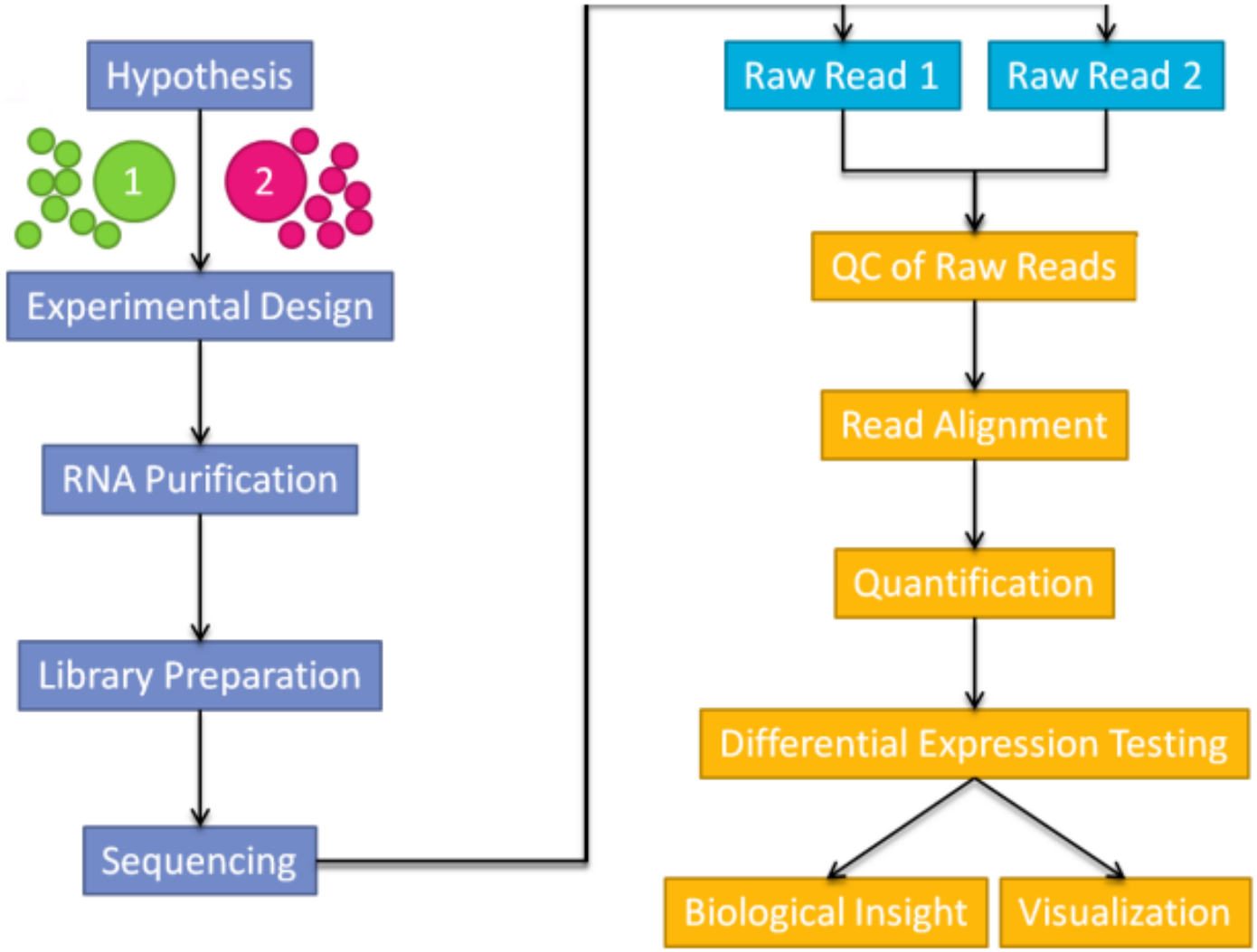
Differential Gene Expression analysis

TRANSCRIPTOMICS

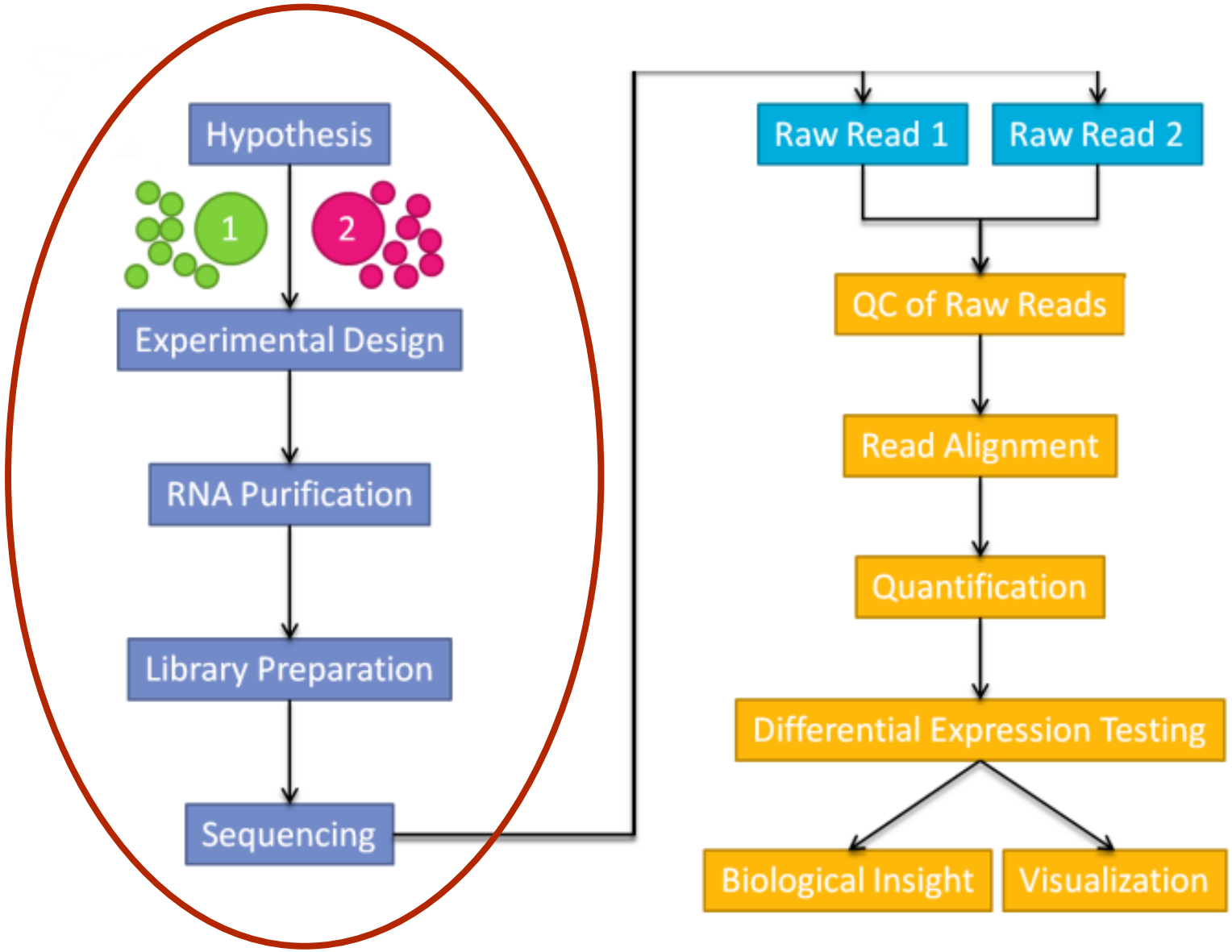
BIGNAUD AMAURY
30/03/2023

RNA-seq pipeline

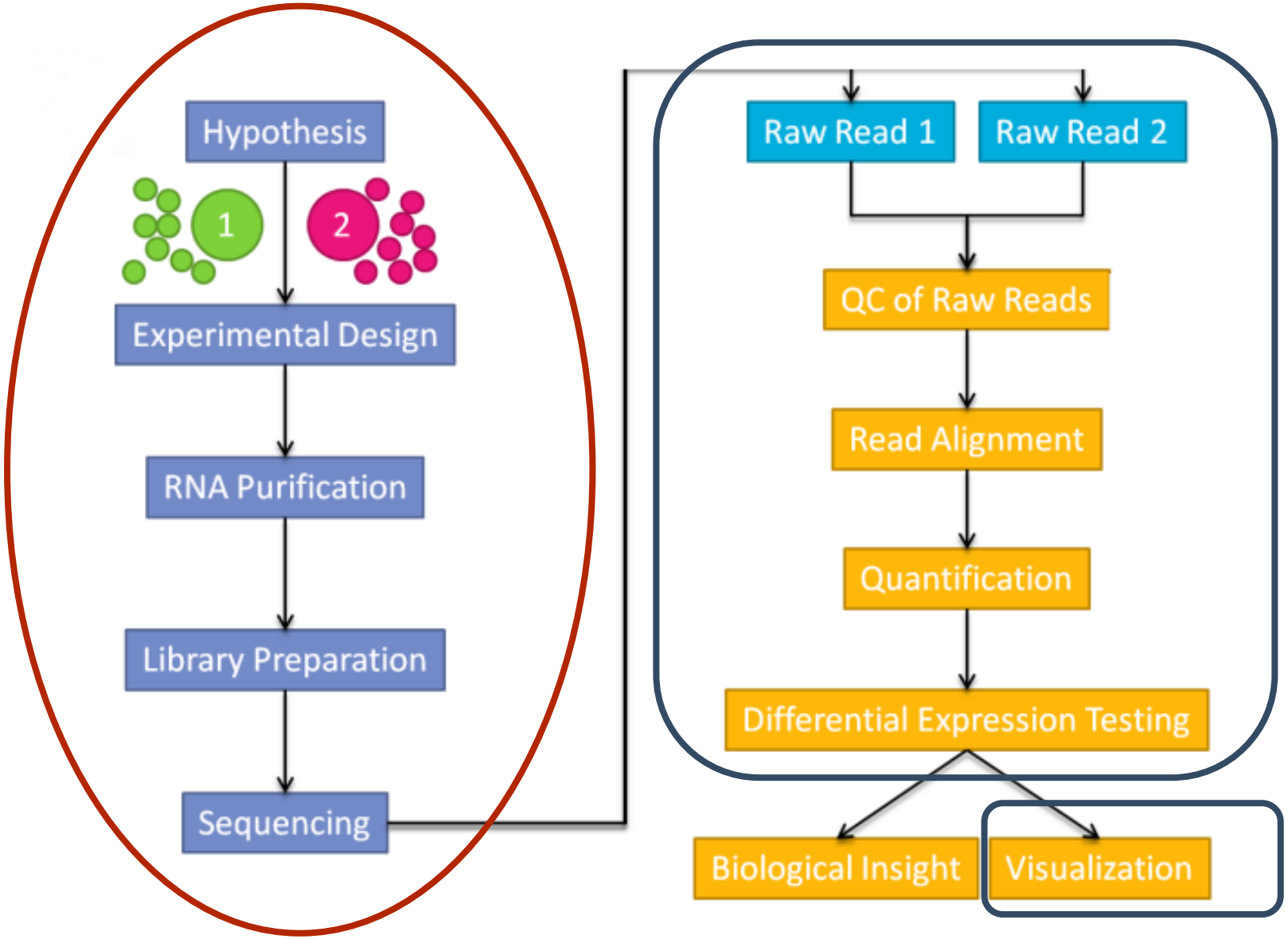


RNA-seq pipeline

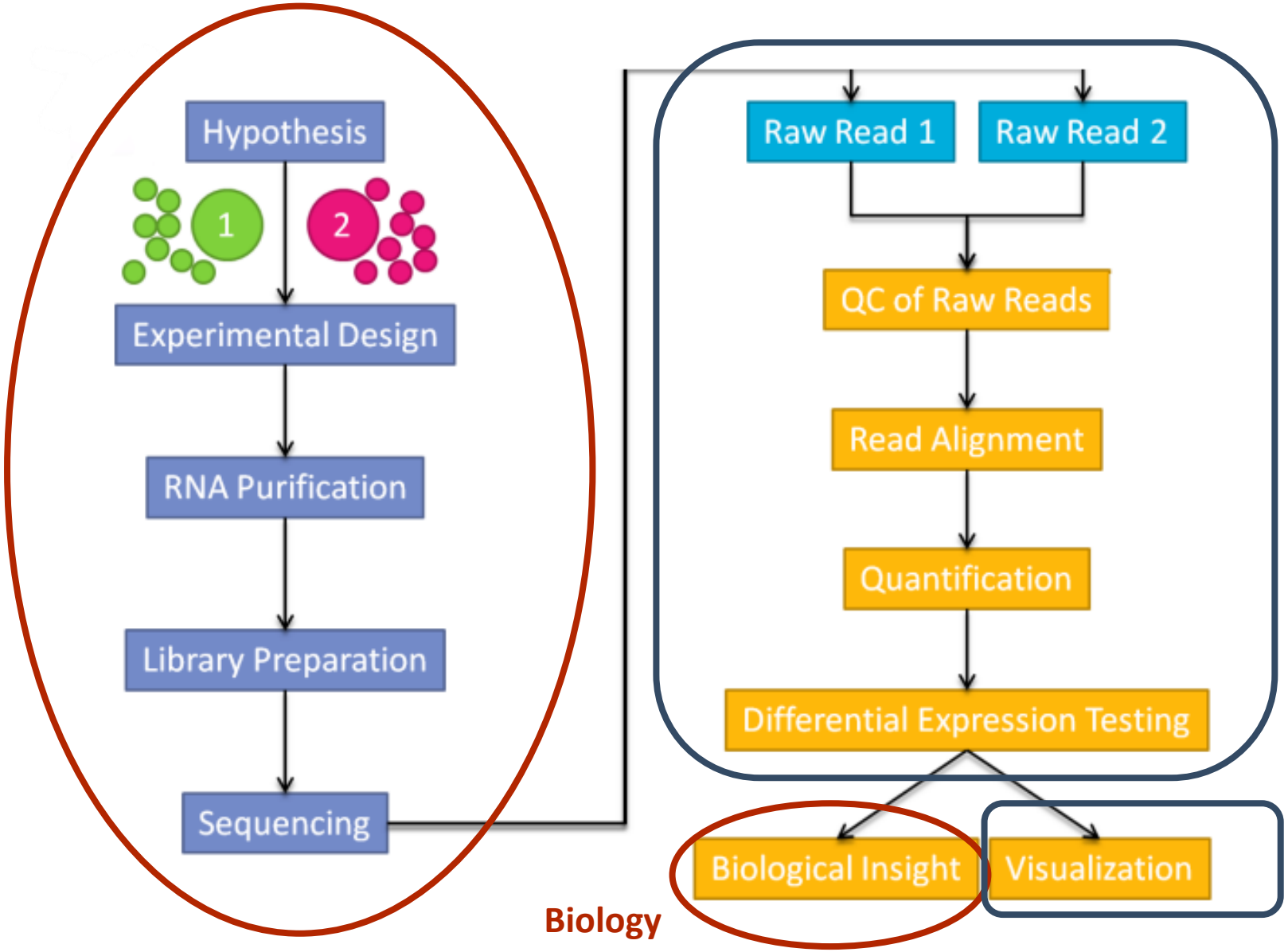
Biological experiment



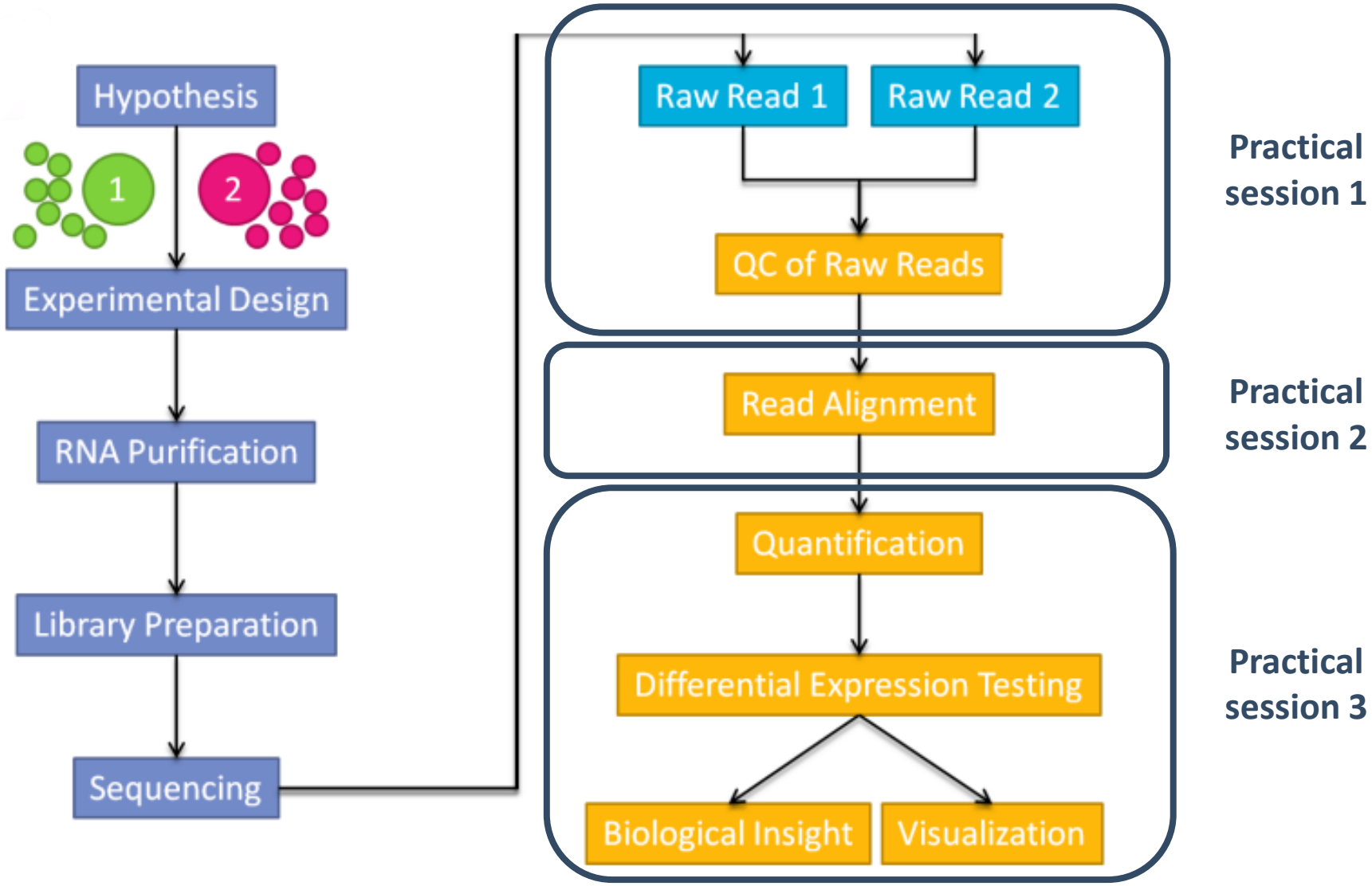
RNA-seq pipeline



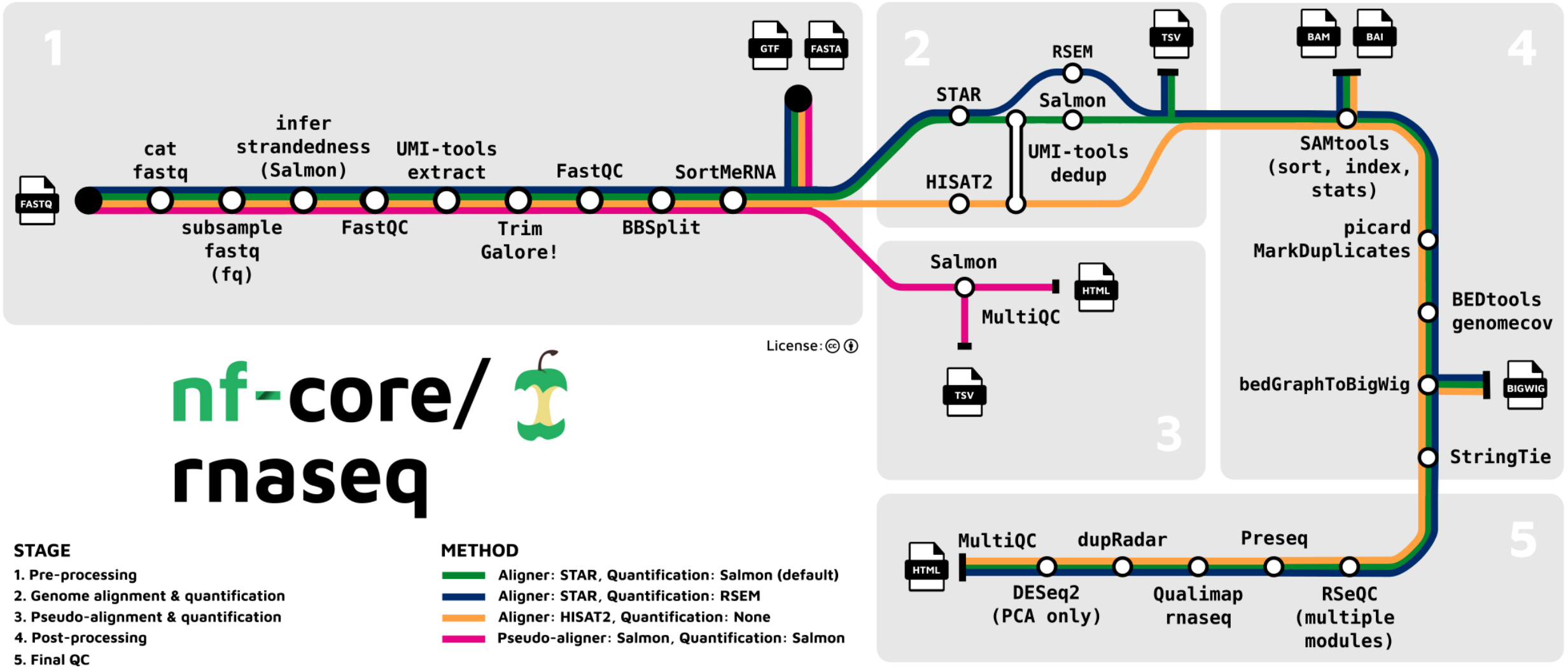
RNA-seq pipeline



RNA-seq pipeline

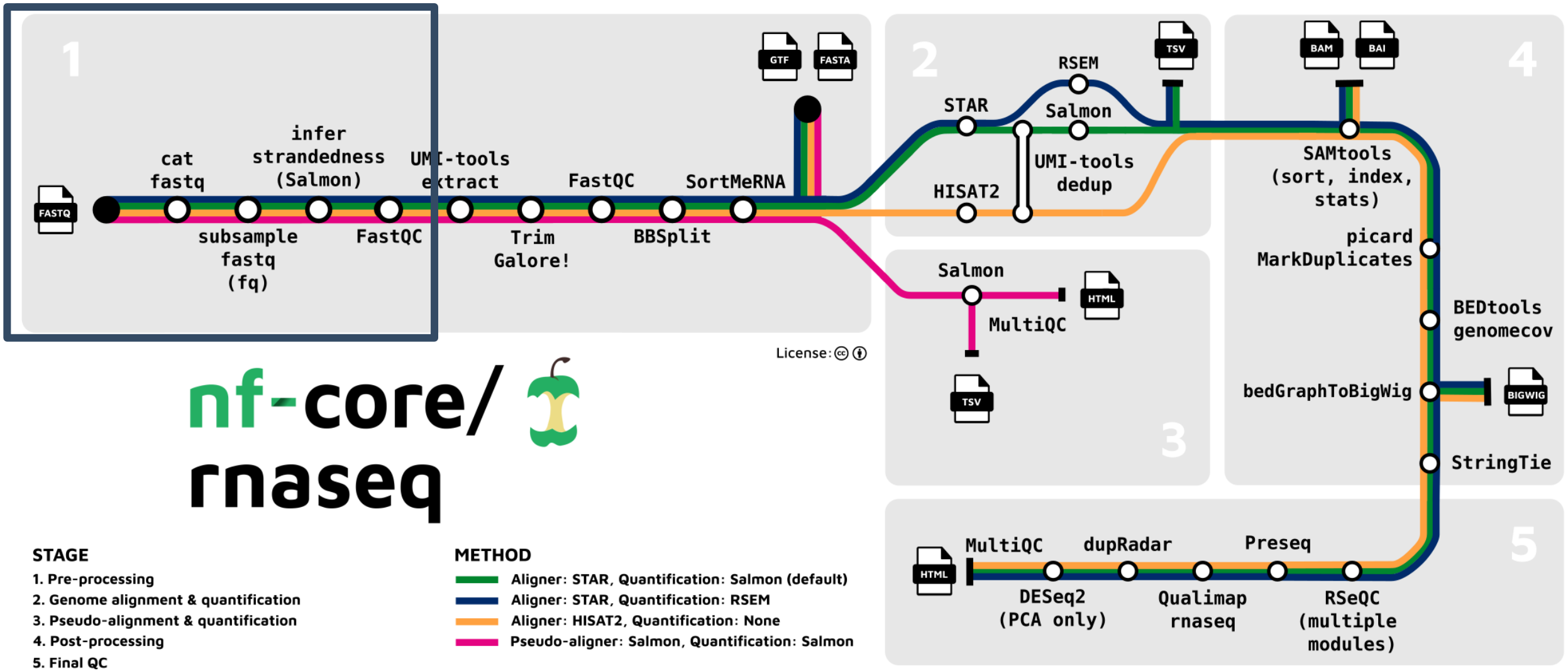


The RNAseq processing pipeline



The RNAseq processing pipeline

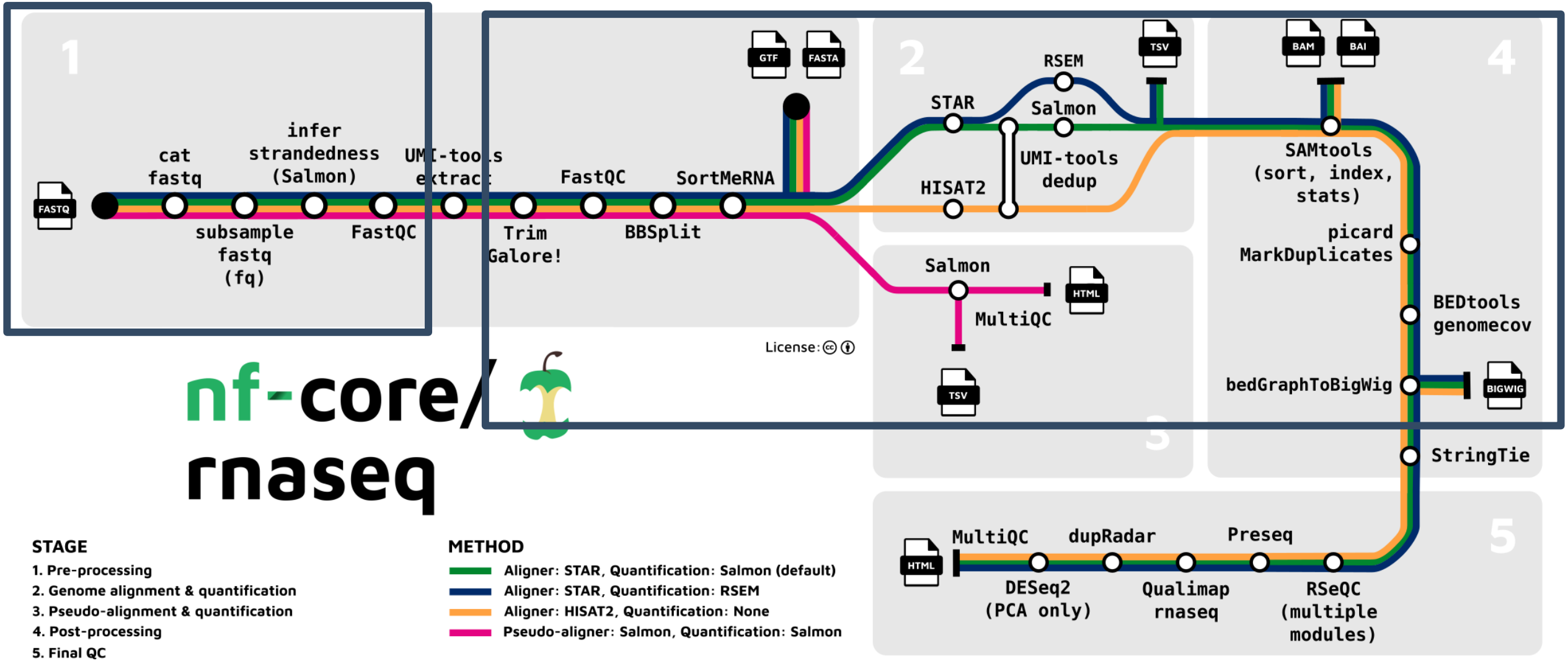
Practical session 1



The RNAseq processing pipeline

Practical session 1

Practical session 2



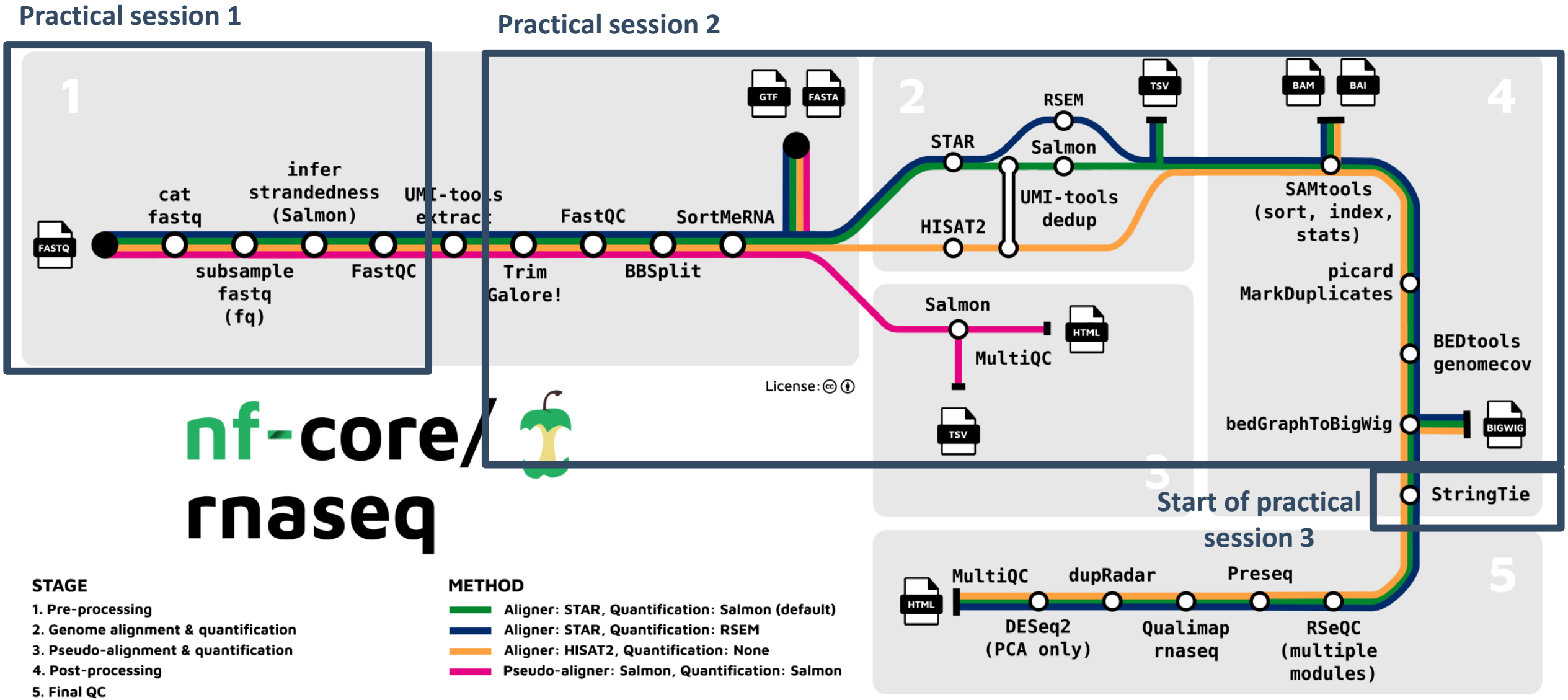
STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

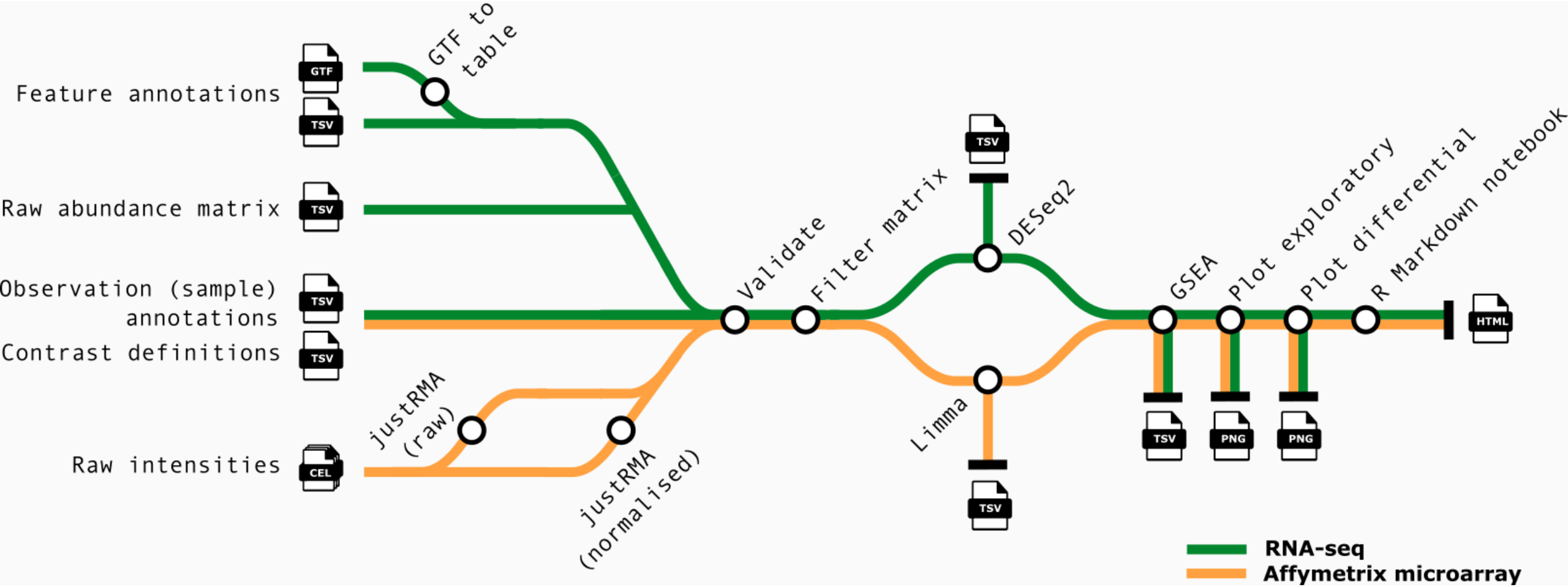
METHOD

- Aligner: STAR, Quantification: Salmon (default)
- Aligner: STAR, Quantification: RSEM
- Aligner: HISAT2, Quantification: None
- Pseudo-aligner: Salmon, Quantification: Salmon

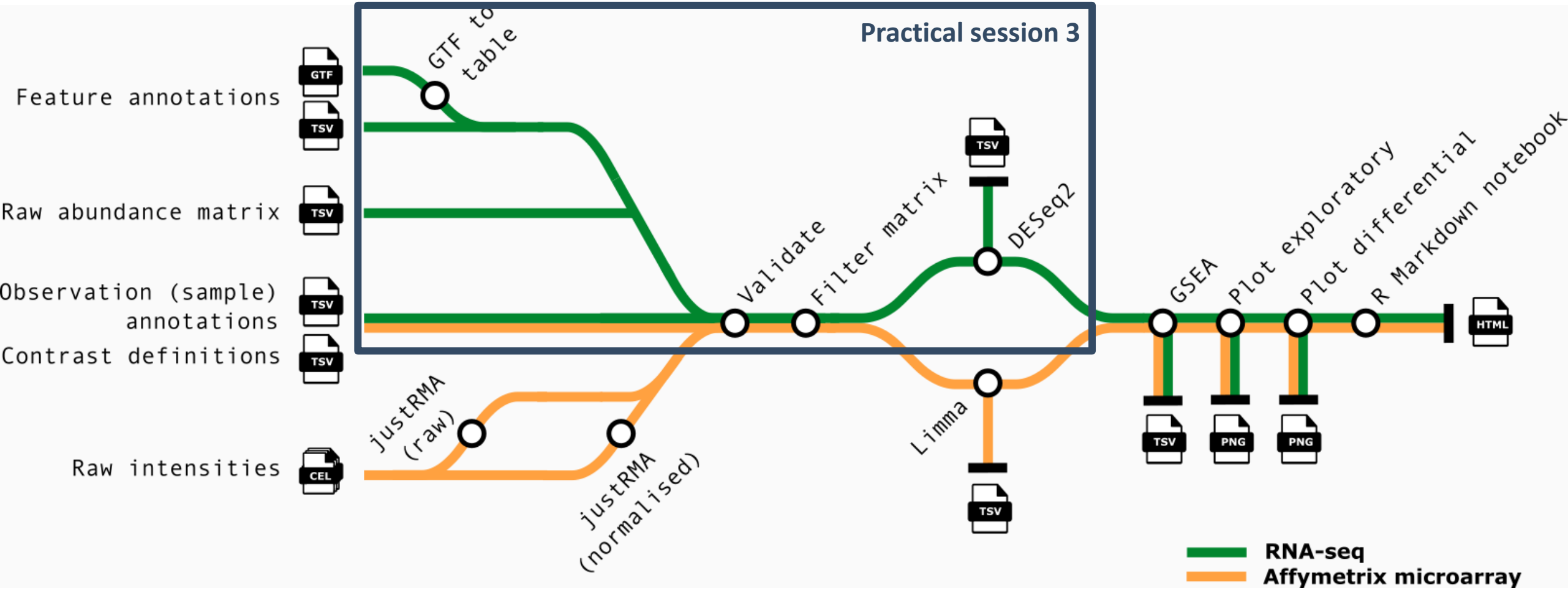
The RNAseq processing pipeline



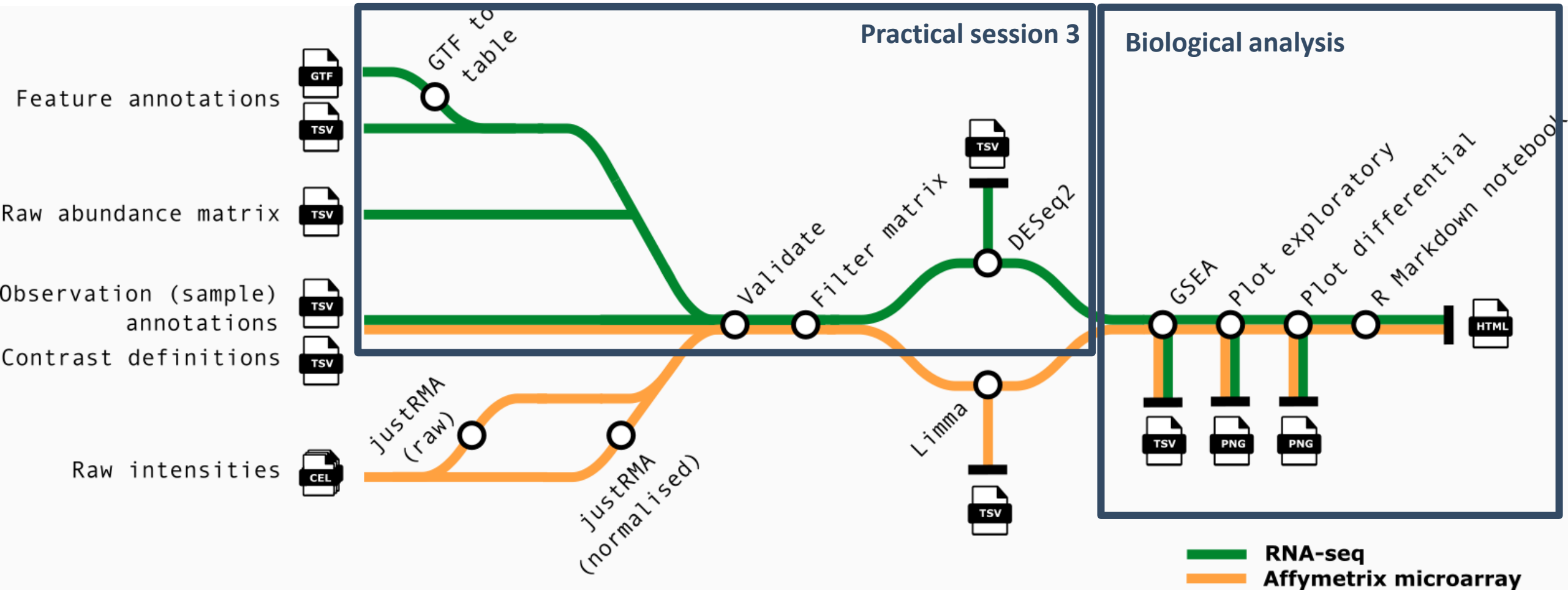
Differential gene expression pipeline



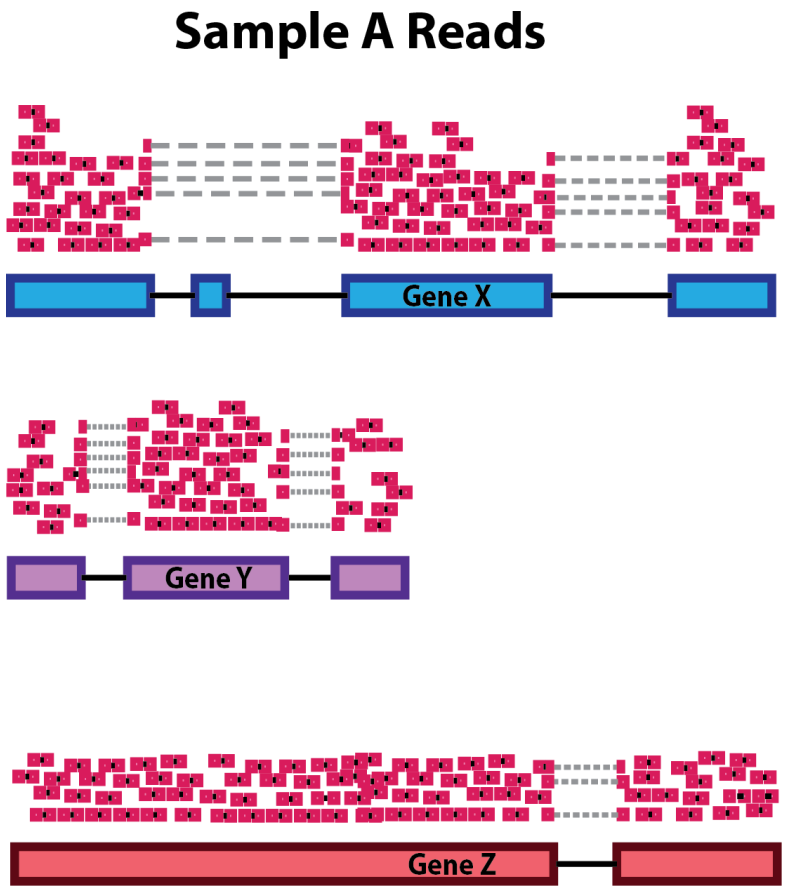
Differential gene expression pipeline



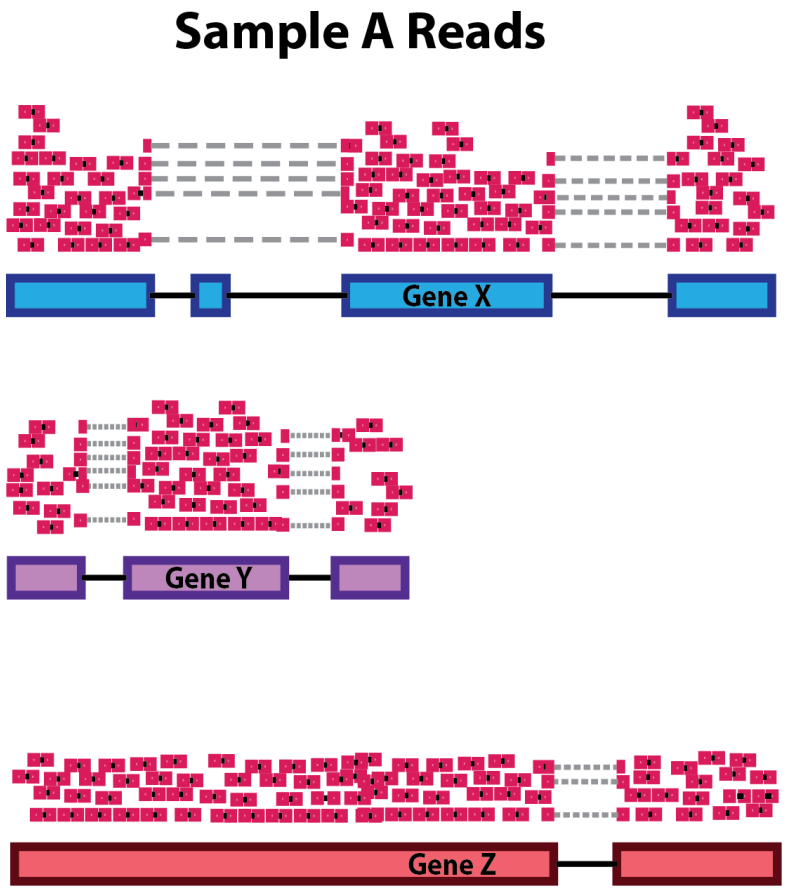
Differential gene expression pipeline



Count reads on the genes



Count reads on the genes



What do we need to do ?

- **Gene annotation**.
- Counting number of reads on the genes.

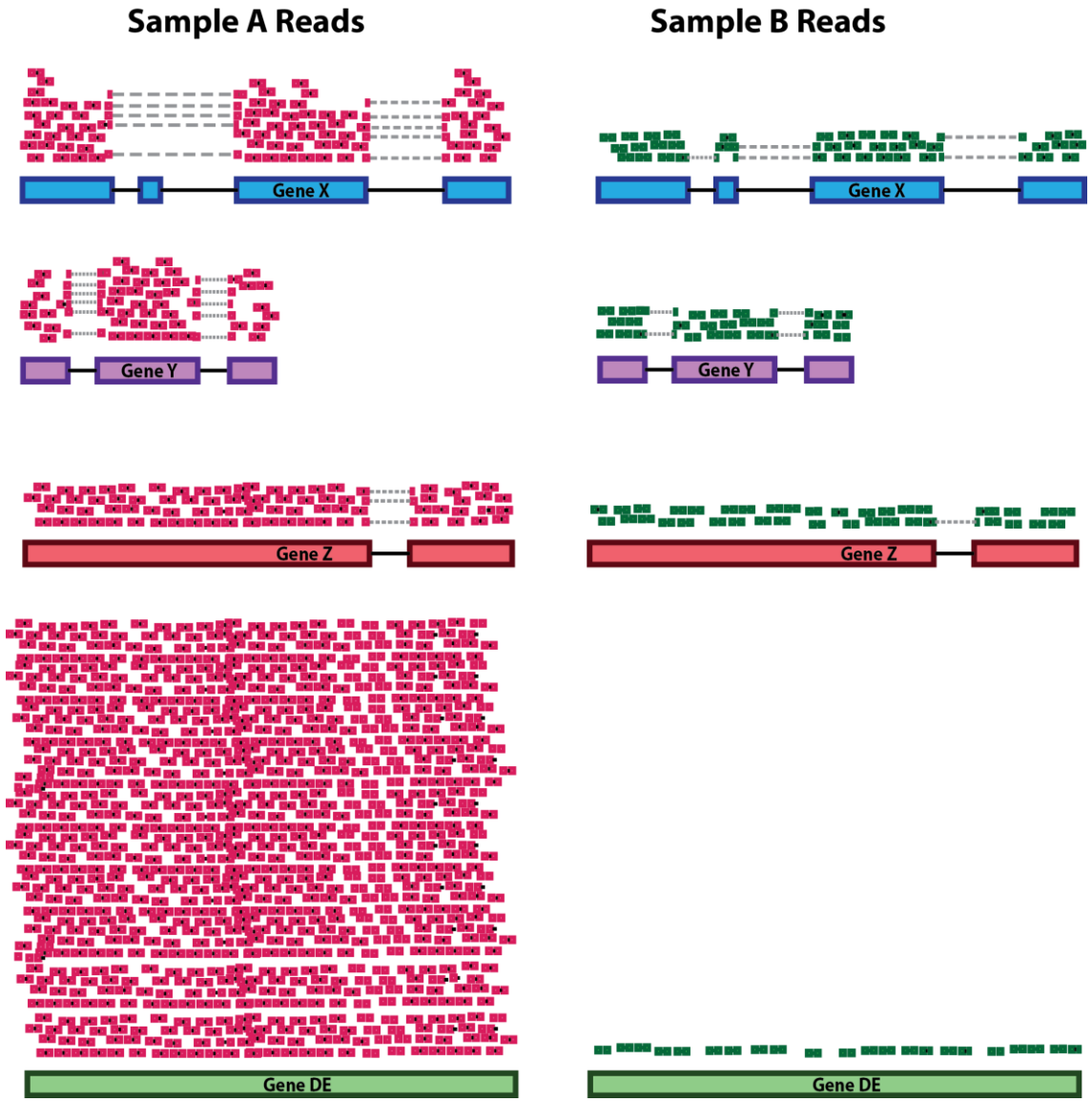
The annotation file format gff

General GFF3 structure

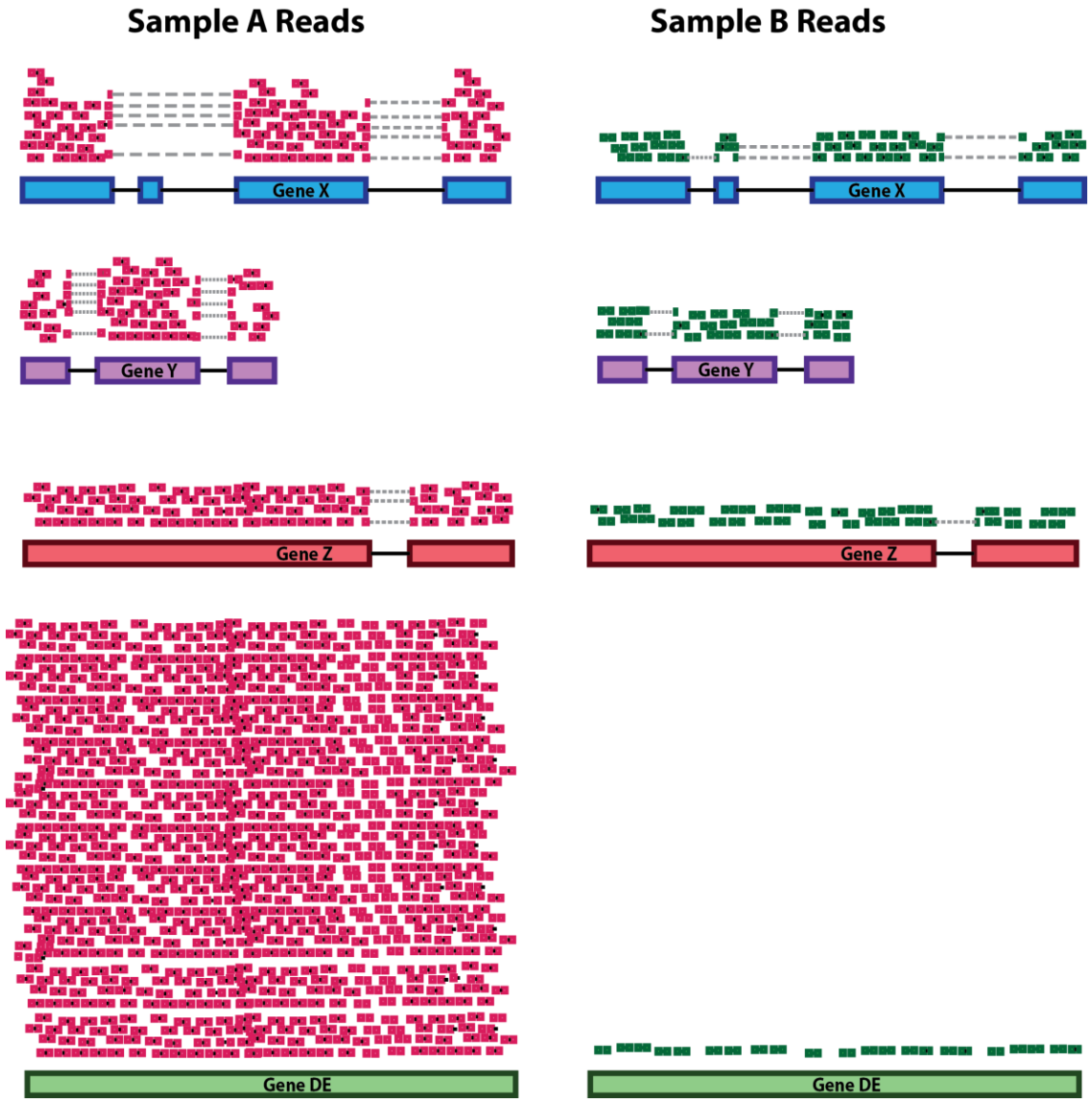
Position index	Position name	Description
1	seqid	The name of the sequence where the feature is located.
2	source	The algorithm or procedure that generated the feature. This is typically the name of a software or database.
3	type	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project.
4	start	Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED.
5	end	Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED.[*citation needed*]
6	score	Numeric value that generally indicates the confidence of the source in the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the strand of the feature. This can be "+" (positive, or 5'->3'), "-", (negative, or 3'->5'), "." (undetermined), or "?" for features with relevant but unknown strands.
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	attributes	A list of tag-value pairs separated by a semicolon with additional information about the feature.

```
manual_scaffold_1	Prodigal:002006	CDS	12197	14662	.	-	0	gene_id=FGBKPADF_00001;eC_number=5.6.2.2;Name=gyrA;db_xref=COG:COG0188;gene=gyrA;inference=ab in
manual_scaffold_1	Prodigal:002006	CDS	14873	16795	.	-	0	gene_id=FGBKPADF_00002;eC_number=5.6.2.2;Name=gyrB;db_xref=COG:COG0187;gene=gyrB;inference=ab in
manual_scaffold_1	Prodigal:002006	CDS	16844	17089	.	-	0	gene_id=FGBKPADF_00003;inference=ab initio prediction:Prodigal:002006;locus_tag=FGBKPADF_00003;p
```

Count the features

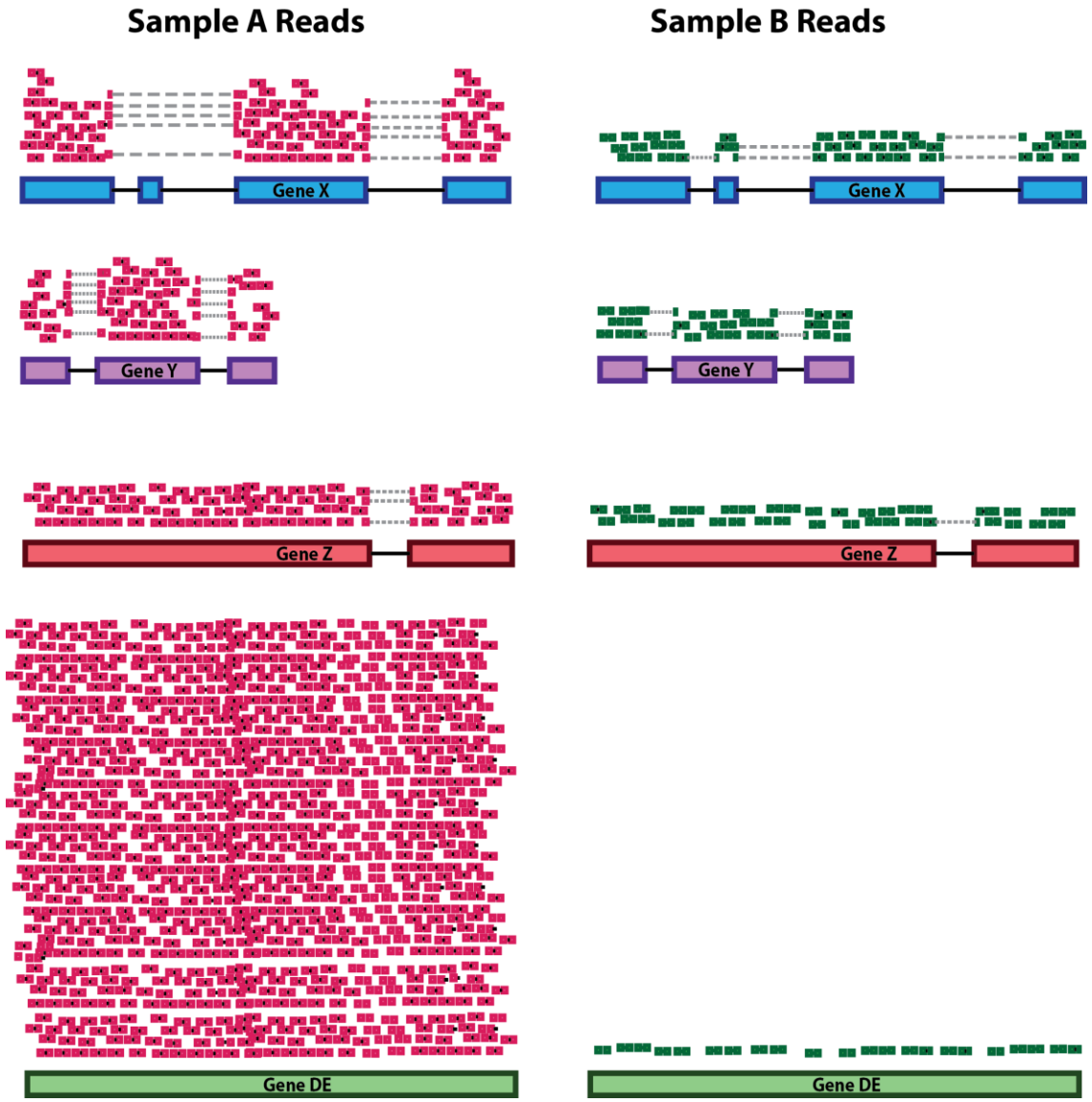


Count the features



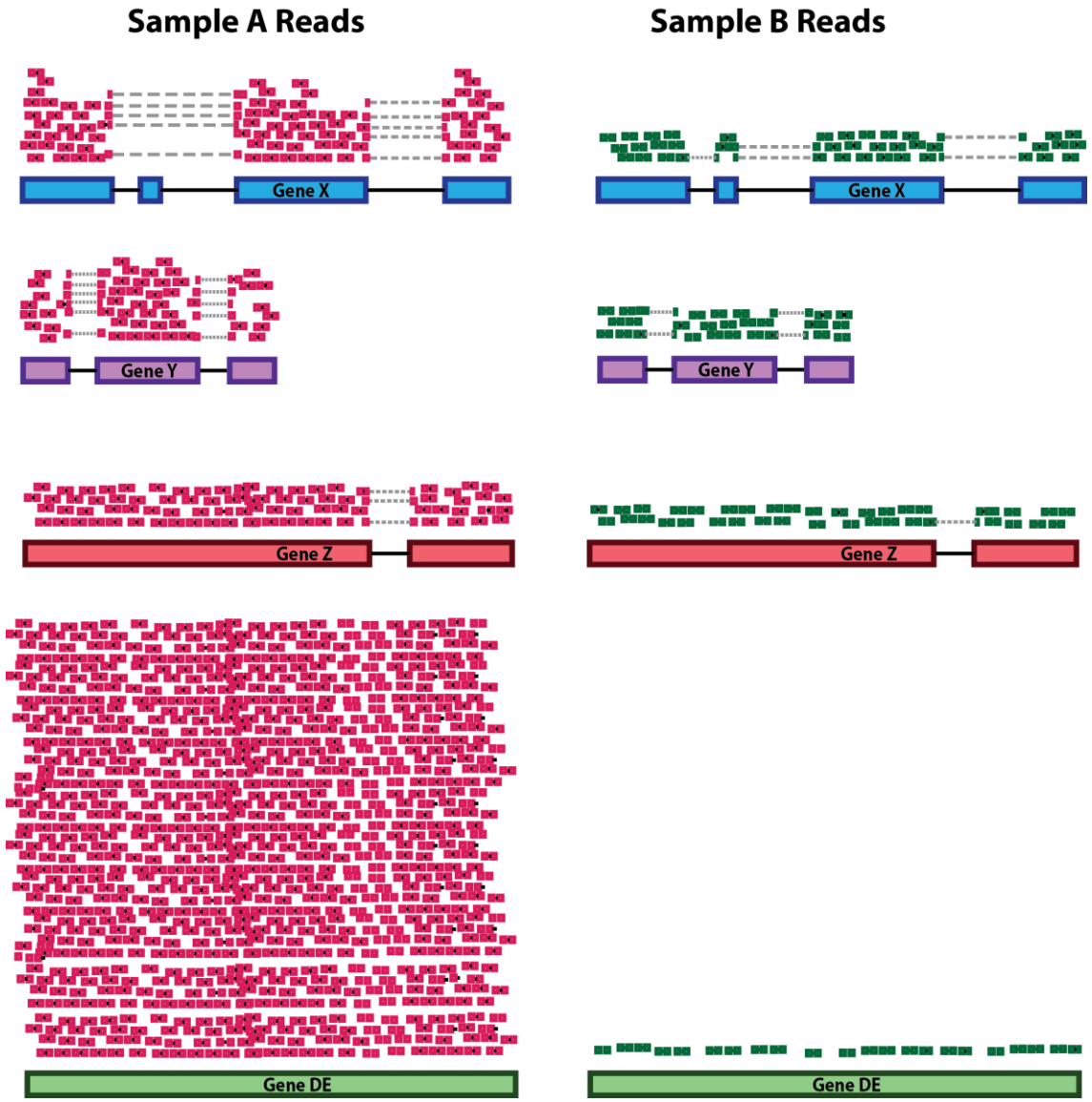
	Count sample A	Count sample B
Gene X	100	50
Gene Y	50	25
Gene Z	50	25
Gene DE	400	50

How to compare samples ?



From HBC training

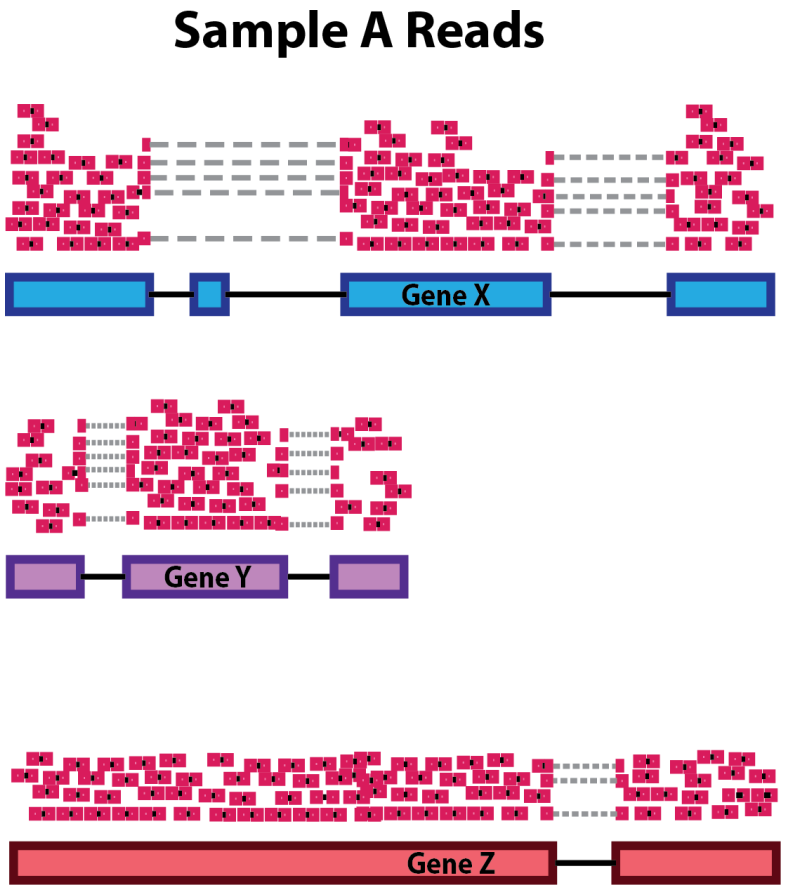
How to compare samples ?



	Count sample A	Count sample B	Fold change
Gene X	100	50	2
Gene Y	50	25	2
Gene Z	50	25	2
Gene DE	400	50	8

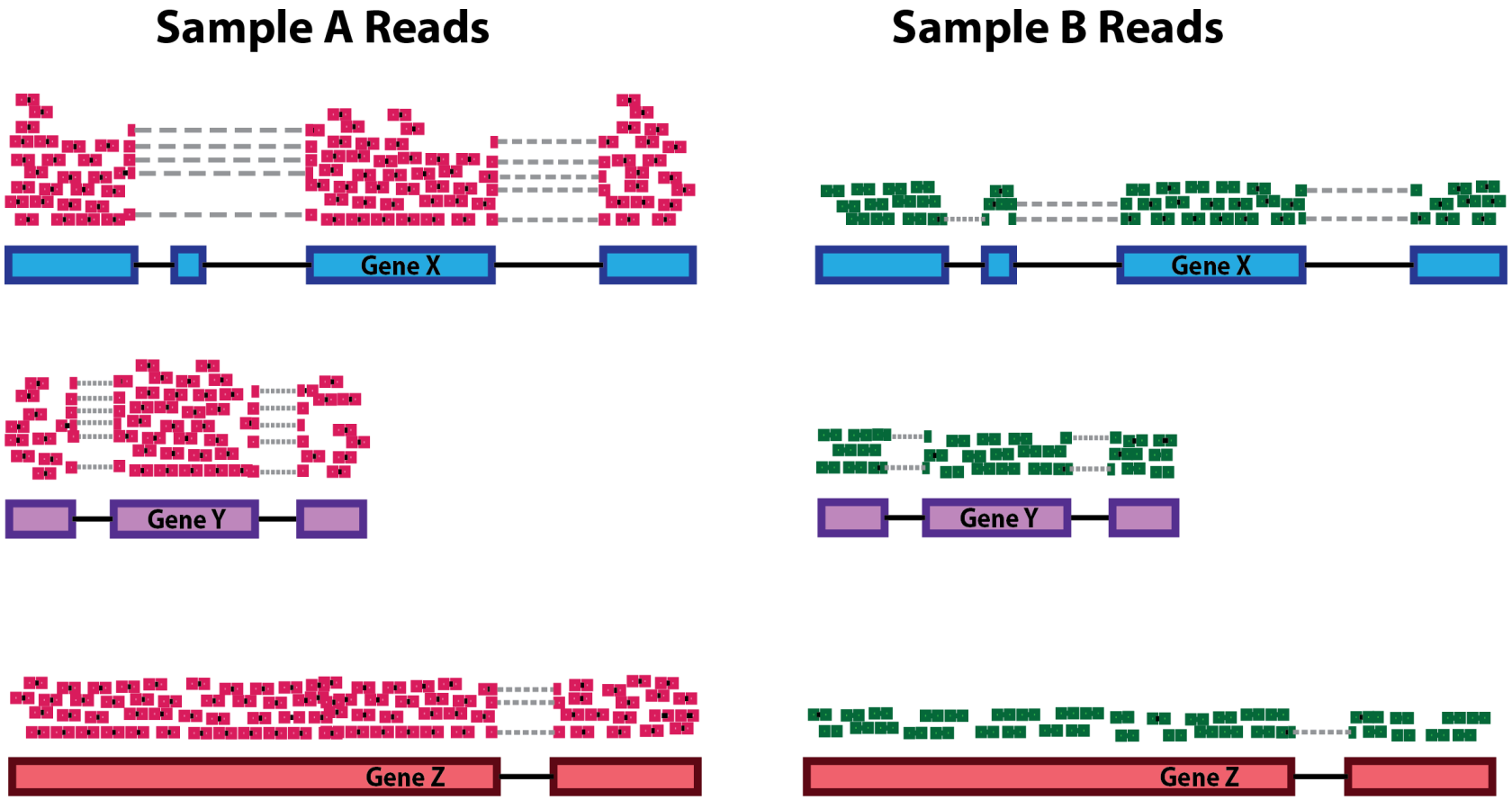
From HBC training

Normalization – Gene length



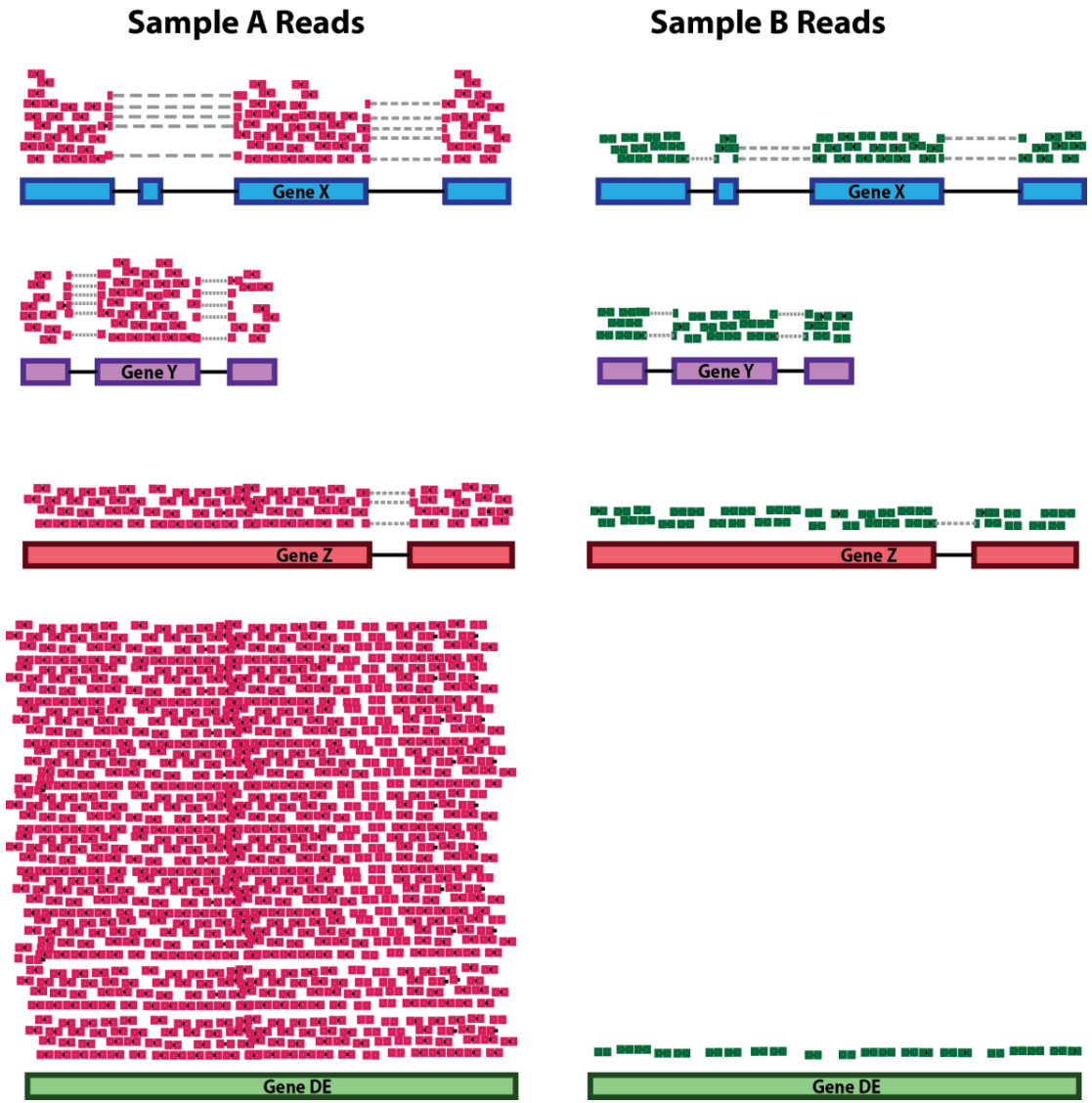
- **Gene length:** Accounting for gene length is necessary for comparing expression between different genes within the same sample.

Normalization – Sequencing depth



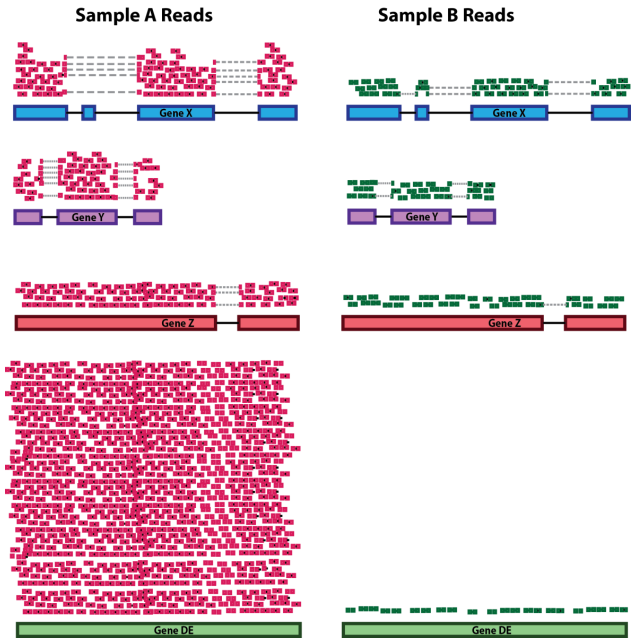
- **Sequencing depth:** Accounting for sequencing depth is necessary for comparison of gene expression between samples.

Highly transcribed genes...

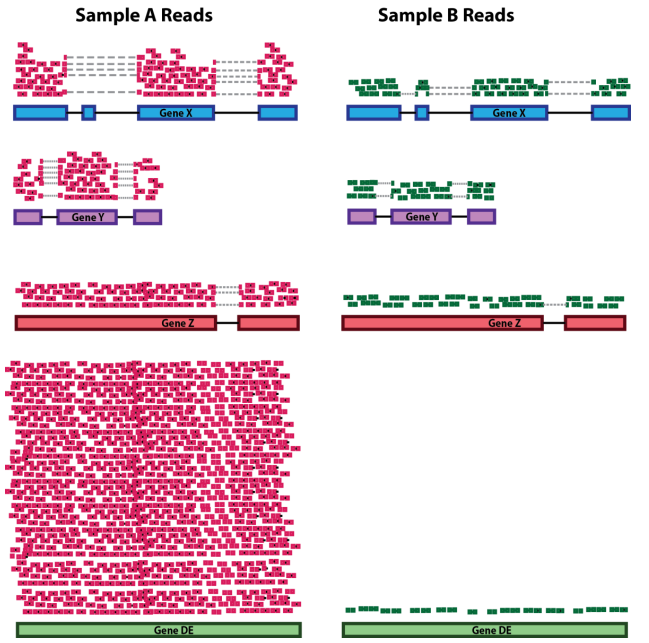


From HBC training

Highly transcribed genes…

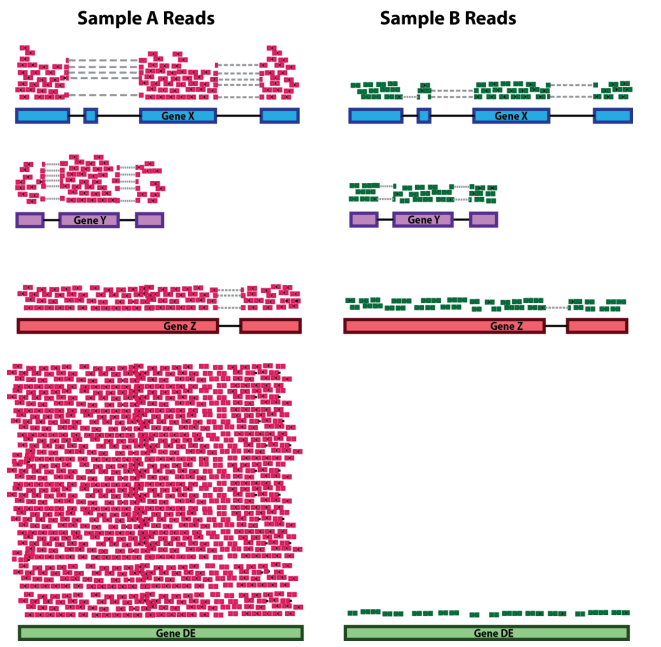


Highly transcribed genes…



	Count sample A	Count sample B	Gene length
Gene X	100	50	100
Gene Y	50	25	50
Gene Z	50	25	100
Gene DE	400	50	100
Total	600	150	

Highly transcribed genes…



	Count sample A	Count sample B	Gene length
Gene X	100	50	100
Gene Y	50	25	50
Gene Z	50	25	100
Gene DE	400	50	100
Total	600	150	

Normalized count sample A	Normalized count sample B	Initial fold change	Fold change normalized
1/600	1/300	2	**½**
1/600	1/300	2	**½**
1/1200	1/600	2	**½**
1/150	1/300	8	**2**

Highly transcribed genes…

	Count sample A	Count sample B	Gene length
Gene X	100	50	100
Gene Y	50	25	50
Gene Z	50	25	100
Gene DE	400	50	100
Total	600	150	

Normalized count sample A	Normalized count sample B	Initial fold change	Fold change normalized
1/600	1/300	2	½
1/600	1/300	2	½
1/1200	1/600	2	½
1/150	1/300	8	**2**

- **RNA composition:** A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods. Accounting for RNA composition is recommended for accurate comparison of expression between samples and is particularly important when performing differential expression analyses.

From HBC training

Common normalization methods

Normalization method	Description	Accounted errors	Recommendations for use
CPM (Count Per Million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same sample group; **NOT for within sample comparisons or DE analysis**
RPKM/FPKM (Reads/Fragments Per Kilobase per million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios**	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)**	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

From HBC training

Why not using RPKM to compare samples

RPKM-normalized counts table

Gene	Sample A	Sample B
XCR1	5.5	5.5
WASHC1	73.4	21.8
…	…	…
Total RPKM-normalized counts	**1,000,000**	**1,500,000**

Why not using RPKM to compare samples

RPKM-normalized counts table

Gene	Sample A	Sample B
XCR1	5.5	5.5
WASHC1	73.4	21.8
…	…	…
Total RPKM-normalized counts	**1,000,000**	**1,500,000**

XRC1 case:

$$\frac{5.5}{1{,}000{,}000} \neq \frac{5.5}{1{,}500{,}000}$$

- Using RPKM/FPKM normalization, the total number of RPKM/FPKM **normalized counts for each sample will be different**. Therefore, you **cannot compare the normalized counts** for each gene equally between samples.

DESeq2 normalization principle – Step 1

- **Step 1:** Create a pseudo-reference sample (row-wise geometric mean). $\bar{c} = \sqrt[n]{c_1 x c_2 x \ldots x c_n}$

Gene	Sample A	Sample B	Pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = 1161.5
ABCD1	22	13	sqrt(22 * 13) = 16.9
MEFV	793	410	570.2
BAG1	76	42	56.5
MOV10	521	1196	883.7

DESeq2 normalization principle – Step 2

- **Step 2:** Calculate ratio of each samples to the pseudo reference.

Gene	Sample A	Sample B	Pseudo reference sample	Ratio of sample A / ref	Ratio of sample B / ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**

From HBC training

DESeq2 normalization principle – Step 2

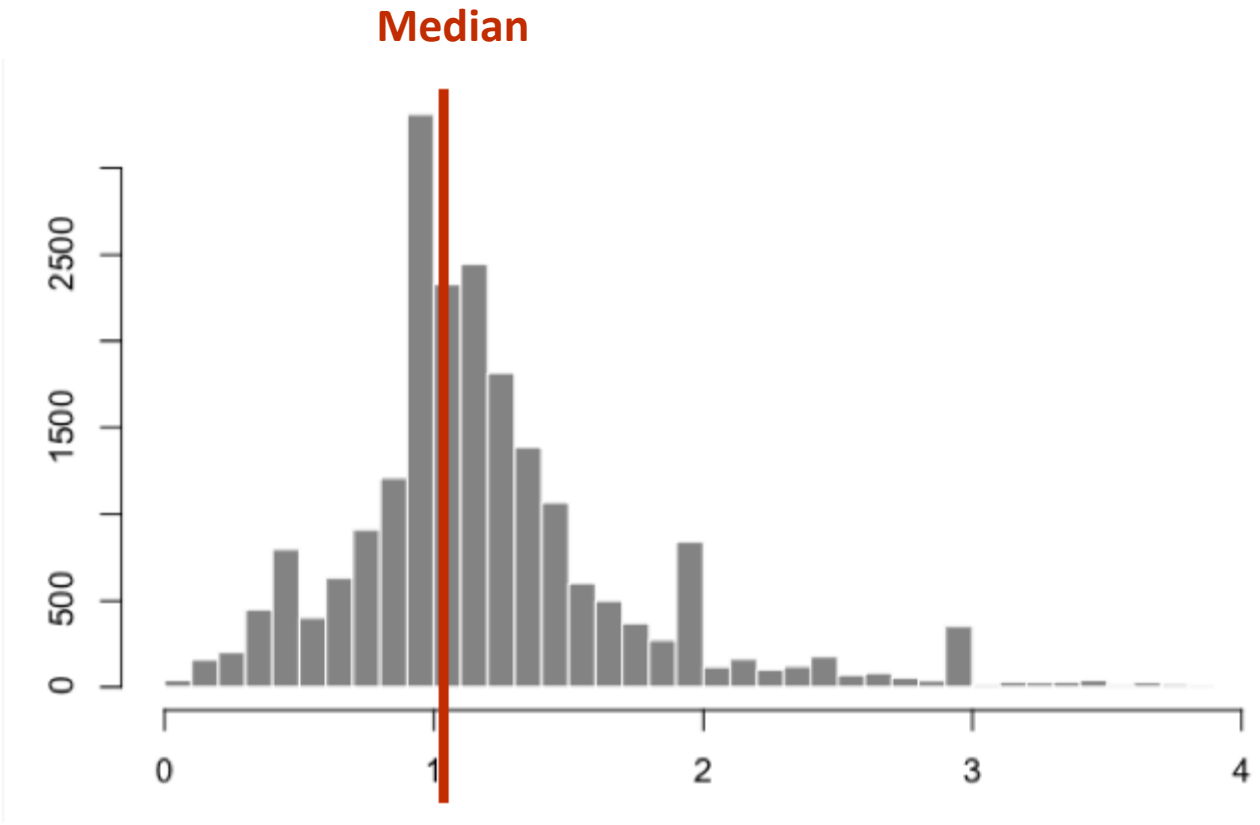
- **Step 3:** Calculate the normalization factor for each factor.

Ratio of sample A / ref	Ratio of sample B / ref
1.28	0.78
1.30	0.77
1.39	0.72
1.35	0.74
0.590	1.35

DESeq2 normalization principle – Step 2

- **Step 3:** Calculate the normalization factor for each factor.

Ratio of sample A / ref	Ratio of sample B / ref
1.28	0.78
1.30	0.77
1.39	0.72
1.35	0.74
0.590	1.35
1,3	**0,77**

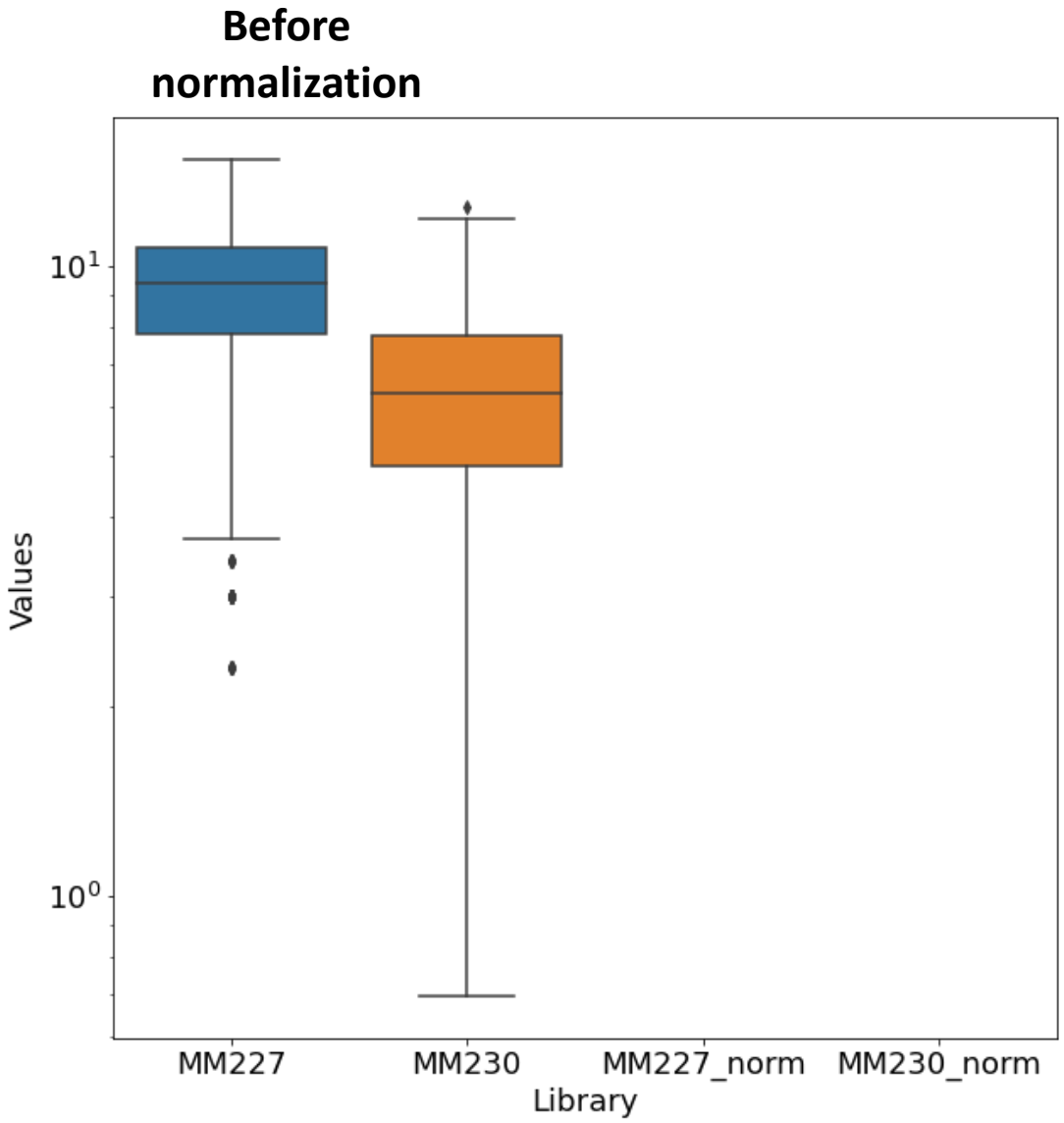


DESeq2 normalization principle – Step 4

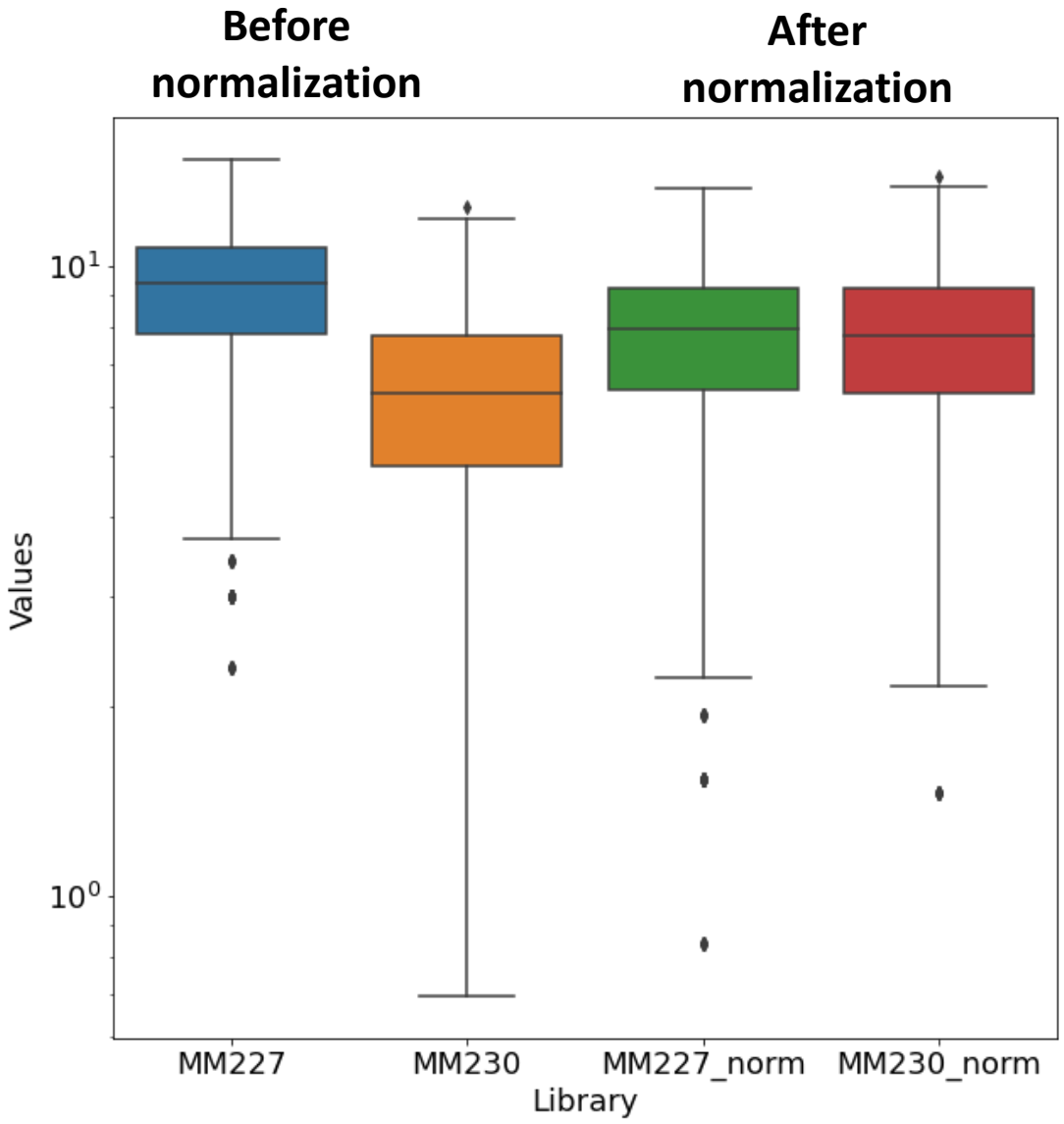
- **Step 4:** Calculate the normalized count values using the normalization factor.

Gene	Sample A	Sample B	Normalized sample A	Normalized sample B
EF2A	1489	906	1489/1.3 = **1145.39**	906/0.77 = **1176.62**
ABCD1	22	13	22/1.3 = **16.92**	13/0.77 = **16.88**
MEFV	793	410	793/1.3 = **610**	410/0.77 = **532.47**
BAG1	76	42	76/1.3 = **58.46**	42/0.77 = **54.54**
MOV10	521	1196	521/1.3 = **400.77**	1196/0.77 = **1553.24**

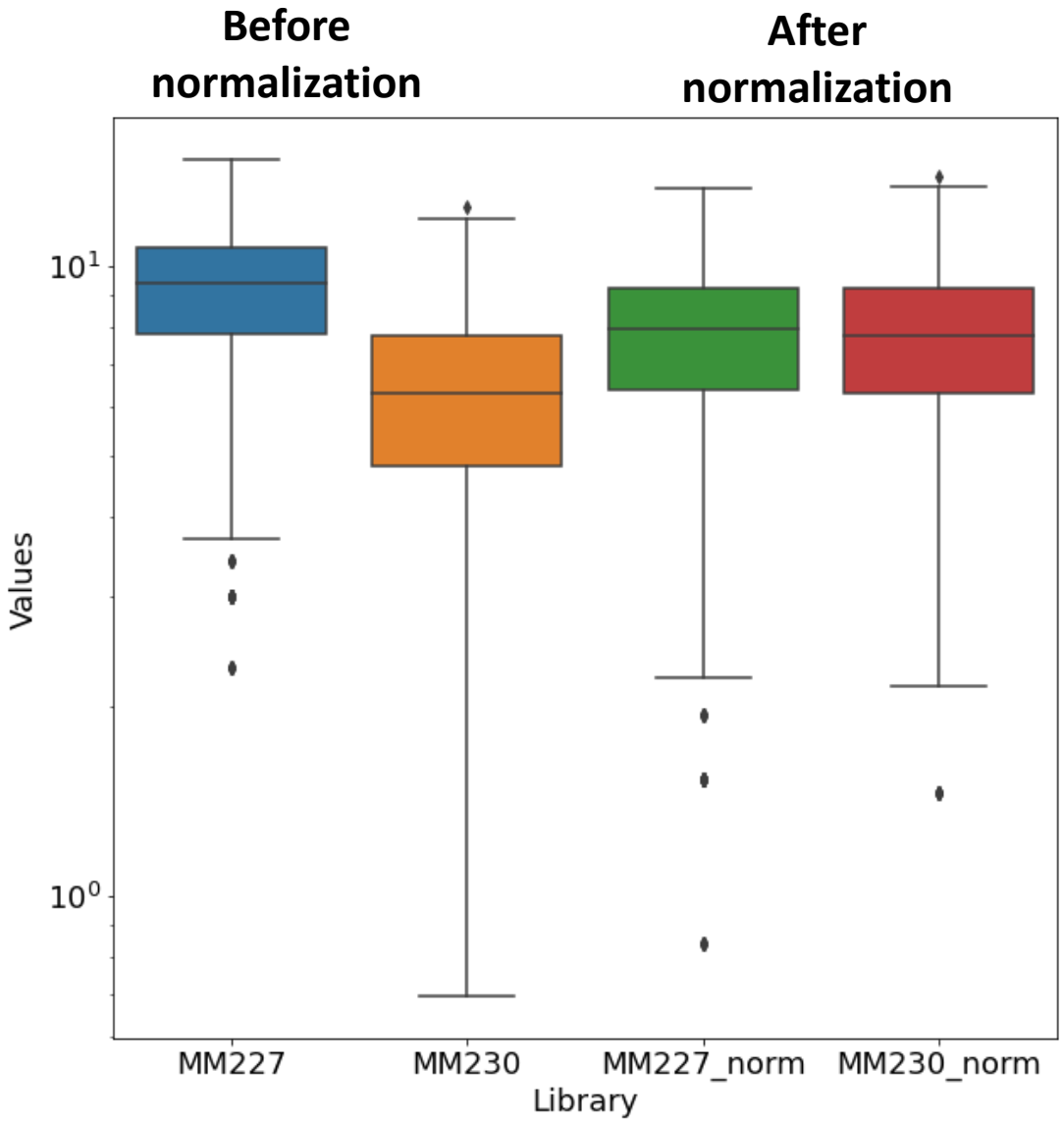
Impact of the normalization



Impact of the normalization



Impact of the normalization



The DESeq2 normalization yields to two distributions with the **same mean and same variance.**

One crucial hypothesis

One crucial hypothesis

- To be able to normalize using a pseudo-reference, we start with the hypothesis all the genes have the same expression in all the conditions… to search for differentially expressed genes…

One crucial hypothesis

- To be able to normalize using a pseudo-reference, we start with the hypothesis all the genes have the same expression in all the conditions… to search for differentially expressed genes…

Why it works ?

One crucial hypothesis

- To be able to normalize using a pseudo-reference, we start with the hypothesis all the genes have the same expression in all the conditions… to search for differentially expressed genes…

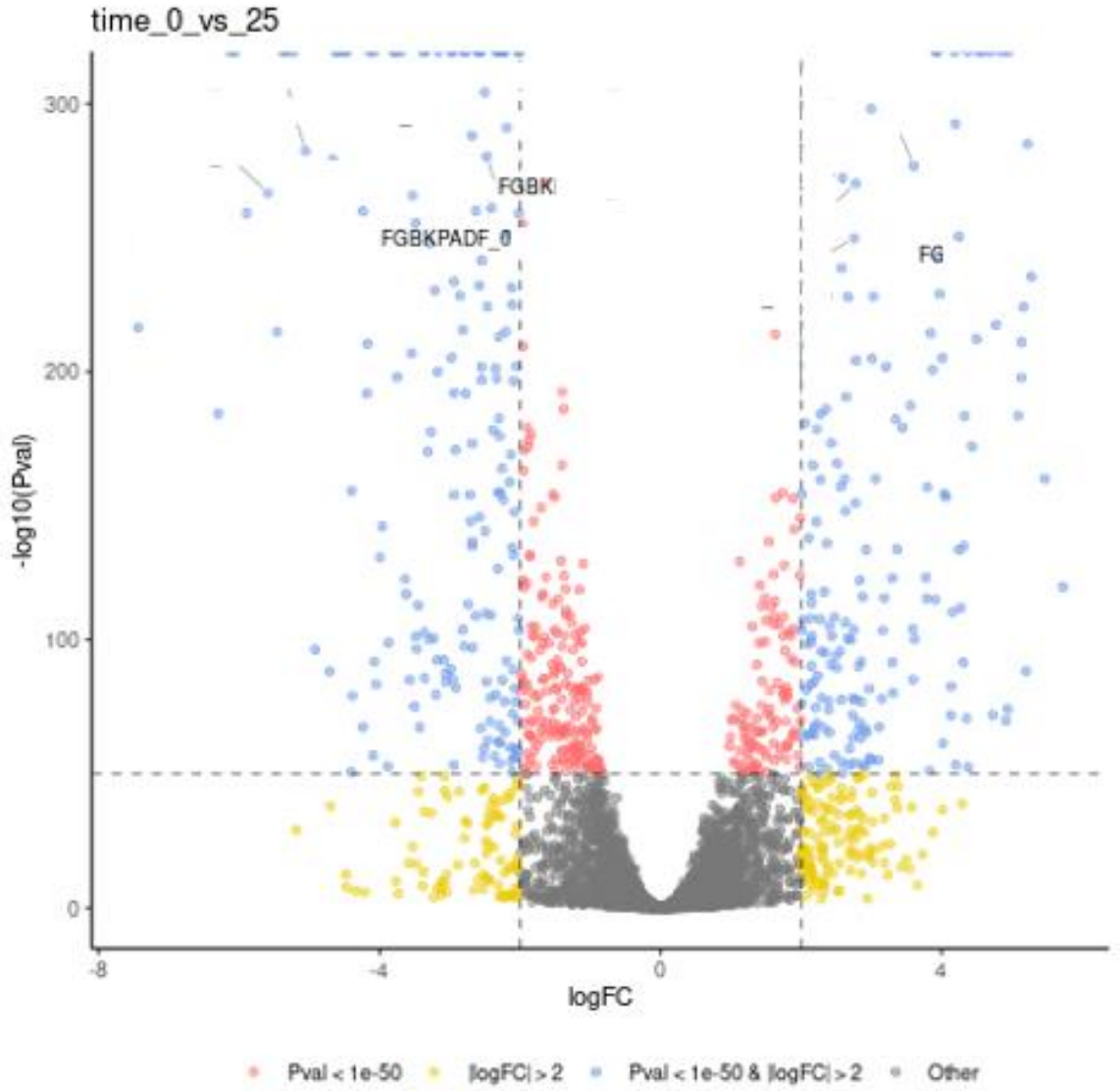
Why it works ?

- Most of genes don't move usually during a gene expression analysis (few percent). However be careful that normalization won't work if you disturb the system to the extent where all the expression is disturbed (inhibition of the RNA polymerase).

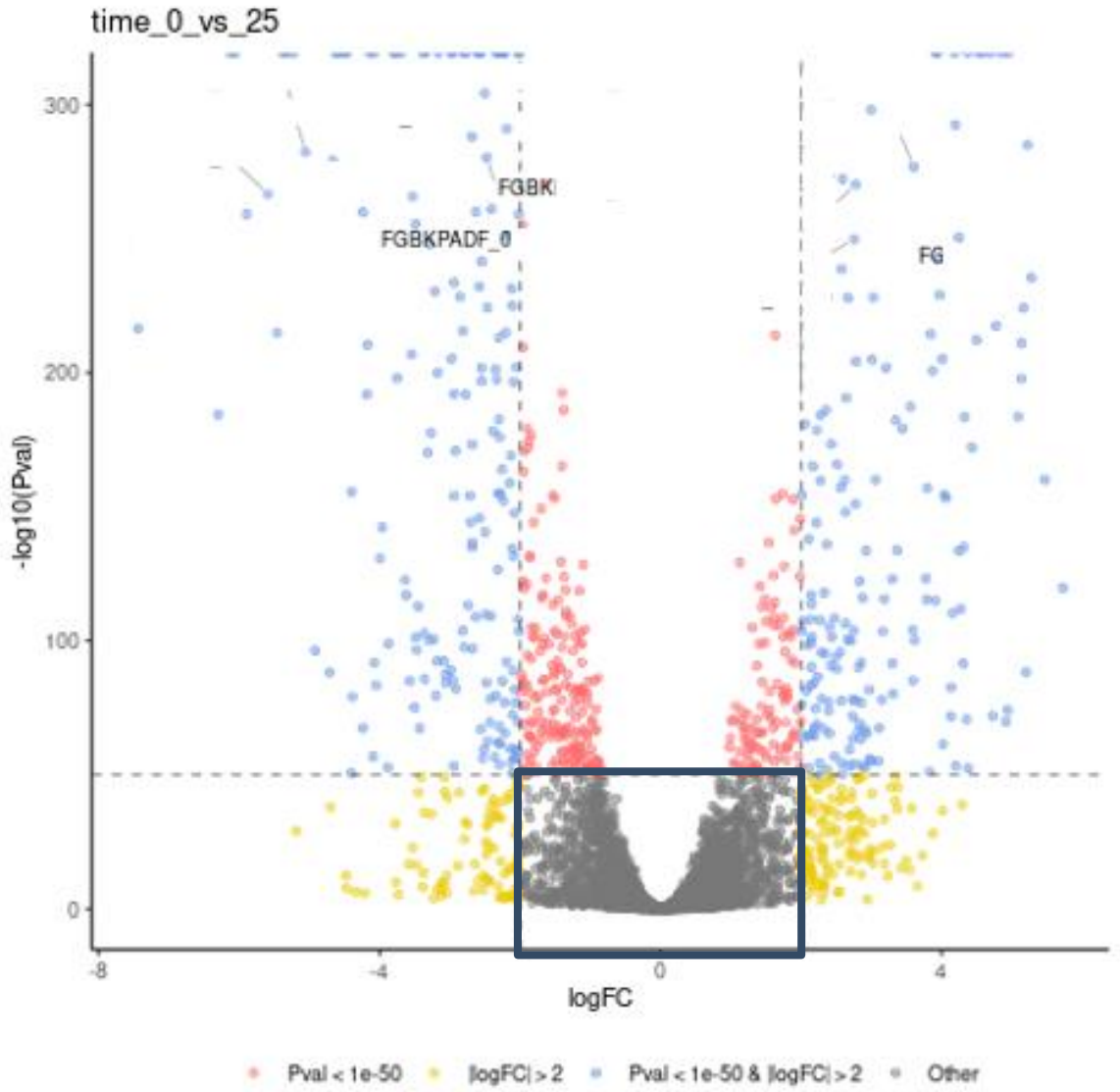
Log2 fold change

Gene	Normalized sample A	Normalized sample B	Fold change (B / A)	Log2 fold change	p-value
EF2A	1489/1.3 = **1145.39**	906/0.77 = **1176.62**	1.02	0.03	0.67
ABCD1	22/1.3 = **16.92**	13/0.77 = **16.88**	0.99	-0.003	0.99
MEFV	793/1.3 = **610**	410/0.77 = **532.47**	0.87	-0.2	$9x10^{-4}$
BAG1	76/1.3 = **58.46**	42/0.77 = **54.54**	0.93	-0.1	$4x10^{-3}$
MOV10	521/1.3 = **400.77**	1196/0.77 = **1553.24**	**3.88**	**1.96**	$2x10^{-75}$

Volcano plot

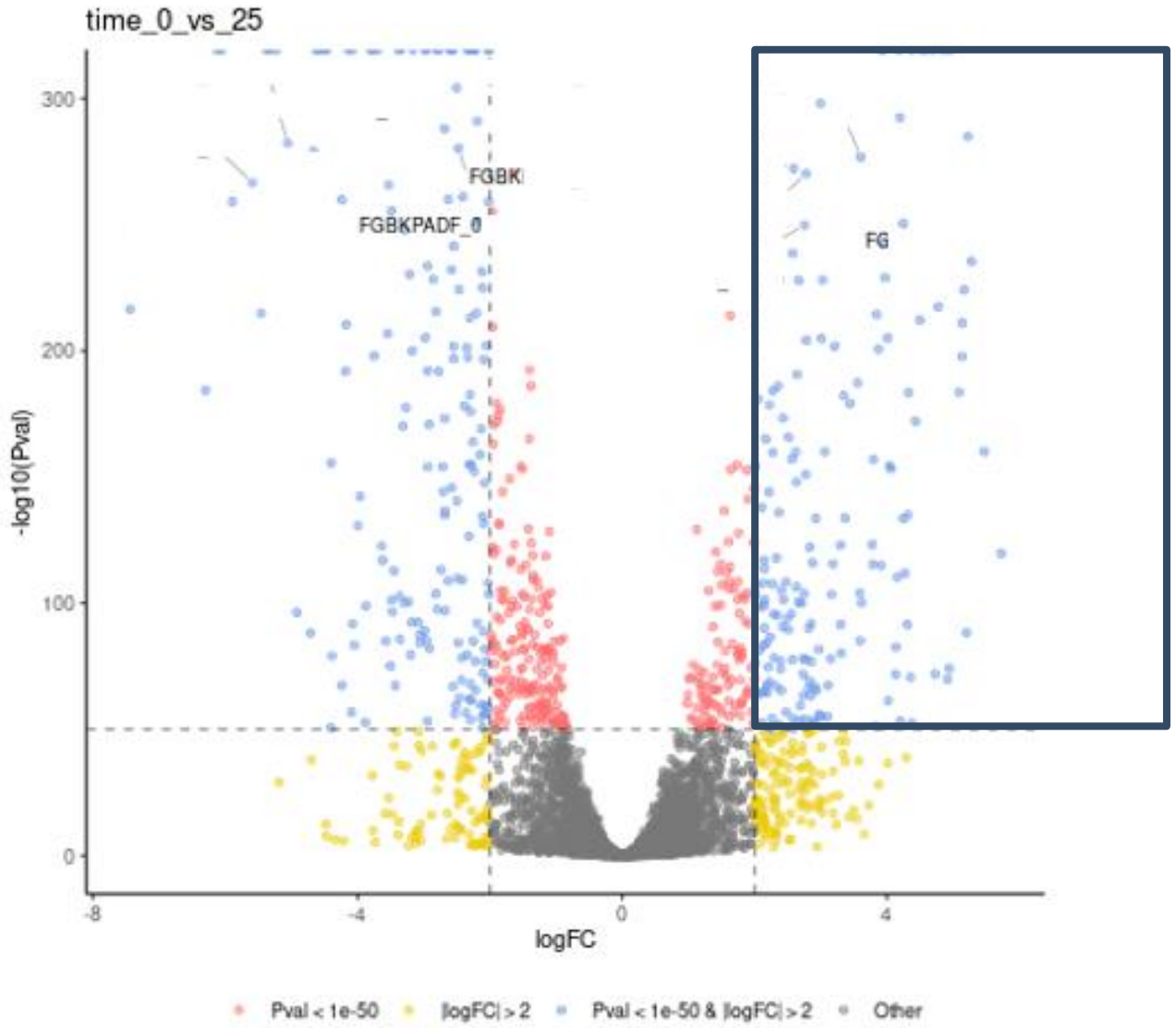


Volcano plot



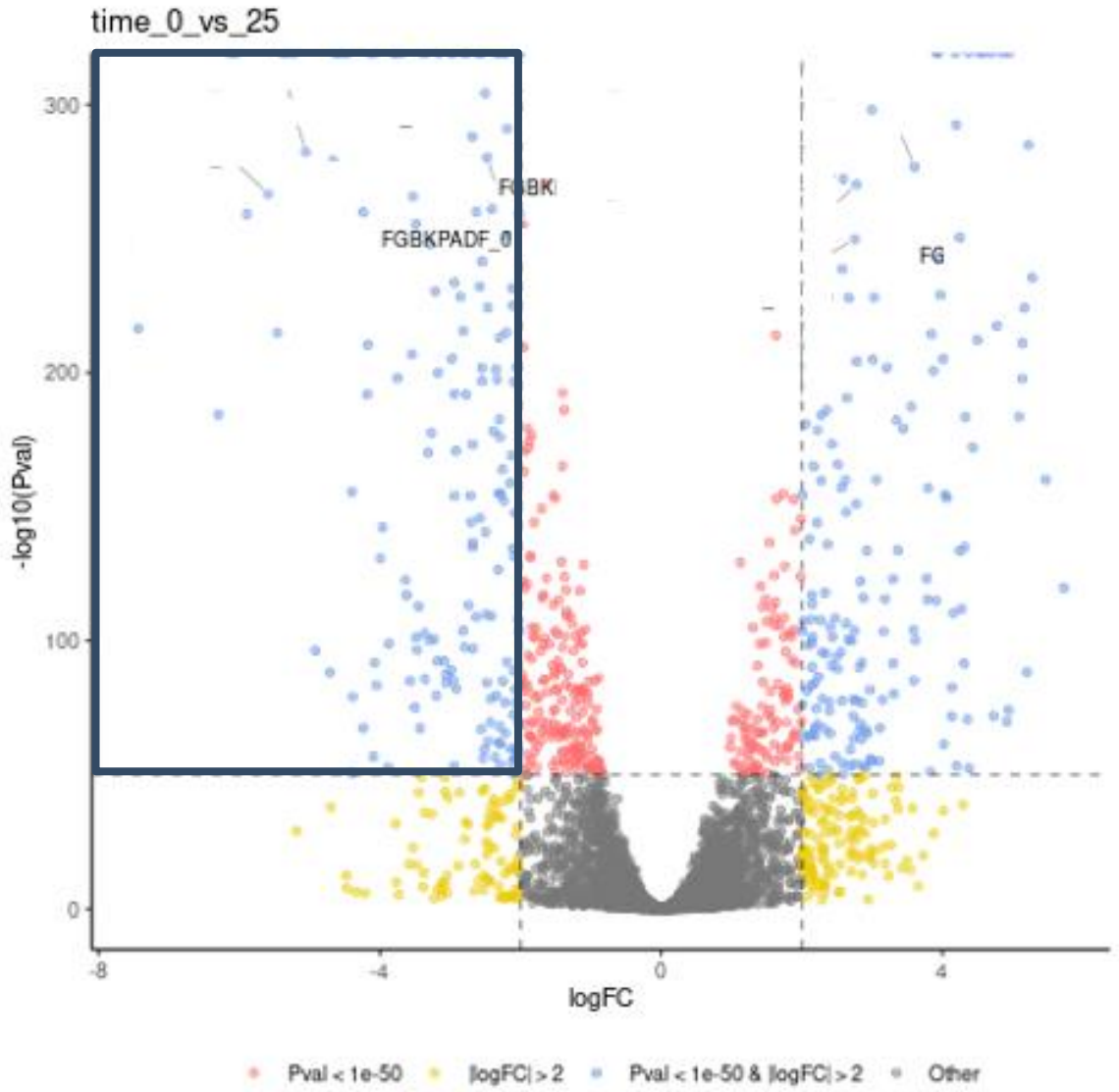
- Genes not differentially expressed.

Volcano plot



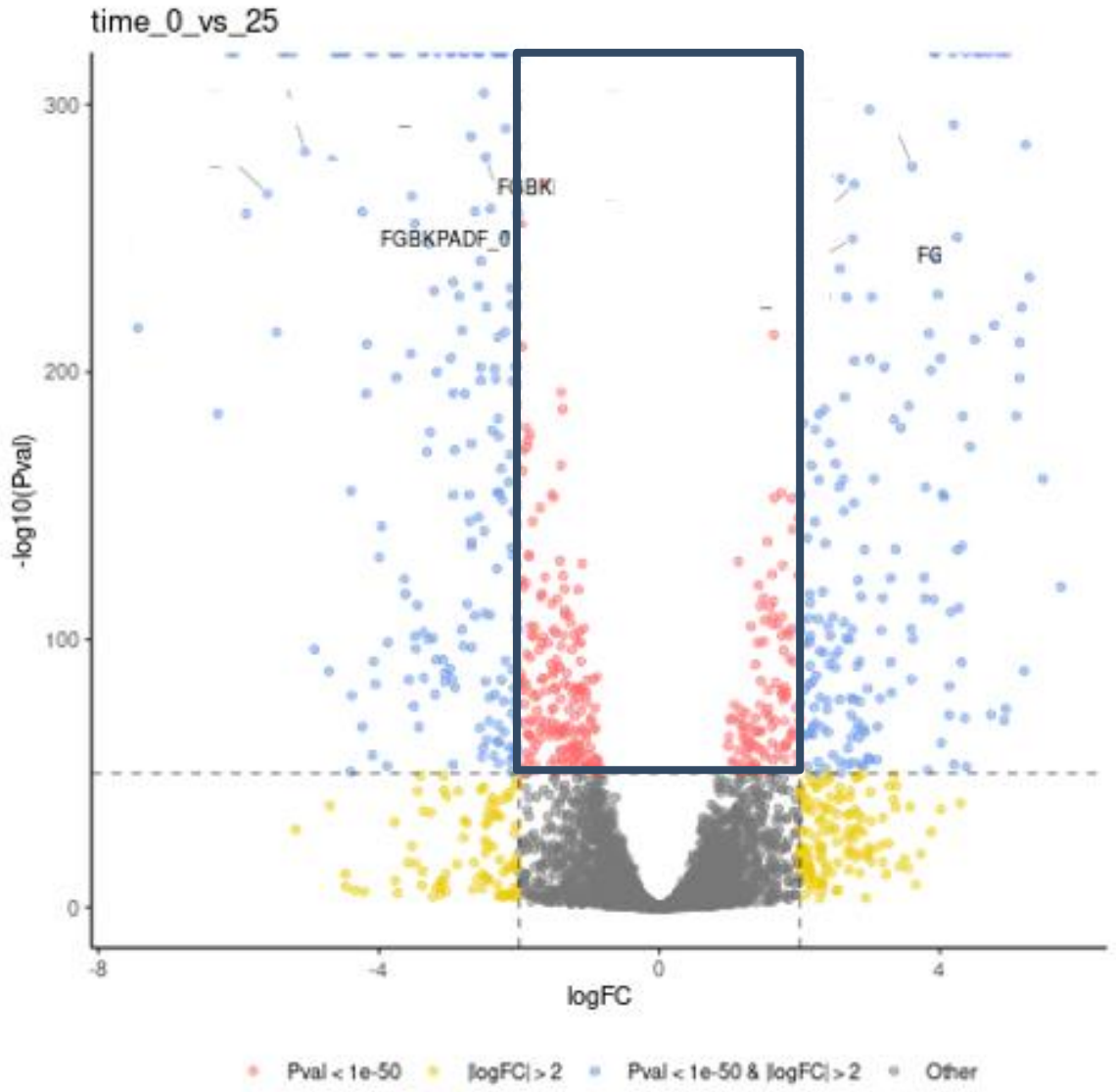
- Genes not differentially expressed.
- Genes up-regulated.

Volcano plot



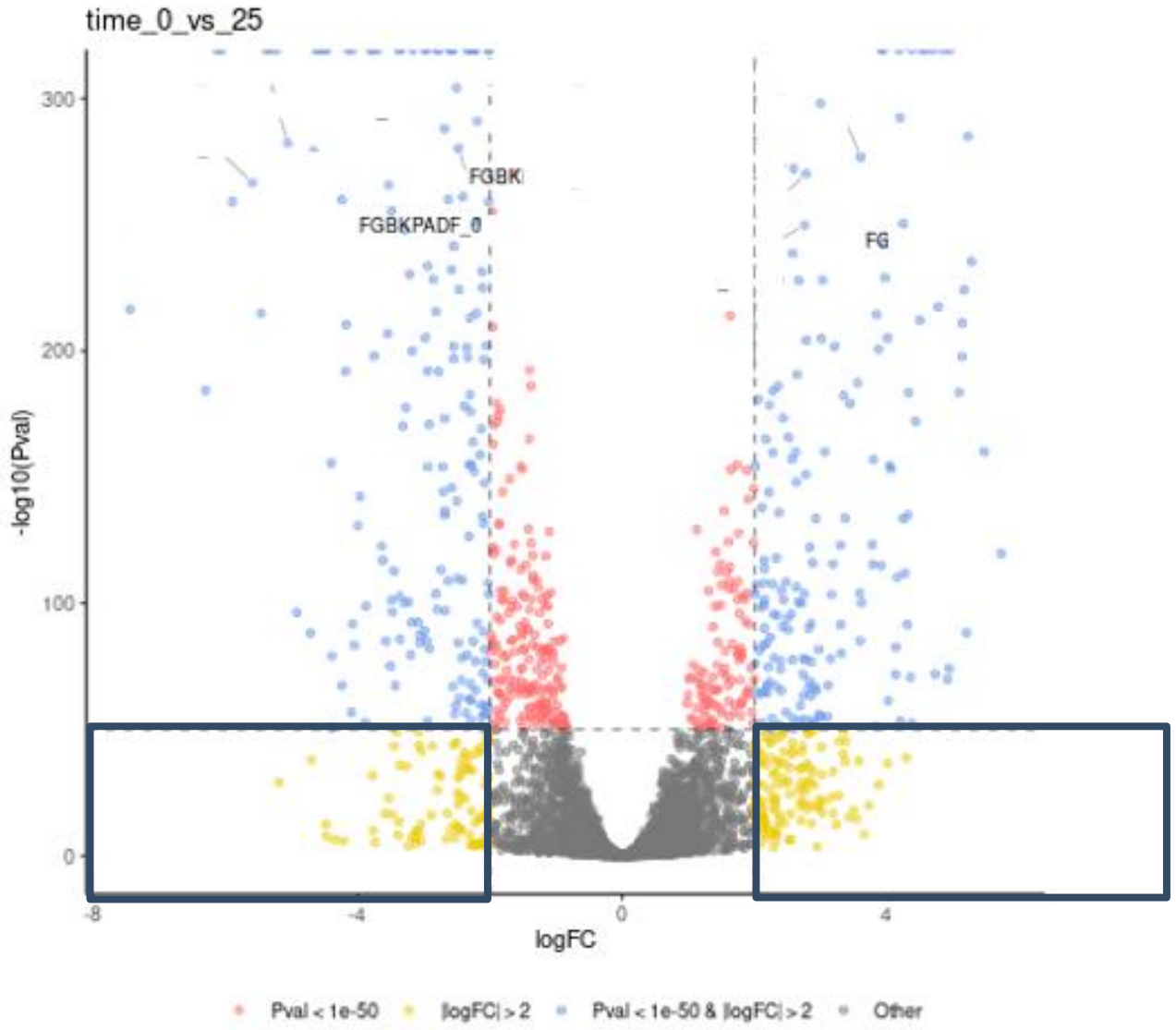
- Genes not differentially expressed.
- Genes up-regulated.
- Genes down-regulated.

Volcano plot



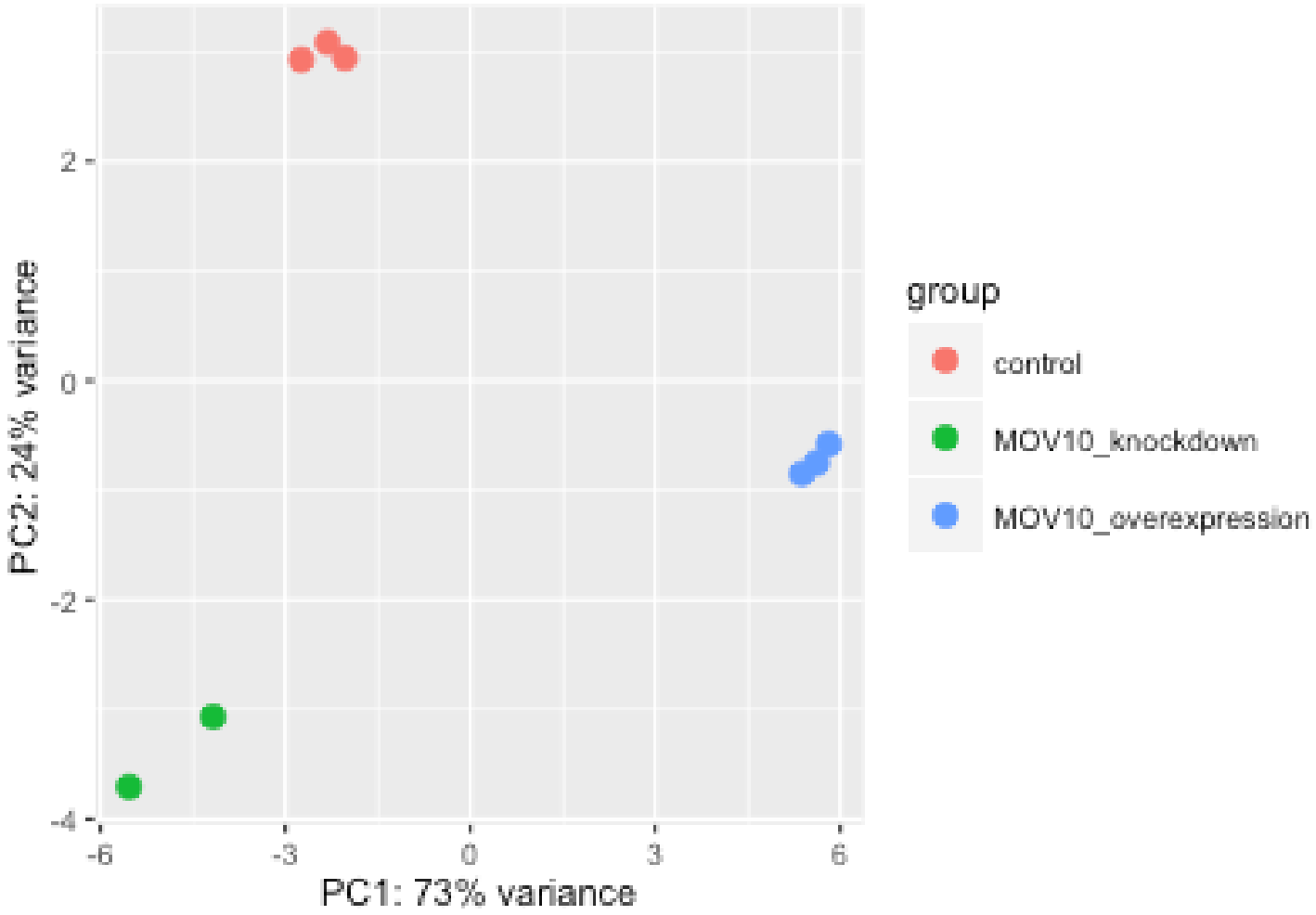
- Genes not differentially expressed.
- Genes up-regulated.
- Genes down-regulated.
- Genes with a low fold change but significant.

Volcano plot



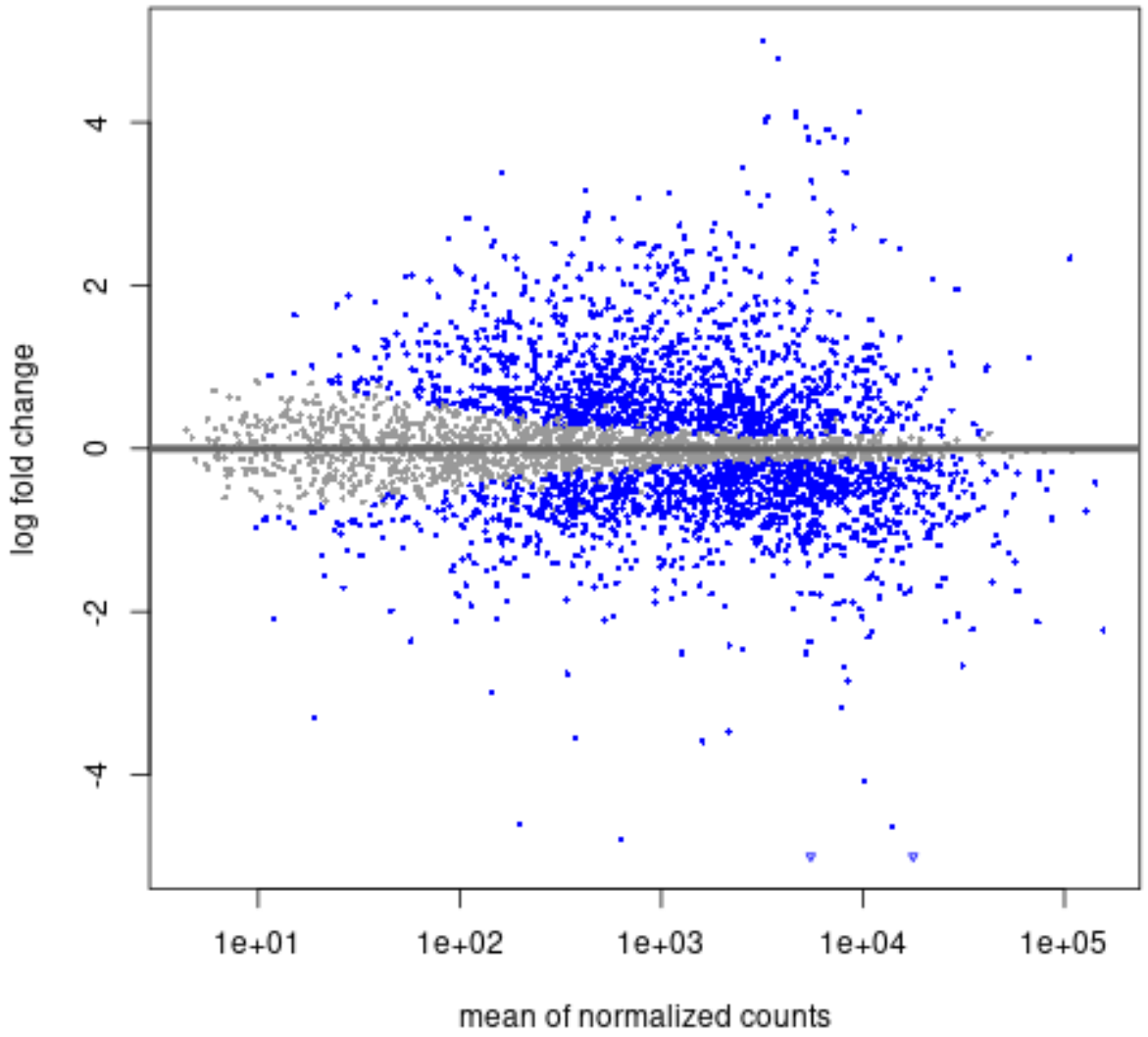
- Genes not differentially expressed.
- Genes up-regulated.
- Genes down-regulated.
- Genes with a low fold change but significant.
- Genes up or down regulated but not significant.

PCA plot – checking replicates quality



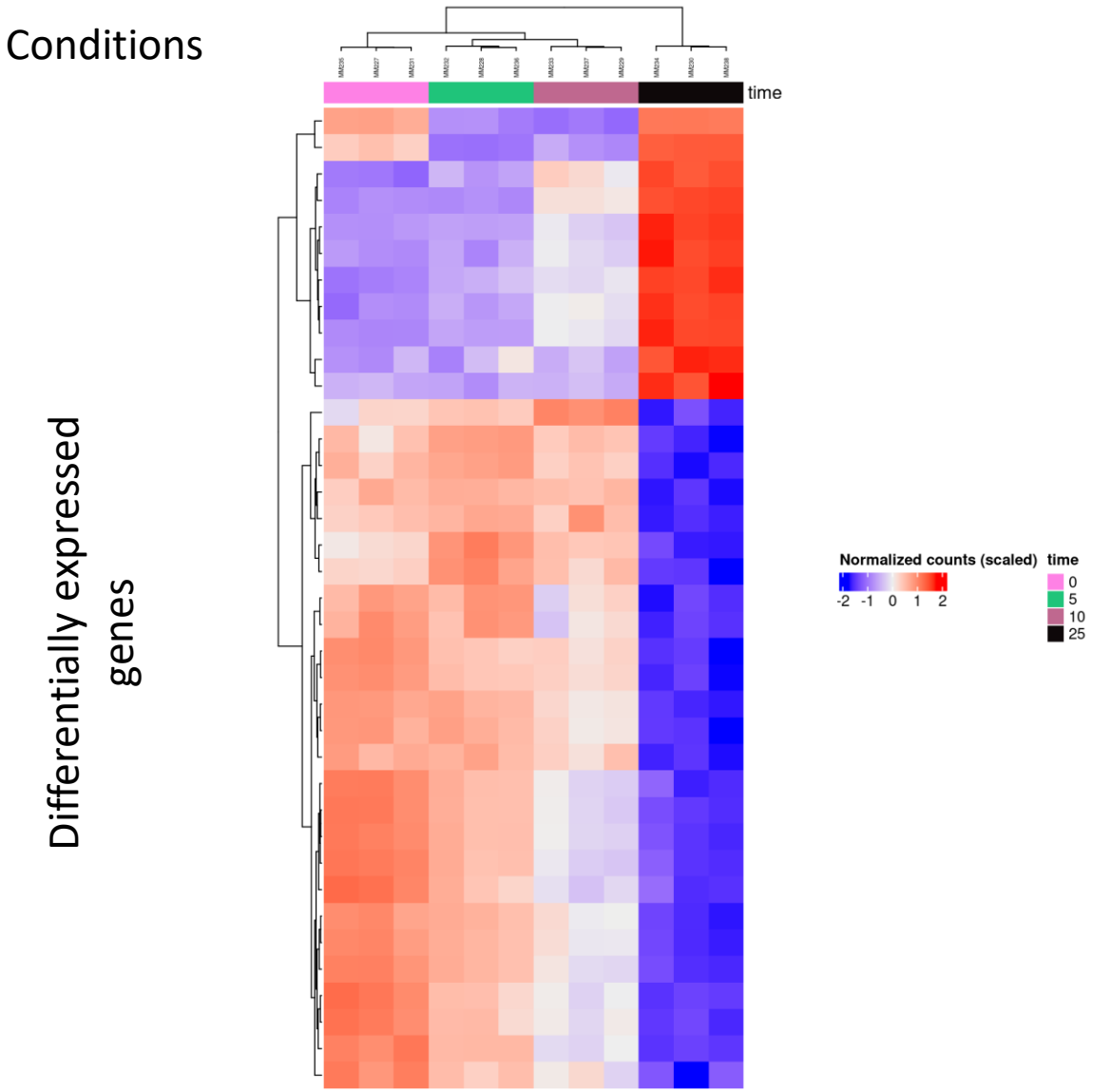
- We reduce based on variance. To build new spaced with fewer dimension explaining most of the variation.
- We expect replicates to group together and conditions to be far away.

MA plot – checking replicates quality

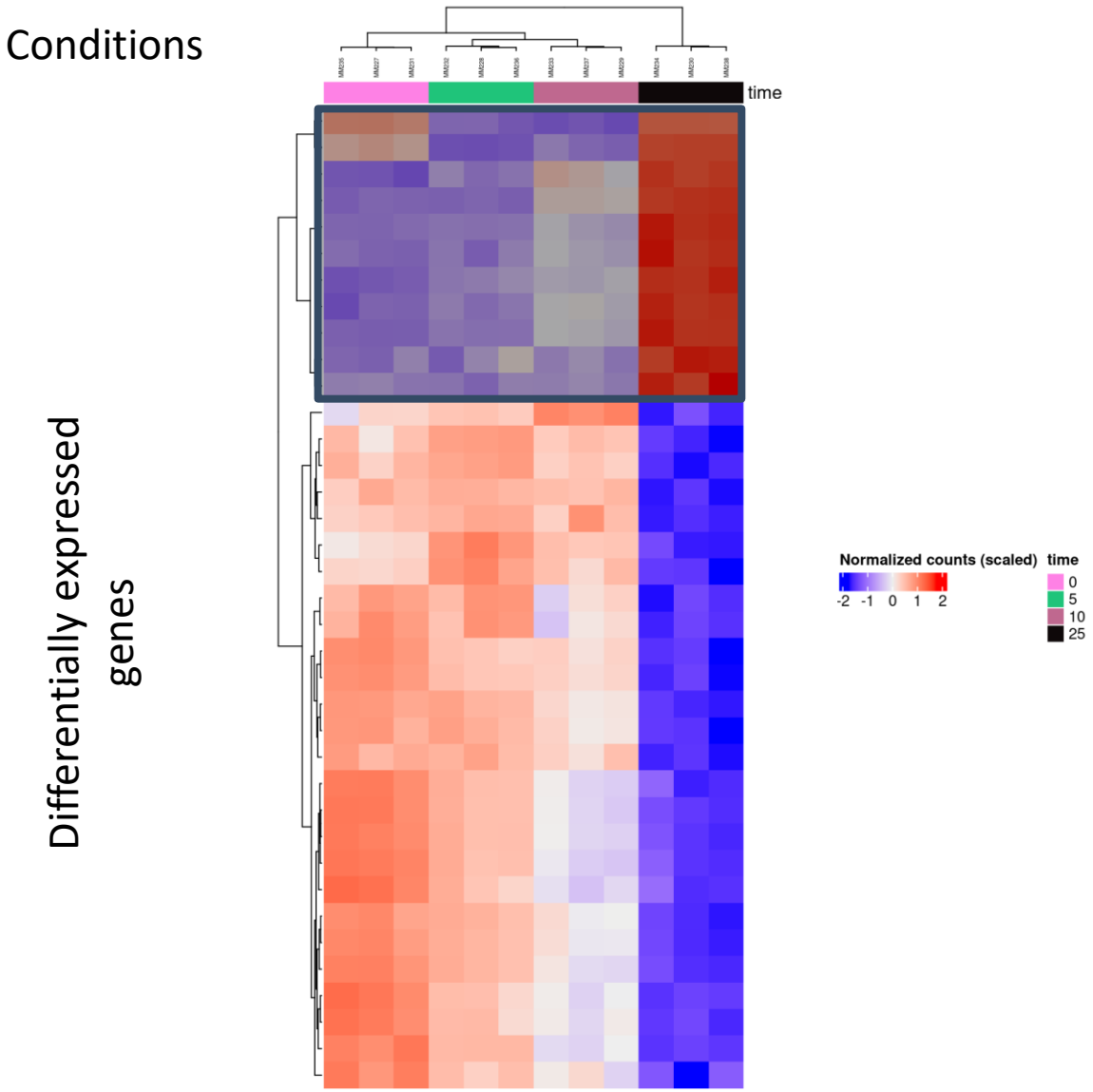


- Grey dots are unsignificant fold change based on the replicates and conditions. It's due to either noise on these genes, or due to significant changes between two replicates.
- The p-value should be more significant if there are high count genes. Indeed, the noise will be lower compared to the biological signal.
- A good signal is a signal where there is no mix between blue and grey signal.

Heatmap plot

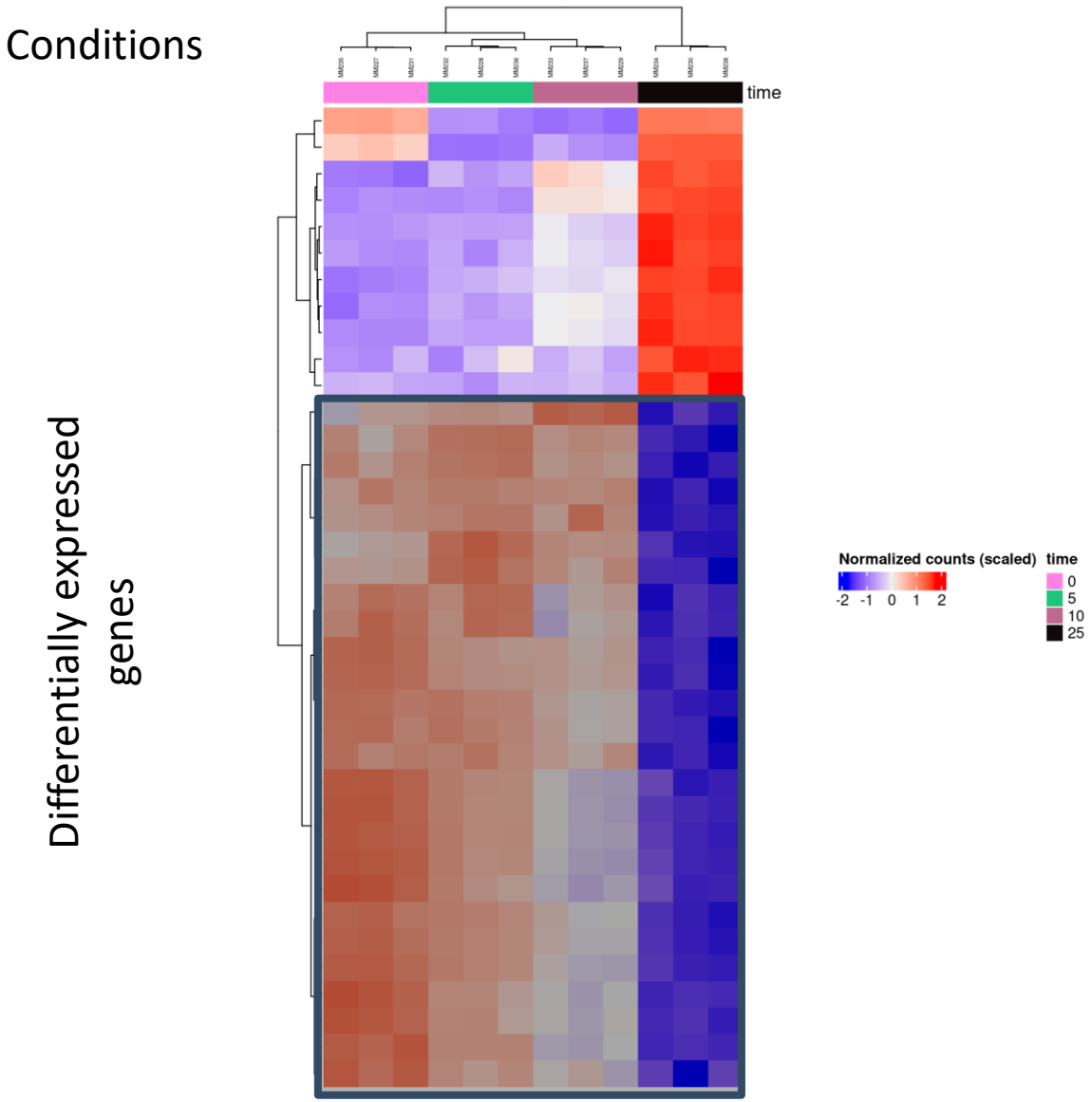


Heatmap plot



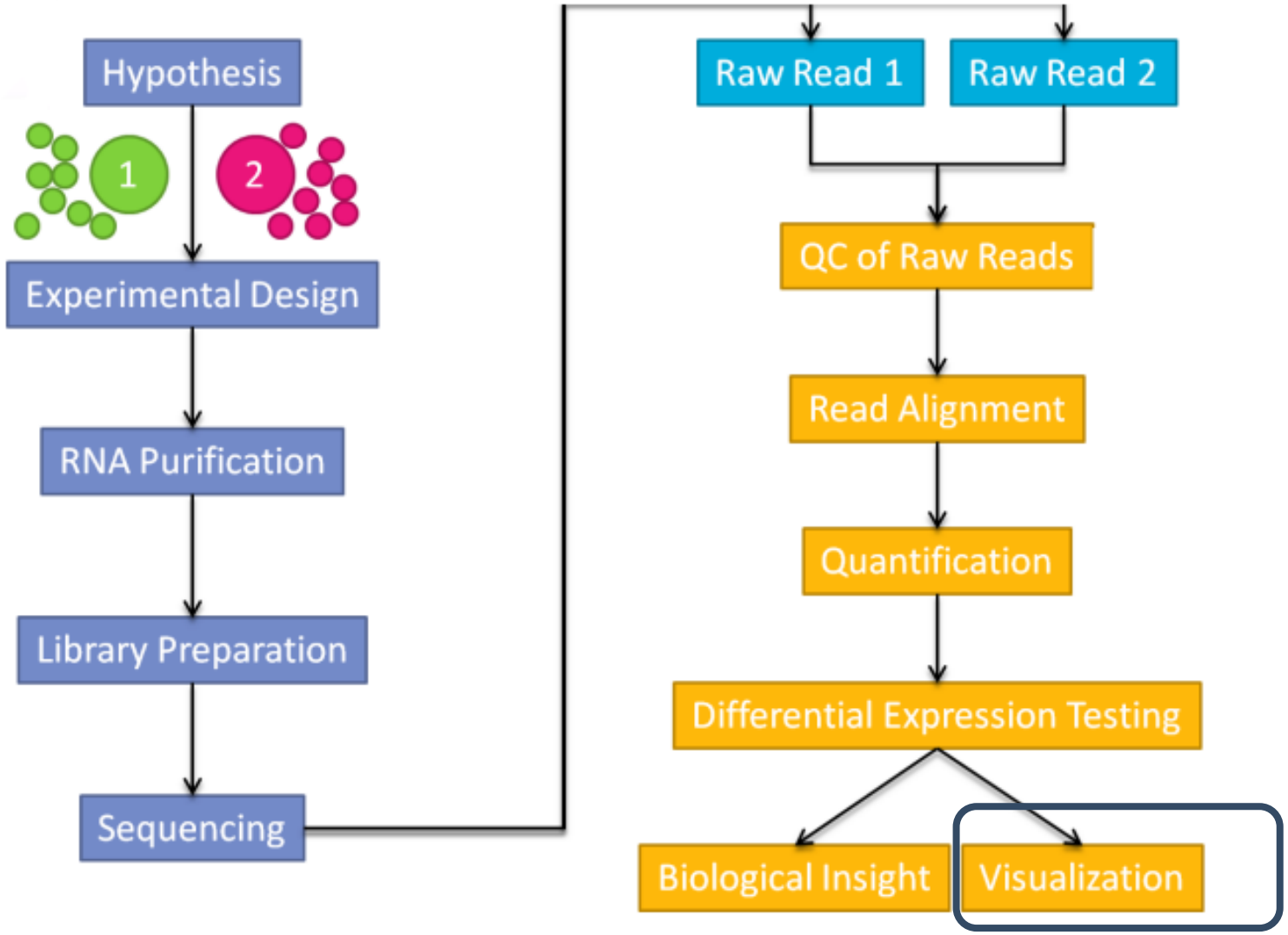
- Gene up-regulated.

Heatmap plot



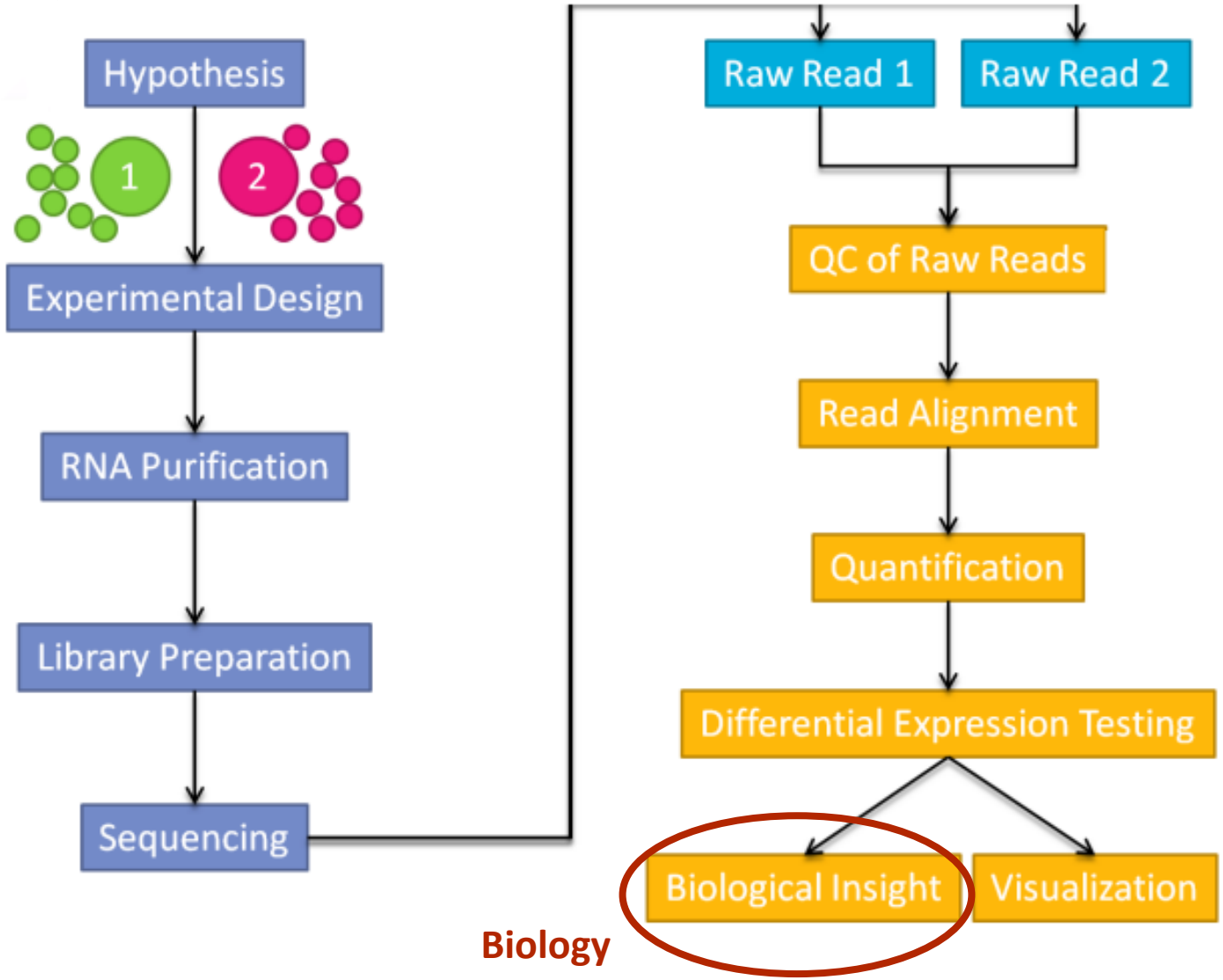
- Gene up-regulated.
- Genes down-regulated.

Where biology came back

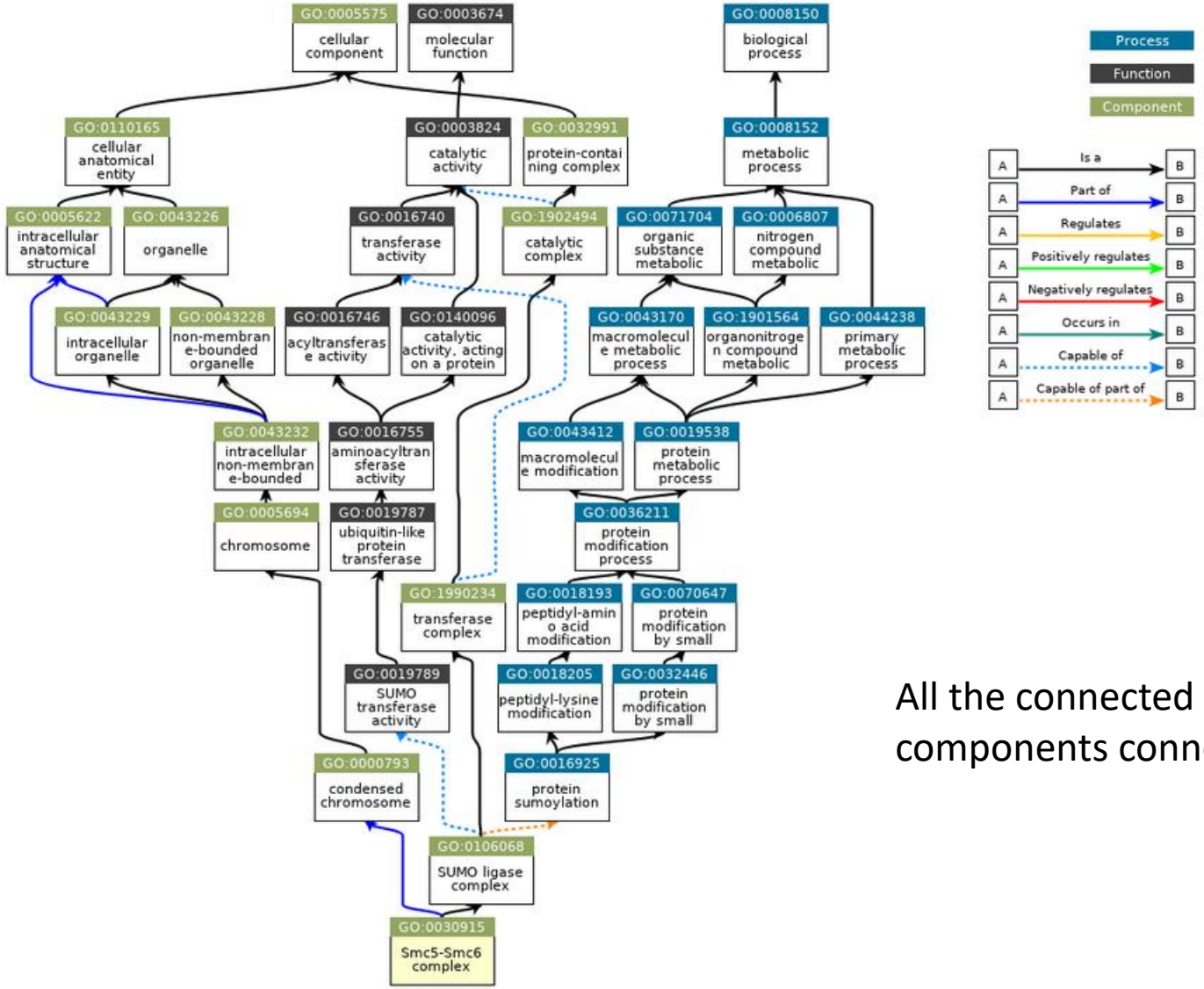


Bioinformatics

Where biology came back



Gene ontology

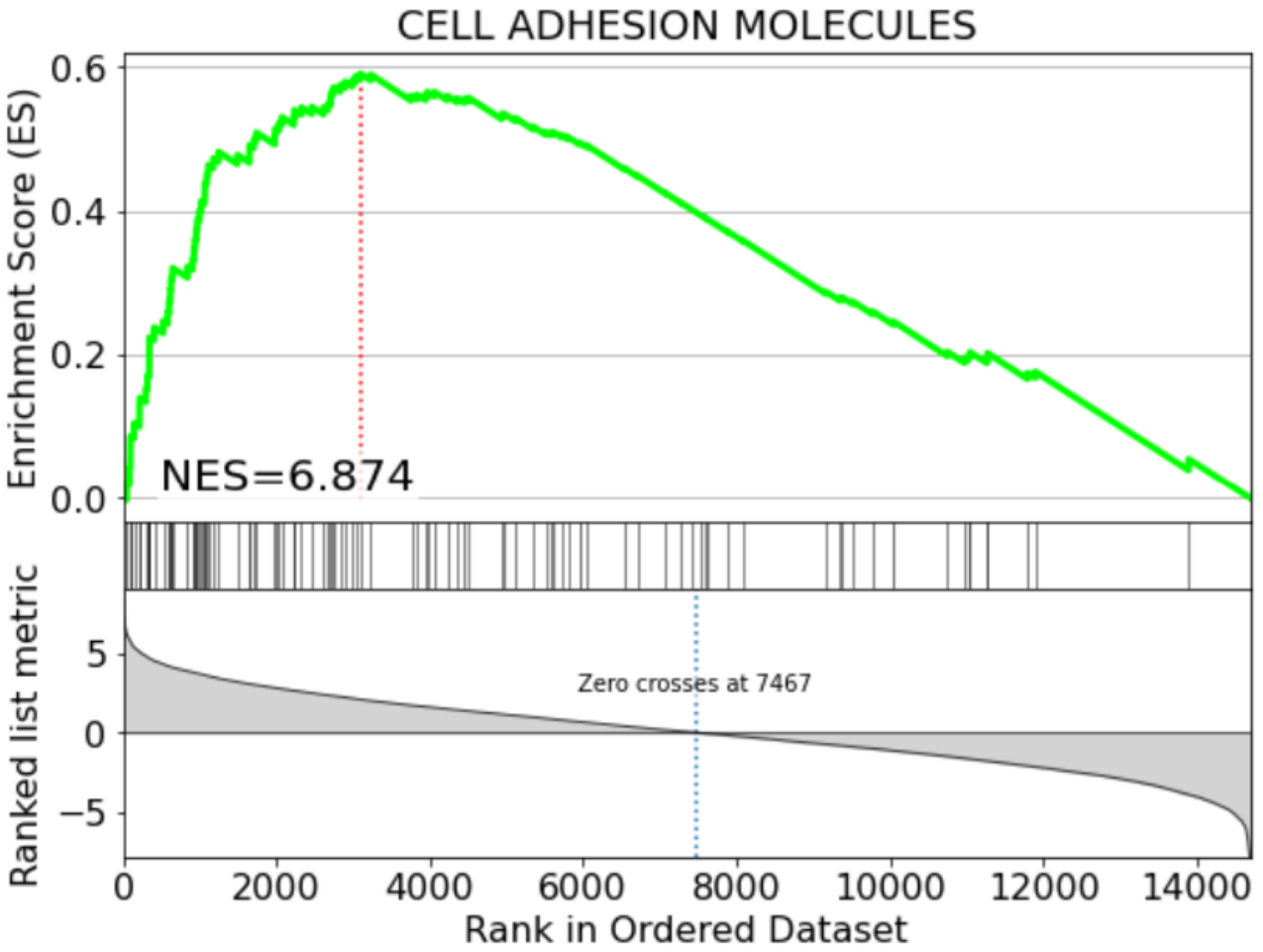


All the connected functions, processes and components connected to one protein

Gene set enrichment analysis

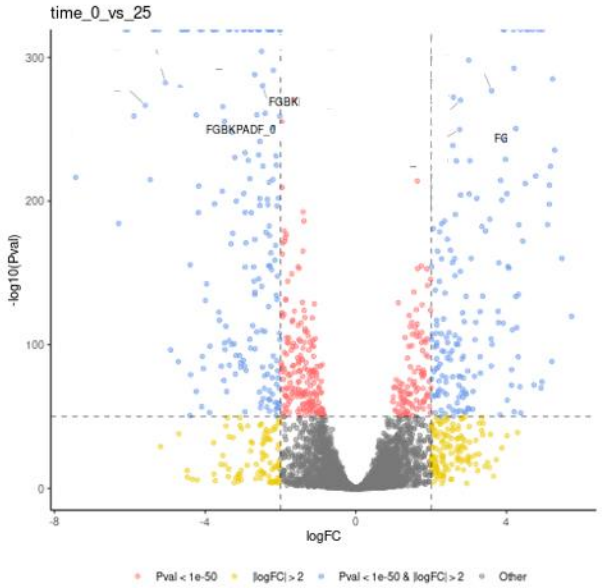
GSEA

NES	SET
6.874	CELL ADHESION MOLECULES
-6.047	PROTEIN PROCESSING IN ENDOPLASMIC RETICULUM
6.039	ECM-RECEPTOR INTERACTION
5.300	CALCIUM SIGNALING PATHWAY
5.297	STAPHYLOCOCCUS AUREUS INFECTION
5.189	PROTEIN DIGESTION AND ABSORPTION
-5.086	SPINOCEREBELLAR ATAXIA
4.876	COMPLEMENT AND COAGULATION CASCADES
-4.787	RIBOSOME BIOGENESIS IN EUKARYOTES
-4.690	UBIQUITIN MEDIATED PROTEOLYSIS
-4.674	AMYOTROPHIC LATERAL SCLEROSIS
-4.647	PROTEASOME
4.619	SYSTEMIC LUPUS ERYTHEMATOSUS
4.584	NEUROACTIVE LIGAND-RECEPTOR INTERACTION
4.512	FOCAL ADHESION

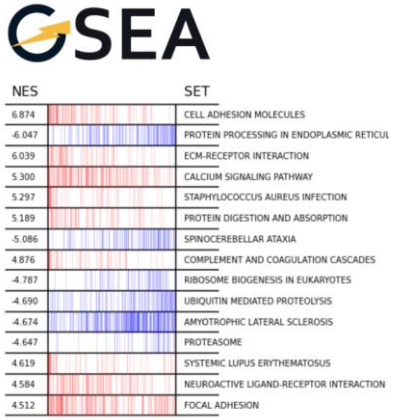


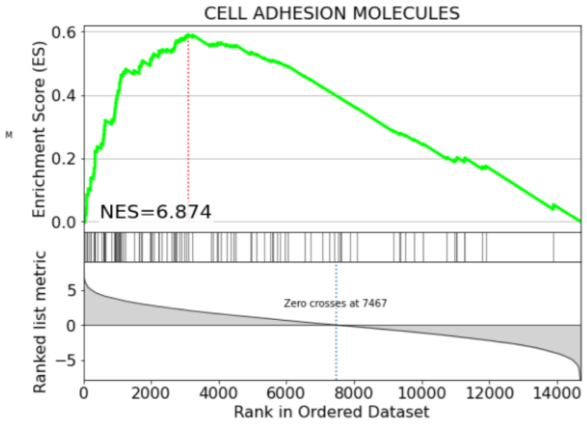
Choosing your genes for mutation

A lot of differentially expressed genes

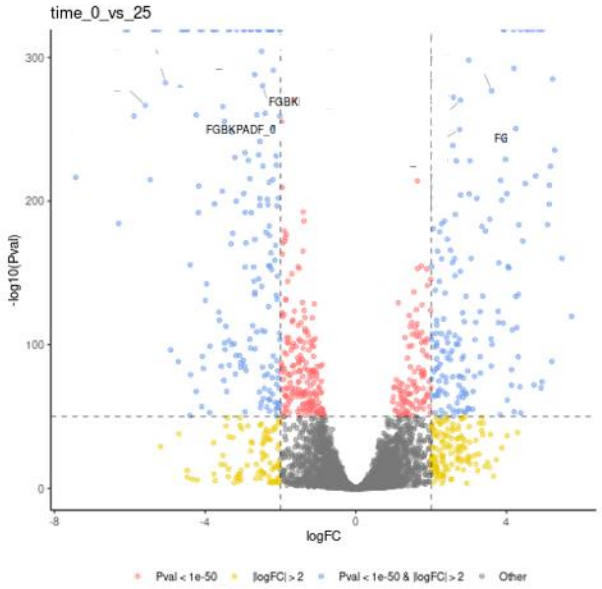


Choosing your genes for mutation



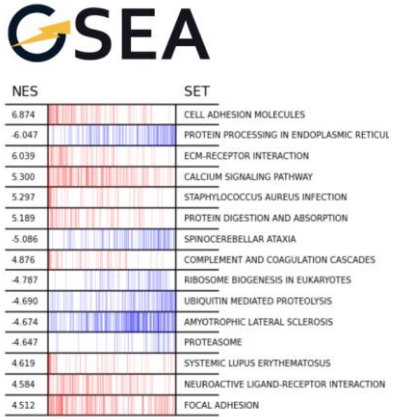


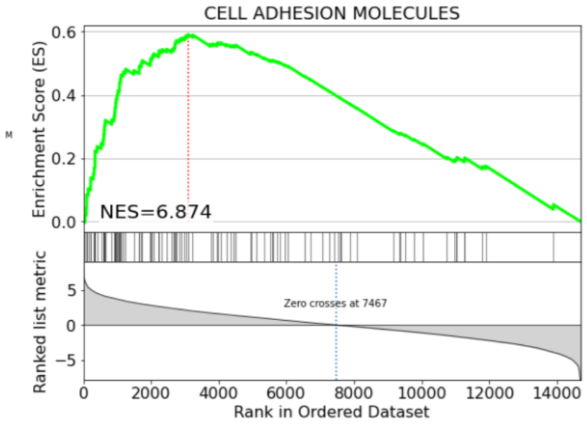
A lot of differentially expressed genes



Search for pathway enrichment

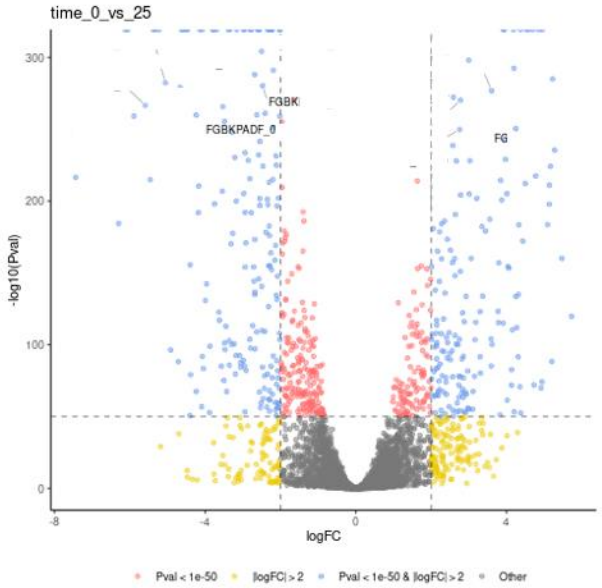
Choosing your genes for mutation

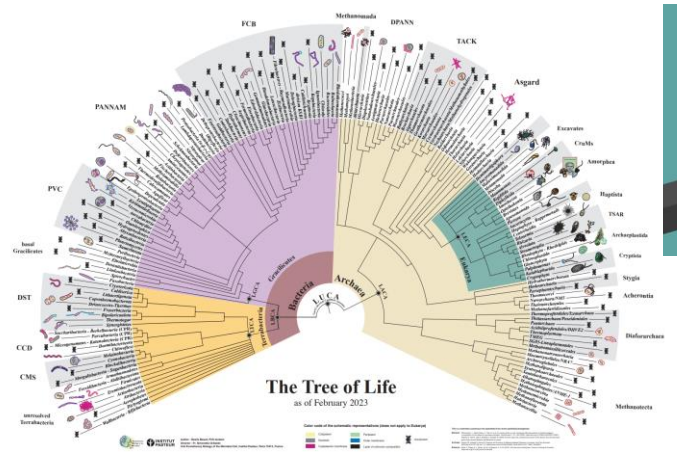


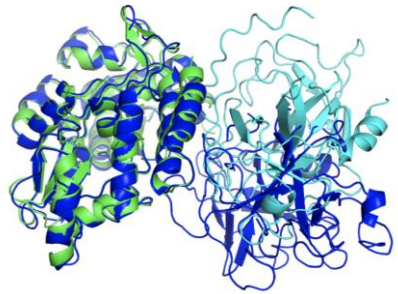


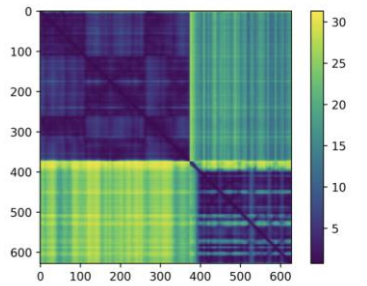
Search for pathway enrichment

A lot of differentially expressed genes









Search for putative function of an unknown highly differentially expressed protein

Let’s try it with our own dataset

https://abignaud.github.io/Transcriptomics_SupBioTech_2023/Session03/DESeq2