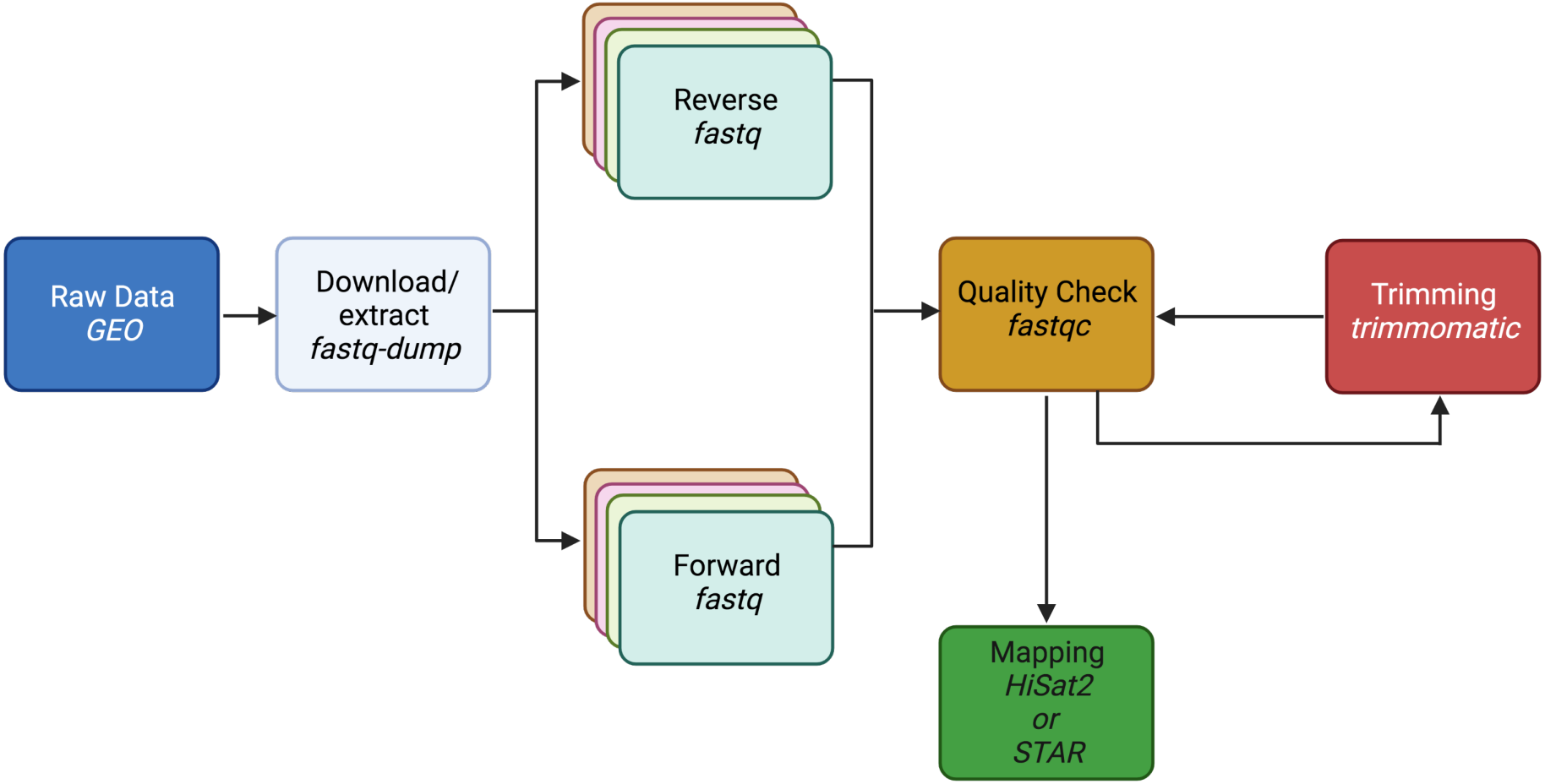
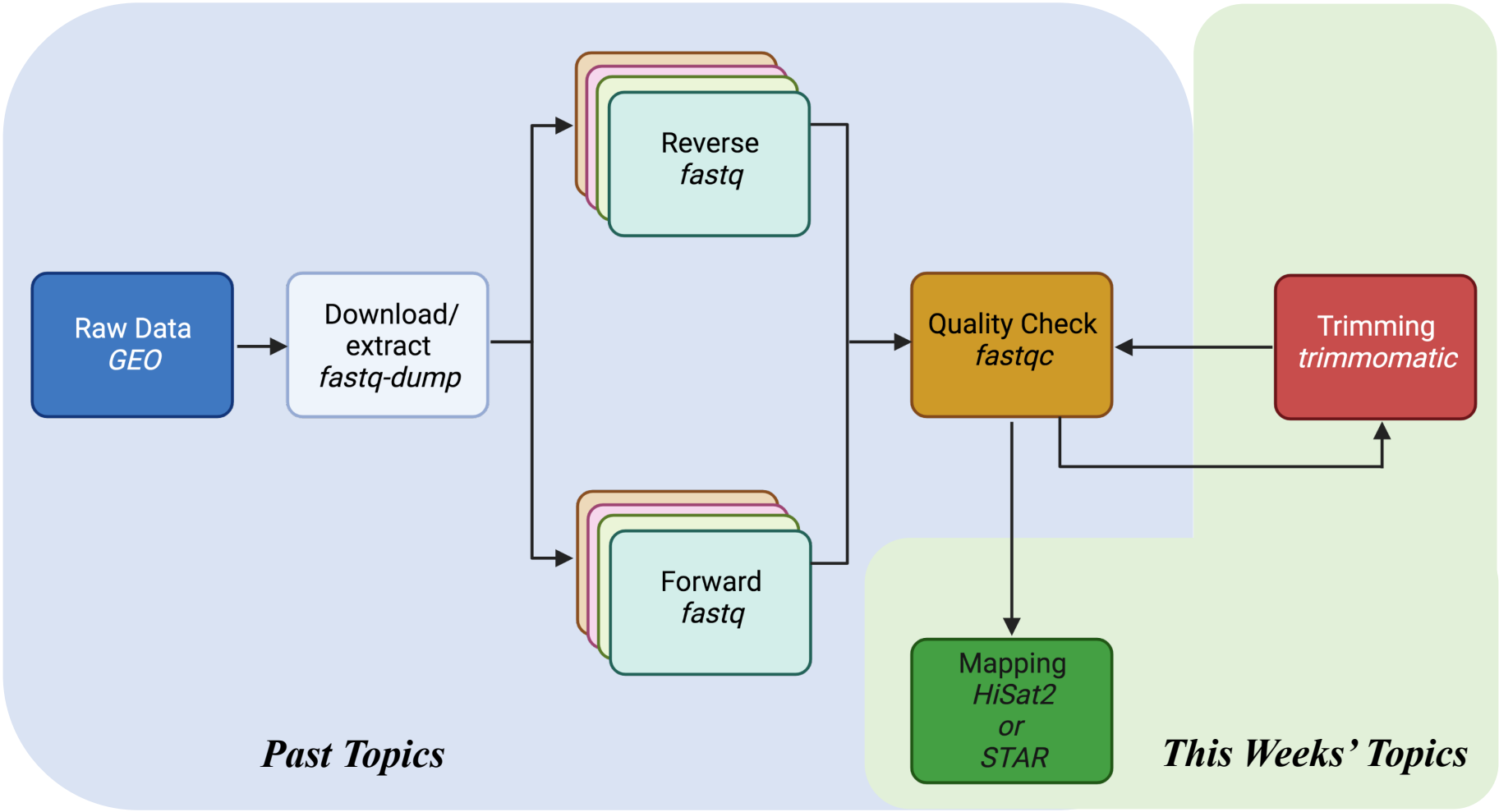
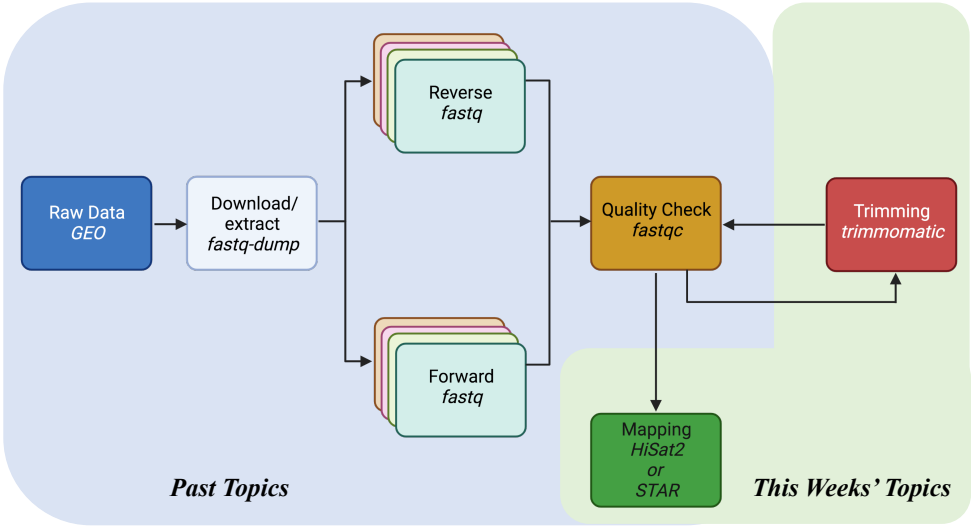


Lecture 8: Mapping
February 21, 2023


Raw Data
GEO
Download/
extract
fastq-dump
Reverse
fastq
Forward
fastq
Quality Check
fastqc
Trimming
trimmomatic
Mapping
HiSat2
or
STAR
Past Topics
This Weeks' Topics

FASTQC recap



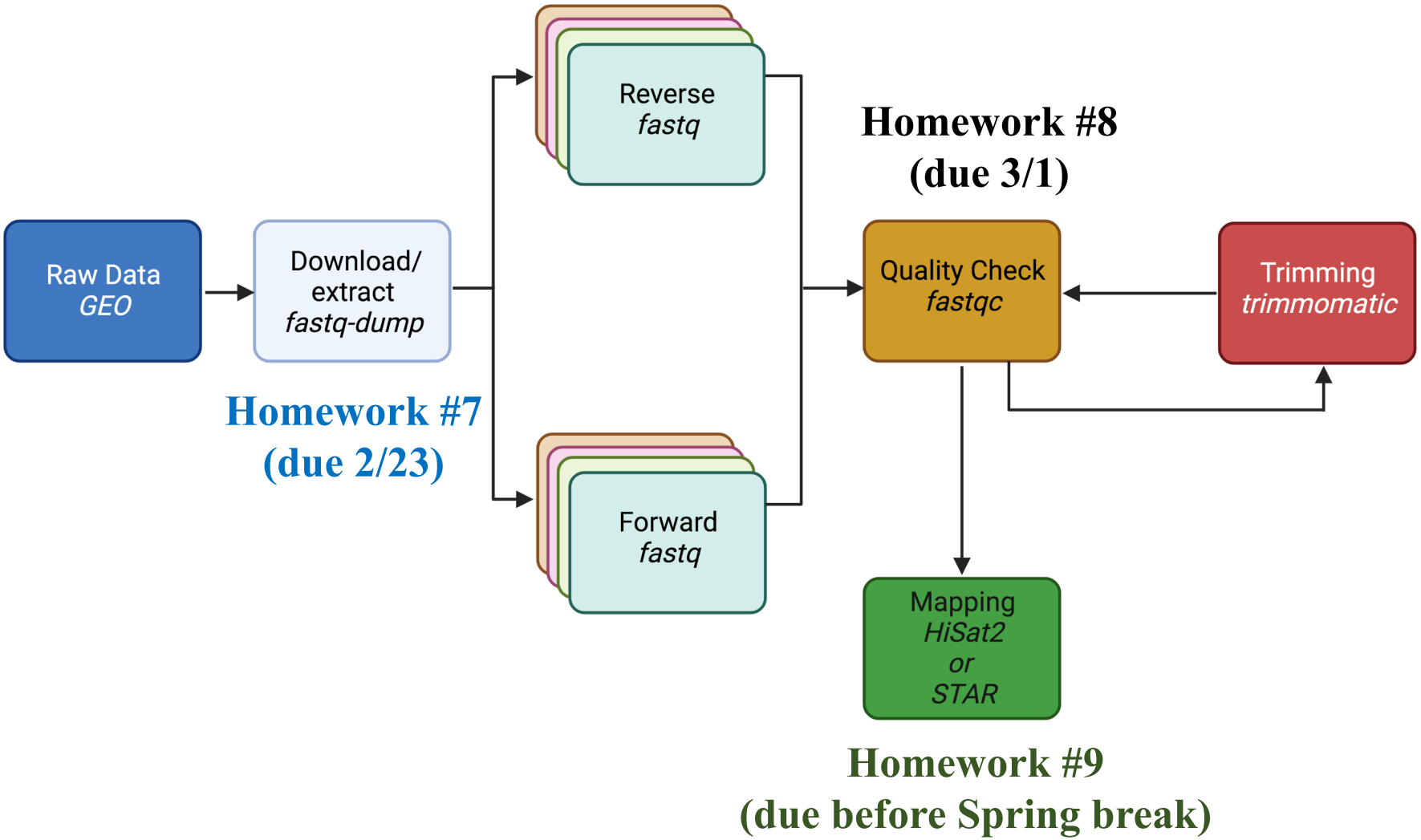
Step 1: FASTQC

Two basic questions of QC

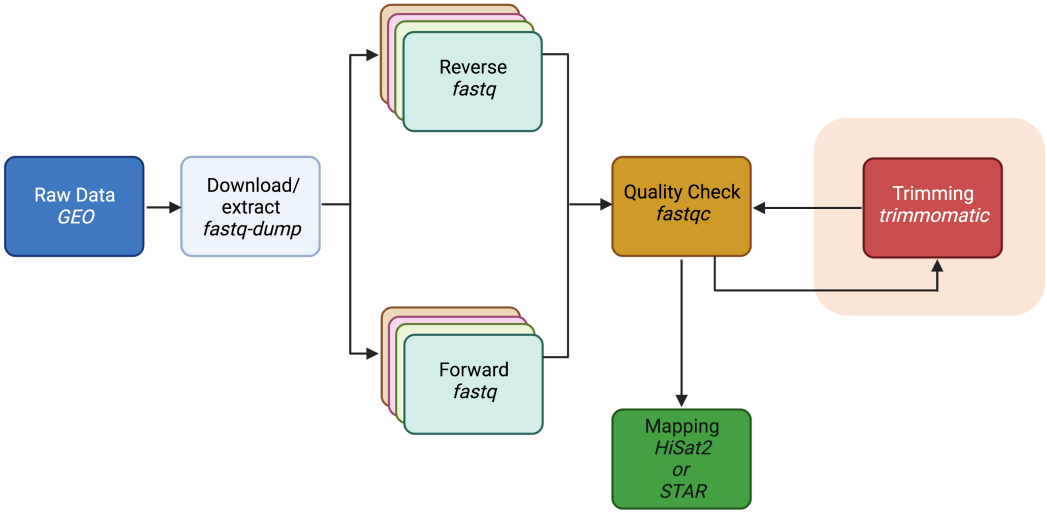
- How successful was the actual sequencing?
  - High base call confidence <30 phred score

- Did the library prep generate a faithful representation of the DNA/RNA molecules in the sample?
  - No contaminations (rRNA, adapters, primers, etc.)
  - No bias towards fragments of certain GC contents/sizes

Assignment overview



Trimming


Raw Data
GEO
Download/
extract
fastq-dump
Reverse
fastq
Forward
fastq
Quality Check
fastqc
Trimming
trimmomatic
Mapping
HiSat2
or
STAR

Step 2: Clean the FASTQ files

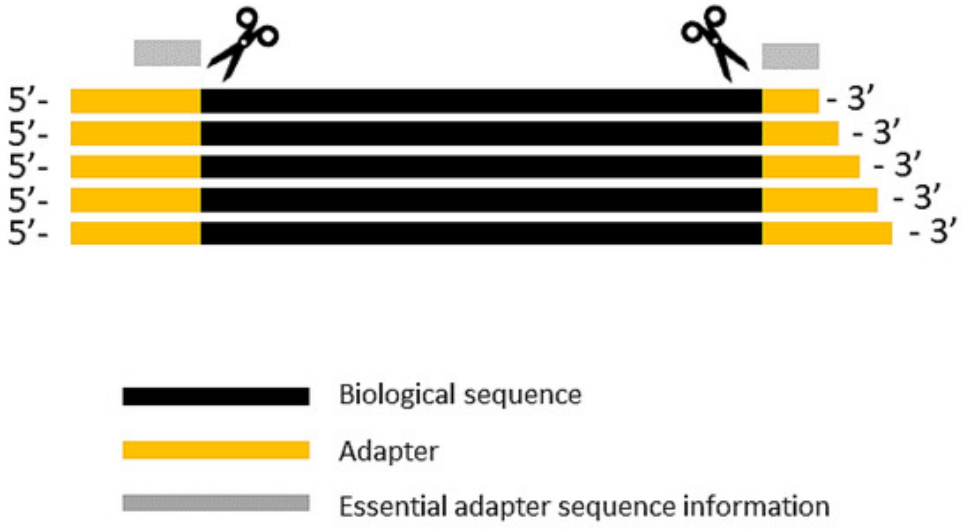
- Raw reads may suffer from the following:
  - Presence of adaptors
  - Low quality reads
  - Or other issues that result in low base call
- Trimming is optional
- Avoid aggressive clipping as this could cause miss-mapping

Why optional?

- Modern “local aligners” like STAR, BWA-MEM, HISAT2, will “soft-clip” non-matching sequences
- Other pseudo-aligners like Salmon or Kallisto will not have a problem with reads containing adapter sequences
- However, if the data is being used for variant analysis or genome assembly, reads should be trimmed

Common programs used for trimming

Preprocessing of FASTQ can be performed by a variety of programs



fastp

Trimmomatic

Trim galore

Cutadapt

How would you run a program on the VACC?

1) Install it in your personal VACC account and then be able to call it (*sratoolkit*)

2) Or load it from the shared computing cluster using `module` package (*fastqc*)

Use module to load

`module avail` – ***to see what programs are available***

`module load` **`trimmomatic-0.38-gcc-7.3.0-jwjzei2`**

Use tab completion!

Usage:

Code

```
$ trimmomatic
```

Which will give you the following output:

Paired-end

Output

```
Usage:
       PE [-version] [-threads <threads>] [-phred33|-phred64] [-trimlog <trimLogFile>] [-summary <statsSummaryFile>] [-quiet] [-validatePair
s] [-basein <inputBase> | <inputFile1> <inputFile2>] [-baseout <outputBase> | <outputFile1P> <outputFile1U> <outputFile2P> <outputFile2U>] <t
rimmer1>...
   or:
       SE [-version] [-threads <threads>] [-phred33|-phred64] [-trimlog <trimLogFile>] [-summary <statsSummaryFile>] [-quiet] <inputFile> <ou
tputFile> <trimmer1>...
   or:
       -version
```

single-end

Command parameters

step	meaning
`ILLUMINACLIP`	Perform adapter removal.
`SLIDINGWINDOW`	Perform sliding window trimming, cutting once the average quality within the window falls below a threshold.
`LEADING`	Cut bases off the start of a read, if below a threshold quality.
`TRAILING`	Cut bases off the end of a read, if below a threshold quality.
`CROP`	Cut the read to a specified length.
`HEADCROP`	Cut the specified number of bases from the start of the read.
`MINLEN`	Drop an entire read if it is below a specified length.
`TOPHRED33`	Convert quality scores to Phred-33.
`TOPHRED64`	Convert quality scores to Phred-64.

Typical command

```
trimmomatic input.fastq.gz
output.trim.fastq.gz
ILLUMINACLIP:~/software/Trimmomatic-
0.36/adapters/pick_adapters LEADING:3
TRAILING:1 SLIDINGWINDOW:4:20 MINLEN:60
```

https://github.com/timflutre/trimmomatic/tree/master/adapters

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- Drop reads below the 60 bases long (MINLEN:60)

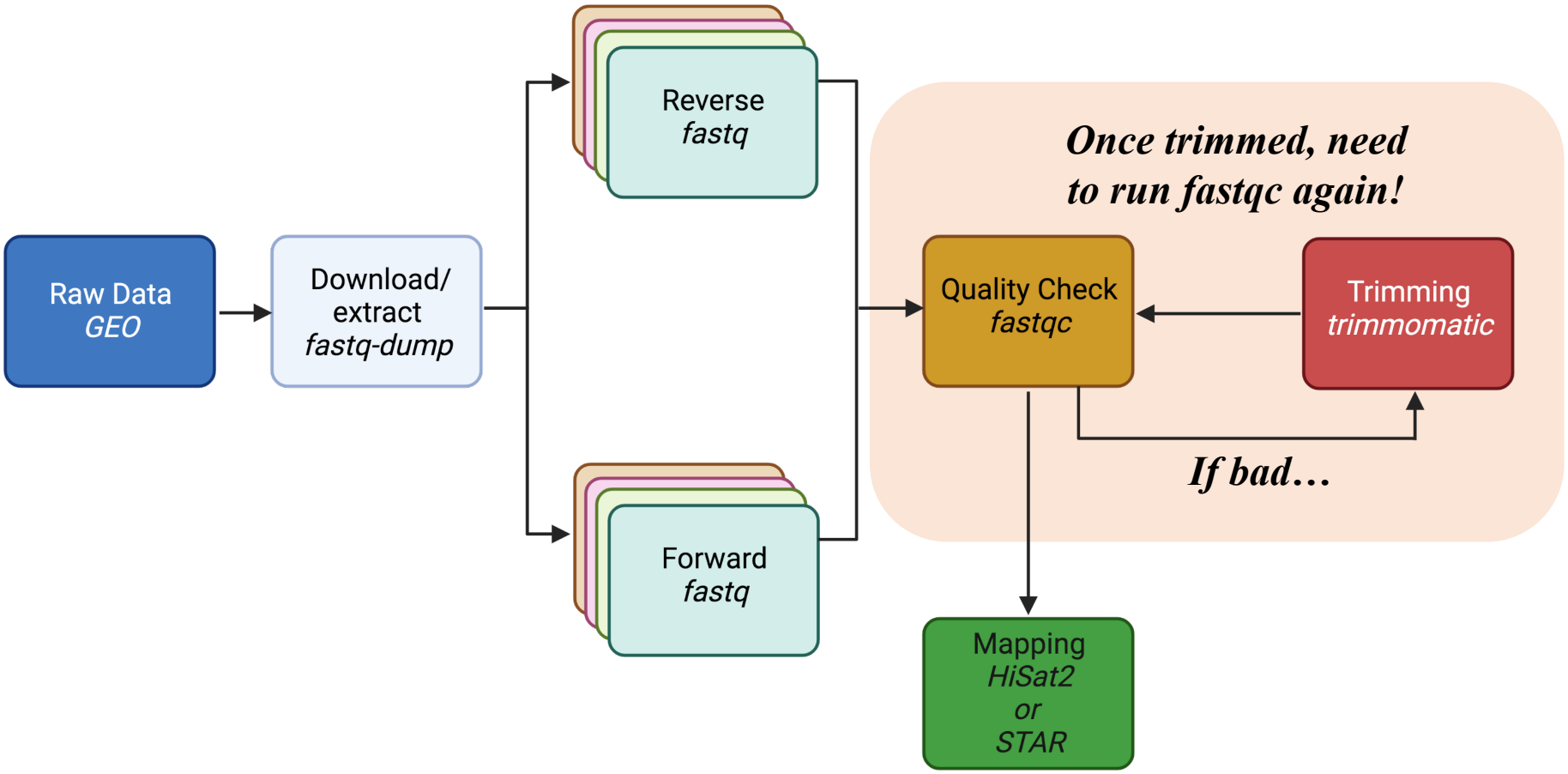
Output

SRR258199.fastq.qz

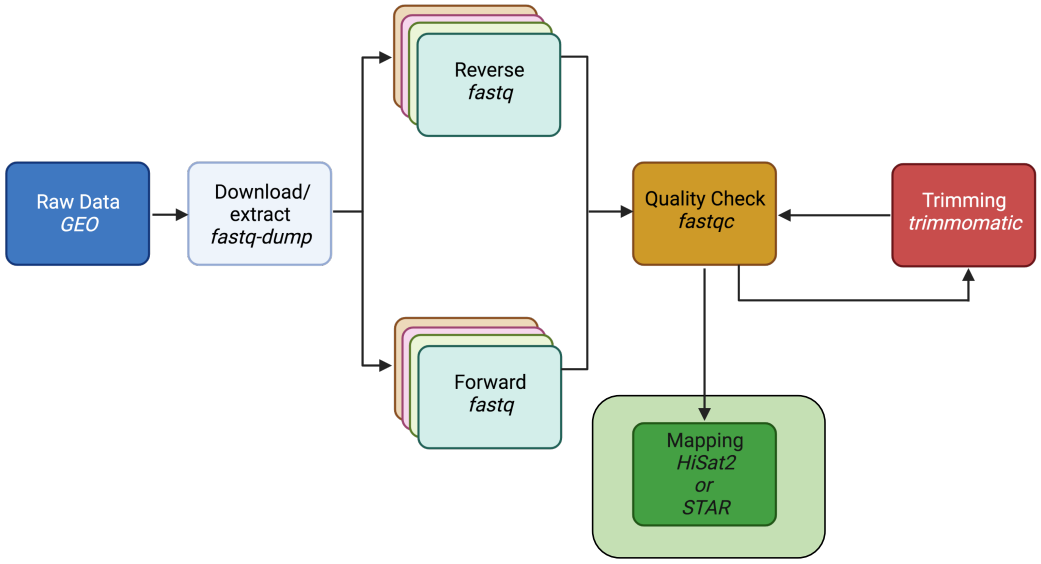
SRR258199.trim.fastq.qz

Should be smaller, because you have removed reads

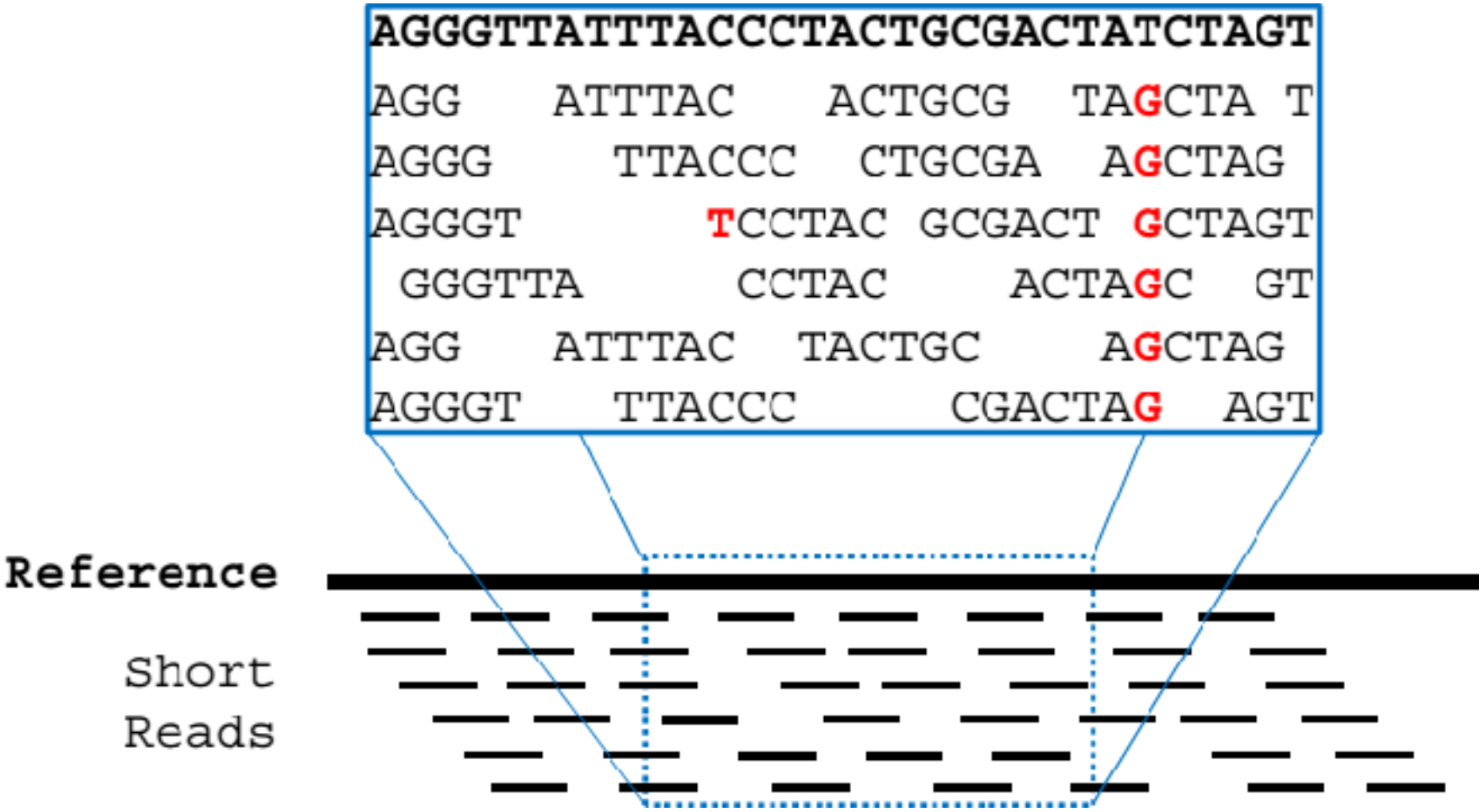
```
trimmomatic input.fastq.gz output.trim.fastq.gz
ILLUMINACLIP:~/software/Trimmomatic-
0.36/adapters/pick_adapters LEADING:3
TRAILING:1 SLIDINGWINDOW:4:20 MINLEN:60
```


Raw Data
GEO
Download/
extract
fastq-dump
Reverse
fastq
Forward
fastq
Once trimmed, need
to run fastqc again!
Quality Check
fastqc
Trimming
trimmomatic
If bad…
Mapping
HiSat2
or
STAR

Mapping to Reference



Read alignment / “mapping”



we are identifying the genomic origin of the sequenced cDNA fragment

The general challenge

The general challenge of alignment following high-throughput sequencing is to map millions of reads **accurately** and in a reasonable **time**, despite the presence of sequencing errors, genomic variation, and repetitive elements.

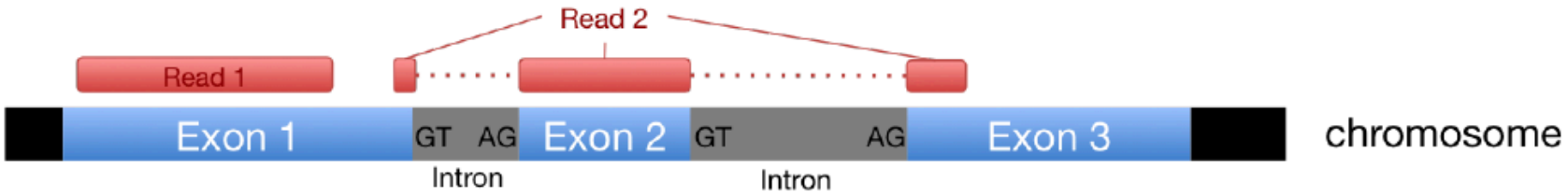
The challenge with RNA-Seq specifically…

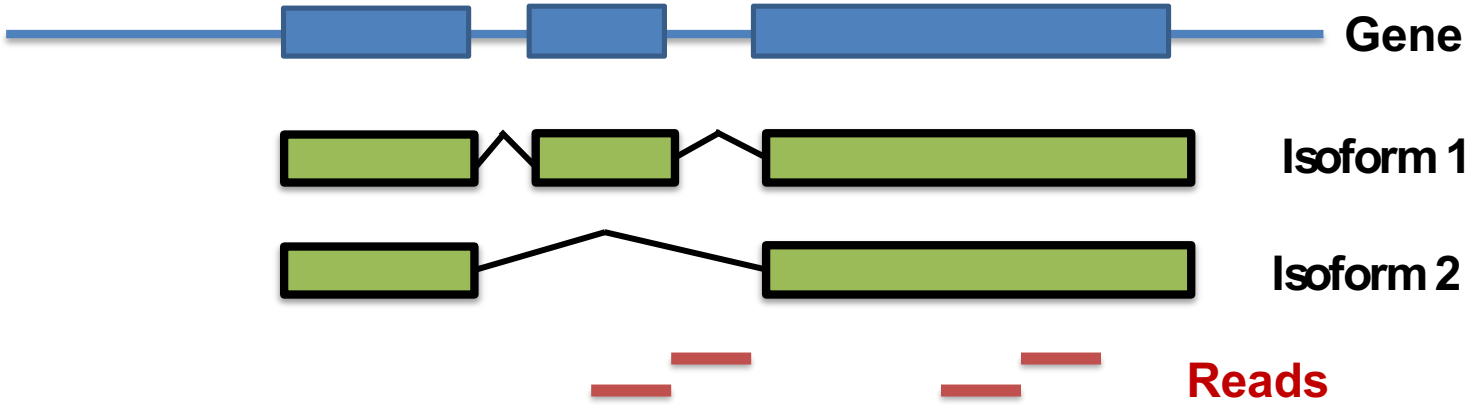
The sequencing library was constructed from transcribed RNA, intronic sequence are not present, and these sequenced molecule natively spanned exon boundaries.

Two categories of reads:

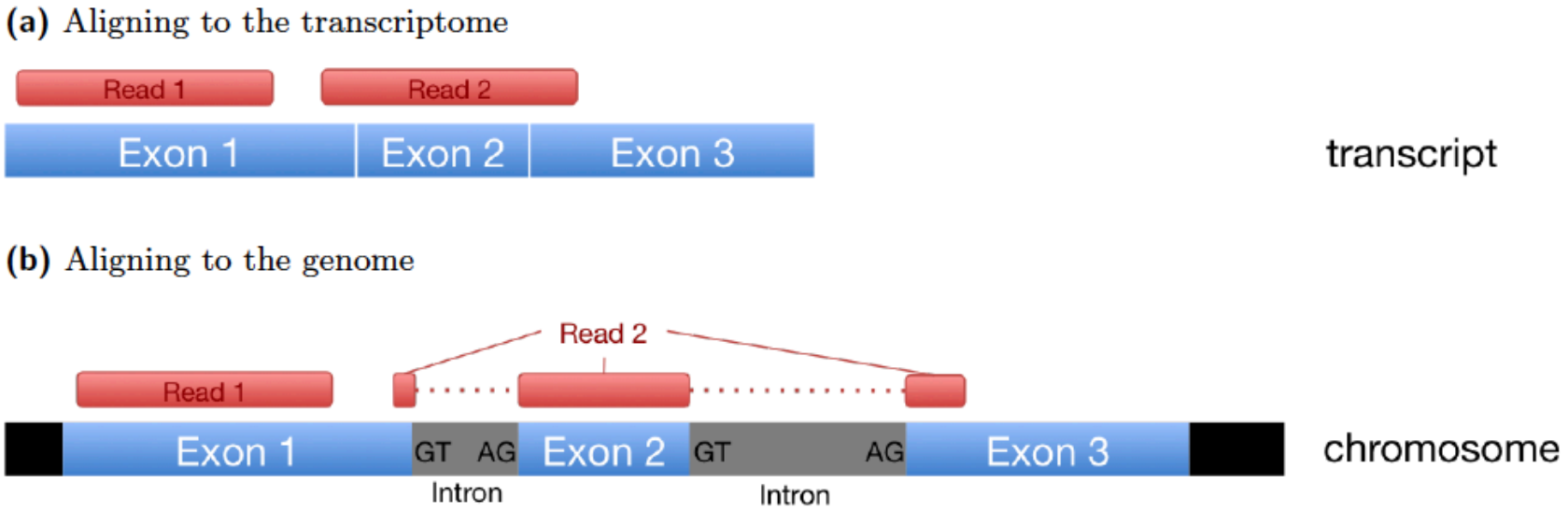
1. Reads that map entirely within exons
2. Reads that span two or more exons

Exon-exon spanning reads



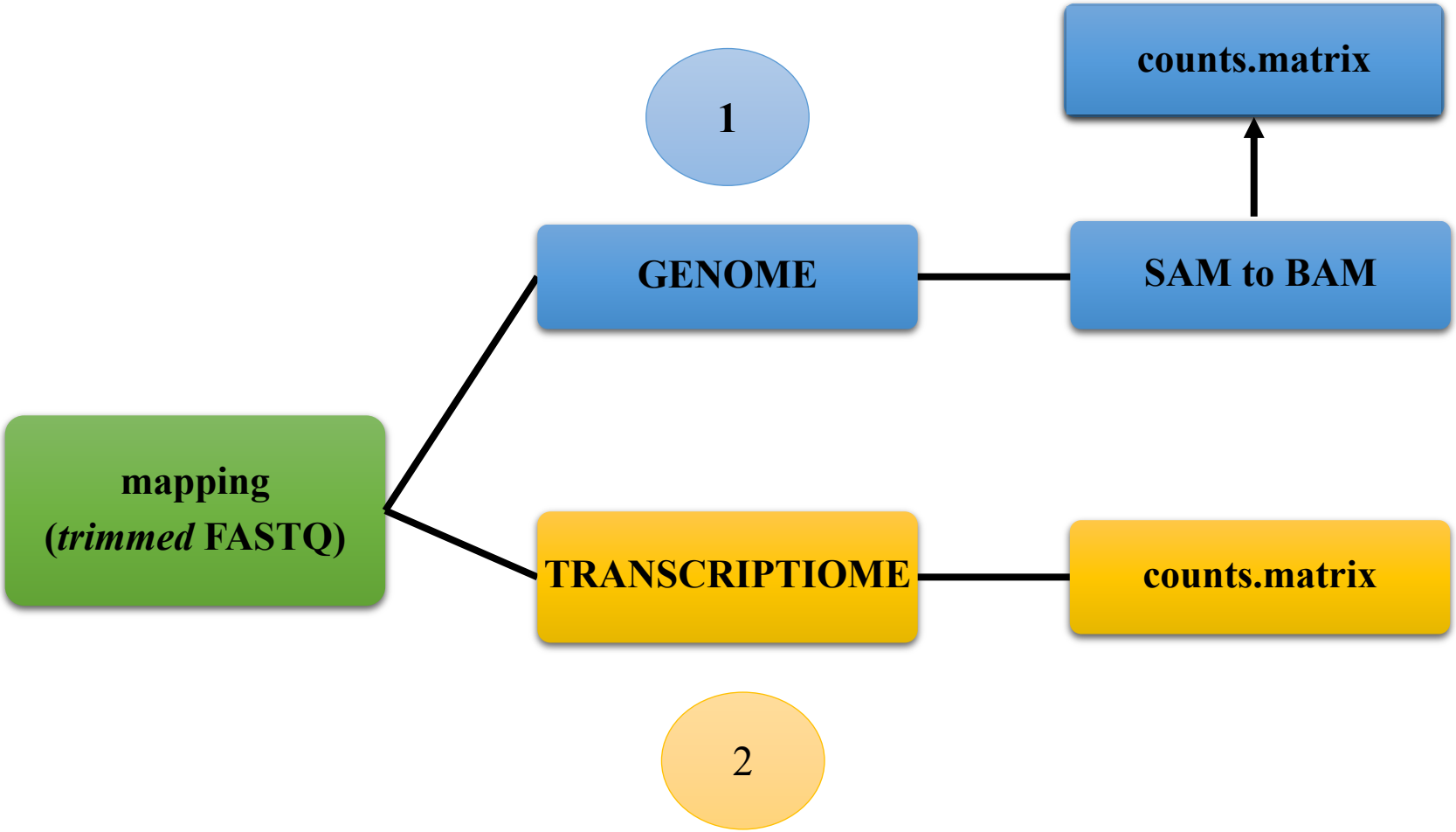


Some opt to align to transcriptome instead



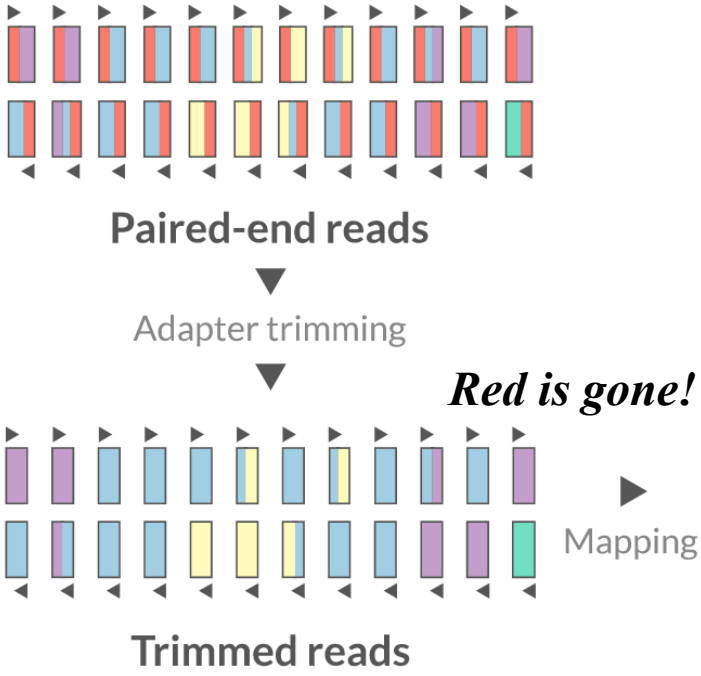
If you are mapping reads to a transcriptome intron/exon boundaries are irrelevant

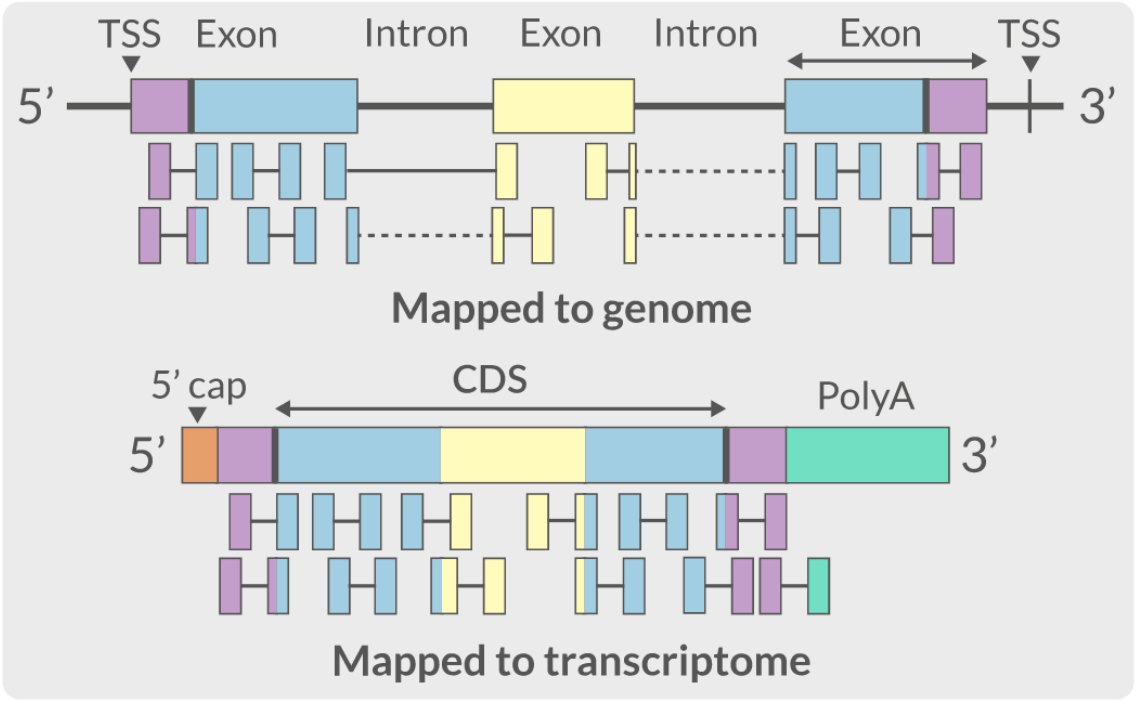
Mapping


counts.matrix
1
GENOME
SAM to BAM
mapping
(*trimmed* FASTQ)
TRANSCRIPTIOME
counts.matrix
2

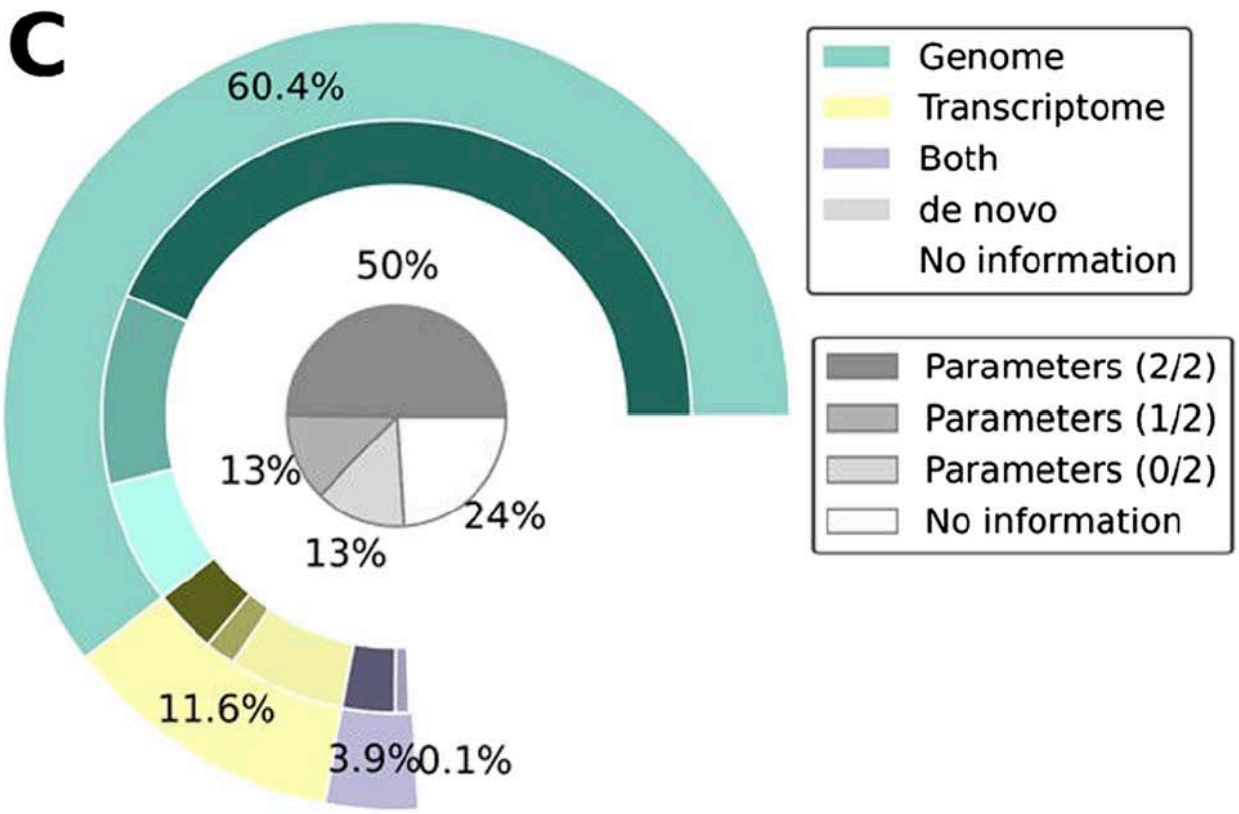
Genome vs Transcriptome?

Red = adapter





What the scientific community does



Simoneau et al. 2021

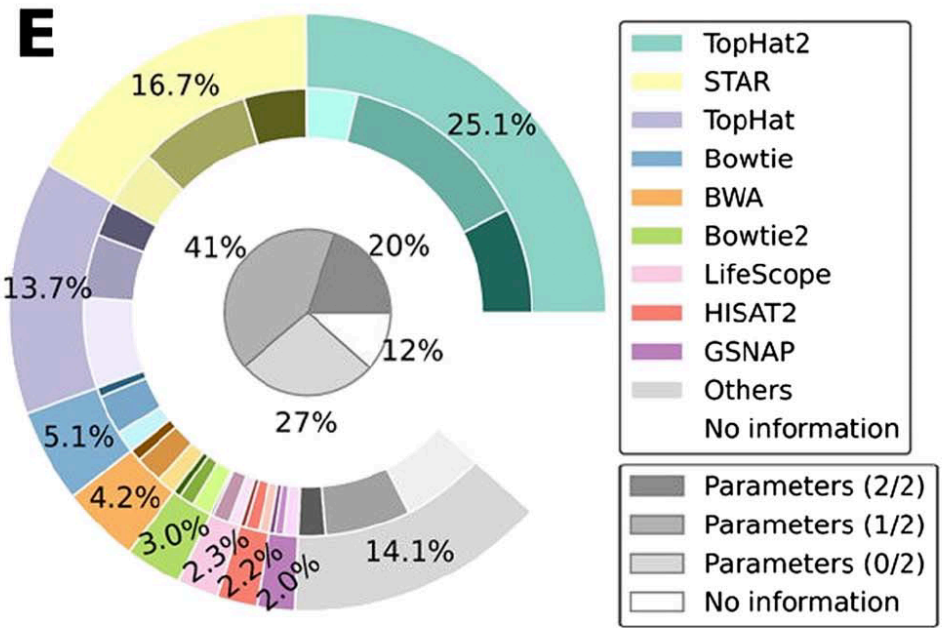
Programs available

Genome

- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

Transcriptome

- Salmon
- Kallisto
- Sailfish



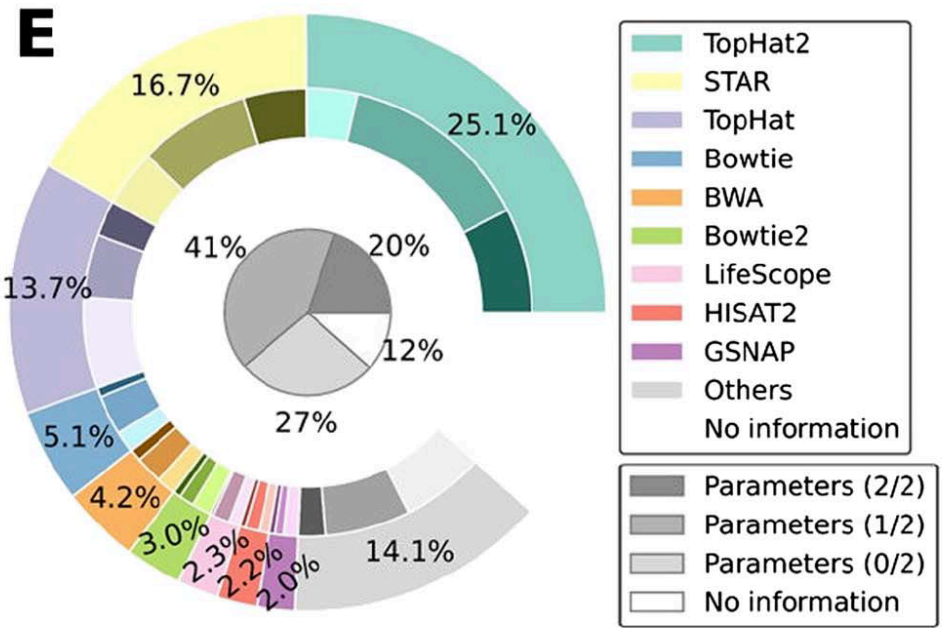
Programs available

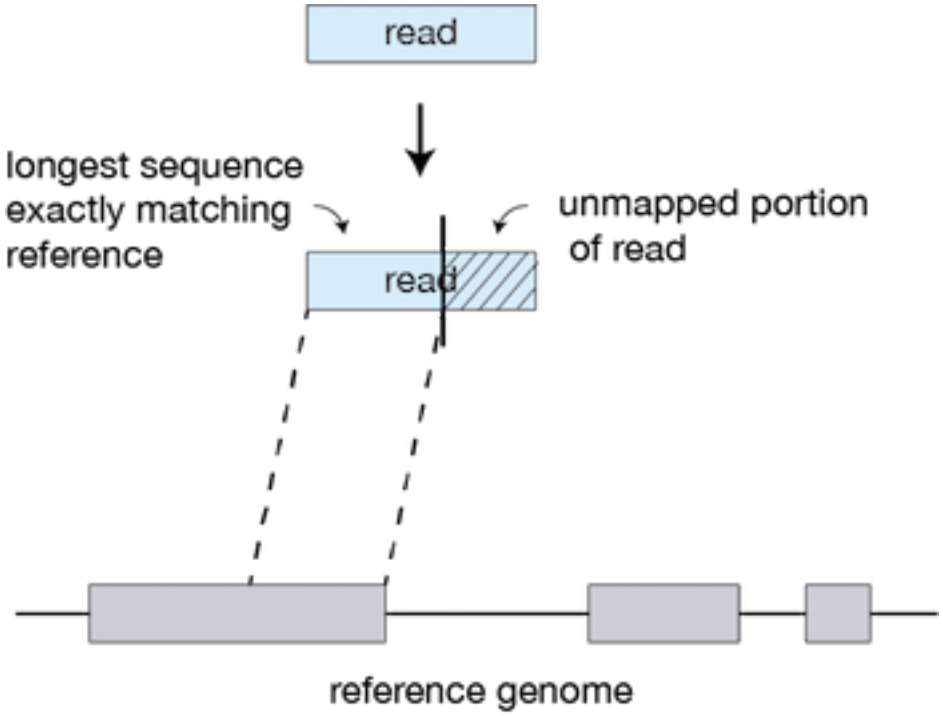
Genome

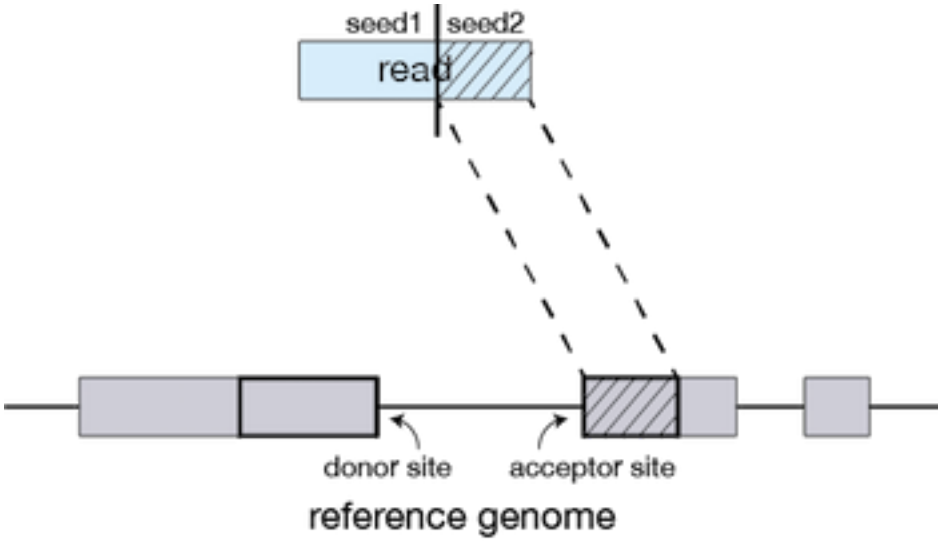
- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

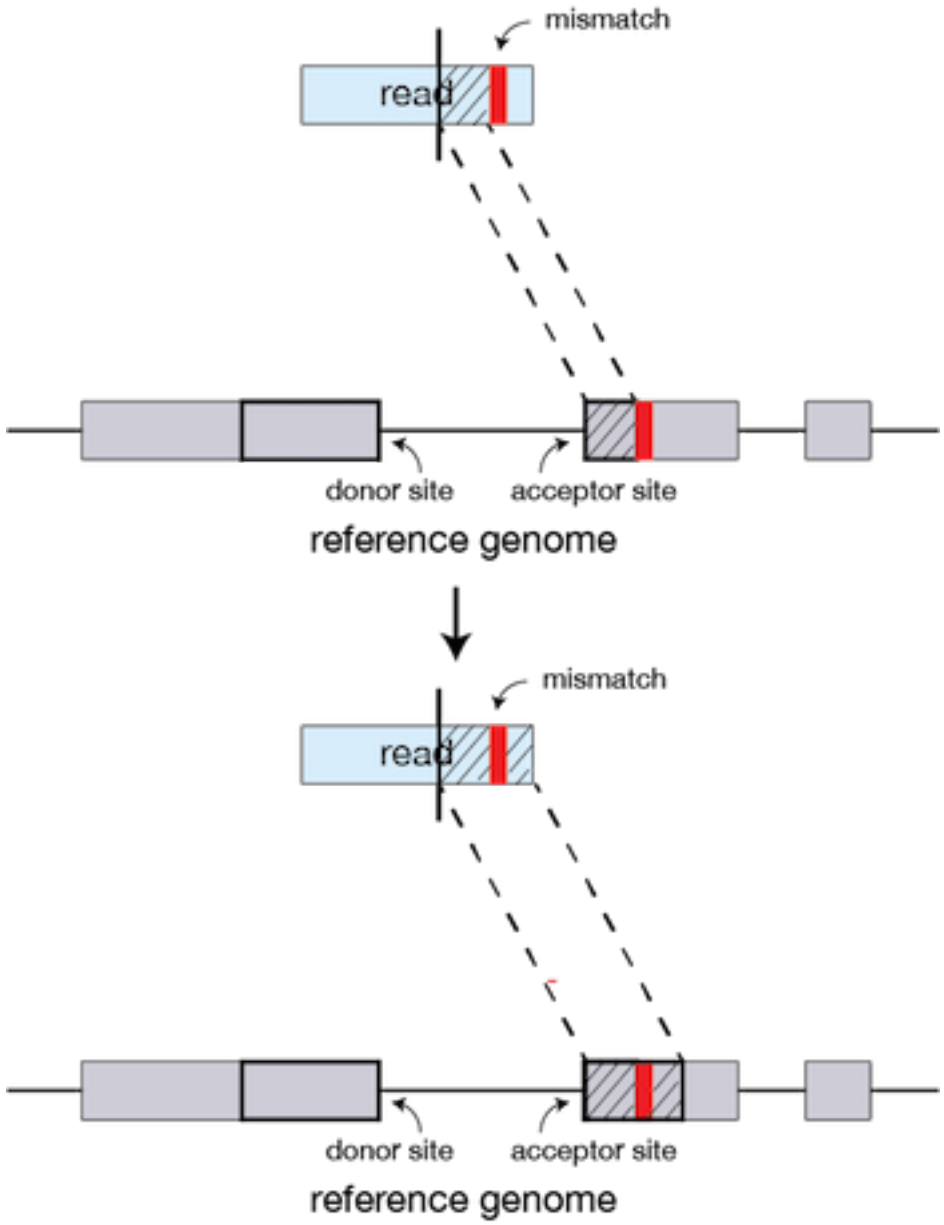
Transcriptome

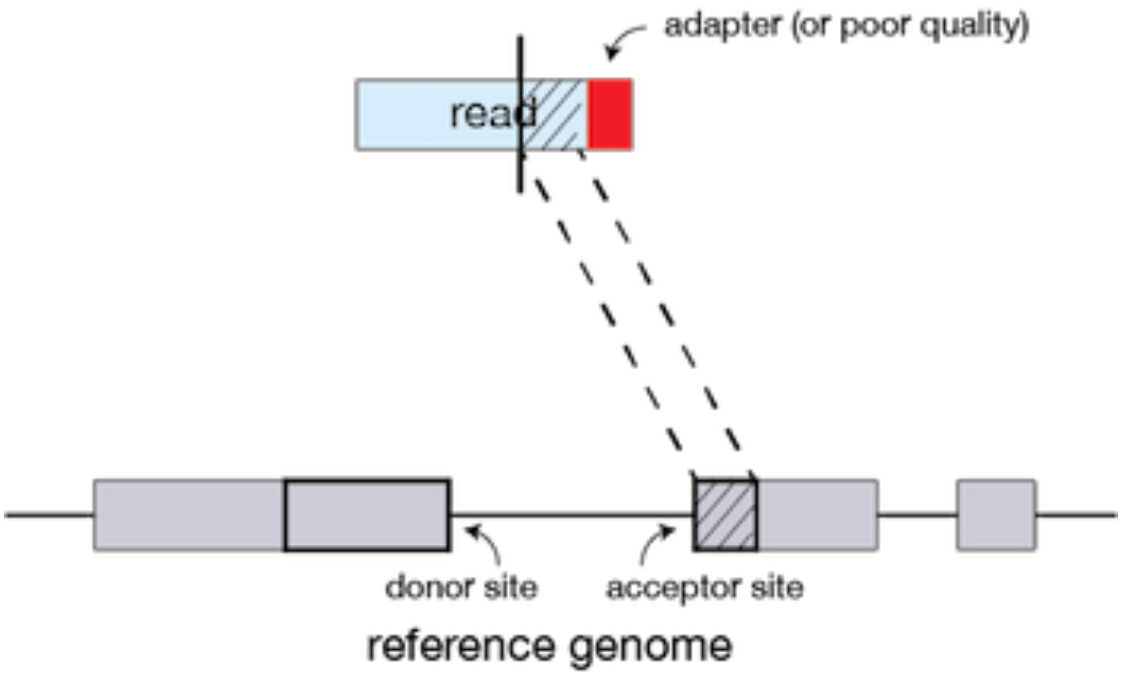
- Salmon
- Kallisto
- Sailfish




read
longest sequence
exactly matching
reference
unmapped portion
of read
read
reference genome


seed1
seed2
read
donor site
acceptor site
reference genome


mismatch
read
donor site
acceptor site
reference genome
mismatch
read
donor site
acceptor site
reference genome

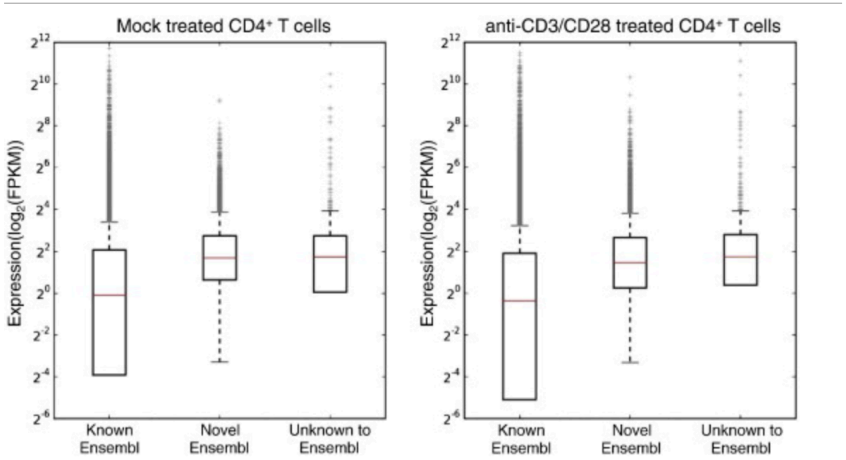

adapter (or poor quality)
read
donor site
acceptor site
reference genome

Genome

- Creation of BAM-visualization file
- 70-90% of RNA-seq reads to map to human genome
- Transcript discovery & counting
  - Mice - Transcripts unknown to Ensembl represent as much as 5% of the transcripts that are robustly expressed

Transcriptome

- BAM not created
- Mapping is only as good as the reference – slightly lower reads mapped
- Does not allow for transcript discovery



Download : Download high-res image (233KB)

Download : Download full-size image

Fig. 3. Unknown transcripts are robustly expressed. Box and whisker plots for the FPKM expression values ($\log_2$ (FPKM)) for the Known, Novel and Unknown transcript categories for both mock treated $CD4^+$ T cells and anti-CD3/CD28 treated cells.

Hutchins et al.

Aligners - Speed and Memory

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

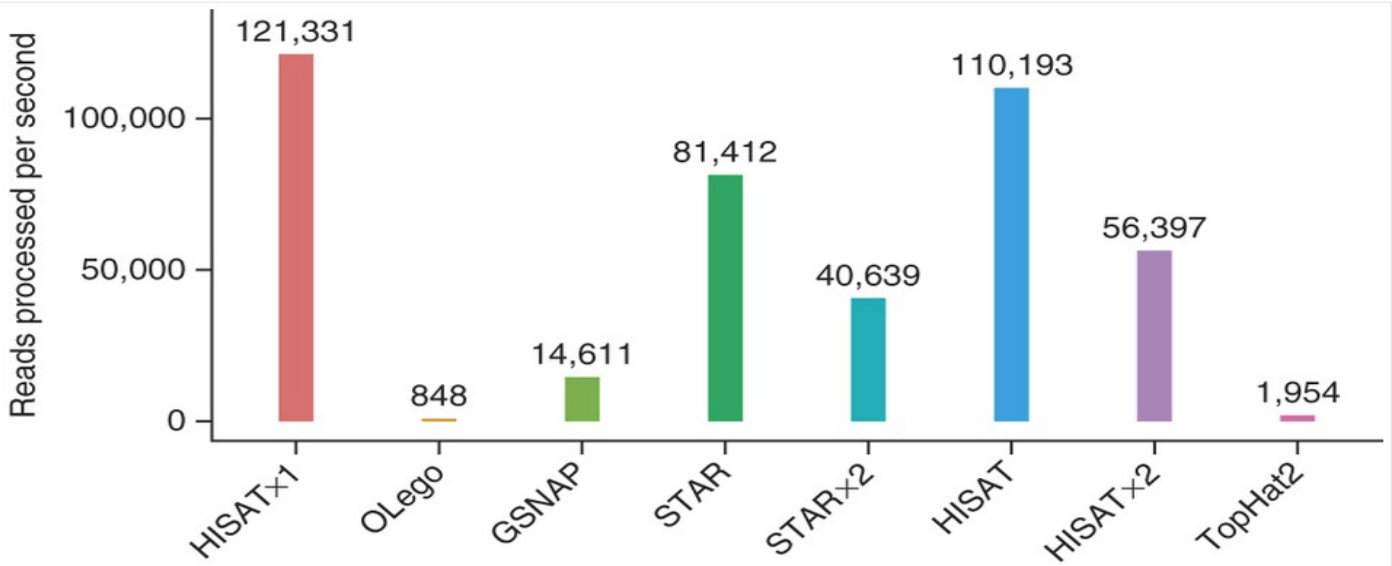
From

HISAT: a fast spliced aligner with low memory requirements

Daehwan Kim, Ben Langmead & Steven L Salzberg

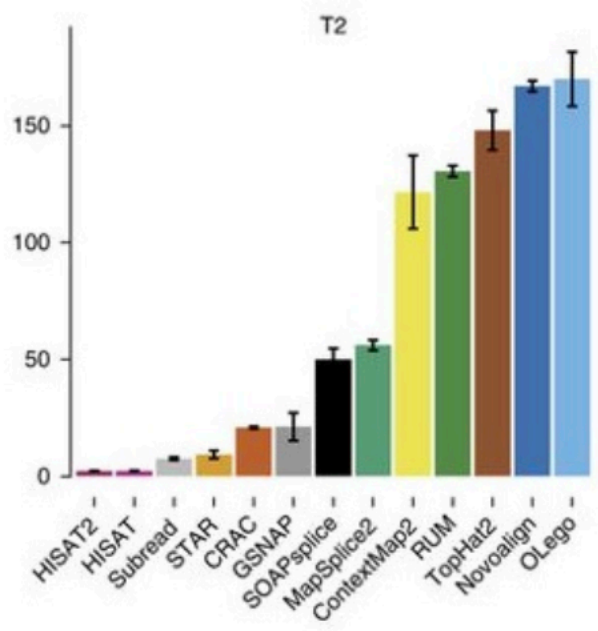
Nature Methods **12**, 357–360 (2015) | doi:10.1038/nmeth.3317

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

Reference genomes and annotation

- Irrespective of the type of read-mapping, the presence of a reference sequence as well as gene annotation are fundamental

- Reference sequence = what are you aligning to?

- Gene annotation = which parts of the reference sequence correspond to genes?

Reference Genome

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

  - Reads (FASTQ)

```
@ST-E00274:179:HHYMLALXX:8:1101:1641:1309 1:N:0:NGATGT
NCATCGTGGTATTTGCACATCTTTTCTTATCAAATAAAAAGTTTAACCTACTCAGTTATGCGCATACGTTTTTTGATGGCATTTCCATA
+
#AAAFAFA<-AFFJJJAFA-FFJJJJFFFAJJJJ-<FFJJJ-A-F-7--FA7F7-----FFFJFA<FFFFJ<AJ--FF-A<A-<JJ-7-
```

```
@instrument:runid:flowcellid:lane:tile:xpos:ypos read:isfiltered:controlnumber:sampleid
```

  - Reference Genome/Transcriptome (FASTA)

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

Annotation

- Used to quantify the number of reads which align to different genome features

- In the form of a GTF or GFF file

GTF files

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

Preparation for next class

I. An understanding of where to find GTF and FASTA files and how they were prepared for use

II. UNIX command tutorial

   a. cut
   b. grep
   c. paste
   d. sed
   e. awk
   f. Introduction to loops

Where to find 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)

General biological databases

- UCSC and Ensembl try to organize, unify, and provide access to a wide range of genomes and annotation data

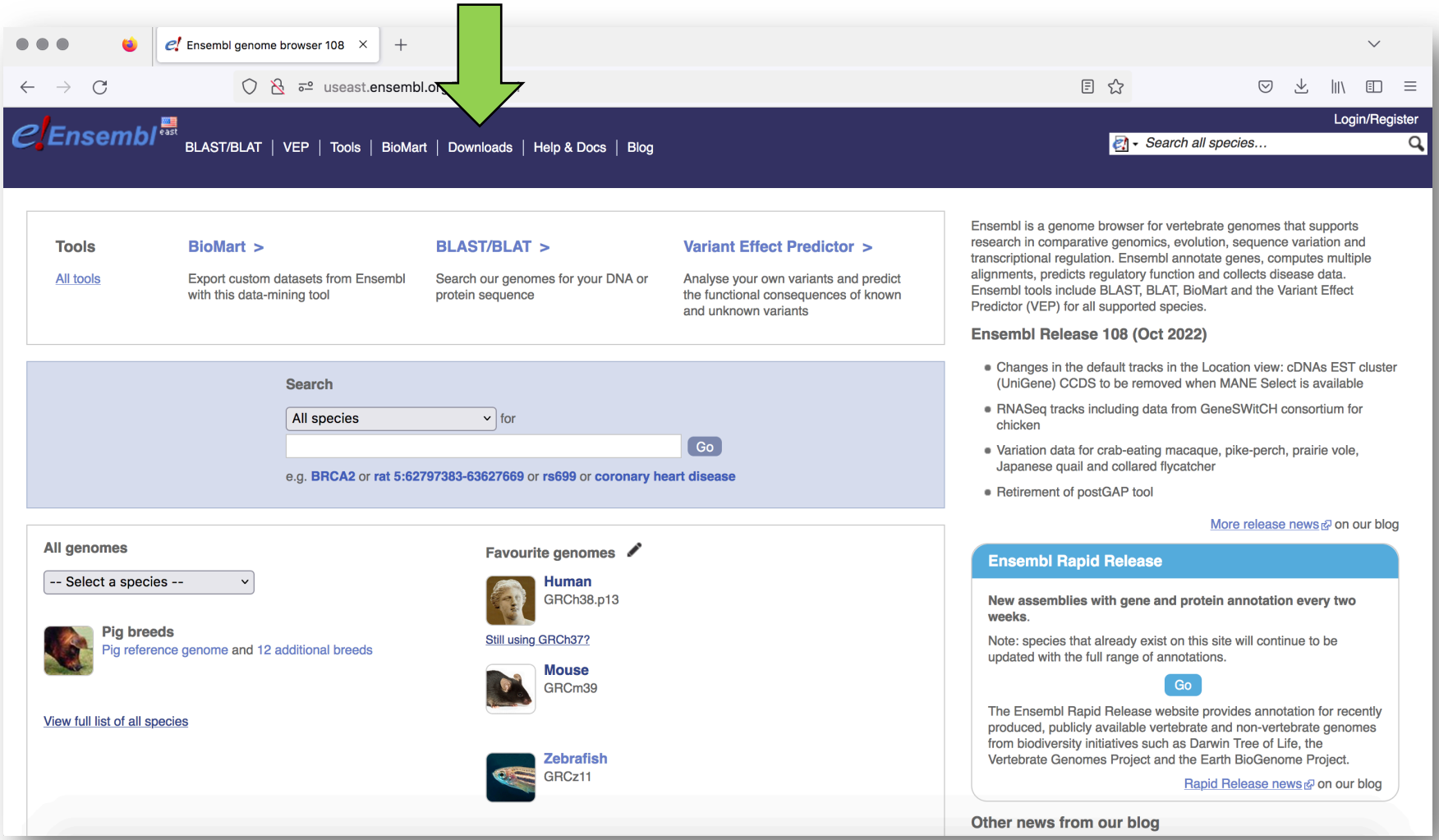
- The advantage of downloading from these databanks is that even if you work with different species, the file formats and naming conventions remain consistent

Good practice

- 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

Ensembl


Ensembl genome browser 108
useast.ensembl.org
Ensembl east
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search all species…
Tools
All tools
BioMart >
Export custom datasets from Ensembl with this data-mining tool
BLAST/BLAT >
Search our genomes for your DNA or protein sequence
Variant Effect Predictor >
Analyse your own variants and predict the functional consequences of known and unknown variants
Search
All species
for
Go
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease
All genomes
-- Select a species --
Pig breeds
Pig reference genome and 12 additional breeds
View full list of all species
Favourite genomes
Human
GRCh38.p13
Still using GRCh37?
Mouse
GRCm39
Zebrafish
GRCz11
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
More release news on our blog
Ensembl Rapid Release
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.
Go
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.
Rapid Release news on our blog
Other news from our blog

Accessing Ensembl Data

useast.ensembl.org/info/data/index.html

Ensembl east

BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search all species...

Using this website | Annotation and prediction | Data access | API & software | About us

In this section

- Exporting data via website
- API data access
- Public MySQL Serv
- FTP Download
  - Downloading with
- BioMart
  - BiomaRt Bioc R package
  - BioMart Perl API
  - BioMart RESTful access
  - Combining species datasets
  - How to use BioMart
- Virtual Machine

Search documentation...

Go

Help & Documentation > Accessing Ensembl Data

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.

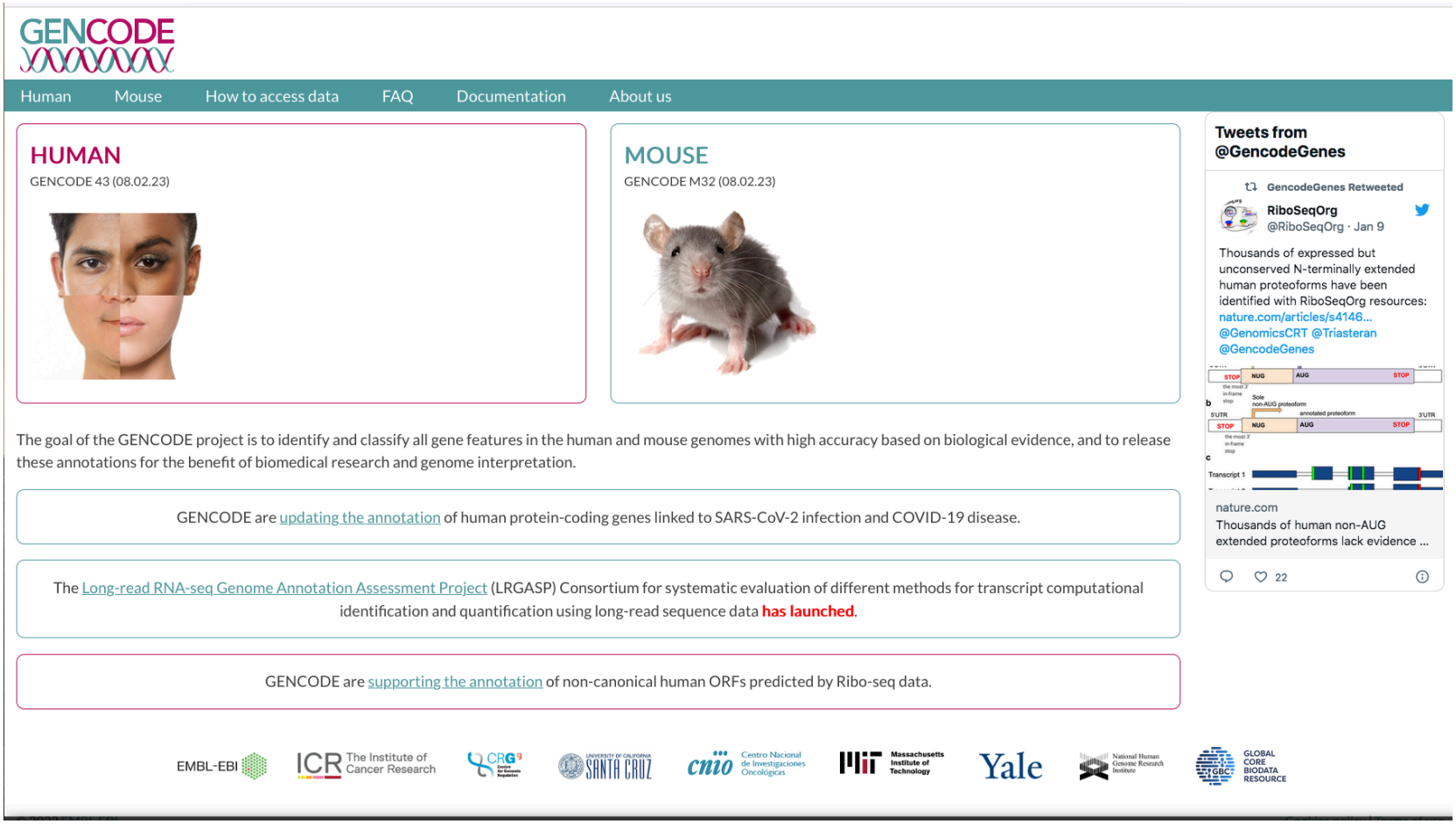
Entire databases are also available via FTP as MySQL dumps.

Complex cross-database queries

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

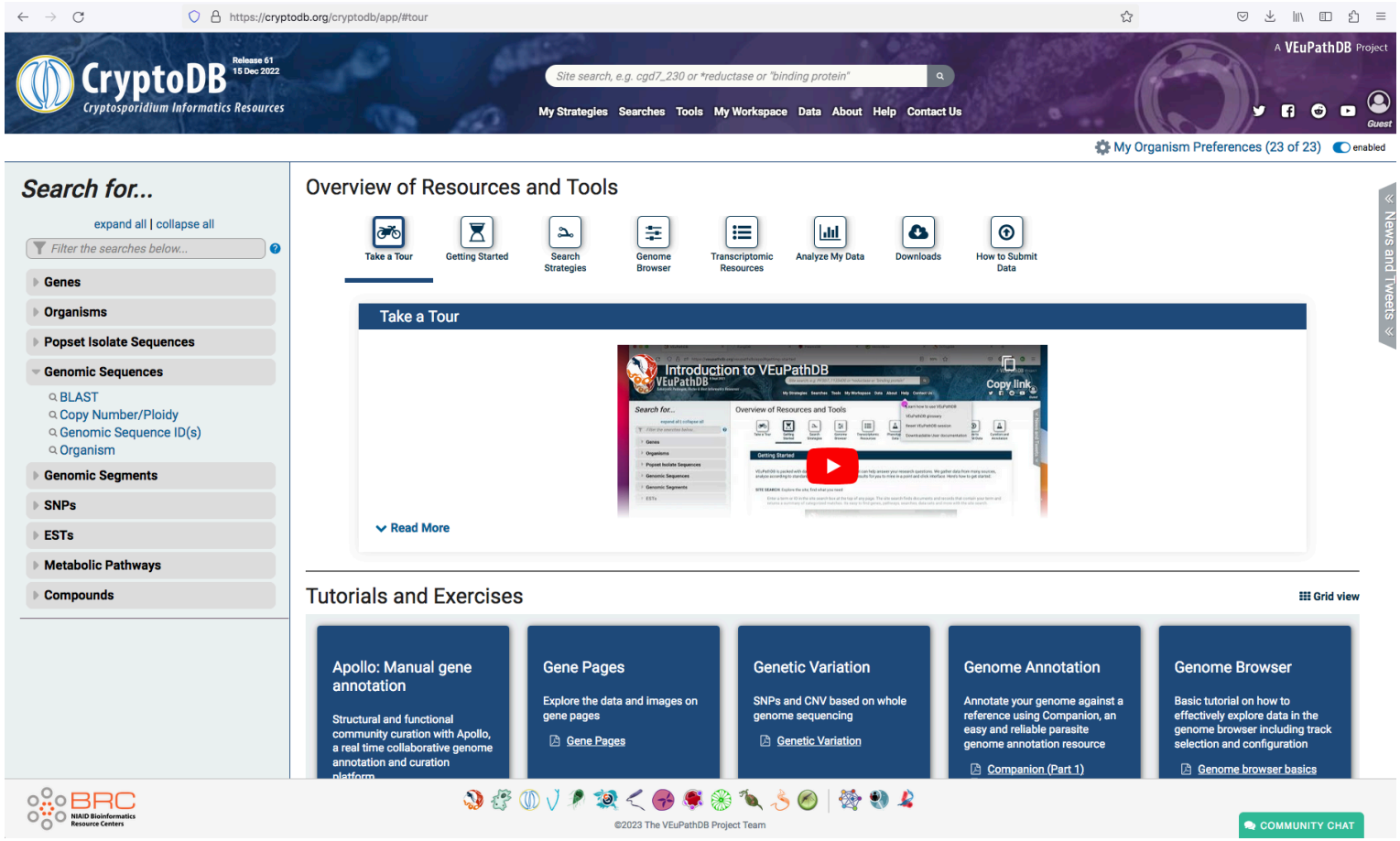
All data produced by the Ensembl project is freely available for your own use.

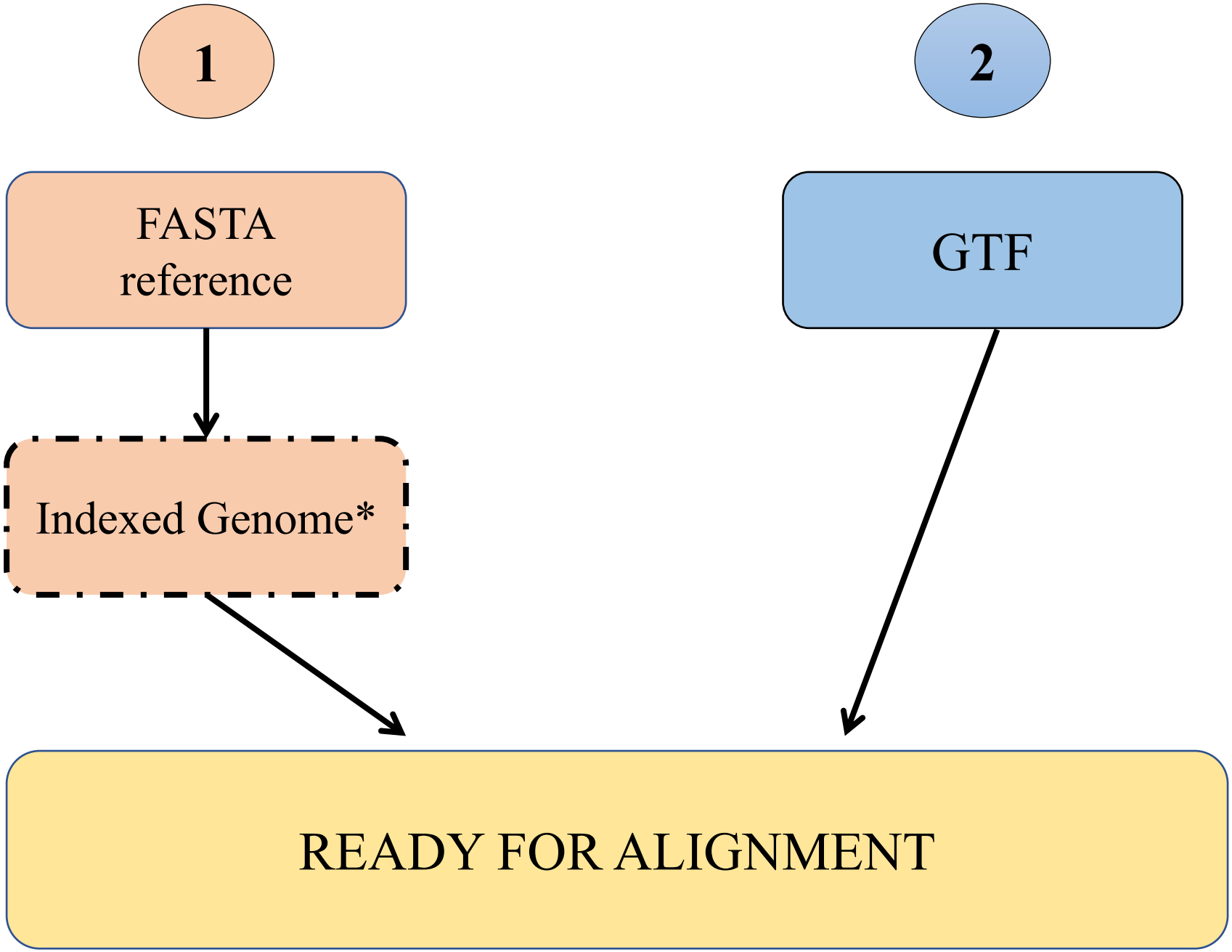
Gencode



https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_43/

Organism-specific biological databases: CryptoDB




1
2
FASTA reference
GTF
Indexed Genome*
READY FOR ALIGNMENT