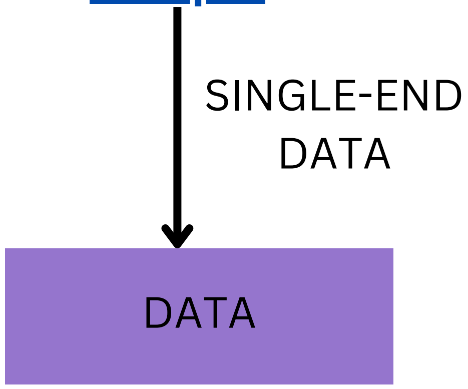
REPORT

SUBMITTED BY:

Vaishnavi Goyal

Table of Contents

Sample



SINGLE-END DATA

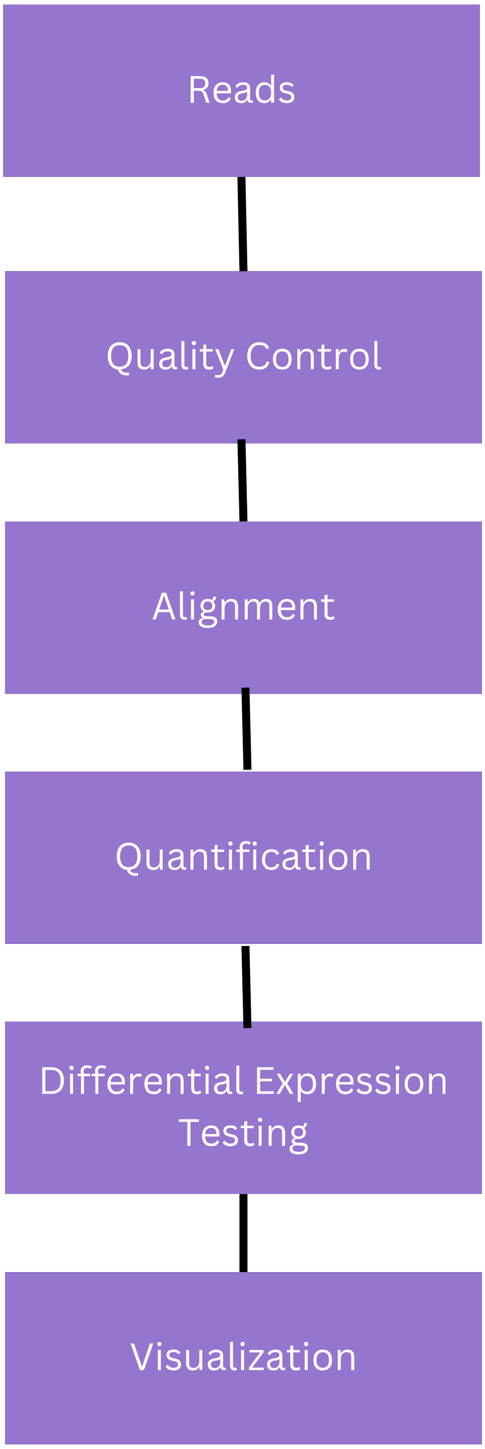
DATA

1. SRR1552444.fastq.gz
2. SRR1552445.fastq.gz
3. SRR1552446.fastq.gz
4. SRR1552447.fastq.gz
5. SRR1552448.fastq.gz
6. SRR1552449.fastq.gz
7. SRR1552450.fastq.gz
8. SRR1552451.fastq.gz
9. SRR1552452.fastq.gz
10. SRR1552453.fastq.gz
11. SRR1552454.fastq.gz
12. SRR1552455.fastq.gz

TO UNZIP DATA

```
gunzip foldername/SRR*
```

Workflow


Reads
Quality Control
Alignment
Quantification
Differential Expression Testing
Visualization

Tools Used

1. COMMAND-Based In UBUNTU

- **Quality Control ->** fastqc reads/SRR* -o fastqc-output
- **Multiqc Report ->** multiqc .
- **Quality Trimming ->** fastp -i reads/SRR1552444.fastq -o trim/SRR1552444_trim.fastq -j SRR1552444.json -h SRR1552444.html (PERFORM FOR ALL READS)
- **Mapping ->** hisat2 --phred33 --dta -x ./reference/grcm38/genome -U ./trim/SRR1552444_trim.fastq -S mapped/SRR1552444.sam (PERFORM FOR ALL READS)
- **Mapping Summary ->** samtools flagstat mapped/SRR1552444.sam>summary/SRR1552444_alignment_summary.txt (PERFORM FOR ALL READS)
- **Count File ->** featureCounts -a reference/ GTF_File -o feature-count/counts.txt -t exon -g gene_id mapped/SRR*.sam

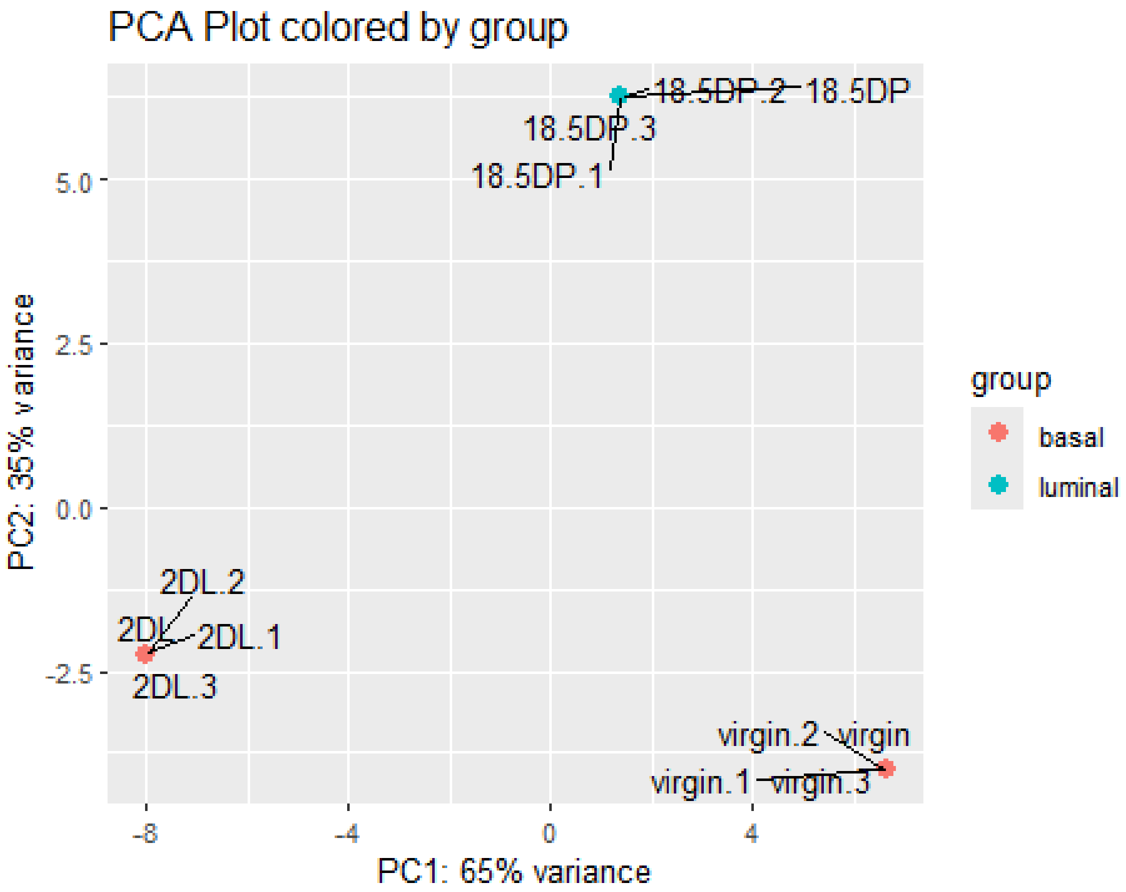
2. RStudio

- **Step1 ->** Upload Libraries
- **Step2 ->** Prepare Data
- **Step3 ->** Creat DESeq Object
- **Step4 ->** Add Annotation File
- **Step5 ->** Visualisation
- **Step6 ->** Differential Gene Analysis
- **Step7 ->** Differential Gene Annotation and Summary
- **Step8 ->** Further Analysis

- Generated various plots - heatmap, variable genes heatmap, PCA, MA Plot, Volcano Plot, Volcano Plot with Significant genes.

Results & Discussion

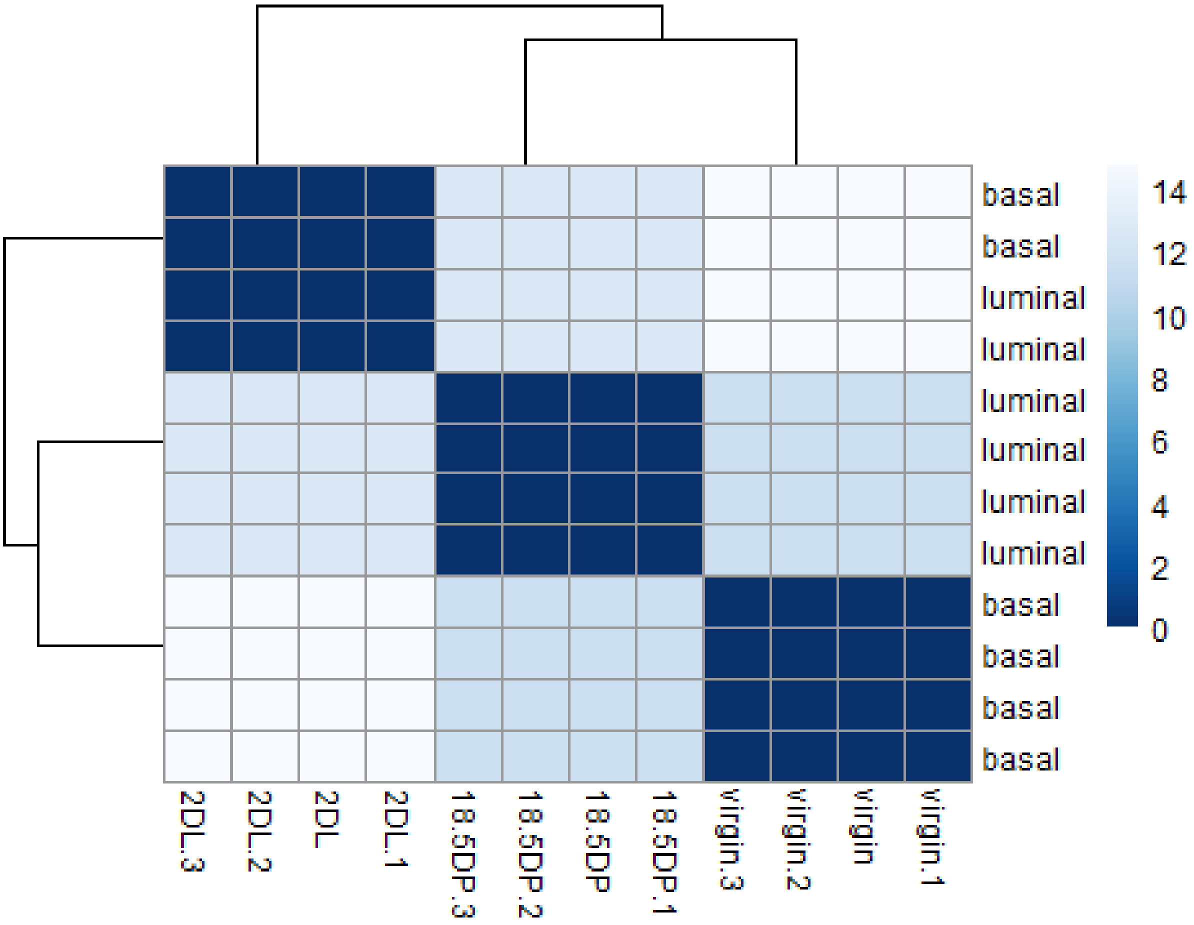
1



Principal Component Analysis (PCA):

- The PCA plot shows sample clustering along PC1 (explaining 65% of variance) and PC2 (explaining 35% of variance).
- Basal and luminal groups are clearly distinct, meaning the two subtypes have significantly different gene expression profiles.
- The separation along PC1 suggests that basal and luminal groups are fundamentally different at the transcriptional level.

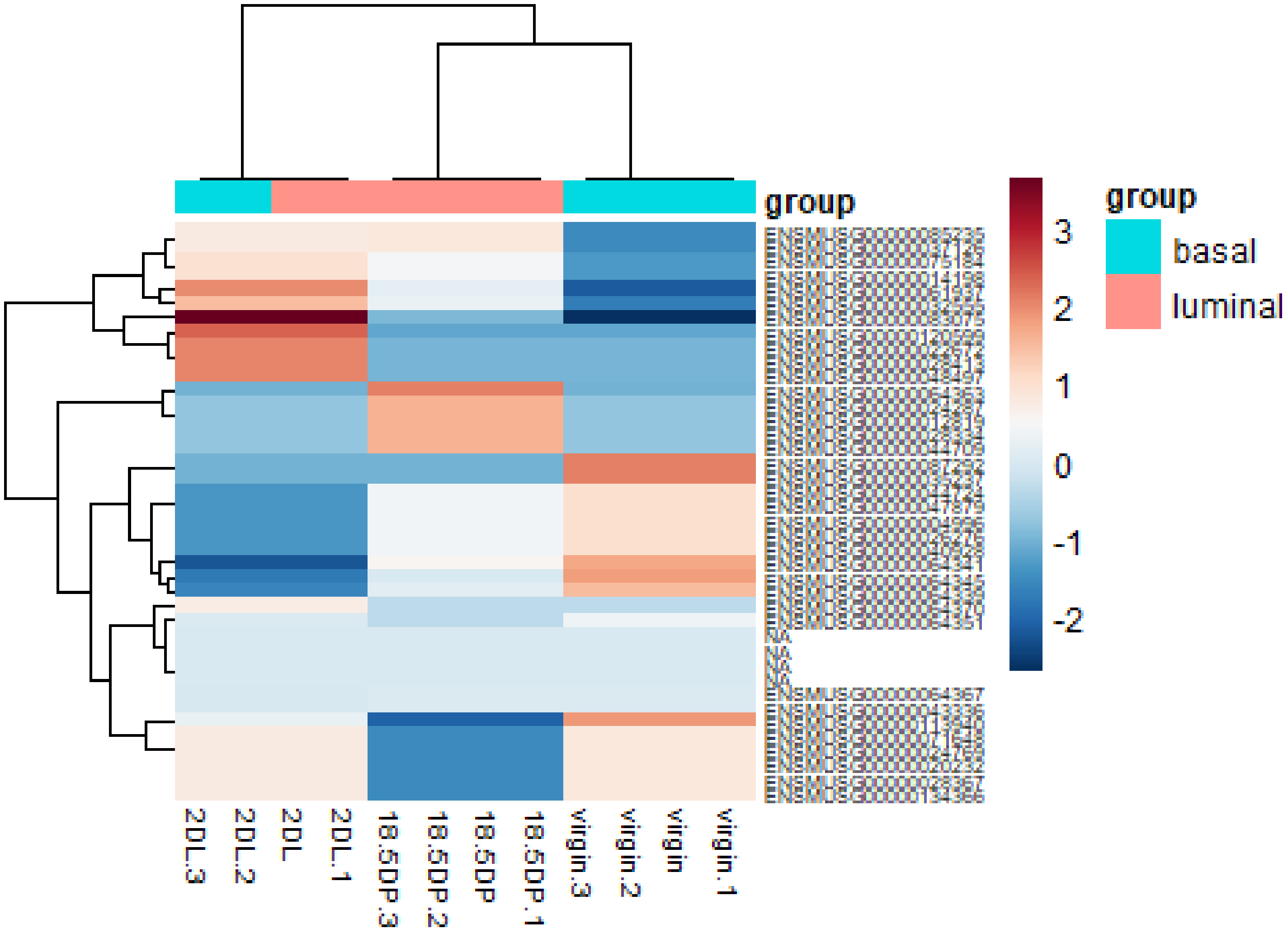
2



Sample-to-Sample Distances:

- A heatmap illustrates distances between samples, suggesting consistency within groups (e.g., luminal vs basal)
- Dark blue represents smaller distances (higher similarity between samples).
- Light blue/white represents larger distances (lower similarity between samples).
- Basal samples (on the right) cluster together.
- Luminal samples (in the middle) form a separate cluster.

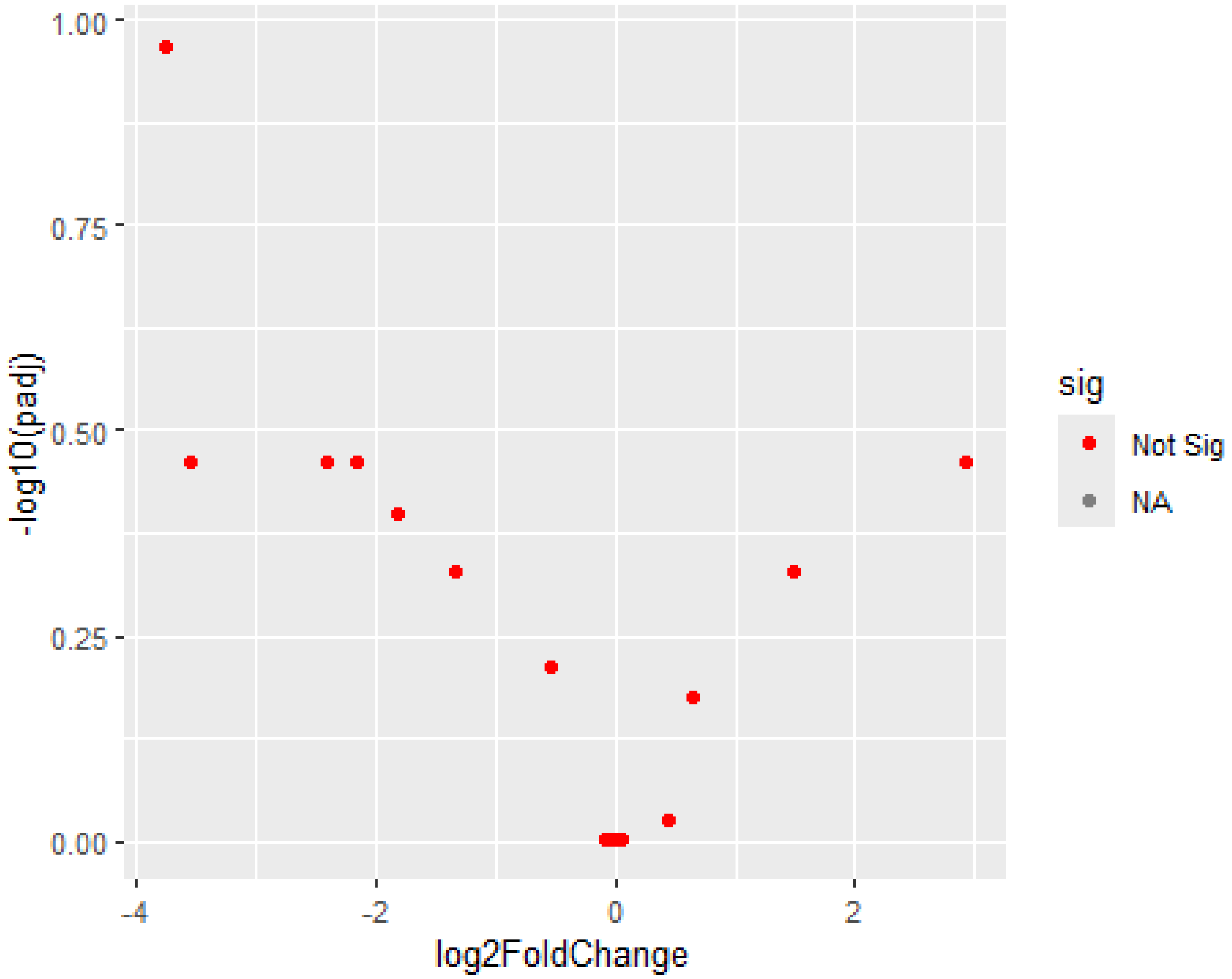
3



Variable Gene Heatmap:

- The color scale ranges from blue (low expression) to red (high expression).
- The dendrogram on the left shows hierarchical clustering, indicating similarity between gene expression patterns.

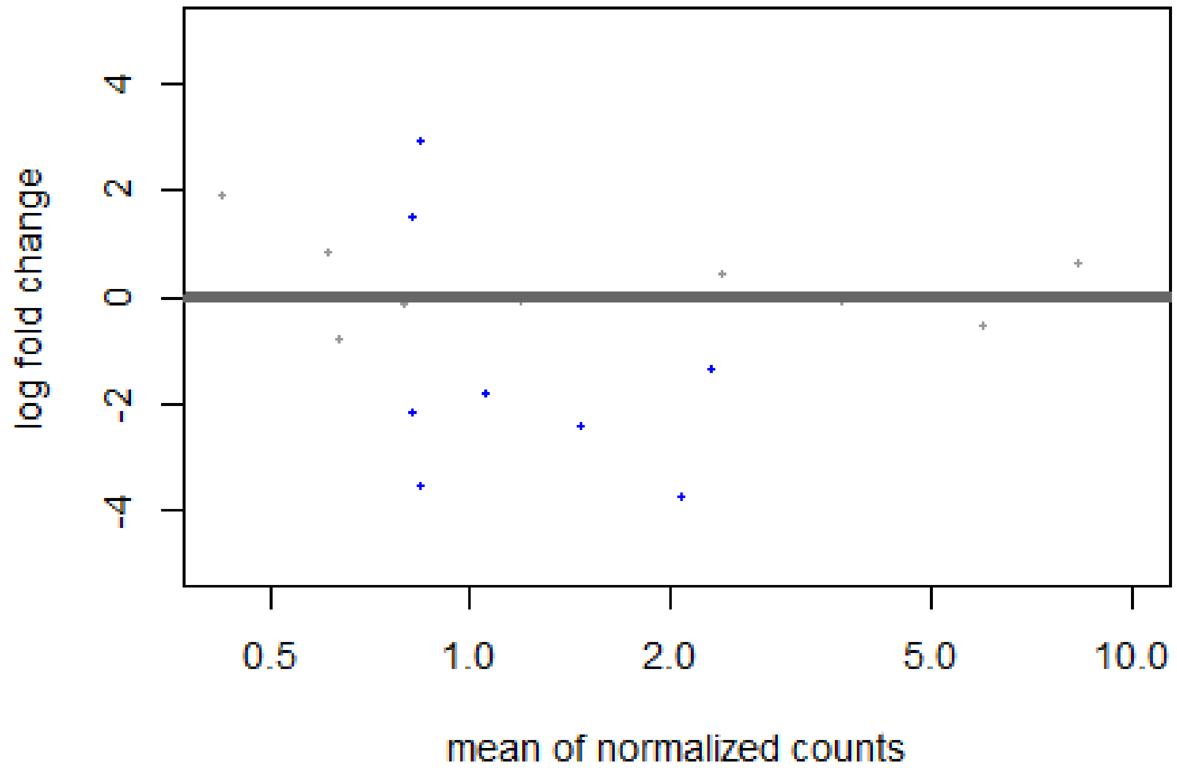
4



Volcano Plot:

- The x-axis represents the log2 fold change (log2FC), indicating how much a gene's expression has increased or decreased. The y-axis represents the negative log10 of the adjusted p-value (-log10(padj)), measuring statistical significance.
- The volcano plot suggests that there are no significantly differentially expressed genes in the dataset.

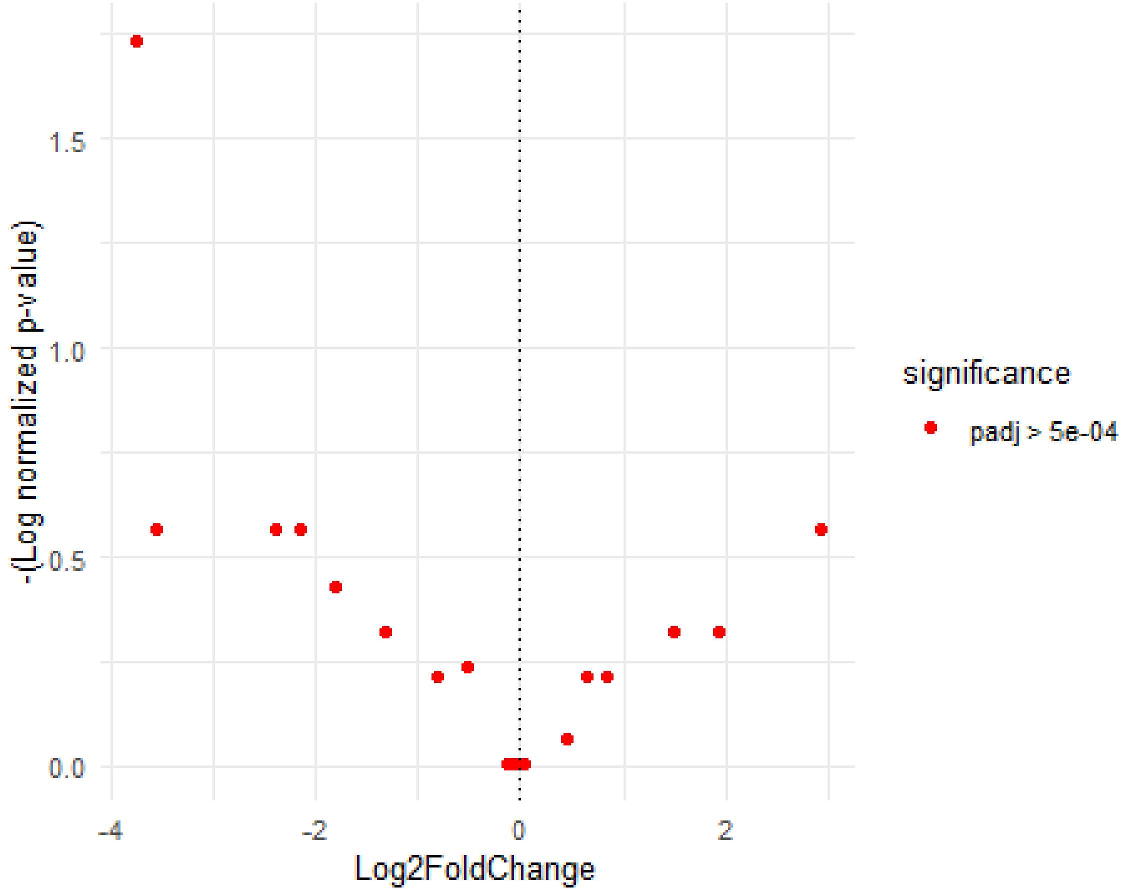
5



MA Plot:

- The MA plot visualizes the log2 fold changes (M) against the mean normalized counts (A).
- X-axis: Mean of Normalized Counts
  1. Represents the average expression level of genes across samples.
  2. Genes with low expression are on the left, and genes with high expression are on the right.
- **Y-axis: Log Fold Change**
  1. Represents the magnitude of change in gene expression.
  2. Points above 0 indicate upregulated genes, and points below 0 indicate downregulated genes.

6



Volcano Plot for significant genes:

- The data points are evenly spread around log2FC = 0, suggesting a mix of upregulated and downregulated genes.
- No significantly differentially expressed genes based on the chosen cutoff.
- Negative values indicate downregulated genes.
- Positive values indicate upregulated genes.