Exploring Transcriptomic Alterations in Human Respiratory Cells Post SARS-CoV-2 Infection : A Differential Gene Expression and Enrichment Study

Objective: To align, quantify, and construct transcriptome profiles for control and SARS-CoV-2 infected respiratory cells at each time point (24 and 72 hours), followed by the identification of differentially expressed genes (DEGs) between time points (24H vs. 72H) and conditions (control vs. SARS-CoV-2). Additionally, to perform Gene Ontology (GO) term enrichment analysis on the identified DEGs to uncover key biological processes influenced by SARS-CoV-2 infection and time progression.

Data: Data for this project is available through the National Center for Biotechnology Information (NCBI) under the accession BioProjectID: PRJNA901149.

Data Preprocessing: The files are single-ended, the .sra format files were converted to .fastq format using fastq-dump for all sample files.

Workflow:
FastQC was used to analyze the quality of the reads, and adapter trimming was performed using Cutadapt with the specified adapter sequence. Following trimming, HISAT2 was used to map the reads to the reference genome. SAMtools was then employed to convert the .sam format files to .bam format for more efficient storage and downstream analysis. For counting features, Subread was utilized. Initial data preprocessing was conducted using Python's Pandas library, and finally, DESeq2 was used for differential gene expression analysis.

Step	**Tool Used**
1. Data Conversion	fastq-dump
2. Quality Control	FastQC
3. Adapter Trimming	Cutadapt
4. Mapping	HISAT2
5. SAM to BAM Conversion	SAMtools
6. Feature Counting	Subread (featureCounts)
7. Initial Data Preprocessing	Python (Pandas)
8. Differential Expression Analysis	DESeq2 (R)
9. GO Enrichment Analysis	clusterProfiler (R)

FASTQC: The FastQC report indicated that both the per base sequence quality and per sequence quality scores were within acceptable ranges across all files. However, the report revealed the presence of Illumina Universal Adapter content in all samples, suggesting a need for adapter trimming.

ADAPTER TRIMMING: Based on the FastQC analysis, the Illumina Universal adapter sequence was trimmed from all files using the Cutadapt tool. After trimming, FastQC was rerun on the processed FASTQ files to confirm the removal of adapter content. The resulting FastQC reports confirmed that the trimmed files were free from adapter contamination and suitable for further downstream analysis.

MAPPING The human reference genome, GRCh38, was downloaded from Ensembl and indexed before mapping. HISAT2, a widely-used tool for mapping transcriptome reads, was employed to align the reads to the reference genome. SAM-tools was then used to convert files from .sam format to .bam format to facilitate downstream analysis.
The mapping statistics for the samples indicate a very good alignment rate to the reference genome, with overall mapping percentages ranging from approximately 70% to 80% across samples. Most samples have alignment rates above 72%, which is considered satisfactory for miRNA-seq experiments.

Primary mapped reads (unique alignments) typically comprise around 50-61% of the total reads, while a few samples have lower mapping percentages, around 22%. These mapping percentages suggest good quality alignment, indicating that a significant portion of the reads correspond well to the reference genome. Moving forward, these mapped reads will be used for feature counting, which is the next step in the analysis.

FeatureCounts: Using Subread's featureCounts to quantify the number of reads mapped to each gene based on GRCh38 gene annotation file. Key parameters, such as strand specificity and alignment quality filtering, were set to ensure accurate and reproducible counts per gene. The output of featureCounts is the matrix containing each gene's read count across samples, serving as the input for DESeq2 in differential expression analysis.

Differential Expression of Genes: DESeq2 package in R was used for identifying differentially expressed genes using the count matrix generated from the Subread.

Results

- **Volcano Plots** The volcano plots illustrate the distribution of differentially expressed genes (DEGs) between two experimental conditions by plotting the statistical significance against the magnitude of expression changes.

72H vs 24H

The volcano plot comparing the 72-hour (T72) and 24-hour (T24) time points (Fig1) genes are plotted based on their $\log_2$ fold change (x-axis) and $-\log_{10}$ p-value (y-axis). Genes highlighted in red indicate both a statistically significant p-value (typically $p < 0.05$) and a substantial $\log_2$ fold change, suggesting strong differential expression between the two time points. Upregulated genes (positive $\log_2$ fold change) appear on the right side, while downregulated genes (negative $\log_2$ fold change) appear on the left. There are both upregulated and downregulated genes. However, a notable observation is the presence of more downregulated genes (genes with negative $\log_2$ fold change on the left side of the plot) compared to upregulated genes. This suggests that, over time, certain genes may have decreased their expression levels, potentially indicating regulatory processes that may be turning off or downregulating specific pathways or responses after the initial exposure period.

SARS-Cov2 vs Mock

The volcano plot, compare gene expression between SARS-CoV-2-infected cells (SARS-CoV-2) and control cells (Mock) (Fig-2). The x-axis represents the $\log_2$ fold change, with upregulated genes in the infected condition on the right and downregulated genes on the left. The y-axis shows the $-\log_{10}$ p-value, highlighting genes with statistical significance. We see a relatively balanced distribution of upregulated and downregulated genes, but there is a slight tendency towards more downregulated genes in the SARS-CoV-2 condition. Genes marked in red are those that have significant differential expression, indicating strong changes in response to SARS-CoV-2 infection. These genes could be crucial in understanding the host's cellular response to the virus.

Both the 72H vs 24H and SARS-CoV-2 vs Mock comparisons show significant differential gene expression. The 72H vs 24H plot has more downregulated genes over time, suggesting reduced activity in certain genes, while the SARS-CoV-2 vs Mock plot shows a slight bias toward downregulated genes, indicating gene inactivation in response to infection.

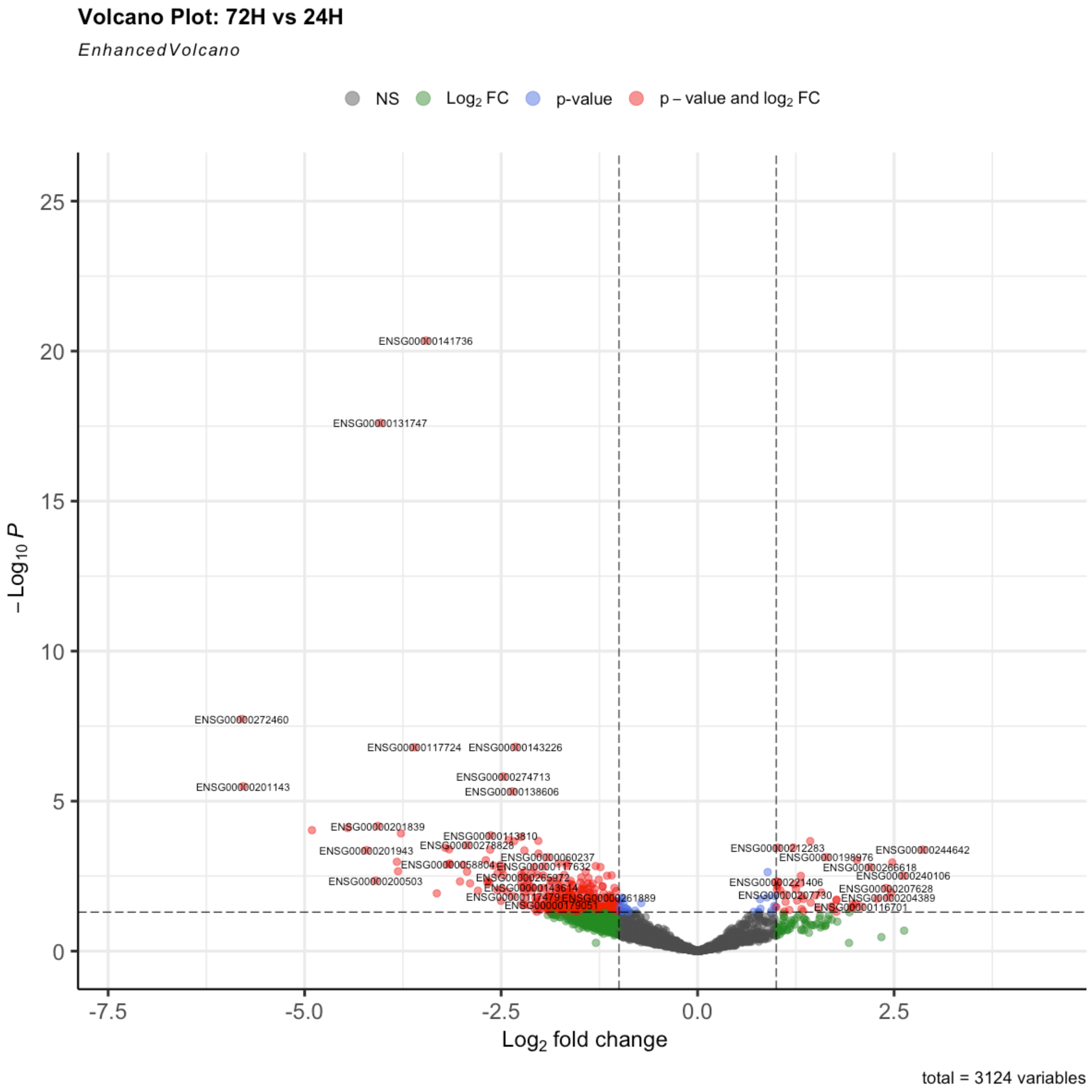

Fig-1: Volcano Plot: 72H vs 24H

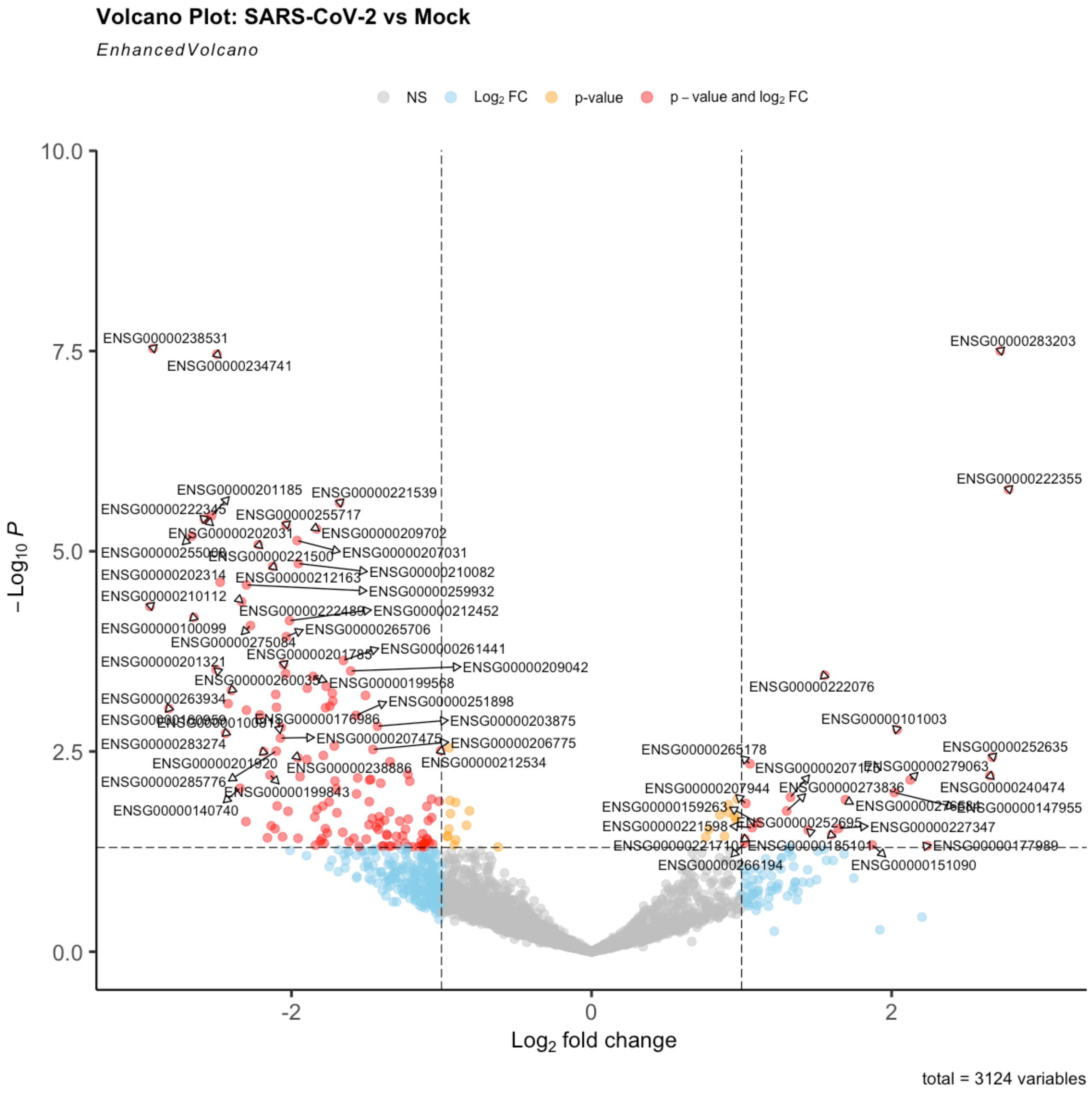

Fig-2: Volcano Plot: SARS-CoV-2 vs Mock

MA plots An MA plot visualizes the relationship between the mean expression level (x-axis) and the log fold change (y-axis) of genes in two conditions. Each dot represents a gene, indicating its expression change between conditions. The significance of an MA plot lies in its ability to quickly identify genes with substantial fold changes in expression, highlighting differentially expressed genes.

The plot in Fig-3 shows changes in gene expression over time, with blue dots representing significant differentially expressed genes between the 72-hour and 24-hour time points. More downregulated genes (negative fold change) indicate reduced expression at the 72-hour mark. In Fig-4, the plot compares gene expression between infected (SARS-CoV-2) and control (Mock) cells, with blue dots marking genes that significantly respond to infection, showing a mix of upregulated and downregulated expression changes.

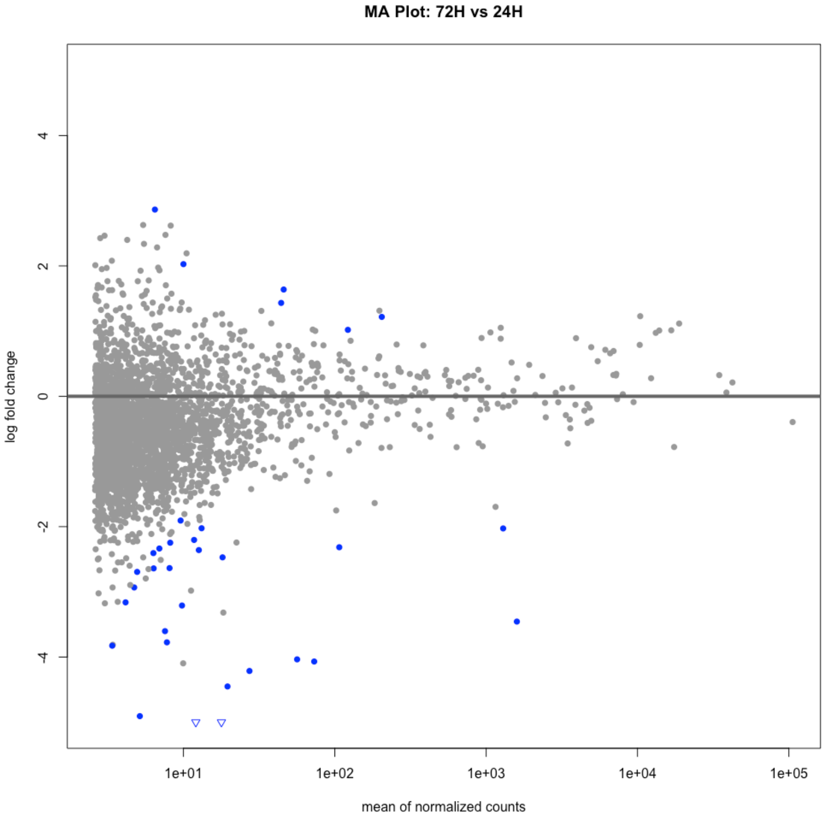
MA Plot: 72H vs 24H
log fold change
4
2
0
-2
-4
1e+01
1e+02
1e+03
1e+04
1e+05
mean of normalized counts

Fig-3

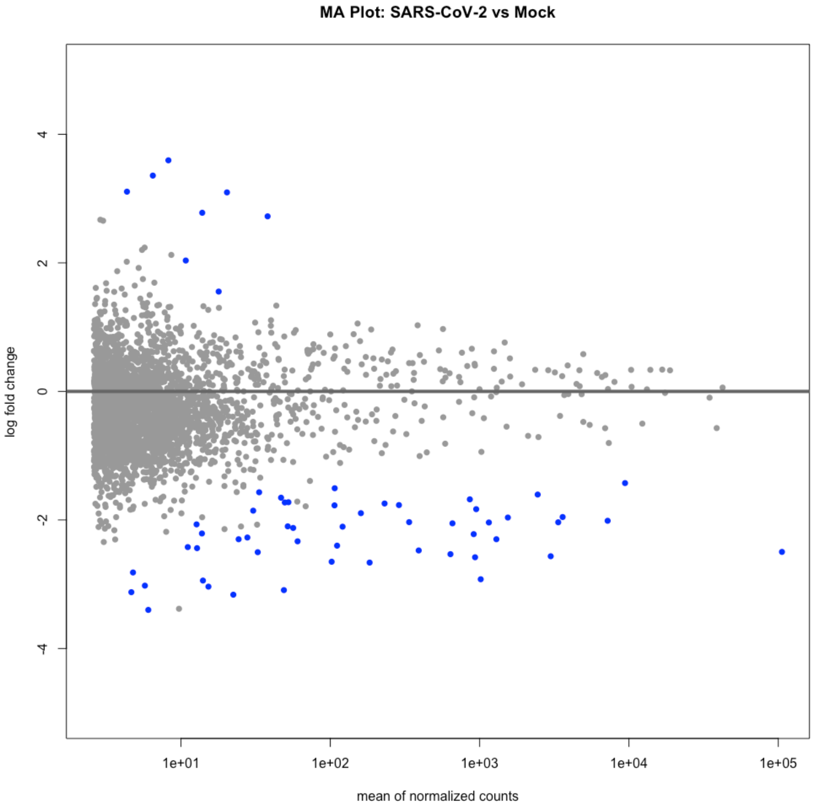
MA Plot: SARS-CoV-2 vs Mock
log fold change
4
2
0
-2
-4
1e+01
1e+02
1e+03
1e+04
1e+05
mean of normalized counts

Fig-4

- **PCA plot** This PCA (Principal Component Analysis) (Fig-5) plot visualizes the variance in gene expression data across different samples. Each point represents a sample, with colors distinguishing the experimental conditions (Mock vs. SARS-CoV-2 infection) and shapes representing the time points (T24 and T72)

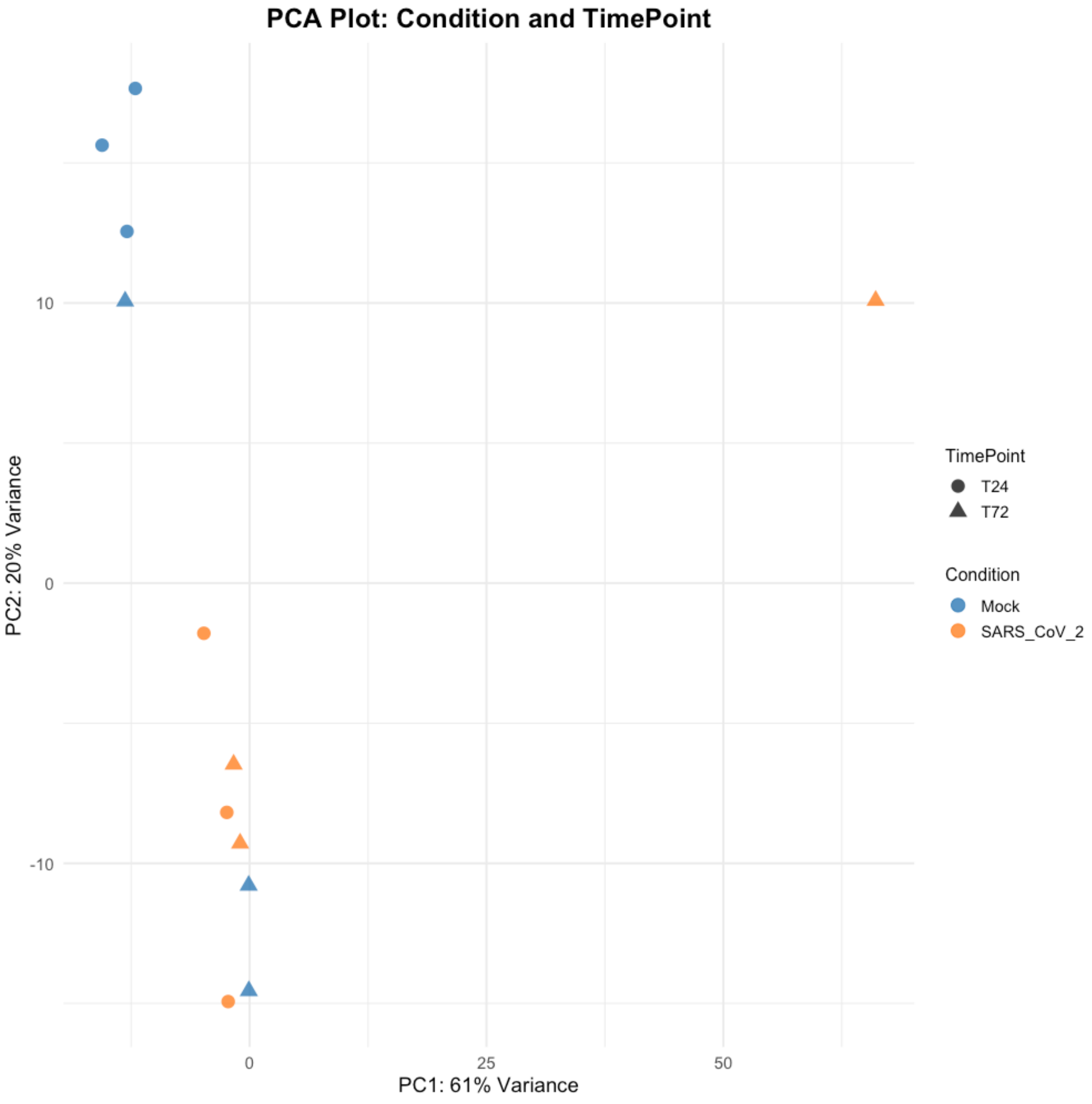

Fig-5

Each point represents a sample, with colours distinguishing the experimental conditions (Mock vs. SARS-CoV-2 infection) and shapes representing the time points (T24 and T72).
The x-axis (PC1) explains 61% of the variance, and the y-axis (PC2) explains 20% of the variance, meaning that these two components capture the majority of variation in the data.
Key observations:

- Samples tend to cluster based on condition (Mock vs. SARS-CoV-2), indicating clear differences in gene expression between infected and control samples.
- There is also a separation based on time points, with each time point (T24 and T72) showing a distinct pattern, suggesting time-dependent changes in gene expression.

GO Enrichment Analysis: GO (Gene Ontology) enrichment analysis is to identify biological processes, cellular components, or molecular functions that are significantly represented among a set of differentially expressed genes. GO enrichment analysis was performed using the enrichGO function from the clusterProfiler package in R, with the list of differentially expressed genes as input. The analysis was configured to identify enriched biological processes, with results filtered based on adjusted p-value thresholds for significance.

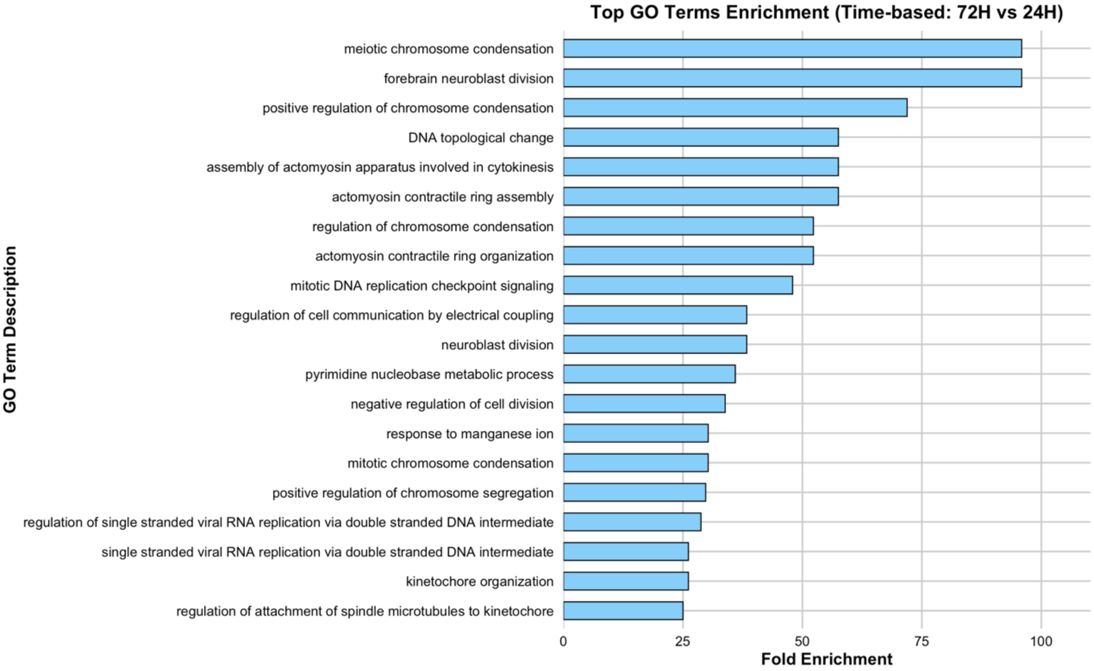

Fig-6

Time-based GO Term Enrichment (72H vs 24H): This (Fig-6) plot shows the top 20 GO terms enriched when comparing samples collected at 72 hours versus 24 hours.These terms highlight the biological processes that change over time post-infection, reflecting the temporal dynamics in the response.

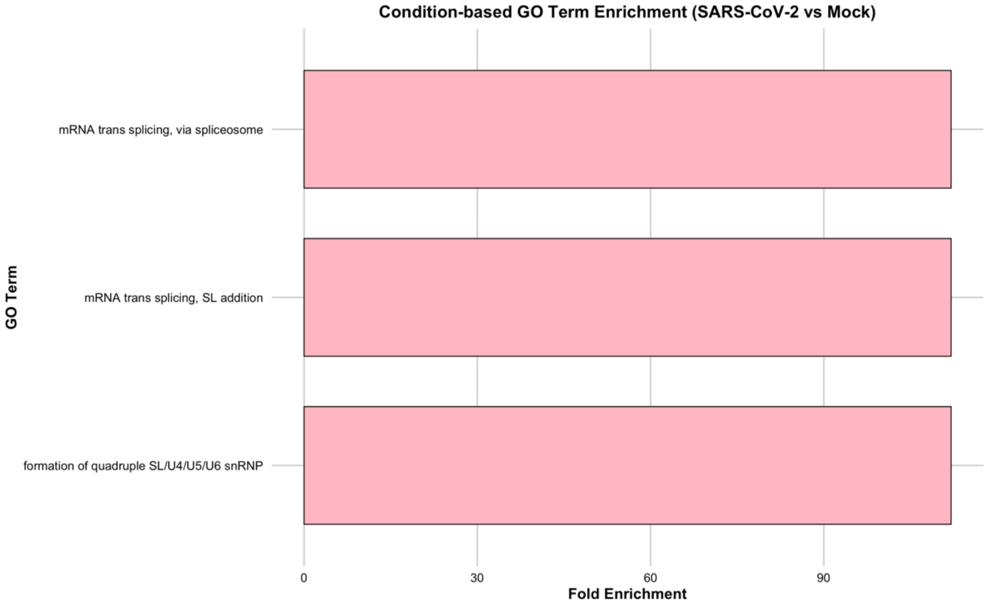

Fig-7

Condition-based GO Term Enrichment (SARS-CoV-2 vs Mock): This plot (Fig-7) displays the top GO terms enriched in SARS-CoV-2-infected cells compared to mock-treated cells. Since this is miRNA GO enrichment, which has very limited annotation compared to regular RNA-seq GO enrichment, we are seeing fewer results in the GO enrichment analysis.

Conclusion: This analysis successfully identified differentially expressed genes (DEGs) between SARS-CoV-2-infected and control (Mock) samples as well as across time points (24H vs. 72H). Through quality control, adapter trimming, and mapping steps, we ensured high-quality data for downstream analysis. Differential expression analysis revealed key changes in gene expression, with distinct patterns of upregulation and downregulation in response to infection and over time. The GO enrichment analysis provided insights into the biological processes impacted by SARS-CoV-2 infection and temporal progression, though limited by the sparse miRNA annotation available.