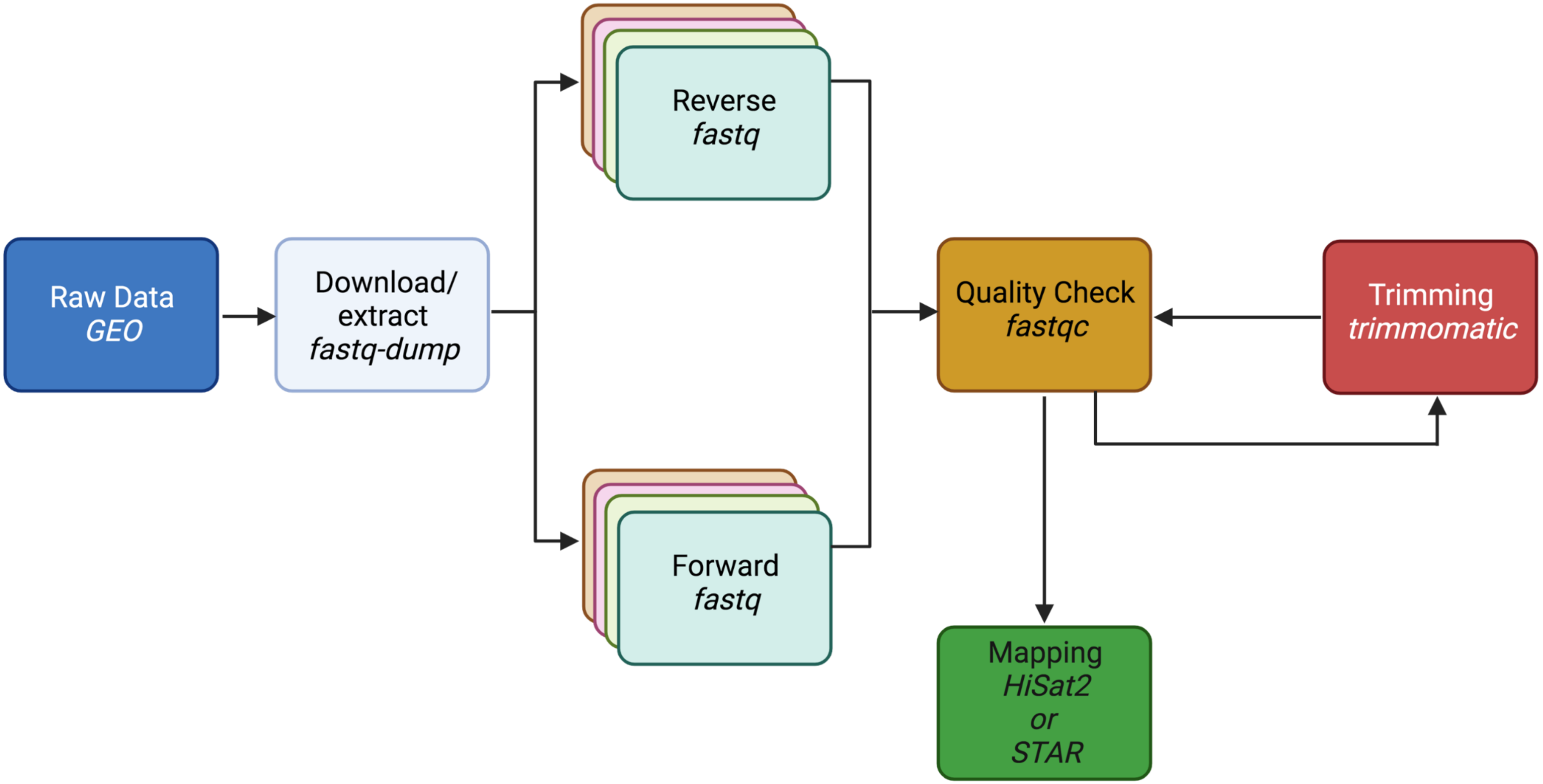
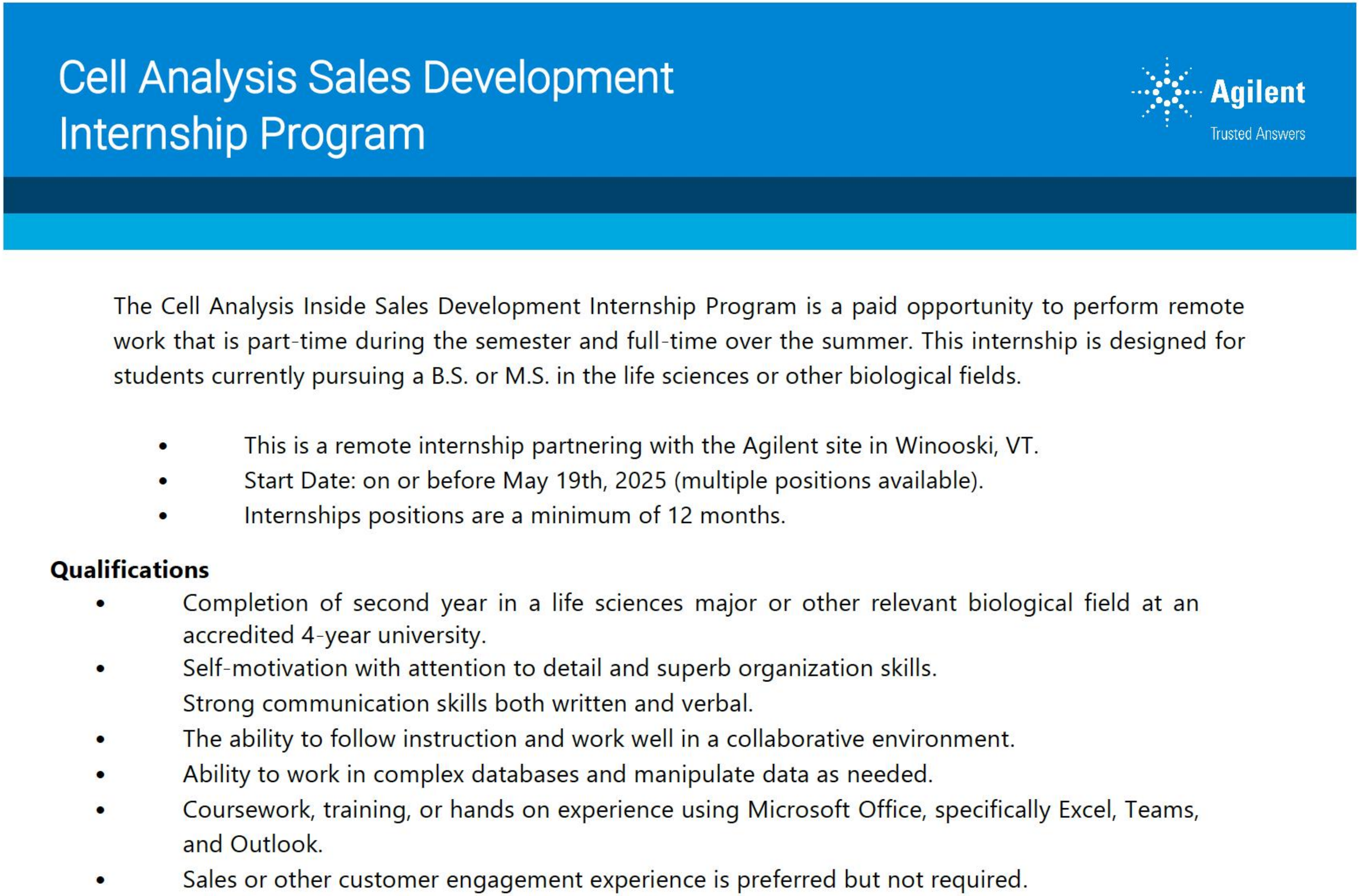


Read Mapping
February 19, 2025



Cell Analysis Sales Development Internship Program

Agilent
Trusted Answers

The Cell Analysis Inside Sales Development Internship Program is a paid opportunity to perform remote work that is part-time during the semester and full-time over the summer. This internship is designed for students currently pursuing a B.S. or M.S. in the life sciences or other biological fields.

- This is a remote internship partnering with the Agilent site in Winooski, VT.
- Start Date: on or before May 19th, 2025 (multiple positions available).
- Internships positions are a minimum of 12 months.

Qualifications

- Completion of second year in a life sciences major or other relevant biological field at an accredited 4-year university.
- Self-motivation with attention to detail and superb organization skills.
  Strong communication skills both written and verbal.
- The ability to follow instruction and work well in a collaborative environment.
- Ability to work in complex databases and manipulate data as needed.
- Coursework, training, or hands on experience using Microsoft Office, specifically Excel, Teams, and Outlook.
- Sales or other customer engagement experience is preferred but not required.

https://agilent.wd5.myworkdayjobs.com/en-US/Agilent_Student_Careers/details/Cell-Analysis-Sales-Development-Intern_4030427

Email Andrea.Lee@Agilent.com
Internship positions will be posted at **https://careers.agilent.com**

Learning objectives

- Describe the types of data formats encountered during alignment
- Identify challenges associated with read alignment and understand strategies to address them
- Explain the importance of genome indexing and outline the steps to perform it
- Explore the features of the splice-aware aligner HISAT2

Outline

- Class Activity #1 = HISAT2_exercise = 10 minutes

- Lecture for ~20 mins

- Class Activity #2 = indexed_genomes_example = 10 minutes

- Lecture for ~5 minutes

- Class Activity #3 = HISAT2_modify = 20 minutes

Class activity #1
Script Submission

HISAT2_example

General Bioinformatic Workflow

1. Experimental Design
   - What scientific research question am I asking?
2. Sample Preparation
   - Sample Prep
   - Library Prep
3. Sequencing
   - Technology/Platform
4. Data Analysis (Computation)

> **"You have to go back to the beginning to understand the end"**

What Question am I asking?

- What genes are differentially expressed between two conditions?
- Does this gene undergo alternative splicing?
- Is there a fusion gene present in this dataset or other structural variants, such as large deletions?
- Can we identify novel isoforms or unannotated genes in a newly sequenced organism?

All of these questions will have a slightly different approach!

General Bioinformatic Workflow

1. Experimental Design
   - What scientific research question am I asking?
2. Sample Preparation
   - Sample Prep
   - Library Prep
3. Sequencing
   - Technology/Platform
4. Data Analysis (Computation)

> **"You have to go back to the beginning to understand the end"**

The Question will guide the sample/sequence prep

- Read Depth
  - More depth is needed for lowly expressed genes
- Read Length
  - The longer the length the more likely to map uniquely
  - PE helps in mapping and junctions
- Stranded Protocols
  - Aids in identifying reads that map uniquely
- Biological Replicates
  - Aids in detecting novel genes or alternative isoforms

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%

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%

FASTQC will aid in identifying if minimum requirements are met

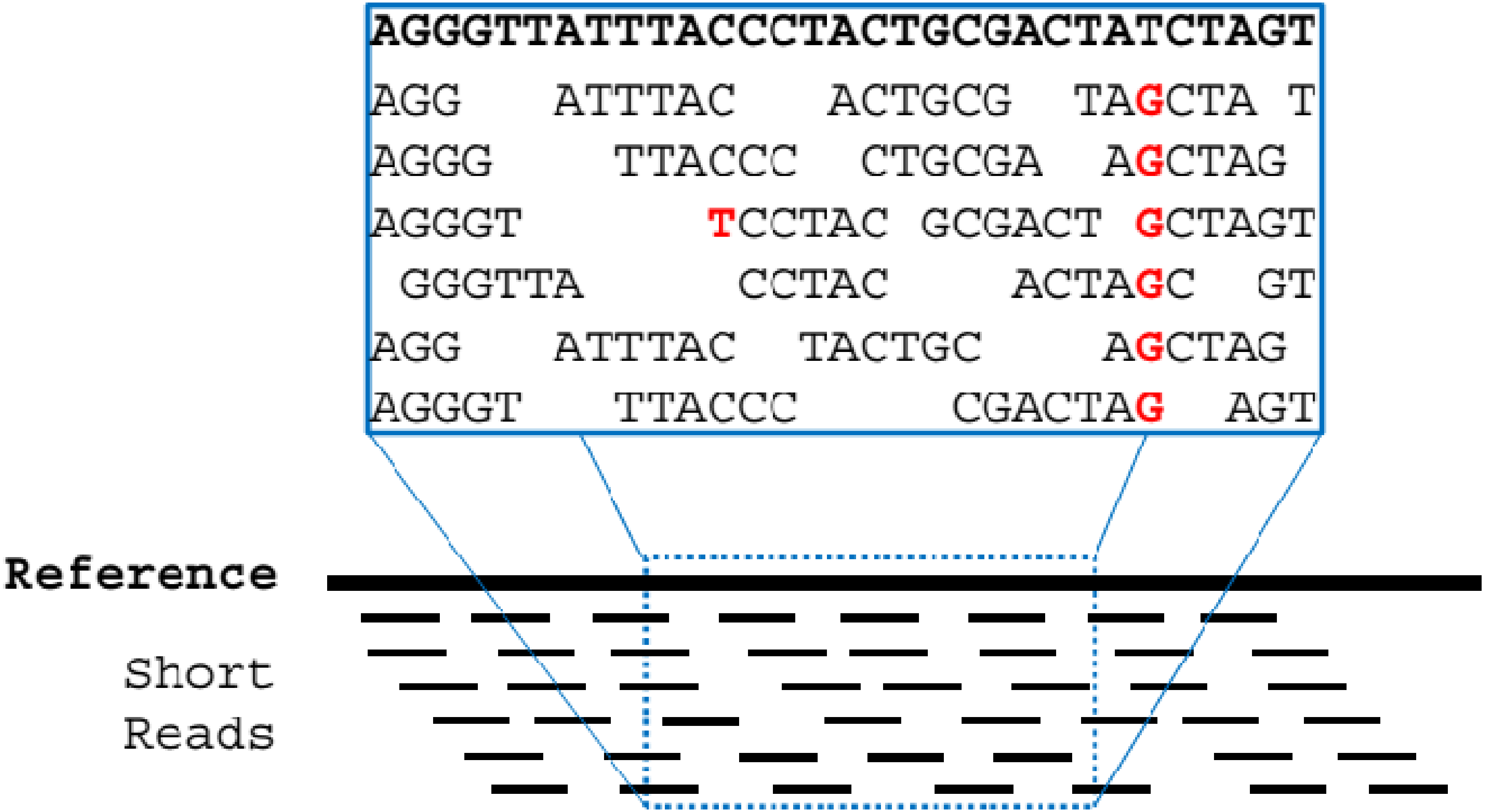
	Question 1: **Which genes are differentially expressed?**	**Question 2:** **Are different splicing isoforms expressed?**	**Question 3:** **Are you interested in non-coding RNAs? Novel transcripts?**
Reads	> 10M	> 25-50M	**> 25-50M**
Biological replicates	3 replicates	> 3 replicates	> 3 replicates
SE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

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**

Alignment

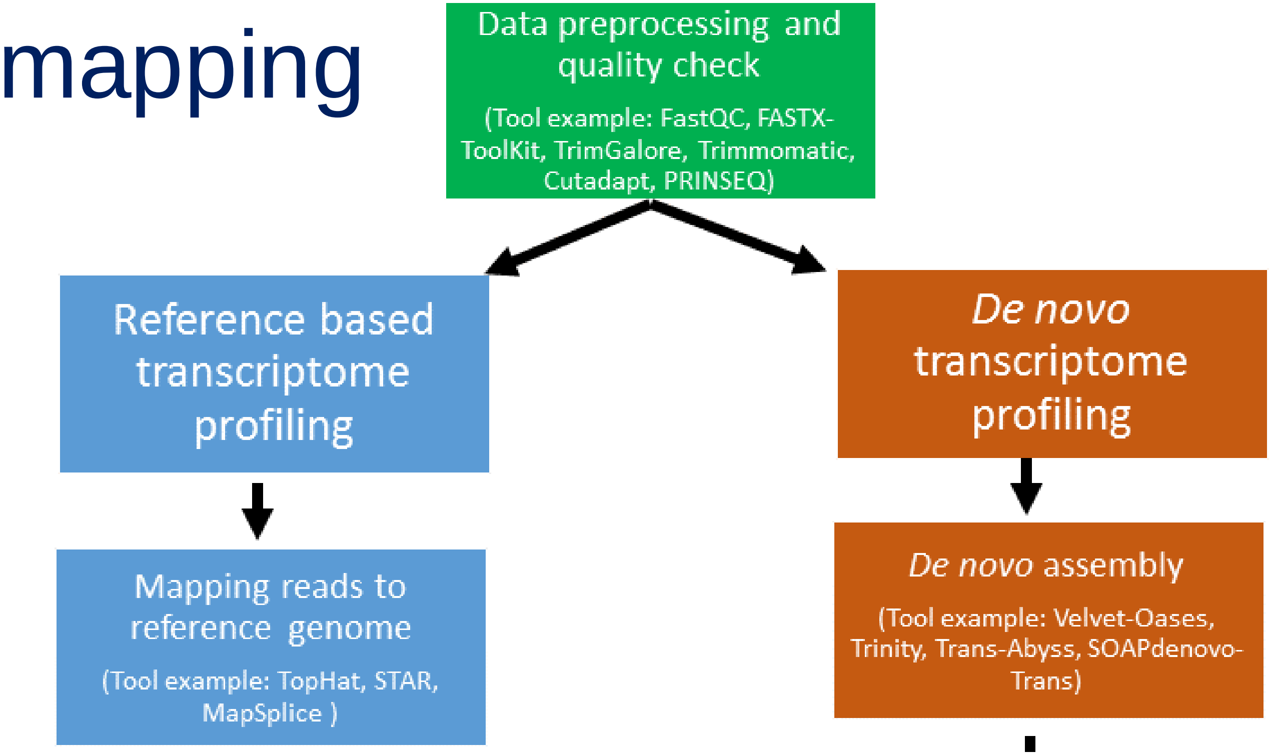
Read alignment / “mapping”



```
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
```

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

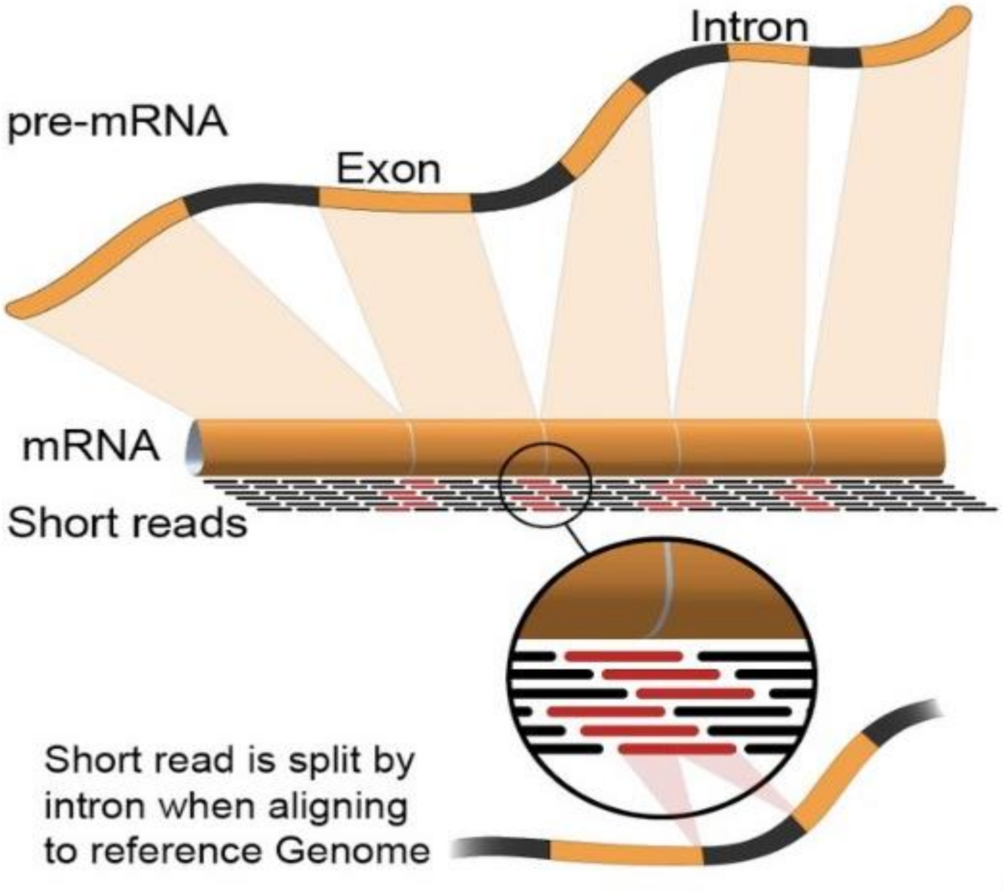
RNAseq Mapping Challenges/Considerations

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*

RNAseq Mapping Challenges/Considerations

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


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

Introns
Exons

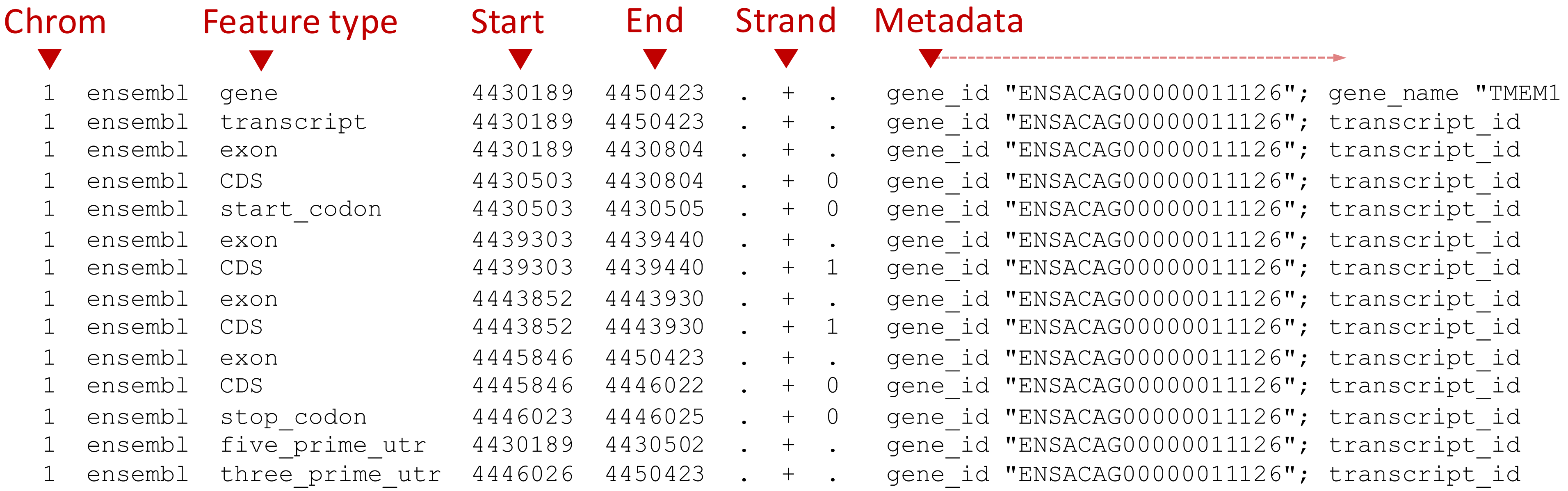
We have to account for reads that may be split by potentially thousands of bases of intronic sequences

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

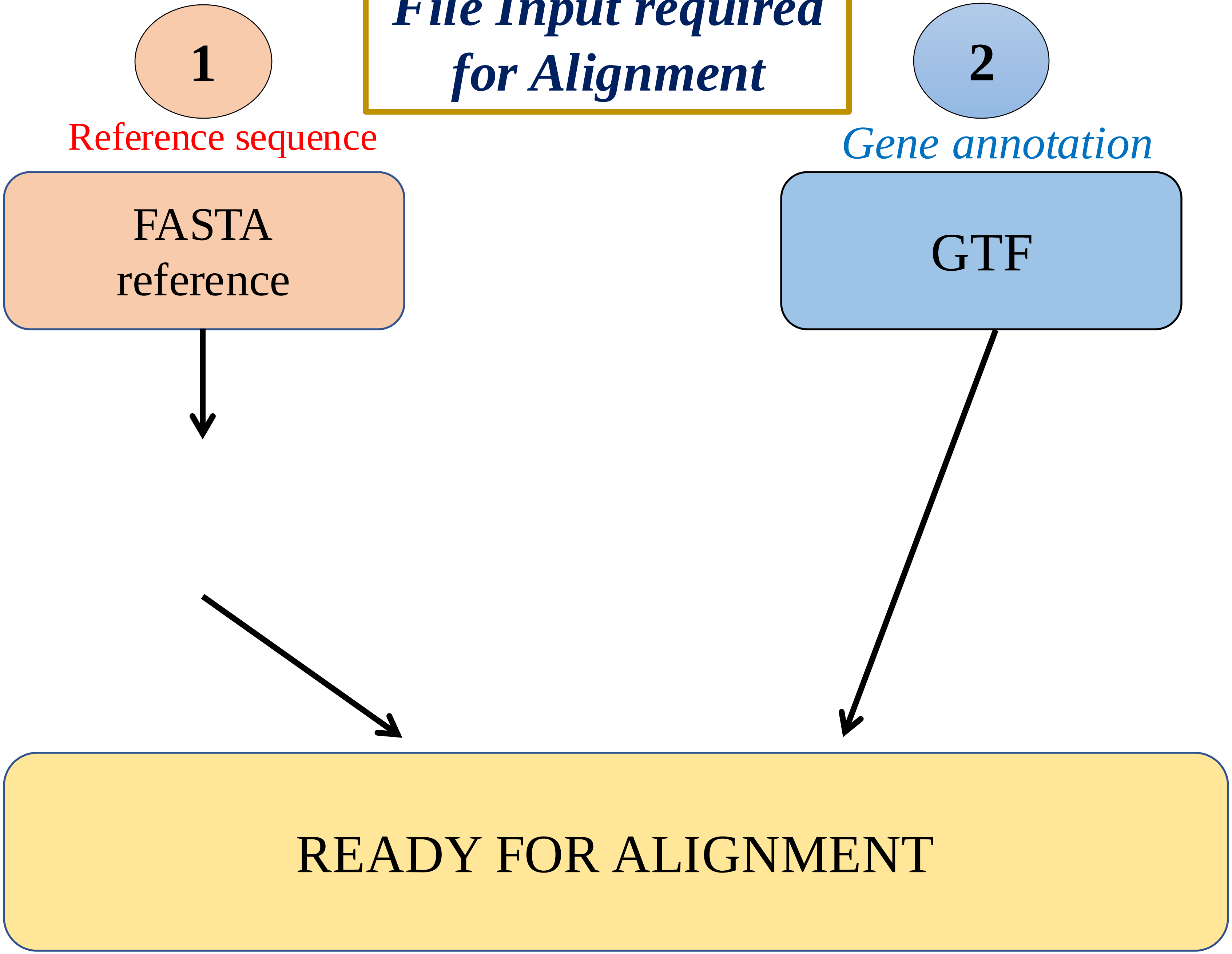
```
1  ensembl  gene             4430189  4450423  .  +  .   gene_id "ENSACAG00000011126"; gene_name "TMEM1
1  ensembl  transcript       4430189  4450423  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
1  ensembl  exon             4430189  4430804  .  +  .   gene_id "ENSACAG00000011126"; transcript_id
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

File Inputs required for Alignment

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

File Input required for Alignment



1

Reference sequence

FASTA reference

2

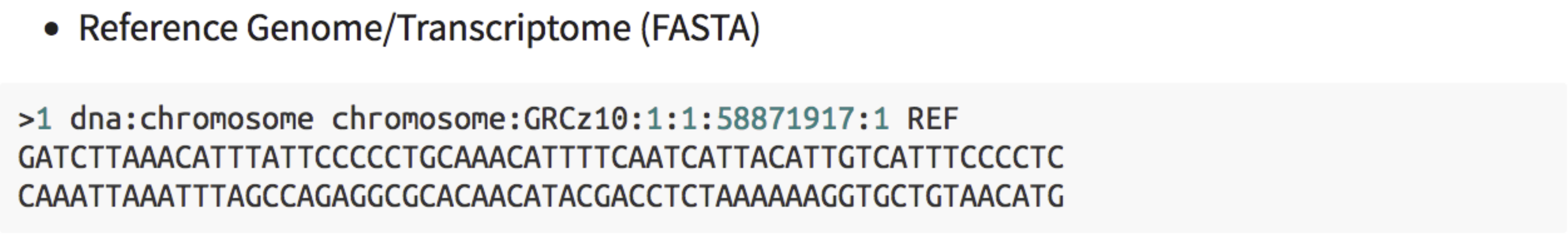
Gene annotation

GTF

READY FOR ALIGNMENT

Reference Genome

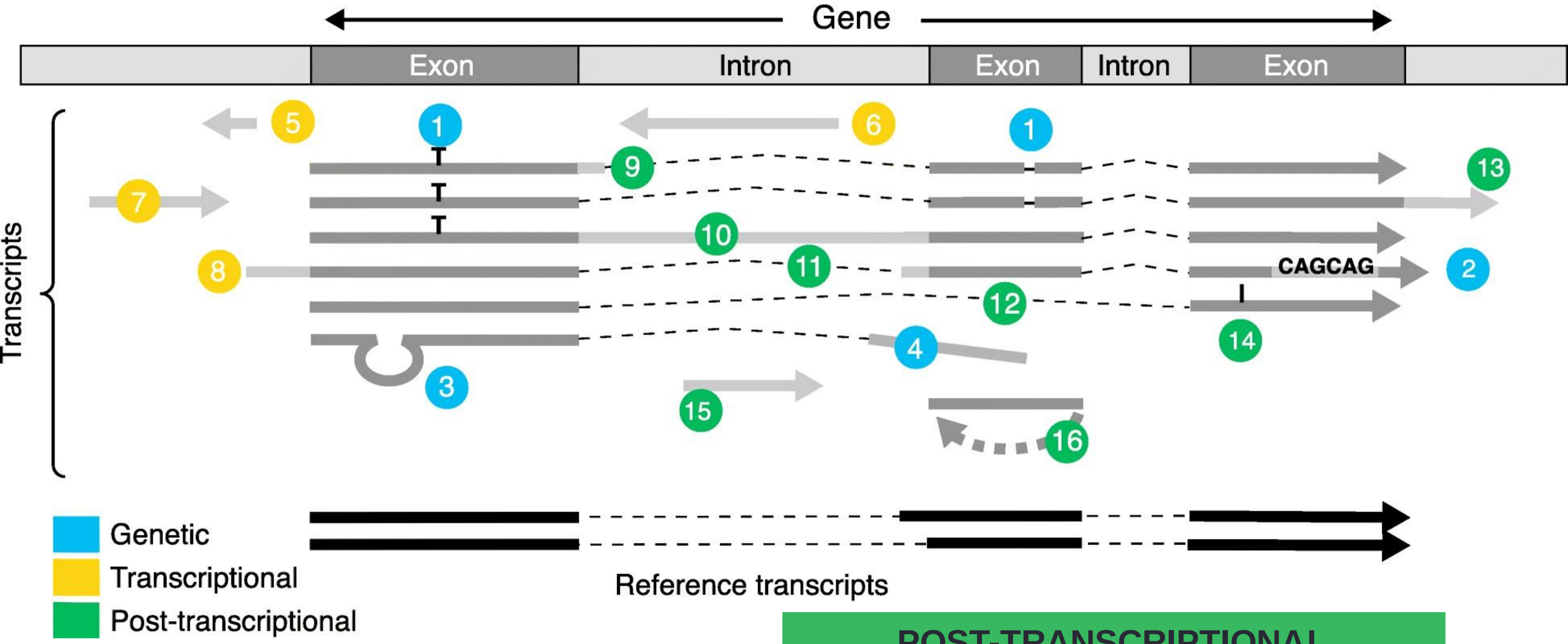
- The reference genome are usually stored in a plain text **FASTA file**



- Reference Genome/Transcriptome (FASTA)

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

Most individual RNA variations do not find their way into the reference sequences



GENETIC

1 SNP or short InDel
2 microsatellite variation
3 transposition
4 gene fusion

TRANSCRIPTIONAL

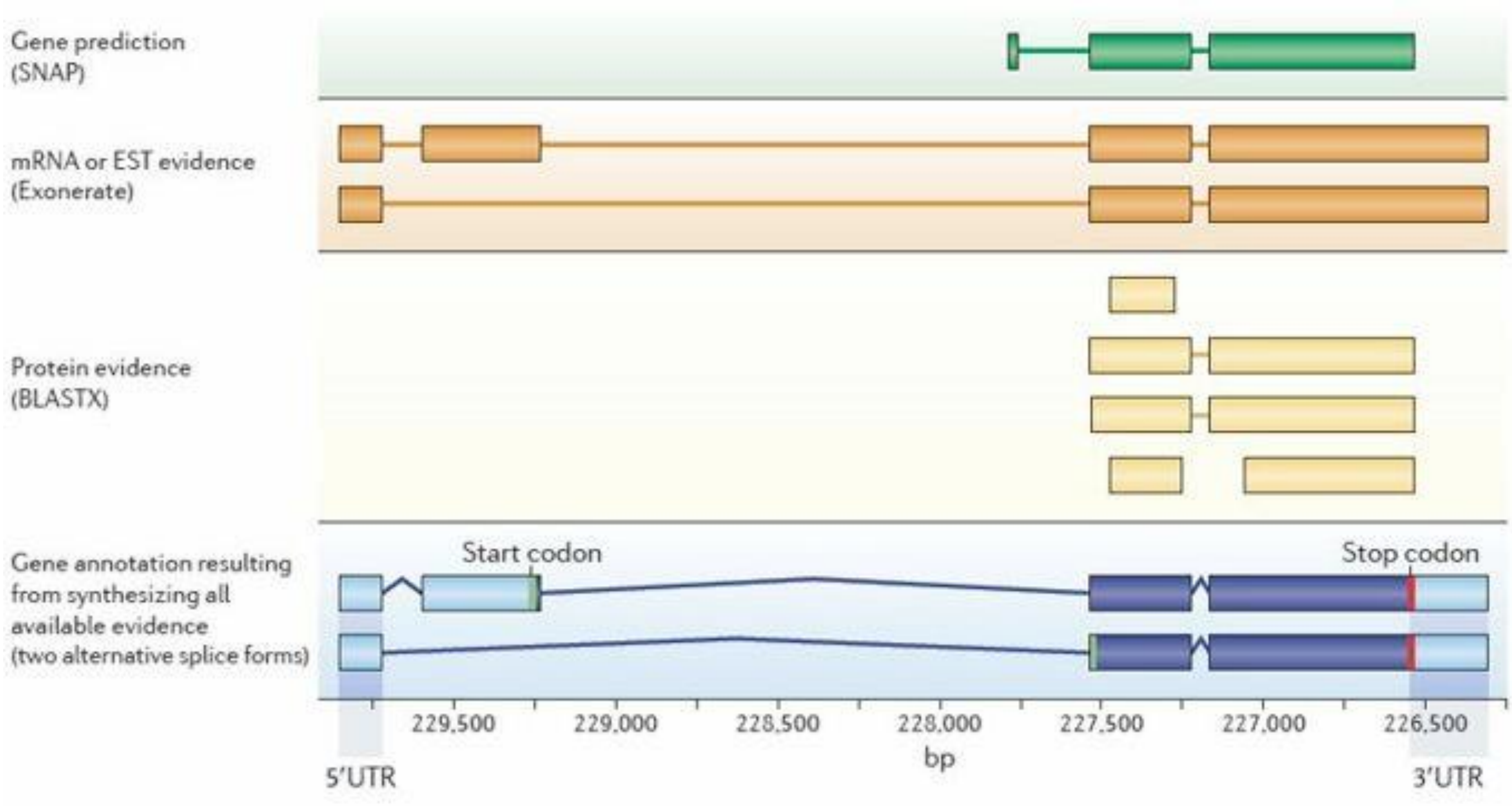
5 bidirectional TSS
6 antisense transcript
7 enhancer RNA
8 alternative TSS

POST-TRANSCRIPTIONAL

9 alternative 5' splice site (SS)
10 alternative 3' SS
11 alternative 3' SS
12 skipped exon

13 alternative poly(A) site
14 editing and modification
15 processed pre-mi/ snoRNA
16 circular RNA

Gene annotation



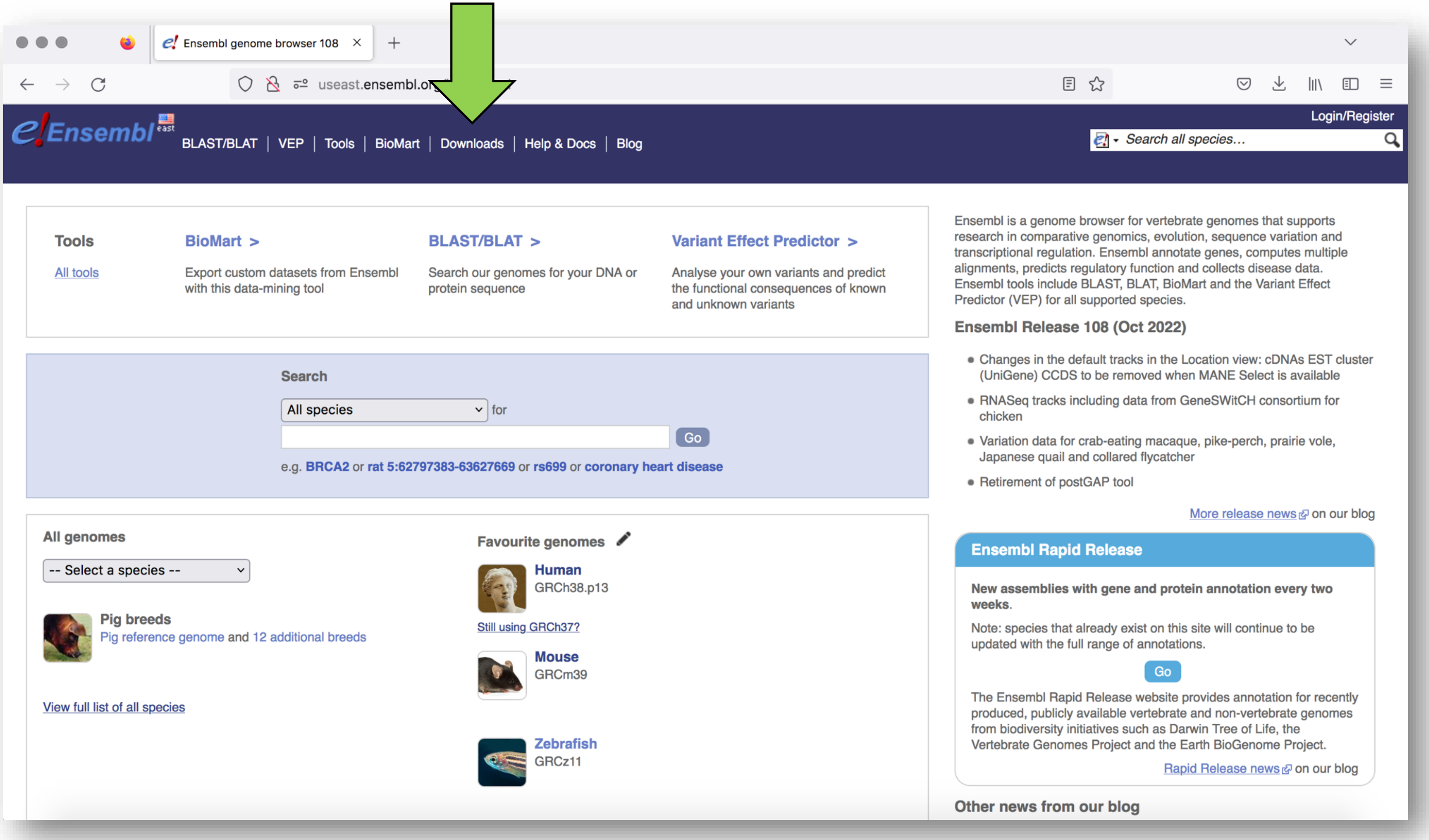
Gene annotations generally include UTRs, alternative splice isoforms and have attributes such as evidence trails.

Where can I 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)

Ensembl



Export custom datasets from Ensembl with this data-mining tool

Search our genomes for your DNA or protein sequence

Analyse your own variants and predict the functional consequences of known and unknown variants

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 108 (Oct 2022)

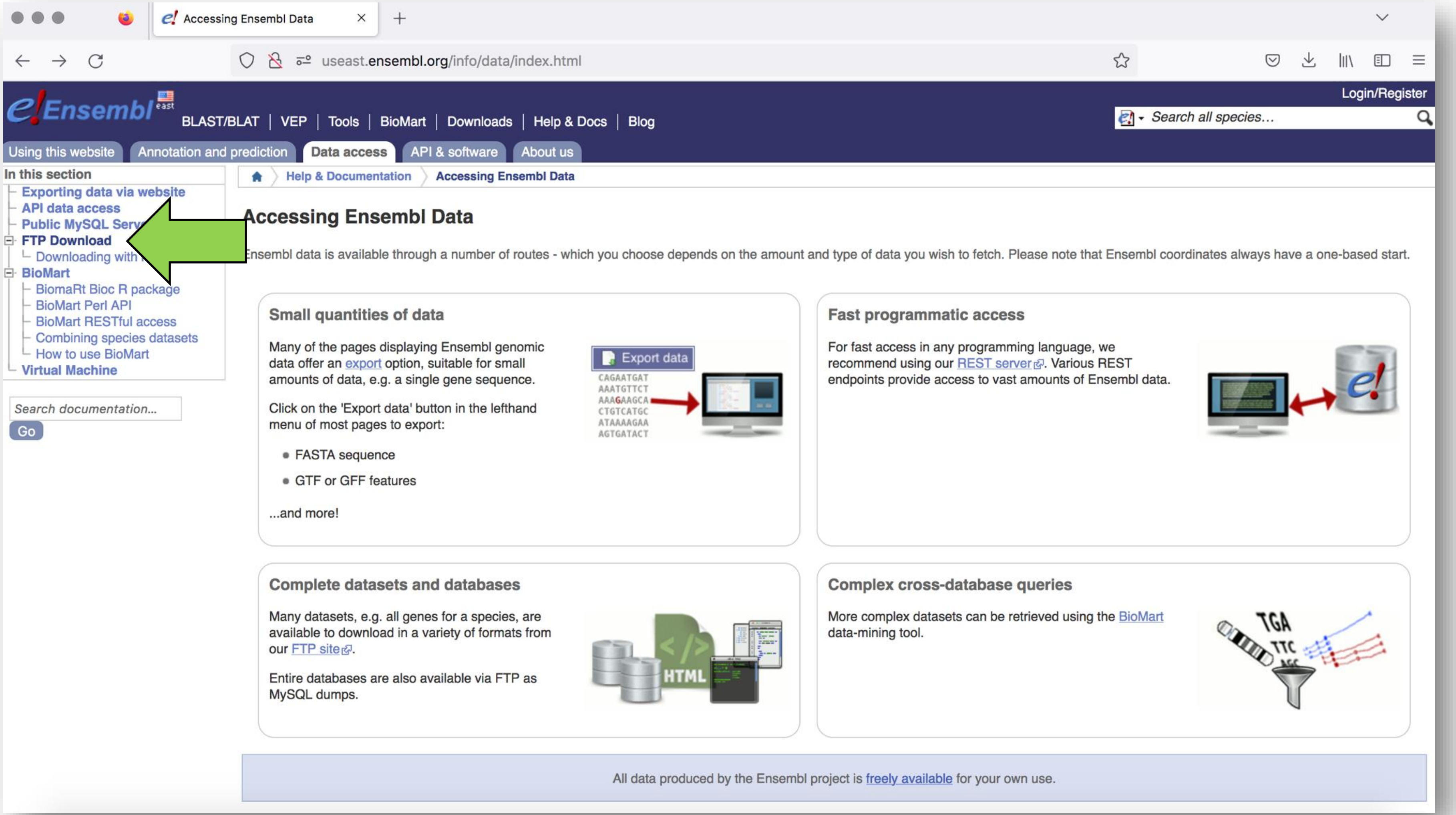
- Changes in the default tracks in the Location view: cDNAs EST cluster (UniGene) CCDS to be removed when MANE Select is available
- RNASeq tracks including data from GeneSWitCH consortium for chicken
- Variation data for crab-eating macaque, pike-perch, prairie vole, Japanese quail and collared flycatcher
- Retirement of postGAP tool

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

The Ensembl Rapid Release website provides annotation for recently produced, publicly available vertebrate and non-vertebrate genomes from biodiversity initiatives such as Darwin Tree of Life, the Vertebrate Genomes Project and the Earth BioGenome Project.

Ensembl



Accessing Ensembl Data

Ensembl data is available through a number of routes - which you choose depends on the amount and type of data you wish to fetch. Please note that Ensembl coordinates always have a one-based start.

Small quantities of data

Many of the pages displaying Ensembl genomic data offer an export option, suitable for small amounts of data, e.g. a single gene sequence.

Click on the 'Export data' button in the lefthand menu of most pages to export:

- FASTA sequence
- GTF or GFF features

...and more!

Fast programmatic access

For fast access in any programming language, we recommend using our REST server. Various REST endpoints provide access to vast amounts of Ensembl data.

Complete datasets and databases

Many datasets, e.g. all genes for a species, are available to download in a variety of formats from our FTP site.

Entire databases are also available via FTP as MySQL dumps.

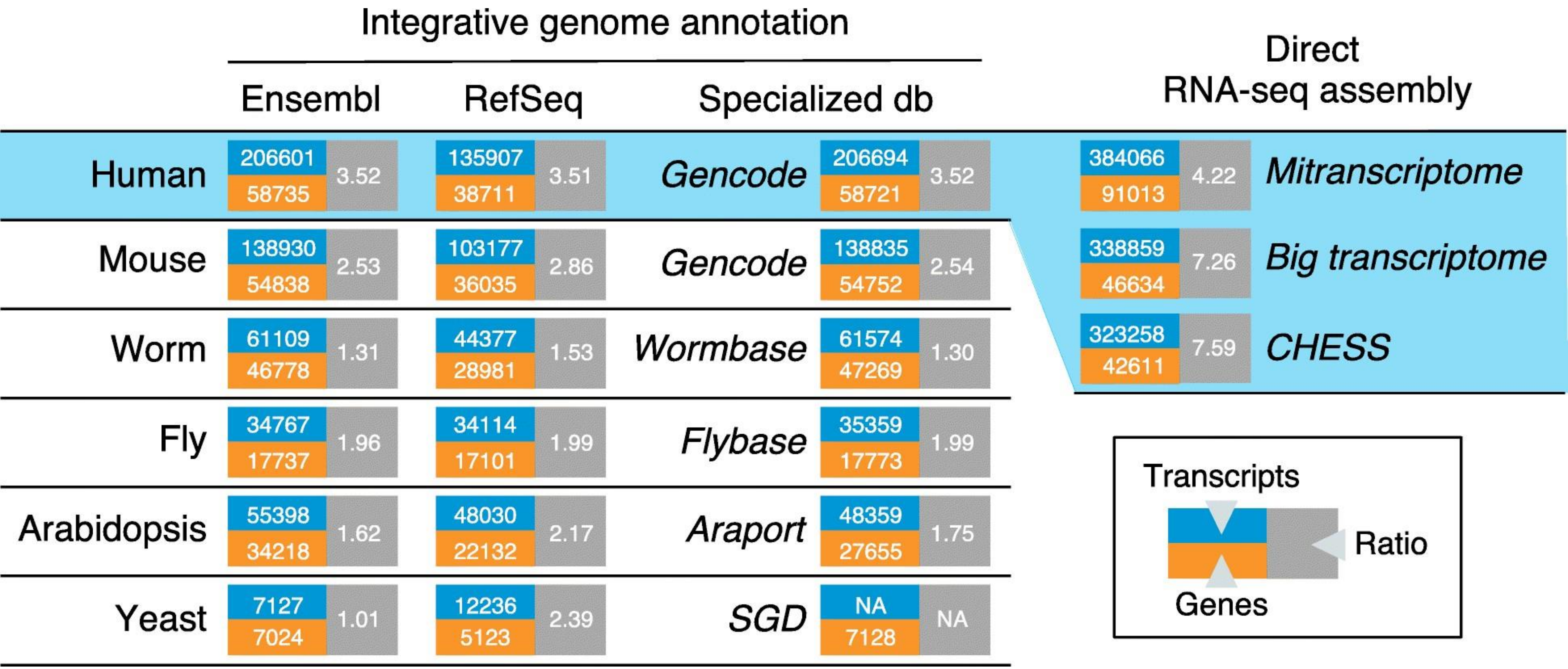
Complex cross-database queries

More complex datasets can be retrieved using the BioMart data-mining tool.

All data produced by the Ensembl project is freely available for your own use.

Good practical advice

- Always use the same biological database for all data files (FASTA + GTF)
- Always ensure you know exactly which version of a genome and annotation you are working with

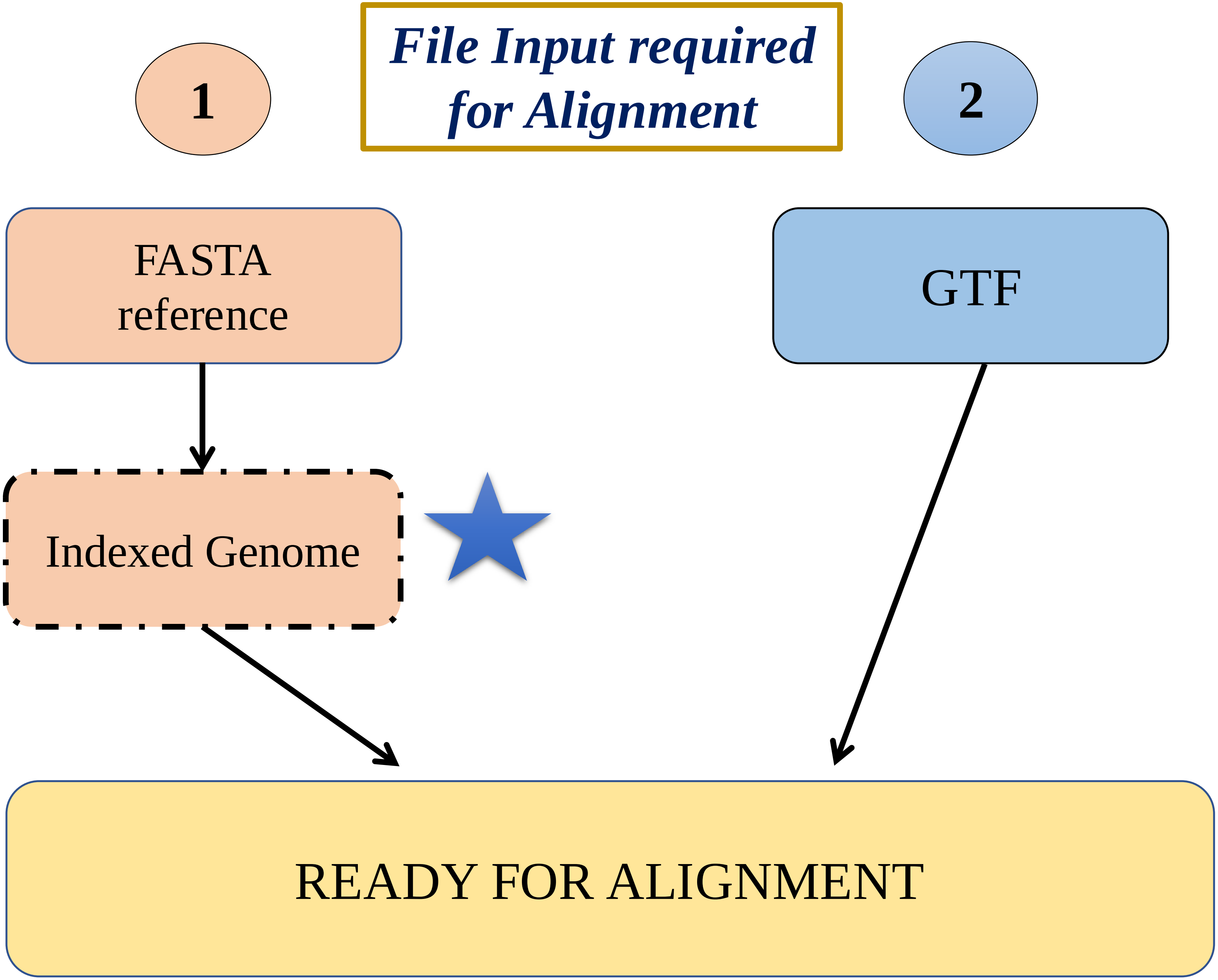


Integrative genome annotation

	Ensembl			RefSeq			Specialized db			
	Transcripts	Genes	Ratio	Transcripts	Genes	Ratio		Transcripts	Genes	Ratio
Human	206601	58735	3.52	135907	38711	3.51	*Gencode*	206694	58721	3.52
Mouse	138930	54838	2.53	103177	36035	2.86	*Gencode*	138835	54752	2.54
Worm	61109	46778	1.31	44377	28981	1.53	*Wormbase*	61574	47269	1.30
Fly	34767	17737	1.96	34114	17101	1.99	*Flybase*	35359	17773	1.99
Arabidopsis	55398	34218	1.62	48030	22132	2.17	*Araport*	48359	27655	1.75
Yeast	7127	7024	1.01	12236	5123	2.39	*SGD*	NA	7128	NA

Direct RNA-seq assembly

Transcripts	Genes	Ratio	
384066	91013	4.22	*Mitranscriptome*
338859	46634	7.26	*Big transcriptome*
323258	42611	7.59	*CHESS*



File Input required for Alignment

1

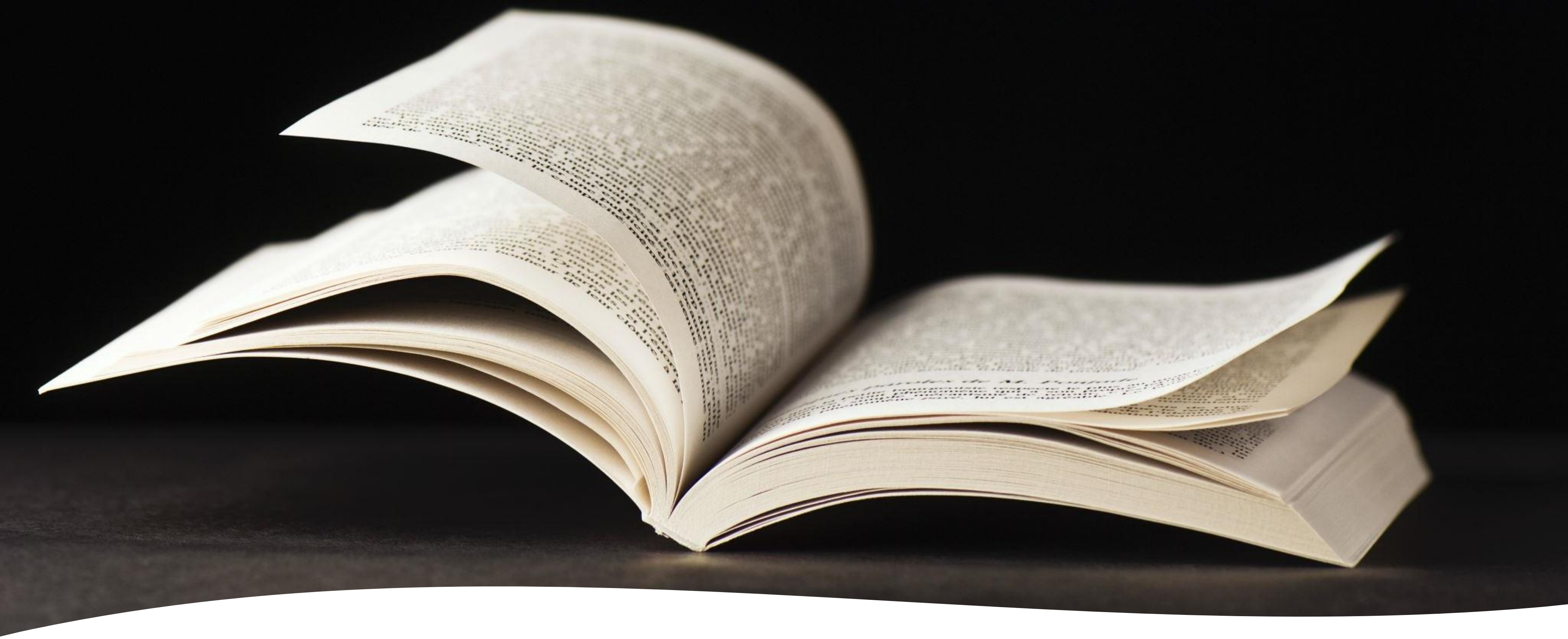
2

FASTA reference

GTF

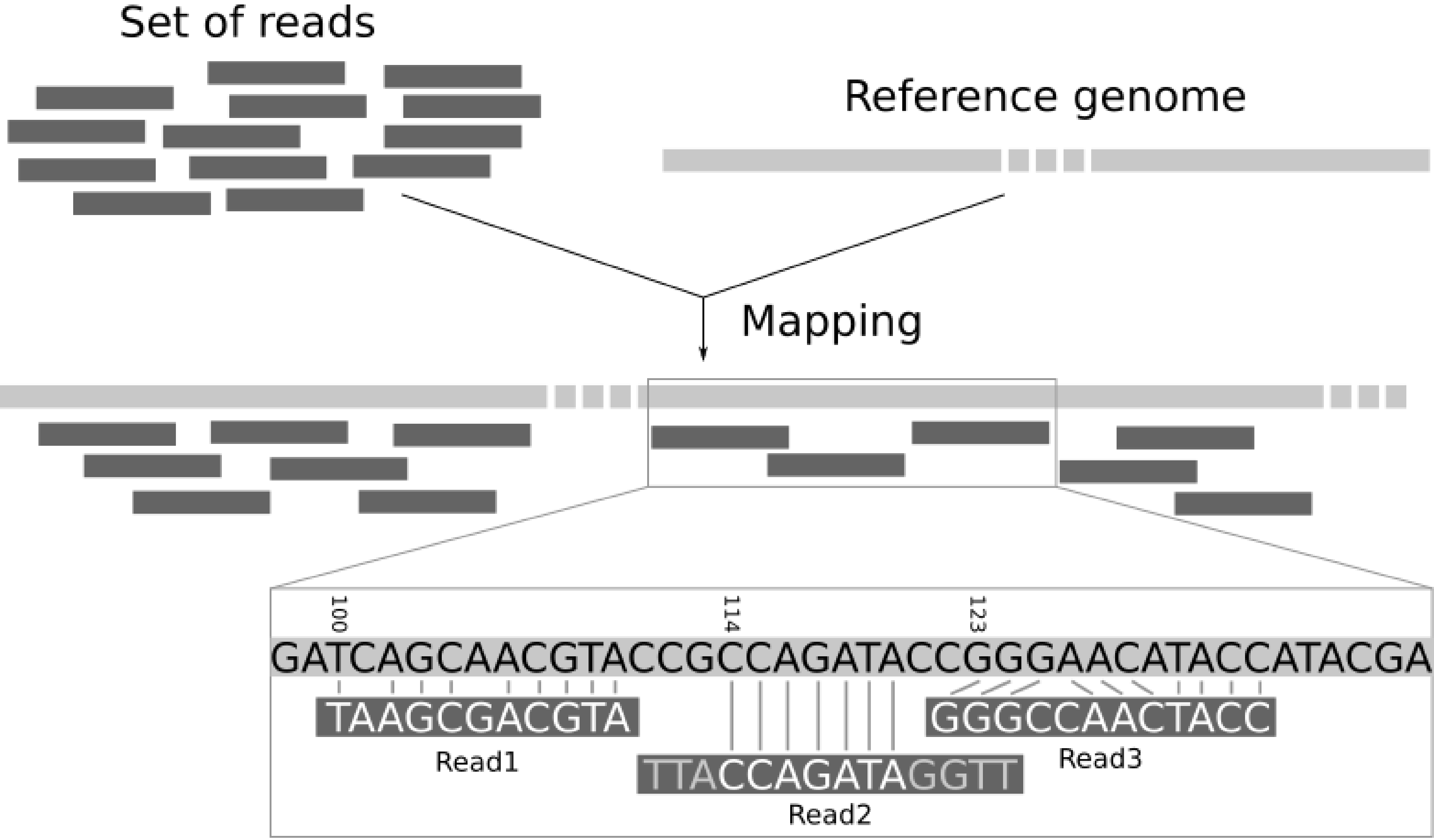
Indexed Genome

READY FOR ALIGNMENT

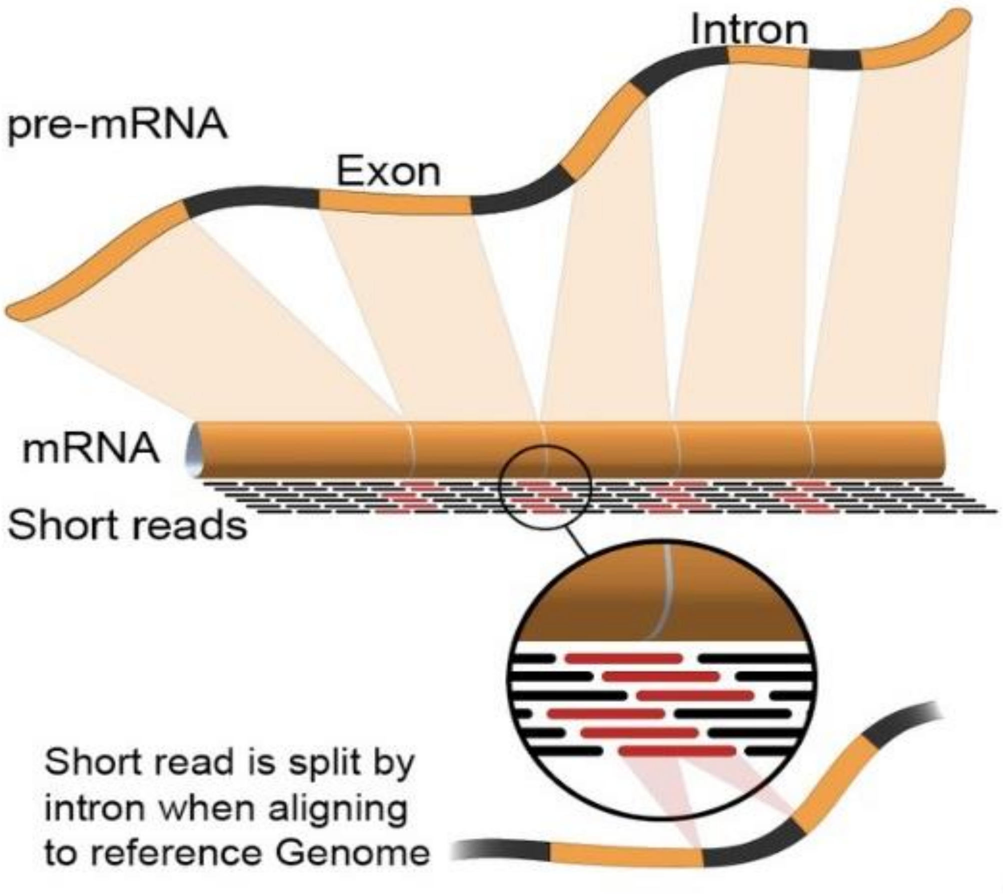


Indexing benefits

- 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

1 RNASeq Mapping Challenges: Intron/Exon Boundaries

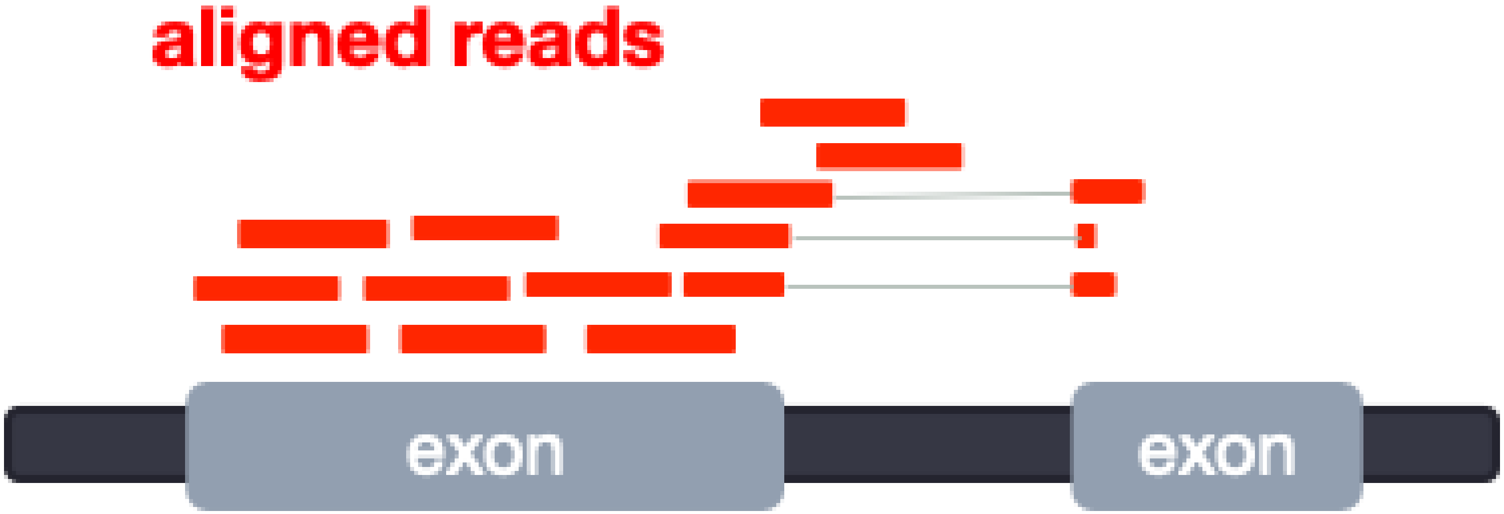

pre-mRNA
Exon
Intron
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

Introns
Exons

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

Two categories of reads:

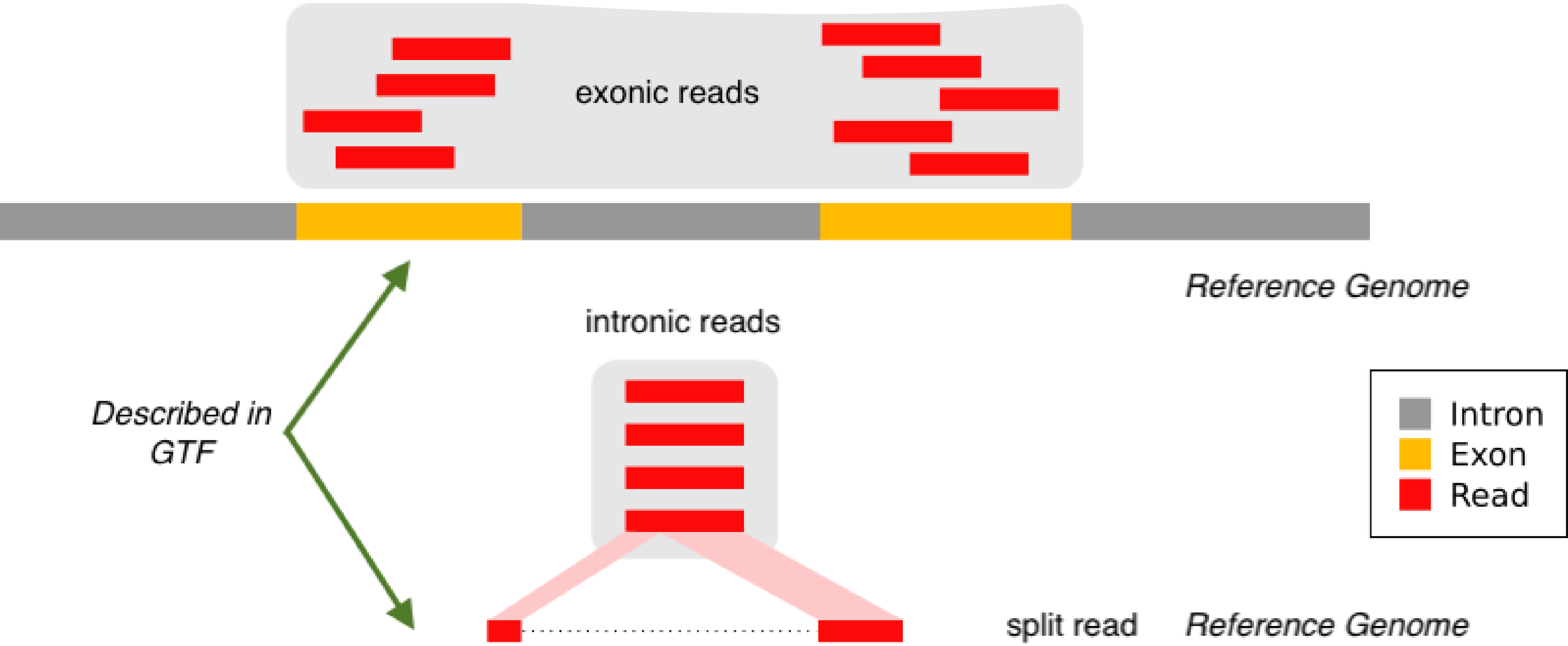
1. Reads that map entirely within exons
2. Reads that span two or more exons



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 the GTF annotation **Splice aware aligners rely heavily on annotations**

greatest downside: it can be resource-intensive!



Splice-aware aligners

HISAT2

STAR

TopHat2

RNA-Seq

Splice-unaware aligner

Bowtie2

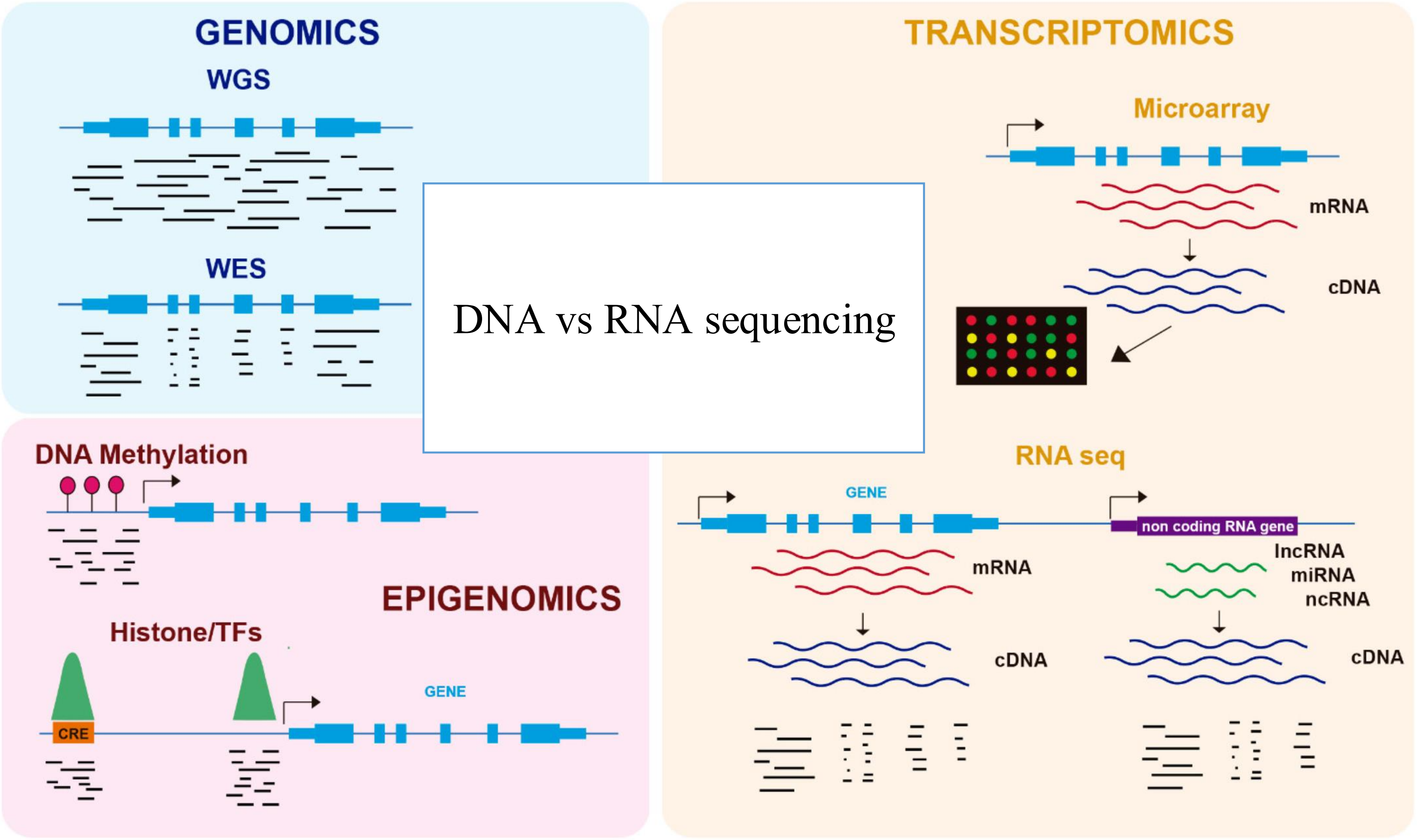
BWA

minimap2

?

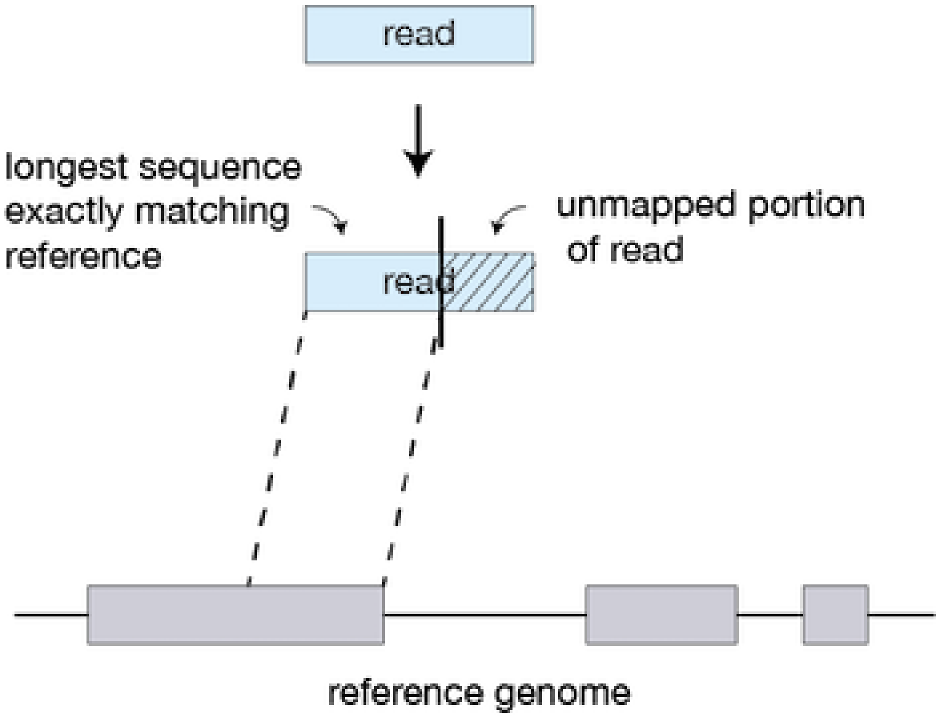
Question: For what applications is it okay to use a splice unaware aligner?

DNA vs RNA sequencing



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

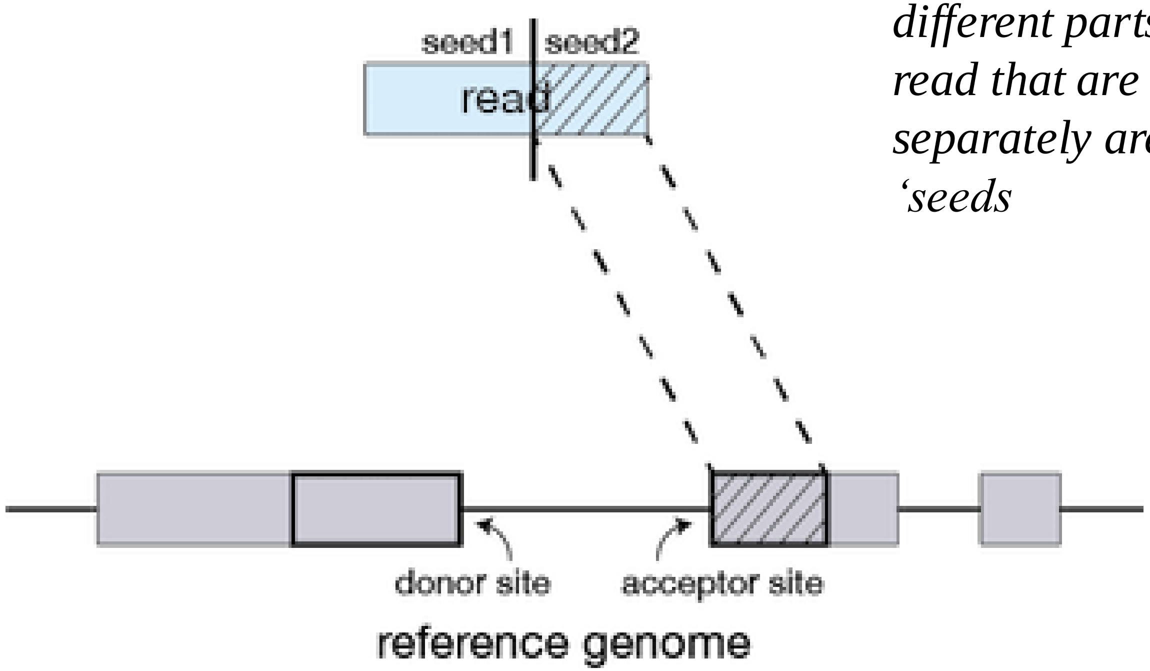
- STAR Alignment Strategy
    - Step 1: Seed Searching



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

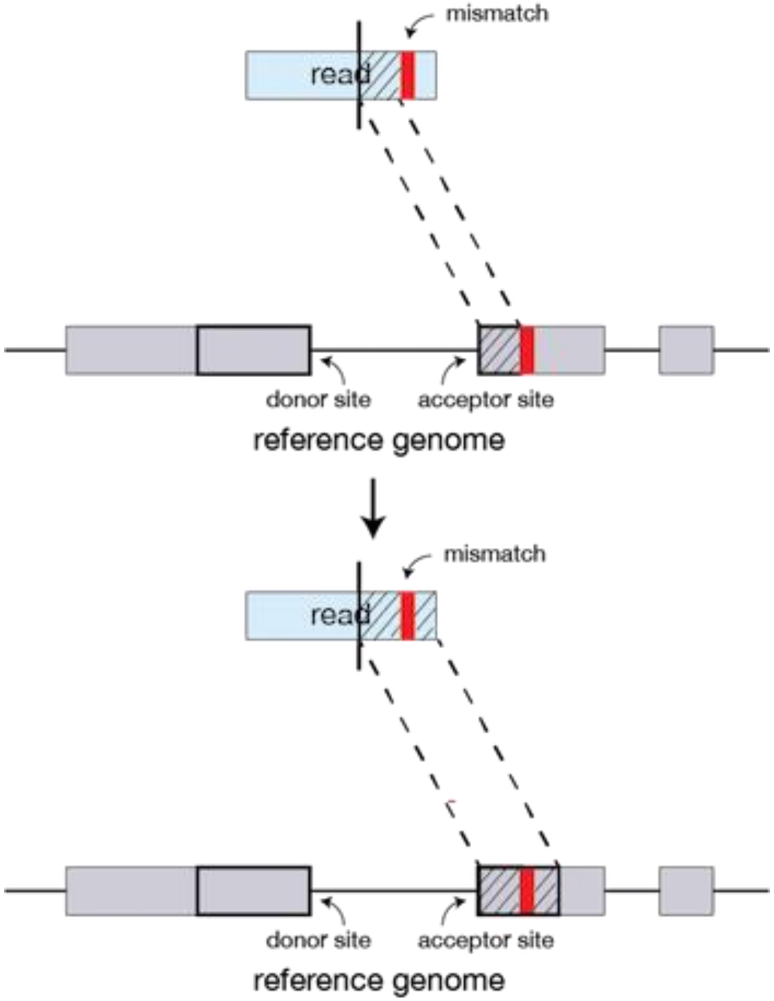


different parts of the read that are mapped separately are called ‘seeds

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

- STAR Alignment Strategy
  - Step 1: Seed Searching

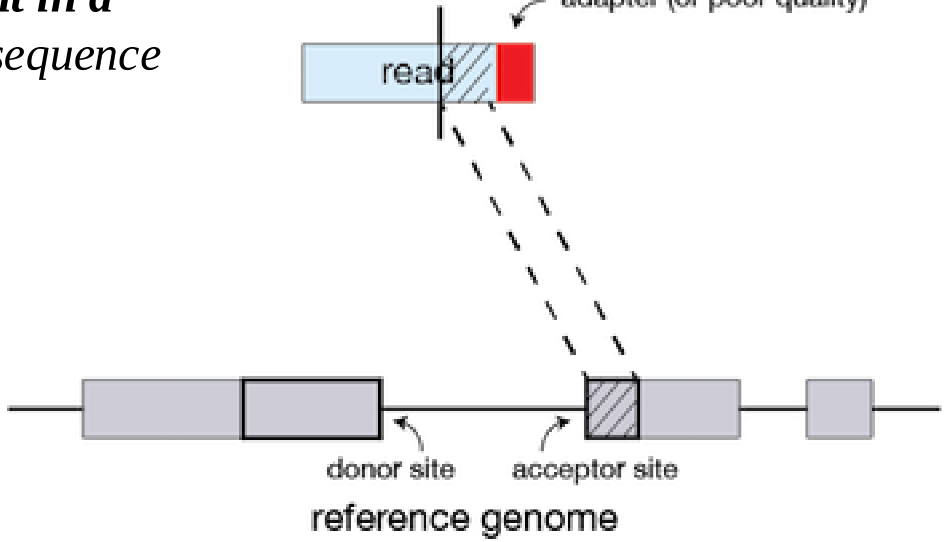
***If STAR does not find an exact matching sequence**, the MMPs will be extended.*



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.

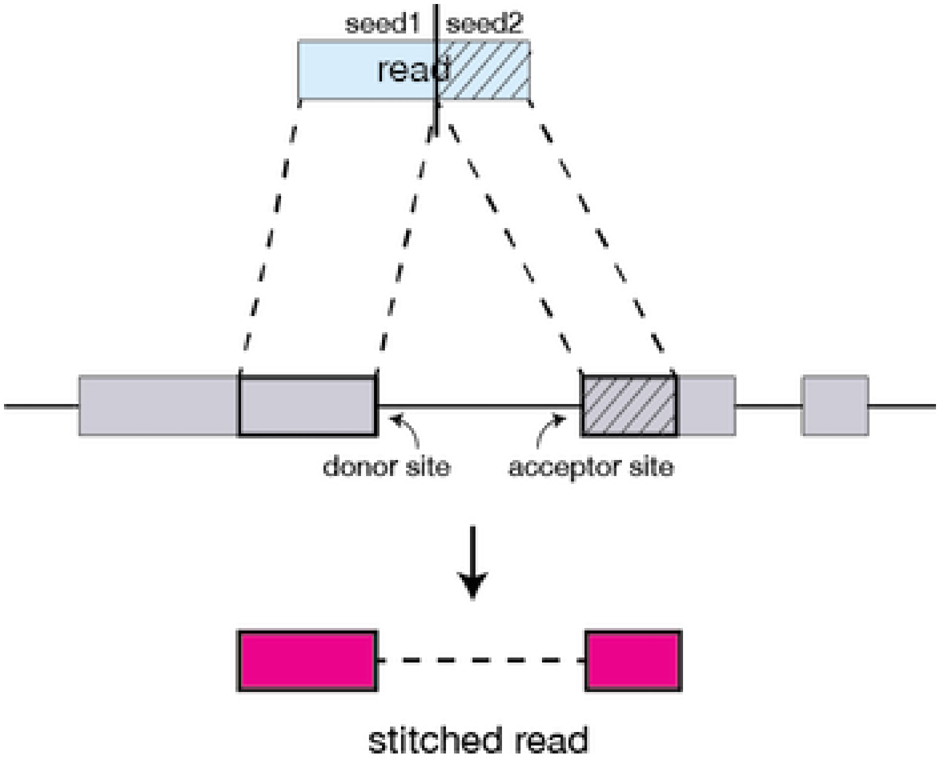


How does STAR (Spliced Transcripts Alignment to a Reference) work?

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

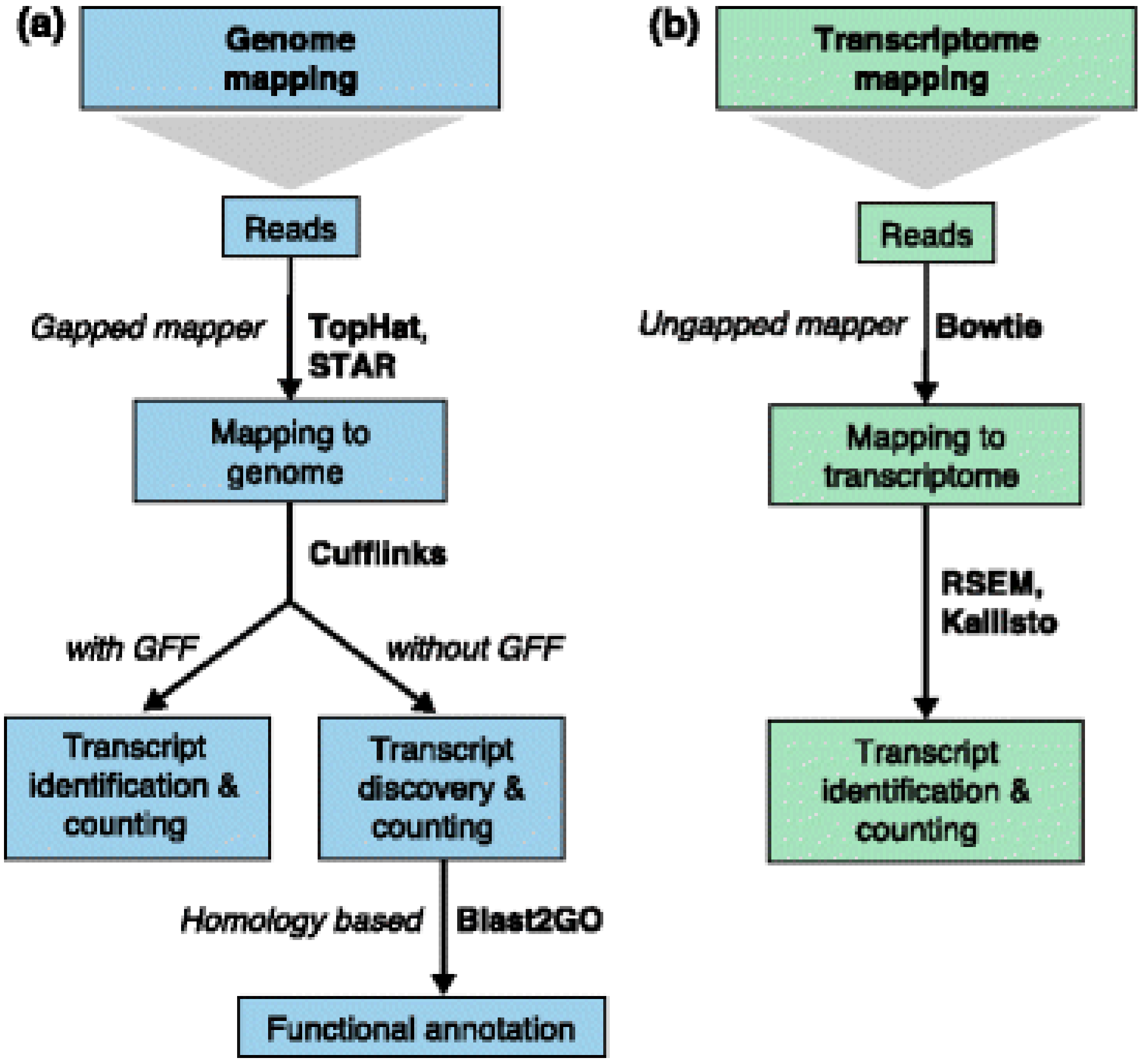
The seeds are stitched together to create a complete read based on the best alignment

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



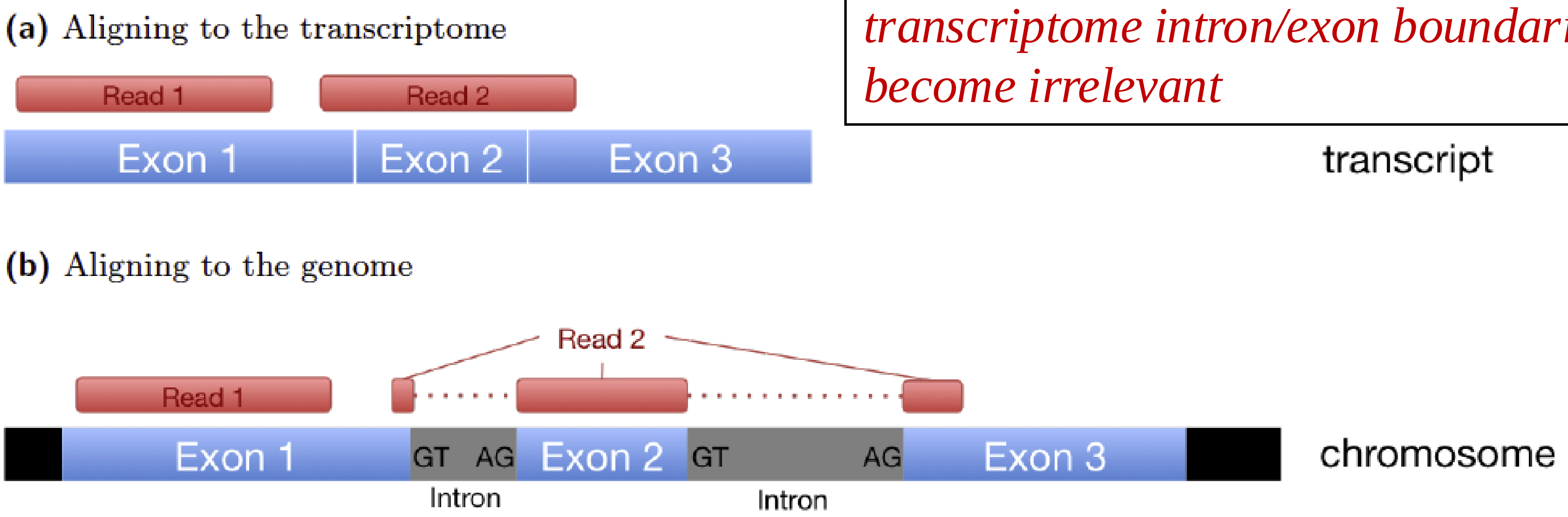
2

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

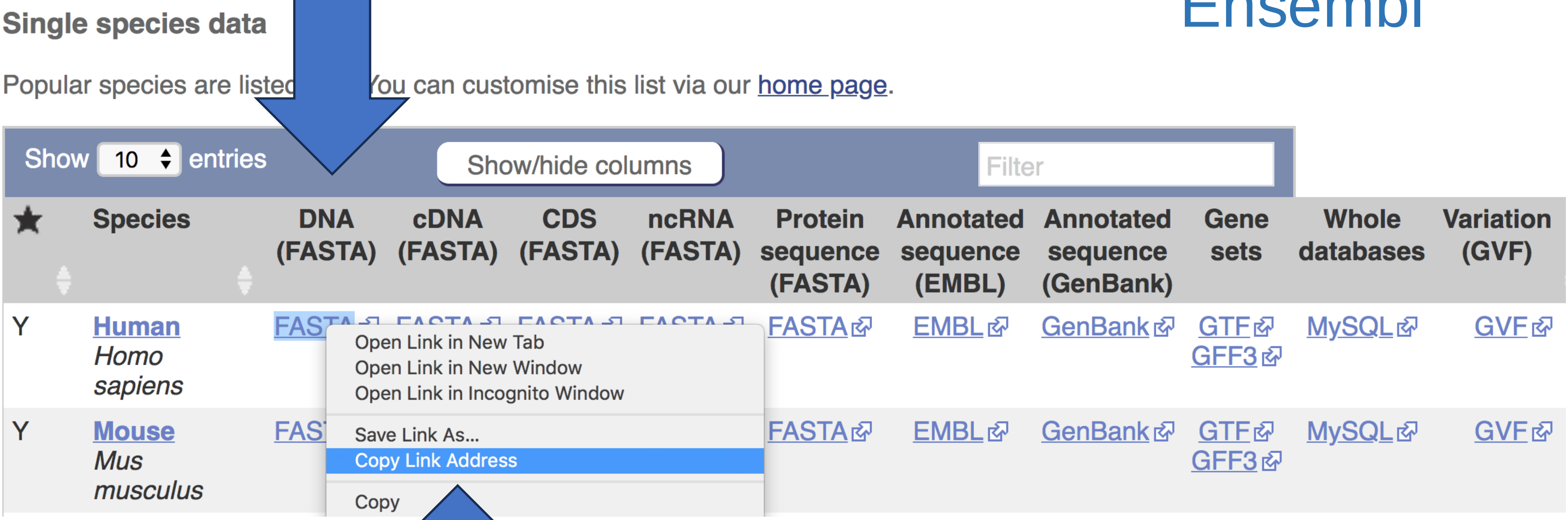


computationally a much harder task

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



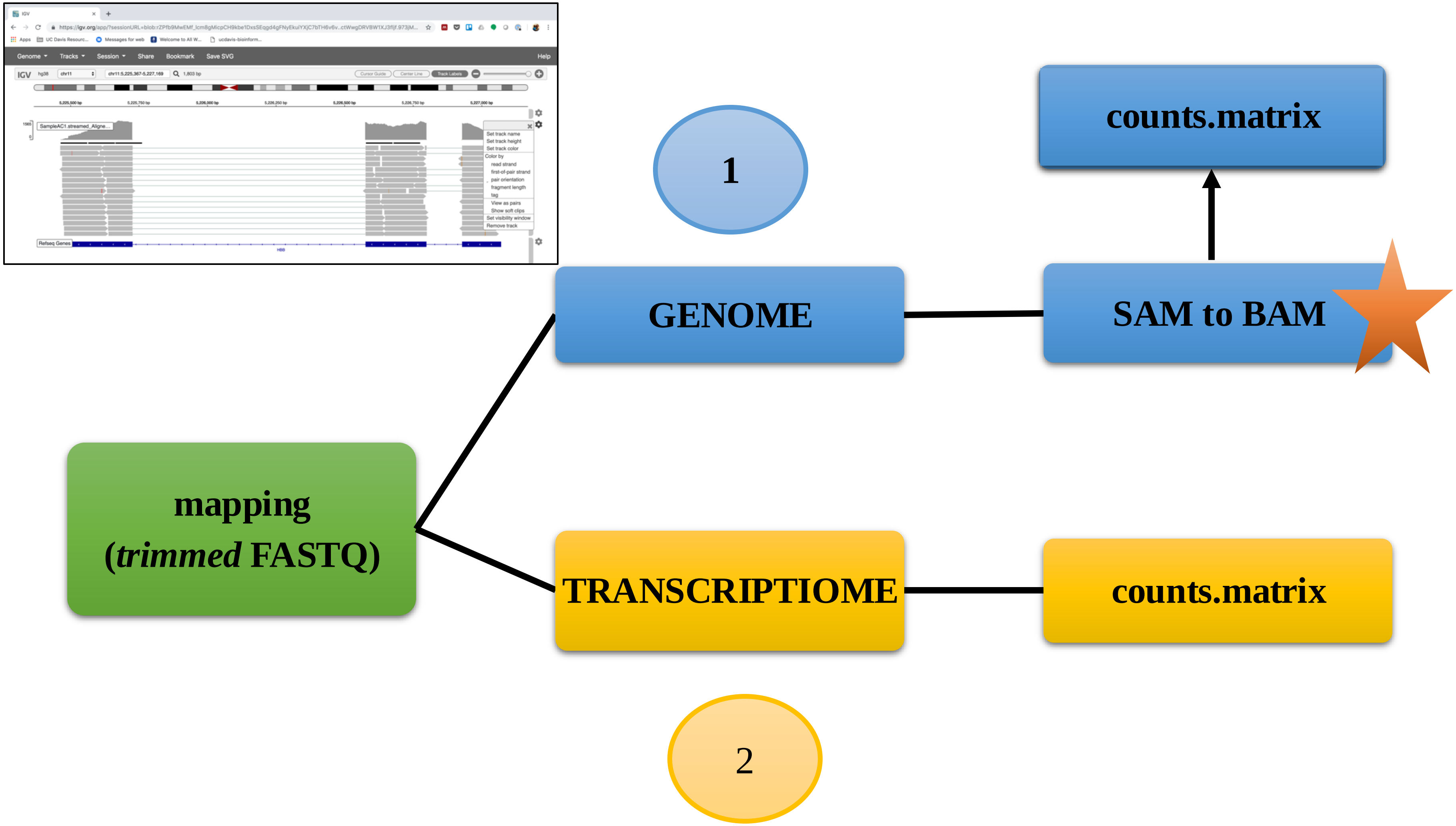
Single species data

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

★	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

Forgo intermediate files with transcriptome alignment



Forgo transcript discovery with transcriptome alignment

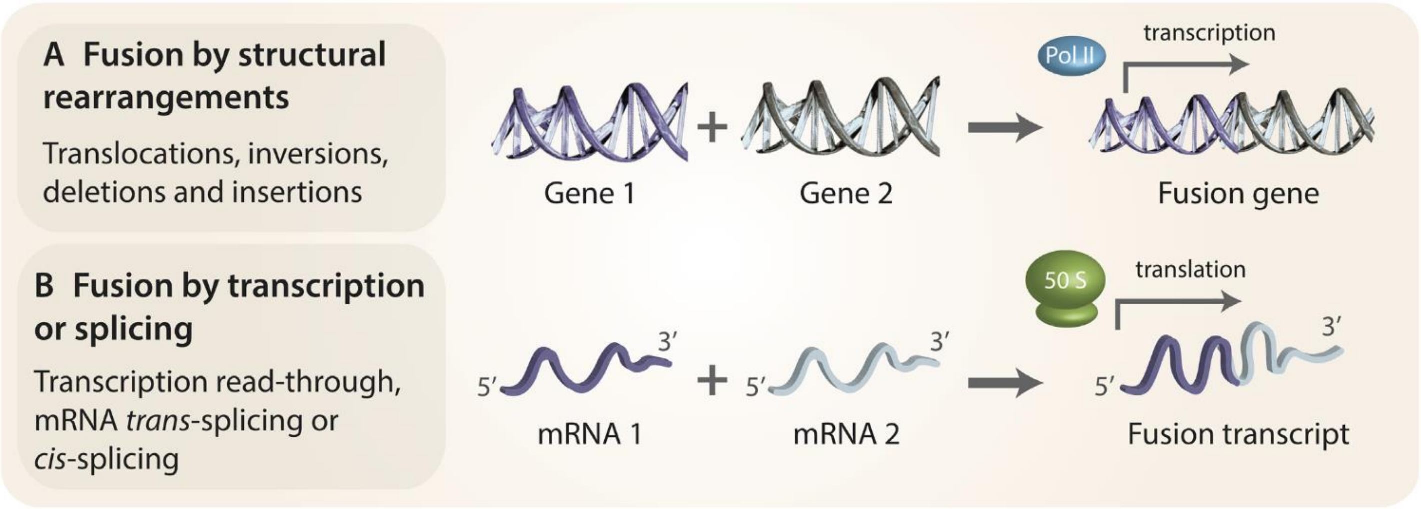
- Refers to allowing researchers to identify new splice variants or transcripts not previously annotated
- Transcriptome alignment is limited because it maps reads **only to known, annotated transcripts** rather than the full genome.

The input FASTA file only contains known protein-coding sequences

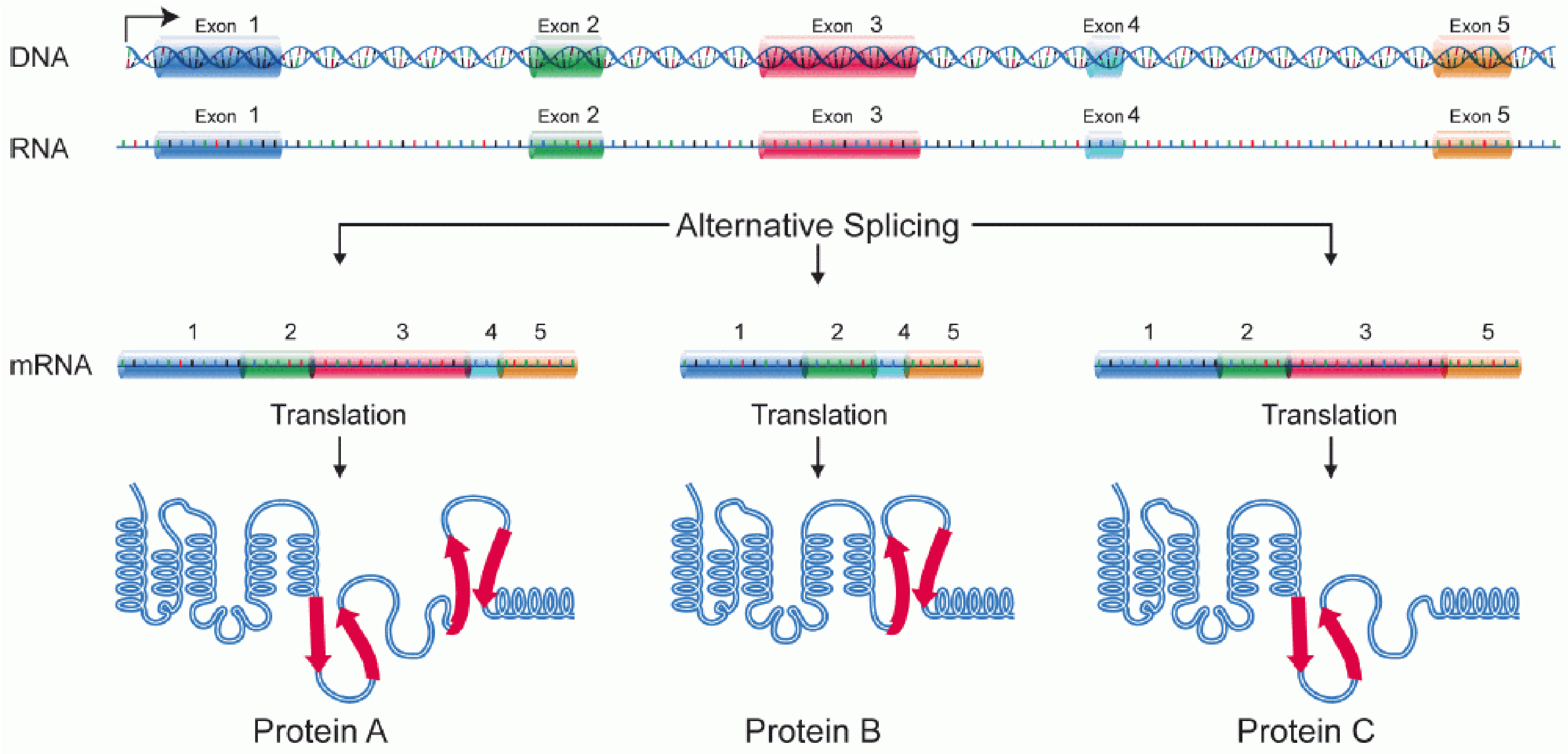
Forgo fusion gene detection with transcriptome alignment

- 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
Gene 1
+
Gene 2
Pol II
transcription
Fusion gene
B Fusion by transcription or splicing
Transcription read-through, mRNA *trans*-splicing or *cis*-splicing
5′
3′
mRNA 1
+
5′
3′
mRNA 2
50 S
translation
5′
3′
Fusion transcript

Forgo detection of novel splice variants with transcriptome alignment



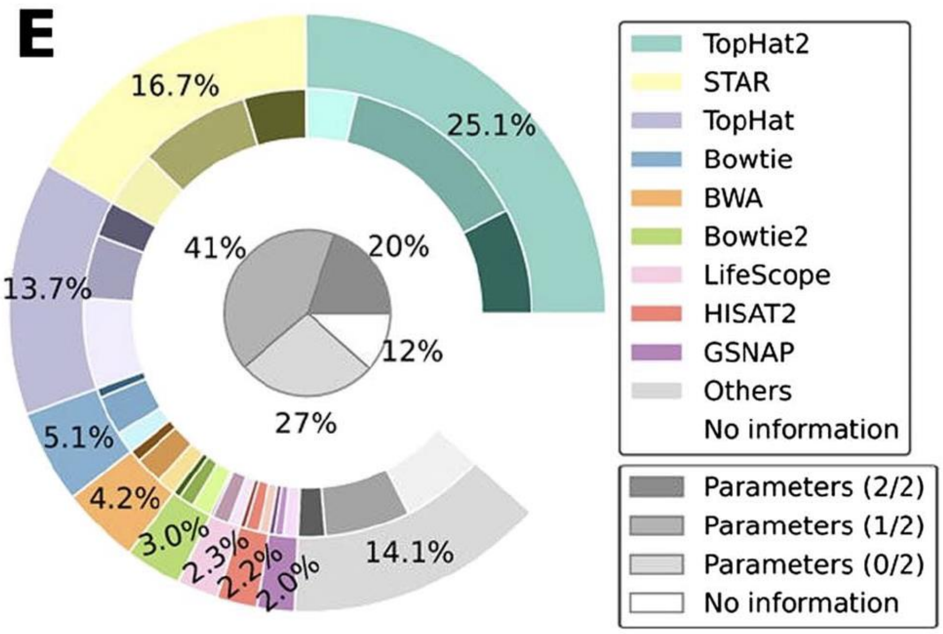
Multiple Alignment Programs available

Genome

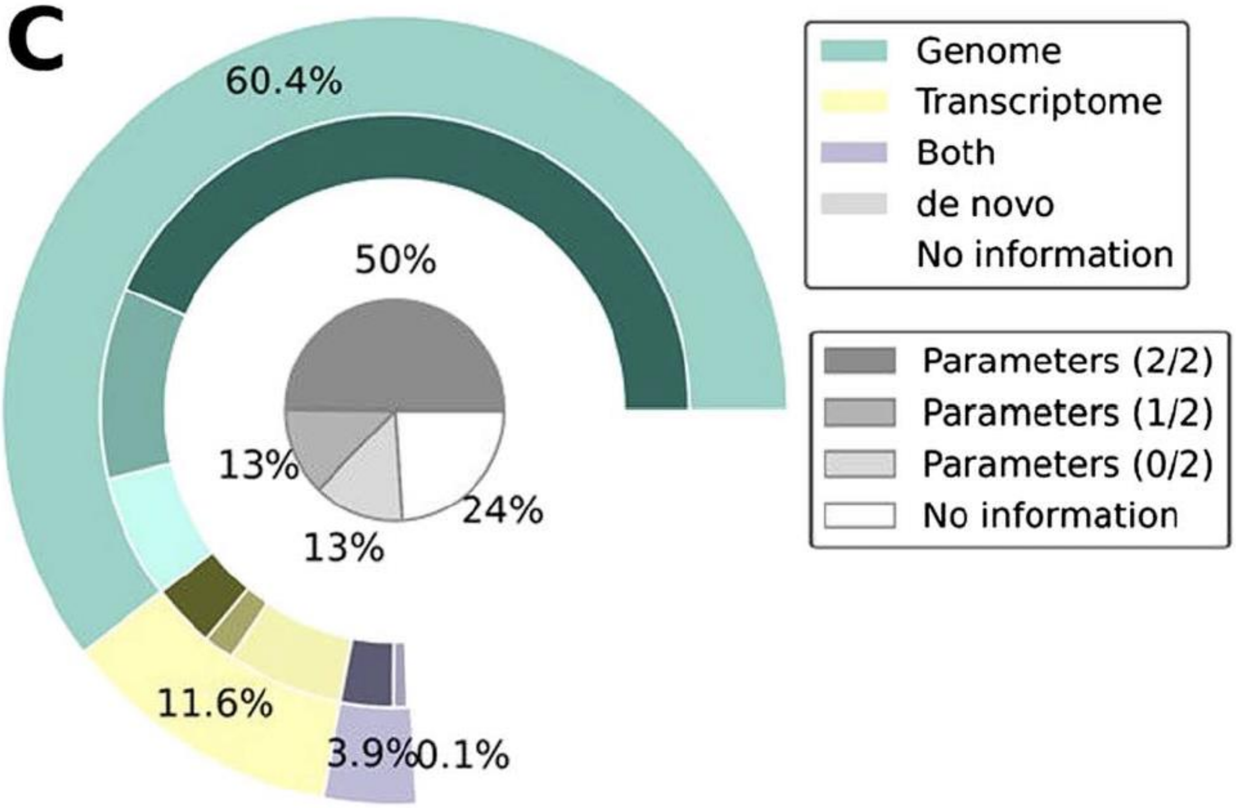
- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

Transcriptome

- Salmon
- Kallisto
- Sailfish



What does the scientific community do?



Simoneau et al. 2021

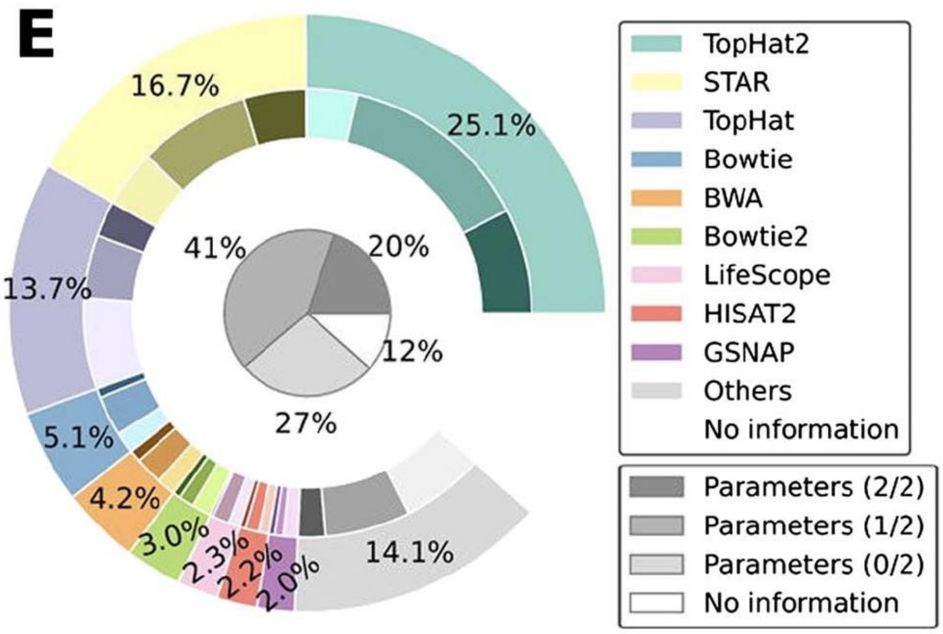
Programs we will use:

Genome

- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

Transcriptome

- Salmon
- Kallisto
- Sailfish



Class activity #2 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

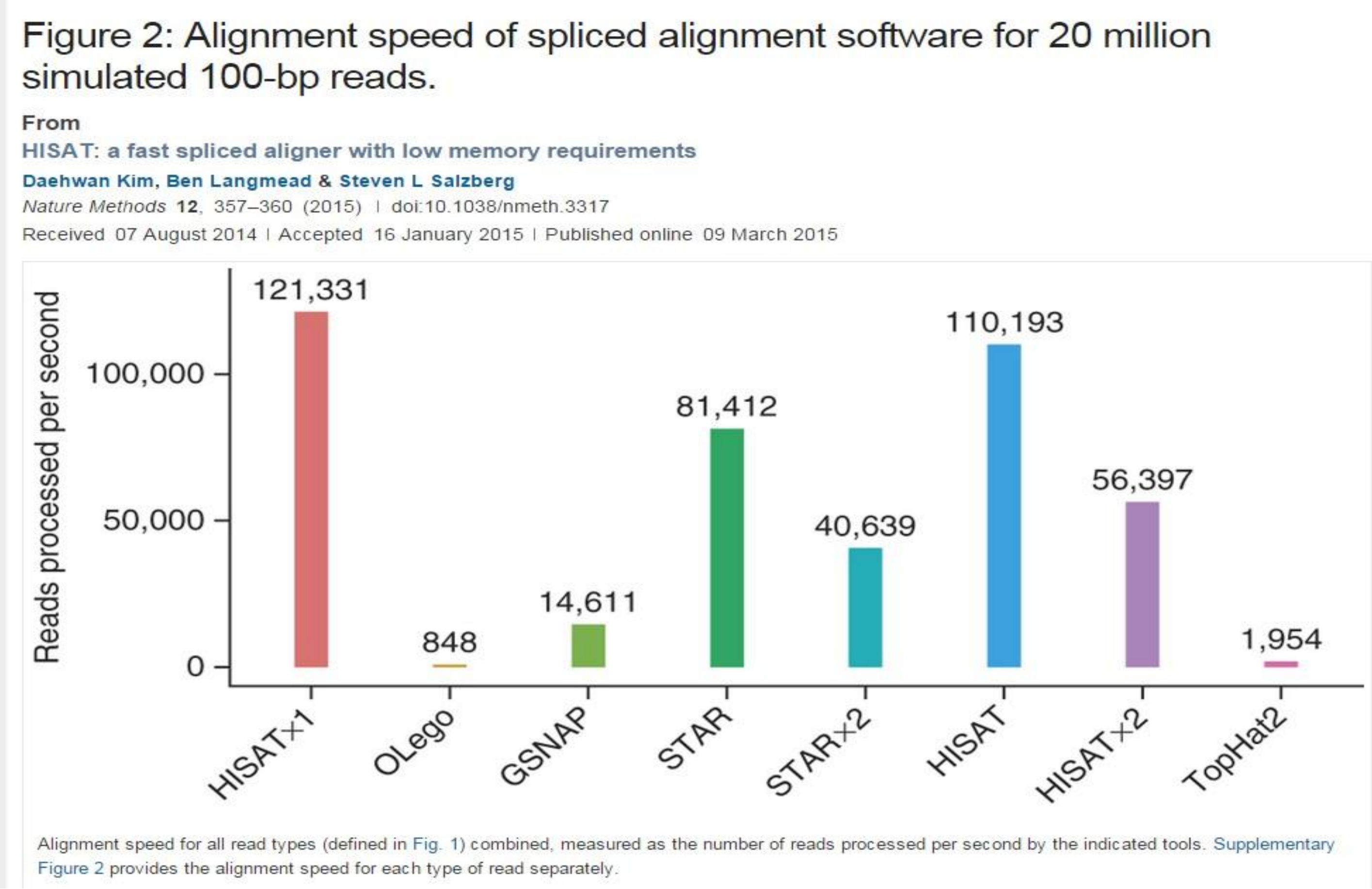


Figure 2: Alignment speed of spliced alignment software for 20 million simulated 100-bp reads.

From

HISAT: a fast spliced aligner with low memory requirements

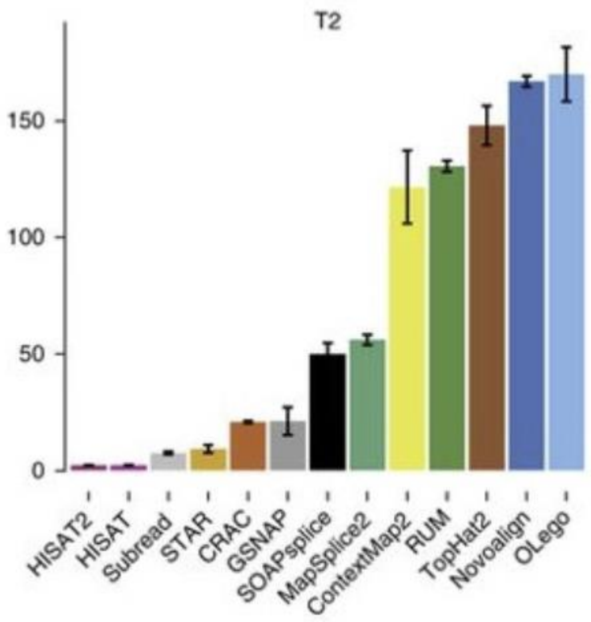
Daehwan Kim, Ben Langmead & Steven L Salzberg

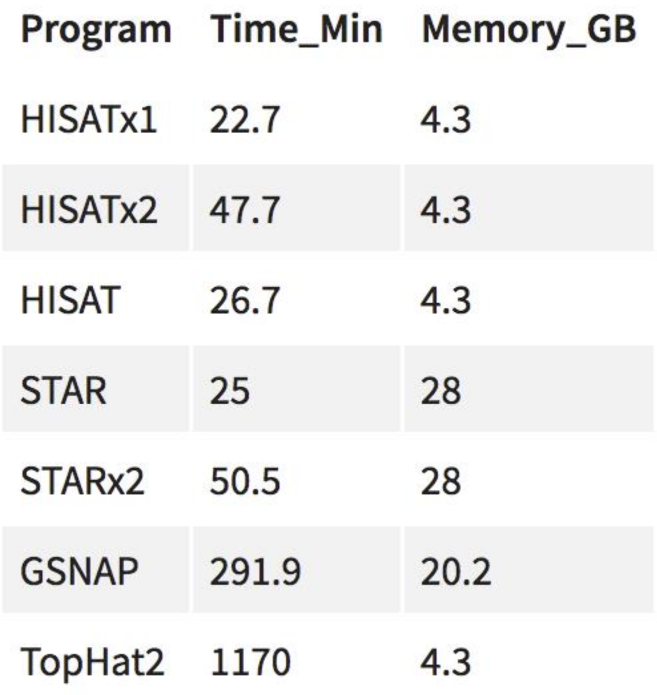
Nature Methods **12**, 357–360 (2015) | doi:10.1038/nmeth.3317

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

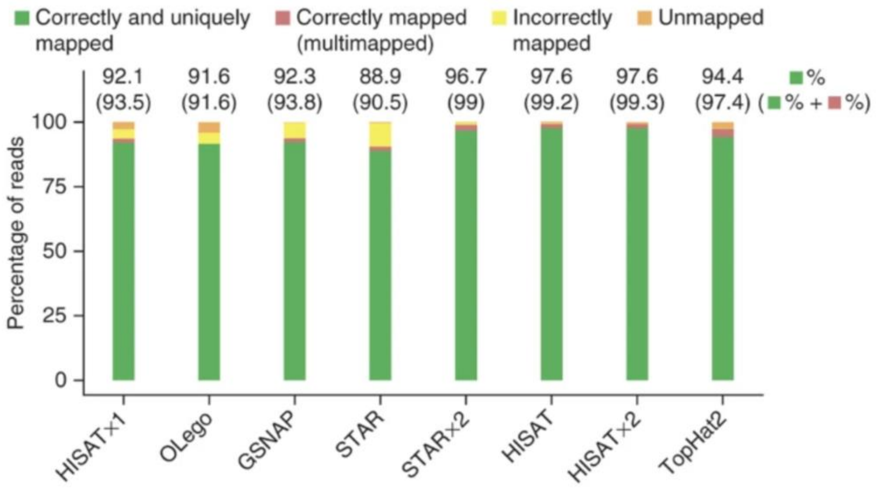
HISAT2

- 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.

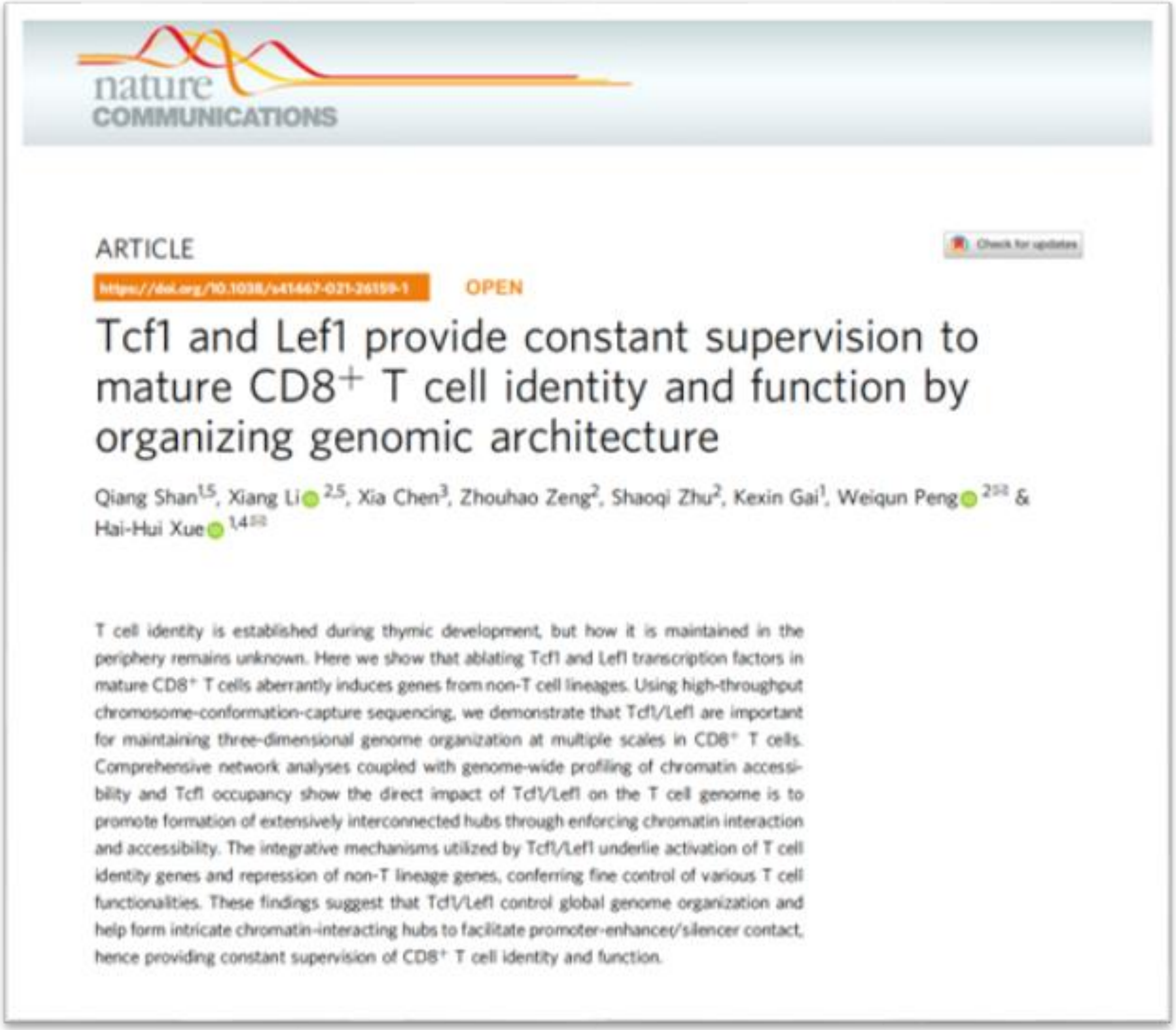


HISAT2 usage

- 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✉]

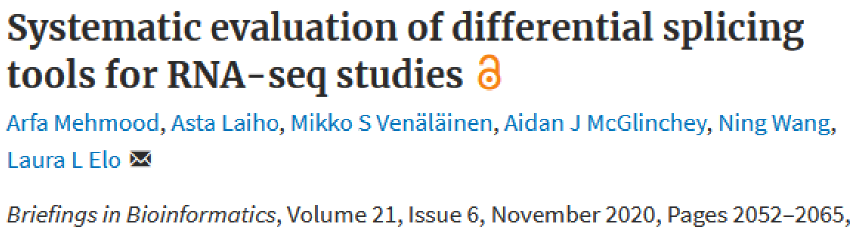
T cell identity is established during thymic development, but how it is maintained in the periphery remains unknown. Here we show that ablating Tcf1 and Lef1 transcription factors in mature CD8$^+$ T cells aberrantly induces genes from non-T cell lineages. Using high-throughput chromosome-conformation-capture 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

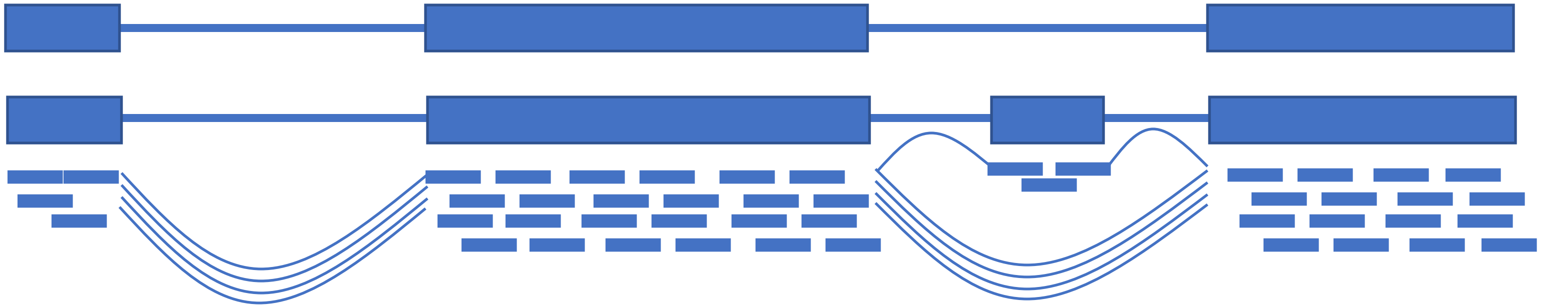
Analysing Splicing



Systematic evaluation of differential splicing tools for RNA-seq studies

Arfa Mehmood, Asta Laiho, Mikko S Venäläinen, Aidan J McGlinchey, Ning Wang, Laura L Elo

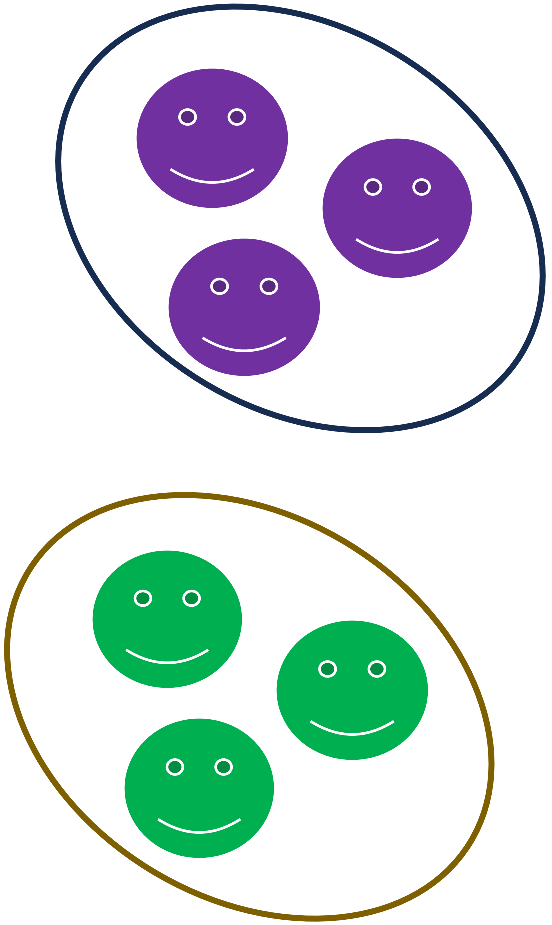
Briefings in Bioinformatics, Volume 21, Issue 6, November 2020, Pages 2052–2065,



- Try to quantitate transcripts (cufflinks, RSEM, bitSeq)
- Quantitate exons and compare to gene (EdgeR, DEXSeq)
- Quantitate splicing events (rMATS, MAJIQ)

Computational Considerations

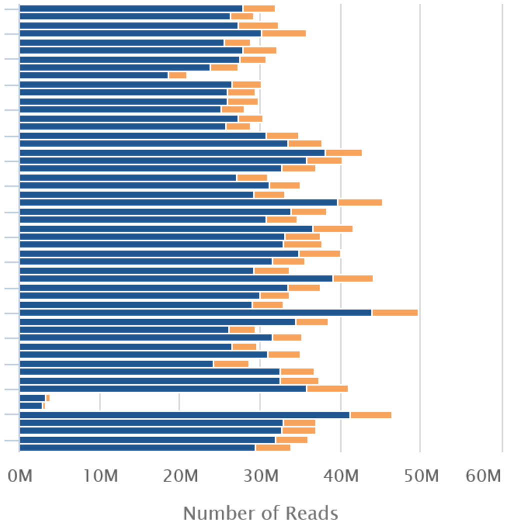
- Two conditions three replicates
- 6-12 FASTQ files
- 6-12 quality control files
- 6-12 FASTQ files post trimming
- 6 BAM files + 6 index BAM files
- 6 Gene count files
- 1 counts matrix



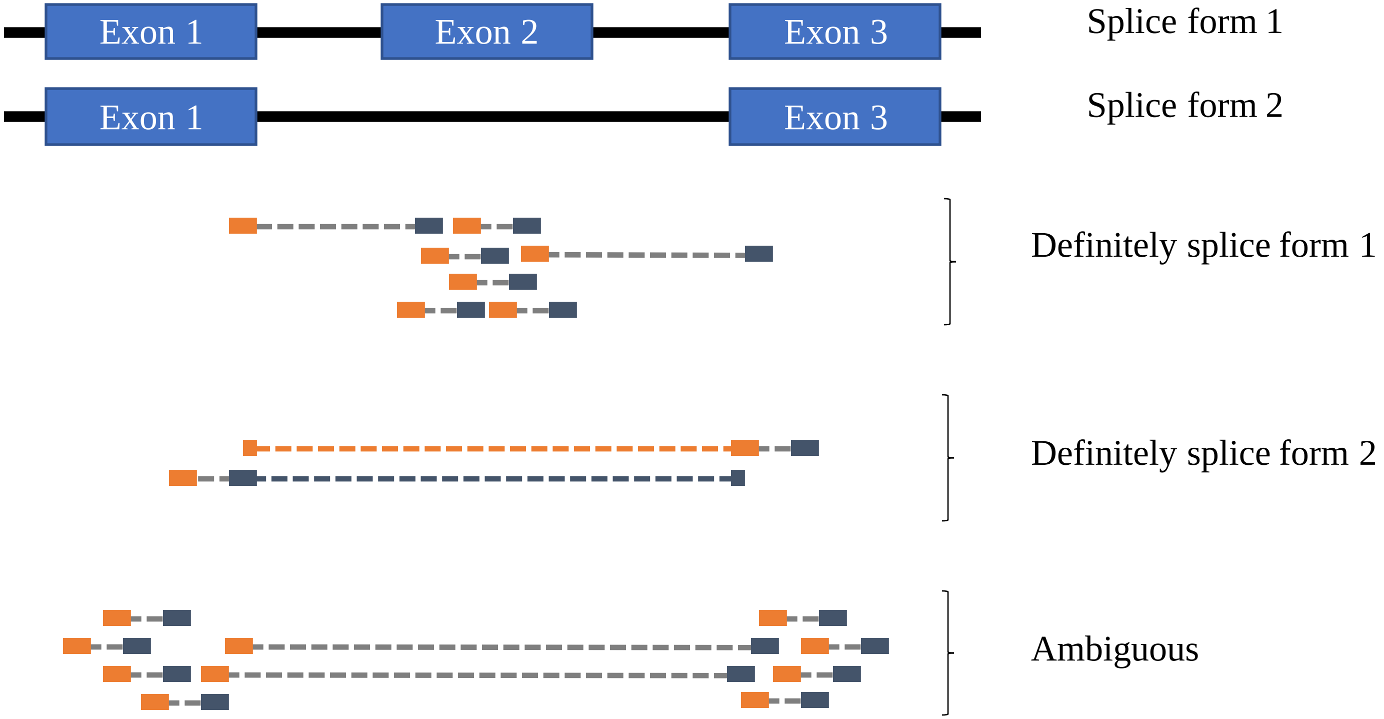
36 - 48 files

Next Week:

- Storing aligned reads: SAM/BAM file formats
- We will review outputs from HISAT2_exercise (class exercise #1) vs HISAT2_modify (class exercise #3); please complete both!
- We will create a MULTIQC output
- We will use RSEQC to QC alignment statistics



Next Week: Quantitation



Class activity #3
Script Submission

HISAT2_modify

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/)