mRNA sequencing Analysis Report

Dr. Harsh Agrawal

IISER THIRUVANANTHAPURAM

15-10-2025

Table of Contents

Executive Data Analysis Summary

We analyzed a total of 4 Mouse samples, as shared by the client. Samples and Grouping information are provided in Table 1 and Table 2. The raw reads were filtered using Cutadapt for quality scores and adapters. Filtered reads were aligned to the Mouse genome (GRCm39.primary_assembly.genome) using splice aware aligner STAR to quantify reads mapped to each gene. Percentage of uniquely aligned reads ranged between 60% – 80% across all the samples. The aligned data quality check was performed using Qualimap. rRNA contamination was screened using RSeqQC package. Total number of uniquely mapped reads were counted using FeatureCounts. The uniquely mapped reads were then subjected to differential gene expression using DESeq2 in R. Total number of differentially expressed genes are given in the Table 5 for the respective comparisons.

Methods – Data Analysis

Bioinformatics Pipeline

The following bioinformatics steps were performed for analysis of the data

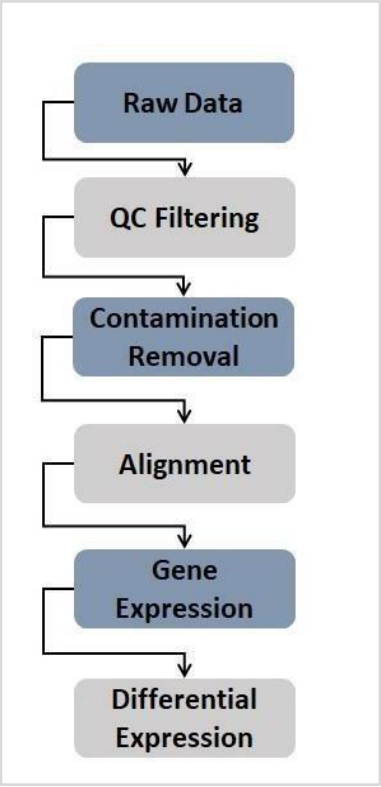


Fig 1. Bioinformatics workflow for RNA seq analysis

Bioinformatic Steps

Read quality check

The following parameters from raw fastq files were checked using fastqc tool (version – 0.11.9) as part of quality check- Base quality score distribution, Sequence quality score distribution, Average base content per read, GC distribution in the reads, PCR amplification issue, Check for over-represented sequences and Adapter trimming.

Based on the quality report of fastq files trimming on raw read was performed to only retain high quality sequence for further analysis. In addition, the low-quality sequence reads were excluded from the analysis. The adapter trimming was performed using Cutadapt (Version 4.9).

Read alignment

The paired-end reads are aligned to the reference Mouse genome (GRCm39. primary_assembly.genome). Alignment of reads was performed using STAR (version - 2.7.11b)

Alignment quality screening

Aligned reads were analyzed for their quality check related statistics; against both mapping quality as well as read alignment distribution against reference transcriptome features using Qualimap. Factors like read distribution, Alignment distribution as well as splice junction distribution were analyzed separately. rRNA contamination was screened using rSeqQC package.

Expression estimation

The aligned reads were used for estimating expression of the genes. The raw read counts were estimated using FeatureCount (version - 2.0.8) across all samples

The raw read count across all samples were then normalized using DESeq2 package in R.

Differential expression analysis

The raw read count data were normalized using DESeq2. The ratio of normalized read counts for HT over Non-HT was taken as the fold change. Genes were first filtered based on the p-value (<= 0.05) for statistical significance. Those genes which were found to have log2(foldchange) <= -2

and log2(foldchange) >= 2 were considered as downregulated or upregulated, respectively.

Functional Gene Set Enrichment Analysis

Functional gene set enrichment analysis was performed for KO vs WT comparison, using the gsea method under clusterprofiler R package, against ontologies such as Biological Process(BP), Molecular Function(MF), Cellular Component(CC) and Reactome Pathway. This provides the activation/suppressed status of the ontology/pathway, in KO group relative to WT group.

Samples and Grouping information

Following samples were analyzed.

Table 1. Samples Information

Sl.no	**Sample ID**	Sample Type
1	WT1	WT
2	WT2	WT
3	KO1	KO
4	KO2	KO

Table 2. Grouping information

Comparision	Grouping Type
1.	KO vs WT

Results

Data QC Summary

Table 3 summarizes the overall data generated and the average read quality observed across the reads.

Table 3. Raw Data QC Summary

Sample Name	Total paired reads (Before adapter trimming)	Data Generated (Gb)	Total paired reads (After adapter trimming)	Percentage of reads retained after trimming(%)	Avg.base quality (phred) after trimming	Total data (reads)>=Q30% after trimming	% of Reads belonging to Mouse(GRCm39)
WT1	9377848	2.83	9318397.00	99.37	39.46	99.60	80.45
WT2	11721292	3.54	11617198.00	99.11	39.33	99.39	78.81
KO1	4758908	1.44	4716386.00	99.11	39.35	99.47	90.83
KO2	13280018	4.01	13151845.00	99.03	39.29	99.40	95.76

Data generation for each sample was in the range of 2 GB to 4 GB and the average % of reads with quality > phred score 30 is ~99%.

Alignment Summary

Samples were aligned to Mouse reference genome. The mapping statistics showed the alignment of uniquely mapped reads were at the range of 87-89%. Unaligned read were at the range of 3-7% and multimapped reads ~7.5%. The below Table 4 summarizes the alignment summary of the sequenced reads.

Sample ID	Uniquely mapped reads %	Exonic (%)	Intronic (%)	Intergenic (%)	Rseqc rrna estimatation (%)	Known Splicing Junctions (%)	Partly Known Splicing Junctions (%)	Novel Splicing Junctions (%)
WT1	77.23	15.04	47.99	36.97	2.32	46.00	6.00	46.00
WT2	87.45	8.87	47.95	43.19	1.42	50.00	5.00	43.00
KO1	65.17	35.12	50.65	14.23	6.06	46.00	6.00	47.00
KO2	63.29	35.29	51.90	12.81	5.24	47.00	6.00	46.00

Table 4. Read Alignment Summary

Sample dimensional reduction and Clustering

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical technique used to identify global patterns in high-dimensional datasets. It is commonly used

to explore the similarity of biological samples in RNA-seq datasets. To achieve this, gene expression values are transformed into Principal Components (PCs), a set of linearly uncorrelated features which represent the most relevant sources of variance in the data, and subsequently visualized using a scatter plot.

The scatter plot of the first two principal components (pcs) of the data for given samples are shown in Figure 2. Each point represents the sample. Samples with similar gene expression profiles should come closer in the two-dimensional space.

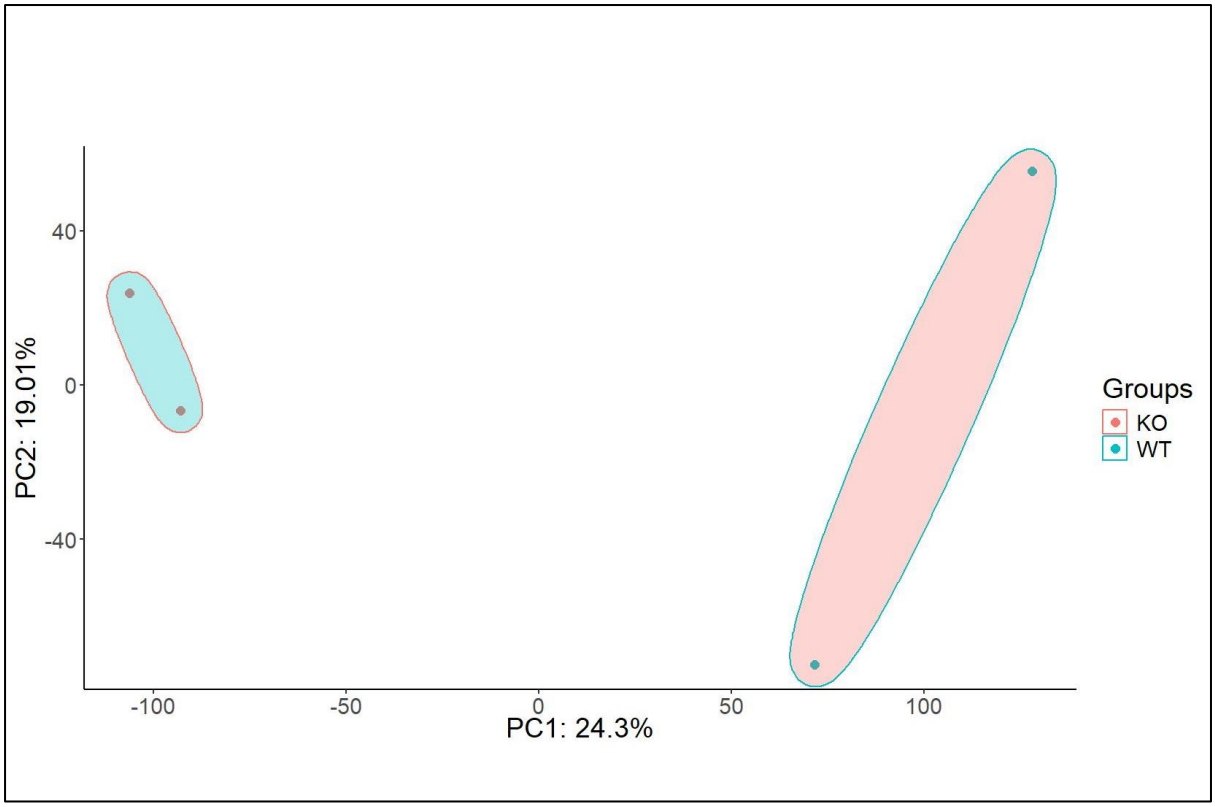


Figure 2. Represents the scatter plot of the first two principal components of the data for given samples.

Hierarchial Clustering

Clustering is a statistical technique used to identify the similarity between data. We used hclust R function together with average linkage and euclidean distance to perform clustering, samples whose gene expression are similar try to cluster together as shown in Figure 3.

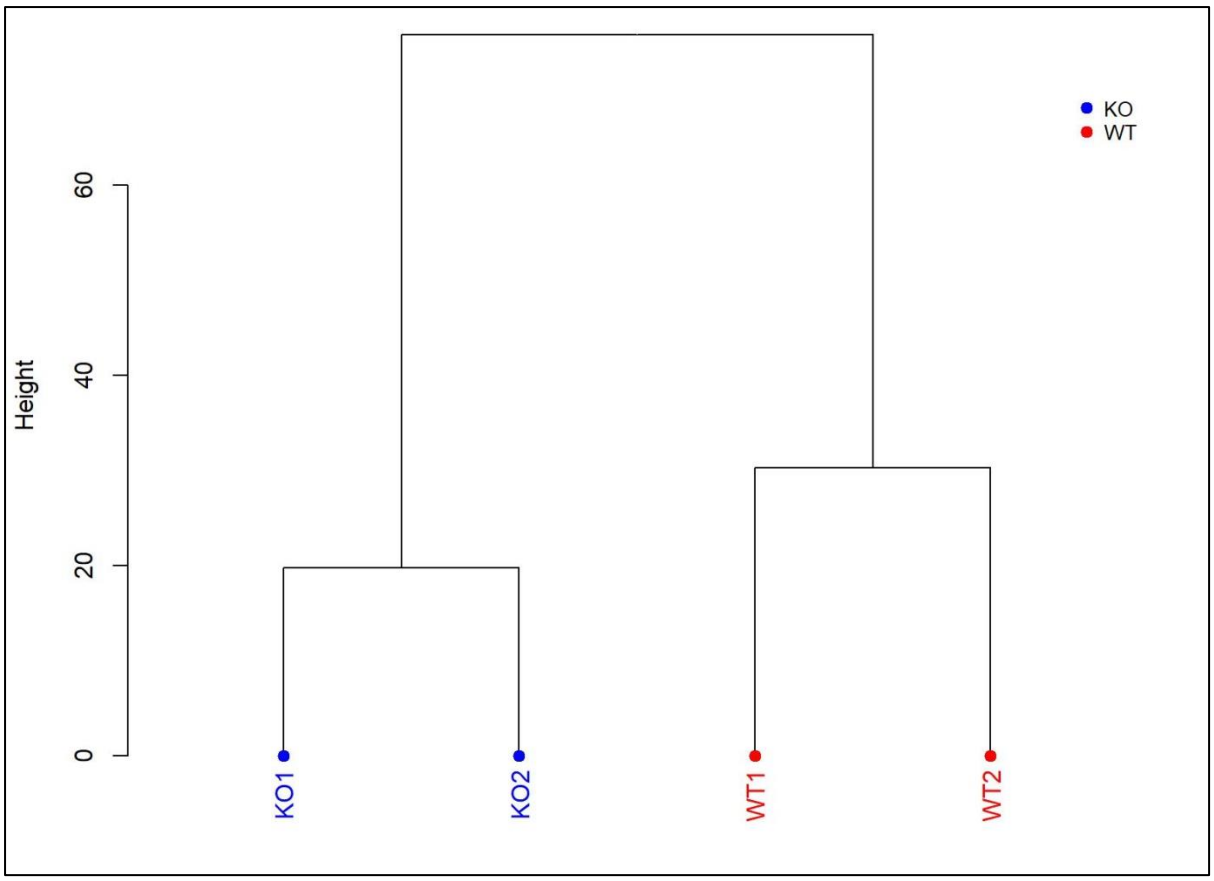


Figure 3. Represents the hierarchical clustering of the data for given samples.

Ideally, samples should cluster according to the group i.e Samples belonging to the same group must be clustered together.

Sample Correlation

A sample-sample correlation heatmap shows how similar samples are based on their gene expression. High correlation among replicates indicates similarity, while low correlation can reveal outliers and clustering of samples based on their type determines their similarity.

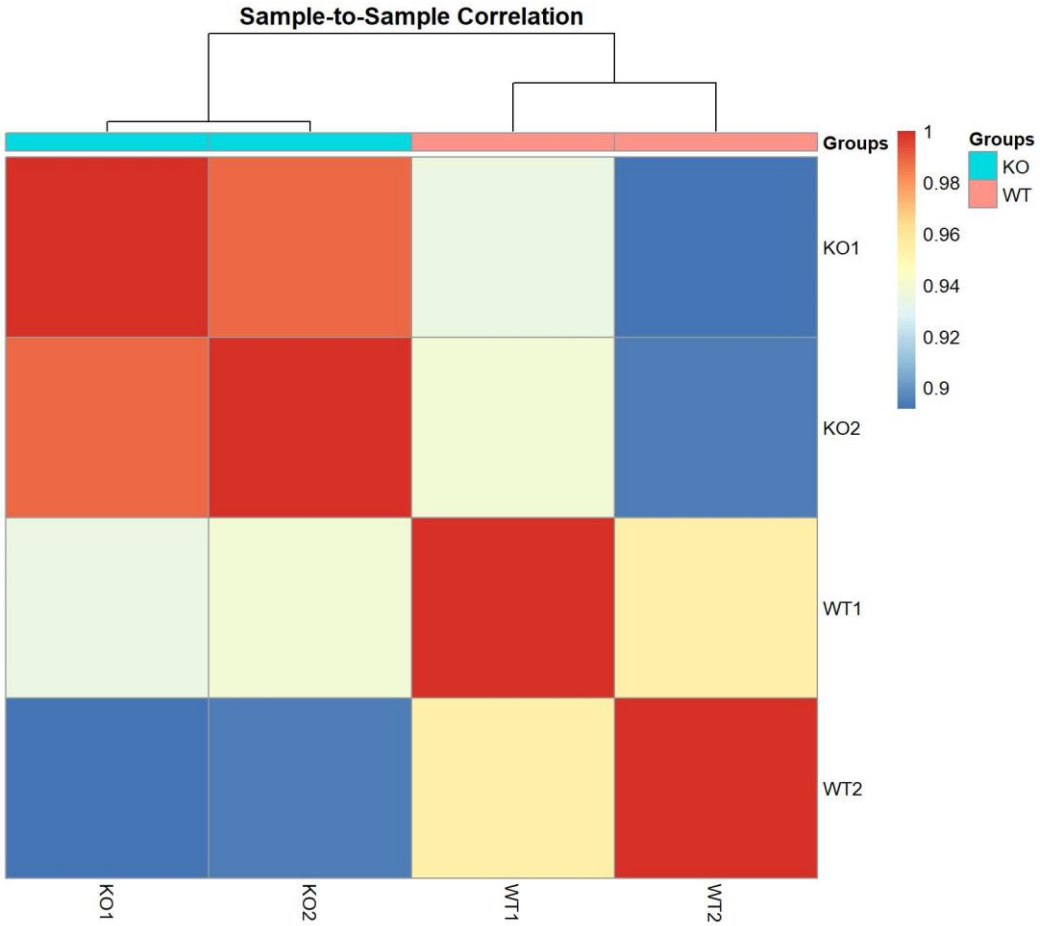


Figure 4 : represent correlation of samples along with its replicates

Differential Gene Expression (DGE) Analysis

Gene expression signatures are alterations in the patterns of gene expression that occur due to cellular perturbations such as drug treatments, gene knockdown or diseases. They can be quantified using differential gene expression (DGE) methods, which compare gene expression between two groups of samples to identify genes whose expression is significantly altered in the perturbation. The signature table is used to display the results of such analyses.

Table 5. Significant differentially expressed genes summary

*Significant – P-value <= 0.05 and Log2Fc ≥ 2 and ≤ -2

Comparisions	Upregulated Genes	Downregulated Genes
KO vs WT	3	6734

Representation of DEGs as plots

Volcano plots

Volcano plots are a type of scatter plot commonly used to display the results of a differential gene expression analysis. They can be used to quickly identify genes whose expression is significantly altered in a perturbation, and to assess the global similarity of gene expression in two groups of biological samples. Each point in the scatter plot represents a gene; the axes display the significance versus fold-change estimated by the differential expression analysis. Green points indicate significantly downregulated and red points indicate significantly upregulated genes and grey ones represent statistically non-significant genes.

Every dot in the plot represents a gene. Green points indicate significantly downregulated genes and red points indicate significantly upregulated genes.

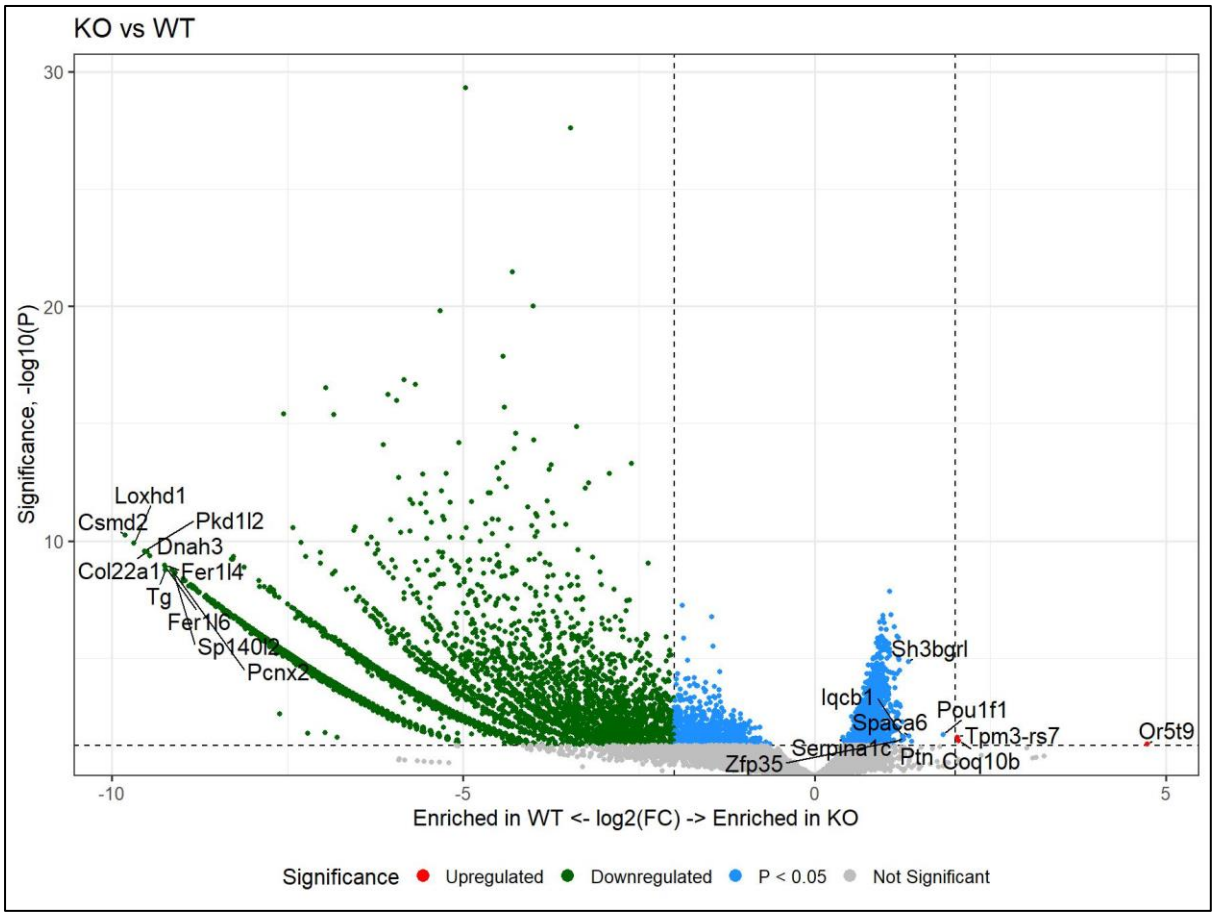


Figure 5: Volcano plot for the given comparison which displays log2 -fold change and statistical significance (p-value*) of each gene calculated based on differential gene expression analysis .

Functional Gene Set Enrichment Analysis

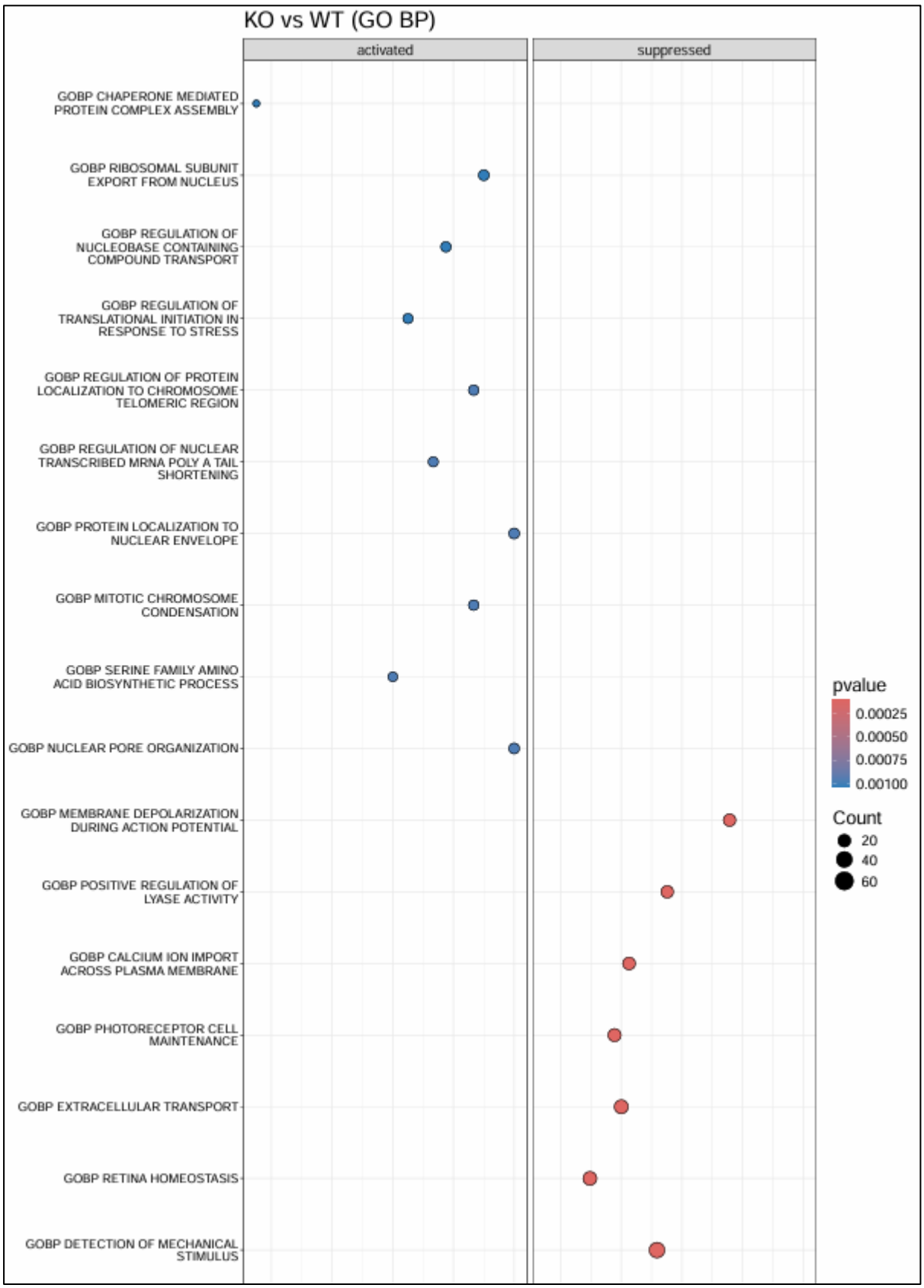
Functional regulation caused by the deregulated genes across different groups of samples was analyzed using the statistical technique, Gene set enrichment analysis (GSEA) which provides the status of regulation (activation or suppression) of the function, per comparison based on the status of regulation of genes involved in the particular function.

These analysis can be performed using overlap with public functional databases like REACTOME, KEGG, Geneontology.org etc.

Currently the GSEA method of pathway enrichment analysis have been performed for each statistical comparison performed, against the Gene ontologies (Biological process, Cellular Component, Molecular function) and REACTOME pathway database. This helps to understand the top Activated & Suppressed pathways within each comparison.

Pathways that are significantly enriched in a gene list are shown as Barplot. Analysis of these functional categories revealed the association of the differentially expressed genes in activation and suppression of multiple functions/pathways.

a)


KO vs WT (GO BP)
activated
suppressed
GOBP CHAPERONE MEDIATED PROTEIN COMPLEX ASSEMBLY
GOBP RIBOSOMAL SUBUNIT EXPORT FROM NUCLEUS
GOBP REGULATION OF NUCLEOBASE CONTAINING COMPOUND TRANSPORT
GOBP REGULATION OF TRANSLATIONAL INITIATION IN RESPONSE TO STRESS
GOBP REGULATION OF PROTEIN LOCALIZATION TO CHROMOSOME TELOMERIC REGION
GOBP REGULATION OF NUCLEAR TRANSCRIBED MRNA POLY A TAIL SHORTENING
GOBP PROTEIN LOCALIZATION TO NUCLEAR ENVELOPE
GOBP MITOTIC CHROMOSOME CONDENSATION
GOBP SERINE FAMILY AMINO ACID BIOSYNTHETIC PROCESS
GOBP NUCLEAR PORE ORGANIZATION
GOBP MEMBRANE DEPOLARIZATION DURING ACTION POTENTIAL
GOBP POSITIVE REGULATION OF LYASE ACTIVITY
GOBP CALCIUM ION IMPORT ACROSS PLASMA MEMBRANE
GOBP PHOTORECEPTOR CELL MAINTENANCE
GOBP EXTRACELLULAR TRANSPORT
GOBP RETINA HOMEOSTASIS
GOBP DETECTION OF MECHANICAL STIMULUS
pvalue
0.00025
0.00050
0.00075
0.00100
Count
20
40
60

b)

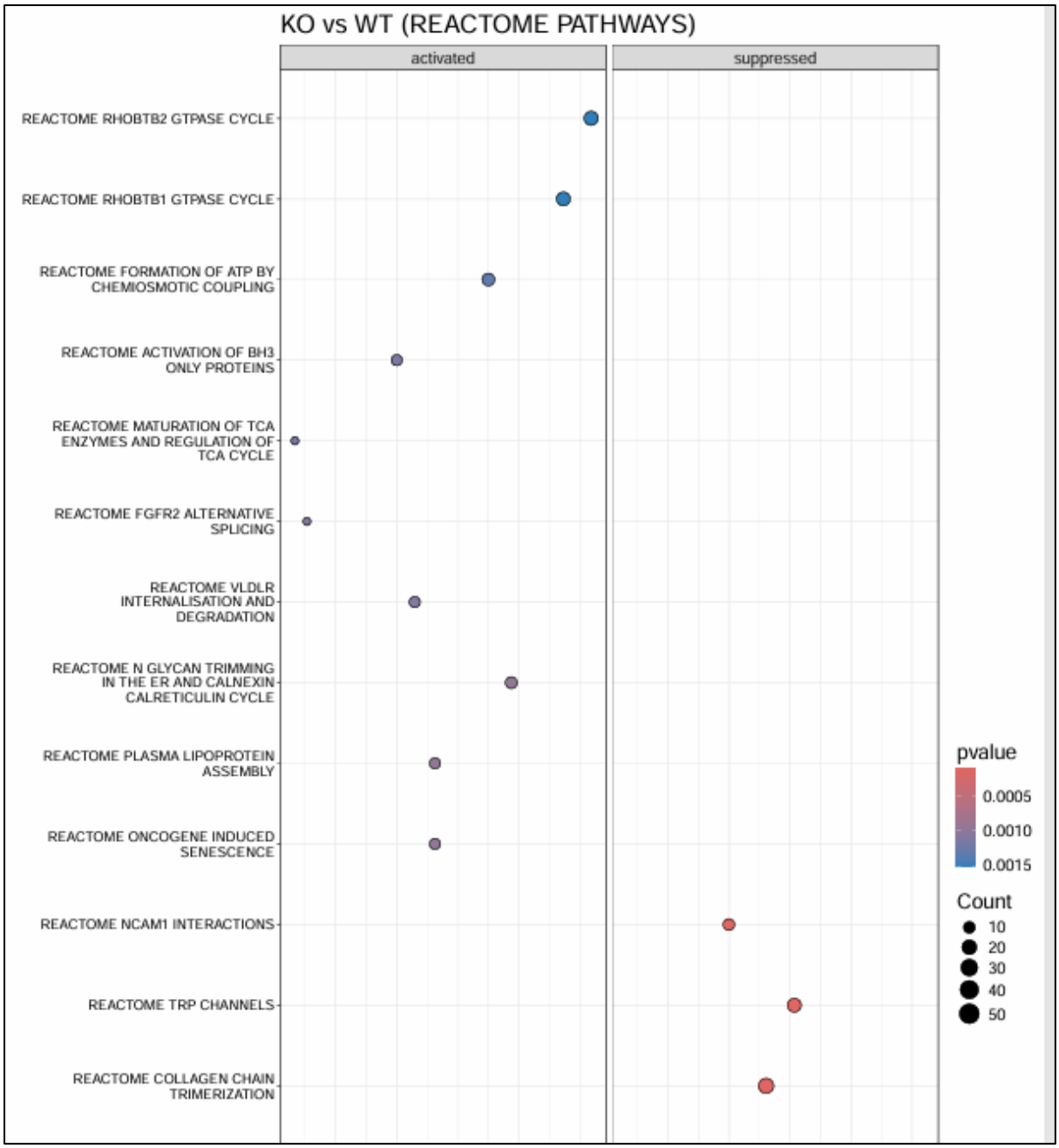


Figure 6: Represent Functional over-representation analysis of KO vs WT Comparision a) Top 20 Activated and Suppressed Biological process b) Top 20 Activated and Suppressed Reactome pathways

*For other functional categories the gradient bubble plots have been provided as supplementary information.

Kegg Pathview

Kegg pathways such as B cell receptor signaling pathway and Hematopetic lineage pathway from KEGG database has been visualized using significant genes (pvalue < 0.05) from KO relative to WT.

a)

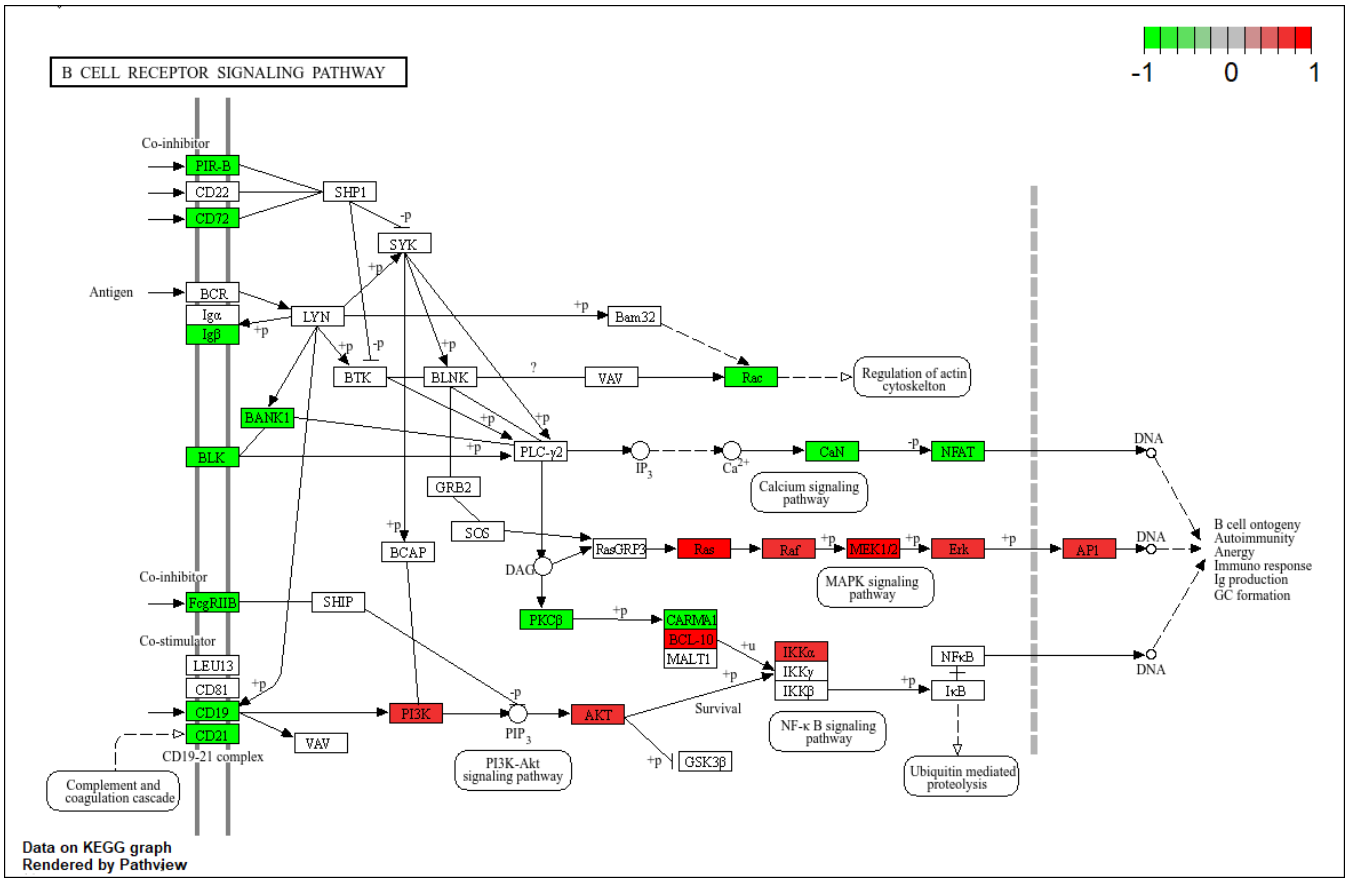


Figure 8: Kegg pathview visualization for a) B cell receptor signaling pathway

b)

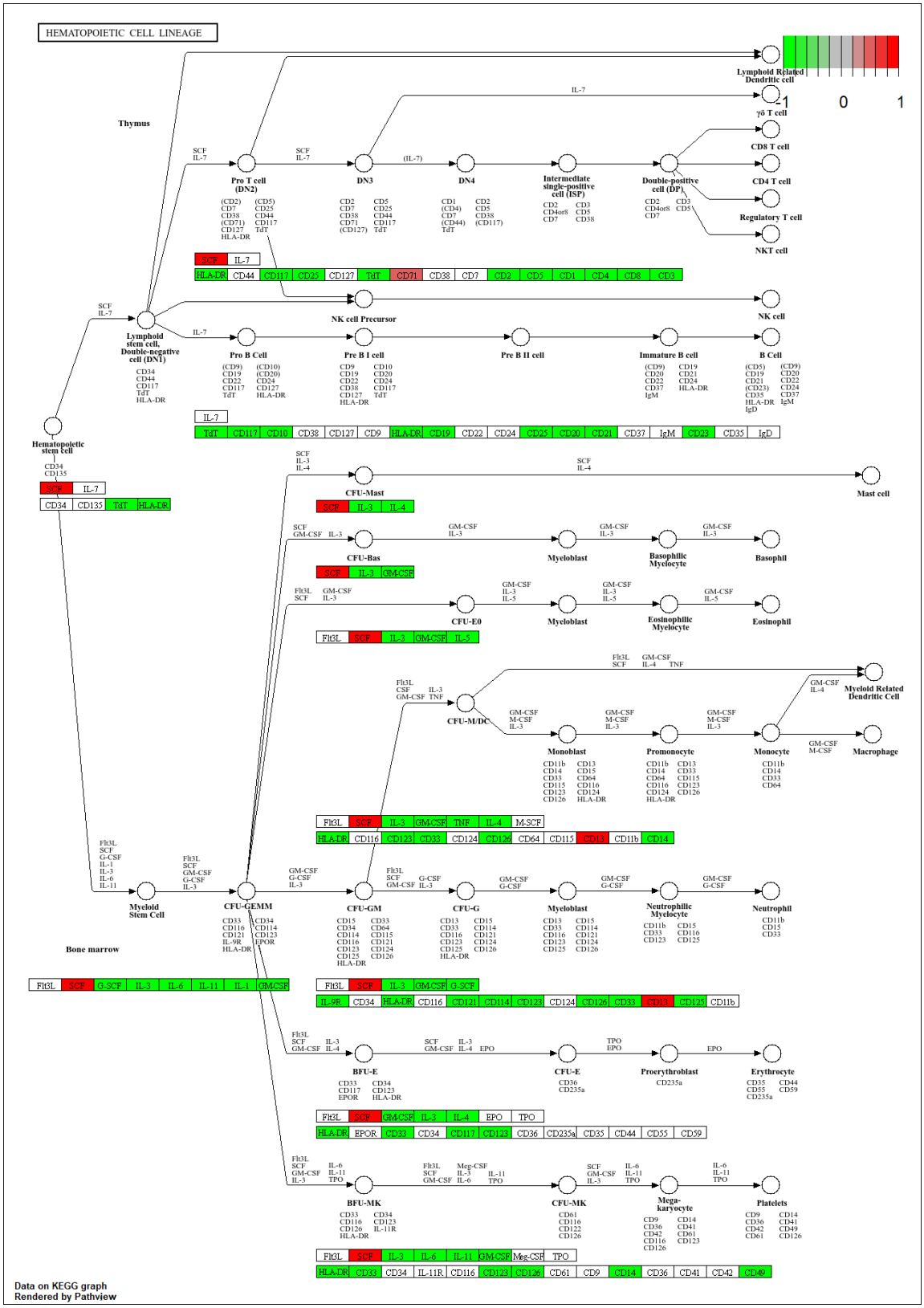


Figure 8: Kegg pathview visualization for b) Hematopoietic system

Deliverables

1. Raw Data QC and Raw Data counts across all samples.
2. Mapping summary and related QC statistics for read distribution, mapping distribution and splice junction distribution.
3. Normalized read count for each samples along with Sample Grouping/Clustering and QC related plots : PCA plot, Heatmap based on top 2000 variant genes, dendrogram and Sample-Sample correlation plot.
4. Differential gene expression reports for each comparison along with corresponding visualization – Volcano plots and Heatmap visualizing top 50 up and downregulated genes.
5. Functional gene set enrichment analysis reports per comparison, along with corresponding visualization – Gradient Bubble charts for top 10 significant activated and suppressed functions (Gene ontology Biological Process, Gene ontology Molecular Function, KeggPathways).
6. Kegg pathview for selected pathway visualization.

Access to all reports:

https://workdrive.zohoexternal.com/external/81ca515b1cb7261f39705c2b887f60d8c095f20897dafa864dfa6a16843ad314

****