Bulk RNA-seq vs DNA-seq Analysis

Ngày 10 tháng 05 năm 2025

TS. Lưu Phúc Lợi

Email: luu.p.loi@googlemail.com

Zalo: 0901802182

Content

- Aim of the methods
- Library prep for RNA-seq
- Upstream Analysis Workflow
- Downstream Analysis Workflow
- From BAM to Count
- Differential Gene Expression
- Enrichment of DEGs with Over Representation Analysis (ORA)
- RNA-seq databases: GTEx and TCGA

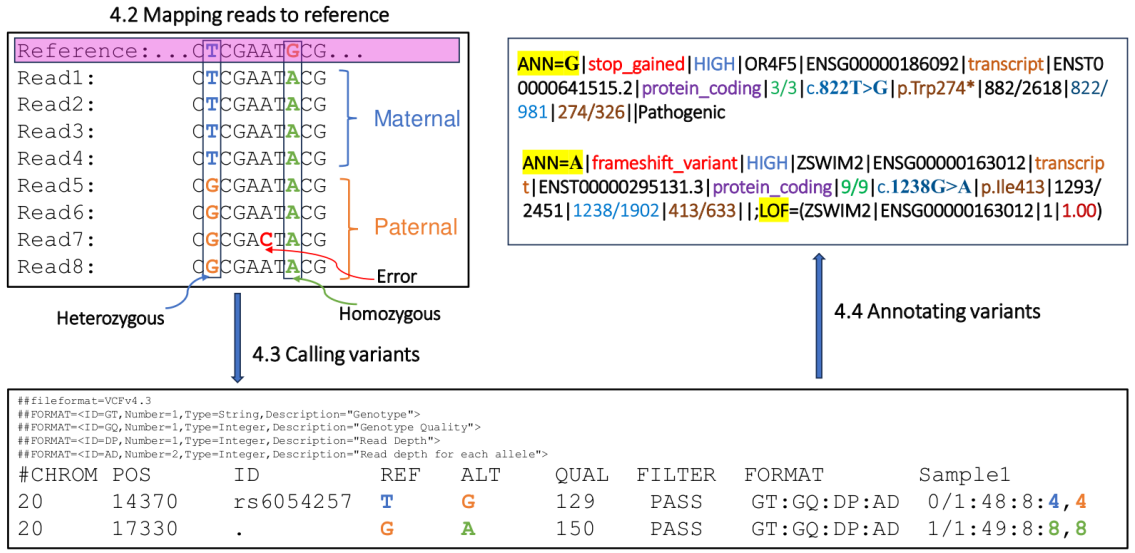
Aim of the methods

Bulk DNA-seq?

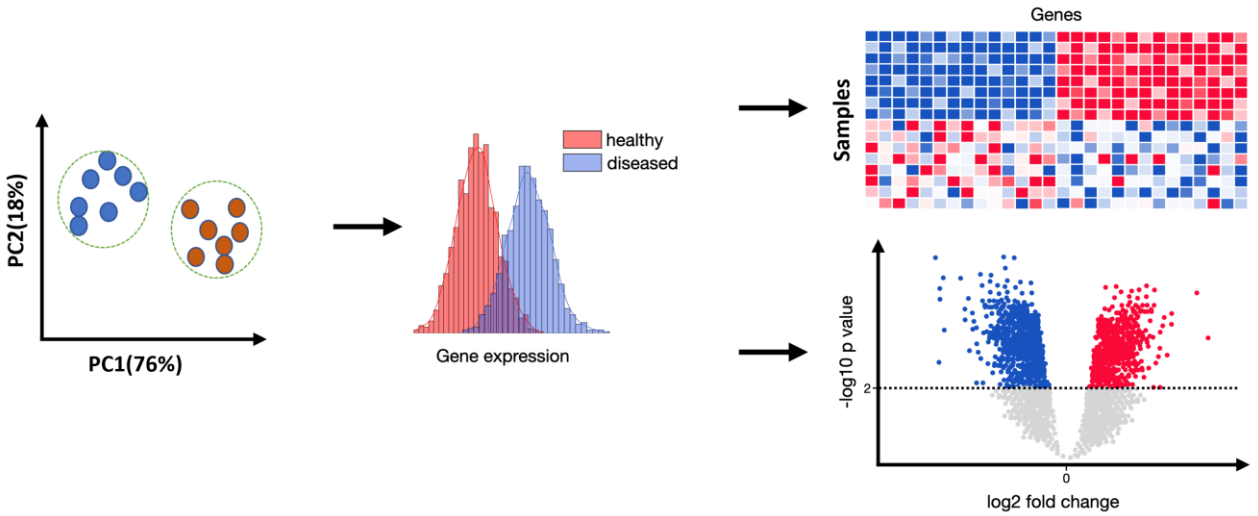
Bulk RNA-seq?

Aim of the methods

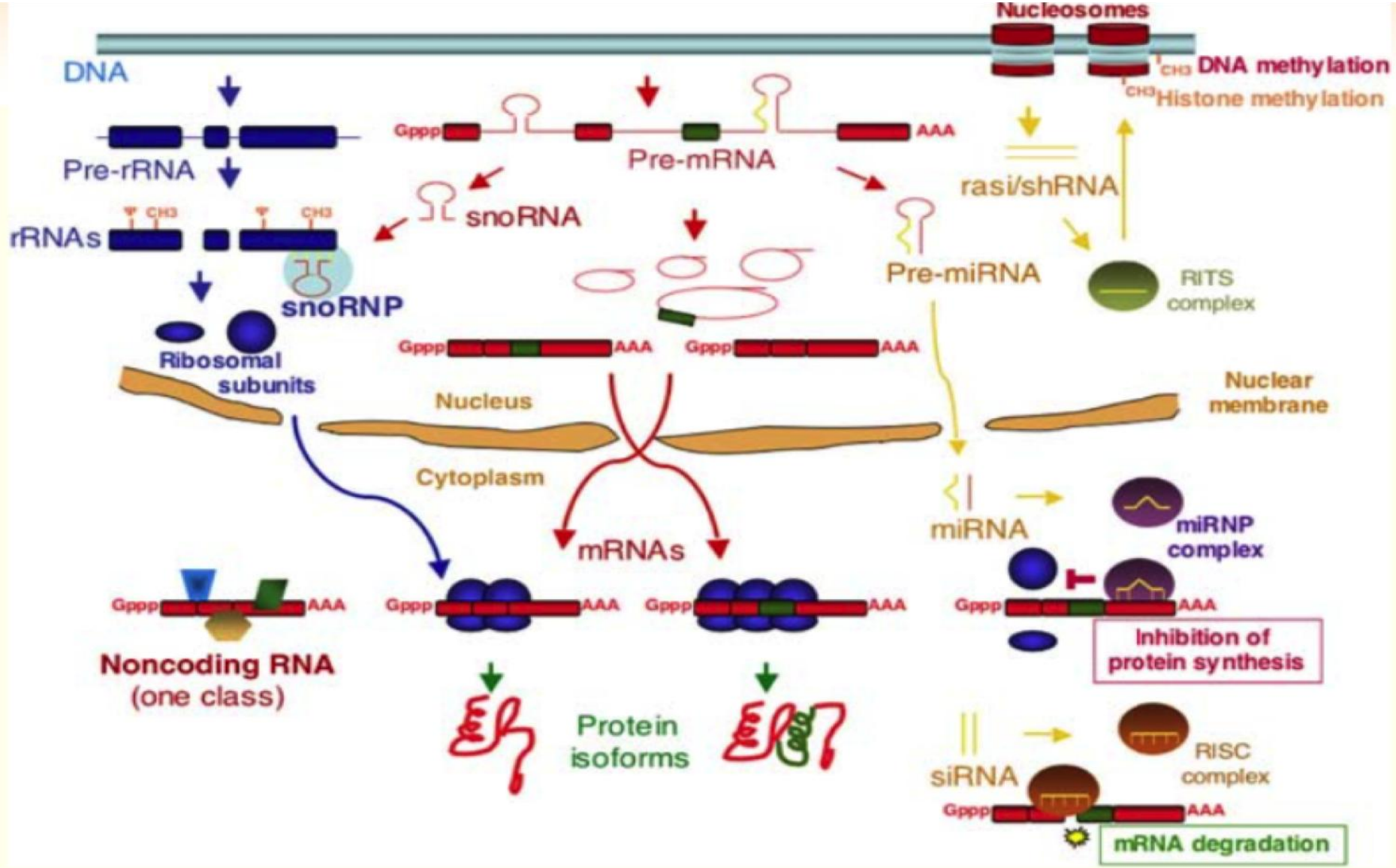
Bulk DNA-seq



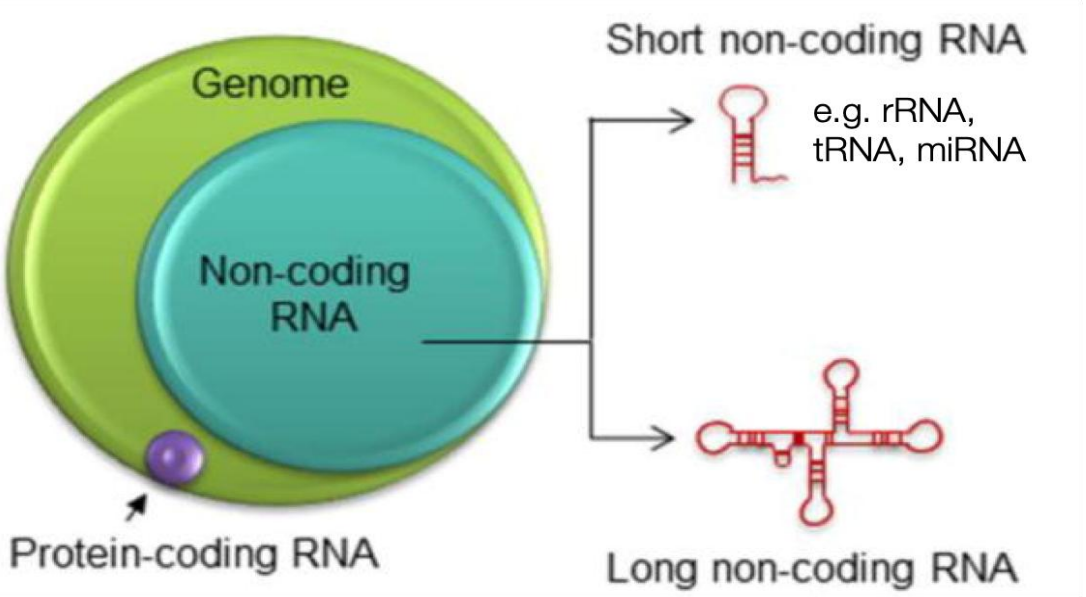
Bulk RNA-seq



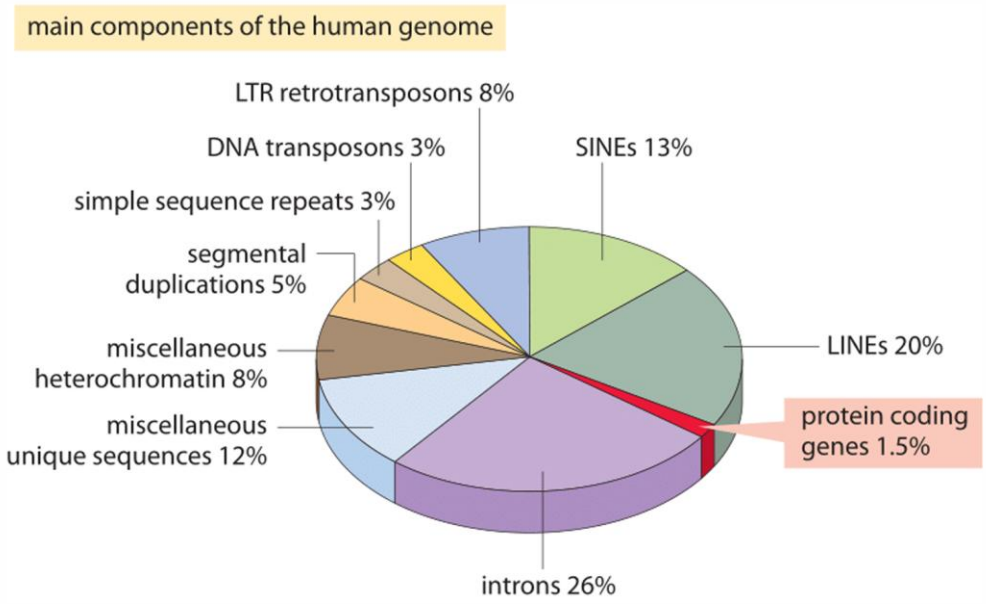
Different types of RNA



Different types of RNA



Parasramka et al. (2016)

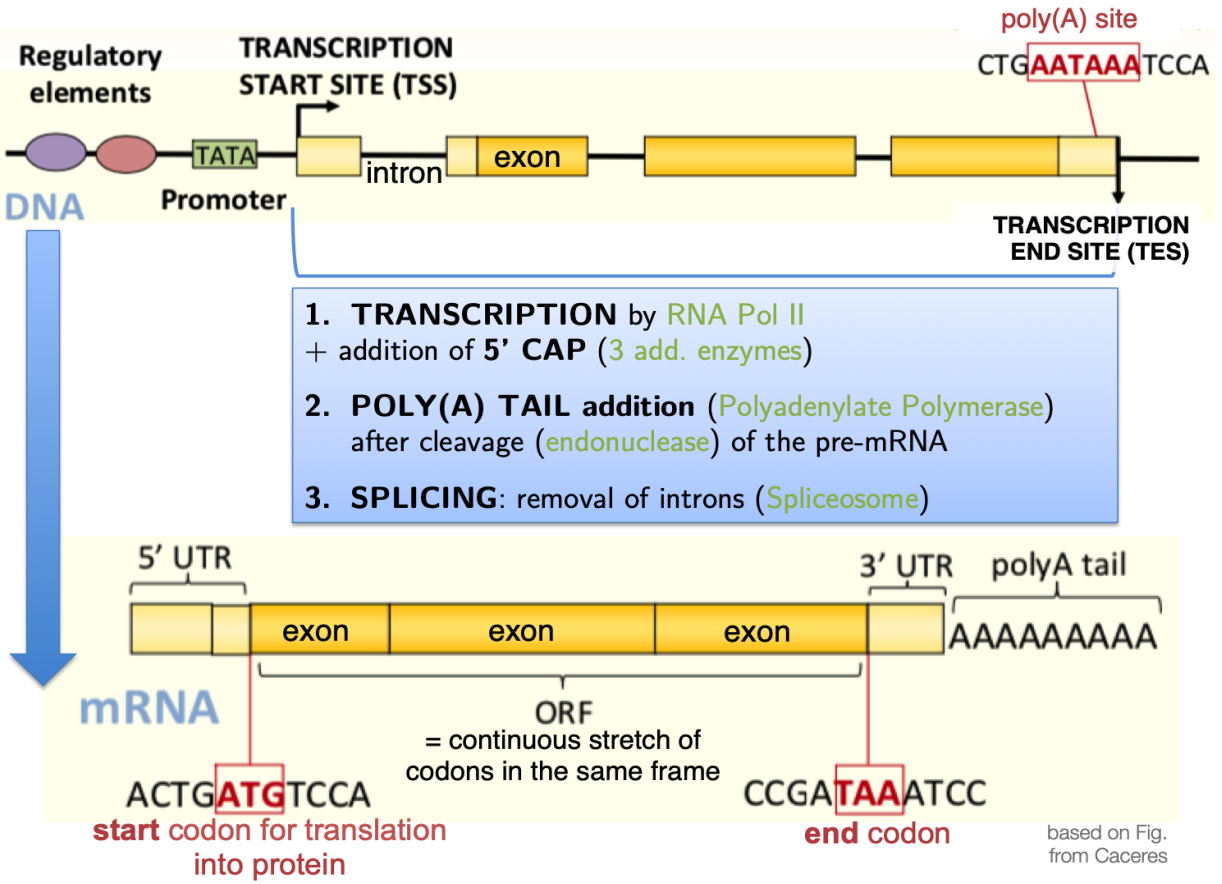


» How many genes are in a genome? (bionumbers.org)

Sequencing library prep protocol depends on the RNA properties

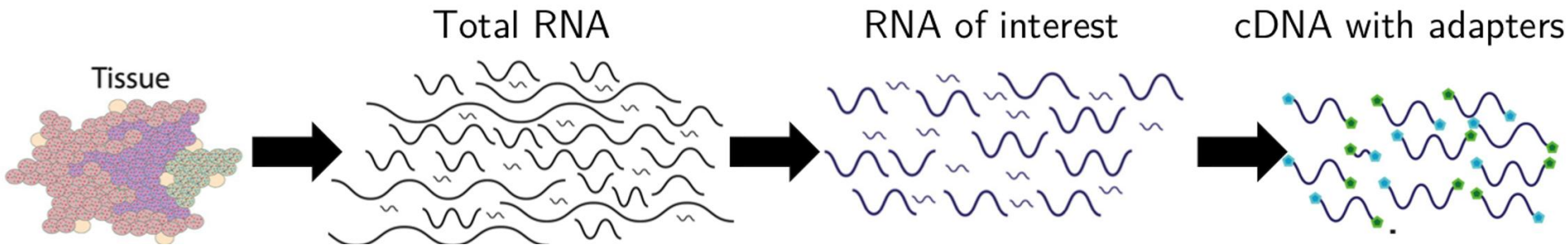
It is not a one-size-fits-all situation!

- Abundance and stability
  - ▶rRNA: 90-95% (!)
  - ▶tRNA: 3-5%
  - ▶mRNA: 2%
  - ▶all other non-coding RNAs: *well* below 1%
- Cellular location
  - ▶most are in the cytoplasm
- Size
  - ▶miRNAs: 18-23bp
  - ▶mRNA: several 100 to 1000 bp
- Specific sequences/modifications
  - ▶poly(A) tails of mRNA
  - ▶2D structure
  - ▶antisense transcripts



General steps of RNA-seq preparation

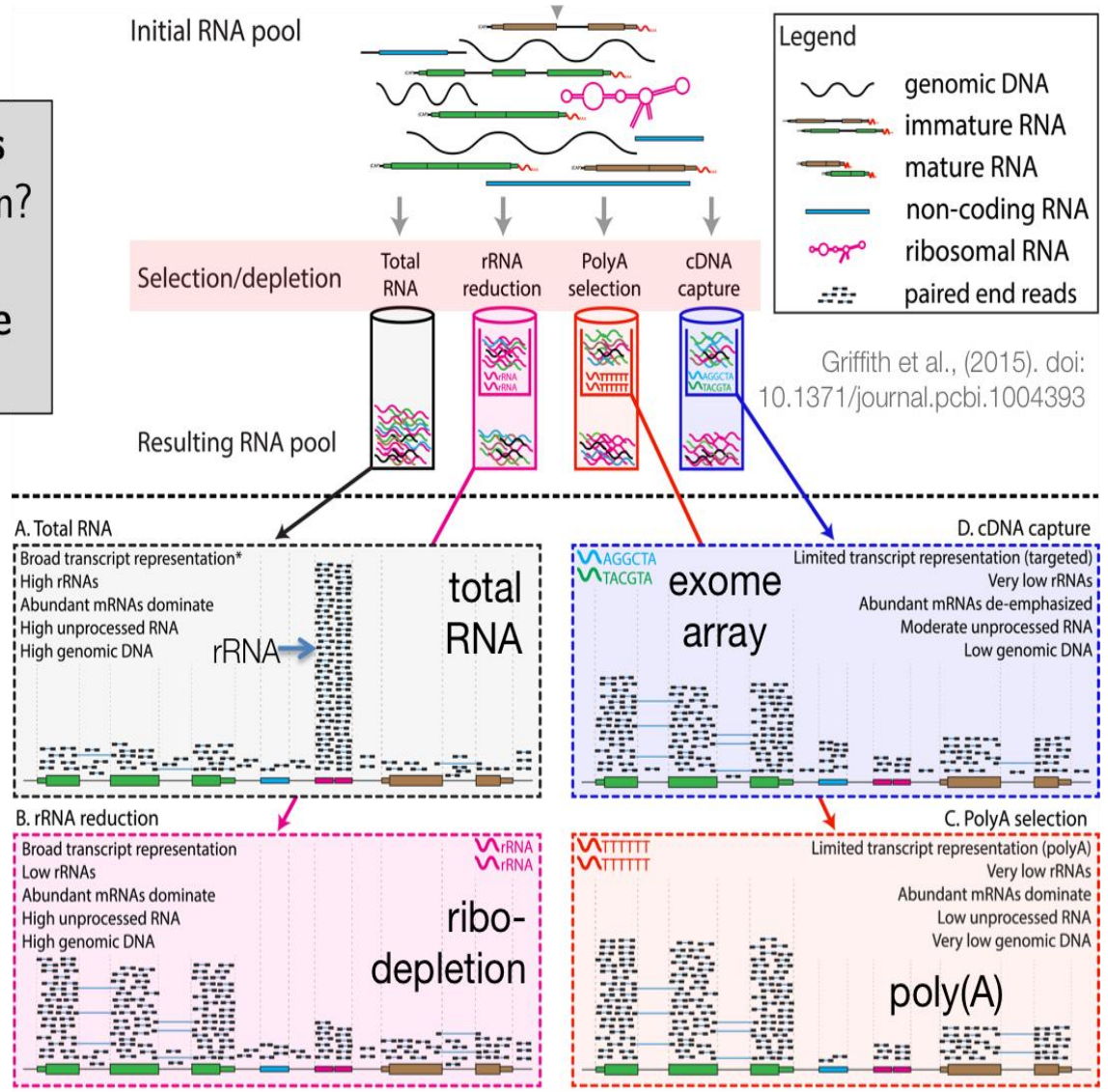
1. RNA extraction (cell lysis, RNA purification)
2. enrichment of the RNA of interest
   - mRNA: poly(A) enrichment vs. ribosomal-depletion
   - small RNAs: size-based enrichment
3. fragmentation (ca. 200 bp)
4. cDNA synthesis
5. library prep to obtain cDNA with adapters for sequencing



General steps of RNA-seq preparation

which **transcripts** are you interested in?

what type of **noise** can you tolerate?



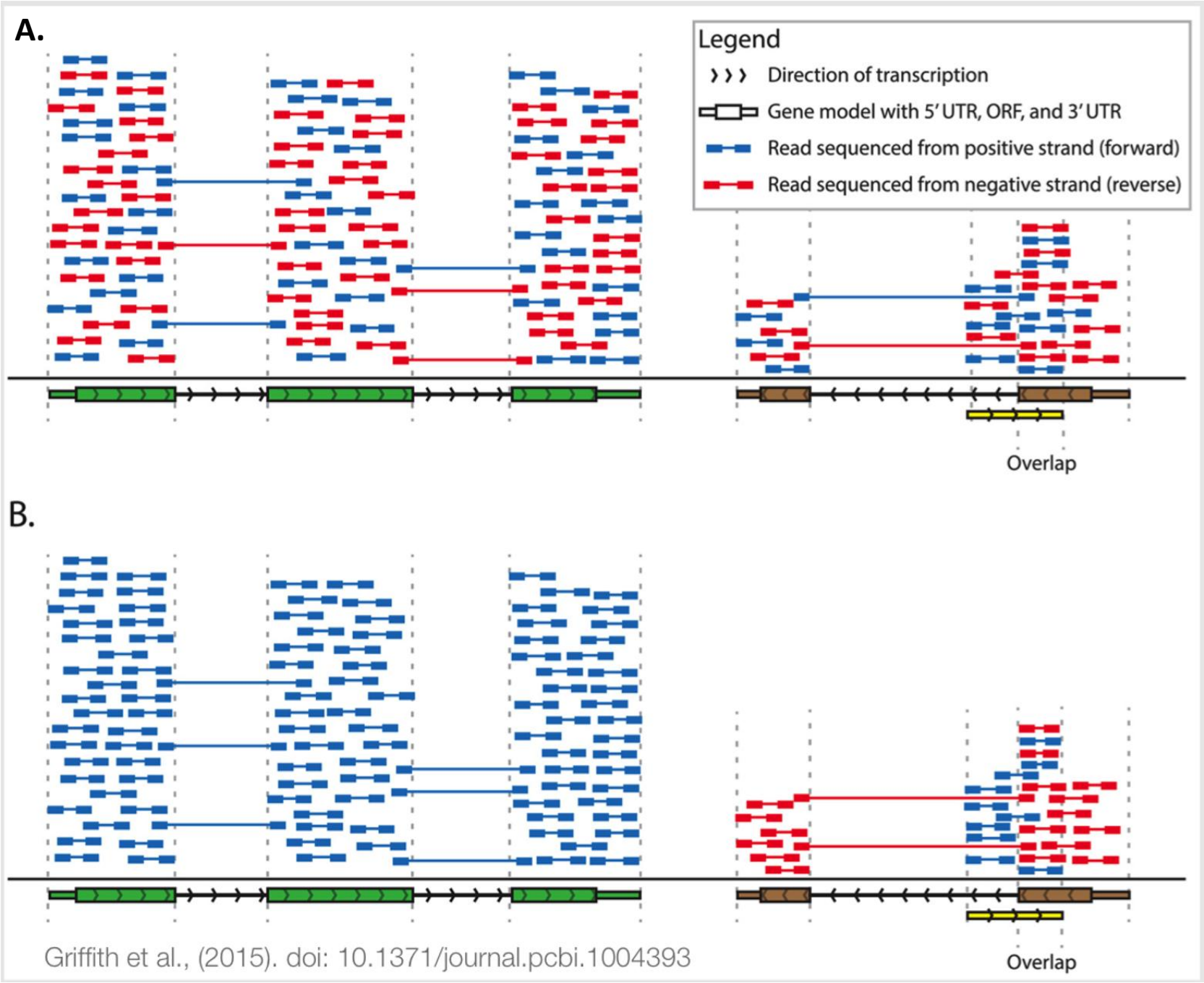
rRNA

protein coding (strongly expressed)

protein coding (lowly expressed)

Non-stranded

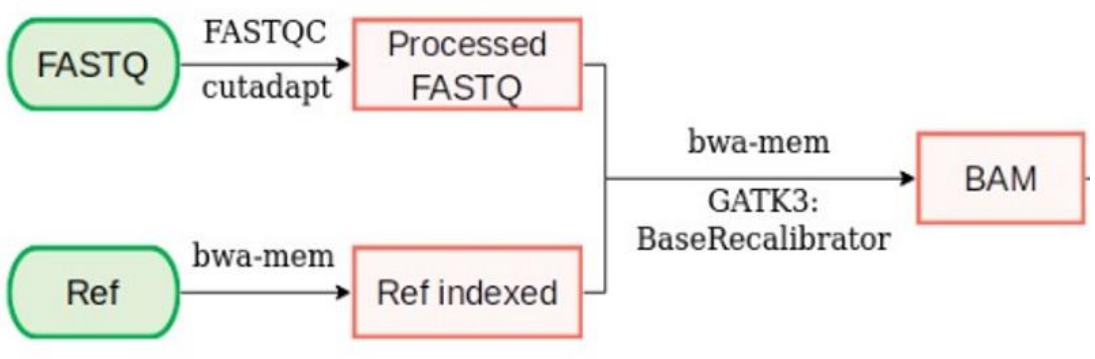
Stranded



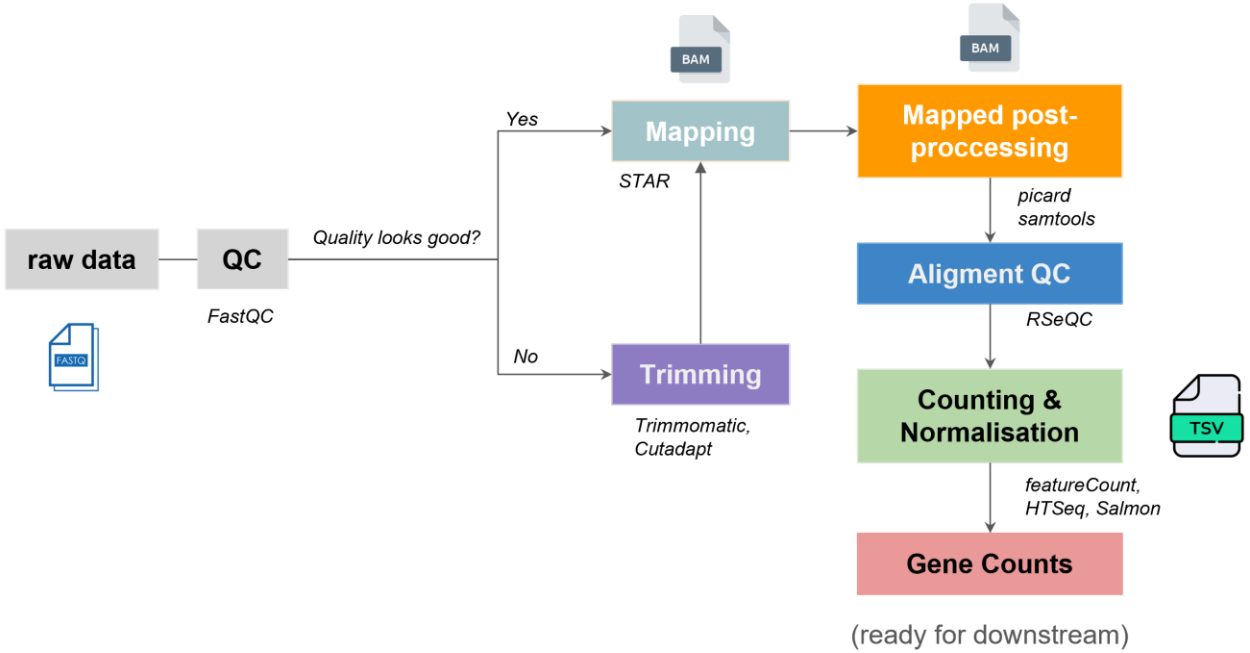
Griffith et al., (2015). doi: 10.1371/journal.pcbi.1004393

Upstream Analysis Workflow

Bulk DNA-seq

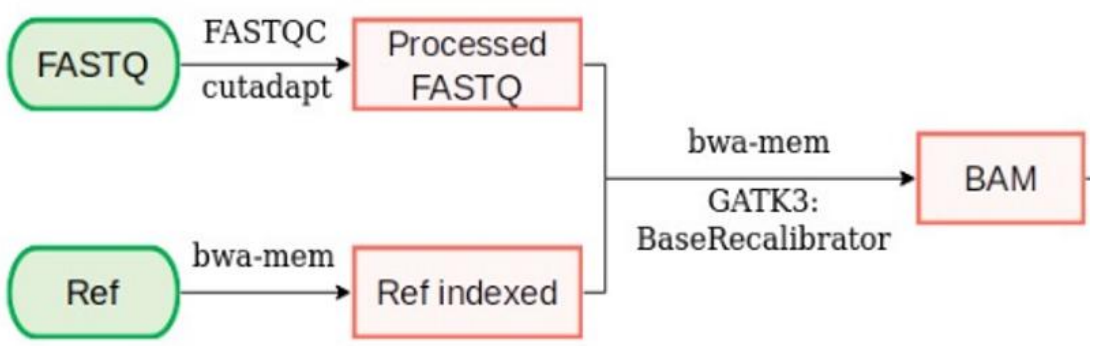


Bulk RNA-seq

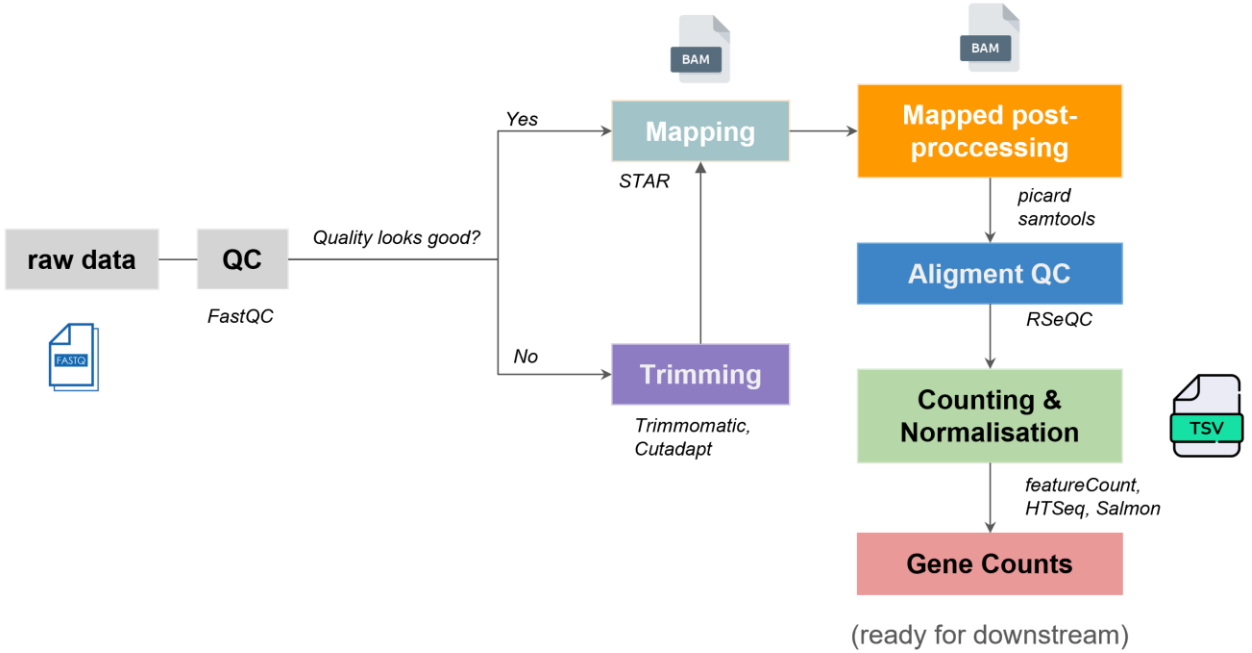


Upstream Analysis Workflow

Bulk DNA-seq



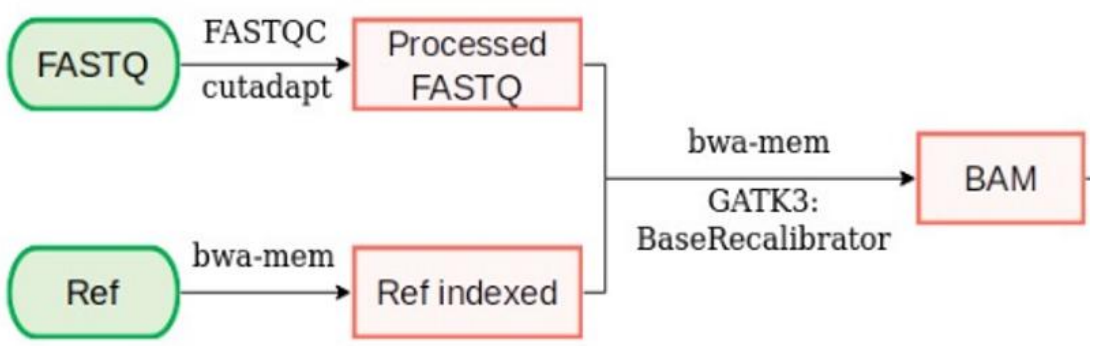
Bulk RNA-seq



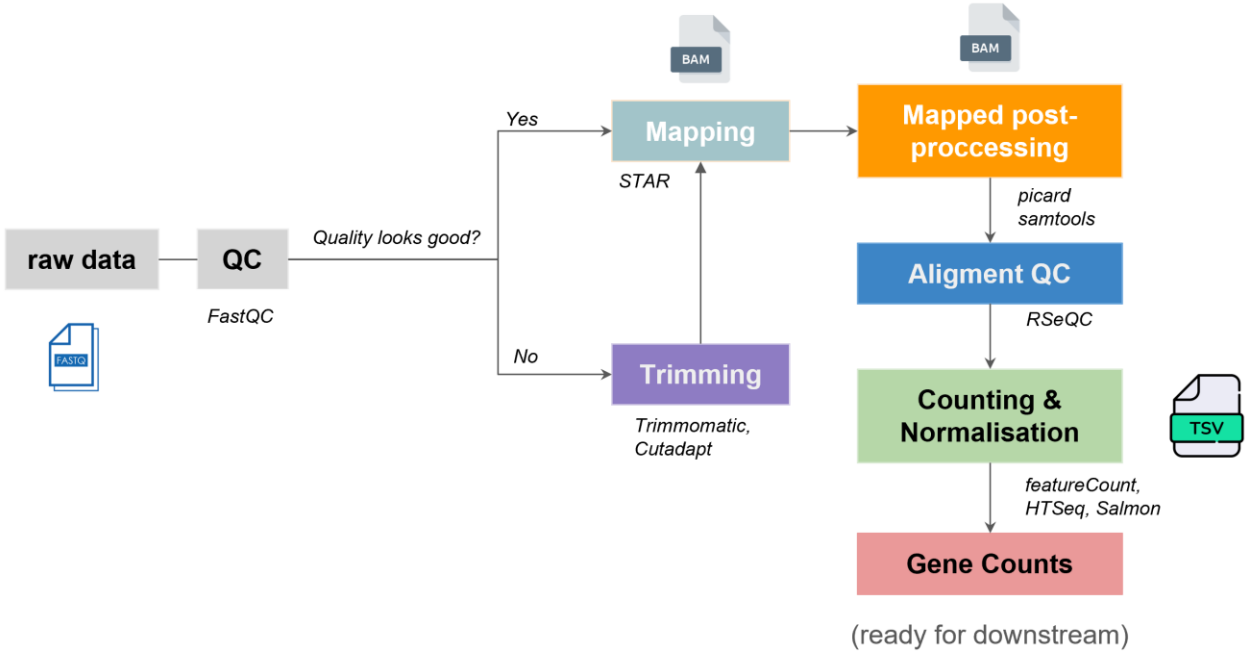
Where are the References need? And what are these?

Upstream Analysis Workflow

Bulk DNA-seq



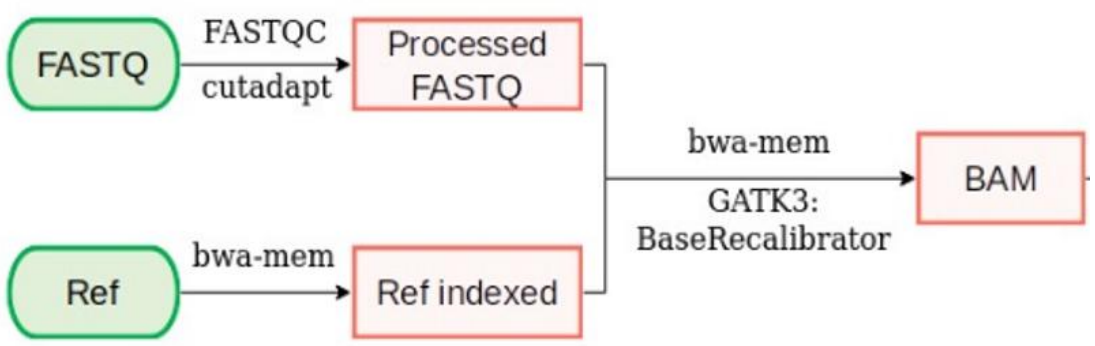
Bulk RNA-seq



Where are the References need? And what are these?

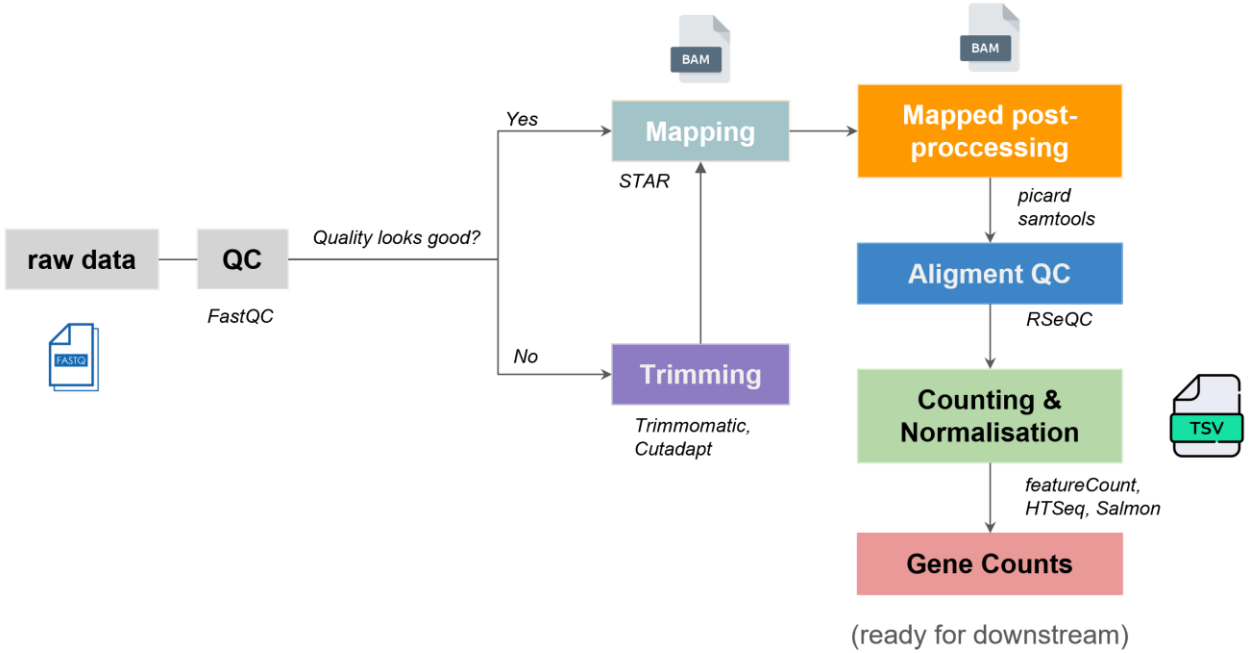
Upstream Analysis Workflow

Bulk DNA-seq



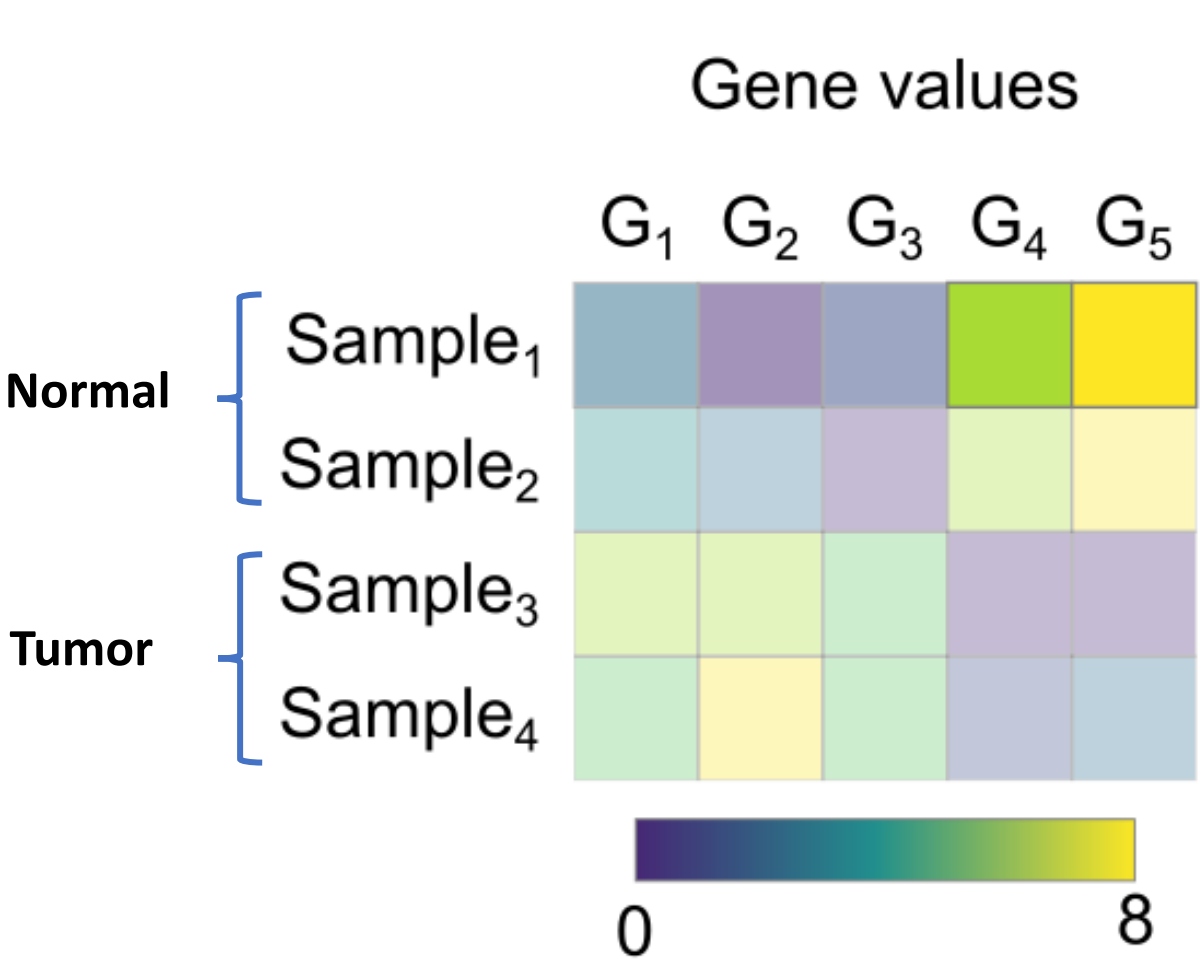
Human Genome Sequences can be found at UCSC, NCBI and Ensembl and GENCODE as fasta format

Bulk RNA-seq

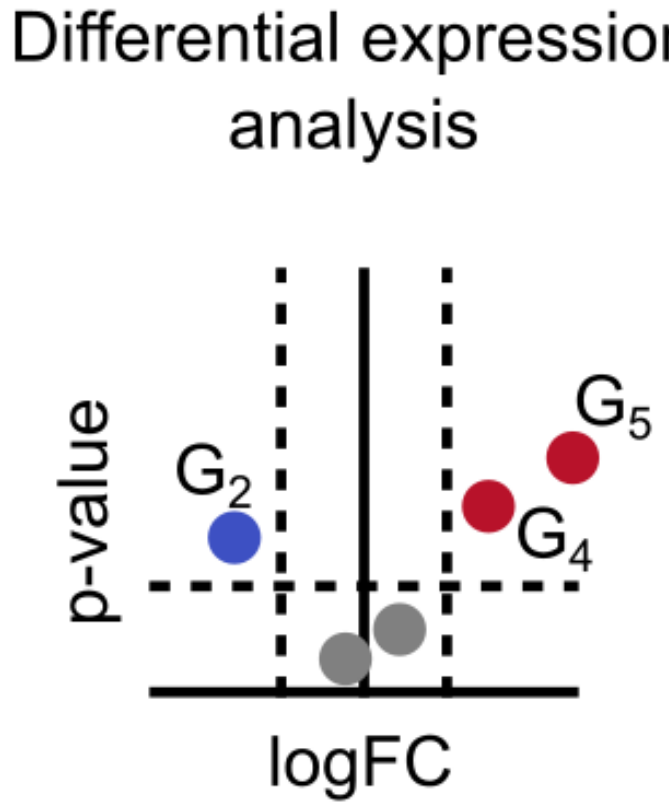


The four most common gene annotation databases are currently RefSeq, UCSC, Ensembl and GENCODE as GTF format

Differential Gene Expression

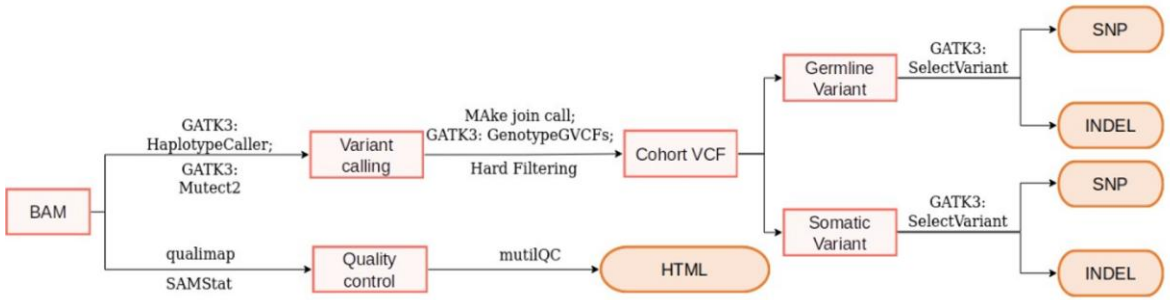


OR

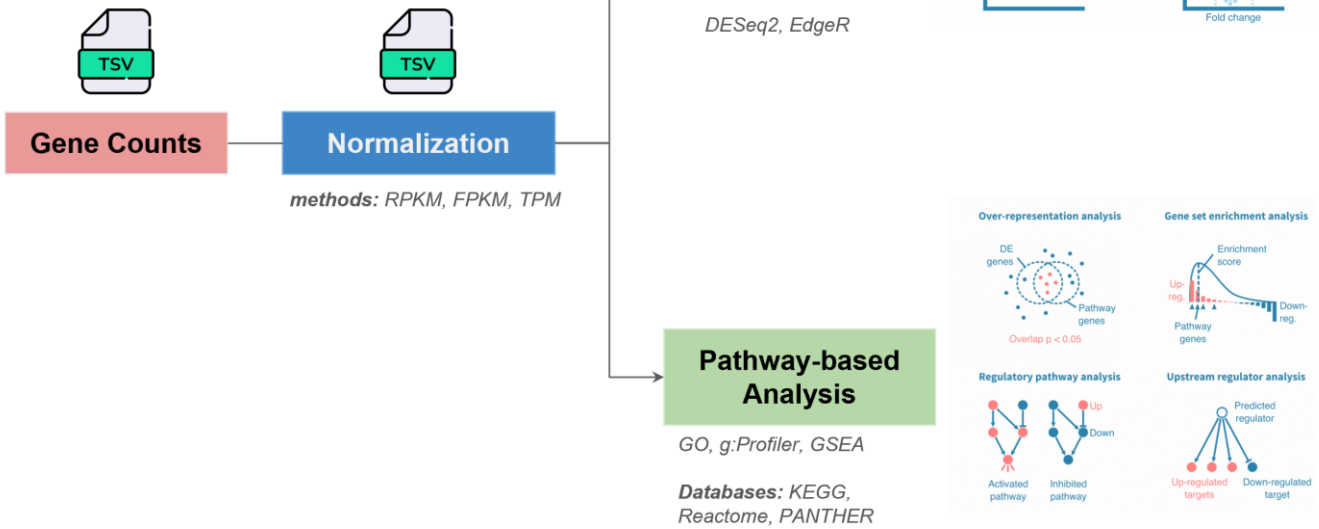


Downstream Analysis Workflow

Bulk DNA-seq

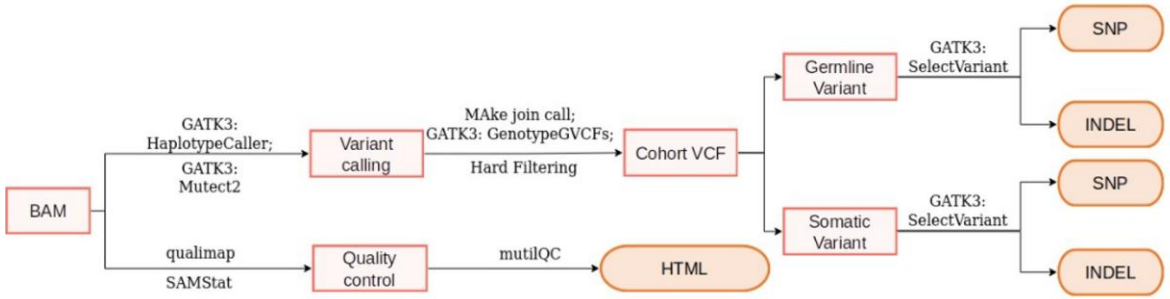


Bulk RNA-seq

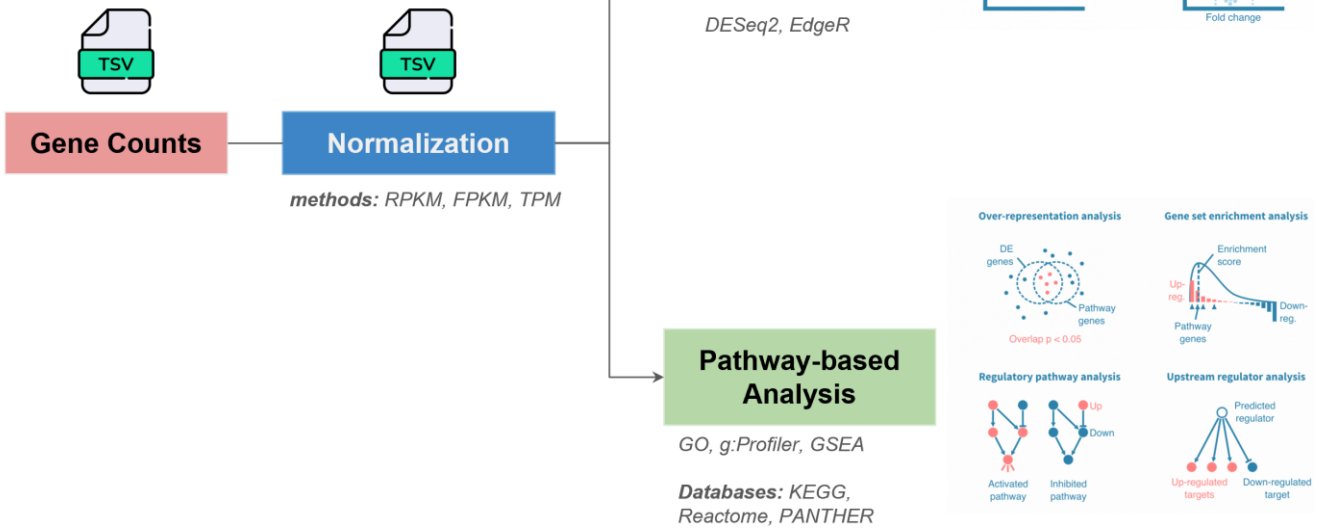


Downstream Analysis Workflow

Bulk DNA-seq



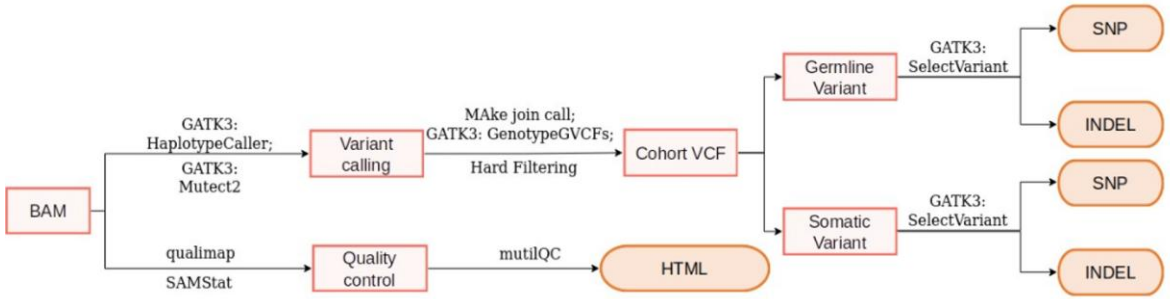
Bulk RNA-seq



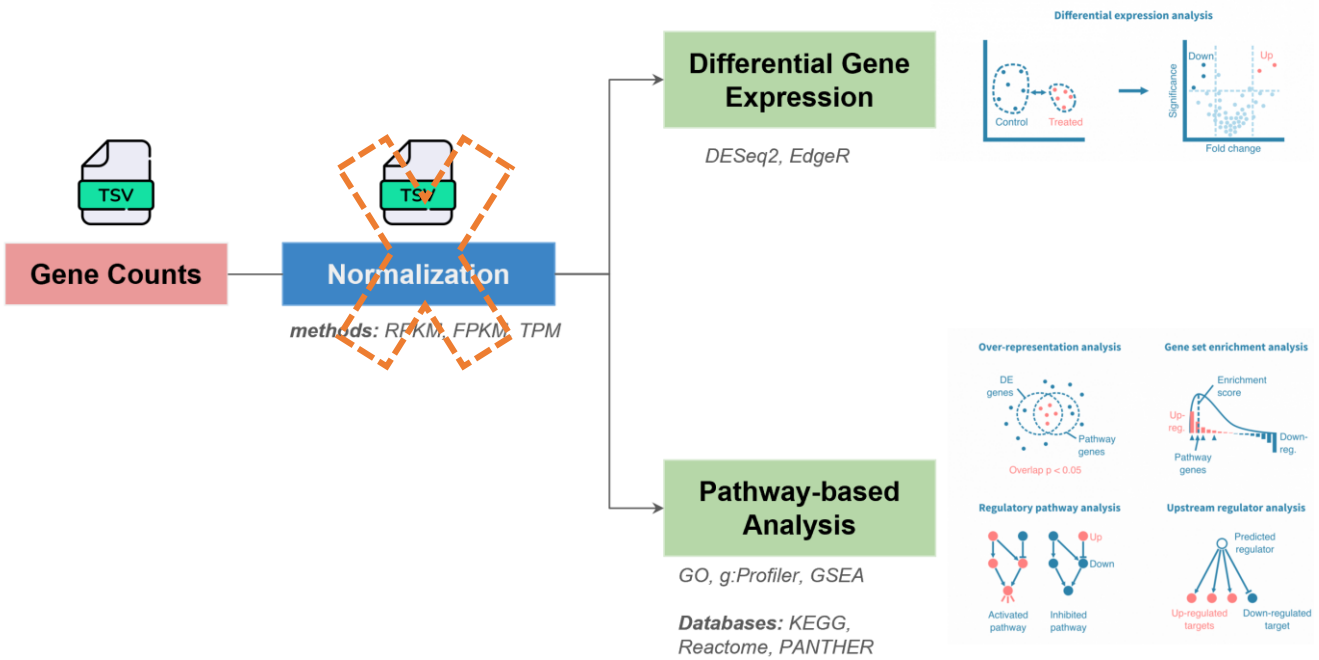
What is the unmatch?

Downstream Analysis Workflow

Bulk DNA-seq



Bulk RNA-seq



What is the unmatch?

From BAM to Count

Gene	Sample 1 Healthy	Sample 2 Healthy	Sample 3 Healthy	Sample 4 Tumor	Sample 5 Tumor	Sample 6 Tumor
A	100	101	105	160	163	154
B	10	8	9	1	2	3
C	45	45	45	46	46	46
D	0.11	0.12	0.13	0.0012	0.0014	0.0013

Differential Gene Expression

Gene	Sample 1 Healthy	Sample 2 Healthy	Sample 3 Healthy	Sample 4 Tumor	Sample 5 Tumor	Sample 6 Tumor
A	100	101	105	160	163	154
B	10	8	9	1	2	3
C	45	44	45	46	47	46
D	0.11	0.12	0.13	0.0012	0.0014	0.0013

```
t.test(c(100, 101, 105), c(160, 163, 154))
```

```
        Welch Two Sample t-test

data:  c(100, 101, 105) and c(160, 163, 154)
t = -18.658, df = 3.2, p-value = 0.0002251
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -66.38765 -47.61235
sample estimates:
mean of x mean of y
      102       159
```

Gene A

```
t.test(c(0.11, 0.12, 0.13), c(0.0012, 0.0014, 0.0013))
```

```
        Welch Two Sample t-test

data:  c(0.11, 0.12, 0.13) and c(0.0012, 0.0014, 0.0013)
t = 20.558, df = 2.0004, p-value = 0.002355
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.09386214 0.14353786
sample estimates:
mean of x mean of y
   0.1200    0.0013
```

Gene D

```
t.test(c(10, 8, 9), c(1, 2, 3))
```

```
        Welch Two Sample t-test

data:  c(10, 8, 9) and c(1, 2, 3)
t = 8.5732, df = 4, p-value = 0.001017
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.733042 9.266958
sample estimates:
mean of x mean of y
        9         2
```

Gene B

```
t.test(c(45, 44, 45), c(46, 45, 46))
```

```
        Welch Two Sample t-test

data:  c(45, 44, 45) and c(46, 45, 46)
t = -2.1213, df = 4, p-value = 0.1012
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3088288  0.3088288
sample estimates:
mean of x mean of y
 44.66667  45.66667
```

Gene C

Fisher's Exact Test: Example

Suppose we want to know whether or not gender is associated with political party preference. We take a simple random sample of 25 voters and survey them on their political party preference. The following table shows the results of the survey:

	Democrat	Republican	Total
Male	4	9	13
Female	8	4	12
Total	12	13	25

- **H_0:** Gender and political party preference are independent.
- **H_1:** Gender and political party preference are *not* independent.

Step 2: Calculated the two-tailed p value.

We can use the Fisher's Exact Test Calculator with the following input:

	Group 1	Group 2
Category 1	4	9
Category 2	8	4

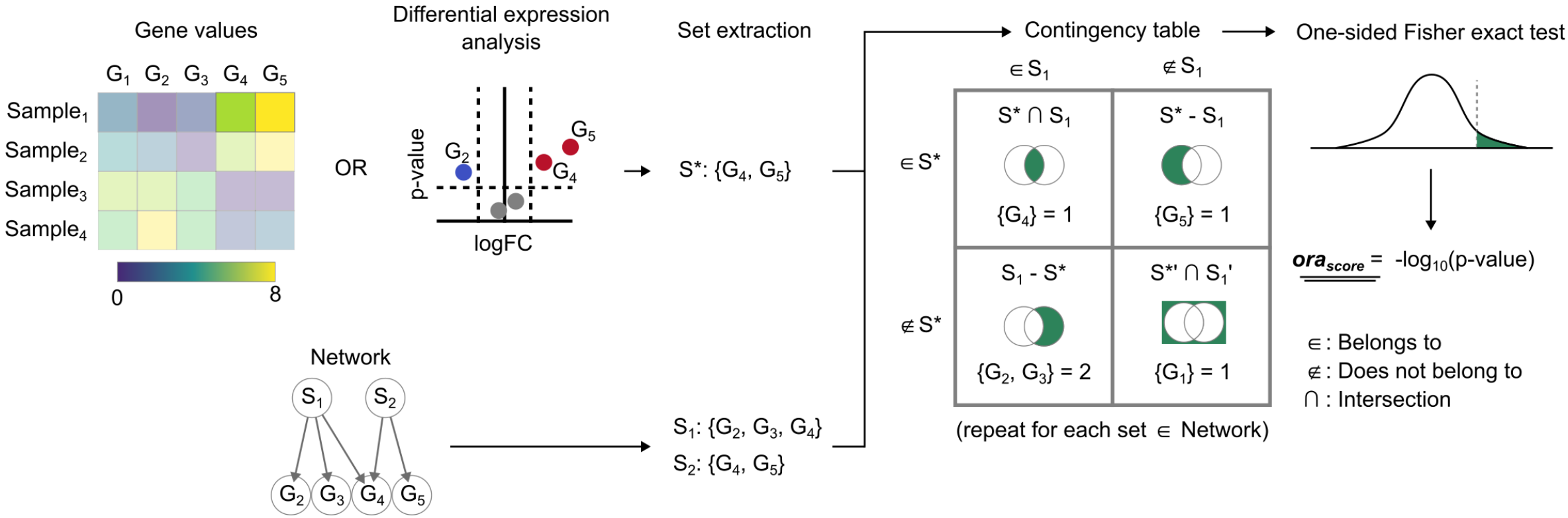
CALCULATE

One-tailed p value: **0.081178**

Two-tailed p value: **0.115239**

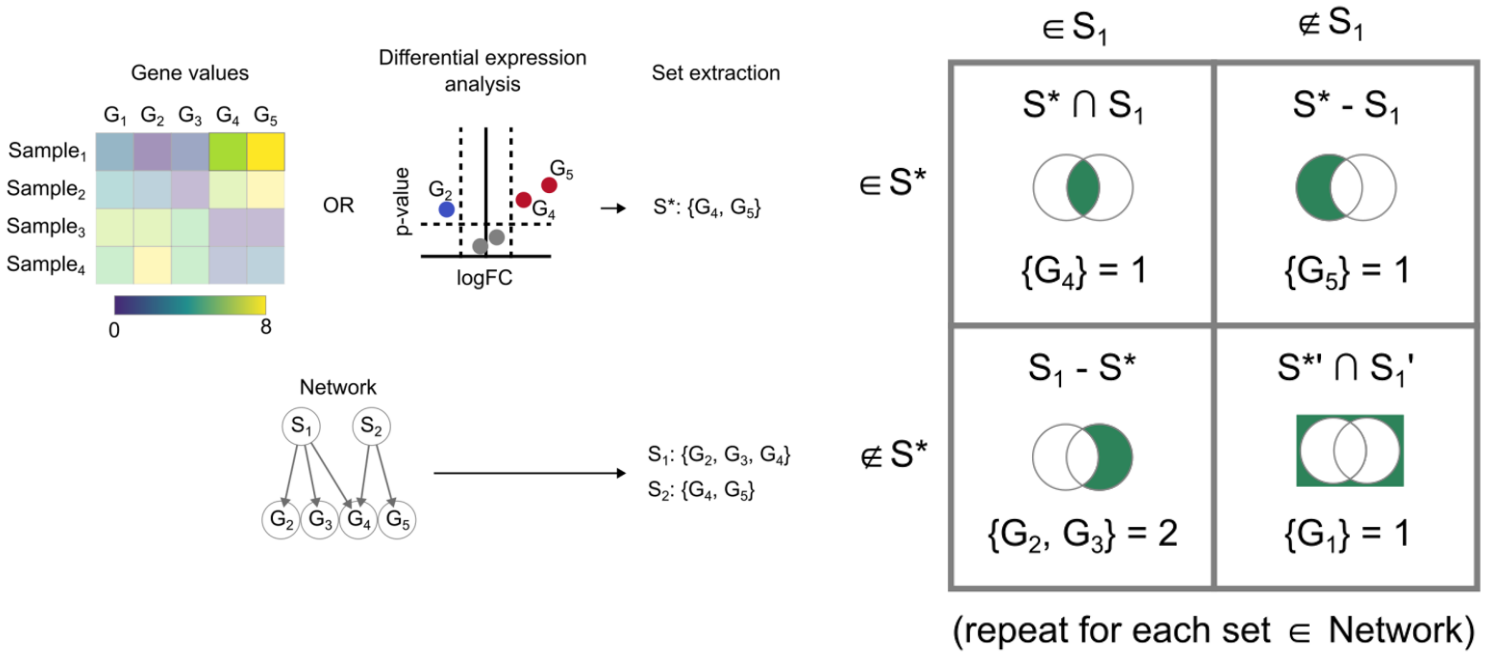
https://www.statology.org/fishers-exact-test/

Enrichment of DEGs with Over Representation Analysis (ORA)



https://decoupler-py.readthedocs.io/en/latest/notebooks/msigdb.html

Enrichment of DEGs with Over Representation Analysis (ORA)



	ID	Description	GeneRatio	BgRatio
	<chr>	<chr>	<chr>	<chr>
GO:0016052	GO:0016052	carbohydrate catabolic process	23/238	112/6476
GO:0044282	GO:0044282	small molecule catabolic process	25/238	145/6476
GO:0005975	GO:0005975	carbohydrate metabolic process	34/238	277/6476
GO:0032787	GO:0032787	monocarboxylic acid metabolic process	25/238	170/6476
GO:0006091	GO:0006091	generation of precursor metabolites and energy	25/238	217/6476
GO:0006979	GO:0006979	response to oxidative stress	16/238	103/6476

$$\text{GeneRatio} = \frac{\text{Number of genes from your input list associated with the GO term}}{\text{Total number of genes in your input list}}$$

$$\text{BgRatio} = \frac{\text{Number of genes associated with the GO term in the background set}}{\text{Total number of genes in the background set}}$$

https://decoupler-py.readthedocs.io/en/latest/notebooks/msigdb.html

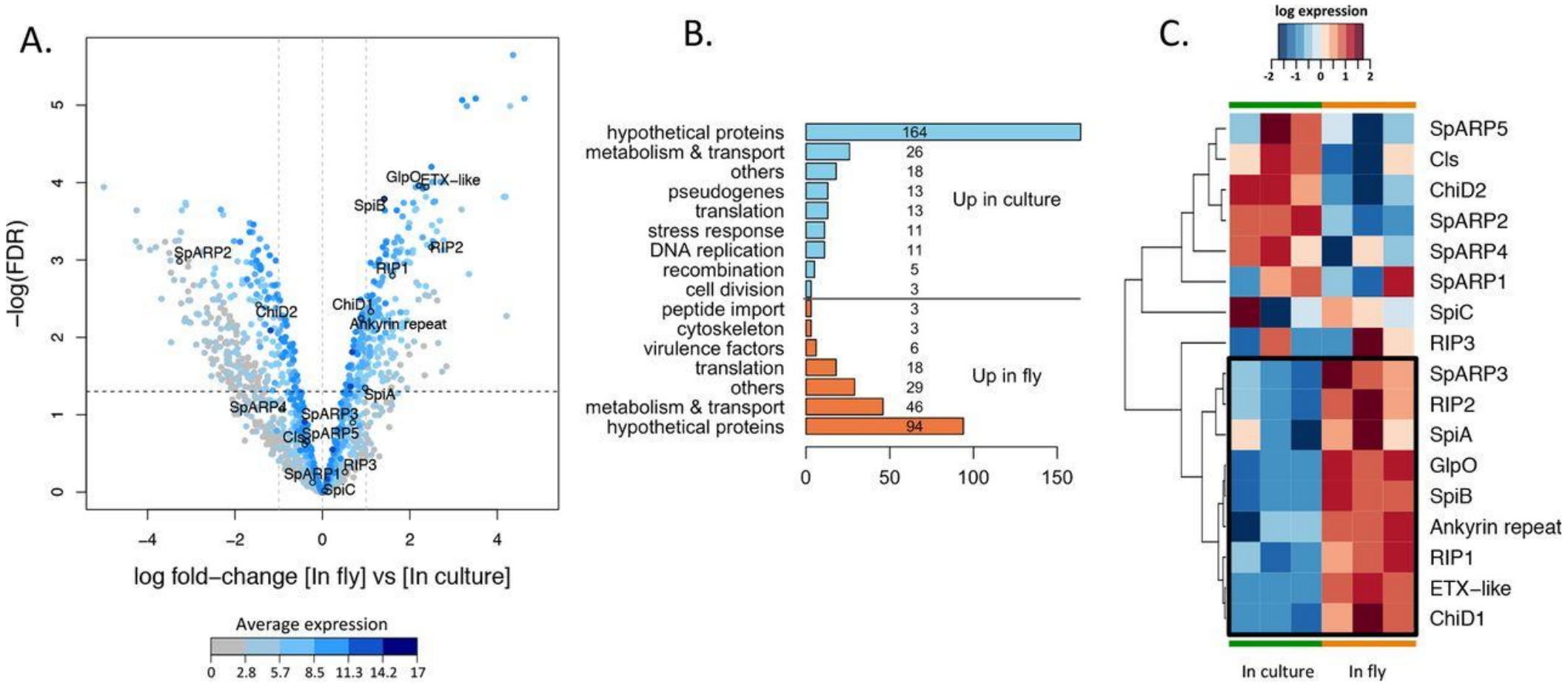
Enrichment of DEGs with Over Representation Analysis (ORA)

	ID	Description	GeneRatio	BgRatio	RichFactor	FoldEnrichment	zScore	pvalue	p.adjust	qvalue
	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
GO:0016052	GO:0016052	carbohydrate catabolic process	23/238	112/6476	0.2053571	5.587785	9.566051	8.859134e-12	6.856970e-09	5.856354e-09
GO:0044282	GO:0044282	small molecule catabolic process	25/238	145/6476	0.1724138	4.691394	8.780594	6.069770e-11	2.349001e-08	2.006219e-08
GO:0005975	GO:0005975	carbohydrate metabolic process	34/238	277/6476	0.1227437	3.339866	7.774200	3.099557e-10	7.996858e-08	6.829902e-08
GO:0032787	GO:0032787	monocarboxylic acid metabolic process	25/238	170/6476	0.1470588	4.001483	7.745849	2.010853e-09	3.891001e-07	3.323199e-07
GO:0006091	GO:0006091	generation of precursor metabolites and energy	25/238	217/6476	0.1152074	3.134802	6.247708	2.930343e-07	4.536170e-05	3.874221e-05
GO:0006979	GO:0006979	response to oxidative stress	16/238	103/6476	0.1553398	4.226809	6.447712	8.643298e-07	9.716547e-05	8.298642e-05

The qvalue in the results table from enrichGO represents the adjusted p-value for multiple testing correction. It is typically calculated using the False Discovery Rate (FDR) method, which controls the proportion of false positives among the list of enriched terms.

https://colab.research.google.com/drive/1uvY0IK2Dl75L61A0KEkXAIad3DDaEOMp?usp=sharing

Exercise 1: Explain the below figure



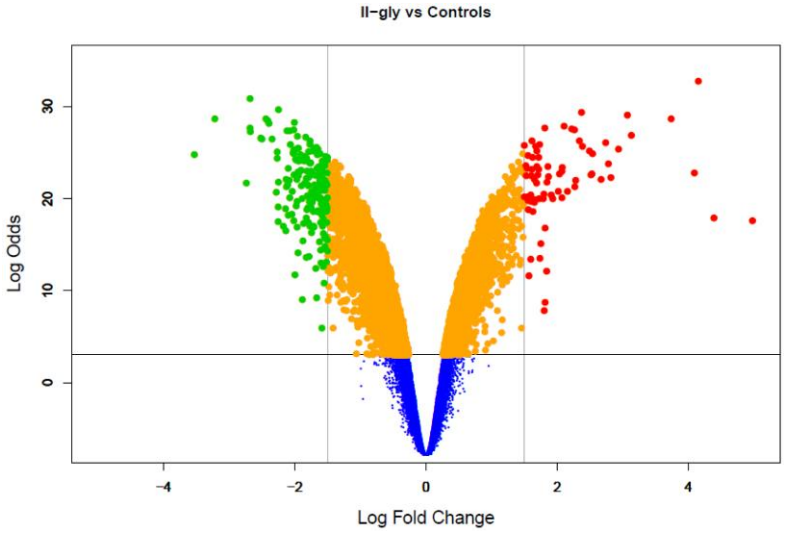
https://journals.asm.org/doi/10.1128/mbio.00024-18

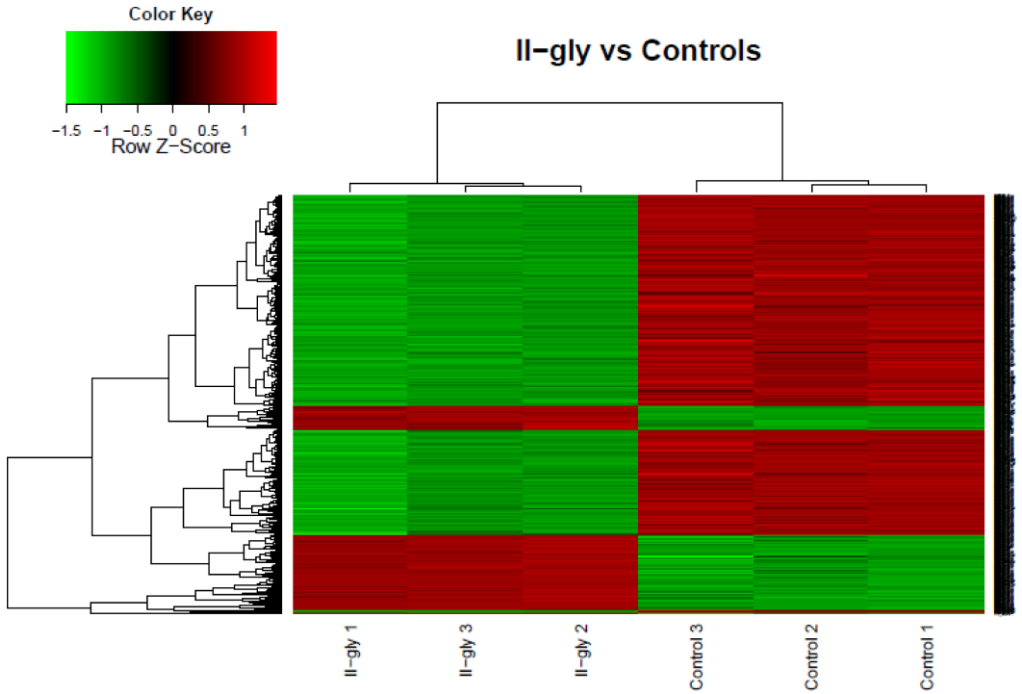
Exercise 1: Explain the below figure

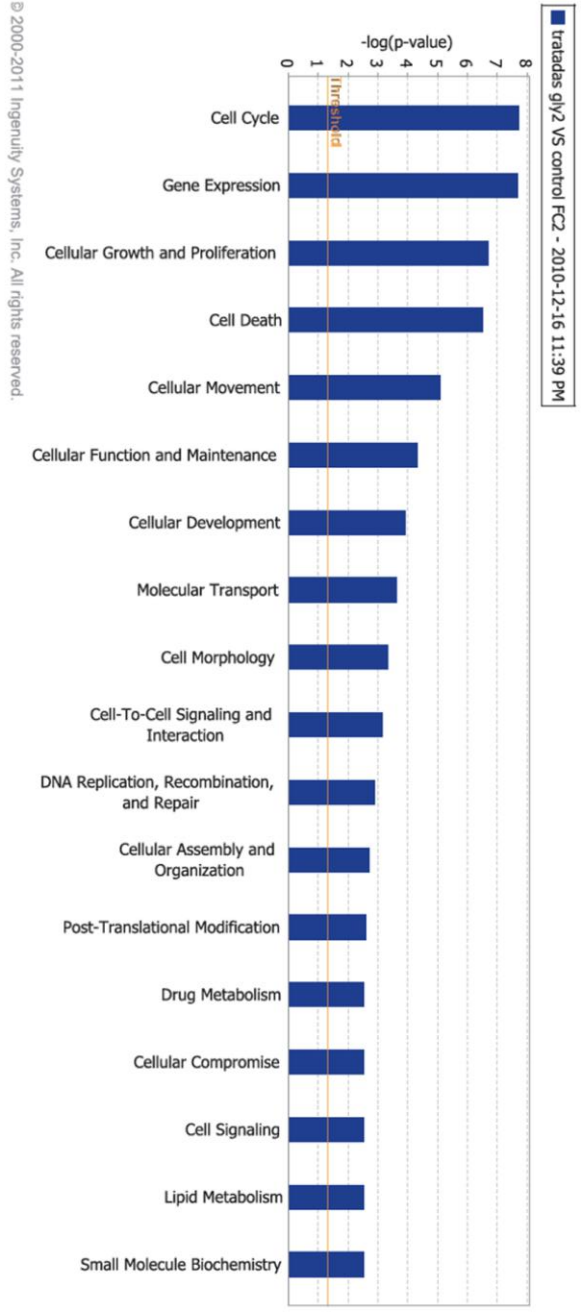
FIG 3 (A) Volcano plot of differential gene expression of S. poulsonii in host versus in culture. Each point represents the average value of one transcript in three replicate experiments. The expression difference is considered significant for a log2 fold change of ≥1 (outer light gray broken vertical lines) and for a P value of ≤0.05 [−log(FDR) of ≥1.3, dark broken horizontal line]. Points are colored according to their average expression in all data sets. Names and outlined points represent virulence factors. FDR, false-discovery rate. (B) Manual clustering of the transcripts differentially expressed by S. poulsonii in the fly versus in the culture. The numbers of sequences in the different categories are indicated on the bars or to the right of the bars. (C) Heatmap of S. poulsonii virulence gene expression. Each column represents the value for one replicate experiment in culture or in the fly. The colors represent the log10 level of expression in the corresponding experiment. The cluster of genes that are induced when S. poulsonii is in the host (versus in vitro) is shown enclosed in a black box. SpARP5, S. poulsonii ARP5.

https://journals.asm.org/doi/10.1128/mbio.00024-18

Exercise 2: Explain the below figure

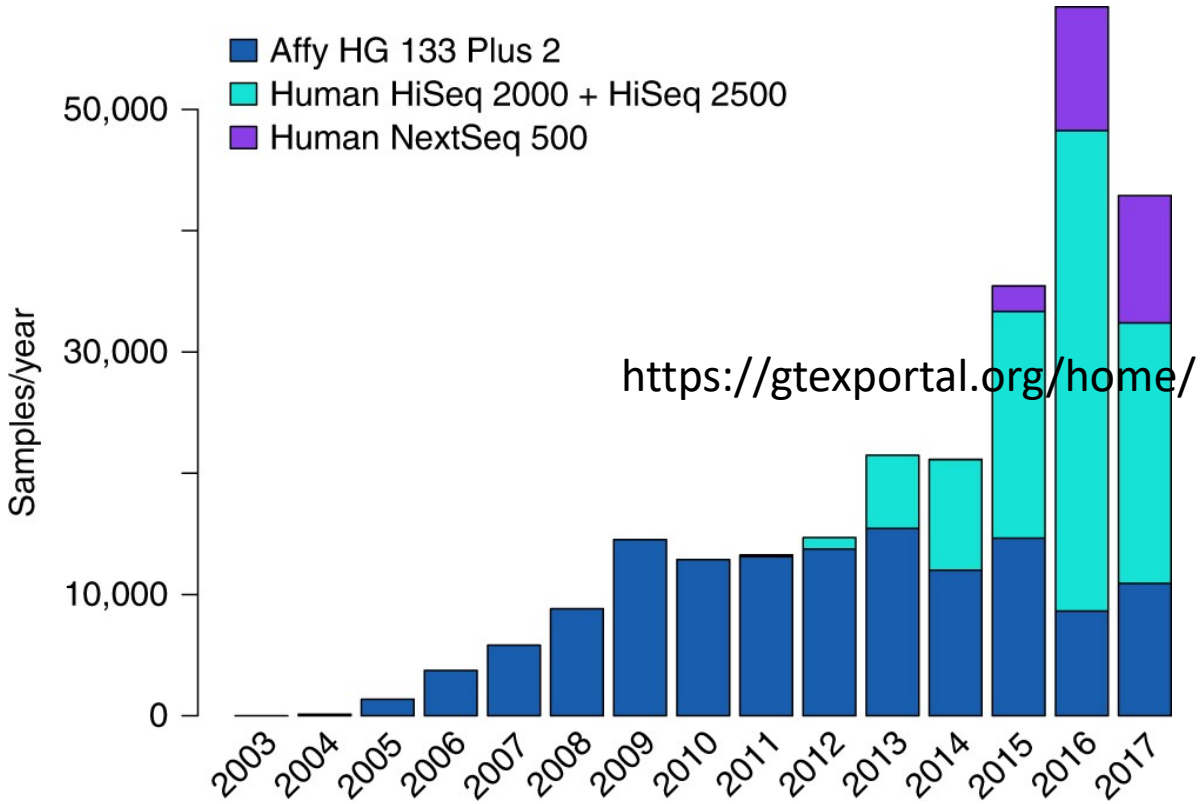




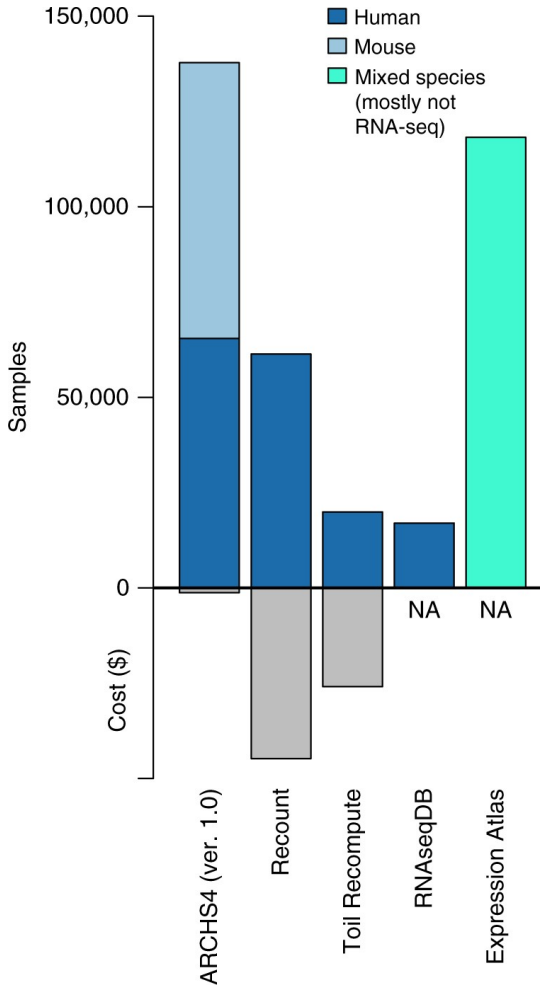


https://doi.org/10.1371/journal.pone.0054664

RNA-seq databases: GTEx and TCGA

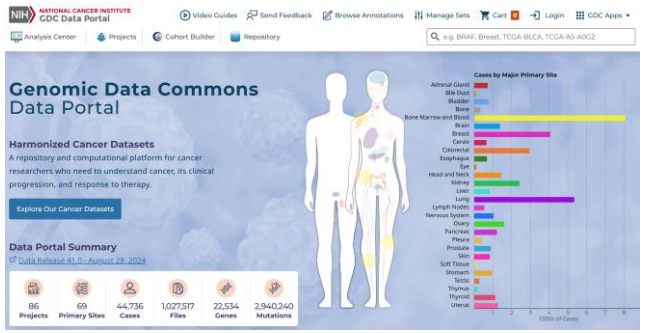


https://gtexportal.org/home/



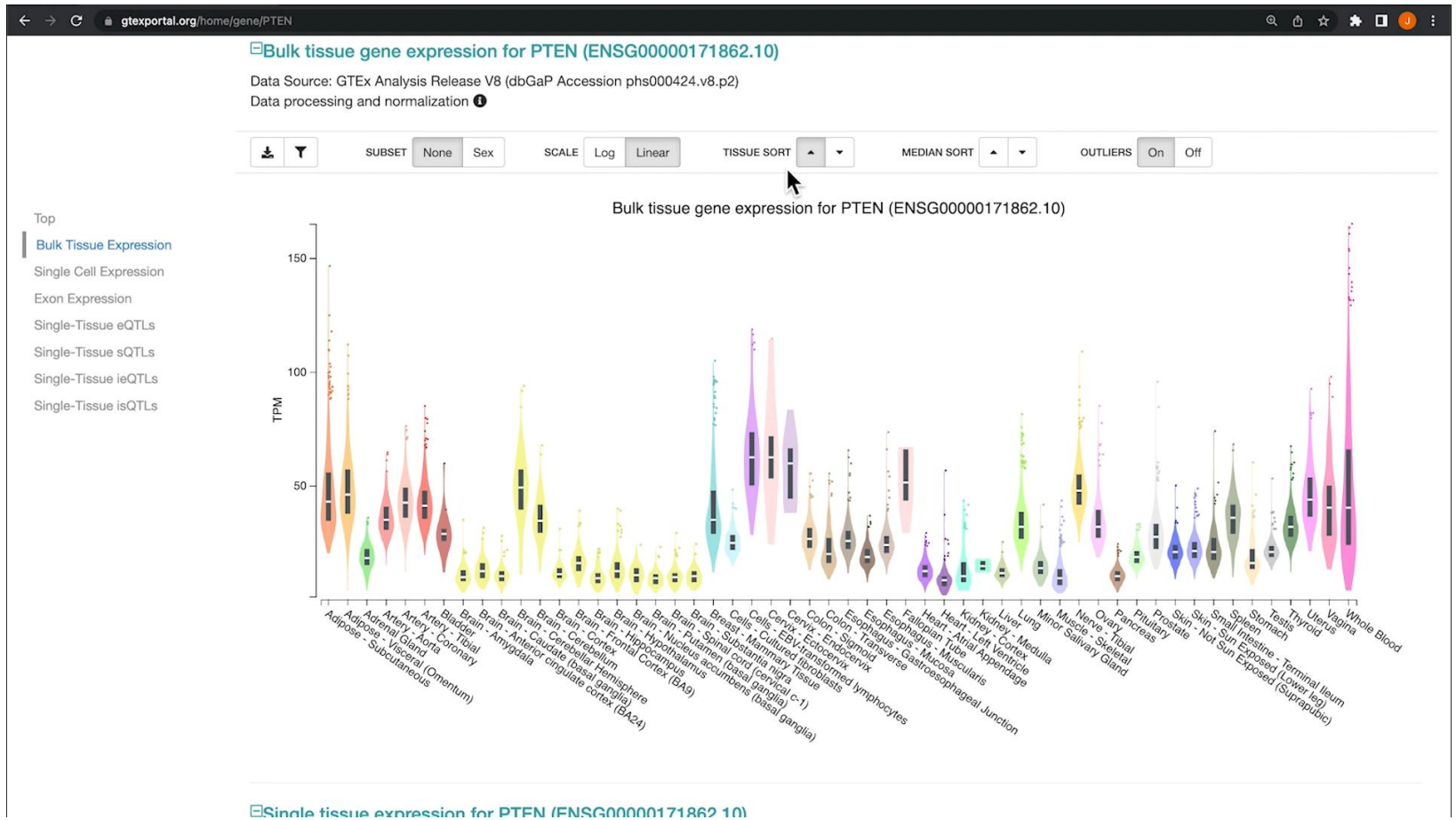
GTEx Portal

THE HUMAN PROTEIN ATLAS



https://www.nature.com/articles/s41467-018-03751-6

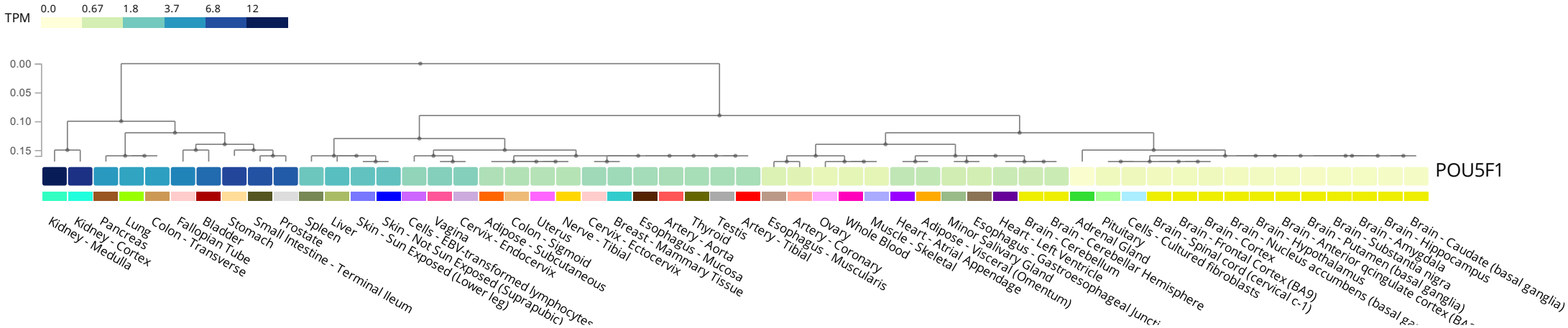
GTEx Portal



https://www.youtube.com/watch?v=WjHkb_y_yFk&t=2s

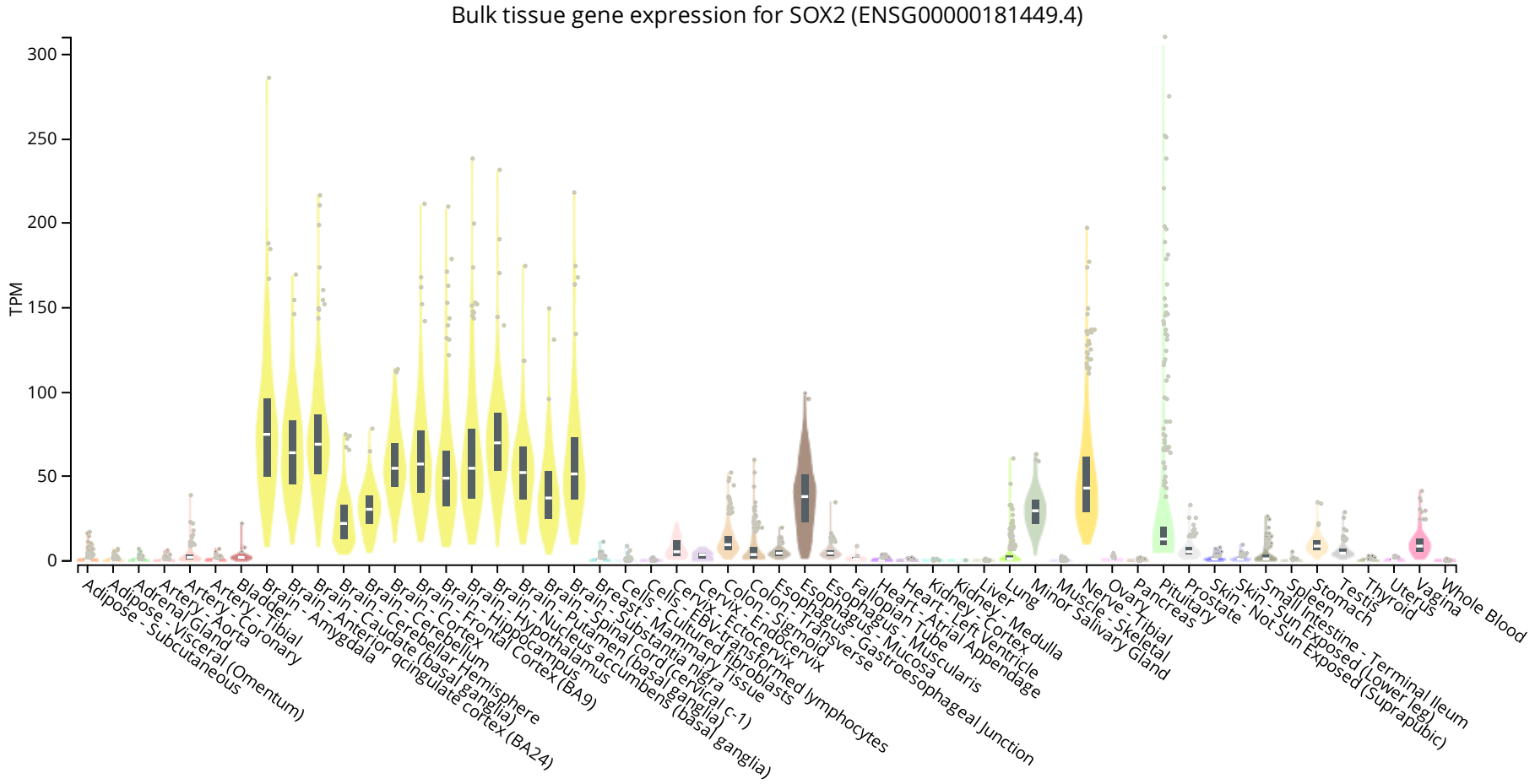
https://gtexportal.org/home/

GTEx Portal



https://www.gtexportal.org/home/multiGeneQueryPage/POU5F1

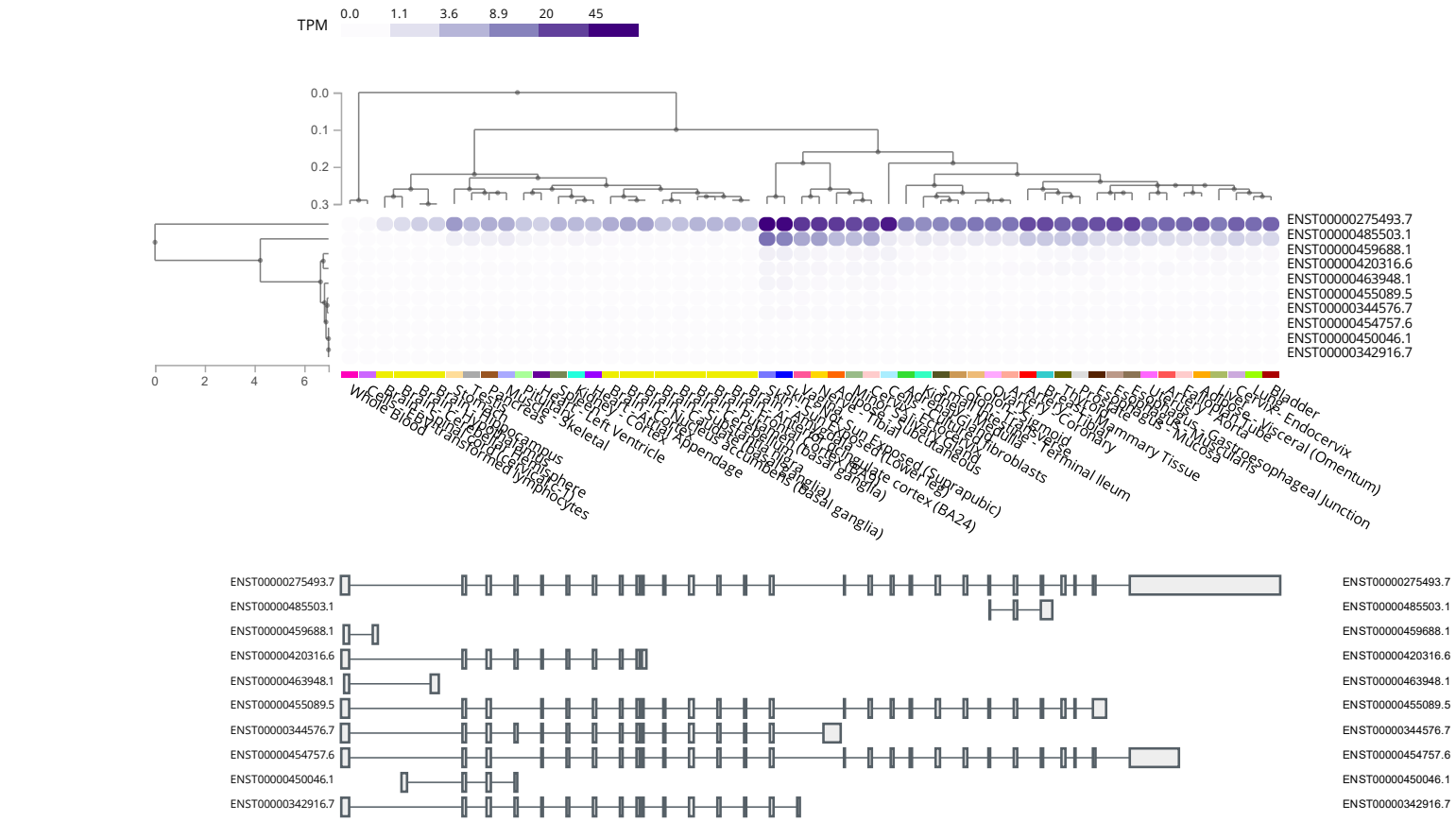
GTEx Portal



https://www.gtexportal.org/home/gene/SOX2#geneExpression

GTEx Portal

Isoform Expression of EGFR: ENSG00000146648.20 epidermal growth factor receptor [Source:HGNC Symbol;Acc:HGNC:3236]



https://www.gtexportal.org/home/transcriptPage

NATIONAL CANCER INSTITUTE
GDC Data Portal

Video Guides | Send Feedback | Browse Annotations | Manage Sets | Cart 0 | Login | GDC Apps

Analysis Center | Projects | Cohort Builder | Repository

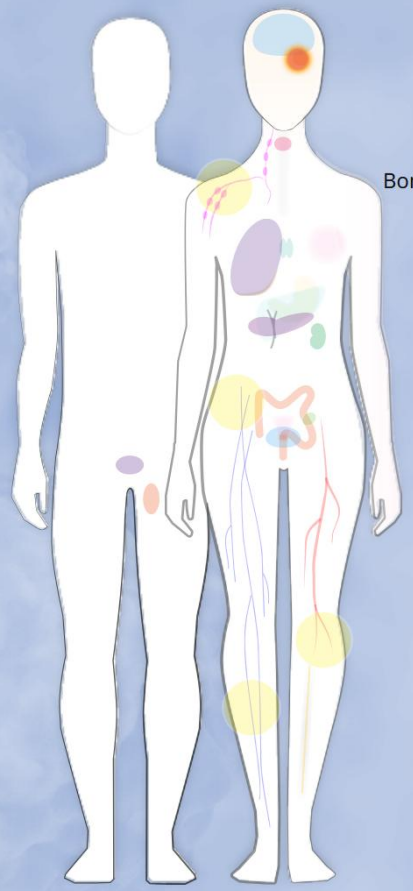
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

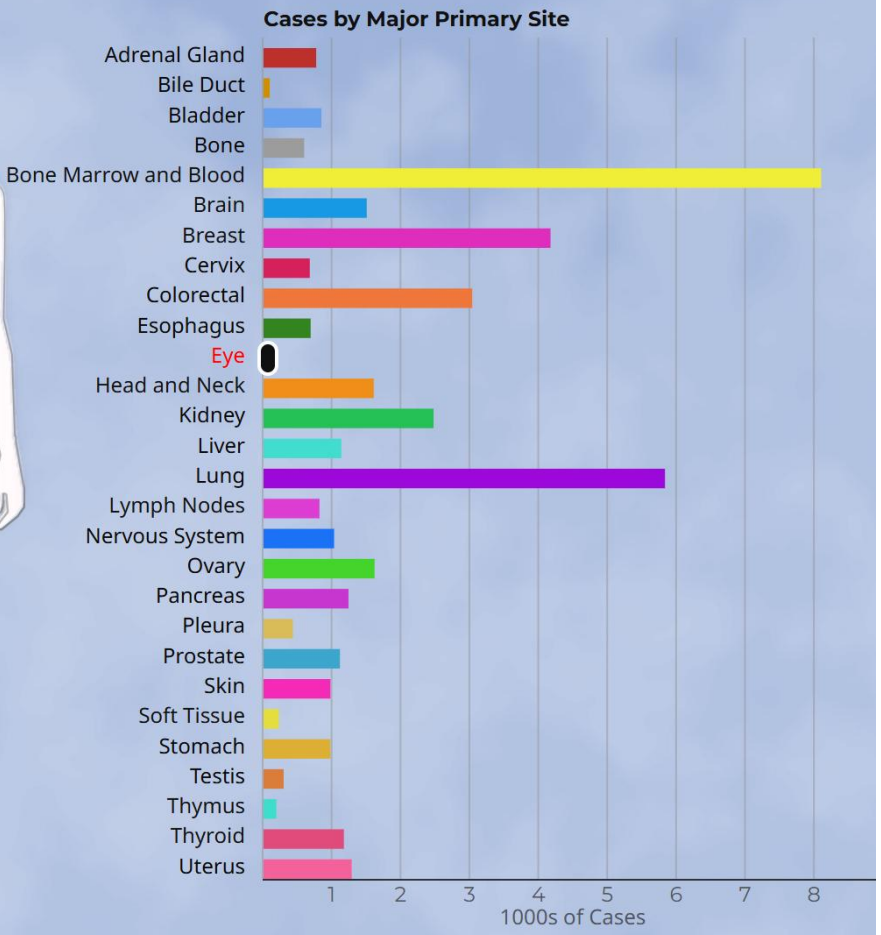
Genomic Data Commons Data Portal

Harmonized Cancer Datasets

A repository and computational platform for cancer researchers who need to understand cancer, its clinical progression, and response to therapy.

Explore Our Cancer Datasets





Data Portal Summary

Data Release 42.0 - January 30, 2025

86	69	44,736	1,121,816	22,534	2,940,240
Projects	Primary Sites	Cases	Files	Genes	Mutations

https://portal.gdc.cancer.gov/

ANALYSIS TOOLS

BAM Slicing Download
80 Cases

Clinical Data Analysis Demo
114 Cases

Cohort Comparison Demo
114 Cases

OncoMatrix Demo
80 Cases

ProteinPaint Demo
80 Cases

Single Cell RNA-seq
0 Cases

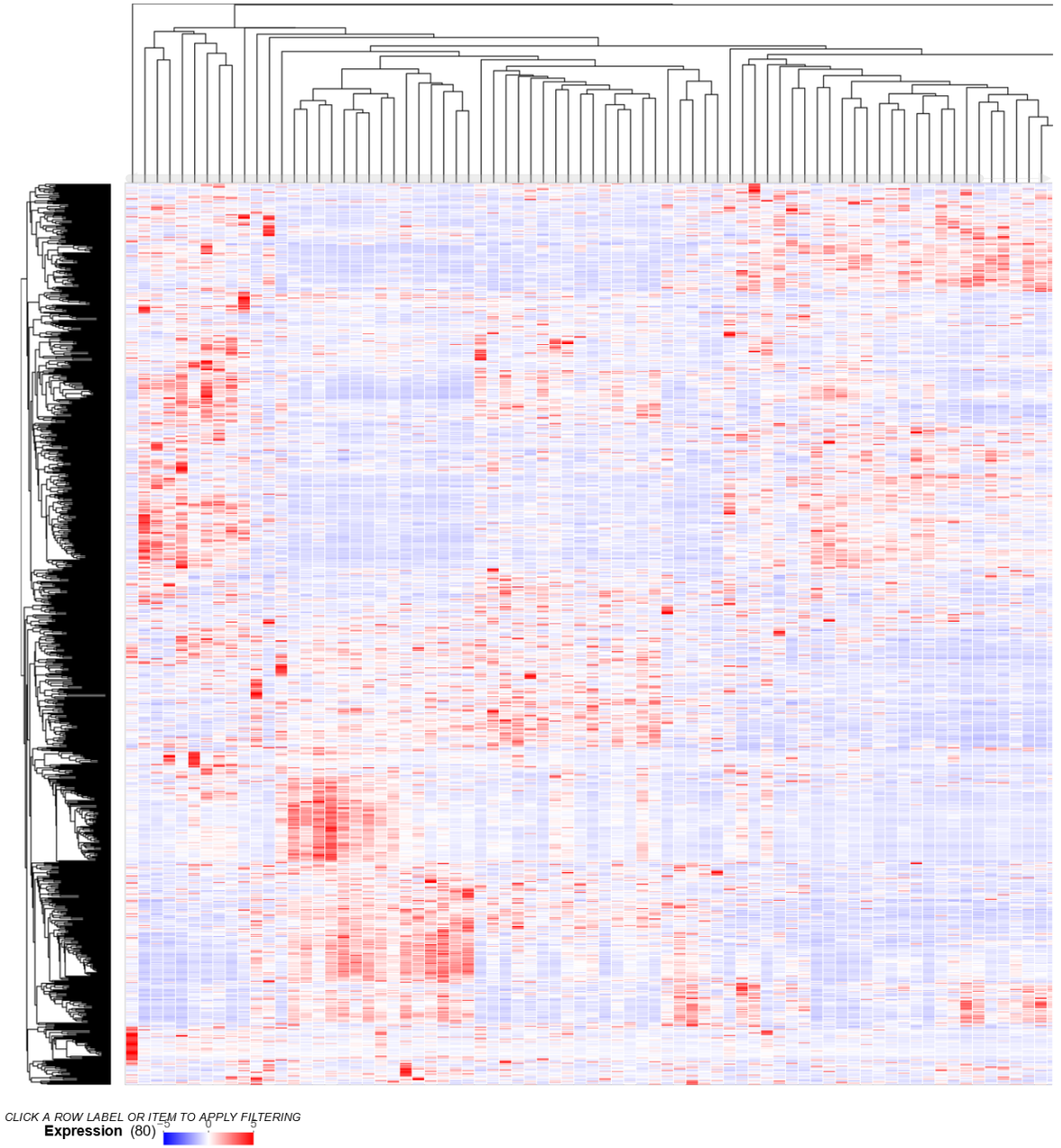
Cohort Level MAF
80 Cases

Gene Expression Clustering Demo
80 Cases

Mutation Frequency Demo
80 Cases

Sequence Reads
80 Cases

Set Operations Demo



https://portal.gdc.cancer.gov/analysis_page?app=GeneExpression

Cảm ơn các thầy cô và các bạn!

Ngày 10 tháng 05 năm 2025

TS. Lưu Phúc Lợi

Email: luu.p.loi@googlemail.com

Zalo: 0901802182