Best Practice for Bulk RNA-seq analysis

Ngày 08 tháng 12 năm 2024

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

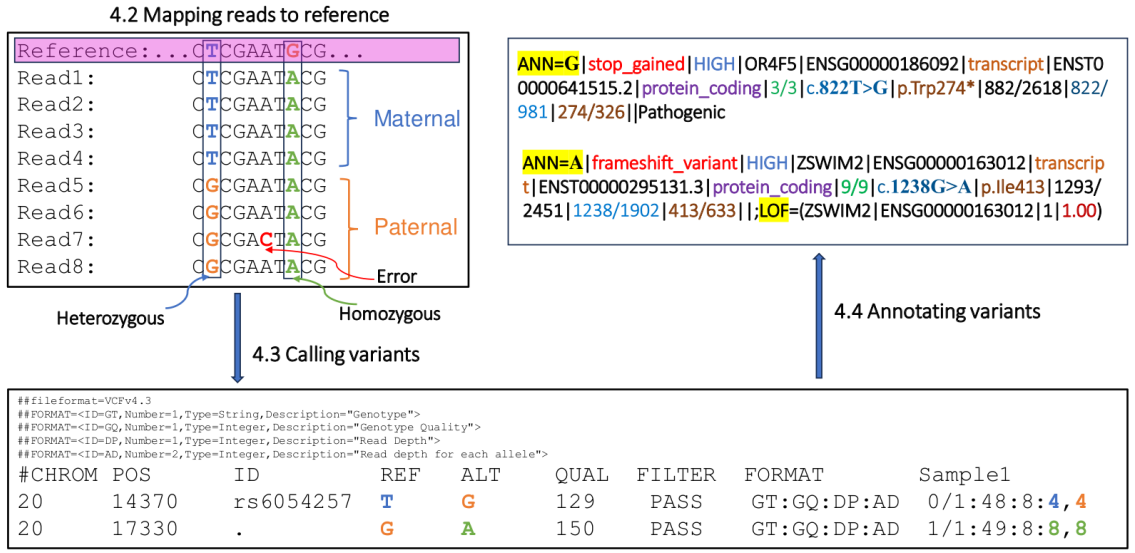
Aim of the methods

Bulk DNA-seq?

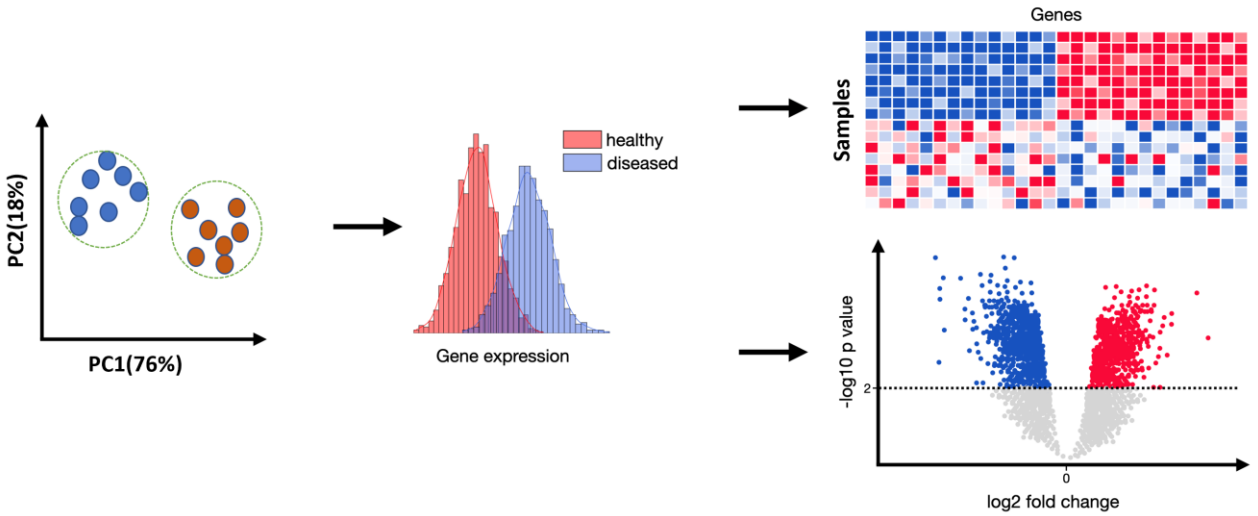
Bulk RNA-seq?

Aim of the methods

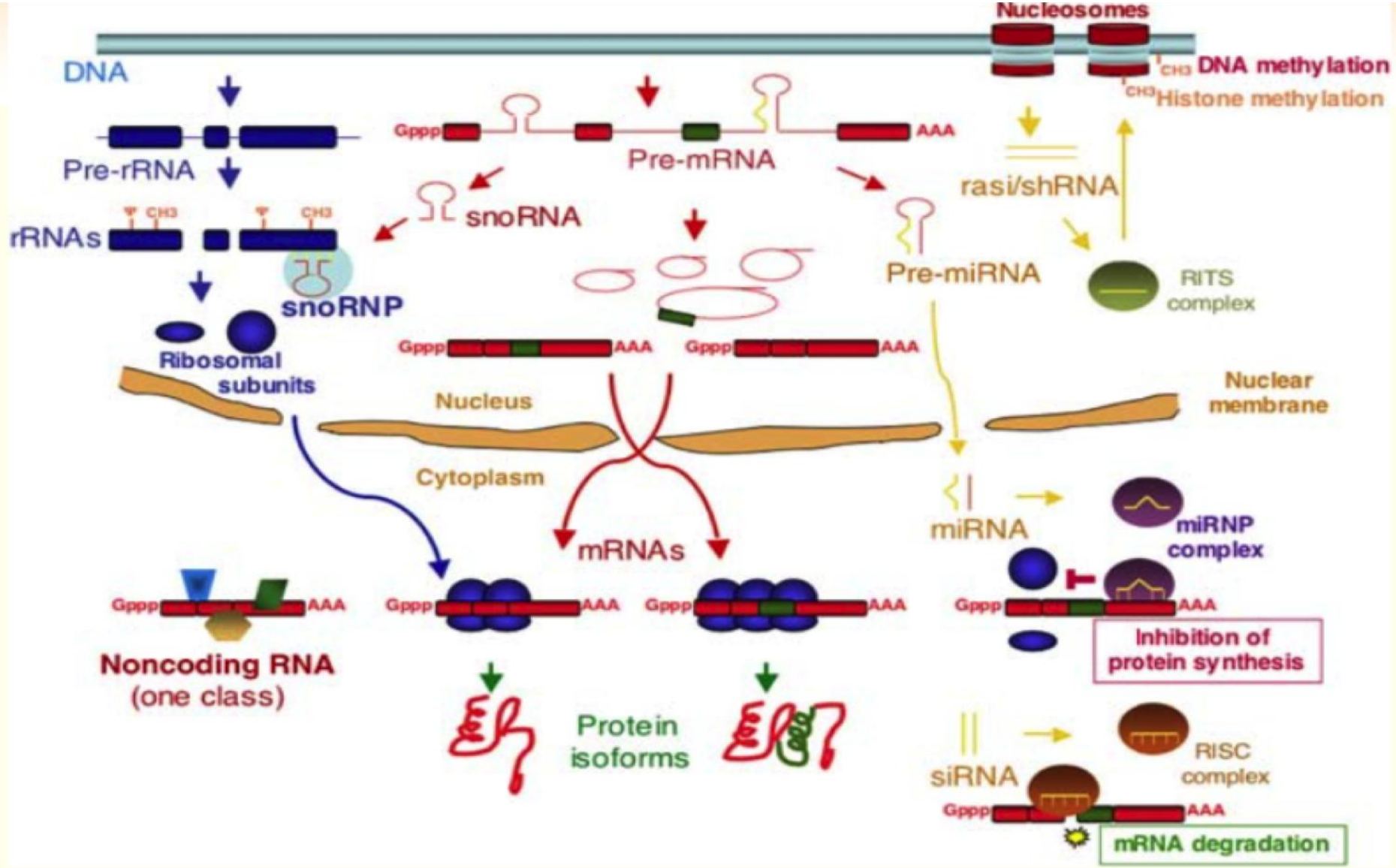
Bulk DNA-seq



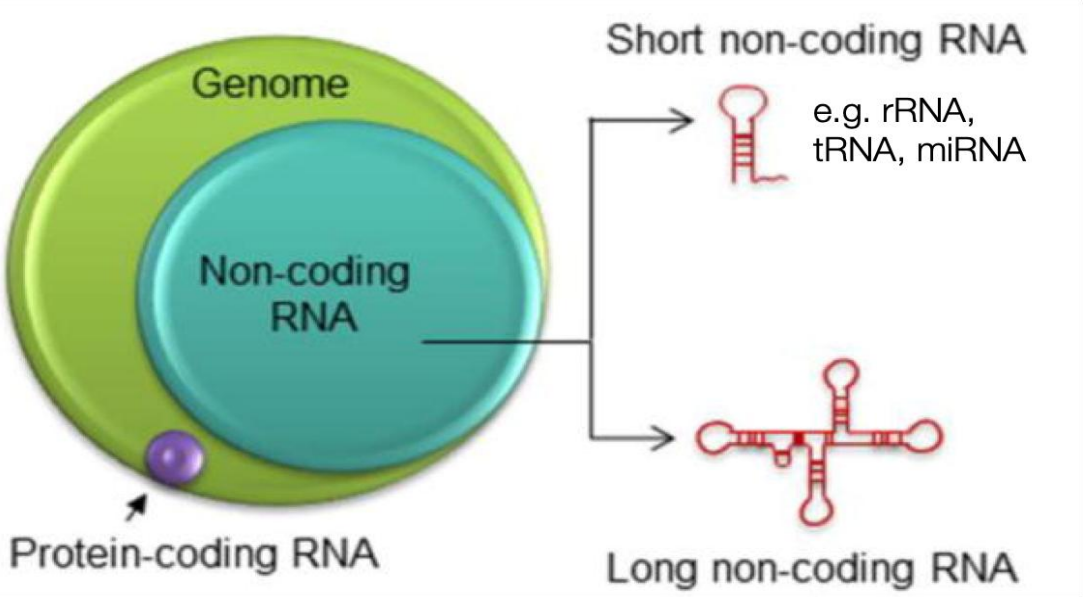
Bulk RNA-seq



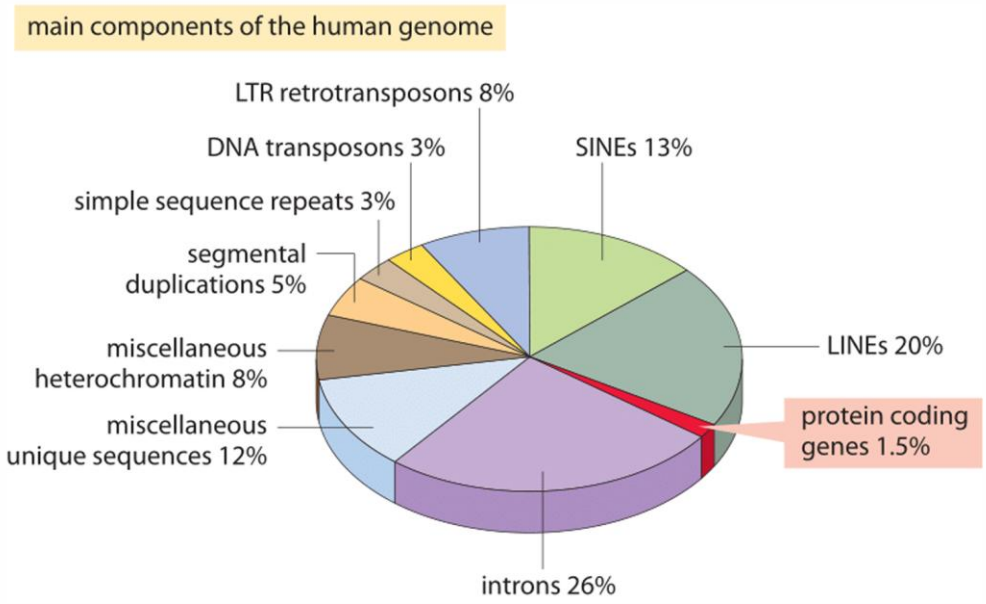
Different types of RNA



Different types of RNA



Parasramka et al. (2016)

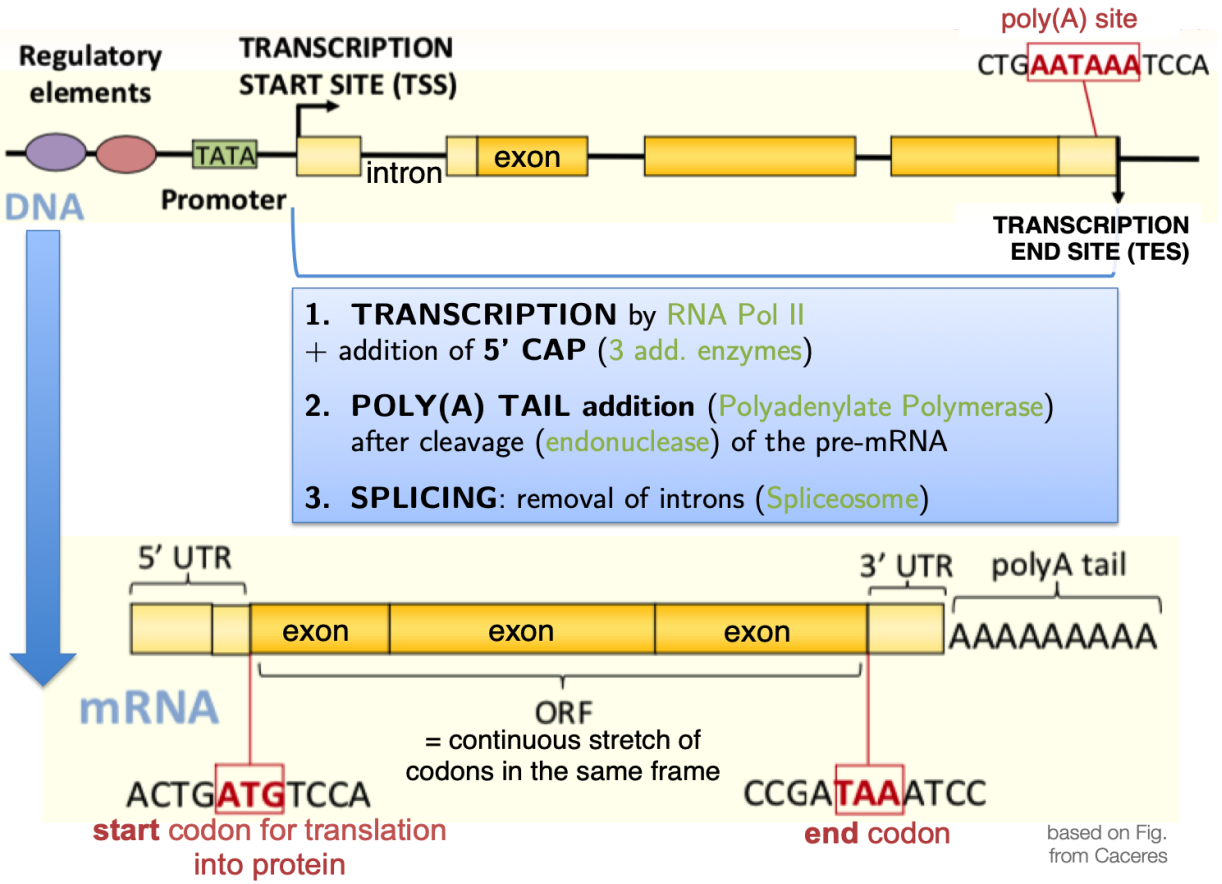


» How many genes are in a genome? (bionumbers.org)

Sequencing library prep protocol depends on the RNA properties

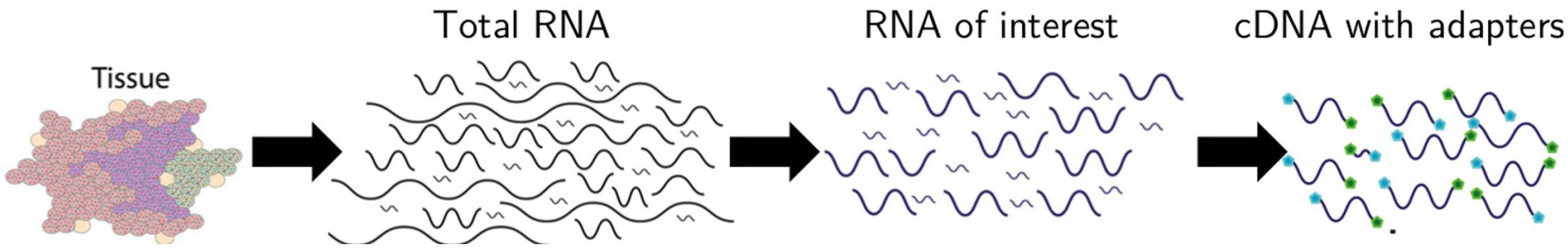
It is not a one-size-fits-all situation!

- 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

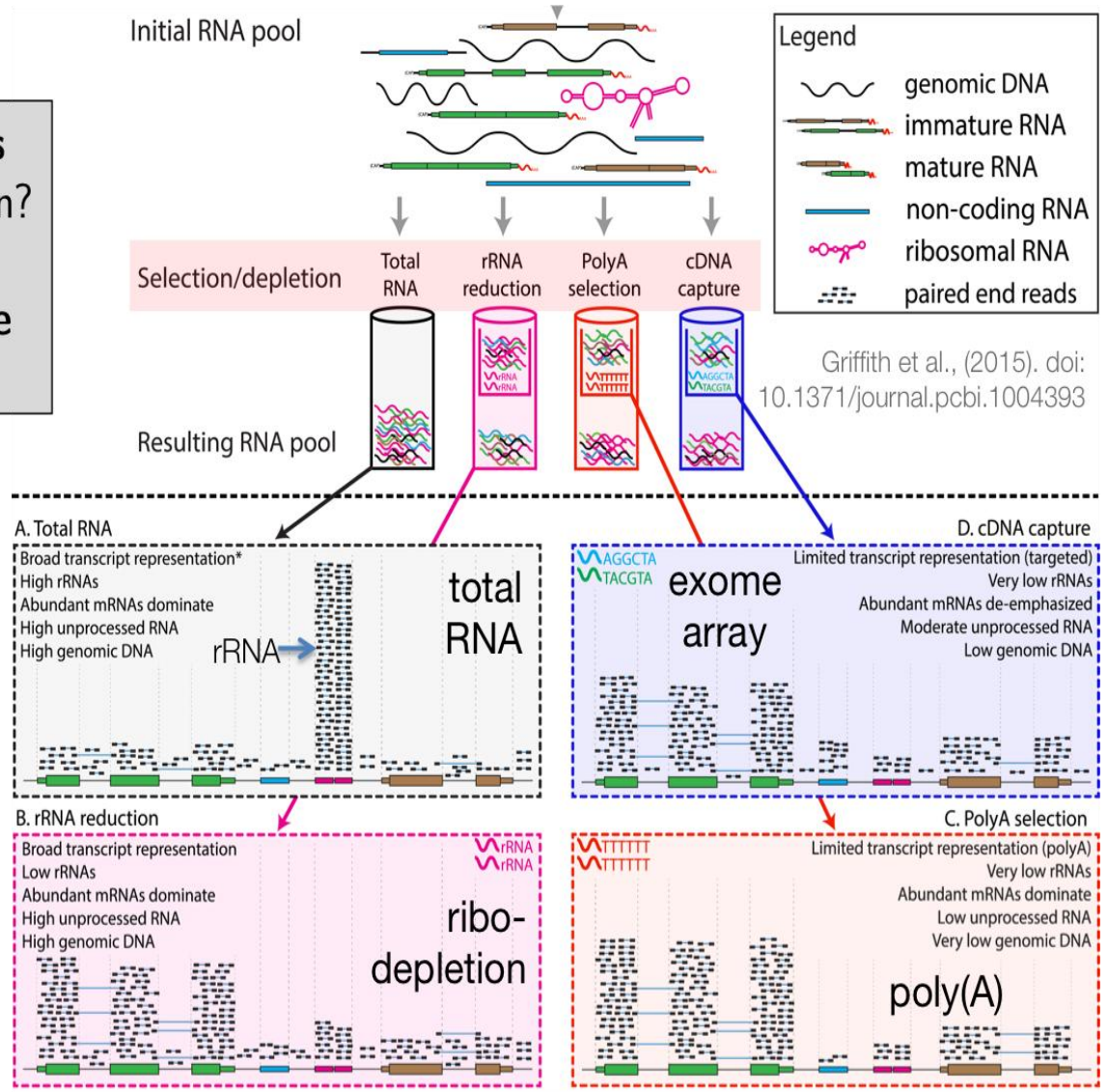
1. RNA extraction (cell lysis, RNA purification)
2. enrichment of the RNA of interest
   - mRNA: poly(A) enrichment vs. ribosomal-depletion
   - small RNAs: size-based enrichment
3. fragmentation (ca. 200 bp)
4. cDNA synthesis
5. library prep to obtain cDNA with adapters for sequencing



General steps of RNA-seq preparation

which **transcripts** are you interested in?

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



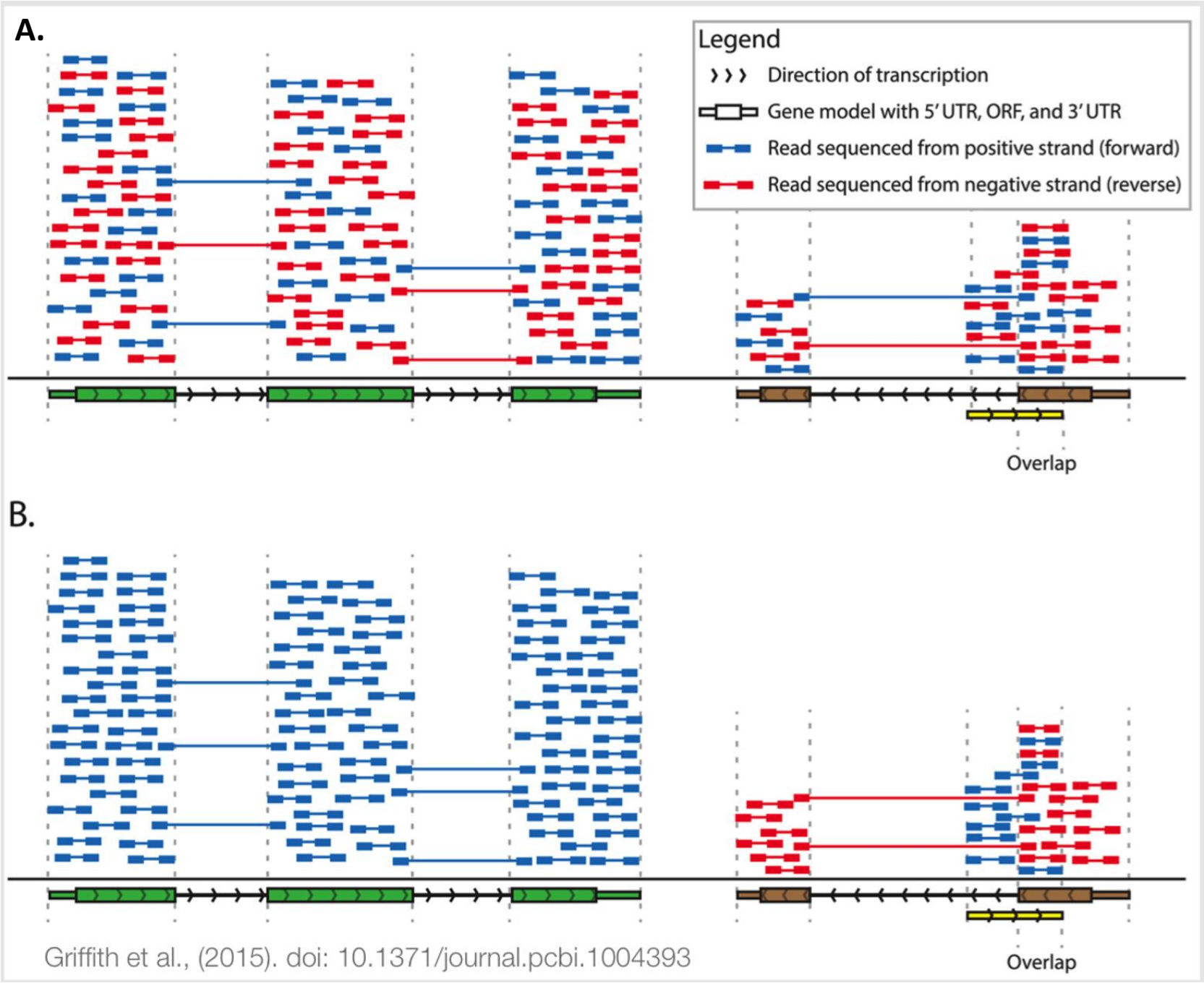
rRNA

protein coding (strongly expressed)

protein coding (lowly expressed)

Non-stranded

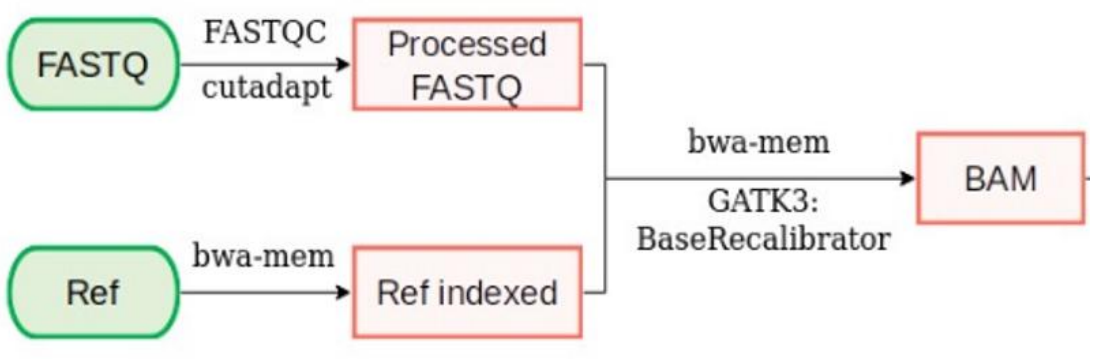
Stranded



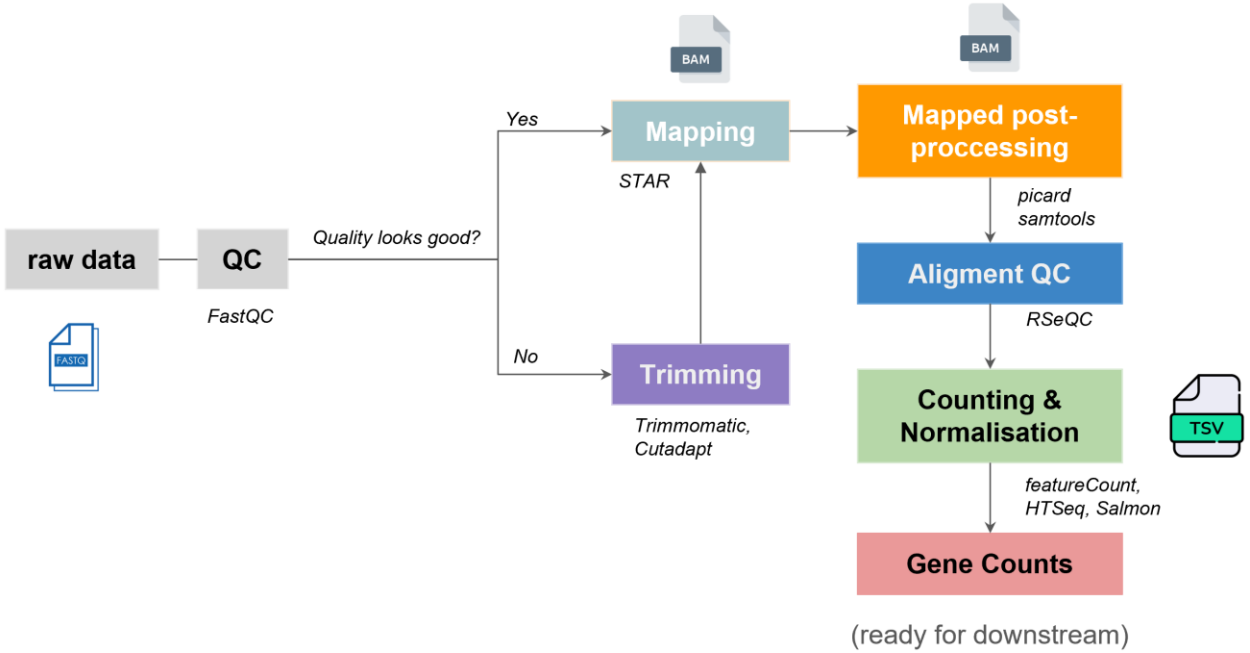
Griffith et al., (2015). doi: 10.1371/journal.pcbi.1004393

Upstream Analysis Workflow

Bulk DNA-seq

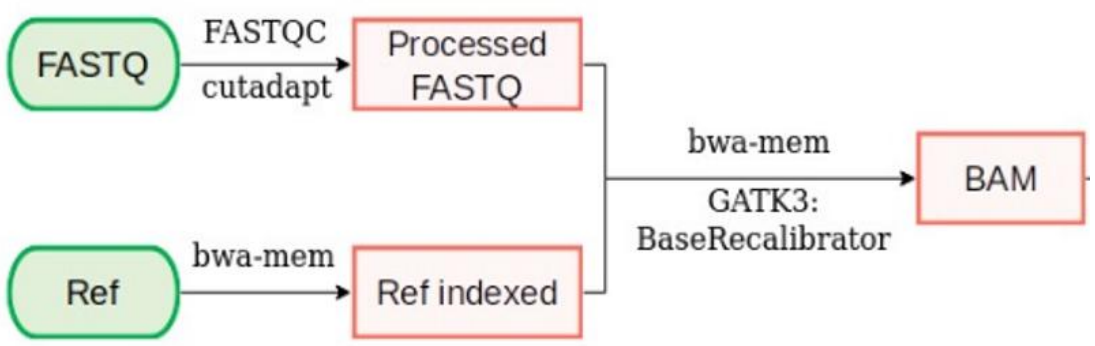


Bulk RNA-seq

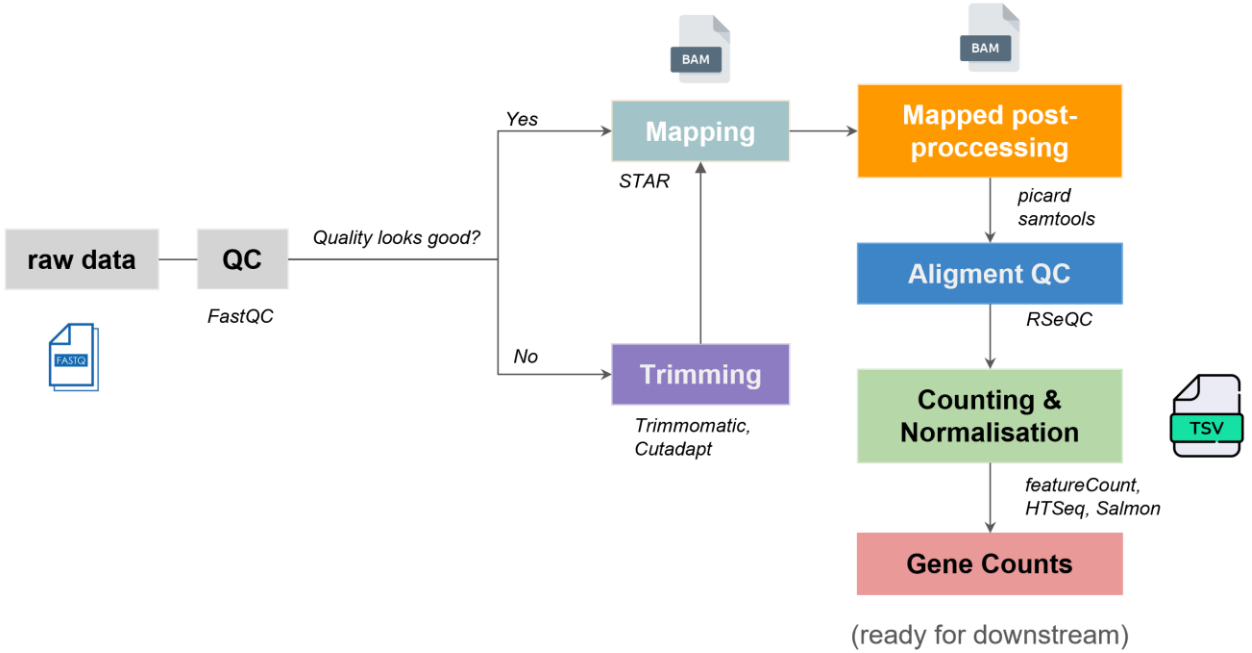


Upstream Analysis Workflow

Bulk DNA-seq



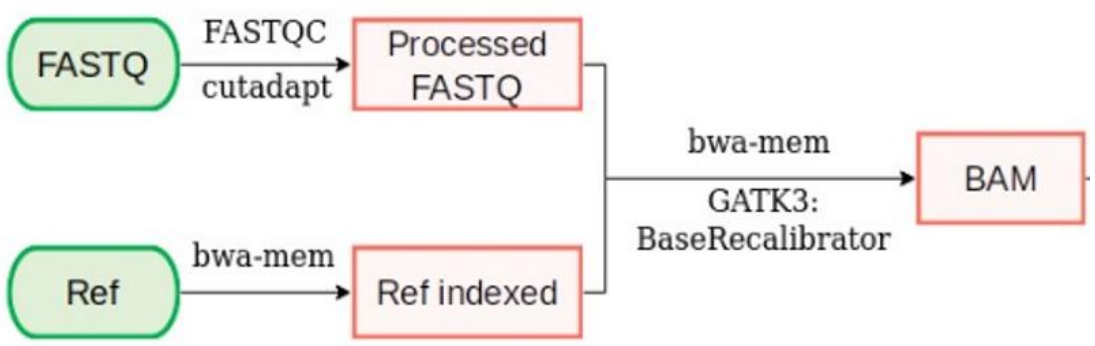
Bulk RNA-seq



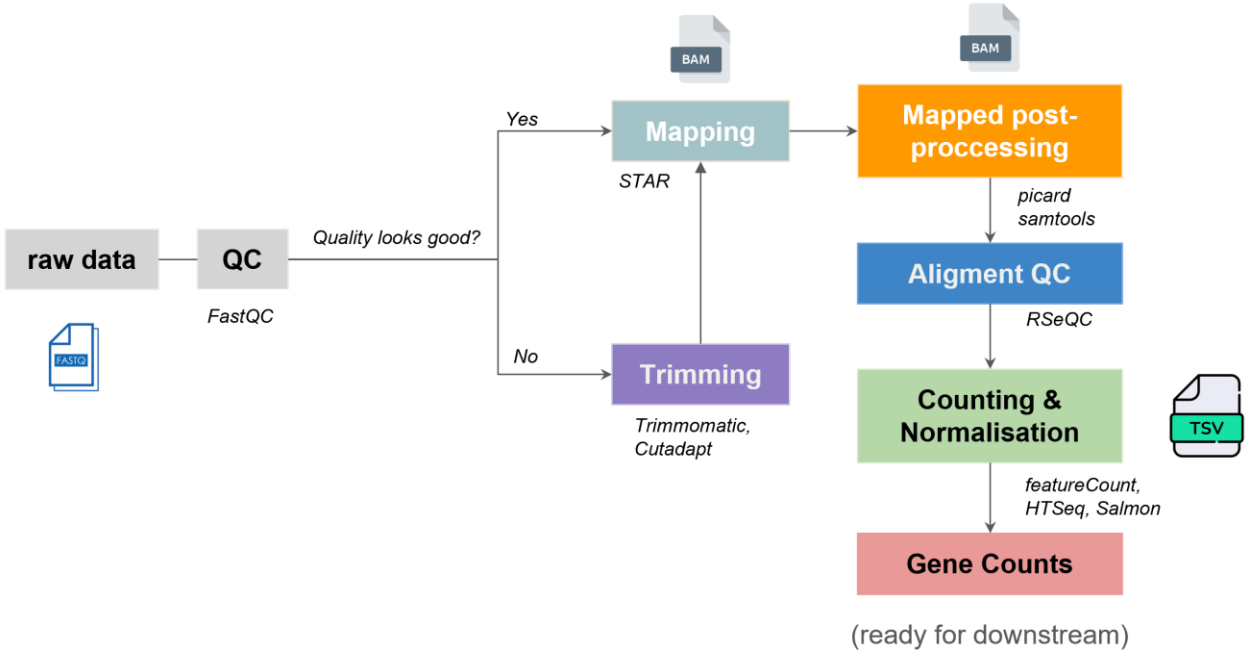
Where are the References need? And what are these?

Upstream Analysis Workflow

Bulk DNA-seq



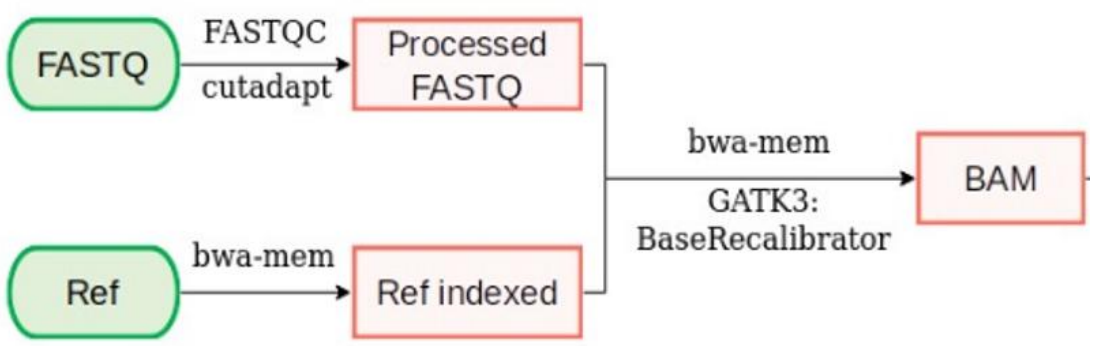
Bulk RNA-seq



Where are the References need? And what are these?

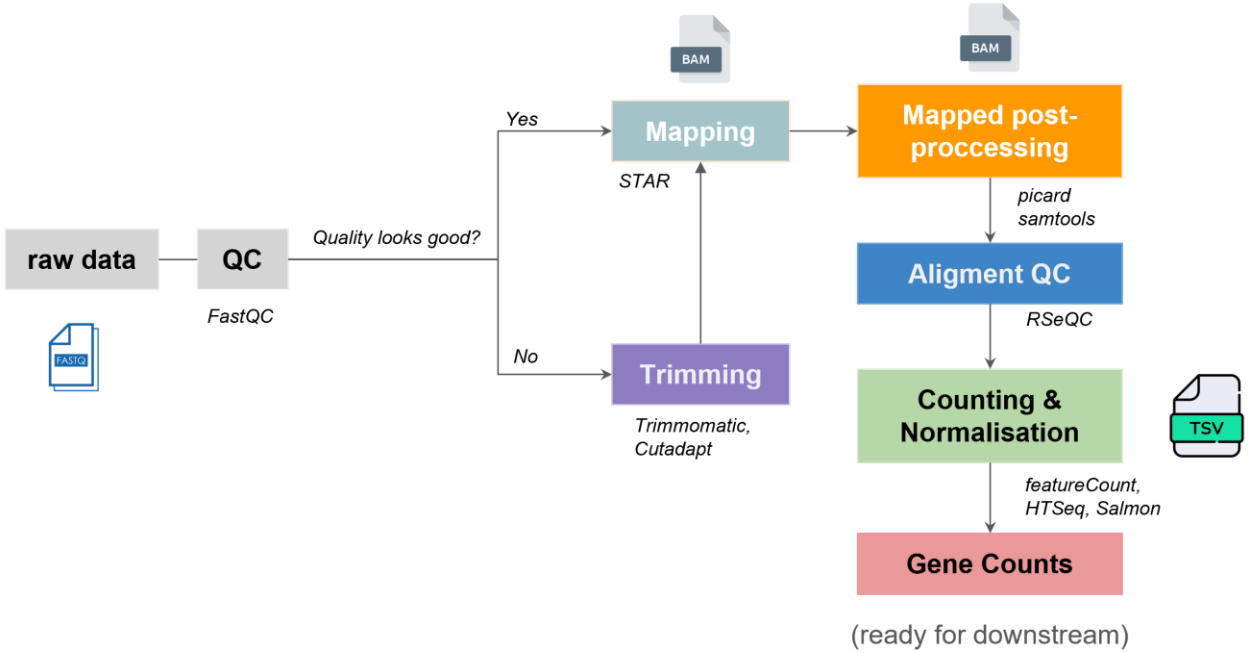
Upstream Analysis Workflow

Bulk DNA-seq



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

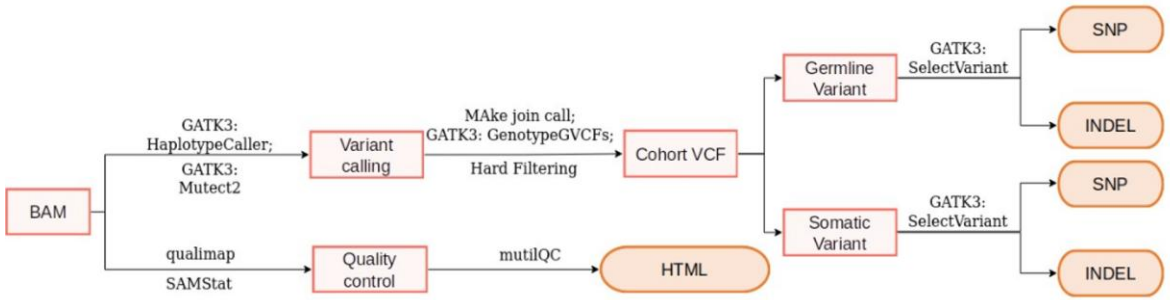
Bulk RNA-seq



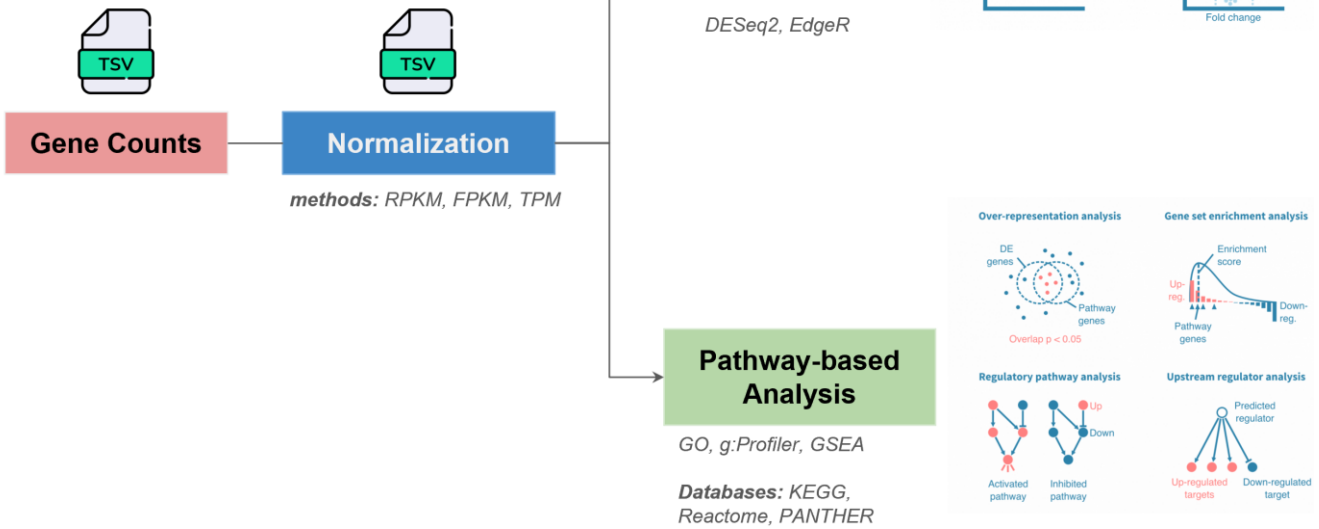
The four most common gene annotation databases are currently RefSeq, UCSC, Ensembl and GENCODE as GTF format

Downstream Analysis Workflow

Bulk DNA-seq

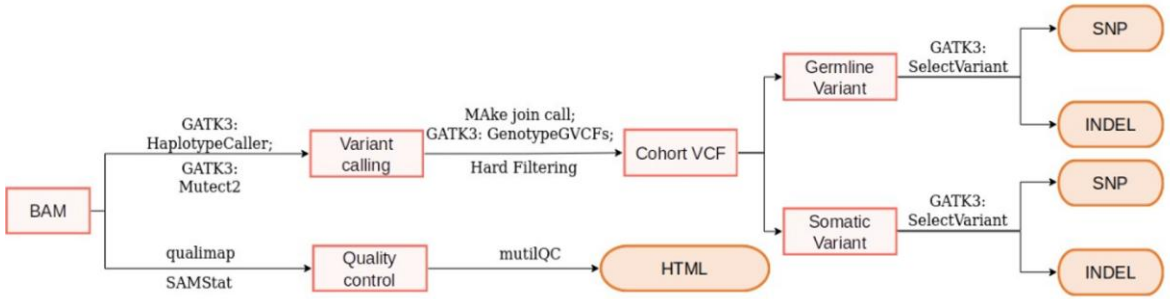


Bulk RNA-seq

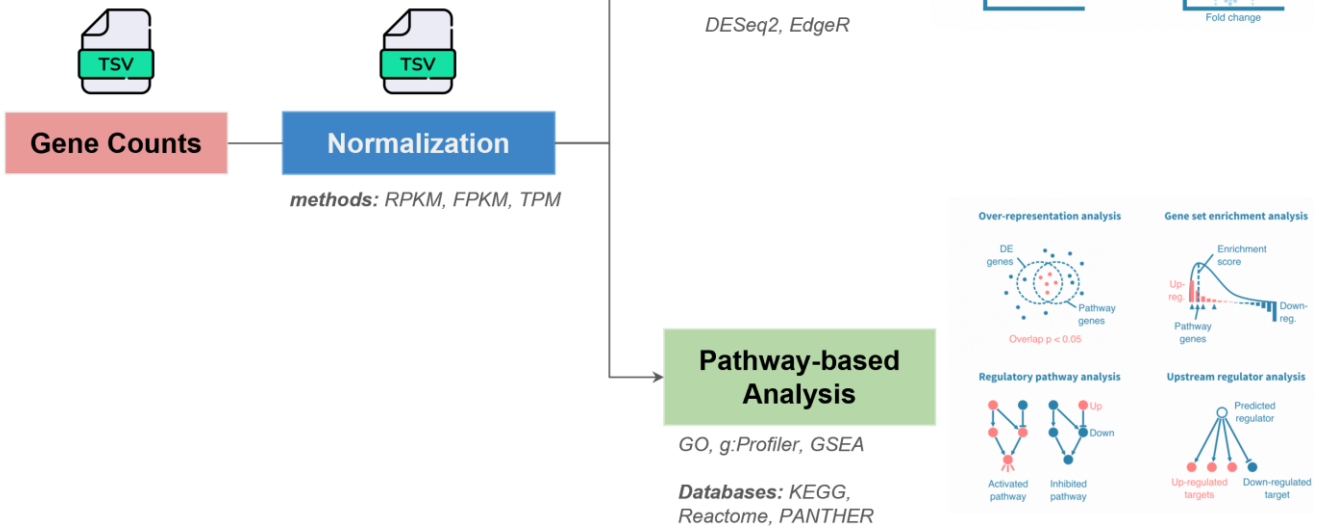


Downstream Analysis Workflow

Bulk DNA-seq



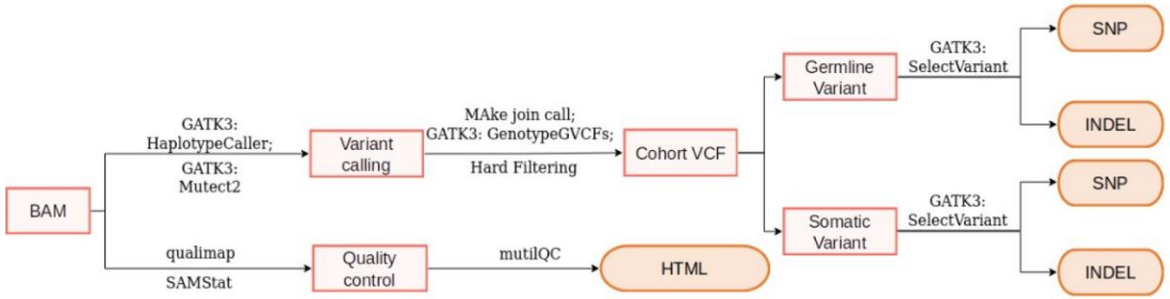
Bulk RNA-seq



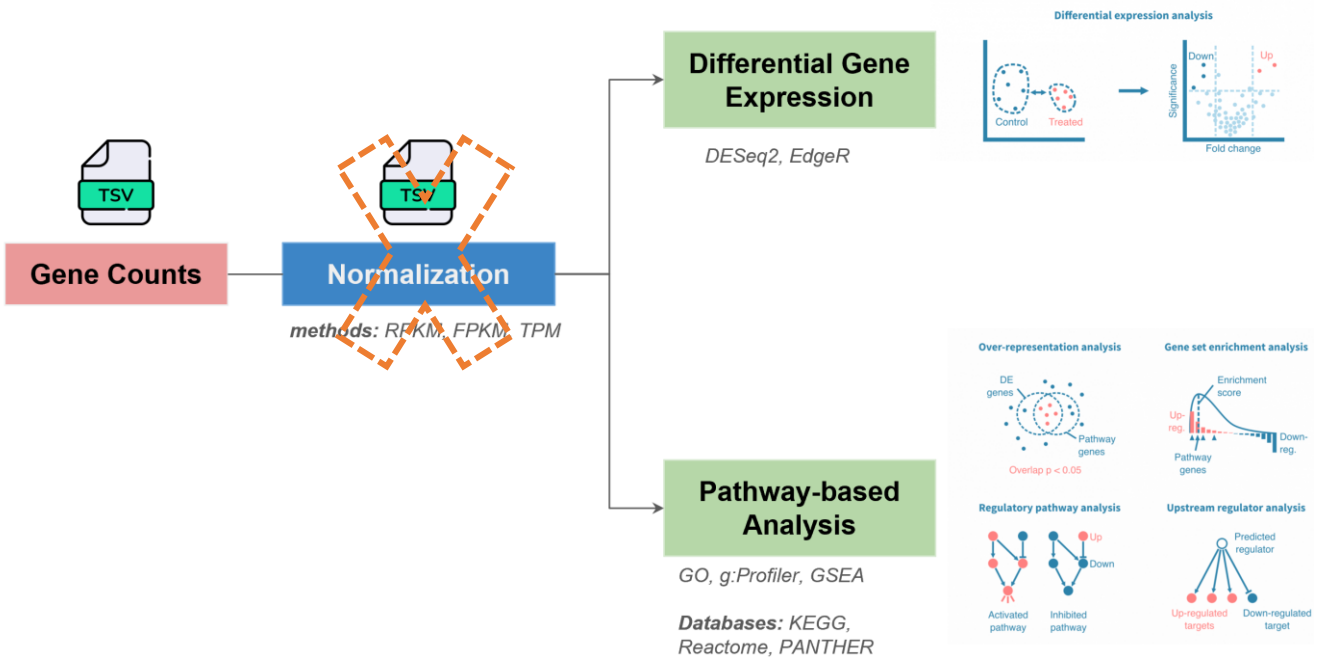
What is the unmatch?

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

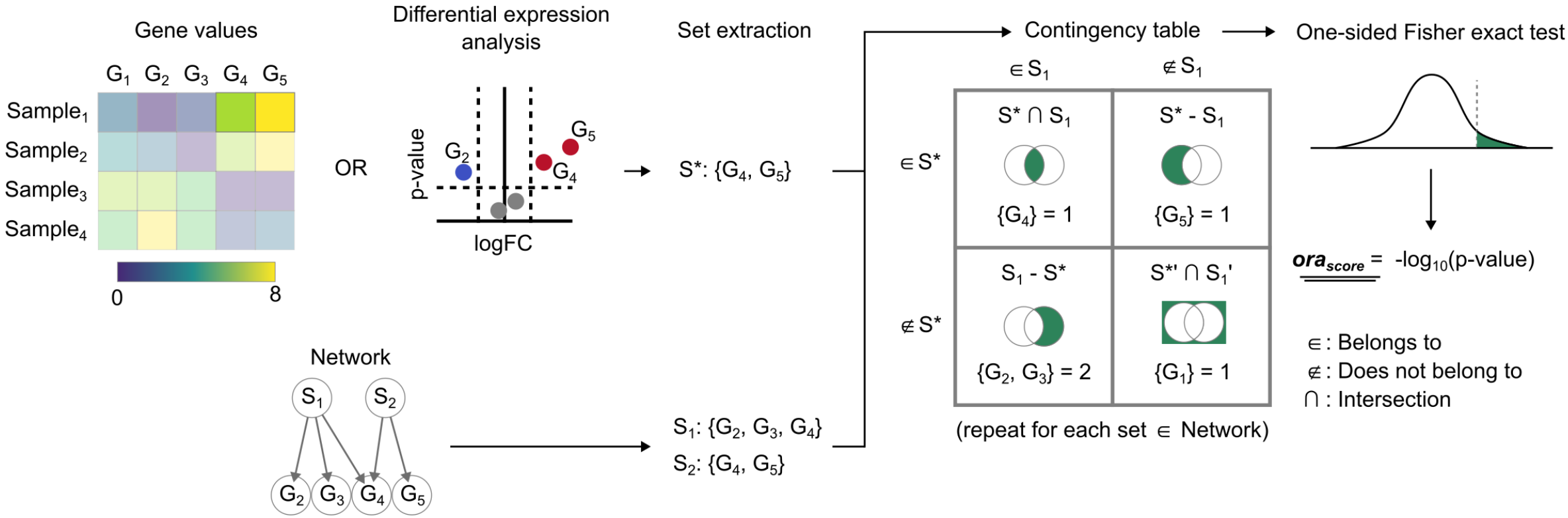
CALCULATE

One-tailed p value: **0.081178**

Two-tailed p value: **0.115239**

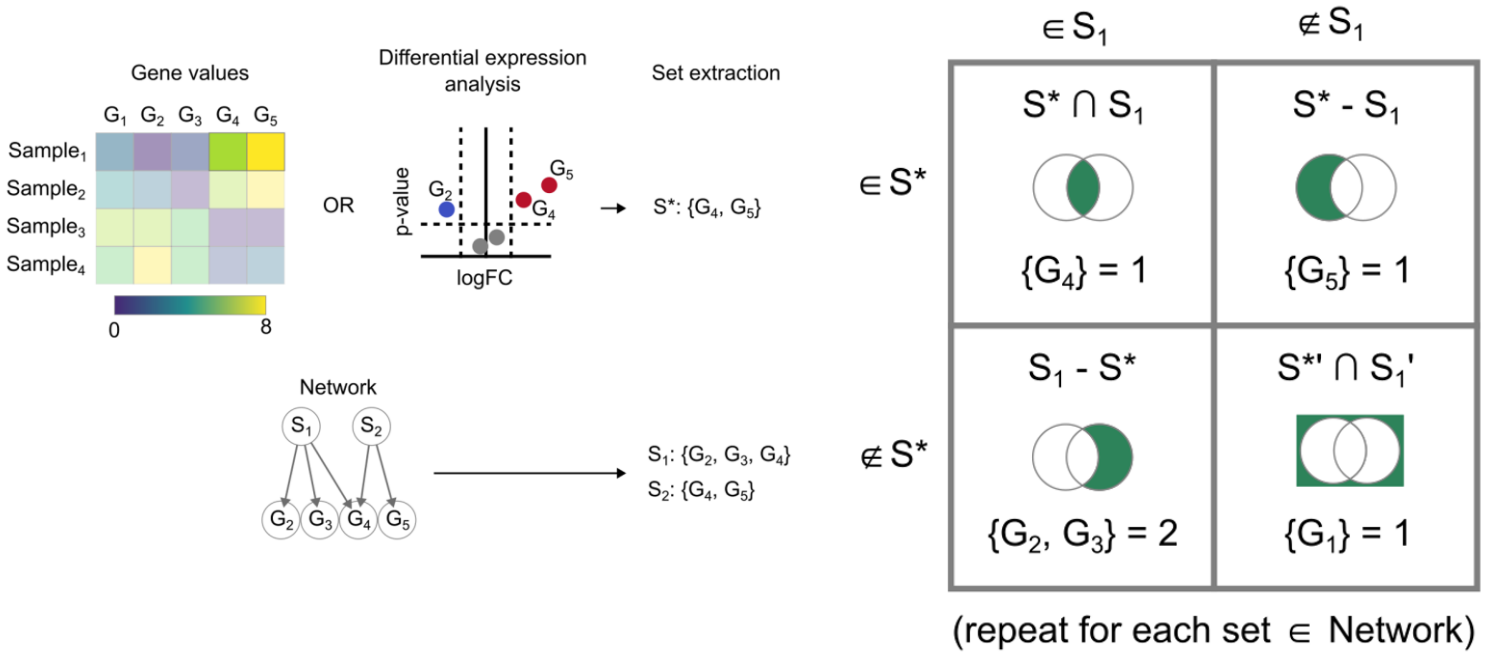
https://www.statology.org/fishers-exact-test/

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

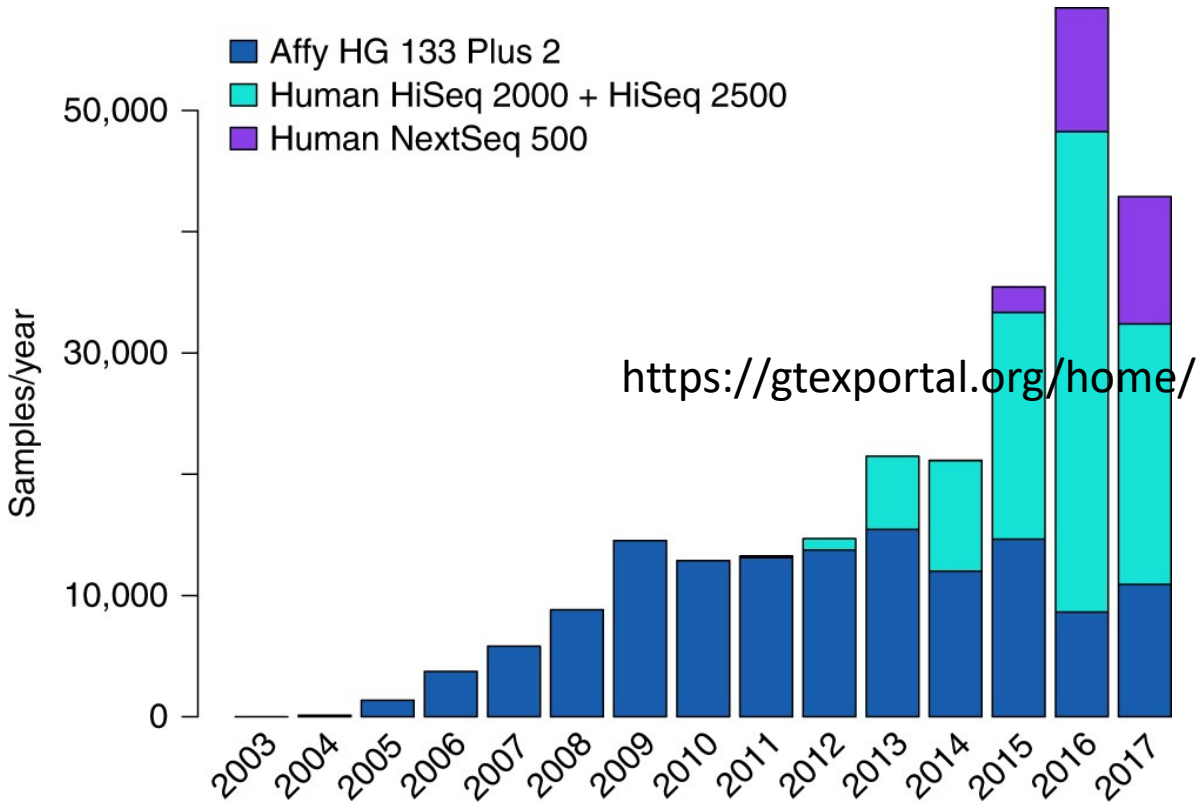
Enrichment of DEGs with Over Representation Analysis (ORA)

	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

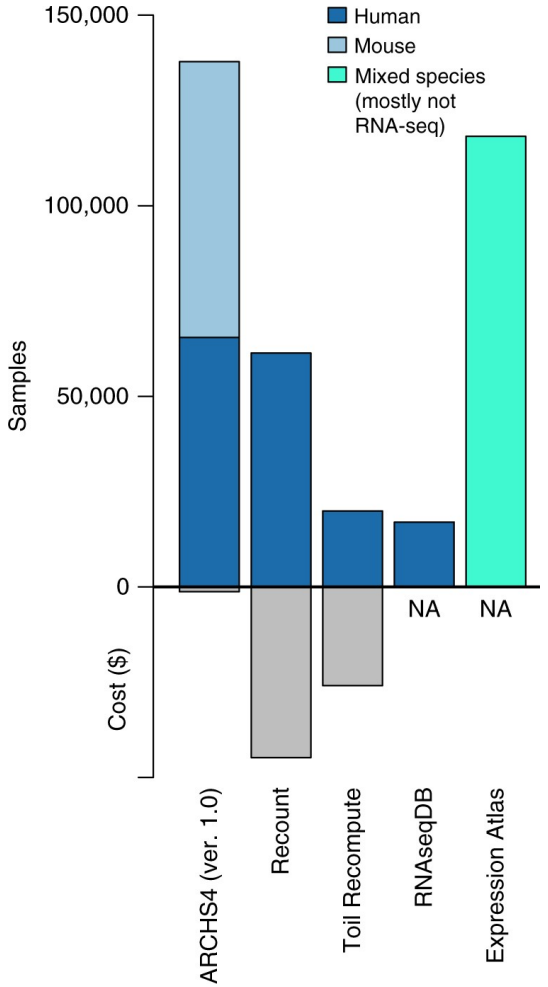
The qvalue in the results table from enrichGO represents the adjusted p-value for multiple testing correction. It is typically calculated using the False Discovery Rate (FDR) method, which controls the proportion of false positives among the list of enriched terms.

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

RNA-seq databases

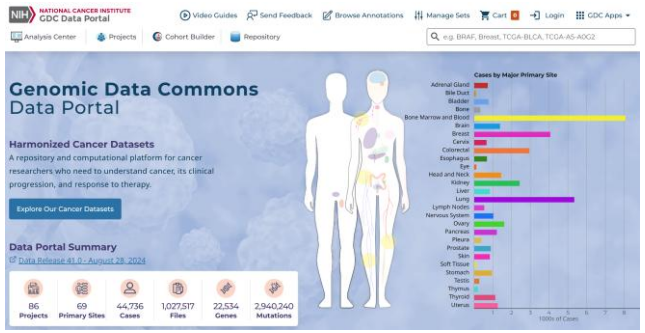


https://gtexportal.org/home/



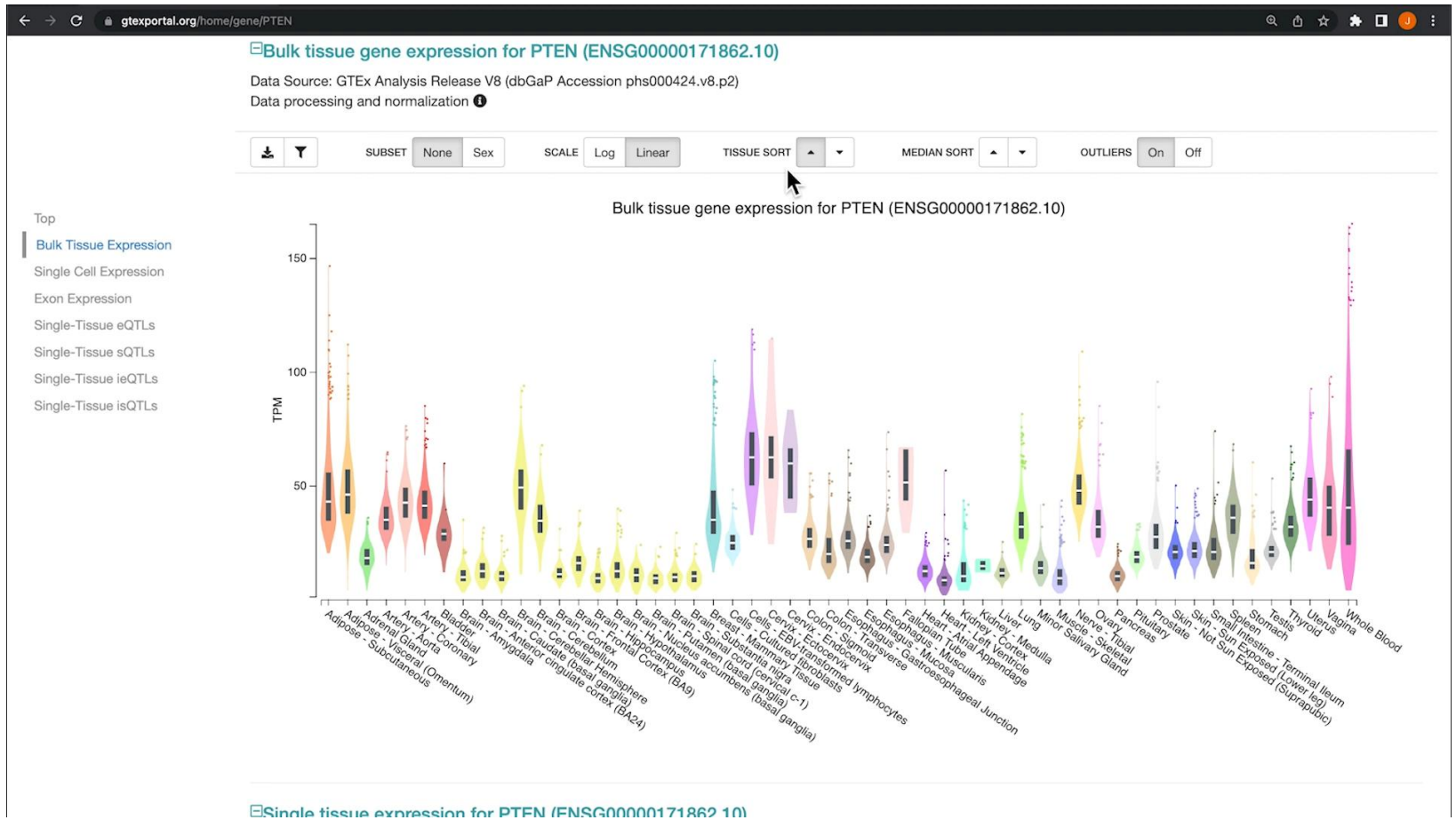
GTEx Portal

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/

Cảm ơn các thầy cô và các bạn!

Ngày 08 tháng 12 năm 2024

TS. Lưu Phúc Lợi

Email: luu.p.loi@googlemail.com

Zalo: 0901802182