

# Lecture 8: Mapping

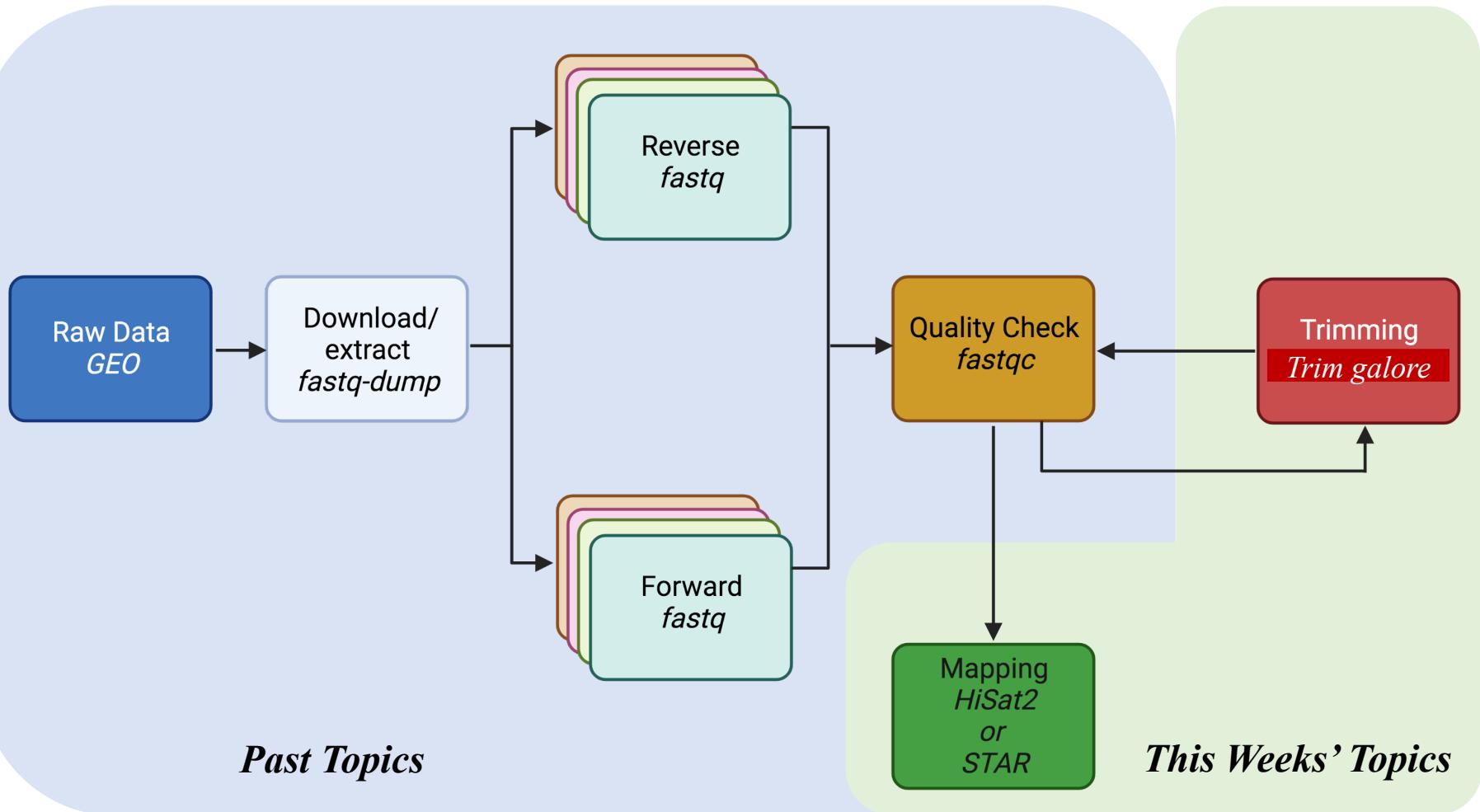
## February 27, 2024

# Outline

- Lecture for ~10 mins
- Class Activity #1 = 10 minutes
- Lecture for ~15 minutes
- Class Activity #2 = 10 minutes
  
- Any remaining time will be time to chat with me

# Learning objectives

- Explore the splice-aware mapper called STAR and HISAT2
- Index genome with HISAT2
- Begin to understand file outputs after alignment



# 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

# Trimming

# 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*
- Caution: Avoid aggressive clipping as this could cause miss-mapping

# Why optional?

- Modern “local aligners” like the STAR aligner, 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

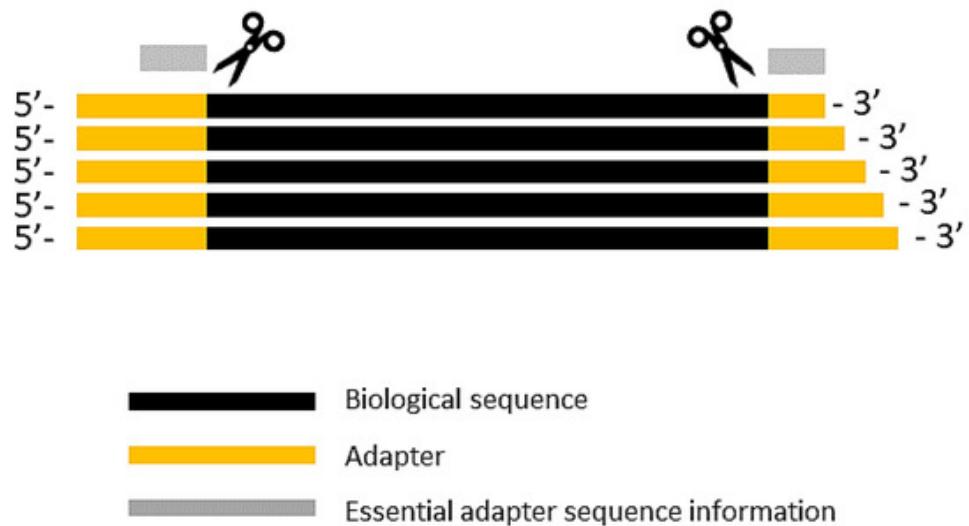
Preprocessing of FASTQ can be performed by a variety of programs

*fastp*

*Trimmomatic*

*Trim galore*

*Cutadapt*



# How would you run “trim galore”?

- 1) Install it in your personal VACC account and then be able to call it (ex. *multiQC*)
- 2) Or load it from the shared computing cluster using `module` package (ex. *fastqc*)

# Use module to load

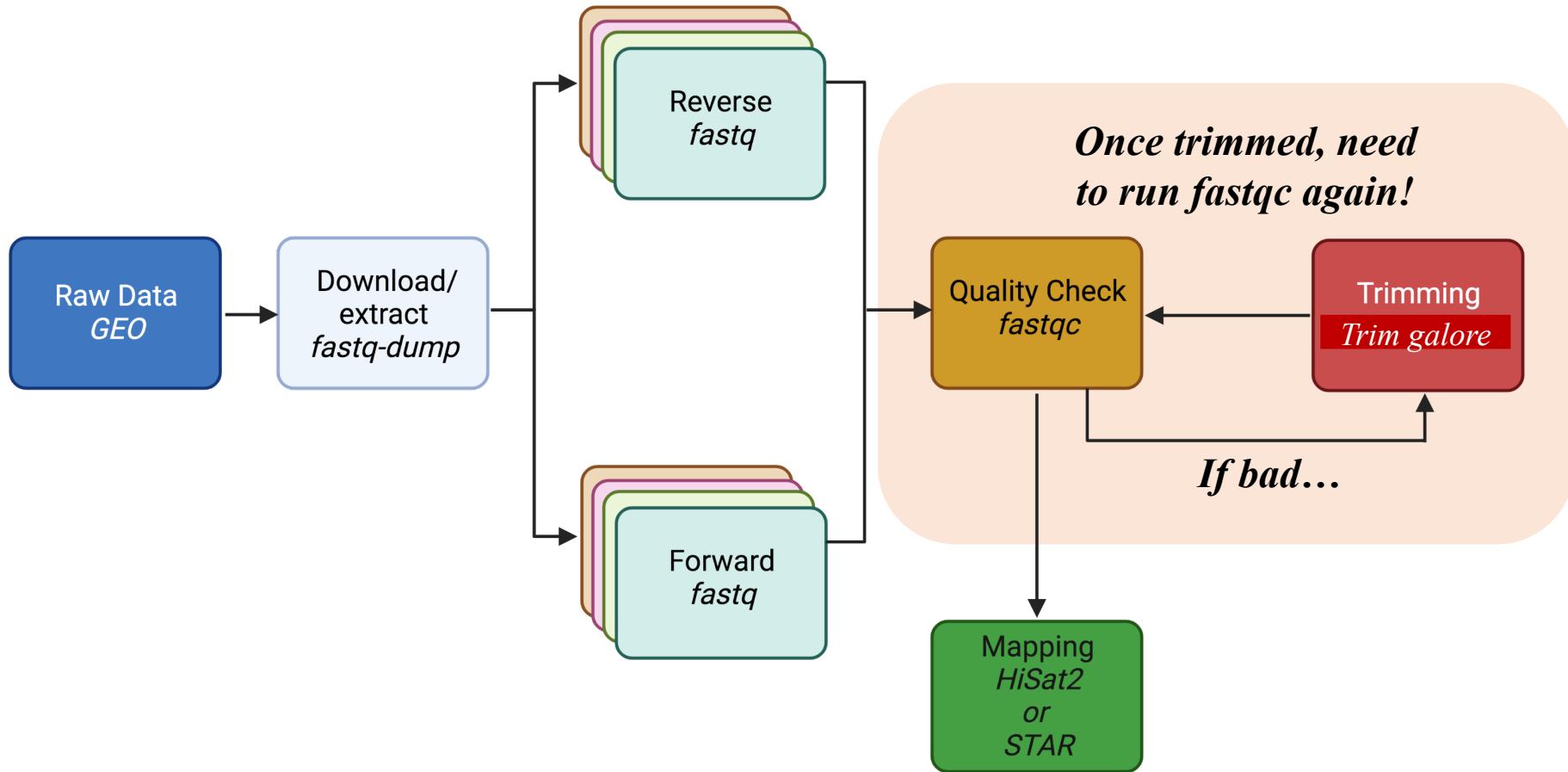
module avail – *to see what programs are available*

```
module load trimgalore-0.6.4-gcc-  
7.3.0-kml55yf
```

**Use tab completion!**

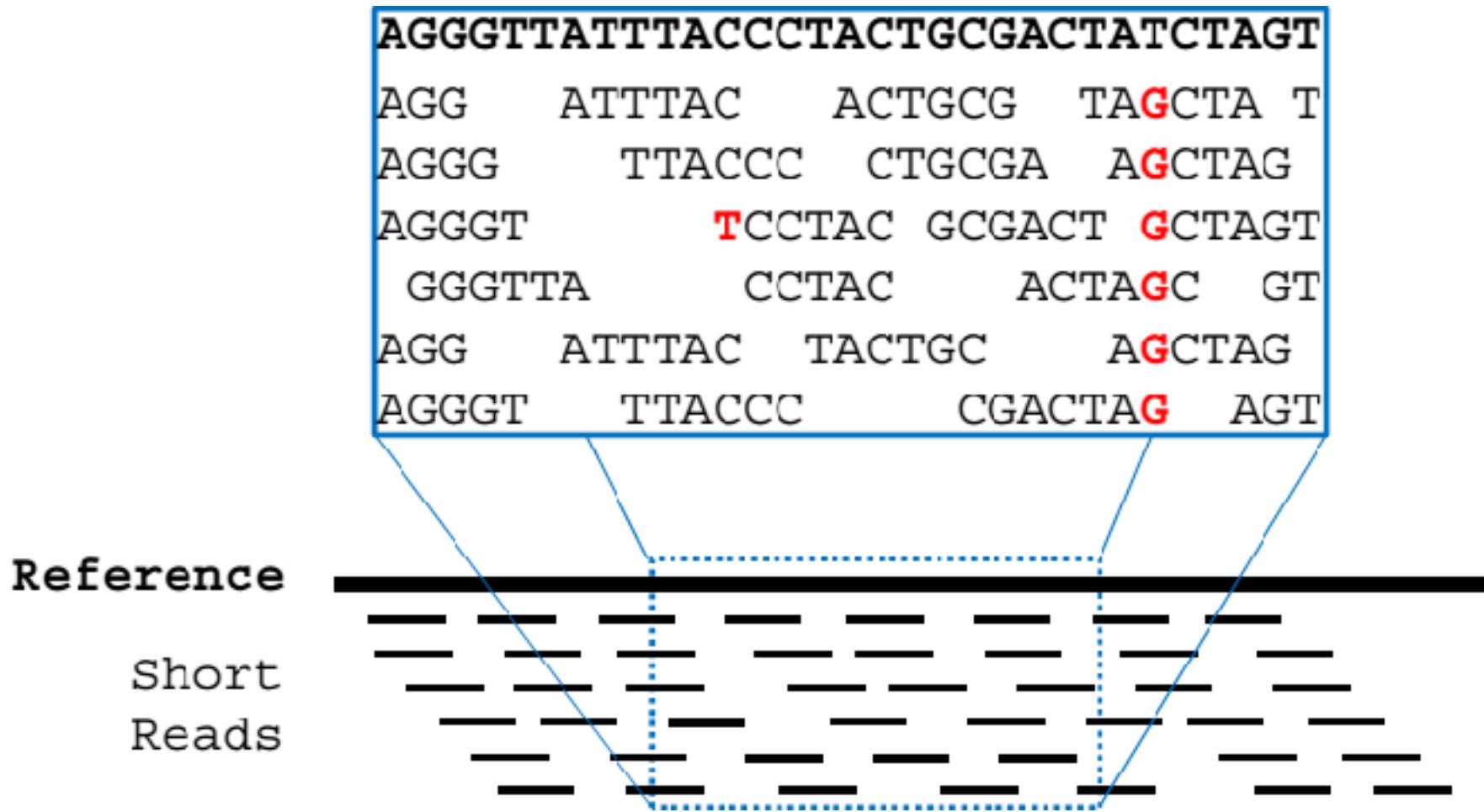
# Class Activity #1

- Open the Trim Galore exercise in the course website
- This will take ~10 minutes



# 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.

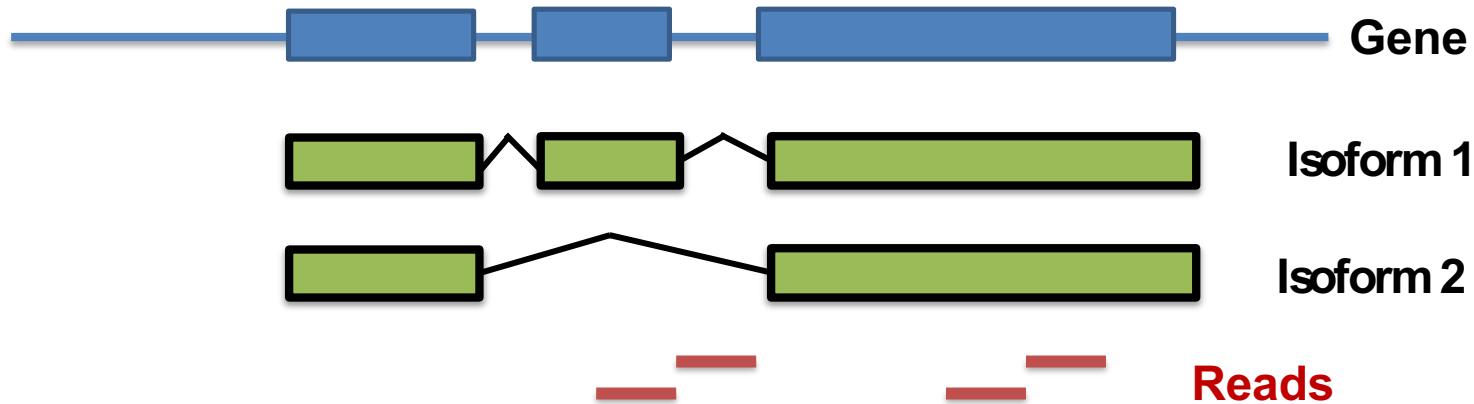
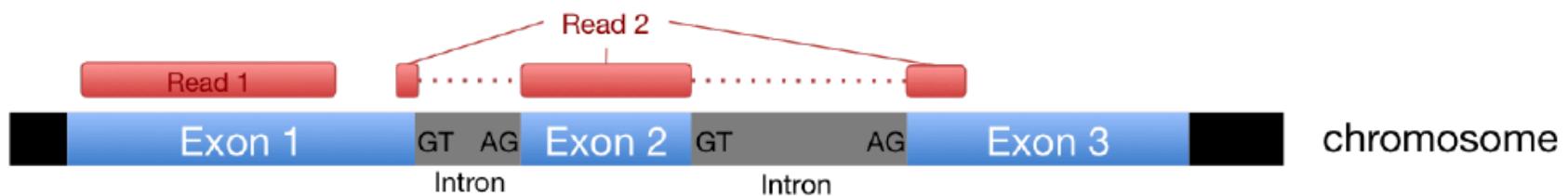
# The challenge with RNA-Seq specifically...

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

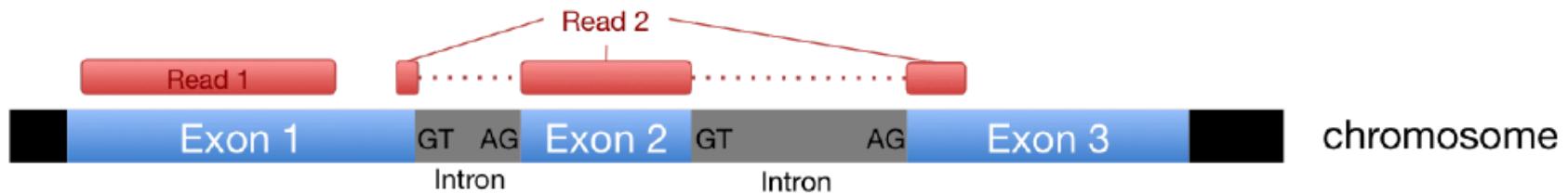


# Some opt to align to transcriptome instead

(a) Aligning to the transcriptome

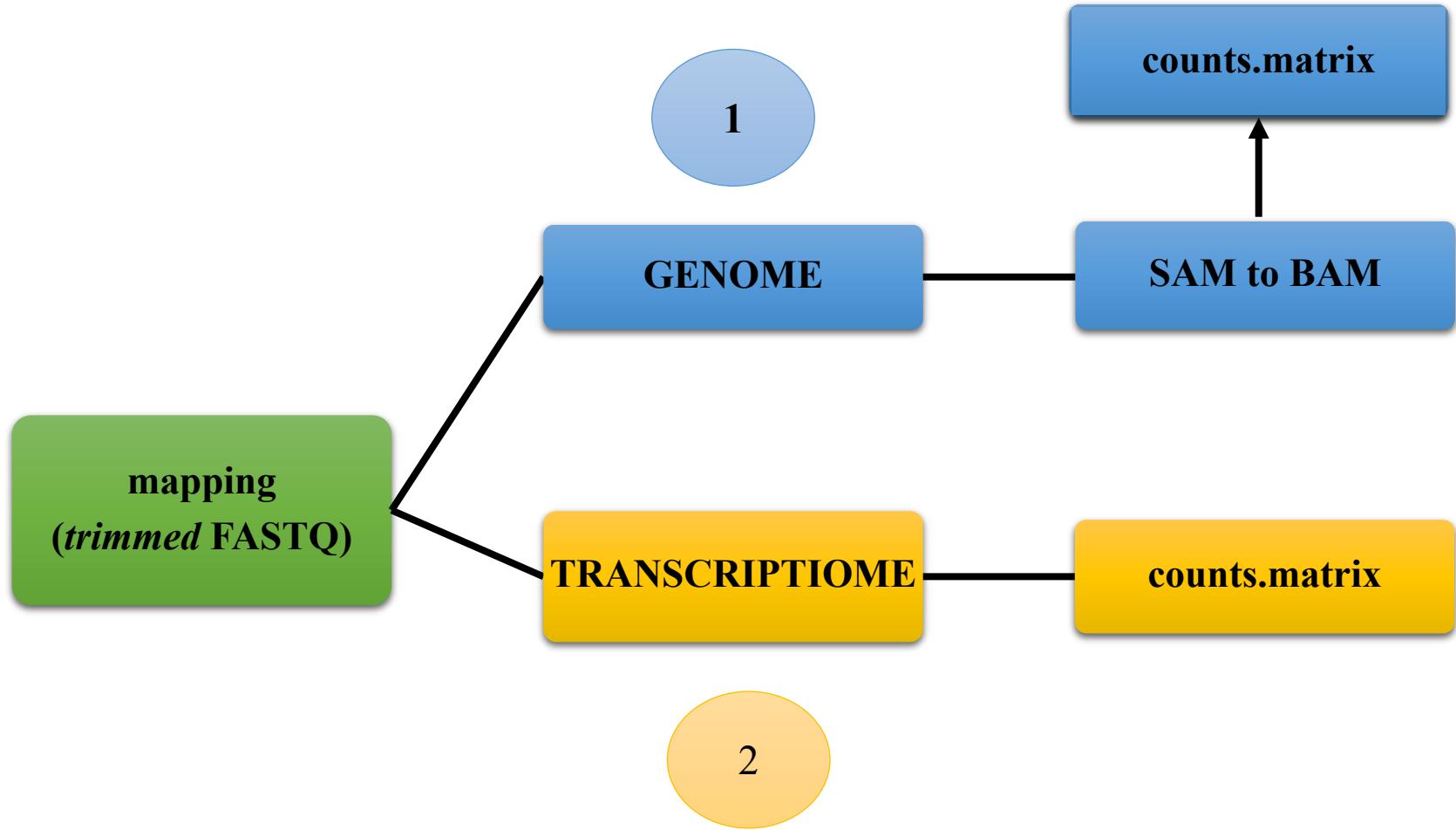


(b) Aligning to the genome



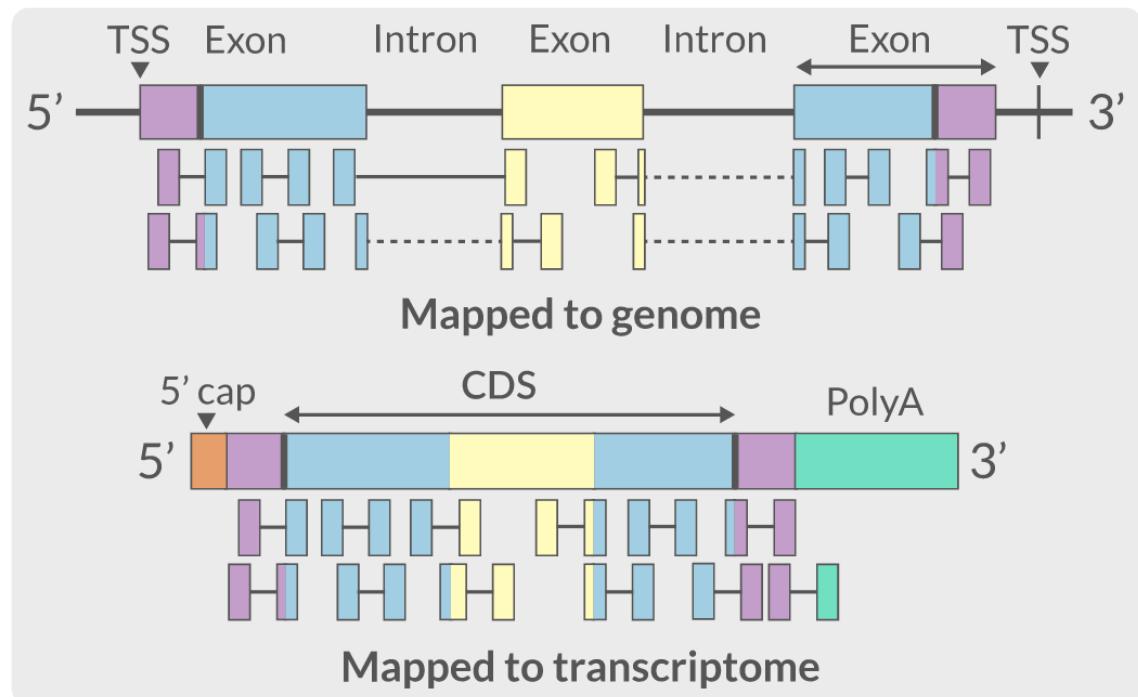
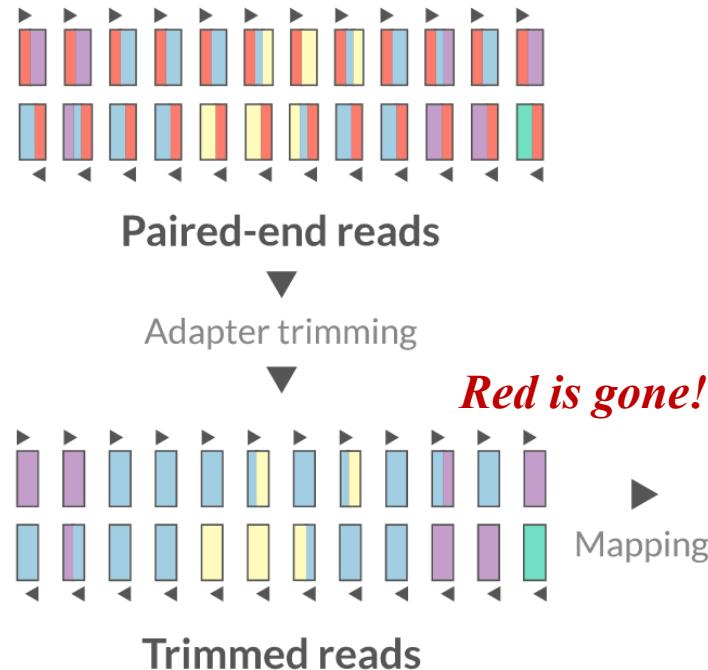
*If you are mapping reads to a transcriptome intron/exon boundaries are irrelevant*

# Mapping

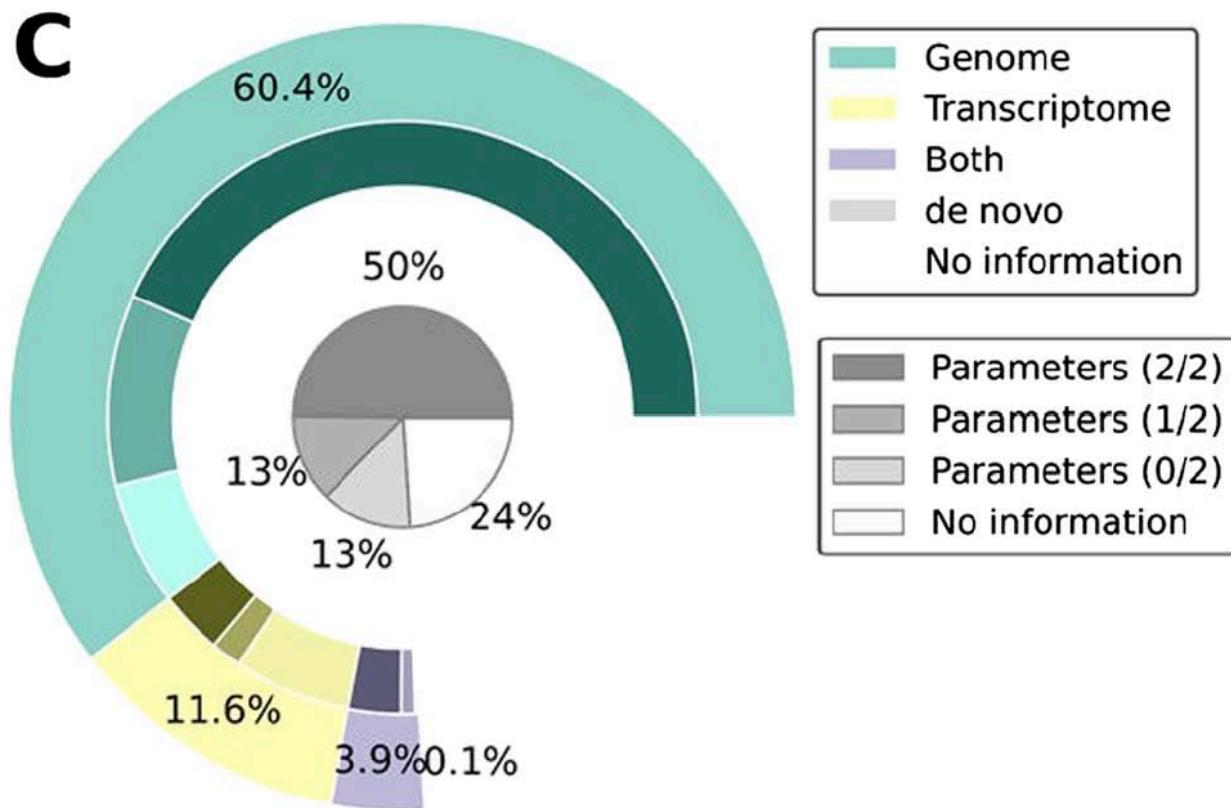


# Genome vs Transcriptome?

Red = adapter



# What the scientific community does



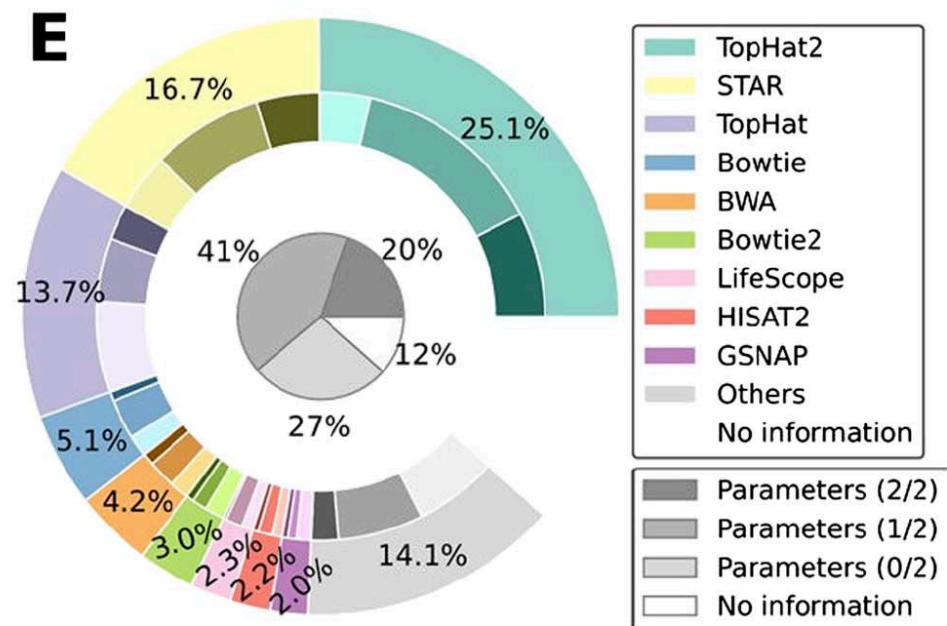
# Programs available

## Genome

- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

## Transcriptome

- Salmon
- Kallisto
- Sailfish



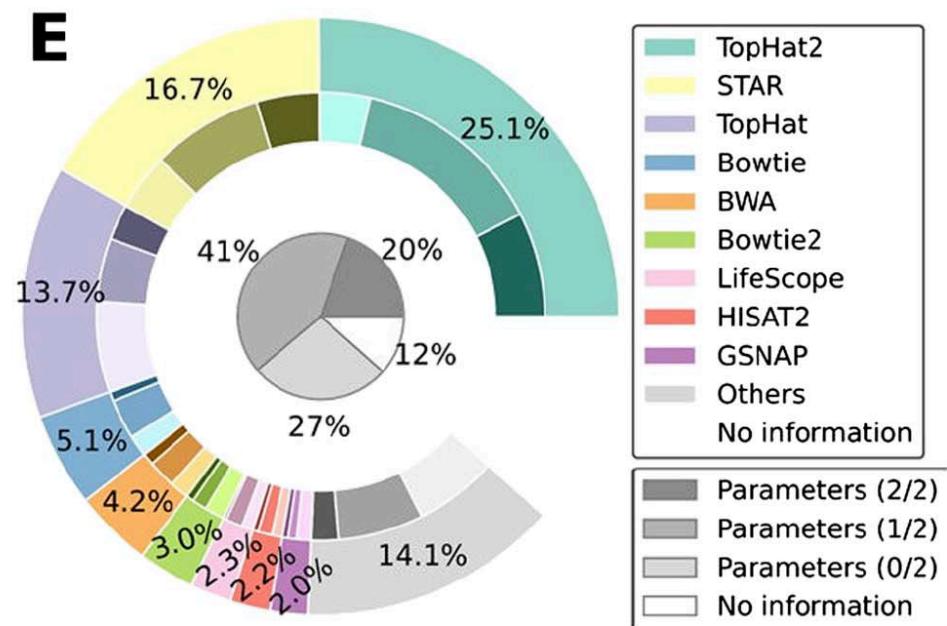
# Programs available

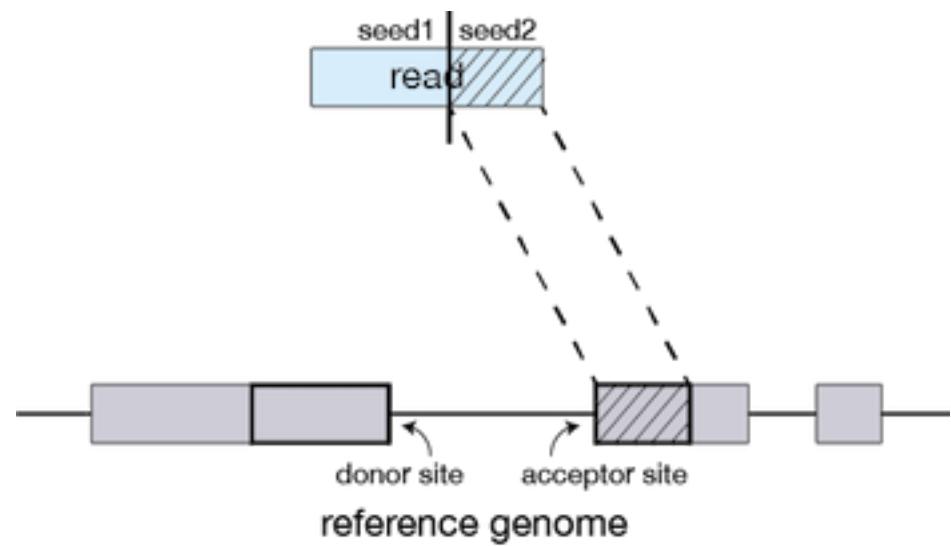
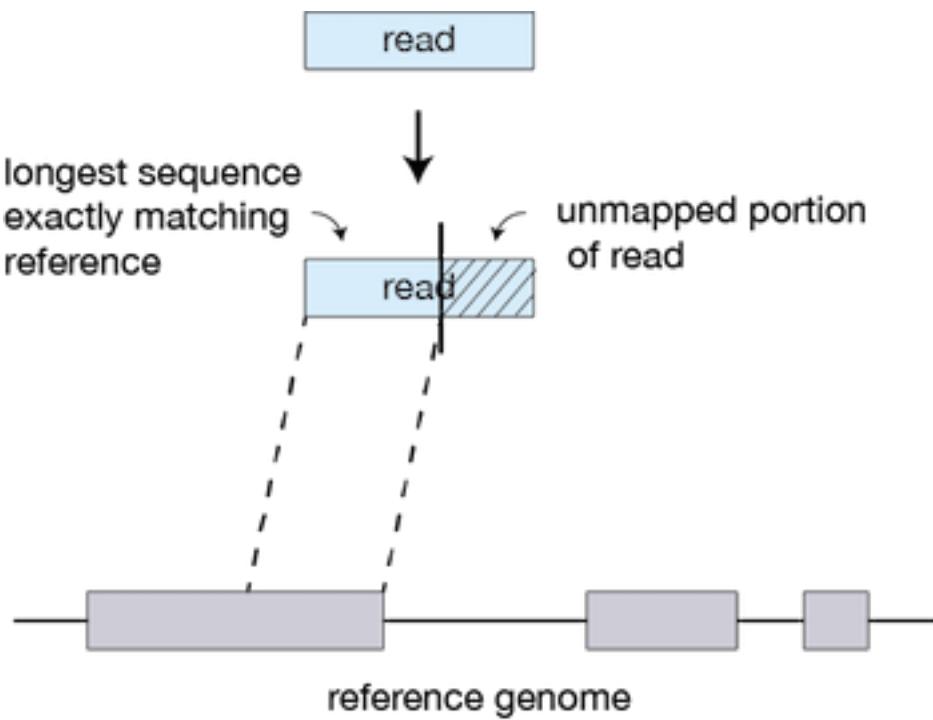
## Genome

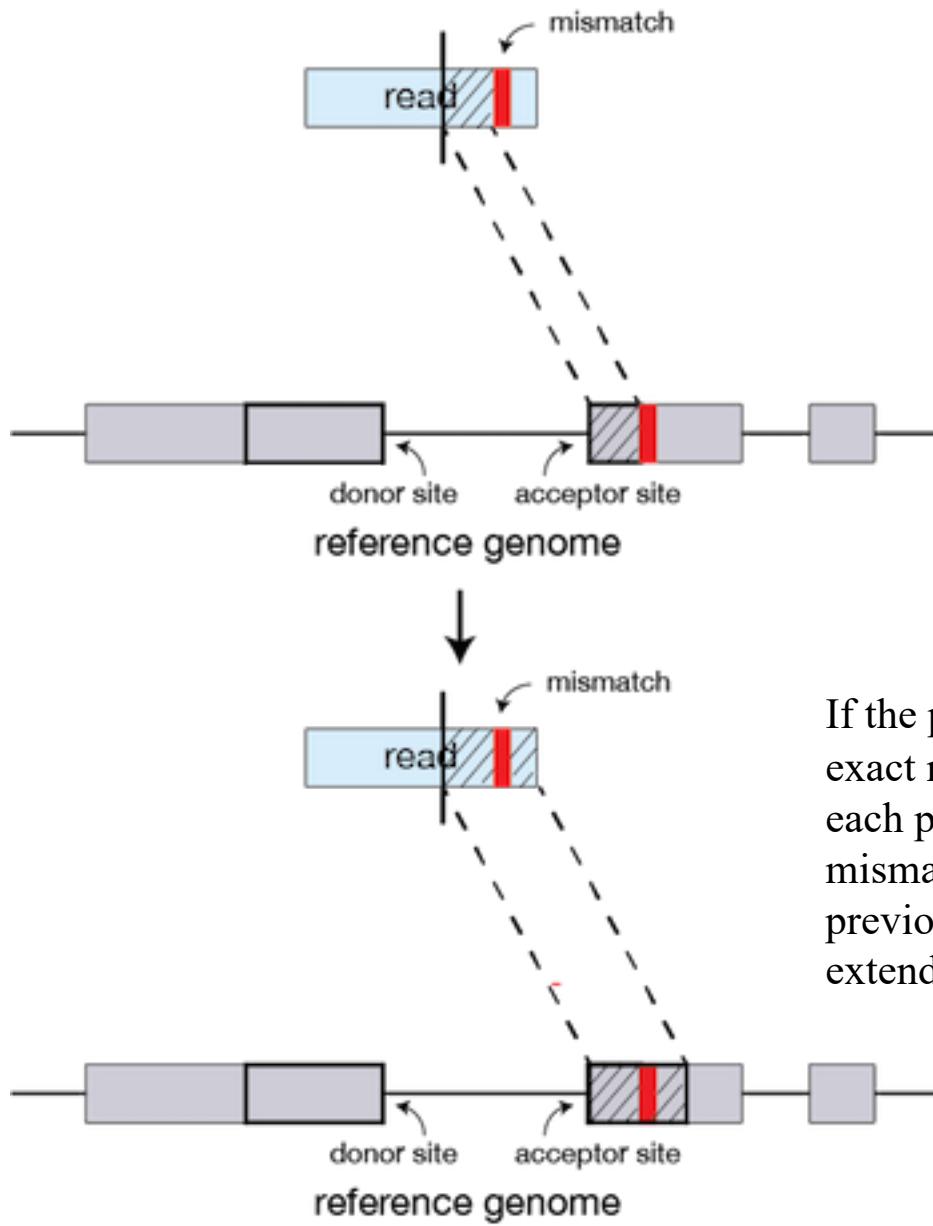
- TopHat2
- STAR
- Bowtie2
- BWA
- HiSat2

## Transcriptome

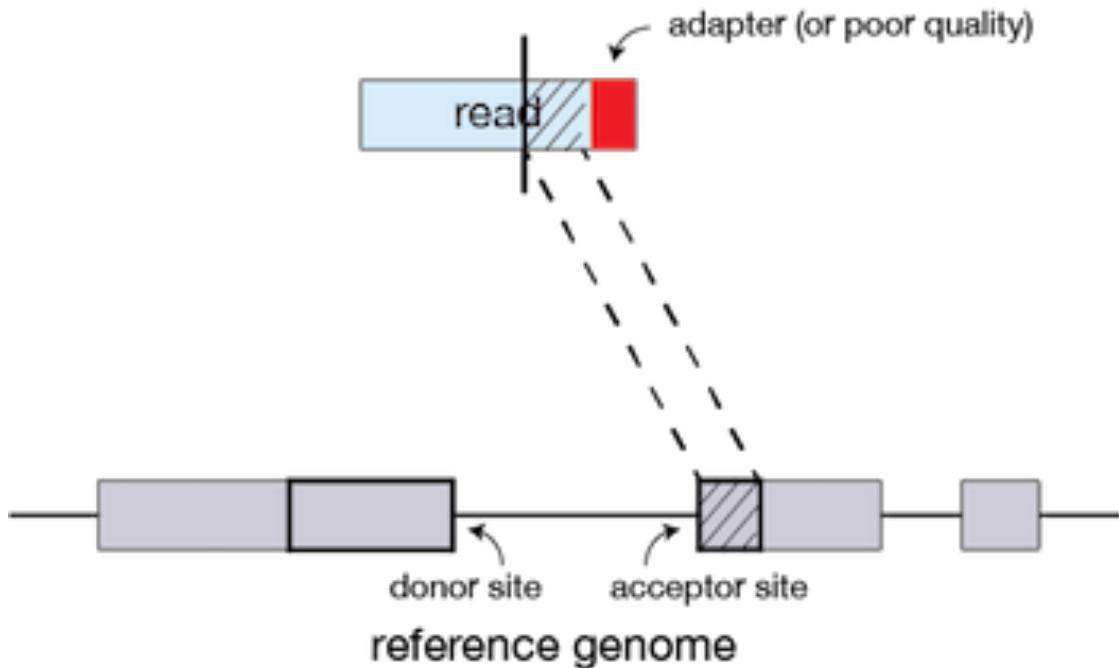
- Salmon
- Kallisto
- Sailfish







If the program does not find an exact matching sequence for each part of the read due to mismatches or indels, the previous mismatch will be extended.



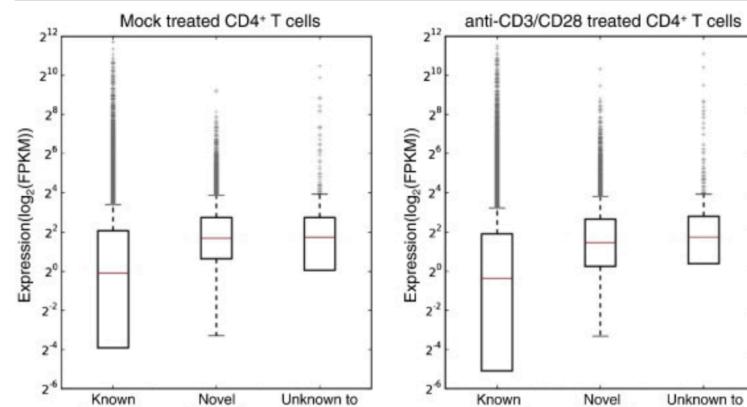
**If extension does not give a good alignment**, then the poor quality or adapter sequence (or other contaminating sequence) will be soft clipped.

# Genome

- Creation of BAM-visualization file
- >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<sup>+</sup> T cells and anti-CD3/CD28 treated cells.

# Aligners - Speed and Memory

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

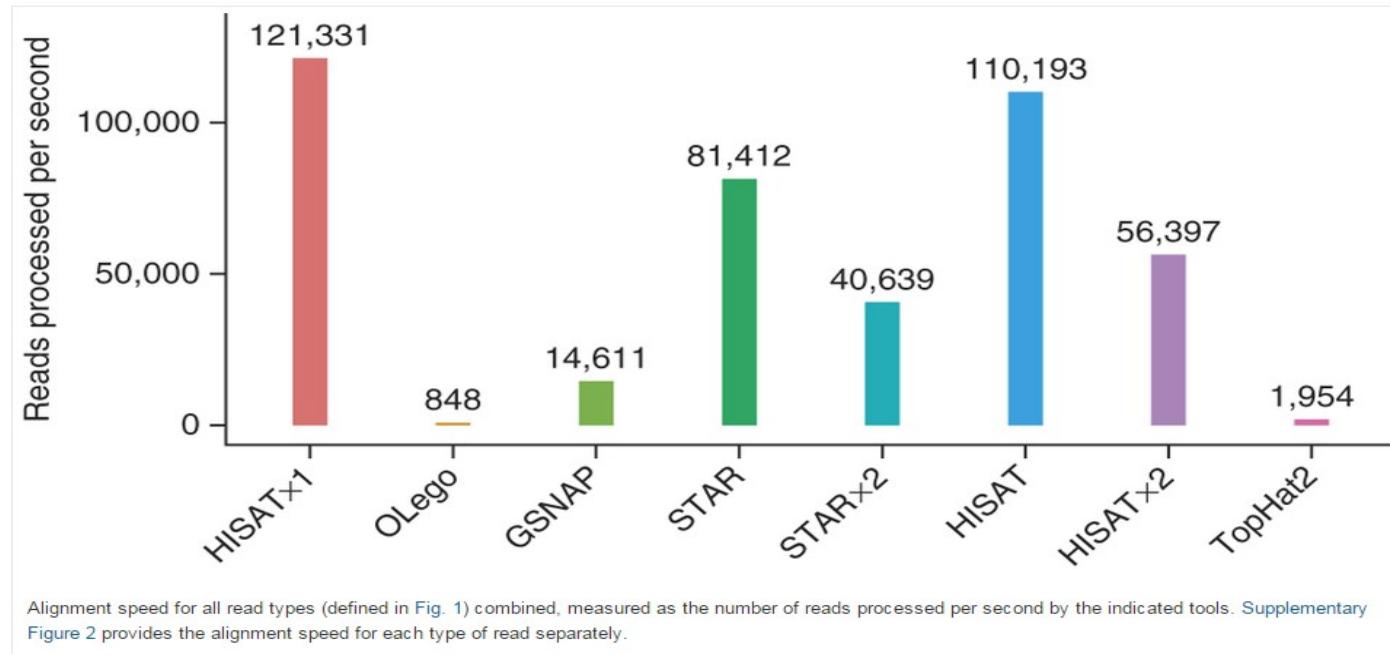
From

HISAT: a fast spliced aligner with low memory requirements

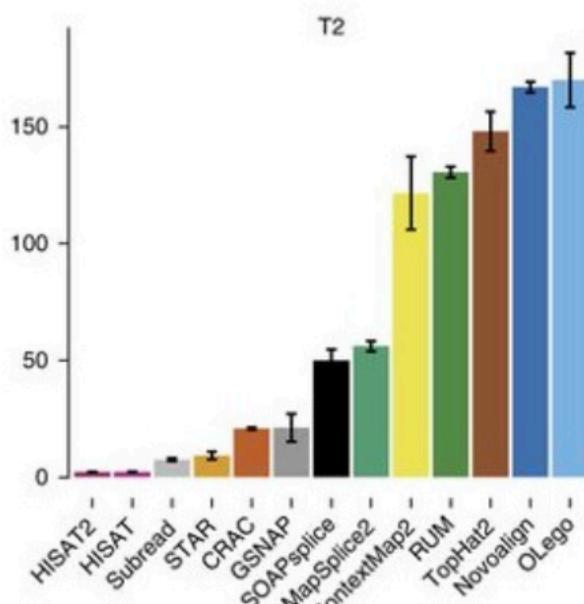
Daehwan Kim, Ben Langmead & Steven L Salzberg

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

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



# 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  
NCATCGTGGTATTCACATCTTTCTTATCAAATAAAAAGTTAACCTACTCAGTTATGCGCATACGTTTTGATGGCATTCCATA  
+  
#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  
GATCTAACATTTATTCCCCCTGCAAACATTTCAATCATTACATTGTCATTCCCCCTC  
CAAATTAAATTAGCCAGAGGCGACAACATACGACCTCTAAAAAAGGTGCTGTAACATG
```

# 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	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl start_codon	4430503	4430505	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl exon	4439303	4439440	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl CDS	4439303	4439440	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl exon	4443852	4443930	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl CDS	4443852	4443930	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl exon	4445846	4450423	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl CDS	4445846	4446022	.	gene_id "ENSACAG00000011126"; transcript_id
1	ensembl stop_codon	4446023	4446025	.	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

# Where to find genomic files

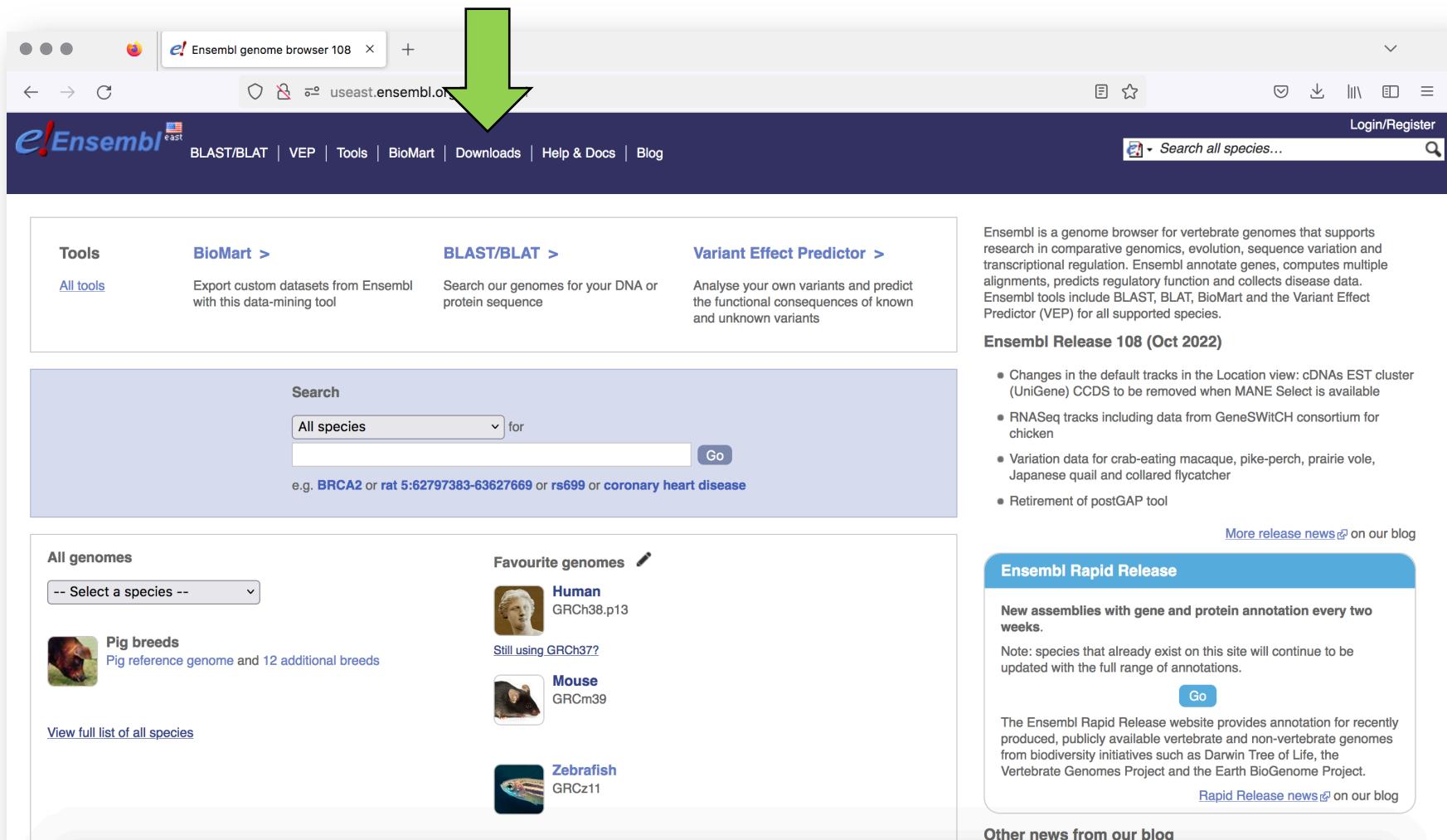
**General biological databases:** Ensembl, GENCODE, and UCSC

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

# 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



A screenshot of the Ensembl genome browser homepage. A large green arrow points downwards from the top center towards the search bar. The page features a dark blue header with the Ensembl logo and navigation links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. The main content area includes sections for Tools (All tools), BioMart, BLAST/BLAT, Variant Effect Predictor, and a general description of Ensembl. Below these are sections for All genomes (with a dropdown menu for species selection) and Favourite genomes (listing Human, Mouse, and Zebrafish). A prominent search bar at the top allows users to search across all species. To the right, there's a 'Ensembl Rapid Release' section with news about new assemblies and a 'Rapid Release news' link. Other news from the blog is also mentioned.

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates 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 GeneSWiCH 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

**e!Ensembl** east

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

Login/Register

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

In this section

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

Search documentation... Go

## 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!

  
Export data  
CAGATGAT  
AAATGTCT  
AAAAGAGCA  
CTGTCATGC  
ATAAAAGAA  
AGTGATACT

### 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.

# Organism-specific biological databases: CryptoDB

The screenshot shows the homepage of the CryptoDB website, which is part of the VEuPathDB Project. The header includes the logo, release information (Release 61, 15 Dec 2022), a search bar, and navigation links for strategies, searches, tools, workspace, data, about, help, and contact us. A sidebar on the left provides a search interface for various biological entities like genes, organisms, and genomic sequences. The main content area features an 'Overview of Resources and Tools' section with links to a tour, getting started, search strategies, genome browser, transcriptomic resources, analyze my data, downloads, and how to submit data. Below this is a 'Take a Tour' section with a video thumbnail. The footer includes logos for BRC (NAID Bioinformatics Resource Centers) and the VEuPathDB Project Team, along with a community chat link.

https://cryptodb.org/cryptodb/app/#tour

CryptoDB  
Cryptosporidium Informatics Resources

Release 61  
15 Dec 2022

Site search, e.g. cgd7\_230 or \*reductase or "binding protein"

My Strategies Searches Tools My Workspace Data About Help Contact Us

A VEuPathDB Project

My Organism Preferences (23 of 23) enabled

Search for...

expand all | collapse all

Filter the searches below...

Genes

Organisms

Popset Isolate Sequences

Genomic Sequences

- BLAST
- Copy Number/Ploidy
- Genomic Sequence ID(s)
- Organism

Genomic Segments

SNPs

ESTs

Metabolic Pathways

Compounds

Overview of Resources and Tools

Take a Tour

Getting Started

Search Strategies

Genome Browser

Transcriptomic Resources

Analyze My Data

Downloads

How to Submit Data

Take a Tour

Read More

Tutorials and Exercises

Apollo: Manual gene annotation

Gene Pages

Genetic Variation

Genome Annotation

Genome Browser

Grid view

BRC NAID Bioinformatics Resource Centers

©2023 The VEuPathDB Project Team

COMMUNITY CHAT

1

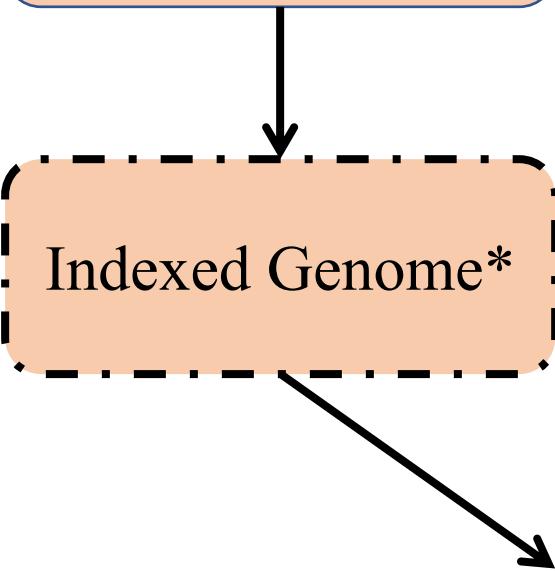
FASTA  
reference

2

GTF

Indexed Genome\*

READY FOR ALIGNMENT



# Class activity #2 indexing genomes

# 1-on-1 Time

- Conda installation – I'm going to triage these for another time. Apologies, but I would like to get everyone ready for HW#8.
- If you want my advice on a paper/dataset, please email the paper and GEO accession number and I will get back to you Wednesday.
- Other questions: HW#8 specific, i.e. How to fill out Metadata? Please stay.
- If you have any other questions, not in these categories, please raise your hand now.