

RNA-SEQ PROFILING DROSOPHILA CELLS AFTER THE DEPLETION OF A REGULATORY GENE

GROUP 6



Background

IN THE STUDY OF BROOKS ET AL. 2011, THE AUTHORS IDENTIFIED GENES AND PATHWAYS REGULATED BY THE PASILLA GENE (THE DROSOPHILA HOMOLOG OF THE MAMMALIAN SPLICING REGULATORS NOVA-1 AND NOVA-2 PROTEINS) USING RNA-SEQ DATA. THEY DEPLETED THE PASILLA (PS) GENE IN DROSOPHILA MELANOGASTER BY RNA INTERFERENCE (RNAI).

TOTAL RNA WAS THEN ISOLATED AND USED TO PREPARE BOTH SINGLE-END AND PAIRED-END RNA-SEQ LIBRARIES FOR TREATED (PS DEPLETED) AND UNTREATED SAMPLES. THESE LIBRARIES WERE SEQUENCED TO OBTAIN RNA-SEQ READS FOR EACH SAMPLE.

Motivation

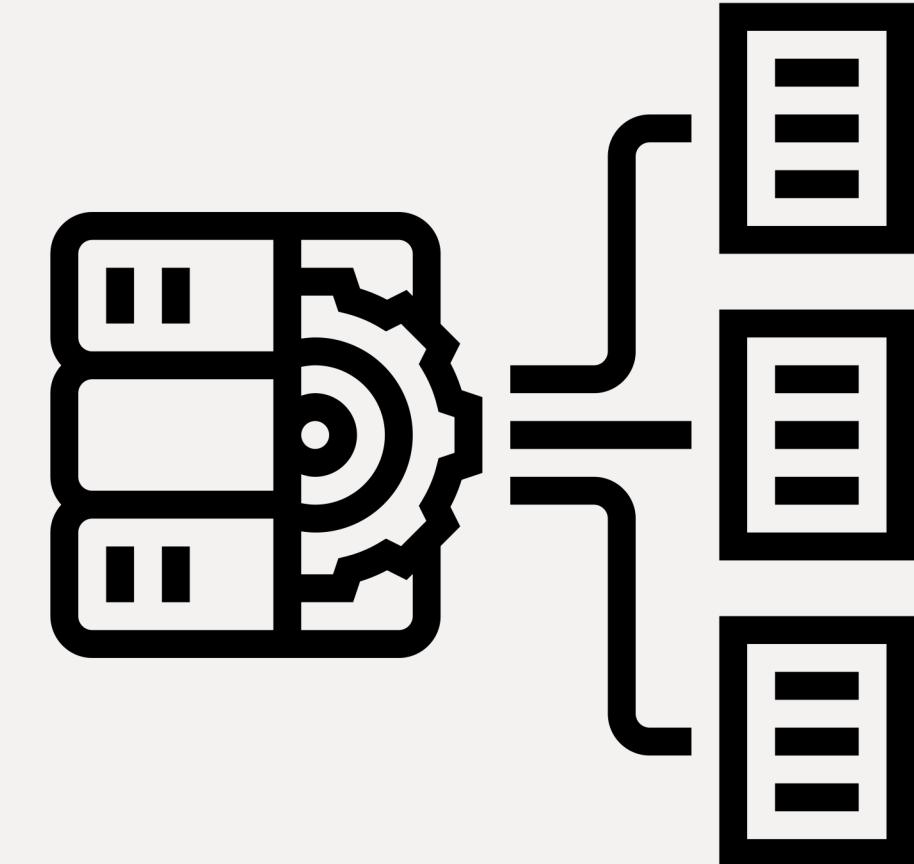
OUR INTEREST IN THE PARTICULAR GENE WAS PIQUED WHEN WE READ THE STUDY OF BROOKS ET AL. THE PAPER EXPLORED THE POSSIBILITY OF CONSERVATION OF REGULATORY CODES AND CONCLUDED THAT THE RNA MAP OF PS AND NOVA1/2 IS HIGHLY CONSERVED BETWEEN MAMMALS AND INSECTS.



Datasets

- 4 UNTREATED SAMPLES: GSM461176, GSM461177, GSM461178, GSM461182
- 3 TREATED SAMPLES (PASILLA GENE DEPLETED BY RNAI): GSM461179, GSM461180, GSM461181

EACH SAMPLE CONSTITUTES A SEPARATE BIOLOGICAL REPLICATE OF THE CORRESPONDING CONDITION (TREATED OR UNTREATED). MOREOVER, TWO OF THE TREATED AND TWO OF THE UNTREATED SAMPLES ARE FROM A PAIRED-END SEQUENCING ASSAY, WHILE THE REMAINING SAMPLES ARE FROM A SINGLE-END SEQUENCING EXPERIMENT.



Galaxy Europe

Workflow Visualize Shared Data Help User Using 0%

Tools

search tools

Upload Data

OBO Ontology manipulation

HCA-Single Cell

Other Tools

Biodiversity data exploration

Sanger Sequencing

Extract Features

QIIME 2

Metagenomic Analysis

GATK Tools

Built-in Converters

WORKFLOWS

All workflows

search datasets in selected histories

RNA_seq_analysis

'DEA_7_genes'

6.35 GB 52 4 12

158 MB 37

65 : MultiQC on data 6 2 and data 59: Webpage #GSM461177 #GSM461180

41 : Pathview on data 3 9: KEGG Pathway (dme03040)

40 : Pathview on data 39: KEGG Pathway (dm e00010)

39 : Genes with significant adj p-value and their Log2 FC

38 : goseq on data 32 and data 30: DE genes

Current History Hide

Switch to Hide

History

search datasets

RNA_seq_analysis

6.35 GB 52 4 12

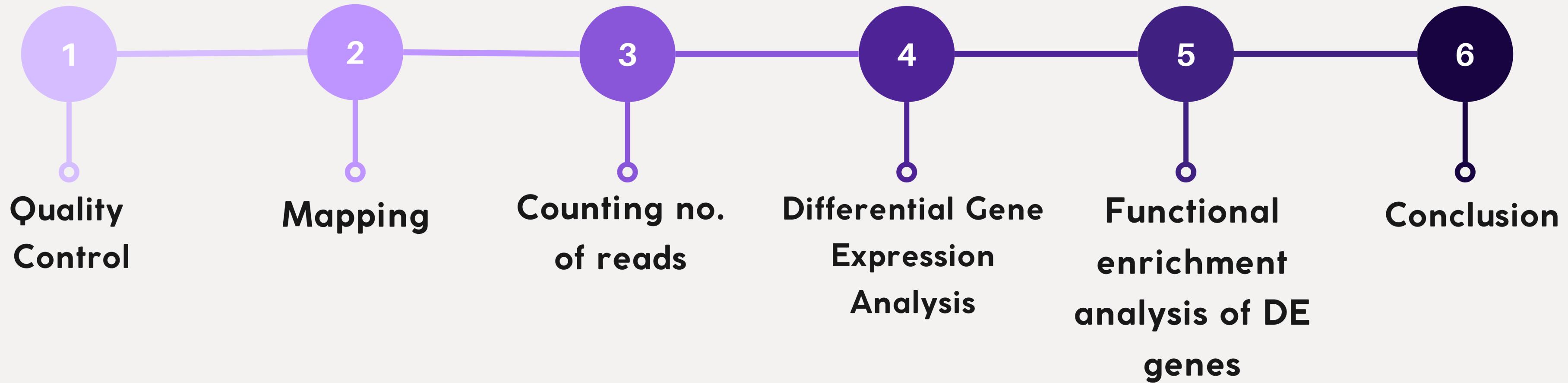
65 : MultiQC on data 62 and d ata 59: Webpage #GSM461177 #GSM461180

64 : MultiQC on data 62 and d ata 59: Stats #GSM461177 #GSM461180

63 : featureCounts on data 29 and data 40: F #GSM461180

Select histories

STEPS

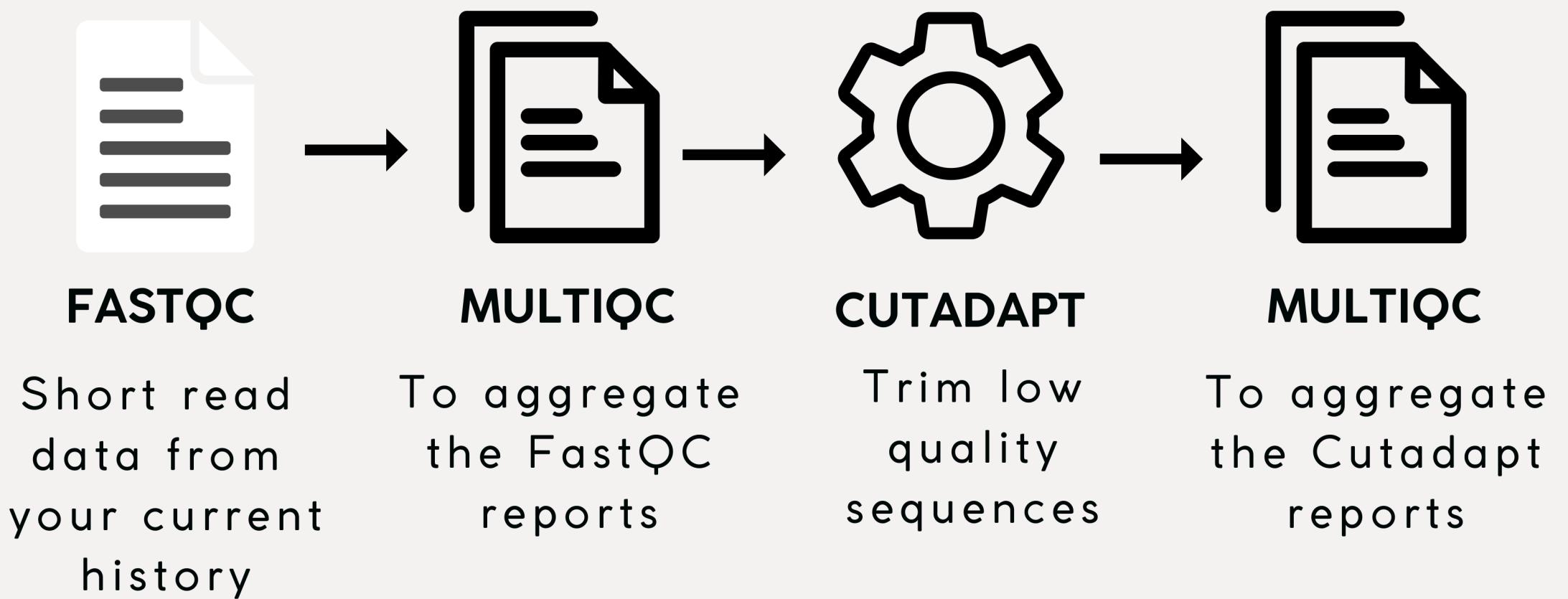


STEP 1 – QUALITY CONTROL

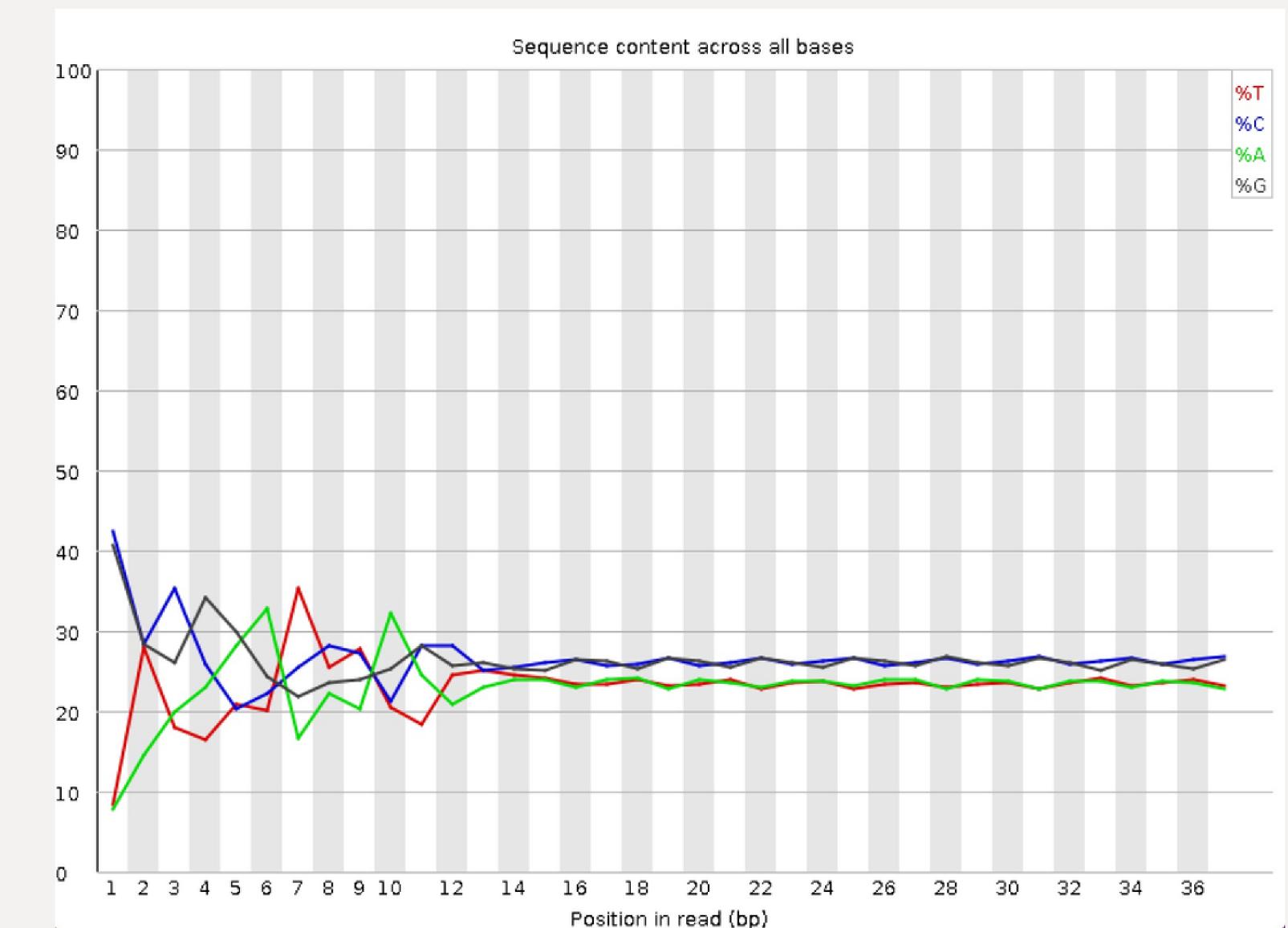
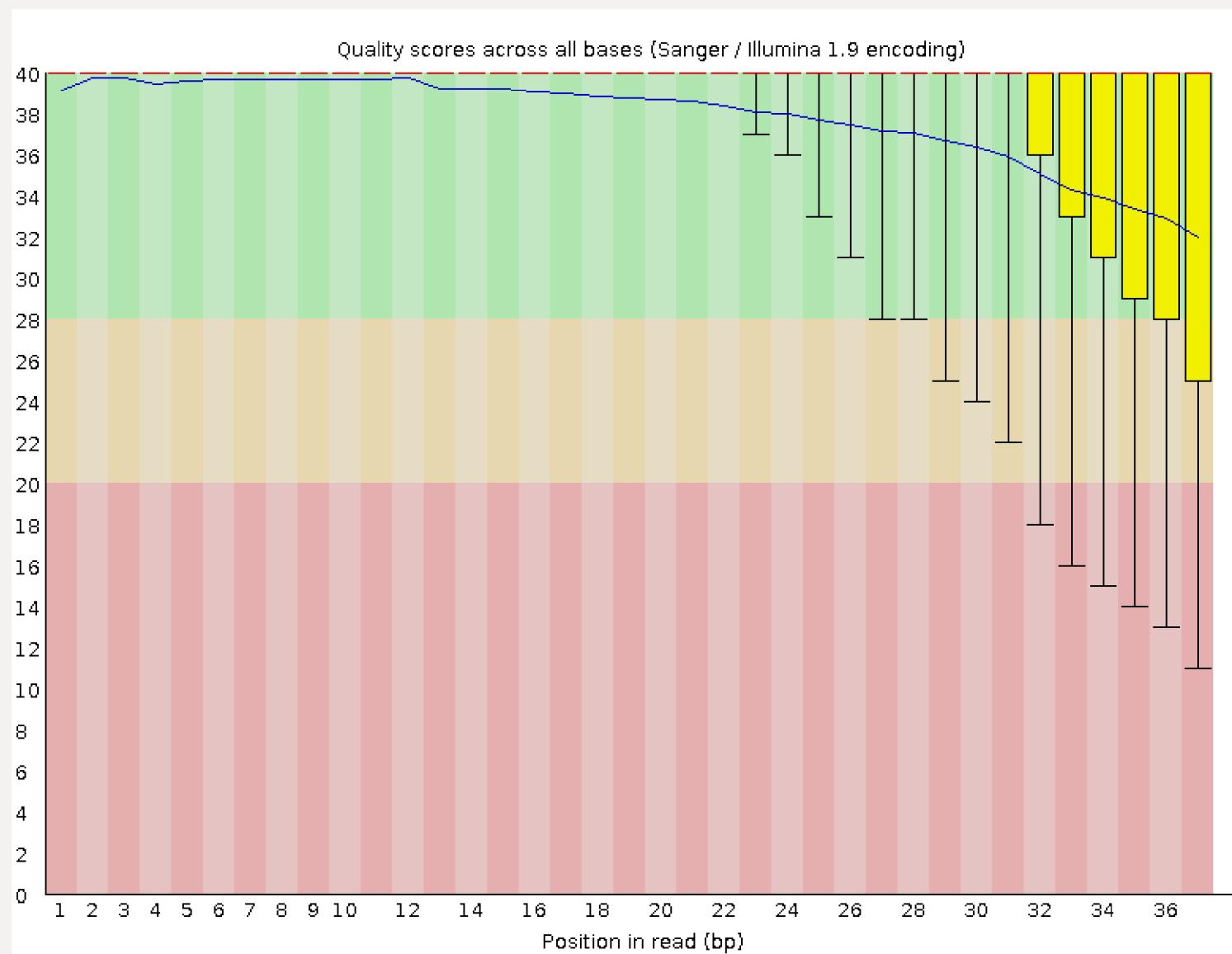
⇒ WHY CONTROL QUALITY?

- Ensures accurate and reliable results
- Poor quality data leads to errors and false conclusions
- High-quality data necessary for downstream analysis

⇒ HOW?



⇒ OUTPUTS:-



STEP 2 – MAPPING

⇒ WHAT IS MAPPING?

PROCESS OF ALIGNING THE SEQUENCE
READS TO A REFERENCE GENOME

⇒ WHAT IS A REFERENCE GENOME?

A REFERENCE GENOME IS A WELL-ANNOTATED GENOME
SEQUENCE THAT SERVES AS A REPRESENTATIVE
EXAMPLE FOR A SPECIES OR GROUP OF RELATED
SPECIES.



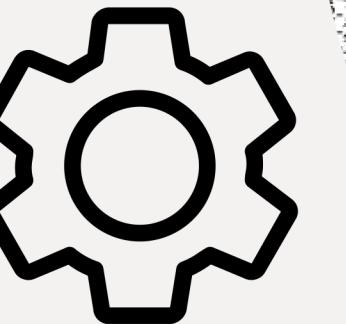
Preparation
of reference
genome



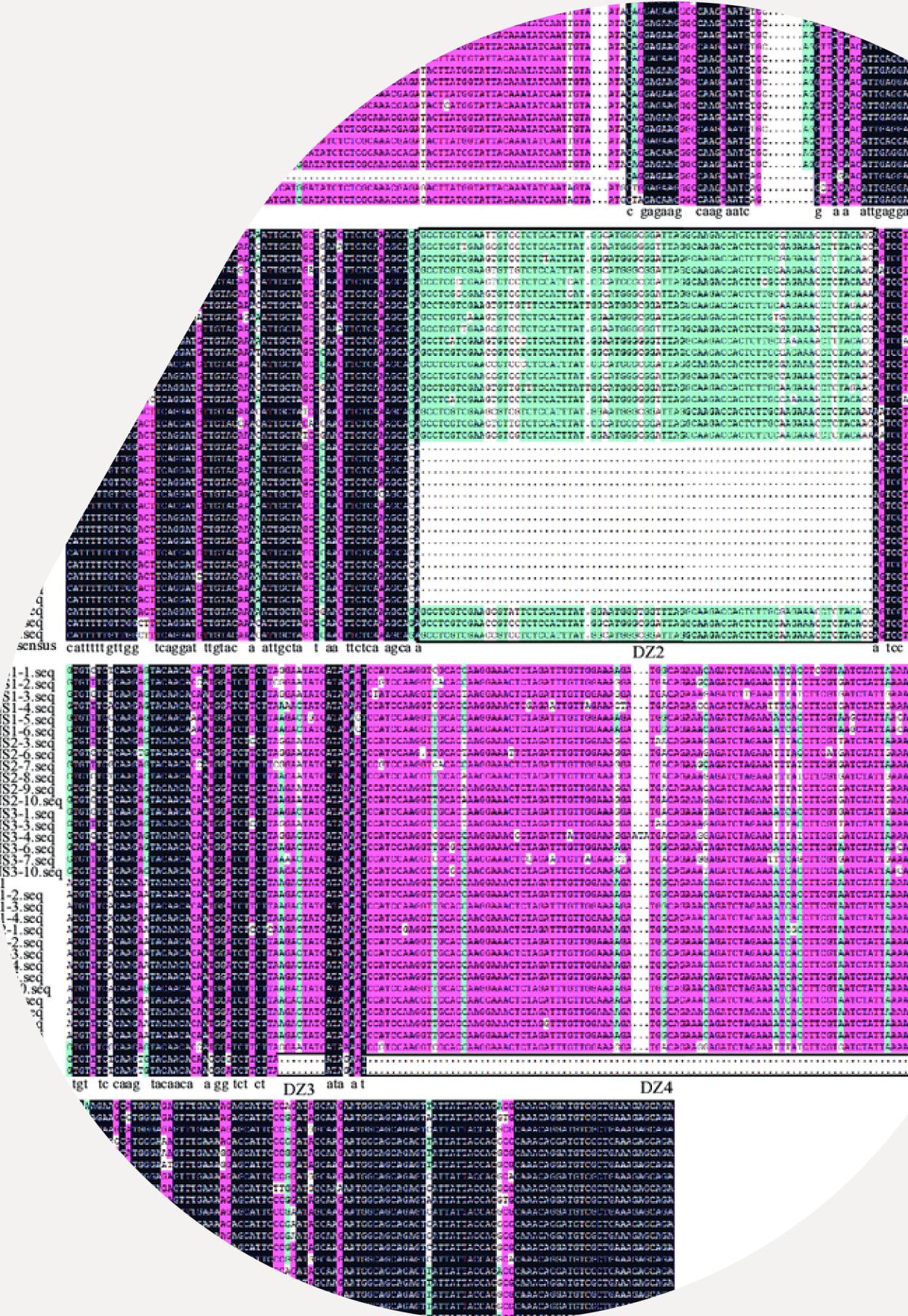
Sequence
Alignment



BAM/SAM
File

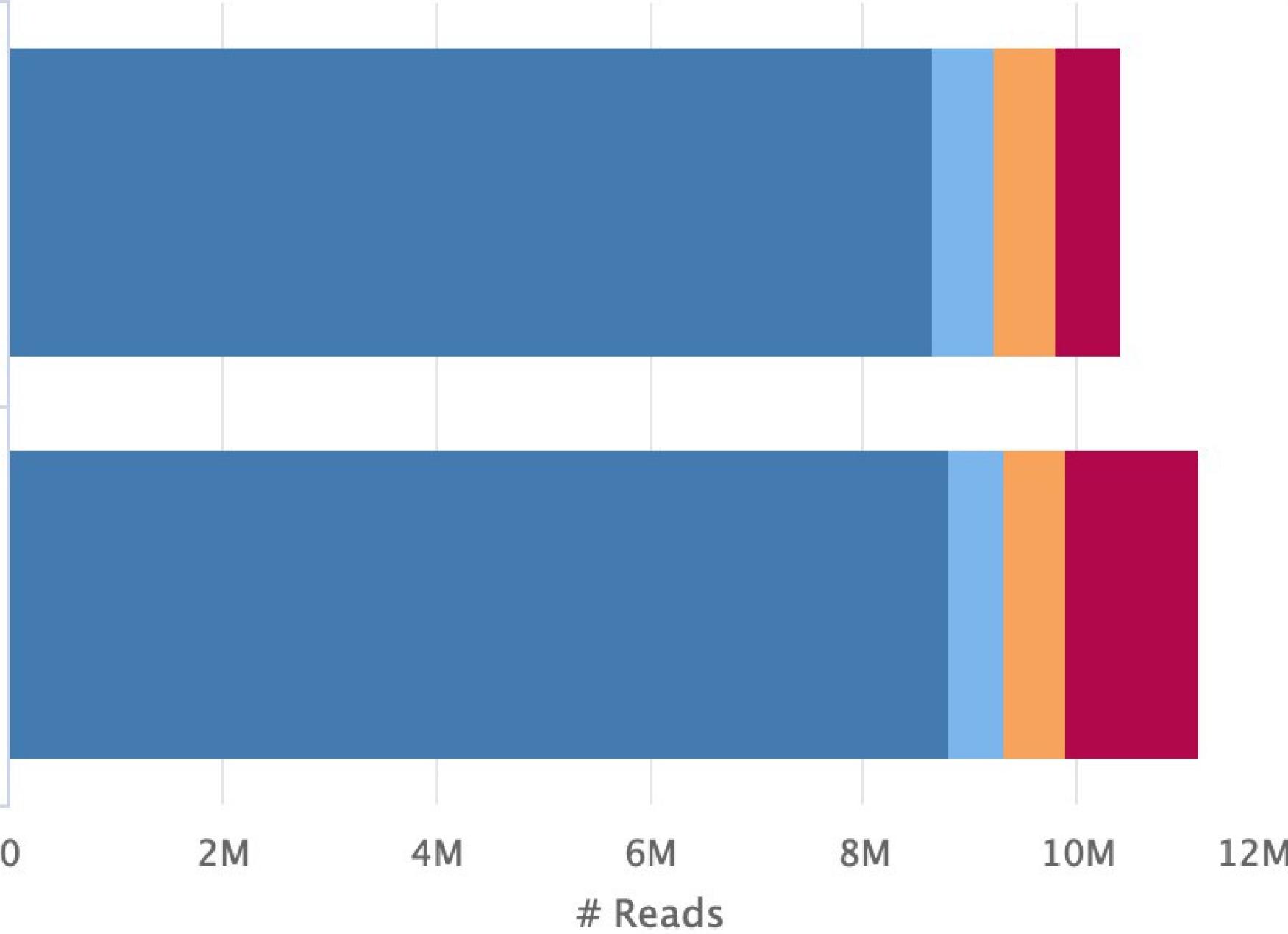


Post
Processing
and Analysis



STAR: Alignment Scores

RNA STAR on data 29_data 19_and data 18_log



● Uniquely mapped
● Unmapped: too short

● Mapped to multiple loci
● Unmapped: other

● Mapped to too many loci

Created with MultiQC

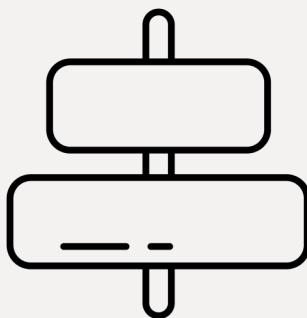
STEP 3 – COUNTING NUMBER OF READS

⇒ WHY?

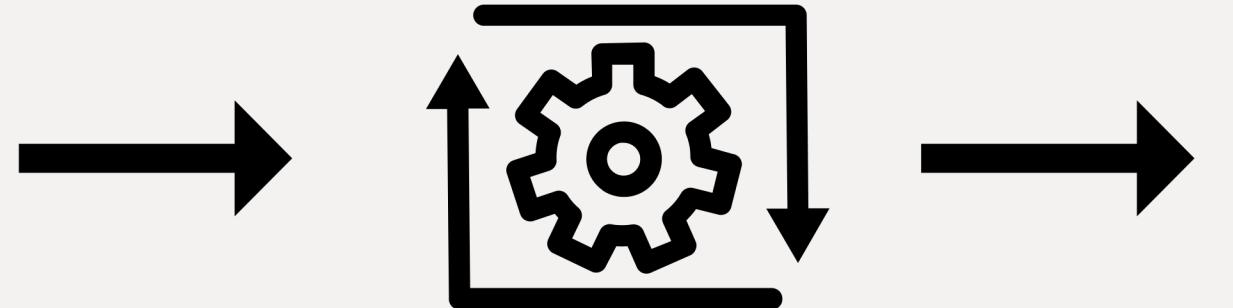
TO COMPARE GENE EXPRESSION BETWEEN CONDITIONS,
WE NEED TO QUANTIFY THE NUMBER OF READS PER
GENE.

⇒ STRANDEDNESS

- STRANDEDNESS OF THE LIBRARY MUST BE DETERMINED FOR ACCURATE COUNTING OF READS OVERLAPPING WITH GENOMIC FEATURES.



ALIGN READS EVALUATE
COUNTS IN THREE
STRANDEDNESS
POSSIBILITIES



DETERMINE LIBRARY
STRANDEDNESS FROM
STAR RESULTS



USE STAR TO COUNT
READS PER GENE AND
GENERATE A TABLE
WITH COUNTS MAPPED



STEP 4 – DIFFERENTIAL EXPRESSION ANALYSIS

⇒ OBJECTIVES

- NORMALISE DATASET AND GENERATE A SINGLE FILE WITH ALL SAMPLE DATA
- PLOT PCA AND SAMPLE-TO-SAMPLE DISTANCE
- GET DIFFERENTIALLY EXPRESSED GENES

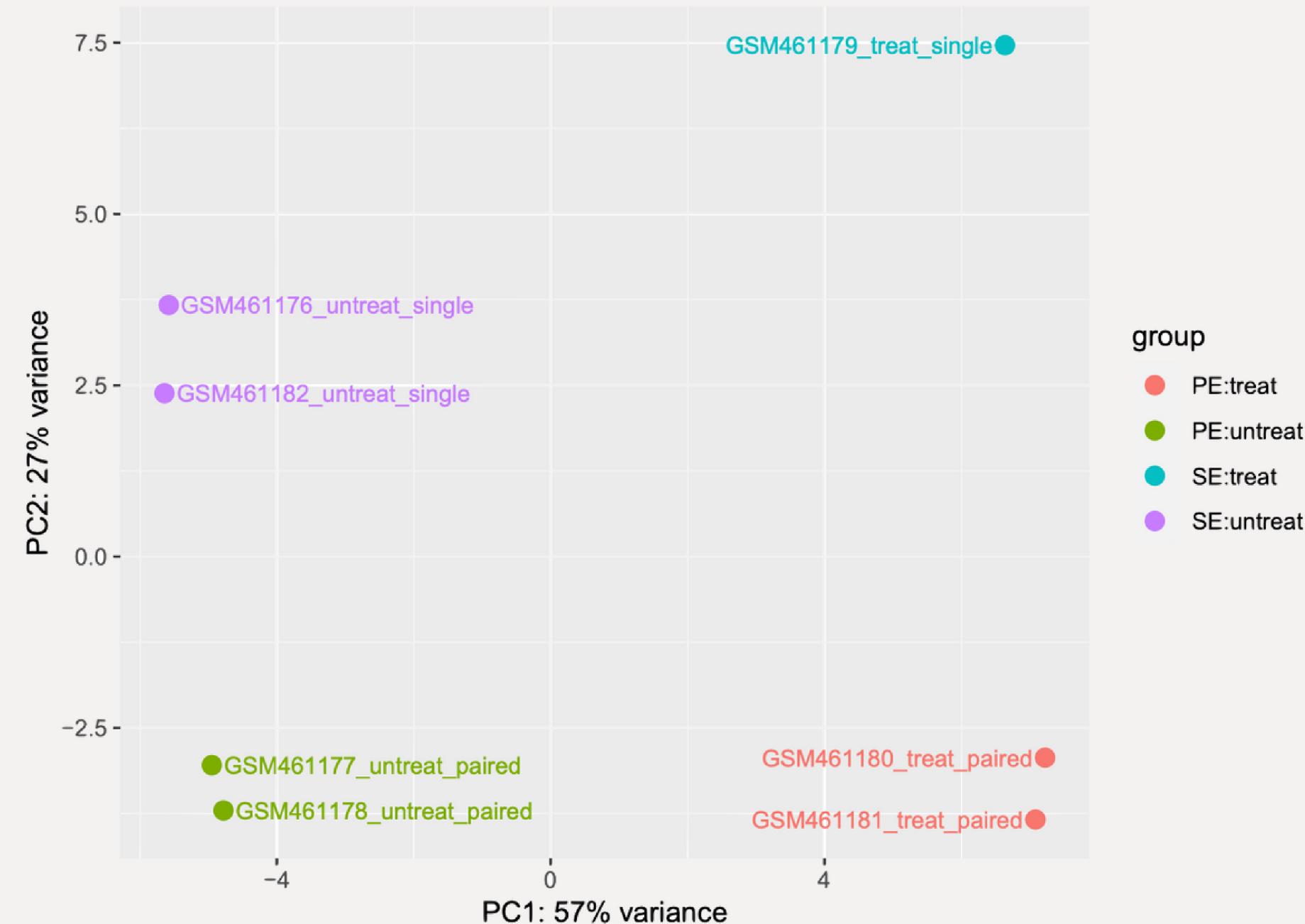


USES DEseq TO NORMALISE DATA OVER CONVENTIONAL PROCEDURES

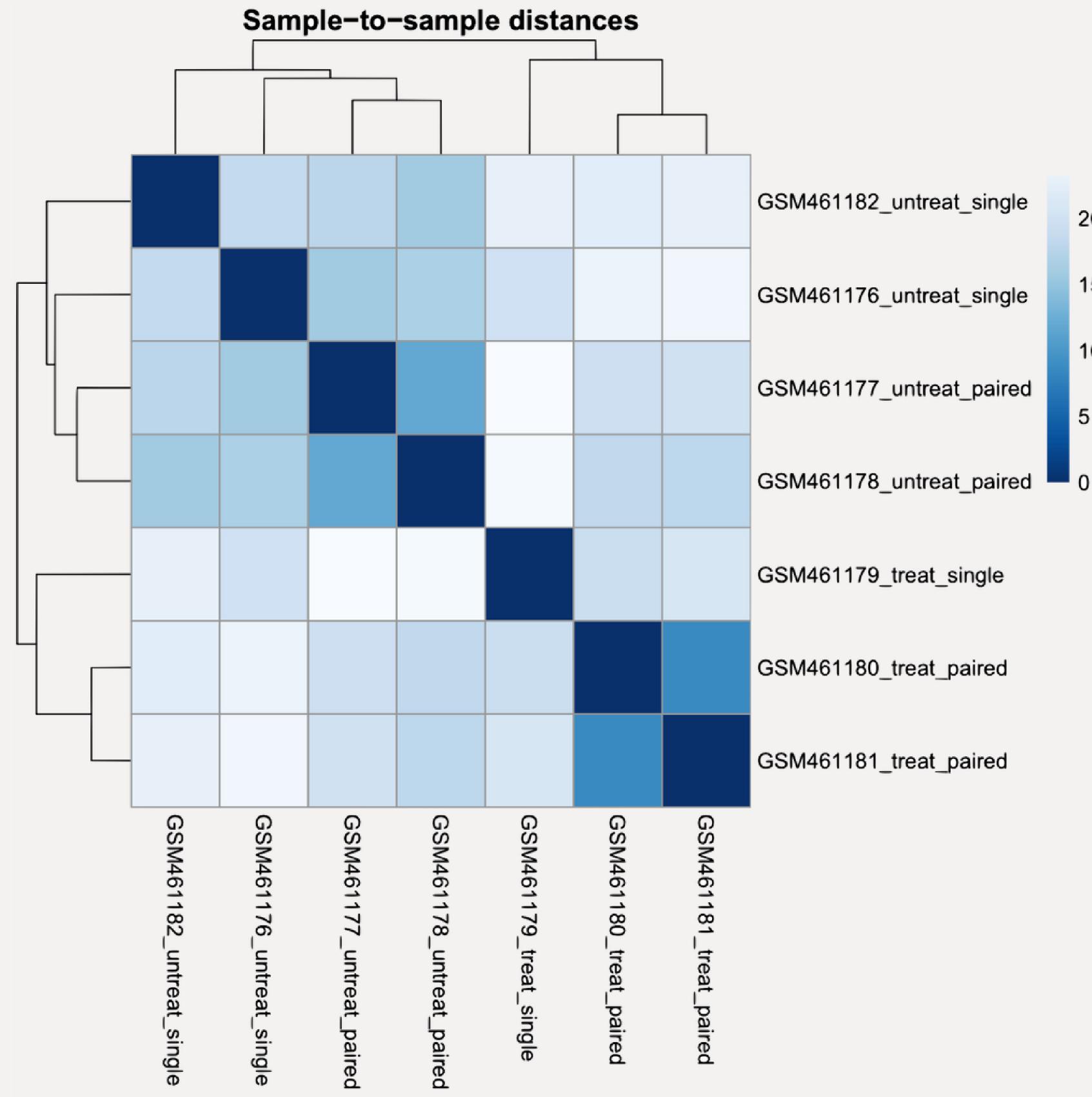
- ⇒ USES SEQUENCING DEPTH AND NORMALISATION AND LIBRARY COMPOSITION
- ⇒ RPKM OR TPM DON'T ACCOUNT OF DIFFERENCE IN LIBRARY COMPOSITION
- ⇒ ABOUT 17000 GENES BETWEEN THE 7 SAMPLES

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7
	GSM461176_untreat_single	GSM461177_untreat_paired	GSM461178_untreat_paired	GSM461182_untreat_single	GSM461179_treat_single	GSM461180_treat
FBgn0085804		0	0	0	0	0
FBgn0267431	2.19988706875227		0	0	0	8.51758304160428
FBgn0039987		0	0	0	0	0
FBgn0058182		0	0	0	0	0
FBgn0267430		0	0	0	0	1.54865146210987
FBgn0266747		0	0	0	0	0
FBgn0086917	1.46659137916818		0	0	0	1.54865146210987

PRINCIPLE COMPONENT ANALYSIS

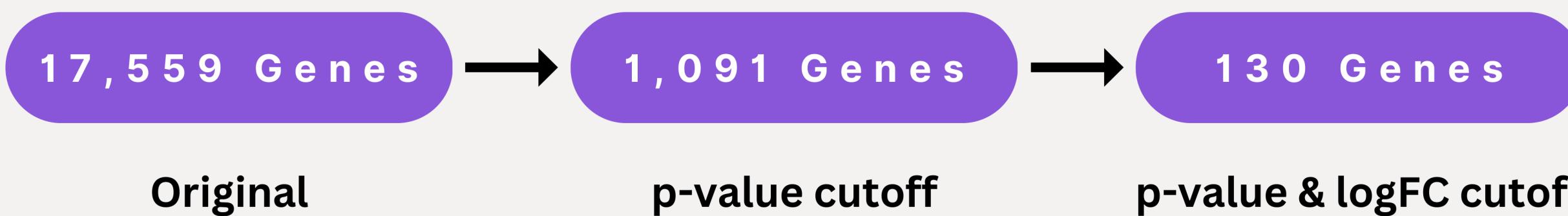


DISTANCE BETWEEN SAMPLES



RESULT AND ANNOTATION

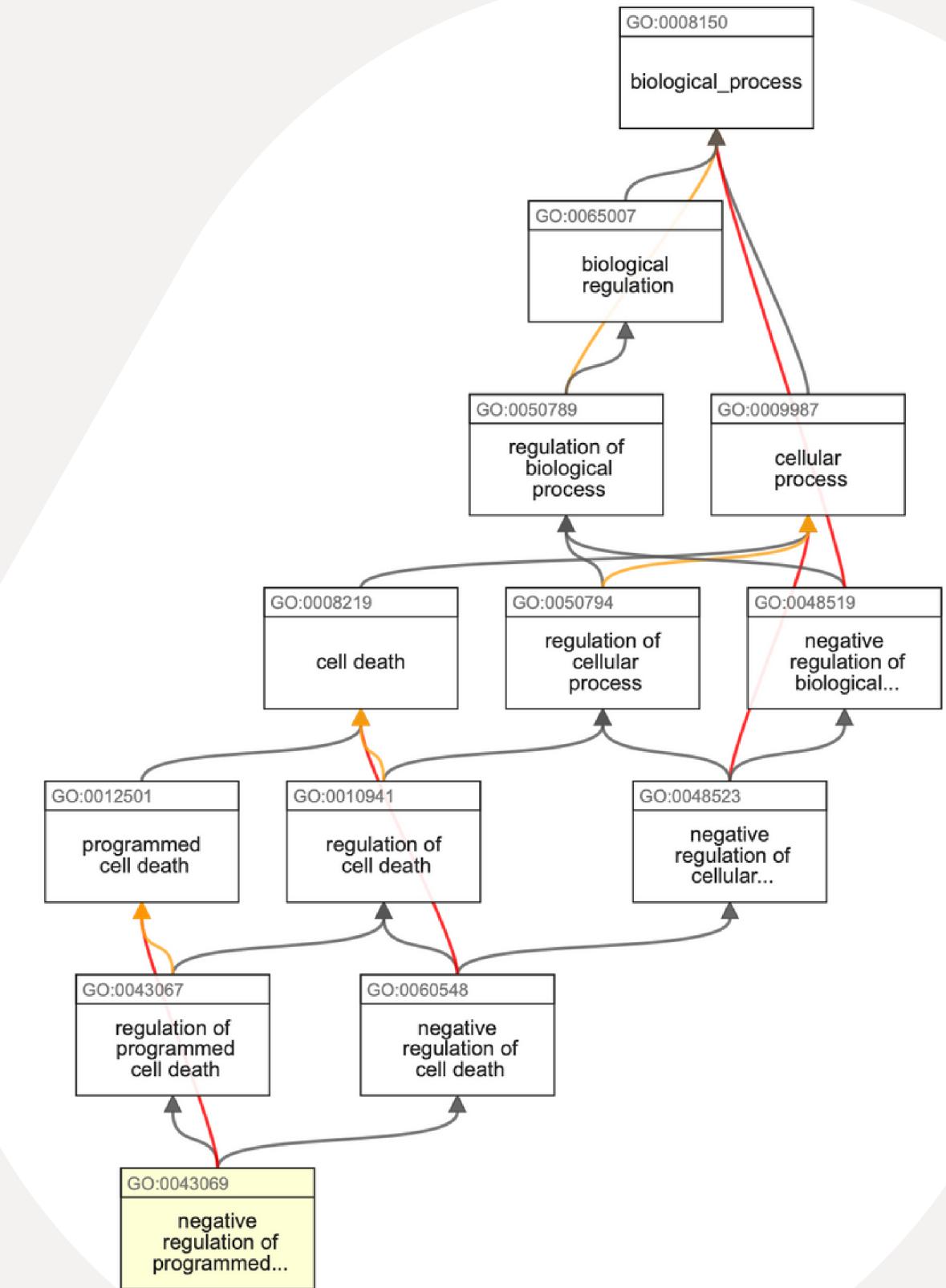
- ⇒ PVALUE ADJUSTED FOR COMPARISON ACROSS GENES
- ⇒ LOG FOLD CHANGE VALUES USED TO DETERMINED EXPRESSION OF GENES
- ⇒ SELECT GENES BASED ON PVALUE <0.05 AND ABSOLUTE LOG FC VALUE > 1



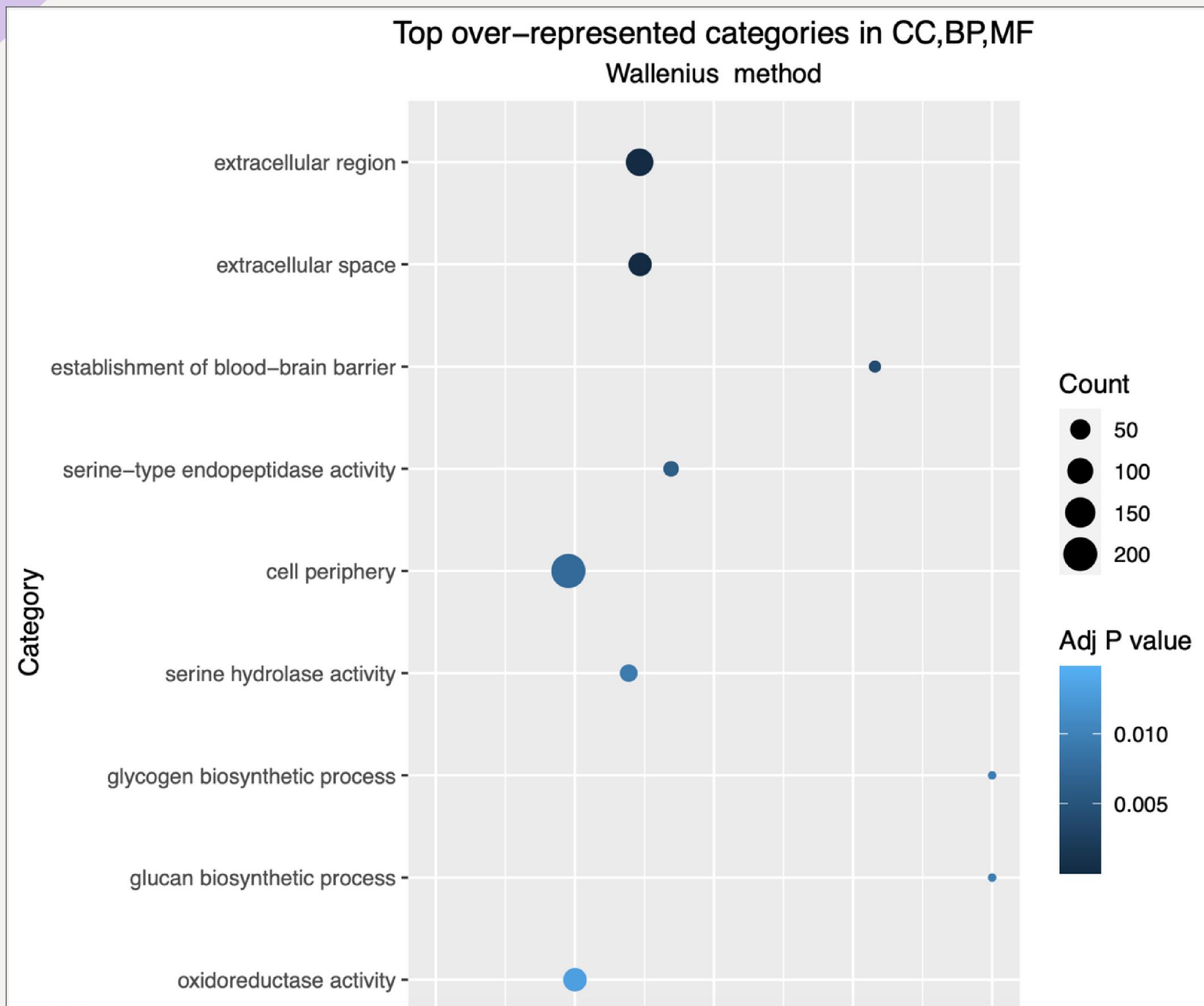
- ⇒ ADD ATTRIBUTES TO GENES TO MAKE DATA INTUITIVE AND DIGESTIBLE

STEP 5 – FUNCTIONAL ENRICHMENT ANALYSIS OF DE GENES

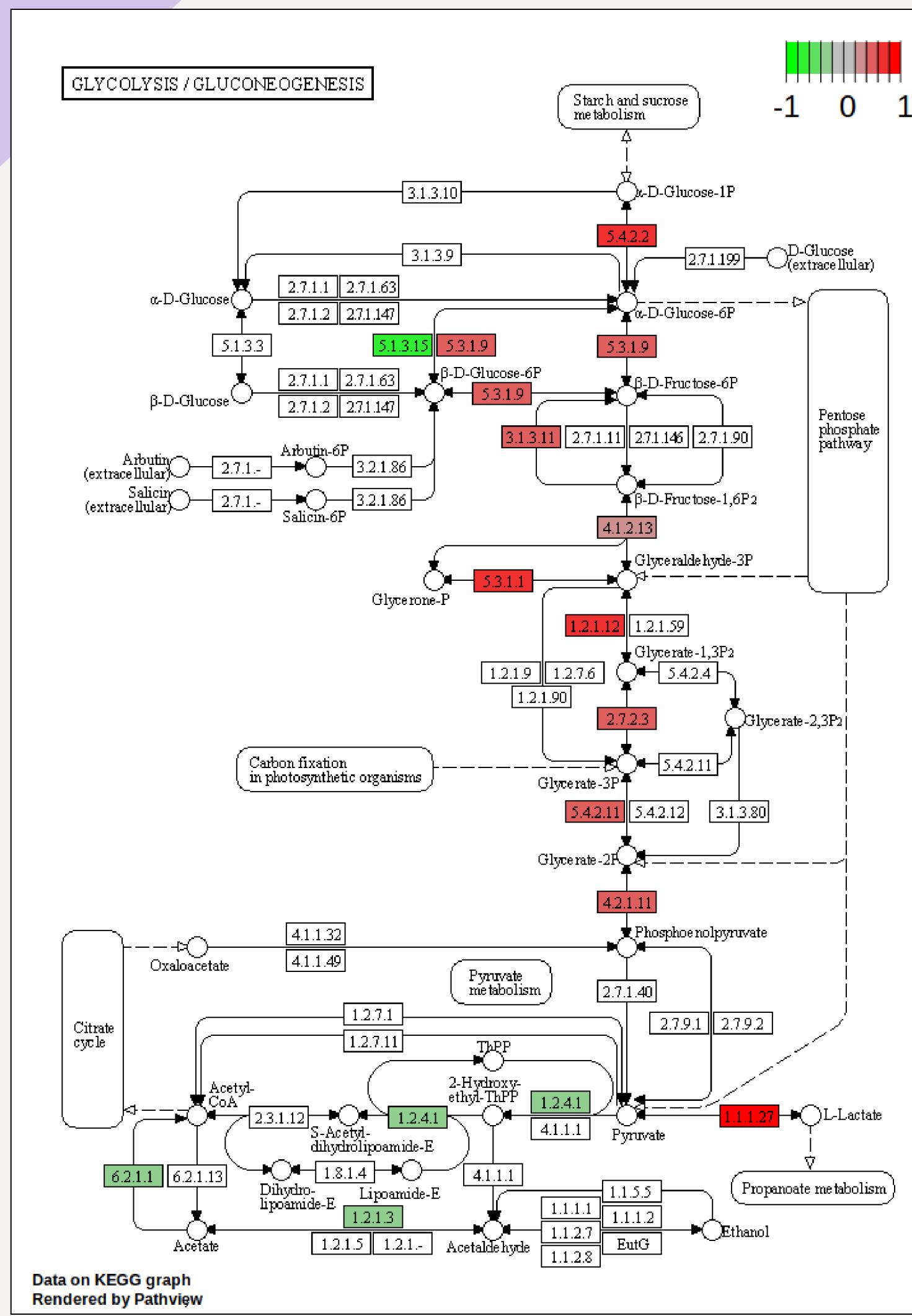
- ⇒ DE GENE FUNCTIONAL ENRICHMENT ANALYSIS IDENTIFIES OVER-REPRESENTED BIOLOGICAL PATHWAYS, FUNCTIONAL CATEGORIES, OR GENE SETS IN A GIVEN LIST OF GENES OR PROTEINS.
- ⇒ GO TERMS DESCRIBE GENE FUNCTIONS IN MF, BP, AND CC CATEGORIES, PROVIDING A STANDARDIZED WAY TO DESCRIBE BIOLOGICAL PATHWAYS AND FUNCTIONS.
- ⇒ ENRICHMENT ANALYSIS TYPICALLY INVOLVES TWO MAJOR STEPS: GENE SET ENRICHMENT ANALYSIS AND PATHWAY VISUALIZATION/INTERPRETATION.



ENRICHMENT ANALYSIS RESULTS WITH WALLENIUS METHOD



THIS GRAPH SHOWS THE TOP 10 OVER-REPRESENTED CATEGORIES IN CC, BP, AND MF AS IDENTIFIED THROUGH ENRICHMENT ANALYSIS USING THE WALLENIUS METHOD. CC STANDS FOR CELLULAR COMPONENT, BP STANDS FOR BIOLOGICAL PROCESS, AND MF STANDS FOR MOLECULAR FUNCTION. THE WALLENIUS METHOD IS A STATISTICAL METHOD USED TO CORRECT FOR GENE SET SIZE BIAS IN ENRICHMENT ANALYSIS. THE PERCENTAGE OF GENES THAT ARE DIFFERENTIALLY EXPRESSED AND THE ENRICHMENT FACTOR ARE ALSO SHOWN.



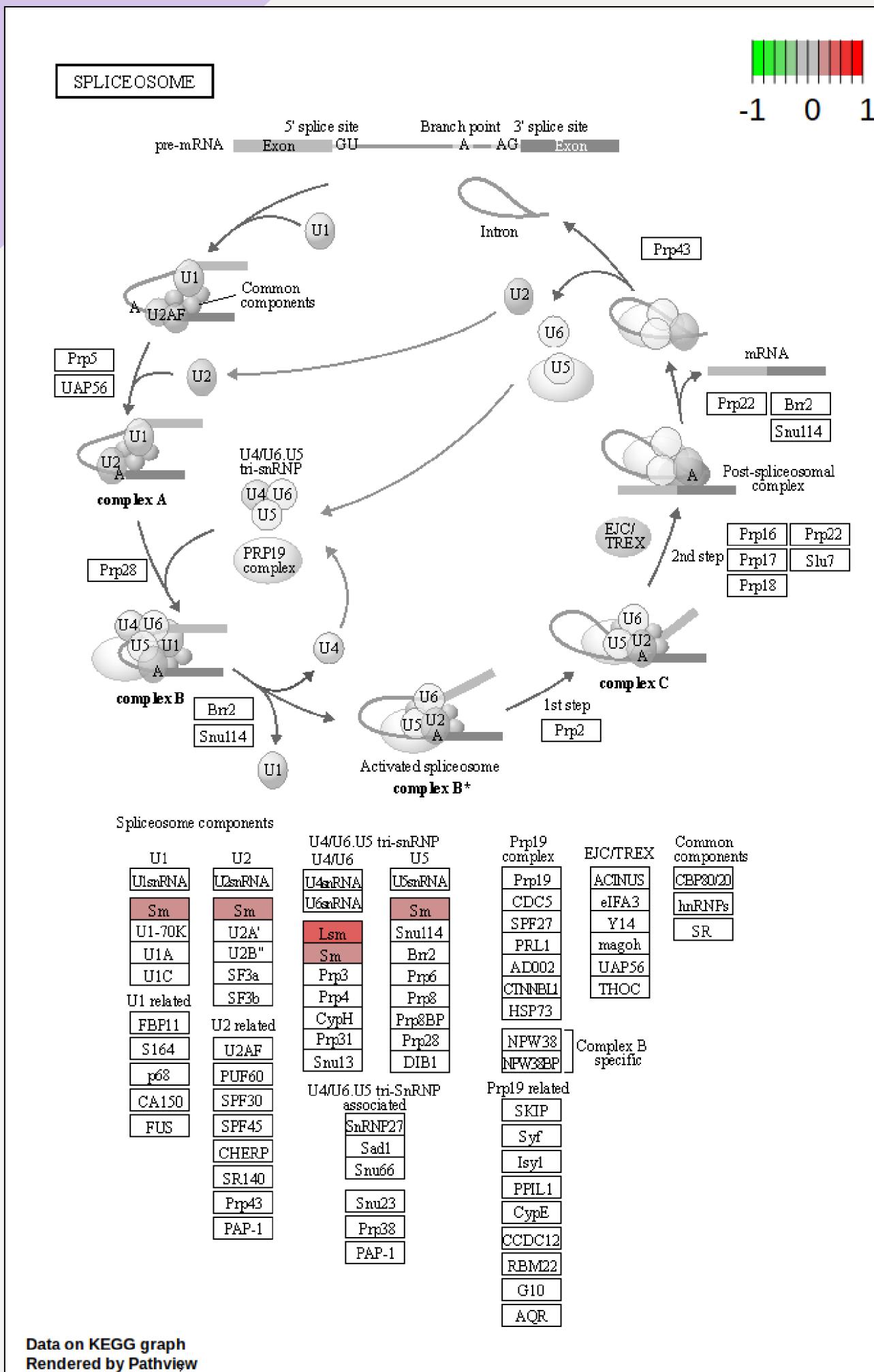
KEGG PATHWAY GRAPH

- REVIEWING THE TABLE CAN BE TIME-CONSUMING AND CUMBERSOME.
- TO BETTER VISUALIZE THIS INFORMATION, WE UTILIZED PATHVIEW.
- PATHVIEW IS A TOOL THAT AUTOMATICALLY GENERATES PATHWAY IMAGES WITH ADDED GENE INFORMATION, SUCH AS GENE EXPRESSION DATA.

KEGG PATHWAY GRAPH

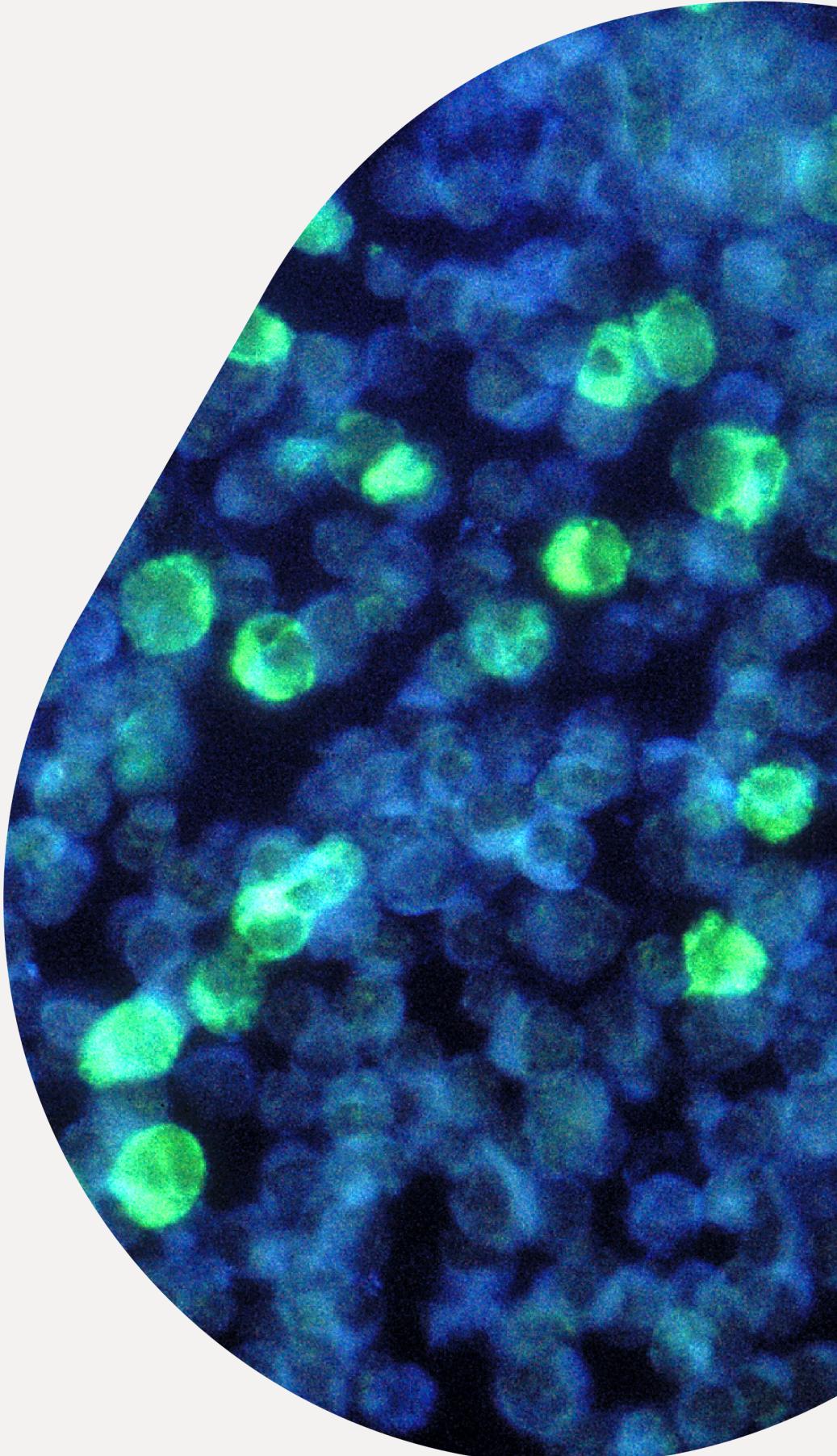
VISUALIZING THE SPLICEOsome PATHWAY:-

- KEGG PATHWAY ID: O3040
- INVOLVED IN RNA PROCESSING FOR REMOVAL OF INTRONS AND SPLICING OF EXONS
- NOT SIGNIFICANTLY OVER-REPRESENTED IN ANALYSIS, BUT STILL INVESTIGATED DE GENES IN PATHWAY
- USED PATHVIEW TO GENERATE PLOT WITH GENE NODES COLORED BY LOG2 FOLD CHANGE
- PLOT REVEALED SEVERAL GENES WITH ALTERED EXPRESSION IN PATHWAY, PROVIDING POTENTIAL INSIGHTS INTO TREATMENT'S IMPACT ON BIOLOGICAL PROCESSES.



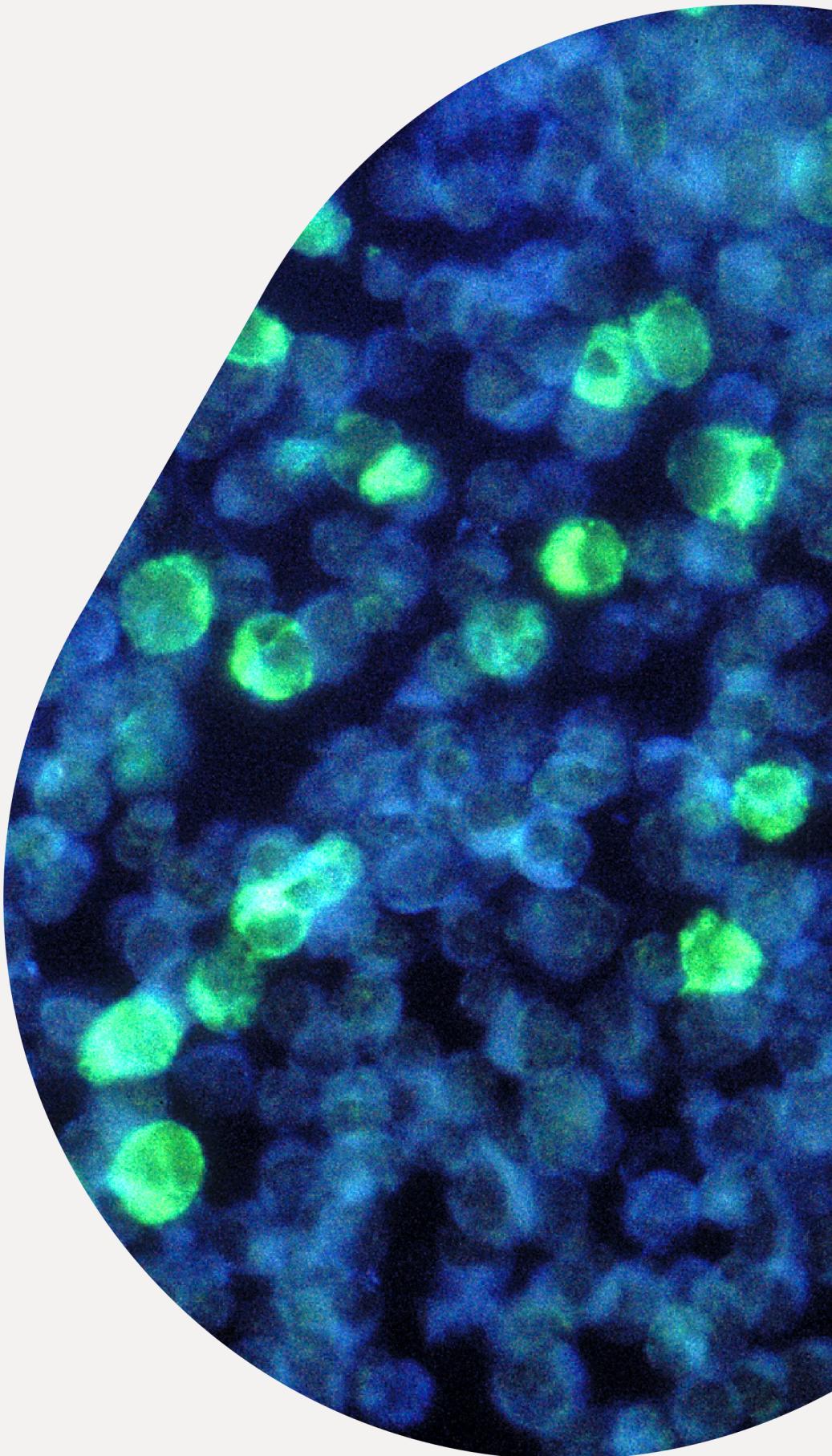
STEP 6 – CONCLUSION

- DEPLETING PASILLA GENE AFFECTS THE SPLICING RATE OF GENE A AND B DIFFERENTLY.
- UPREGULATION OF GENE A AND DOWNREGULATION OF GENE B SUGGEST THAT PASILLA SUPPRESSES AND PROMOTES THEIR SPLICING RATE, RESPECTIVELY.
- PASILLA DEPLETION LEADS TO UNDERREPRESENTATION OF SPLICEOSOME PATHWAY GENES, IMPLYING THAT CORE SPLICEOSOME GENES ARE NOT AFFECTED.
- UNDERREPRESENTATION OF THE SPLICEOSOME PATHWAY DOES NOT MEAN NO CHANGE IN SPLICING EVENTS, AND ANALYZING DIFFERENTIAL SPLICING EVENTS CAN PROVIDE BETTER INSIGHTS.
- PASILLA DEPLETION SIGNIFICANTLY AFFECTS THE EXPRESSION LEVELS OF GENES INVOLVED IN THE GLYCOLYSIS PATHWAY.



CONTD...

- THE OBSERVED DIFFERENTIAL EXPRESSION IN THE GLYCOLYSIS PATHWAY SUGGESTS THAT PASILLA DEPLETION HAS CASCADING EFFECTS ON OTHER GENES, POSSIBLY THROUGH ALTERED ALTERNATIVE SPLICING PATTERNS.
- PASILLA AND NOVA1/2 GENES ARE INVOLVED IN ALTERNATIVE SPLICING REGULATION IN INSECTS AND MAMMALS, RESPECTIVELY.
- THE RNA MAPS OF PASILLA AND NOVA1/2 ARE HIGHLY CONSERVED BETWEEN MAMMALS AND INSECTS, SUGGESTING SIMILAR SPLICING PATTERNS AND CONSERVED ROLES IN BIOLOGICAL PATHWAYS.
- THE STUDY'S RESULTS CAN BE EXTENDED TO STUDYING NOVA GENES IN MAMMALS.



REFERENCES

Conservation of an RNA regulatory map between *Drosophila* and mammals.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032923/>

STAR: ultrafast universal RNA-seq aligner

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032923/>

Pathview: an R/Bioconductor package for pathway-based data integration and visualization.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032923/>

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032923/>

MultiQC: summarize analysis results for multiple tools and samples in a single report.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032923/>

CONTRIBUTORS

- AAYUSH RANJAN (2021003)
- KARTIK GUPTA (2021056)
- LAKSHYA GOEL (2021469)
- LAKSHYA KUMAR RAIKWAR (2021470)
- N NAROTAM (2021477)
- PARAS DHIMAN (2021482)
- SANMAY SOOD (2021095)
- SIDDHARTH RAJPUT (2021102)

THANK YOU !

