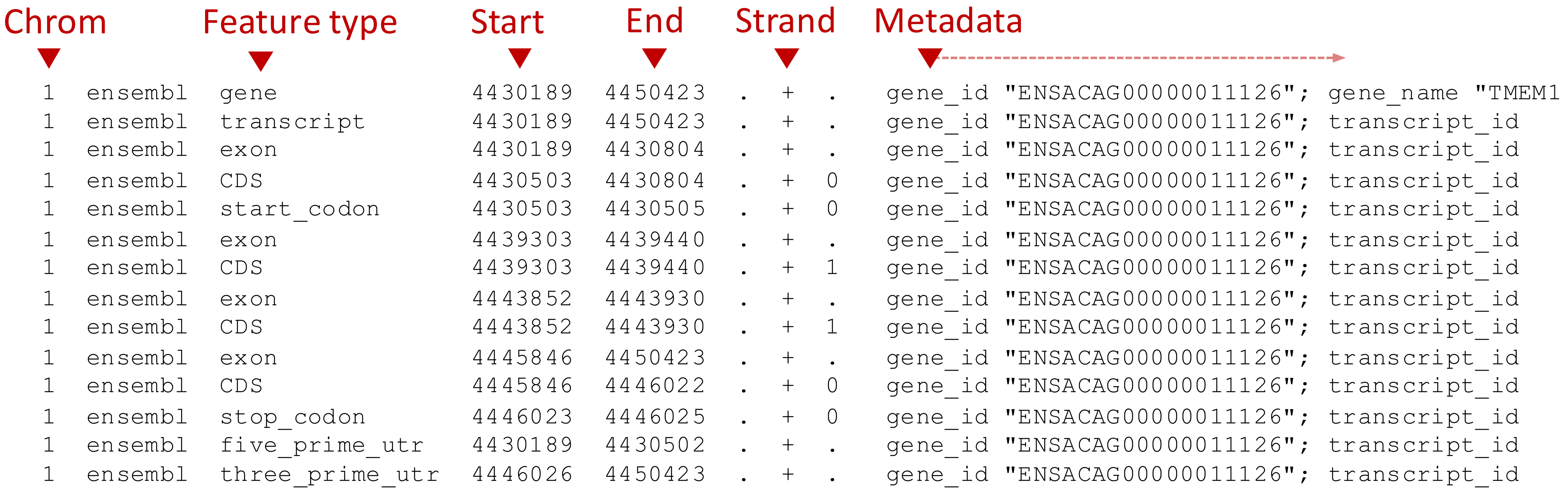


What file type contains coordinates for exons?

chr1	78999	79123
chr1	79699	81423
chr1	88279	89185

Typically, the intron/exon annotations are available here!

GTF file format

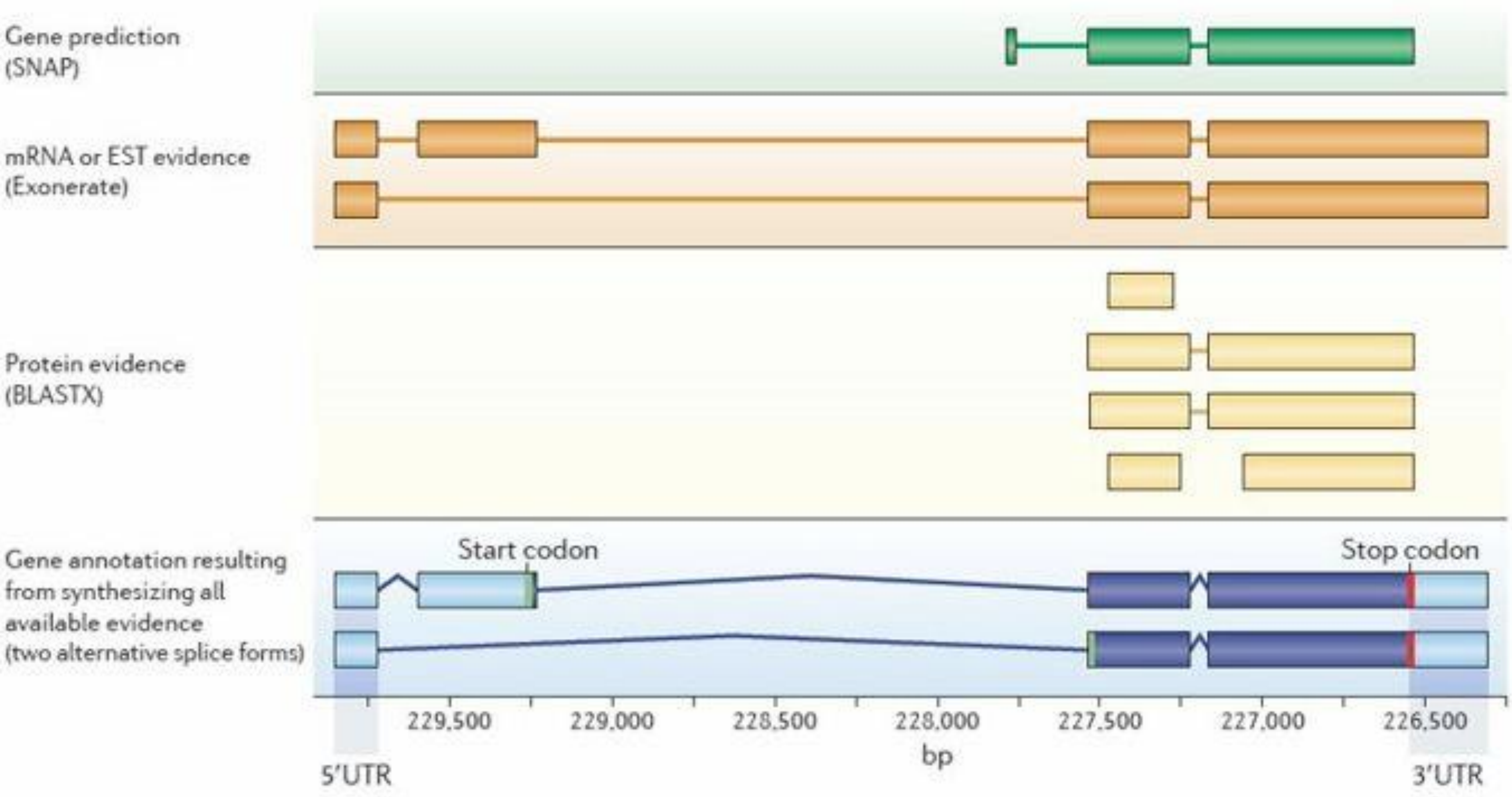


Chrom Feature type Start End Strand Metadata

```
1  ensembl  gene             4430189  4450423  .  +  .   gene_id "ENSACAG00000011126"; gene_name "TMEM1
1  ensembl  transcript       4430189  4450423  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  exon             4430189  4430804  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  CDS              4430503  4430804  .  +  0   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  start_codon      4430503  4430505  .  +  0   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  exon             4439303  4439440  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  CDS              4439303  4439440  .  +  1   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  exon             4443852  4443930  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  CDS              4443852  4443930  .  +  1   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  exon             4445846  4450423  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  CDS              4445846  4446022  .  +  0   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  stop_codon       4446023  4446025  .  +  0   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  five_prime_utr   4430189  4430502  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  three_prime_utr  4446026  4450423  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
```

- Tab-delimited text files
- Used to quantify the number of reads which align to different genome features

Gene annotation



Gene annotations generally include UTRs, alternative splice isoforms and have attributes such as evidence trails.

Splice-aware aligners

HISAT2

STAR

TopHat2

Splice-unaware aligner

Bowtie2

BWA

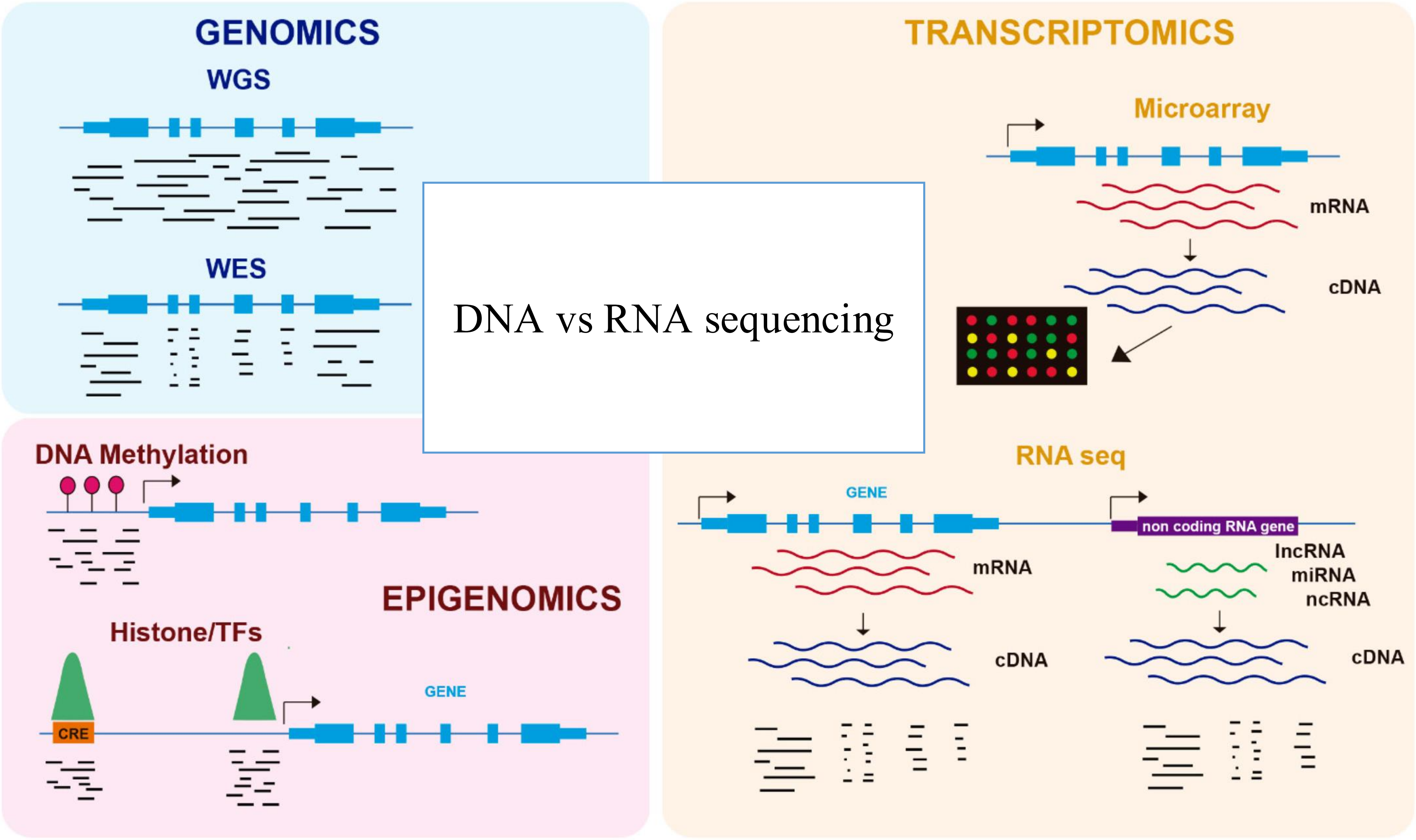
minimap2

RNA-Seq

?

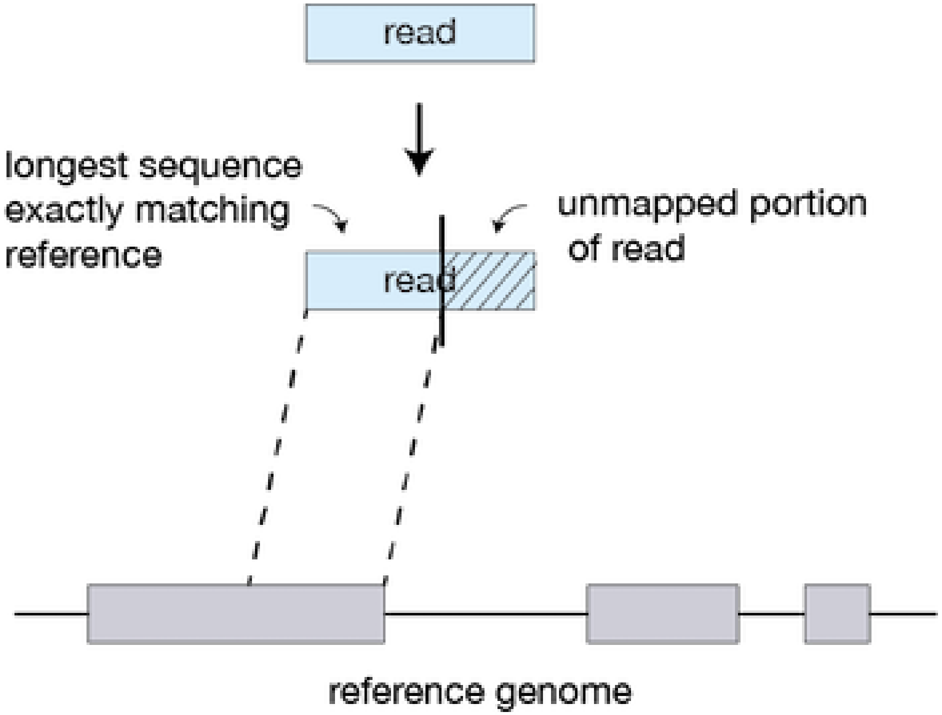
Question: For what applications is it okay to use a splice unaware aligner?

DNA vs RNA sequencing



How does STAR (Spliced Transcripts Alignment to a Reference) work?

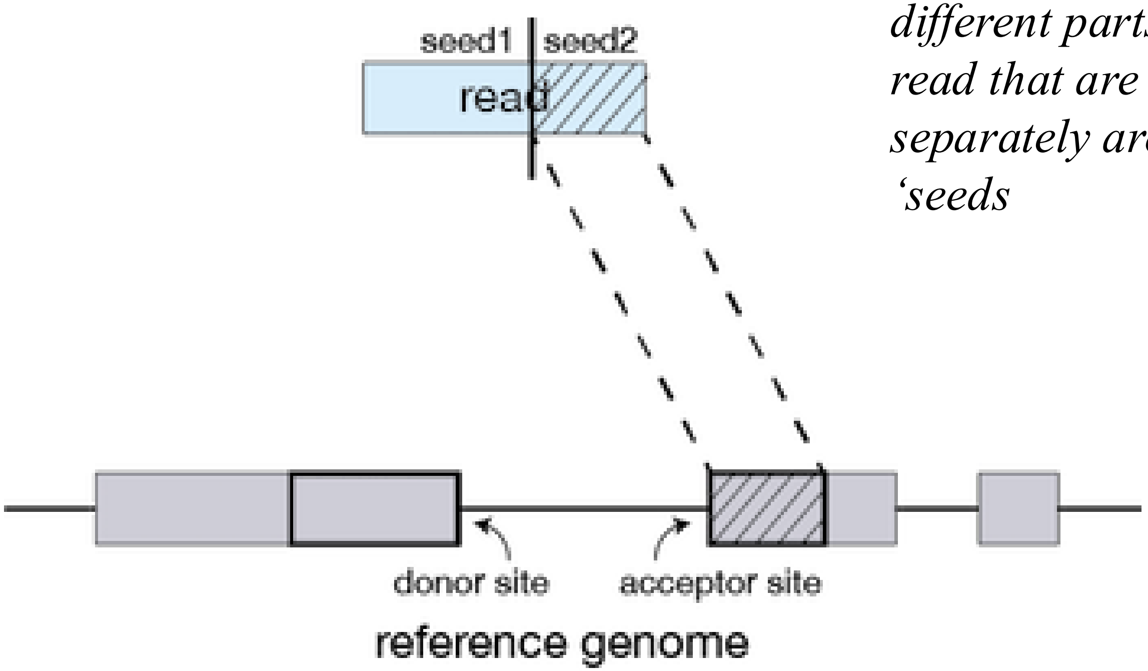
- STAR Alignment Strategy
  - Step 1: Seed Searching



longest matching sequences are called the Maximal Mappable Prefixes (MMPs)

How does STAR (Spliced Transcripts Alignment to a Reference) work?

- STAR Alignment Strategy
  - Step 1: Seed Searching

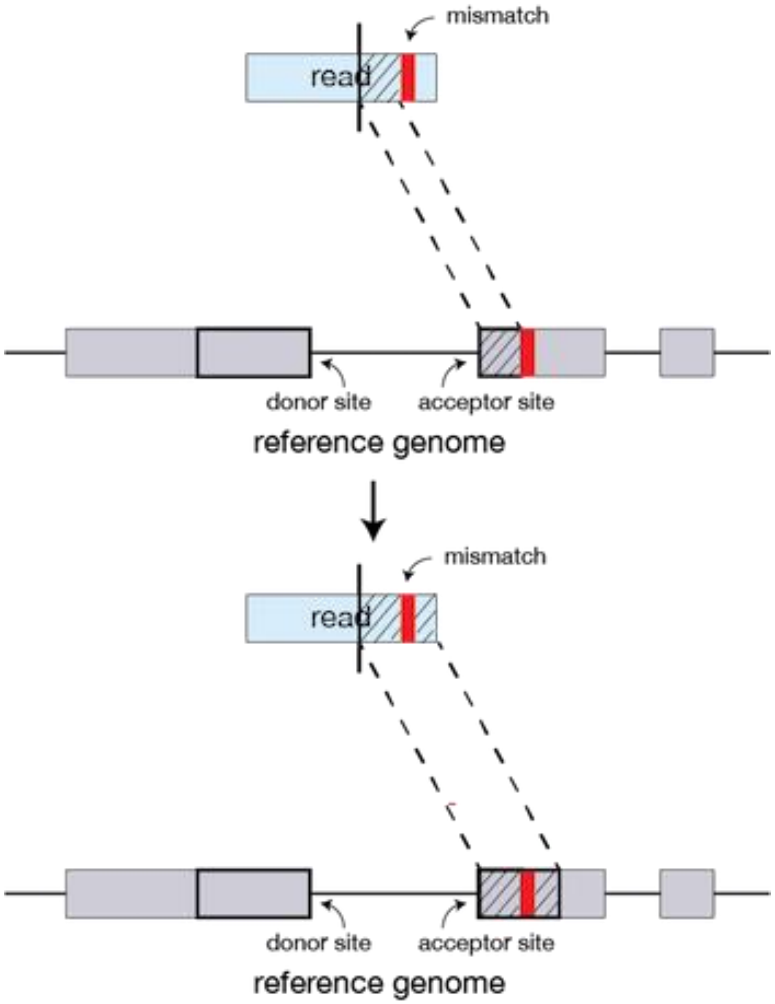


different parts of the read that are mapped separately are called 'seeds

How does STAR (Spliced Transcripts Alignment to a Reference) work?

- STAR Alignment Strategy
  - Step 1: Seed Searching

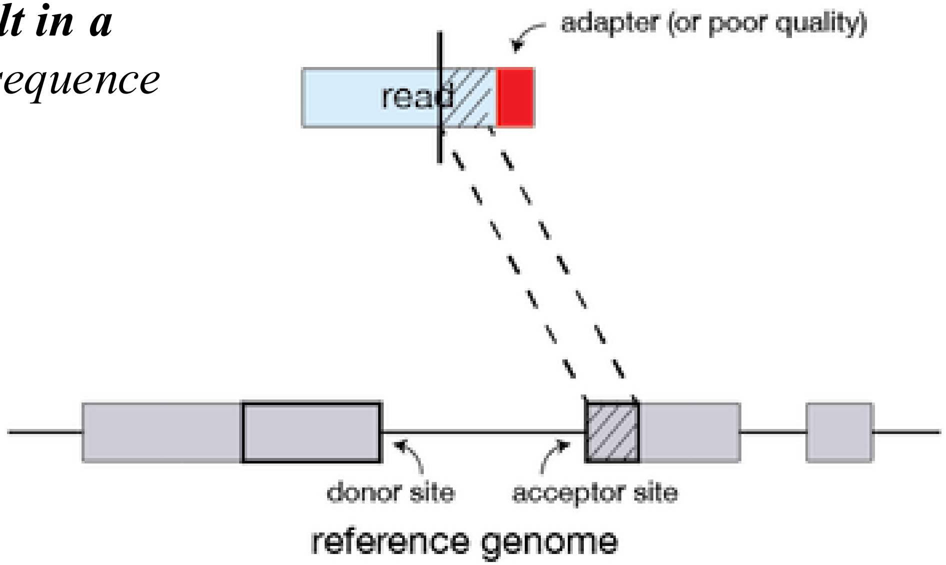
***If STAR does not find an exact matching sequence**, the MMPs will be extended.*



How does STAR (Spliced Transcripts Alignment to a Reference) work?

- STAR Alignment Strategy
  - Step 1: Seed Searching

If extension does not result in a good alignment*, then the sequence will be* ***soft clipped.***

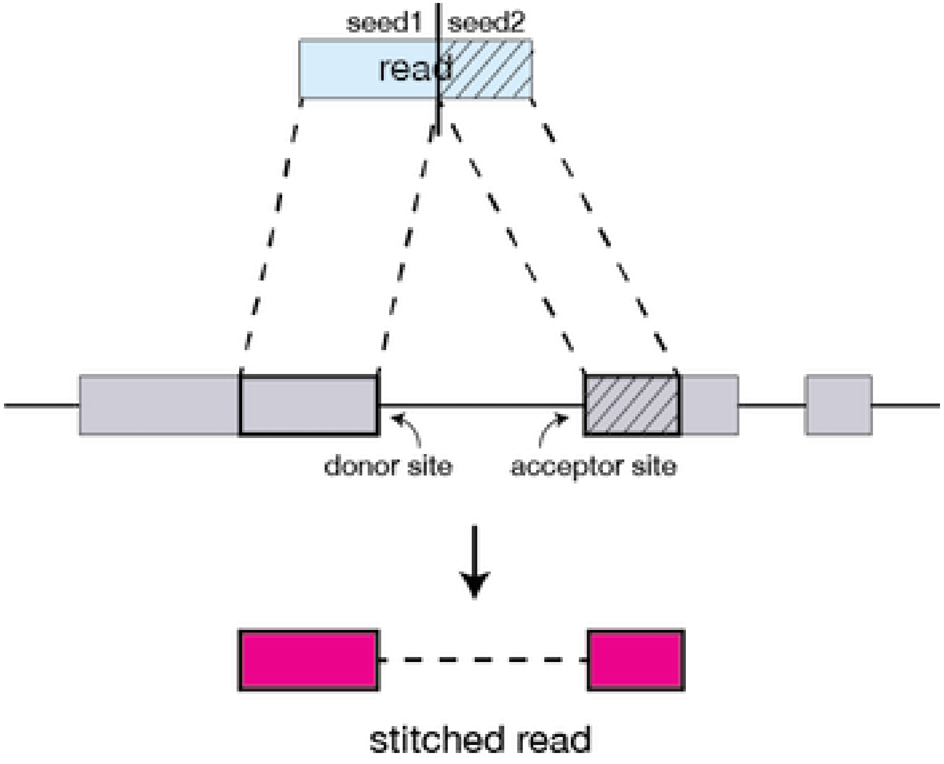


How does STAR (Spliced Transcripts Alignment to a Reference) work?

- STAR Alignment Strategy
  - Step 1: Clustering, stitching, and scoring

The seeds are stitched together to create a complete read based on the best alignment

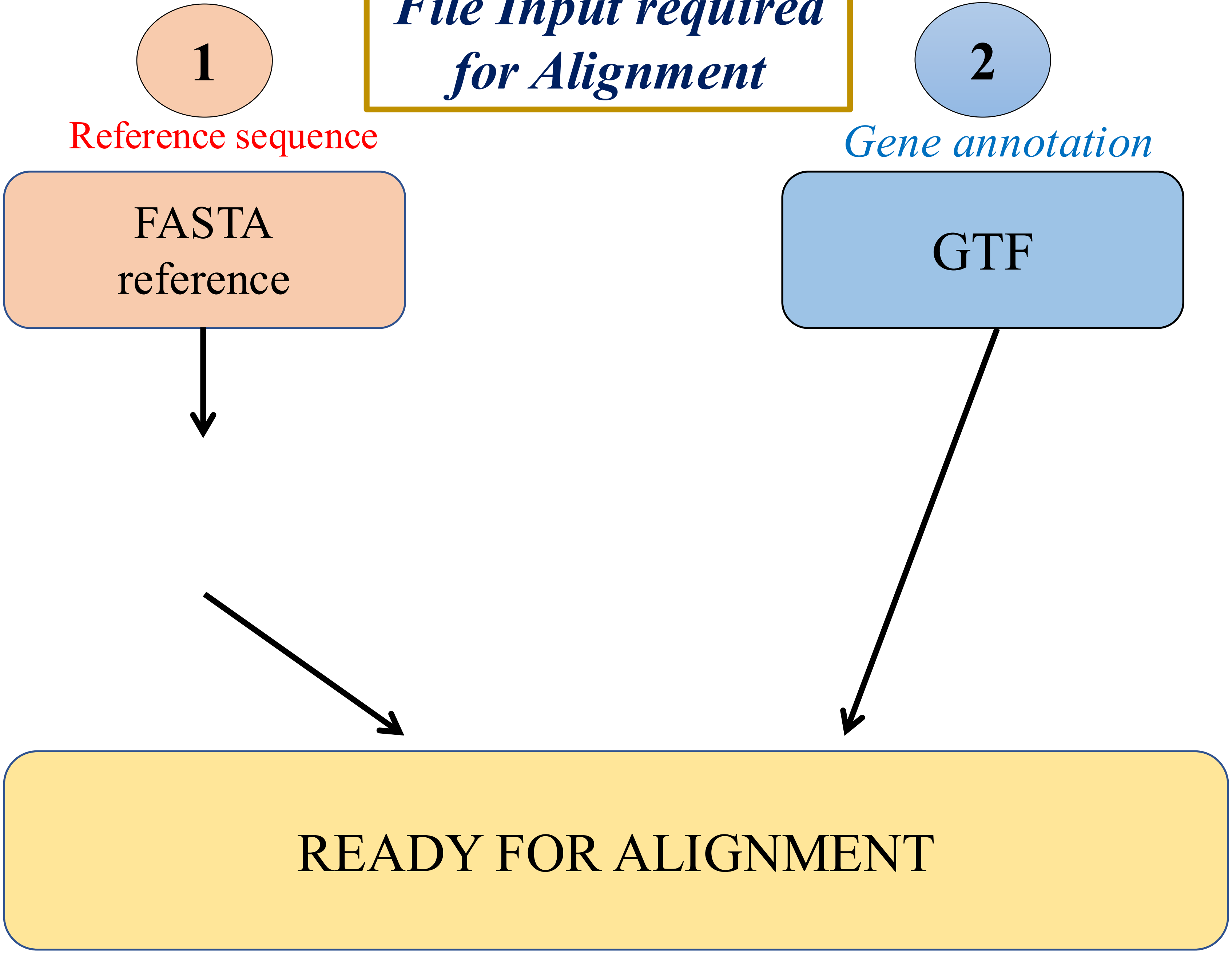
**scoring is based on mismatches, indels, gaps, etc.*



File Inputs required for Alignment

- Gene annotation = which parts of the reference sequence correspond to genes/features/transcripts?
    - GTF file
- Reference sequence = what are you aligning to?

File Input required for Alignment



1

Reference sequence

FASTA reference

2

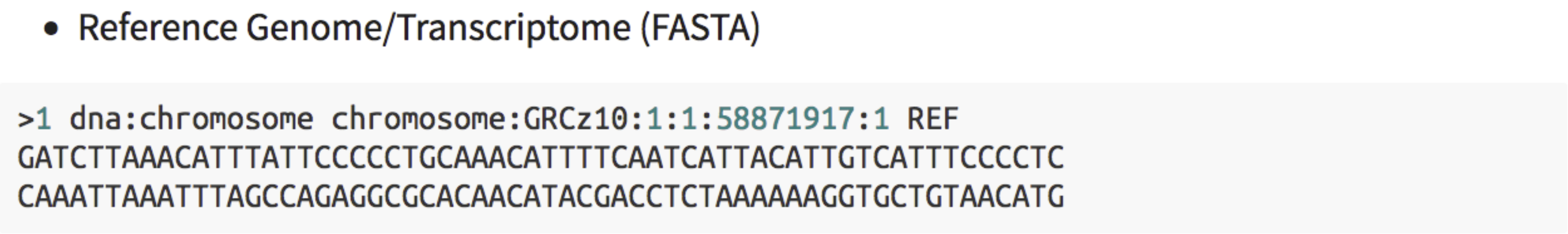
Gene annotation

GTF

READY FOR ALIGNMENT

Reference Genome

- The reference genome are usually stored in a plain text **FASTA file**



- Reference Genome/Transcriptome (FASTA)

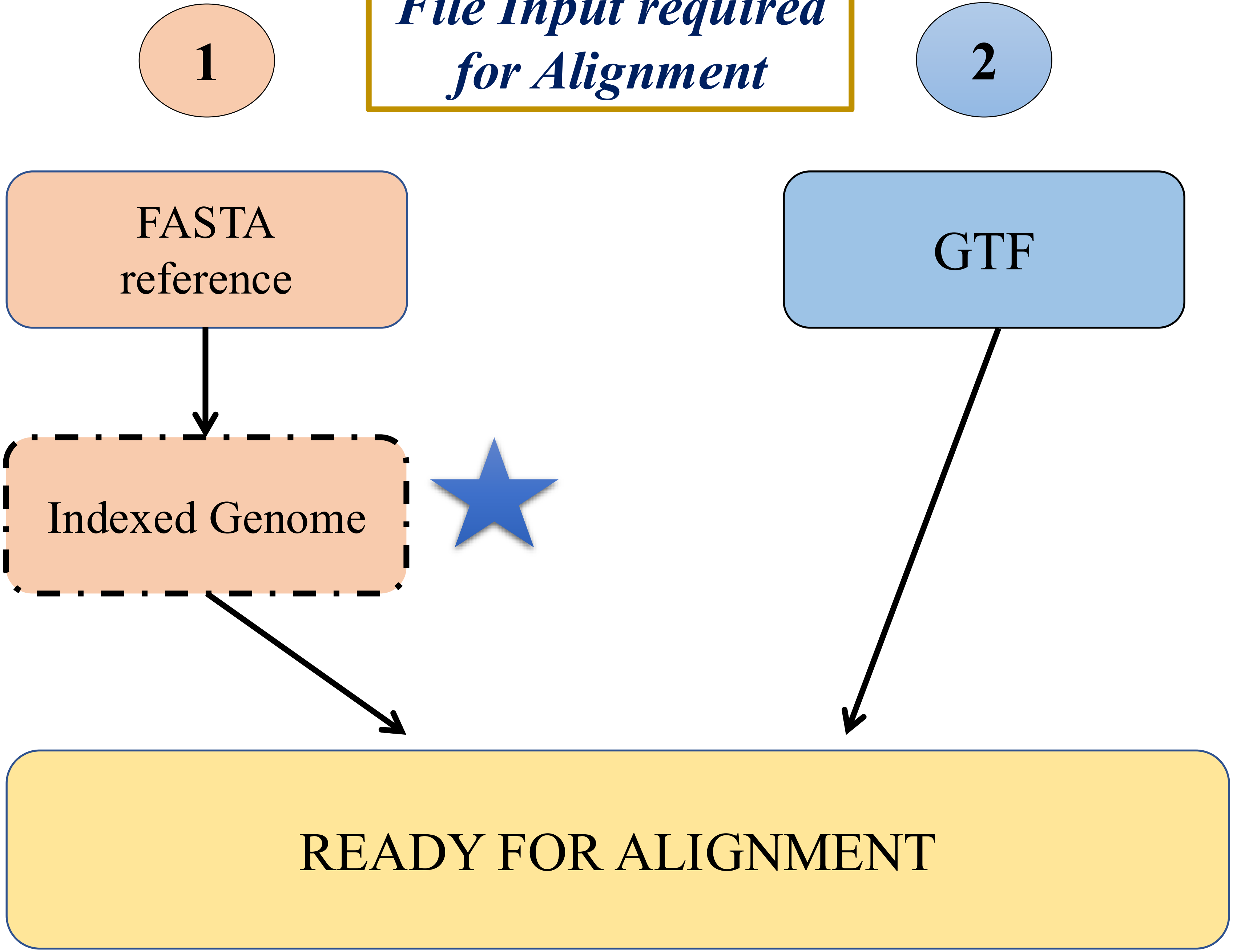
```
>1 dna:chromosome chromosome:GRCz10:1:1:58871917:1 REF
GATCTTAAACATTTATTCCCCCTGCAAACATTTTCAATCATTACATTGTCATTTCCCCTC
CAAATTAAATTTAGCCAGAGGCGCACAACATACGACCTCTAAAAAAGGTGCTGTAACATG
```

Where can you find these genomic files?

General biological databases: Ensembl, GENCODE, and UCSC

Organism-specific biological databases: Wormbase, Flybase, ***CryptoDB***, etc. (often updated more frequently, so may be more comprehensive)

File Input required for Alignment



1

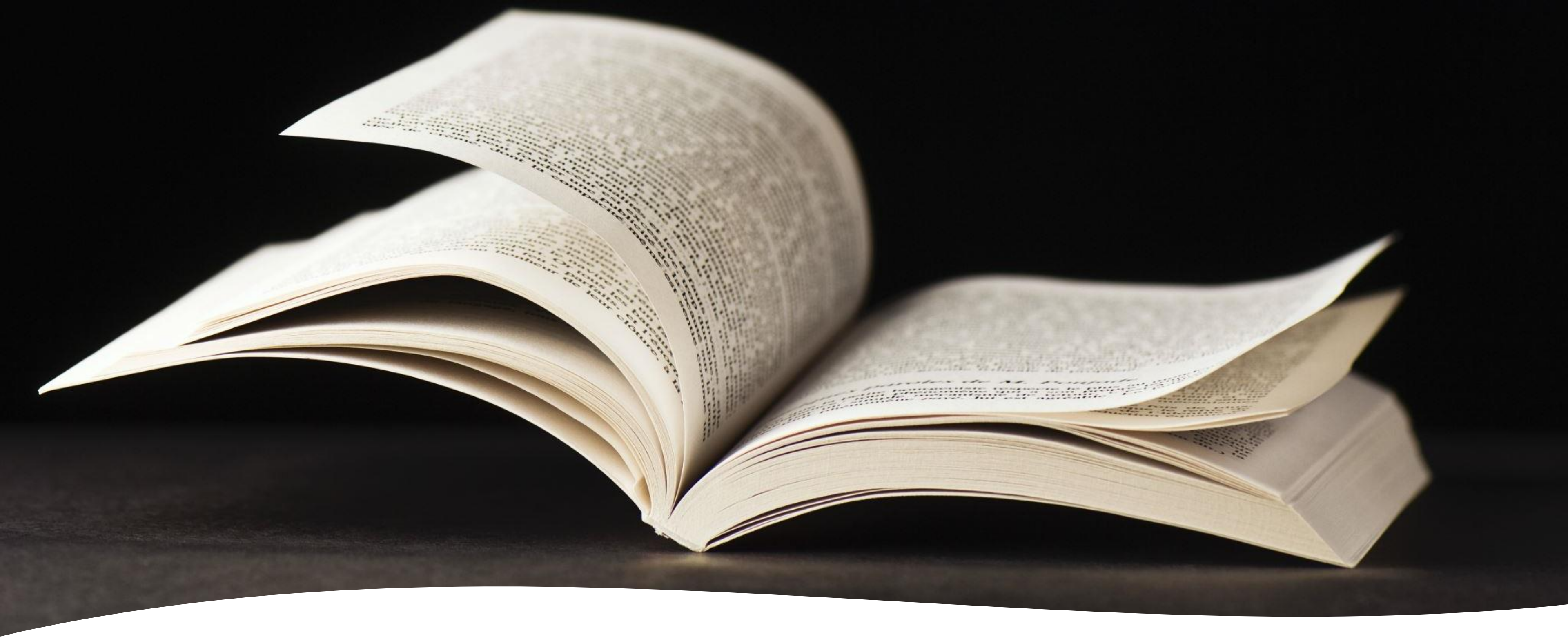
2

FASTA reference

GTF

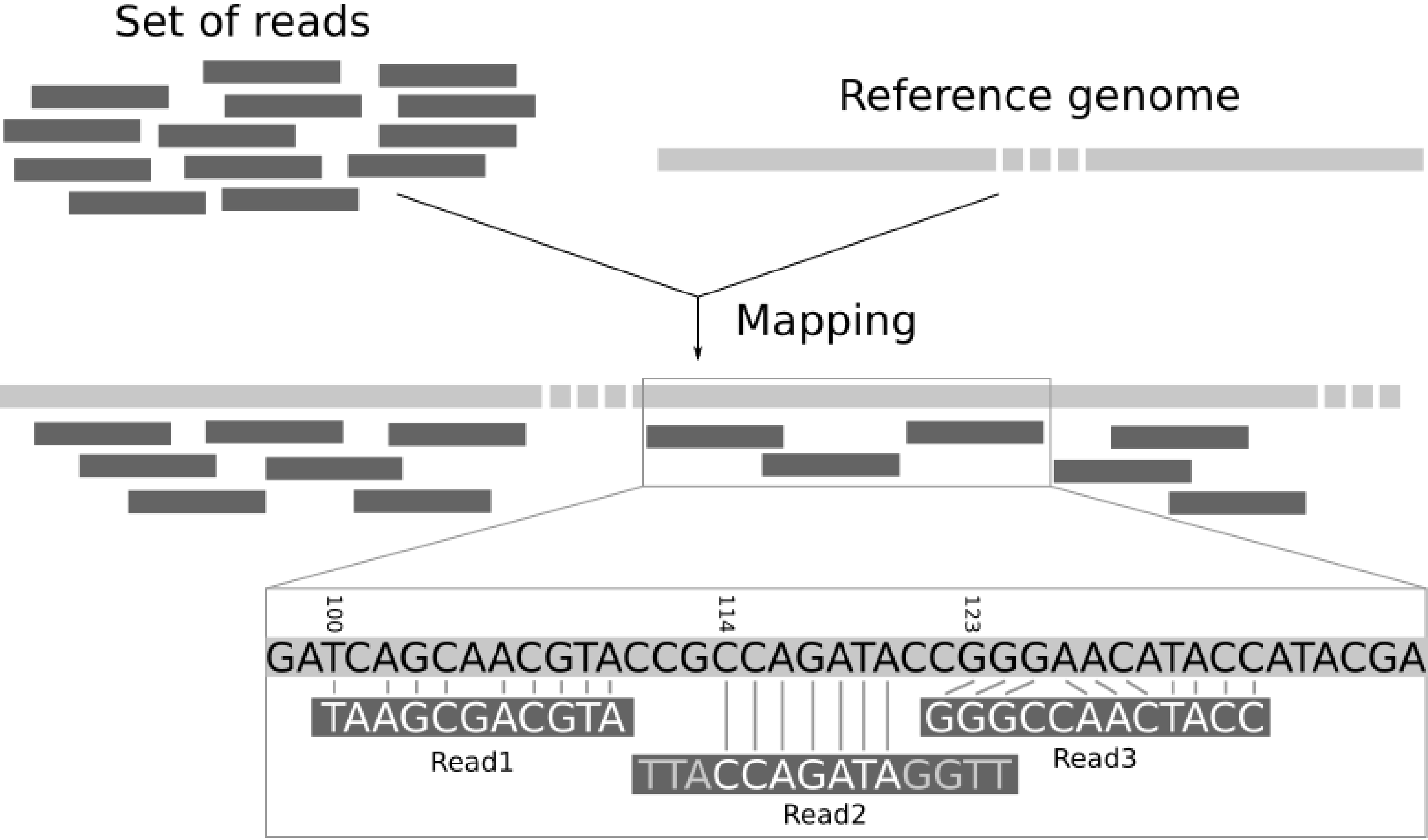
Indexed Genome

READY FOR ALIGNMENT



Indexing benefits

- Think of an index as a table of contents in a book. If we are searching for where chapter 8 starts in a book, we can either search from beginning to end and depending on the size of the book, this could take a long time.
- Alternatively, we could use the table of contents to jump to chapter 8.
- It is much more efficient to look up where the chapter begins using the pre-built index (table of contents) than going through every page.


Set of reads
Reference genome
Mapping
100
114
123
GATCAGCAACGTACCGCCAGATACCGGGAACATACCATACGA
TAAGCGACGTA
Read1
TTACCAGATAGGTT
Read2
GGGCCAACTACC
Read3

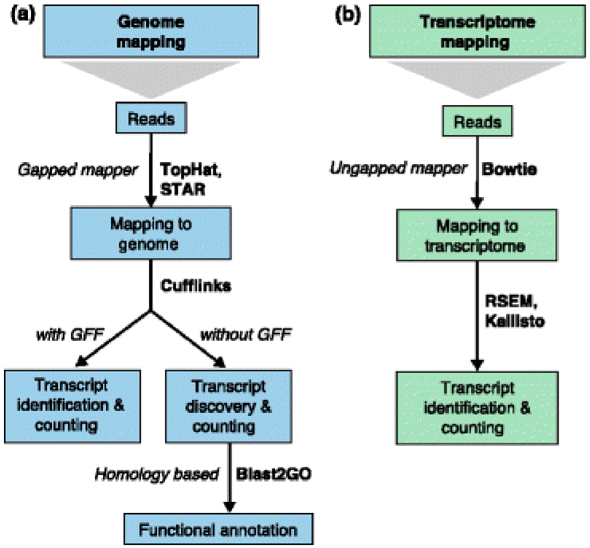
2

RNASeq Mapping Challenges: Genome vs Transcriptome

Genome alignment: maps reads to the full reference genome (including introns)

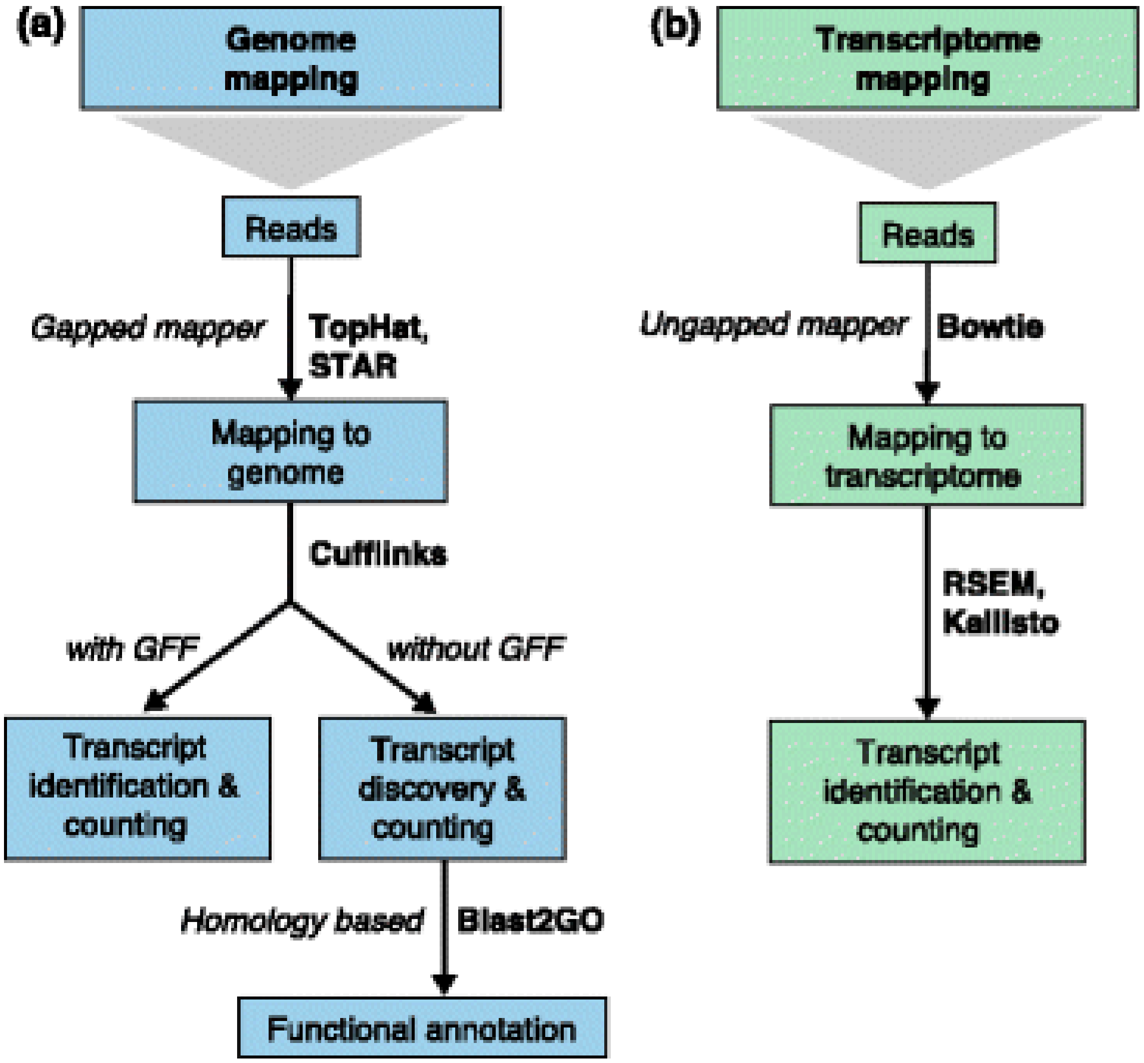
Transcriptome alignment: maps reads directly to a reference set of transcript sequences

These approaches answer different biological questions and involve different computational trade-offs



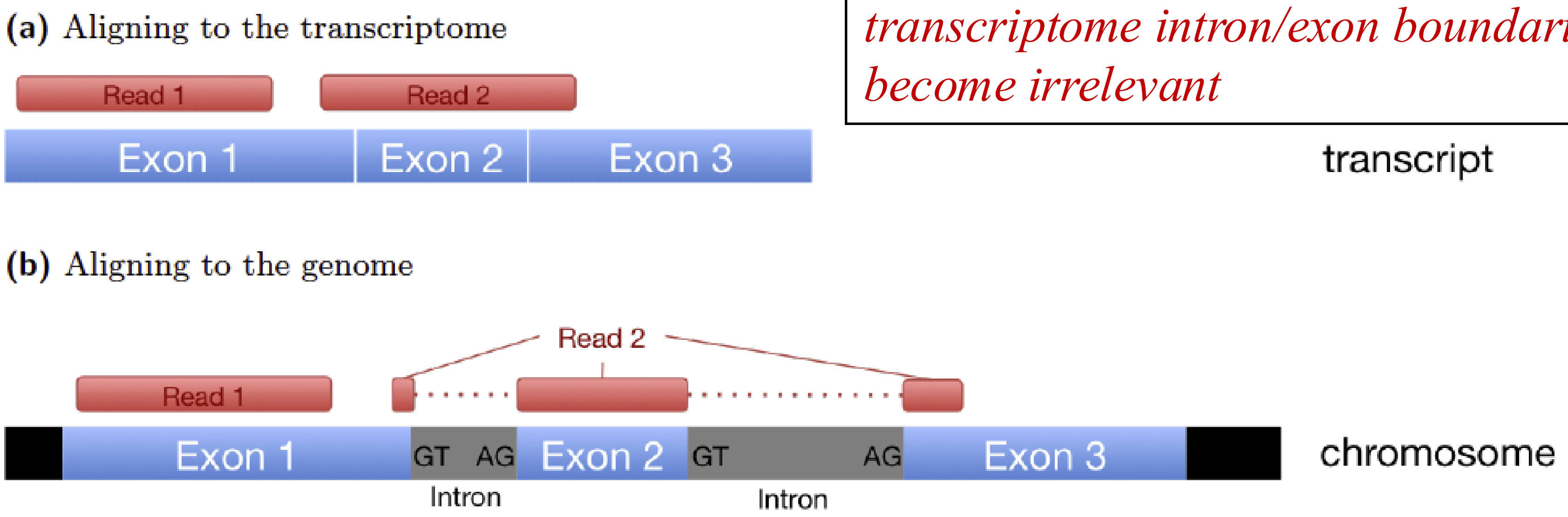
2

RNASeq Mapping Challenges: Genome vs Transcriptome



Benefits of Transcriptome Mapping: intron/exon boundaries

If you are mapping reads to a transcriptome intron/exon boundaries become irrelevant

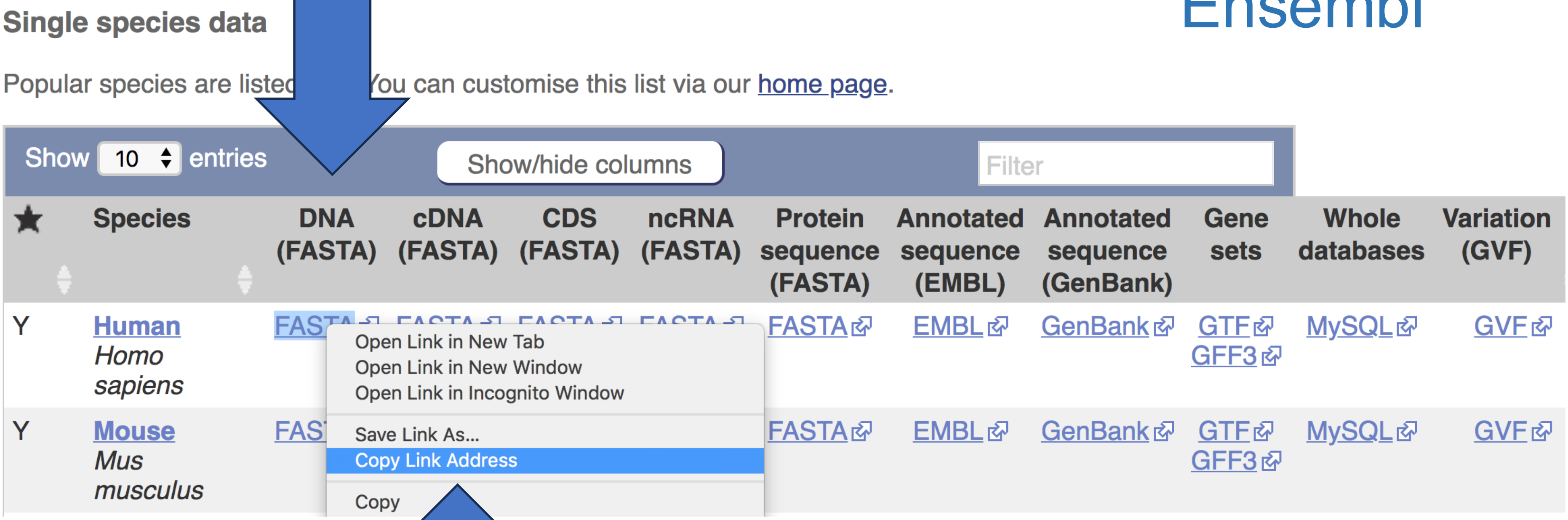


computationally a much harder task

Benefits of Transcriptome Mapping: smaller reference = faster analysis

Genome Reference (DNA): contain complete DNA sequence of organism including coding and noncoding regions

Downloading FASTA from Ensembl



Single species data

Popular species are listed [illegible] You can customise this list via our home page.

Show 10 entries

Show/hide columns

Filter

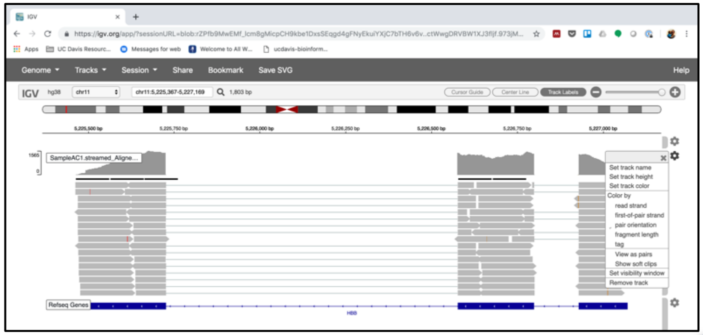
★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)
Y	**Human** *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF
Y	**Mouse** *Mus musculus*	FAS				FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF

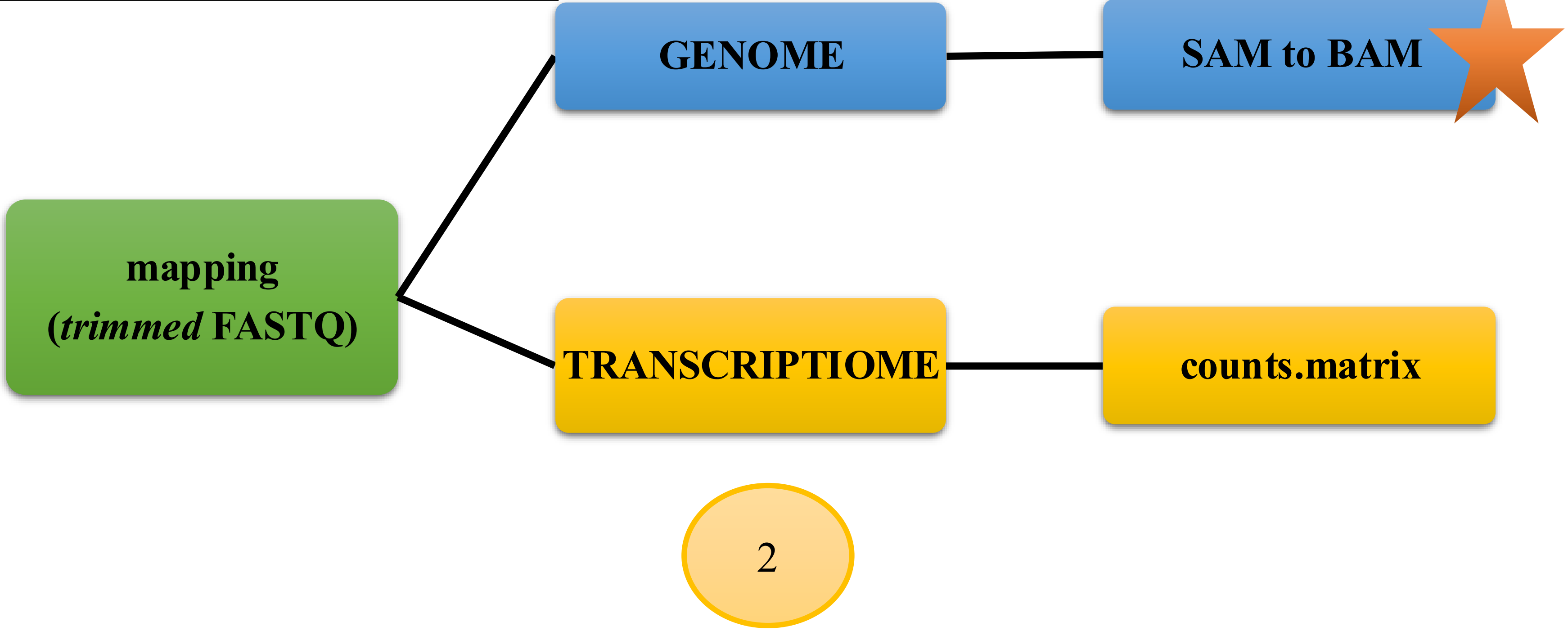
Transcriptome Reference (cDNA): only contains known transcripts

Pros of Transcriptome alignment

- Very fast
- Low memory usage
- Small output files

Cons of transcriptome alignment





Con: Forgo transcript discovery with transcriptome alignment

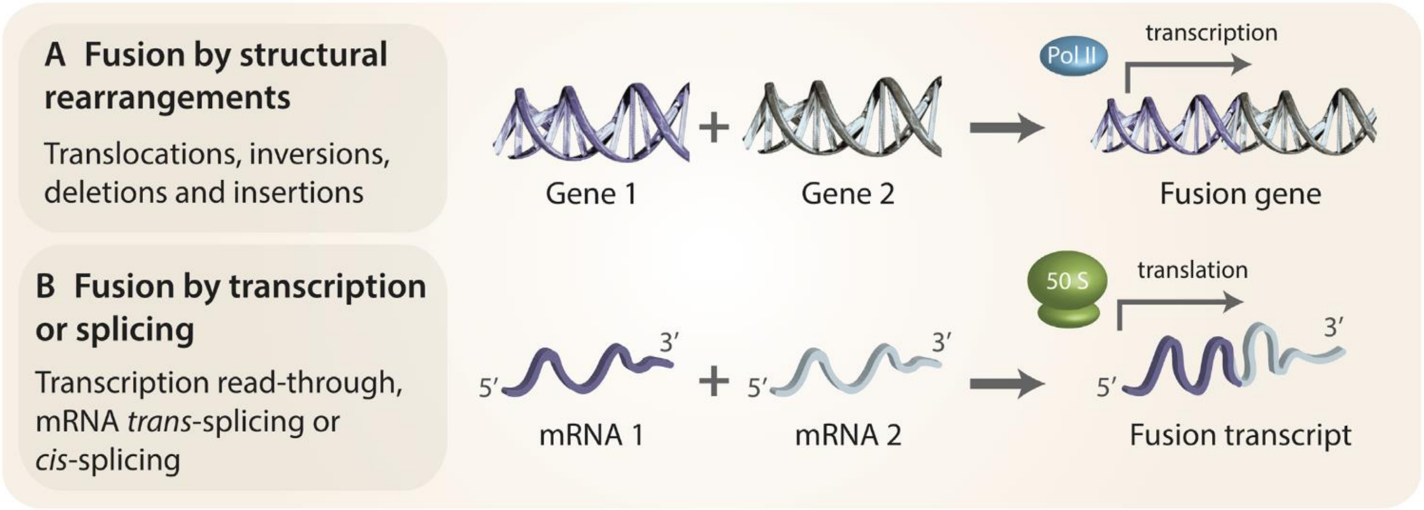
- Refers to allowing researchers to identify new splice variants or transcripts not previously annotated
- Transcriptome alignment is limited because it maps reads **only to known, annotated transcripts** rather than the full genome.

The input FASTA file only contains known protein-coding sequences

Con: Forgo fusion gene detection with transcriptome alignment

- Fusion gene occurs when sequences from two different genes are joined due to genomic rearrangements

Gene fusion formation



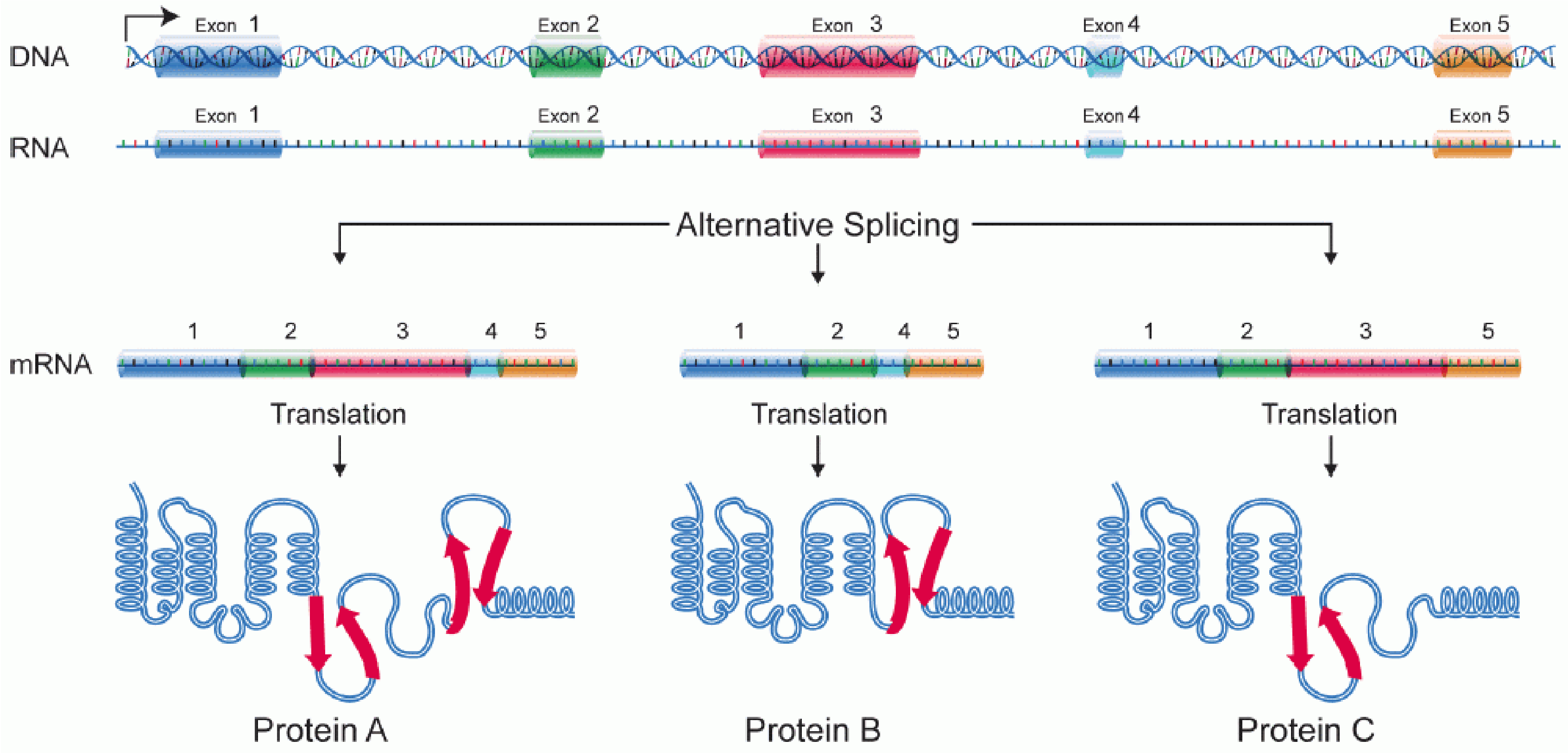
A Fusion by structural rearrangements

Translocations, inversions, deletions and insertions

B Fusion by transcription or splicing

Transcription read-through, mRNA *trans*-splicing or *cis*-splicing

Con: Forgo detection of novel splice variants with transcriptome alignment



Biological Questions

“Are genes differentially expressed between conditions?”

✓ Genome alignment — Yes
✓ Transcriptome alignment — Yes

- Transcriptome alignment is usually preferred for speed and simplicity.

Biological Questions

- “Is there alternative splicing between conditions?”

✓ Genome alignment — Yes
✗ Transcriptome alignment — Only if isoforms already annotated

- Genome alignment is better, especially for novel splicing.

Biological Questions

- “Are there novel transcripts in my dataset?”

✓ Genome alignment — Yes
✗ Transcriptome alignment — No

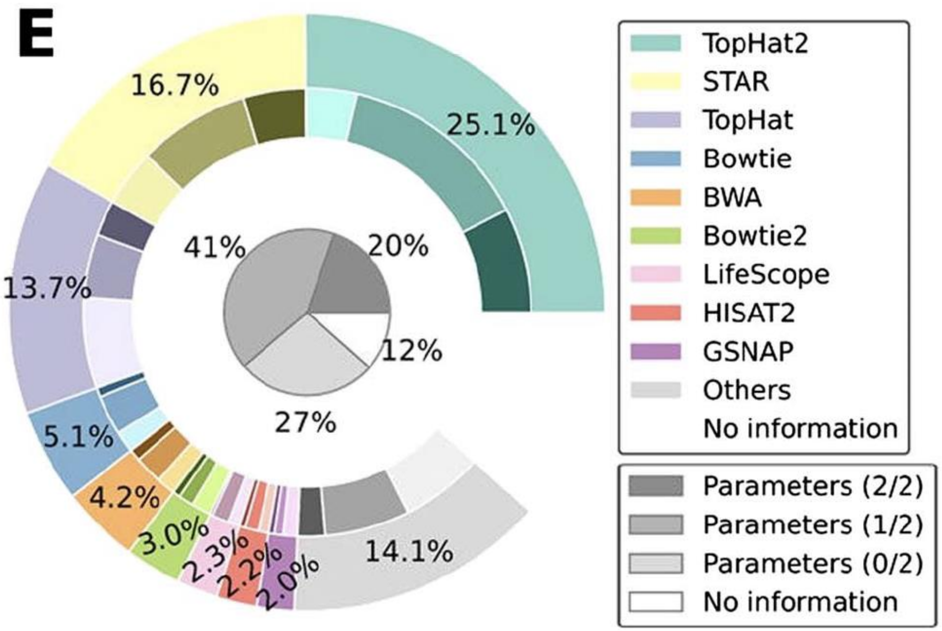
Multiple Alignment Programs available

Genome

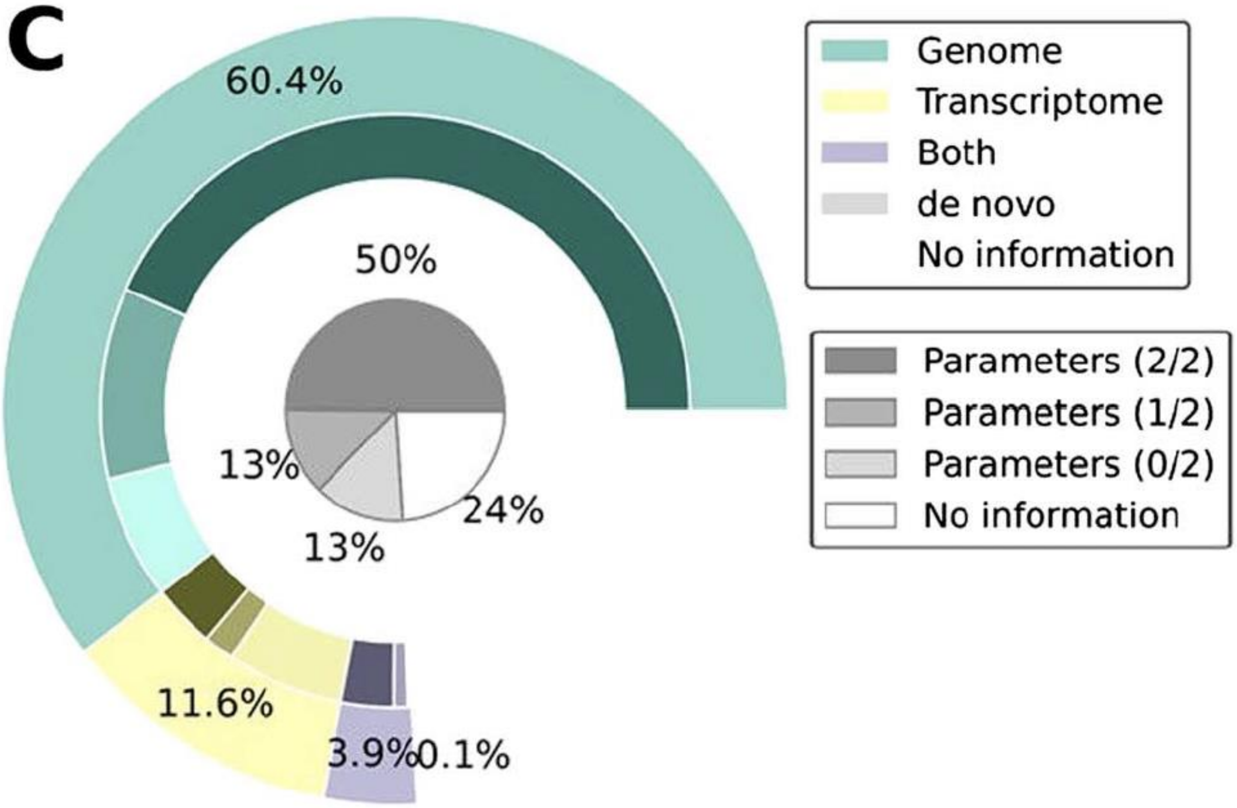
- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

Transcriptome

- Salmon
- Kallisto
- Sailfish



What does the scientific community do?



Simoneau et al. 2021

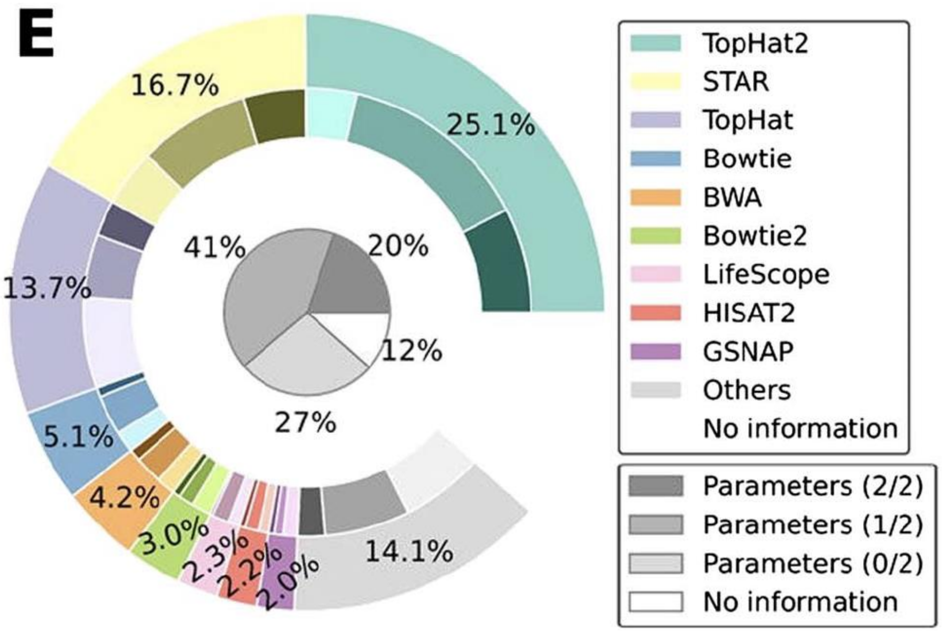
Programs we will use:

Genome

- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

Transcriptome

- Salmon
- Kallisto
- Sailfish



Class activity
Indexing genomes

3

RNASeq Mapping Challenges: Computationally Expensive

Map millions of reads **accurately** and in a reasonable **time**, despite the presence of sequencing errors, genomic variation, and repetitive elements.

Aligners - Speed and Memory

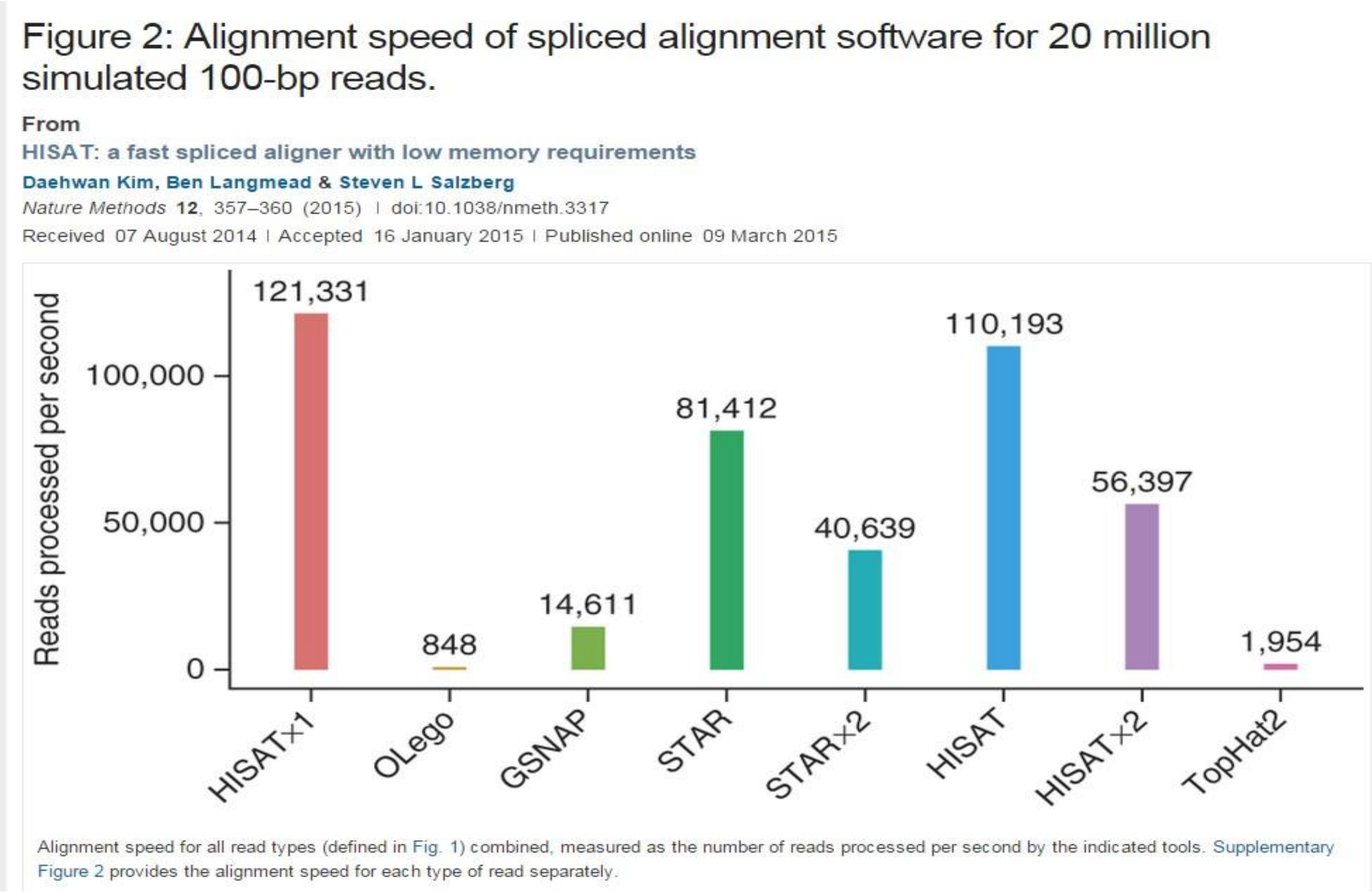


Figure 2: Alignment speed of spliced alignment software for 20 million simulated 100-bp reads.

From

HISAT: a fast spliced aligner with low memory requirements

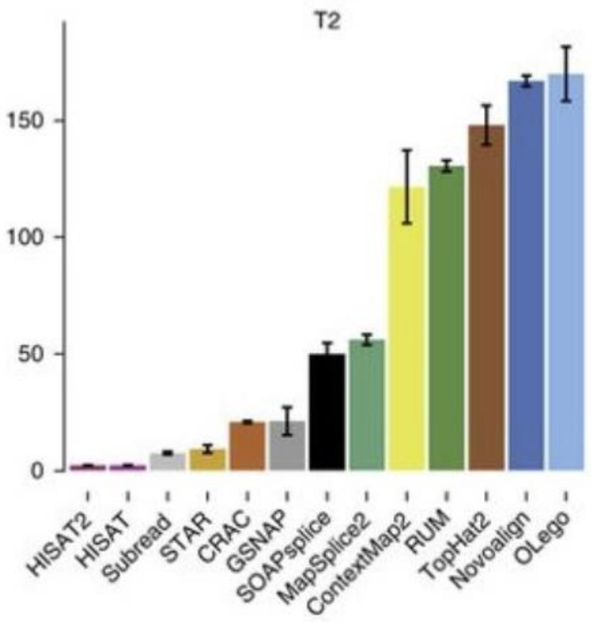
Daehwan Kim, Ben Langmead & Steven L Salzberg

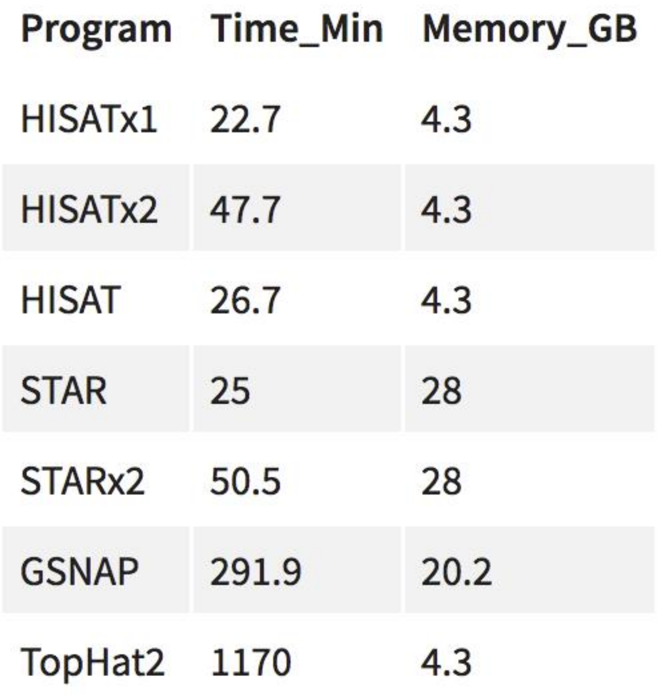
Nature Methods **12**, 357–360 (2015) | doi:10.1038/nmeth.3317

Received 07 August 2014 | Accepted 16 January 2015 | Published online 09 March 2015

Alignment speed for all read types (defined in Fig. 1) combined, measured as the number of reads processed per second by the indicated tools. Supplementary Figure 2 provides the alignment speed for each type of read separately.

Aligners - Speed and Memory





Program	Time_Min	Memory_GB
HISATx1	22.7	4.3
HISATx2	47.7	4.3
HISAT	26.7	4.3
STAR	25	28
STARx2	50.5	28
GSNAP	291.9	20.2
TopHat2	1170	4.3

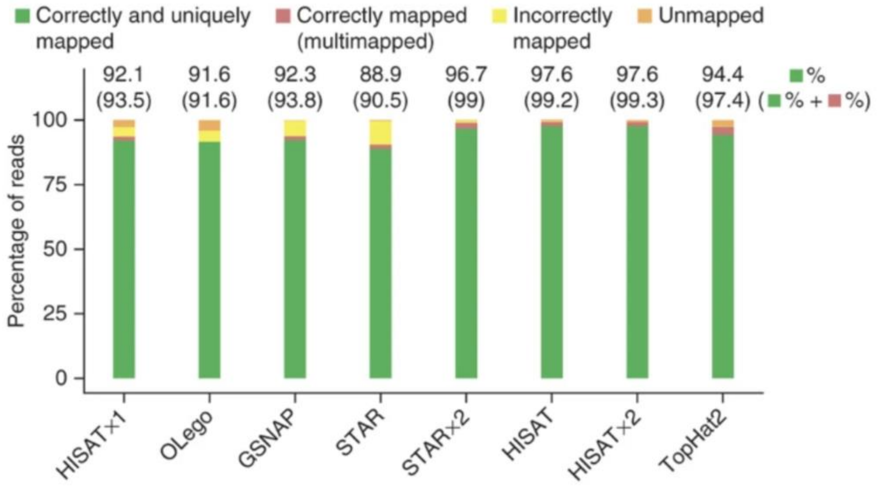
HISAT2

- Stands for **h**ierarchical **i**ndexing for **s**pliced **a**lignment of **t**ranscripts 2
- HISAT2 is an aligner that is used for mapping next-generation sequencing reads
  - Used for whole genome, whole-exome, and transcriptome datasets
  - Is a ‘splice-aware’ aligner
  - Requires a reference genome
  - Is the fastest spliced mapper currently available

HISAT2 has a small memory footprint

- The STAR program runs faster than TopHat2 but both have a memory requirement of ~28GB
- The memory requirement for HISAT2 is ~5GB
  - This makes it possible to do alignments on your laptop!

Figure 3: Alignment accuracy of spliced alignment software for 20 million simulated 100-bp reads.

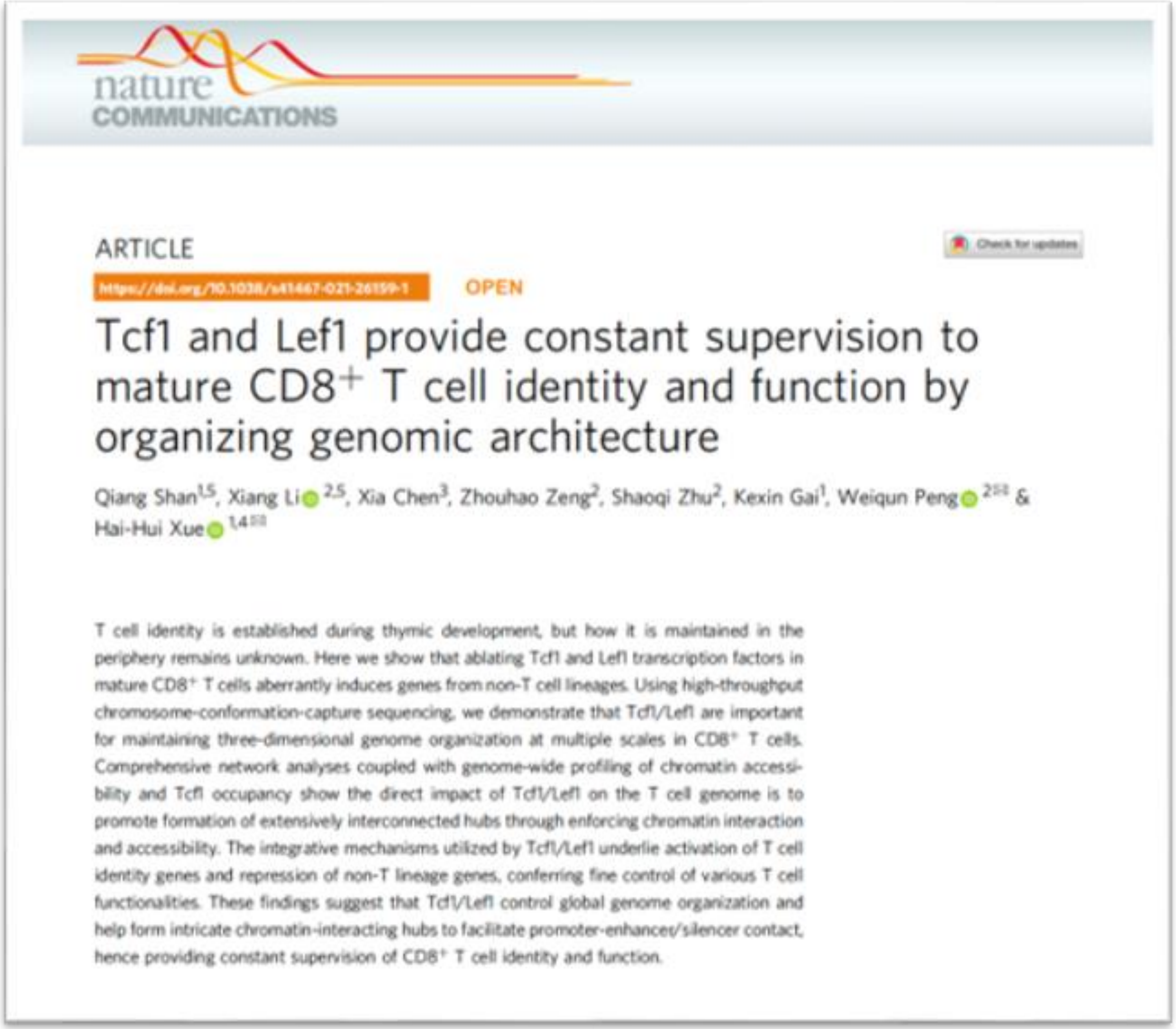


HISAT2 usage

- http://daehwankimlab.github.io/hisat2/

- hisat2 [options]* -x <hisat2-idx> {-1 <m1> -2 <m2> | -U <r> | --sra-acc <SRA accession number>} [-S <hit>]

The dataset



nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-26159-1 OPEN

Tcf1 and Lef1 provide constant supervision to mature CD8$^+$ T cell identity and function by organizing genomic architecture

Qiang Shan[1,5], Xiang Li[2,5], Xia Chen[3], Zhouhao Zeng[2], Shaoqi Zhu[2], Kexin Gai[1], Weiqun Peng[2✉] & Hai-Hui Xue[1,4✉]

T cell identity is established during thymic development, but how it is maintained in the periphery remains unknown. Here we show that ablating Tcf1 and Lef1 transcription factors in mature CD8$^+$ T cells aberrantly induces genes from non-T cell lineages. Using high-throughput chromosome-conformation-capture sequencing, we demonstrate that Tcf1/Lef1 are important for maintaining three-dimensional genome organization at multiple scales in CD8$^+$ T cells. Comprehensive network analyses coupled with genome-wide profiling of chromatin accessibility and Tcf1 occupancy show the direct impact of Tcf1/Lef1 on the T cell genome is to promote formation of extensively interconnected hubs through enforcing chromatin interaction and accessibility. The integrative mechanisms utilized by Tcf1/Lef1 underlie activation of T cell identity genes and repression of non-T lineage genes, conferring fine control of various T cell functionalities. These findings suggest that Tcf1/Lef1 control global genome organization and help form intricate chromatin-interacting hubs to facilitate promoter-enhancer/silencer contact, hence providing constant supervision of CD8$^+$ T cell identity and function.

SRR_number	datatype	treatment	cell	replicate
SRR13423162	RNAseq	WT	CD8 T cell	1
SRR13423163	RNAseq	WT	CD8 T cell	2
SRR13423164	RNAseq	WT	CD8 T cell	3
SRR13423165	RNAseq	TCF1 - KO	CD8 T cell	1
SRR13423166	RNAseq	TCF1 - KO	CD8 T cell	2
SRR13423167	RNAseq	TCF1 - KO	CD8 T cell	3

Overall Recommendations based on Research Question

	Question 1: Differential Expression	**Question 2: Splicing Isoforms**	**Question 3: Novel transcripts**	**Question 4: Transcript Level quantification**
Mapping	STAR, HISAT2, Salmon, Kallisto	STAR, HISAT2 TopHat	STAR, HISAT2	Salmon, Kallisto
Quantification	HTSeq, feature Counts	StringTie, Suppa2, HTSeq, rMATS	StringTie, Cufflinks	Salmon, Kallisto
Comment	*No need to quantify when using Salmon, Kallisto	*Use ballgown or DEXSeq for isoform-level analysis in R		*Not used for transcript discovery

Data Analysis Workflow: File formats

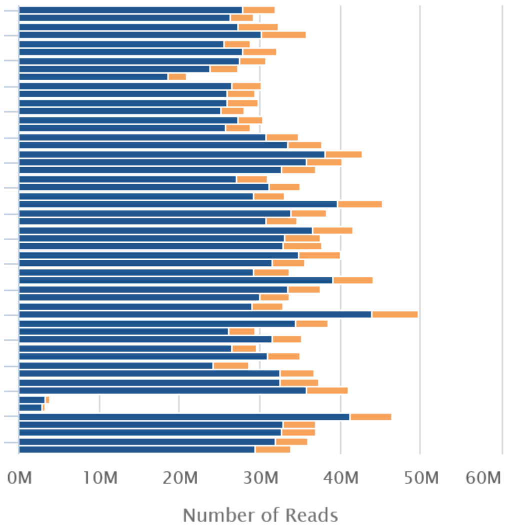
- Quality Control
  - Sample Quality and consistency (FASTQC)
  - Is trimming appropriate - quality/adapters (trimmomatic)
  - **FASTQ file**
- Alignment/Mapping
  - Reference Target (Sequence and annotation files)
  - Alignment programs & parameters (hisat2)
  - **BAM file**
- Quantification (next week)
  - Counting methods and parameters
  - **Count matrices**

FASTQC will aid in identifying if minimum requirements are met

	Question 1: Which genes are differentially expressed?	**Question 2: Are different splicing isoforms expressed?**	**Question 3: Are you interested in non-coding RNAs? Novel transcripts?**
Reads	> 10M	> 25-50M	**> 25-50M**
Biological replicates	3 replicates	> 3 replicates	> 3 replicates
SE or PE	50bp SE (minimum)	100bp SE (minimum)	**150bp PE**
FASTQC	Q30 > 70%	Q30 > 70%	Q30 > 70%

Next Week:

- Storing aligned reads: SAM/BAM file formats
- We will review outputs from HISAT2_exercise1 vs HISAT2_exercise2
- We will create a MULTIQC output
- We will use RSEQC to QC alignment statistics



RNA-Seq Mapping Software

- HiSat2 (https://ccb.jhu.edu/software/hisat2/)
- Star (http://code.google.com/p/rna-star/)
- Tophat (http://tophat.cbcb.umd.edu/)

Class activity
Script Submission for PE

HISAT2_example2